Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit

Song Mei[*], Theodor Misiakiewicz[†], Andrea Montanari[‡]

February 19, 2019

Abstract

We consider learning two layer neural networks using stochastic gradient descent. The mean-field description of this learning dynamics approximates the evolution of the network weights by an evolution in the space of probability distributions in $\mathbb{R}^D$ (where D is the number of parameters associated to each neuron). This evolution can be defined through a partial differential equation or, equivalently, as the gradient flow in the Wasserstein space of probability distributions. Earlier work shows that (under some regularity assumptions), the mean field description is accurate as soon as the number of hidden units is much larger than the dimension D. In this paper we establish stronger and more general approximation guarantees. First of all, we show that the number of hidden units only needs to be larger than a quantity dependent on the regularity properties of the data, and independent of the dimensions. Next, we generalize this analysis to the case of unbounded activation functions, which was not covered by earlier bounds. We extend our results to noisy stochastic gradient descent.

Finally, we show that kernel ridge regression can be recovered as a special limit of the mean field analysis.

Contents

[*] Institute for Computational and Mathematical Engineering, Stanford University
[†] Department of Statistics, Stanford University
[‡] Department of Electrical Engineering and Department of Statistics, Stanford University

1 Introduction

Multi-layer neural networks, and in particular multi-layer perceptrons, present a number of remarkable features. They are effectively trained using stochastic-gradient descent (SGD) [LBBH98]; their behavior is fairly insensitive to the number of hidden units or to the input dimensions [SHK+14]; their number of parameters is often larger than the number of samples.

In this paper consider simple neural networks with one layer of N hidden units:

$$\hat{f}_N(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_i)\,,\qquad \sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_i) = a_i\sigma(\boldsymbol{x};\boldsymbol{w}_i)\,, \tag{1}$$

Here $\boldsymbol{x}\in\mathbb{R}^d$ is a feature vector, $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_N)$ comprises the network parameters, $\boldsymbol{\theta}_i=(a_i,\boldsymbol{w}_i)\in\mathbb{R}^D$, and $\sigma:\mathbb{R}^d\times\mathbb{R}^{D-1}\to\mathbb{R}$ is a bounded activation function. The most classical example is $\sigma(\boldsymbol{x};\boldsymbol{w})=\sigma(\langle\boldsymbol{w},\boldsymbol{x}\rangle)$, where $\sigma:\mathbb{R}\to\mathbb{R}$ is a scalar function (and of course $D=d+1$), but our theory covers a broader set of examples. We assume to be given data $(y_i, \boldsymbol{x}_i) \sim \mathbb{P}$, with $\mathbb{P} \in \mathscr{P}(\mathbb{R} \times \mathbb{R}^d)$ a probability distribution over $\mathbb{R} \times \mathbb{R}^d$, and attempt at minimizing the square loss risk:

$$R_N(\boldsymbol{\theta}) = \mathbb{E}\big\{(y - \hat{f}_N(\boldsymbol{x}; \boldsymbol{\theta}))^2\big\}. \tag{2}$$

The risk function R_N can be either understood as population risk or empirical risk, depending on viewing $\mathbb{P}$ as a population distribution or assuming $\mathbb{P} = n^{-1} \sum_{k=1}^n \delta_{(y_k, \boldsymbol{x}_k)}$ is supported on n data points. If R_N is understood as the population risk, we can rewrite

$$R_N(\boldsymbol{\theta}) = R_{\text{Bayes}} + \mathbb{E}\big\{(f(\boldsymbol{x}) - \hat{f}_N(\boldsymbol{x}; \boldsymbol{\theta}))^2\big\}\,, \tag{3}$$

where $f(\boldsymbol{x}) = \mathbb{E}\{y|\boldsymbol{x}\}$ and R_{Bayes} is the Bayes error.

Classical theory of universal approximation provides useful insights into the way two-layers networks capture arbitrary input-output relations [Cyb89, Bar93]. In particular, Barron's theorem [Bar93] guarantees

$$\inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) \le R_{\text{Bayes}} + \frac{1}{N}\left(2r \int \|\boldsymbol{\omega}\|_2 |F(\boldsymbol{\omega})| \mathrm{d}\boldsymbol{\omega}\right)^2 \,, \tag{4}$$

where F is the Fourier transform of f, and r is the supremum of $\|\boldsymbol{x}\|_2$ in the support of $\mathbb{P}$. This result is remarkable in that the minimum number of neurons needed to achieve a certain accuracy depends only on intrinsic regularity properties of f and not on the dimension d. The proof of this and similar results shows that it is more insightful to think of the representation (1) in terms of the empirical distribution of the neurons $\hat{\rho}^{(N)} \equiv N^{-1} \sum_{i \le N} \delta_{\boldsymbol{\theta}_i}$. With a slight abuse of notation, we have $\hat{f}_N(\boldsymbol{x}; \boldsymbol{\theta}) = \hat{f}(\boldsymbol{x}; \hat{\rho}^{(N)})$, where, for a general distribution $\rho \in \mathscr{P}(\mathbb{R}^D)$, we define

$$\hat{f}(\boldsymbol{x}; \rho) = \int \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})\, \rho(\mathrm{d}\boldsymbol{\theta})\,. \tag{5}$$

The universal approximation property is then related to the fact that an arbitrary distribution ρ can be approximated by one supported on N points[1].

Approximation theory provides some insight into the peculiar properties of neural networks. Small population risk is achieved by many networks, since what matters is the distribution ρ, not the parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$. The behavior is insensitive to the number of neurons N, as long as this is large enough for $\hat{\rho}^{(N)}$ to approximate ρ. Finally, the bound (4) is dimension-free.

Of course these insights concern ideal representations, and not necessarily the networks generated by SGD. Recently, an analysis of SGD dynamics has been developed that connects naturally to the theory of universal approximation [MMN18, SS18, RVE18, CB18b]. The main object of study is the empirical distribution $\hat{\rho}_k^{(N)}$ after k SGD steps. For large N, small step size ε and setting $k = t/\varepsilon$, $\hat{\rho}_k^{(N)}$ turns out to be well approximated by a probability distribution $\rho_t \in \mathscr{P}(\mathbb{R}^D)$. The latter evolves according to the following partial differential equation

$$\partial_t \rho_t = 2\xi(t) \nabla_{\boldsymbol{\theta}} \cdot \big(\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)\big)\,, \quad \Psi(\boldsymbol{\theta}; \rho_t) \equiv V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, \rho_t(\mathrm{d}\tilde{\boldsymbol{\theta}})\,, \tag{DD}$$

$$V(\boldsymbol{\theta}) = -\mathbb{E}\{y \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})\}\,, \quad U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_1) \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_2)\}\,. \tag{6}$$

(Here $\xi(t)$ is a function that gauges the evolution of step size and will be defined below. In fact, there is little loss to the following discussion in setting $\xi(t) = 1$.) We will refer to this as the *mean field* description, or *distributional dynamics*. This description has the advantage of being explicitly independent of the number of hidden units N and hence accounts for one of the empirical findings described above (the insensitivity to the number of neurons). Further, it allows to focus on some key elements of the dynamics (global convergence, typical behavior) neglecting others (local minima, statistical noise).

Several papers used this approach over the last year to analyze learning in two-layers networks: this work will be succinctly reviewed in Section 2.

Of course, a crucial question needs to be answered for this approach to be meaningful: In what regime is the distributional dynamics a good approximation to SGD? Quantitative approximation guarantees were

[1] Of course, here we are hiding some important technical issues.

established in [MMN18], under certain regularity conditions on the data distribution $\mathbb{P}$, and for activation functions $\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})$ bounded. Under these conditions, and for time $t \in [0,T]$ bounded, [MMN18] proves that the distributional dynamics solution ρ_t approximates well the actual empirical distribution $\hat{\rho}^{(N)}_{k=t/\varepsilon}$, when the number of neurons is much larger than the problem dimensions $N \gg D$.

The results of [MMN18] present several limitations, that we overcome in the present paper. We briefly summarize our contributions.

Dimension-free approximation. As mentioned above, both classical approximation theory and the mean-field analysis of SGD approximate a certain target distribution ρ by the empirical distributions of the network parameters $\hat{\rho}^{(N)}$. However, while the approximation bound (4) is dimension-free, the approximation guarantees of [MMN18] are explicitly dimension-dependent. Even for very smooth functions $f(\boldsymbol{x})$, and well behaved data distributions, the results of [MMN18] require $N \gg D$.

Here we prove a new bound that is dimension independent and therefore more natural. The proof follows a coupling argument which is different and more powerful than the one of [MMN18]. A key improvement consists in isolating different error terms, and developing a more delicate concentration-of-measure argument which controls the dependence of the error on N.

Let us emphasize that capturing the correct dimension-dependence is an important test of the mean-field theory, and it is crucial in order to compare neural networks to other learning techniques (see Section 4).

Unbounded activations. The approximation guarantee of [MMN18] only applies to activation functions $\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_i)$ that are bounded. This excludes the important case of unbounded second-layer coefficients as in Eq. (1). We extend our analysis to that case. This requires to develop an *a priori* bound on the growth of the coefficients a_i. As in the previous point, our approximation guarantee is dimension-free.

Noisy SGD. Finally, in some cases it is useful to inject noise into SGD. From a practical perspective this can help avoiding local minima. From an analytical perspective, it corresponds to a modified PDE, which contains an additional Laplacian term $\Delta_{\boldsymbol{\theta}}\rho_t$. This PDE has smoother solutions ρ_t that are supported everywhere and converge globally to a unique fixed point [MMN18].

In this setting, we prove a dimension-free approximation guarantee for the case of bounded activations. We also obtain a guarantee for noisy SGD unbounded activations, but the latter is not dimension-free.

Kernel limit. We analyze the PDE (DD) in a specific short-time limit and show that it is well approximated by a linearized dynamics. This dynamics can be thought as fitting a kernel ridge regression[2] model with respect to a kernel corresponding to the initial weight distribution ρ_0. We thus recover –from a different viewpoint– a connection with kernel methods that has been investigated in several recent papers [JGH18, DZPS18, DLL+18, AZLS18]. Beyond the short time scale, the dynamics is analogous to kernel boosting dynamics with a time-varying data-dependent kernel (a point that already appears in [RVE18]).

Mean-field theory allowed to prove global convergence guarantees for SGD in two-layers neural networks [MMN18, CB18b]. Unfortunately, these results do not provide (in general) useful bounds on the network size N. We believe that the results in this paper are a required step in that direction.

The rest of this paper is organized as follows. The next section overviews related work, focusing in particular on the distributional dynamics (DD), its variants and applications. In Section 3 we present formal statements of our results. Section 4 develops the connection with kernel methods. Proofs are mostly deferred to the appendices.

2 Related work

As mentioned above, classical approximation theory already uses (either implicitly or explicitly) the idea of lifting the class of N-neurons neural networks, cf. Eq. (1), to the infinite-dimensional space (5) parametrized

[2] 'Kernel ridge regression' and 'kernel regression' are used with somewhat different meanings in the literature. Kernel ridge regression uses global information and can be defined as ridge regression in reproducing kernel Hilbert space (RKHS), while kernel regression uses local averages. See Remark H.1 for a definition.

by probability distributions ρ, see e.g. [Cyb89, Bar93, Bar98, AB09]. This idea was exploited algorithmically, e.g. in [BRV$^+$06, NS17].

Only very recently (stochastic) gradient descent was proved to converge (for large enough number of neurons) to the infinite-dimensional evolution (DD) [MMN18, RVE18, SS18, CB18b]. In particular, [MMN18] proves quantitative bounds to approximate SGD by the mean-field dynamics. Our work is mainly motivated by the objective to obtain a better scaling with dimension and to allow for unbounded second-layer coefficients.

The mean-field description was exploited in several papers to establish global convergence results. In [MMN18] global convergence was proved in special examples, and in a general setting for noisy SGD. The papers [RVE18, CB18b] studied global convergence by exploiting the homogeneity properties of Eq. (1). In particular, [CB18b] proves a general global convergence result. For initial conditions ρ_0 with full support, the PDE (DD) converges to a global minimum provided activations are homogeneous in the parameters. Notice that the presence of unbounded second layer coefficients is crucial in order to achieve homogeneity. Unfortunately, the results of [CB18b] do not provide quantitative approximation bounds relating the PDE (DD) to finite-N SGD. The present paper fills this gap by establishing approximation bounds that apply to the setting of [CB18b].

A different optimization algorithm was studied in [WLLM18] using the mean-field description. The algorithm resamples a positive fraction of the neurons uniformly at random at a constant rate. This allows the authors to establish a global convergence result (under certain assumed smoothness properties on the PDE solution). Again, this paper does not provide quantitative bounds on the difference between PDE and finite-N SGD. While our theorems do not cover the algorithm of [WLLM18], we believe that their algorithm could be analyzed using the approach developed here. Exponentially fast convergence to a global optimum was proven in [JMM19] for certain radial-basis-function networks, using again the mean-field approach. While the setting of [JMM19] is somewhat different (weights are constrained to a convex compact domain), the technique presented here could be applicable to that problem as well.

Finally, a recent stream of works [JGH18, GJS$^+$19, DZPS18, DLL$^+$18, AZLS18] argues that, as $N \to \infty$ two-layers networks are actually performing a type of kernel ridge regression. As shown in [CB18a], this phenomenon is not limited to neural network, but generic for a broad class of models. As expected, the kernel regime can indeed be recovered as a special limit of the mean-field dynamics (DD), cf. Section 4. Let us emphasize that here we focus on the population rather than the empirical risk.

A discussion of the difference between the kernel and mean-field regimes was recently presented in [DL19]. However, [DL19] argues that the difference between kernel and mean-field behaviors is due to different initializations of the coefficients a_i's. We show instead that, for a suitable scaling of the initialization, kernel and mean field regimes appear at different time scales. Namely, the kernel behavior arises at the beginning of the dynamics, and mean field characterizes longer time scales. It is also worth mentioning that the connection between mean field dynamics and kernel boosting with a time-varying data-dependent kernel was already present (somewhat implicitly) in [RVE18].

3 Dimension-free mean field approximation

3.1 General results

As mentioned above, we assume to be given data $\{(y_k, \boldsymbol{x}_k)\}_{k\ge 1} \sim_{i.i.d.} \mathbb{P} \in \mathscr{P}(\mathbb{R} \times \mathbb{R}^d)$, and we run SGD with step size s_k:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + 2s_k(y_k - \hat{f}_N(\boldsymbol{x}_k; \boldsymbol{\theta}^k))\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}_k; \boldsymbol{\theta}_i^k). \tag{SGD}$$

We will work under a one-pass model, that is, each data point is visited once.

We also consider a noisy version of SGD, with a regularization term:

$$\boldsymbol{\theta}_i^{k+1} = (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k + 2s_k(y_k - \hat{f}_N(\boldsymbol{x}_k; \boldsymbol{\theta}^k))\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}_k; \boldsymbol{\theta}_i^k) + \sqrt{2s_k\tau/D}\, \boldsymbol{g}_i^k\,, \tag{noisy-SGD}$$

where $\boldsymbol{g}_i^k \sim \mathcal{N}(0, \boldsymbol{I}_D)$. The noiseless version is recovered by setting $\tau = 0$ and $\lambda = 0$. The step size is chosen according to : $s_k = \varepsilon\xi(k\varepsilon)$, for a positive function $\xi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{>0}$.

The infinite-dimensional evolution corresponding to noisy SGD is given by

$$\partial_t \rho_t = 2\xi(t)\nabla_{\boldsymbol{\theta}} \cdot \big(\rho_t(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)\big) + 2\xi(t)\tau D^{-1}\Delta_{\boldsymbol{\theta}}\rho_t\,, \tag{diffusion-DD}$$

$$\Psi_\lambda(\boldsymbol{\theta};\rho) = \Psi(\boldsymbol{\theta};\rho) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\,. \tag{7}$$

The function Ψ is defined as in (DD). At this point it is important to note that the PDE (DD) has to be interpreted in weak sense, while, for $\tau > 0$, Eq. (diffusion-DD) has strong solutions i.e. solutions $\rho : (t,\boldsymbol{\theta}) \mapsto \rho_t(\boldsymbol{\theta})$ that are $C^{1,2}(\mathbb{R}\times\mathbb{R}^D)$ (once continuous differentiable in time and twice in space, see [MMN18] and Appendix F).

It is useful to lift the population risk in the space of distributions $\rho \in \mathscr{P}(\mathbb{R}^D)$

$$R(\rho) = \mathbb{E}(y^2) + 2\int V(\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}')\,. \tag{8}$$

We also note that, given the structure of the activation function in Eq. (1), for $\boldsymbol{\theta} = (a,\boldsymbol{w})$, $\boldsymbol{\theta}_i = (a_i,\boldsymbol{w}_i)$, we can write $V(\boldsymbol{\theta}) = a\,v(\boldsymbol{w})$, $U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = a_1a_2\,u(\boldsymbol{w}_1,\boldsymbol{w}_2)$, where $v(\boldsymbol{w}) = -\mathbb{E}\{y\sigma(\boldsymbol{x};\boldsymbol{w})\}$ and $u(\boldsymbol{w}_1,\boldsymbol{w}_2) = \mathbb{E}\{\sigma(\boldsymbol{x};\boldsymbol{w}_1)\sigma(\boldsymbol{x};\boldsymbol{w}_2)\}$.

In order to establish a non-asymptotic guarantee, we will make the following assumptions:

A1. $t \mapsto \xi(t)$ is bounded Lipschitz: $\|\xi\|_\infty, \|\xi\|_{\mathrm{Lip}} \le K_1$.

A2. The activation function $\sigma : \mathbb{R}^d \times \mathbb{R}^{D-1} \to \mathbb{R}$ and the response variables are bounded: $\|\sigma\|_\infty, |y_k| \le K_2$. Furthermore, its gradient $\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x};\boldsymbol{w})$ is K_2-sub-Gaussian (when $\boldsymbol{x} \sim \mathbb{P}$).

A3. The functions $\boldsymbol{w} \mapsto v(\boldsymbol{w})$ and $(\boldsymbol{w}_1,\boldsymbol{w}_2) \mapsto u(\boldsymbol{w}_1,\boldsymbol{w}_2)$ are differentiable, with bounded and Lipschitz continuous gradient: $\|\nabla v(\boldsymbol{w})\|_2 \le K_3$, $\|\nabla u(\boldsymbol{w}_1,\boldsymbol{w}_2)\|_2 \le K_3$, $\|\nabla v(\boldsymbol{w}) - \nabla v(\boldsymbol{w}')\|_2 \le K_3\|\boldsymbol{w}-\boldsymbol{w}'\|_2$, $\|\nabla u(\boldsymbol{w}_1,\boldsymbol{w}_2) - \nabla u(\boldsymbol{w}_1',\boldsymbol{w}_2')\|_2 \le K_3\|(\boldsymbol{w}_1,\boldsymbol{w}_2)-(\boldsymbol{w}_1',\boldsymbol{w}_2')\|_2$.

A4. The initial condition $\rho_0 \in \mathscr{P}(\mathbb{R}^D)$ is supported on $|a_i| \le K_4$ for a constant K_4.

We will consider two different cases for the SGD dynamics:

General coefficients. We initialize the parameters $\boldsymbol{\theta}_i^0 = (a_i^0,\boldsymbol{w}_i^0)$ as $(\boldsymbol{\theta}_i^0)_{i\le N} \sim_{iid} \rho_0$. Both the a_i^0 and $\boldsymbol{w}_i^0$ are updated during the dynamics.

Fixed coefficients. We use the same initialization as described above, but the coefficients a_i are not updated by SGD. The corresponding PDE is given by Eq. (DD) (or (diffusion-DD)), except that the space derivatives are to be interpreted only with respect to $\boldsymbol{w}$, i.e. replace $\nabla_{\boldsymbol{\theta}}$ by $(0,\nabla_{\boldsymbol{w}})$, and $\Delta_{\boldsymbol{\theta}}$ by $\Delta_{\boldsymbol{w}}$.

While the second setting is less relevant in practice, it is at least as interesting from a theoretical point of view, and some of our guarantees are stronger in that case.

Theorem 1. *Assume that conditions* A1-A4 *hold, and let $T \ge 1$. Let $(\rho_t)_{t\ge 0}$ be the solution of the PDE (DD) with initialization ρ_0, and let $(\boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ to be the trajectory of SGD (SGD) with initialization $\boldsymbol{\theta}_i^0 \sim \rho_0$ independently.*

(A) Consider noiseless SGD with fixed coefficients. *Then there exists a constant K (depending uniquely on the constants K_i of assumptions* A1-A4*) such that*

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\Big|R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon})\Big| \le Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log N} + z] + Ke^{KT}[\sqrt{D+\log(N)} + z]\sqrt{\varepsilon} \tag{9}$$

with probability at least $1 - e^{-z^2}$.

(B) Consider noiseless SGD with general coefficients. *Then there exists constants K and K_0 (depending uniquely on the constants K_i of assumptions* A1-A4*) such that if $\varepsilon \le 1/[K_0(D+\log N + z^2)e^{K_0T^3}]$, we have*

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\Big|R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon})\Big| \le Ke^{KT^3}\frac{1}{\sqrt{N}}[\sqrt{\log N} + z] + Ke^{KT^3}[\sqrt{D+\log N} + z]\sqrt{\varepsilon} \tag{10}$$

with probability at least $1 - e^{-z^2}$.

Remark 3.1. As anticipated in the introduction, provided $T, K = O(1)$, the error terms in Eqs. (9), (10), are small as soon as $N \gg 1$. In other words, the minimum number of neurons needed for the mean-field approximation to be accurate is independent of the dimension D, and only depends on intrinsic features of the activation and data distribution.

On the other hand, the dimension D appears explicitly in conjunction with the step size ε. We need $\varepsilon \ll 1/D$ in order for mean field to be accurate. This is the same trade-off between step size and dimension that was already achieved in [MMN18].

We next consider noisy SGD, cf. Eq. (noisy-SGD), and the corresponding PDE in Eq. (diffusion-DD). We need to make additional assumptions on the initialization in this case.

A5. The initial condition ρ_0 is such that, for $\boldsymbol{\theta}_i^0 = (a_i^0, \boldsymbol{w}_i^0) \sim \rho_0$, we have that $\boldsymbol{w}_i^0$ is K_5^2/D-sub-Gaussian.

A6. $V \in C^4(\mathbb{R}^D)$, $U \in C^4(\mathbb{R}^D \times \mathbb{R}^D)$, and $\nabla_1^k u(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is uniformly bounded for $0 \leq k \leq 4$.

Remark 3.2. The last condition ensures the existence of strong solutions for Eq. (diffusion-DD). The existence and uniqueness of solution of the PDE (DD) and the PDE (diffusion-DD) are discussed in Appendix F.

Theorem 2. *Assume that conditions* A1 - A6 *hold. Let $(\rho_t)_{t\geq 0}$ be the solution of the PDE (diffusion-DD) with initialization ρ_0, and let $(\boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ to be the trajectory of noisy SGD (noisy-SGD) with initialization $\boldsymbol{\theta}_i^0 \sim \rho_0$ independently. Finally assume that $\lambda \leq K_6$, $\tau \leq K_6$, $T \geq 1$.*

(A) Consider noisy SGD with fixed coefficients*. Then there exists a constant K (depending uniquely on the constants K_i of assumptions* A1-A5 *and K_6) such that*

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \Big| R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon}) \Big| \leq K e^{KT} \frac{1}{\sqrt{N}} [\sqrt{\log N} + z] + K e^{KT} [\sqrt{D + \log(N/\varepsilon)} + z] \sqrt{\varepsilon} \tag{11}$$

with probability at least $1 - e^{-z^2}$.

(B) Consider noisy SGD with general coefficients*. Then there exists a constant K (depending uniquely on the constants K_i of assumptions* A1-A5 *and K_6) such that*

$$\begin{aligned}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \Big| R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon}) \Big| \leq & K e^{e^{KT[\sqrt{\log N} + z^2]}} [\sqrt{D \log N} + \log^{3/2}(NT) + z^5]/\sqrt{N} \\ & + K e^{e^{KT[\sqrt{\log N} + z^2]}} [\sqrt{D} \log(N(T/\varepsilon \vee 1)) + \log^{3/2} N + z^6] \sqrt{\varepsilon}\end{aligned} \tag{12}$$

with probability at least $1 - e^{-z^2}$.

Remark 3.3. Unlike the other results in this paper, part (B) of Theorem 2 does not establish a dimension-free bound. Further, while previous bounds allow to control the approximation error for any $T = o(\log N)$, Theorem 2.(B) requires $T = o(\log\log N)$. The main difficulty in part (B) is to control the growth of the coefficients a_i. This is more challenging than in the noiseless case, since we cannot give a deterministic bound on $|a_i|$.

Despite these drawbacks, Theorem 2 (B) is the first quantitative bound approximating noisy SGD by the distributional dynamics, for the case of unbounded coefficients. It implies that the mean field theory is accurate when $N \gg D$.

3.2 Example: Centered anisotropic Gaussians

To illustrate an application of the theorems, we consider the problem of classifying two Gaussians with the same mean and different covariance. This example was studied in [MMN18], but we restate it here for the reader's convenience.

Consider the joint distribution of data $(y, \boldsymbol{x})$ given by the following:

With probability 1/2: $y = +1$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{\Sigma}_+)$, With probability 1/2: $y = -1$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{\Sigma}_-)$,

where $\boldsymbol{\Sigma}_\pm = \boldsymbol{U}^\mathsf{T} \mathrm{diag}((1 \pm \Delta)^2 \boldsymbol{I}_{s_0}, \boldsymbol{I}_{d-s_0})\boldsymbol{U}$ for $\boldsymbol{U}$ to be an unknown orthogonal matrix. In other words, there exists a subspace $\mathcal{V}$ of dimension s_0, such that the projection of $\boldsymbol{x}$ on the subspace $\mathcal{V}$ is distributed according to an isotropic Gaussian with variance $\tau_+^2 = (1+\Delta)^2$ (if $y = +1$) or $\tau_-^2 = (1 - \Delta^2)$ (if $y = -1$). The projection orthogonal to $\mathcal{V}$ has instead the same variance in the two classes.

We choose an activation function without offset or output weights, namely $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$. While qualitatively similar results are obtained for other choices of σ, we will use a simple piecewise linear function (truncated ReLU) as a running example: take $t_1 < t_2$,

$$\sigma(t) = \begin{cases} s_1, & \text{if } t \le t_1, \\ s_2, & \text{if } t \ge t_2, \\ s_1 + (s_2 - s_1)(t - t_1)/(t_2 - t_1), & \text{if } t \in (t_1, t_2). \end{cases}$$

We introduce a class of good uninformative initializations $\mathscr{P}_{\rm good} \subseteq \mathscr{P}(\mathbb{R}_{\ge 0})$ for which convergence to the optimum takes place. For $\bar{\rho} \in \mathscr{P}(\mathbb{R}_{\ge 0})$, we let

$$\overline{R}_d(\bar{\rho}) \equiv R(\bar{\rho} \times \mathrm{Unif}(\mathbb{S}^{d-1})), \qquad \overline{R}_\infty(\bar{\rho}) \equiv \lim_{d \to \infty} \overline{R}_d(\bar{\rho}).$$

We say that $\bar{\rho} \in \mathscr{P}_{\rm good}$ if: (i) $\bar{\rho}$ is absolutely continuous with respect to Lebesgue measure, with bounded density; (ii) $\overline{R}_\infty(\bar{\rho}) < 1$.

The following theorem is an improvement of [MMN18, Theorem 2] using Theorem 1, whose proof is just by replacing the last step of proof of [MMN18, Theorem 2] using the new bounds developed in 1 (A).

Theorem 3. *For any $\eta, \Delta, \delta > 0$, and $\bar{\rho}_0 \in \mathscr{P}_{\rm good}$, there exists $d_0 = d_0(\eta, \bar{\rho}_0, \Delta, \gamma)$, $T = T(\eta, \bar{\rho}_0, \Delta, \gamma)$, and $C_0 = C_0(\eta, \bar{\rho}_0, \Delta, \delta, \gamma)$, such that the following holds for the problem of classifying anisotropic Gaussians with $s_0 = \gamma d$, $\gamma \in (0, 1)$ fixed. For any dimension parameters $s_0 = \gamma d \ge d_0$, number of neurons $N \ge C_0$, consider SGD initialized with initialization $(\boldsymbol{w}_i^0)_{i \le N} \sim_{iid} \bar{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \le 1/(C_0 d)$. Then we have $R_N(\boldsymbol{\theta}^k) \le \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times d}} R_N(\boldsymbol{\theta}) + \eta$ for any $k \in [T/\varepsilon, 10T/\varepsilon]$ with probability at least $1 - \delta$.*

Comparing to [MMN18, Theorem 2], here we require $N = O(1)$ neuron rather than previously $N = O(d)$ neurons. The number of data used $k = O(d)$ is still on the optimal order.

4 Connection with kernel methods

As discussed above, mean-field theory captures the SGD dynamics of two layers neural networks when the number of hidden units N is large. Several recent papers studied a different description, that approximates the neural network as performing a form of kernel ridge regression [JGH18, DZPS18]. This behavior also arises for large N: we will refer to this as to the 'kernel regime', or 'kernel limit'. As shown in [CB18a] the existence of a kernel regime is not specific to neural networks but it is a generic feature of overparameterized models, under certain differentiability assumptions.

4.1 A coupled dynamics

We will focus on noiseless gradient flow, and assume $y = f(\boldsymbol{x})$ (a general joint distribution over $(y, \boldsymbol{x})$ is recovered by setting $f(\boldsymbol{x}) = \mathbb{E}\{y|\boldsymbol{x}\}$). As in [CB18a], we modify the model (1) by introducing an additional scale parameter α:

$$\hat{f}_{\alpha,N}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\alpha}{N} \sum_{i=1}^N \boldsymbol{\sigma}_\star(\boldsymbol{x}; \boldsymbol{\theta}_i)\,, \tag{13}$$

In the case of general coefficients a_i, this amounts to rescaling the coefficients $a_i \to a_i/\alpha$. Equivalently, this corresponds to a different initialization for the a_i's (larger by a factor α).

We first note that the theorems of the previous section obviously hold for the modified dynamics, with the PDE (DD) generalized to

$$\partial_t \rho_t = \alpha \nabla_{\boldsymbol{\theta}} \cdot \big(\rho_t \nabla_{\boldsymbol{\theta}} \Psi_\alpha(\boldsymbol{\theta}, \rho_t)\big)\,, \tag{14}$$

$$\Psi_\alpha(\boldsymbol{\theta}, \rho) = \mathbb{E}_{\boldsymbol{x}}\big\{\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})\,(\hat{f}_\alpha(\boldsymbol{x};\rho) - f(\boldsymbol{x}))\big\} = V(\boldsymbol{\theta}) + \alpha \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')\,, \tag{15}$$

where $\hat{f}_\alpha(\boldsymbol{x};\rho) = \alpha \int \sigma_\star(\boldsymbol{x};\boldsymbol{\theta})\,\rho(\mathrm{d}\boldsymbol{\theta})$. It is convenient to redefine time units by letting $\rho_t^\alpha \equiv \rho_{\alpha^{-2}t}$. This satisfies the *rescaled distributional dynamics*

$$\partial_t \rho_t^\alpha = \frac{1}{\alpha} \nabla_{\boldsymbol{\theta}} \cdot \big(\rho_t^\alpha \nabla_{\boldsymbol{\theta}} \Psi_\alpha(\boldsymbol{\theta}, \rho_t^\alpha)\big)\,. \tag{Rescaled-DD}$$

We next consider the residuals $u_t^\alpha(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x};\rho_t^\alpha)$ which we view as an element of $L^2 = L^2(\mathbb{R}^d;\mathbb{P})$. As first shown in [RVE18], this satisfies the following *mean field residual dynamics* (for further background, we refer to Appendix H):

$$\partial_t u_t^\alpha(\boldsymbol{x}) = -\int \mathcal{H}_{\rho_t^\alpha}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) u_t^\alpha(\tilde{\boldsymbol{x}})\,\mathbb{P}(\mathrm{d}\tilde{\boldsymbol{x}}) \equiv -(\mathcal{H}_{\rho_t^\alpha} u_t^\alpha)(\boldsymbol{x})\,, \tag{RD}$$

$$\mathcal{H}_\rho(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \equiv \int \langle \nabla_{\boldsymbol{\theta}} \sigma_\star(\boldsymbol{x};\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} \sigma_\star(\tilde{\boldsymbol{x}};\boldsymbol{\theta})\rangle\,\rho(\mathrm{d}\boldsymbol{\theta})\,. \tag{16}$$

Coupling the dynamics (Rescaled-DD) and (RD) suggests the following point of description. Gradient flow dynamics of two-layers neural network is a kernel boosting dynamics with a time-varying kernel. The scaling parameter α controls the speed that the kernel evolves.

The mean field residual dynamics (RD) implies that

$$\partial_t R_\alpha(\rho_t^\alpha) = \partial_t(\|u_t^\alpha\|_{L^2}^2) = -2\langle u_t^\alpha, \mathcal{H}_{\rho_t^\alpha} u_t^\alpha\rangle_{L^2},$$

so that the risk will be non-increasing along the gradient flow dynamics. However, since the kernel $\mathcal{H}_{\rho_t^\alpha}$ is not fixed, it is hard to analyze when the risk converges to 0 (see [MMN18, Theorem 4], [CB18b, Theorem 3.3 and 3.5] for general convergence results).

4.2 Kernel limit of residual dynamics

The kernel regime corresponds to large α and allows for a simpler treatment of the dynamics. Heuristically, the reason for such a simplification is that the time derivative of ρ_t^α is of order $1/\alpha$, cf. (Rescaled-DD). We are therefore tempted to replace $\mathcal{H}_{\rho_t^\alpha}$ in Eq. (RD) by $\mathcal{H}_{\rho_0}$. Formally, we define the following *linearized residual dynamics*

$$\partial_t u_t^* = -\mathcal{H}_{\rho_0} u_t^*\,. \tag{17}$$

We can also define the corresponding predictors by $f_t^* = f - u_t^*$. The operator $\mathcal{H}_{\rho_0}$ is bounded and standard semigroup theory [Eva09] implies the following.

Lemma 1. *We have* $\lim_{t\to\infty} u_t^* = u_\infty^* = \boldsymbol{P}_{\rho_0} u_0^*$*, where* $\boldsymbol{P}_{\rho_0}$ *is the orthogonal projector onto the null space of* $\mathcal{H}_{\rho_0}$*. In particular, if the null space of* $\mathcal{H}_{\rho_0}$ *is empty, then* $\lim_{t\to\infty} \|u_t^*\|_{L^2} \to 0$*. Correspondingly* $f_\infty^* = \boldsymbol{P}_{\rho_0}^\perp f + \boldsymbol{P}_{\rho_0} f_0^*$ *(where* $\boldsymbol{P}_{\rho_0}^\perp = \boldsymbol{I} - \boldsymbol{P}_{\rho_0}$*).*

The next theorem shows that the above intuition is correct. For $\alpha \geq t^2 D^{3/2}$, the linearized dynamics is a good approximation to the mean field dynamics. Below, we denote the population risk by $R_\alpha(\rho)$: $R_\alpha(\rho) \equiv \mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_\alpha(\boldsymbol{x};\rho))^2]$.

Theorem 4. *Let* u_t^α *and* u_t^* *be the residues in the mean-field dynamics (RD) and linearized dynamics (17), respectively. Let assumptions* A1, A3, A4 *hold, and additionally assume the following*

- $|y_i|$, $\|\sigma\|_\infty \leq K_2$*, and* $\boldsymbol{\theta} \mapsto \sigma_\star(\boldsymbol{x};\boldsymbol{\theta})$ *is differentiable.*

- $\|\nabla^3 u(\boldsymbol{w}, \boldsymbol{w}')\|_{\rm op}, \|\nabla^4 u(\boldsymbol{w}, \boldsymbol{w}')\|_{\rm op} \le \kappa$.
- $R_\alpha(\rho_0) \le B$.

Then there exists a constant K depending on $\{K_i\}_{i=1}^4$, such that

(A) For SGD with fixed coefficients, we have

$$\|u_t^\alpha - u_t^*\|_{L_2} \le K\kappa^{1/2} B\frac{D^{3/2}t^2}{\alpha}\,, \tag{18}$$

$$R_\alpha(\rho_t^\alpha) \le \Big(\|u_t^*\|_{L^2} + K\kappa^{1/2}B\frac{D^{3/2}t^2}{\alpha}\Big)^2\,. \tag{19}$$

(B) For SGD with general coefficients, we have

$$\|u_t^\alpha - u_t^*\|_{L_2} \le K\kappa^{1/2}(1 + B^{1/2}t/\alpha)^3 B\frac{D^{3/2}t^2}{\alpha}\,, \tag{20}$$

$$R_\alpha(\rho_t^\alpha) \le \Big(\|u_t^*\|_{L^2} + K\kappa^{1/2}(1 + B^{1/2}t/\alpha)^3 B\frac{D^{3/2}t^2}{\alpha}\Big)^2\,. \tag{21}$$

(C) In particular, if under the law $(a, \boldsymbol{w}) \sim \rho_0$, a is independent of $\boldsymbol{w}$ and $|\mathbb{E}(a)| \le K_5/\alpha$. Then $B \le K$ is independent of α. If the null space of $\mathcal{H}_{\rho_0}$ is empty, then under both settings (fixed and variable coefficients)

$$\lim_{\alpha\to\infty} \sup_{t\in[0,T]} \|u_t^\alpha - u_t^*\|_{L_2} = 0, \tag{22}$$

$$\lim_{t\to\infty} \lim_{\alpha\to\infty} R_\alpha(\rho_t^\alpha) = 0\,. \tag{23}$$

Remark 4.1. Unlike in similar results in the literature, we focus here on the population risk rather than the empirical risk. The recent paper [CB18a] addresses both the overparametrized and the underparametrized regime. The latter result (namely [CB18a, Theorem 3.4]) is of course relevant for the population risk. However, while [CB18a] proves convergence to a local minimum, here we show that the population risk becomes close to 0.

Remark 4.2. As stated above, the linearized residual dynamics can be interpreted as performing kernel ridge regression with respect to the kernel $\mathcal{H}_{\rho_0}$, see e.g. [JGH18]. A way to clarify the connection is to consider the case in which $\mathbb{P} = n^{-1}\sum_{i\le n}\delta_{\boldsymbol{x}_i}$ is the empirical data distribution. In this case the linearized dynamics converges to

$$\lim_{t\to\infty} f_t^*(\boldsymbol{z}) = f_\infty^*(\boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{z})^{\mathsf{T}}\boldsymbol{H}^{-1}\boldsymbol{y}$$

where

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{z}) =& [\mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_1), \dots, \mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_n)]^{\mathsf{T}},\\ \boldsymbol{H} =& (\mathcal{H}_{\rho_0}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij=1}^n,\\ \boldsymbol{y} =& [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^{\mathsf{T}}.\end{aligned}$$

For the sake of completeness, we review the connection in Appendix H.7.

References

[AB09] Martin Anthony and Peter L Bartlett, *Neural network learning: Theoretical foundations*, cambridge university press, 2009.

[AZLS18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song, *A convergence theory for deep learning via overparameterization*, arXiv:1811.03962 (2018).

[Bar93] Andrew R Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information theory **39** (1993), no. 3, 930–945.

[Bar98] Peter L Bartlett, *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network*, IEEE transactions on Information Theory **44** (1998), no. 2, 525–536.

[BRV$^+$06] Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte, *Convex neural networks*, Advances in neural information processing systems, 2006, pp. 123–130.

[CB18a] Lenaic Chizat and Francis Bach, *A note on lazy training in supervised differentiable programming*, arXiv:1812.07956 (2018).

[CB18b] ———, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, arXiv:1805.09545 (2018).

[Cyb89] George Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems **2** (1989), no. 4, 303–314.

[DL19] Xialiang Dou and Tengyuan Liang, *Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits*, arXiv:1901.07114 (2019).

[DLL$^+$18] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, *Gradient descent finds global minima of deep neural networks*, arXiv preprint arXiv:1811.03804 (2018).

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, *Gradient descent provably optimizes over-parameterized neural networks*, arXiv:1810.02054 (2018).

[Eva09] Lawrence C. Evans, *Partial differential equations*, Springer, 2009.

[GJS$^+$19] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart, *Scaling description of generalization with number of parameters in deep learning*, arXiv:1901.01608 (2019).

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, arXiv:1806.07572 (2018).

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the fokker–planck equation*, SIAM journal on mathematical analysis **29** (1998), no. 1, 1–17.

[JMM19] Adel Javanmard, Marco Mondelli, and Andrea Montanari, *Analysis of a two-layer neural network via displacement convexity*, arXiv:1901.01375 (2019).

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE **86** (1998), no. 11, 2278–2324.

[LL18] Yuanzhi Li and Yingyu Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, Advances in Neural Information Processing Systems, 2018, pp. 8168–8177.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences (2018).

[NS17] Atsushi Nitanda and Taiji Suzuki, *Stochastic particle gradient descent for infinite ensembles*, arXiv:1712.05438 (2017).

[RVE18] Grant M Rotskoff and Eric Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, arXiv:1805.00915 (2018).

[San15] Filippo Santambrogio, *Optimal transport for applied mathematicians: Calculus of variations, pdes, and modeling*, vol. 87, Birkhäuser, 2015.

[SHK$^+$14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, *Dropout: a simple way to prevent neural networks from overfitting*, The Journal of Machine Learning Research **15** (2014), no. 1, 1929–1958.

[SS18] Justin Sirignano and Konstantinos Spiliopoulos, *Mean field analysis of neural networks*, arXiv:1805.01053 (2018).

[Szn91] Alain-Sol Sznitman, *Topics in propagation of chaos*, Ecole d'été de probabilités de Saint-Flour XIX—1989, Springer, 1991, pp. 165–251.

[WLLM18] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma, *On the margin theory of feedforward neural networks*, arXiv:1810.05369 (2018).

A Notations

- For future reference, we copy the key definitions from the main text:

$$
\begin{aligned}
R_N(\boldsymbol{\theta}) &= \mathbb{E}\{y^2\} + \frac{2}{N}\sum_{i=1}^N V(\boldsymbol{\theta}_i) + \frac{1}{N^2}\sum_{i,j=1}^N U(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j),\\
R(\rho) &= \mathbb{E}\{y^2\} + 2\int V(\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\rho(\mathrm{d}\boldsymbol{\theta}_1)\rho(\mathrm{d}\boldsymbol{\theta}_2),\\
V(\boldsymbol{\theta}) &= -\mathbb{E}\{y\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})\}, \qquad U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1)\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\},\\
\Psi(\boldsymbol{\theta};\rho) &= V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}'),\\
\Psi_\lambda(\boldsymbol{\theta};\rho) &= \Psi(\boldsymbol{\theta};\rho) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2,
\end{aligned}
$$

  where $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i\le N} \in \mathbb{R}^{D\times N}$ or $\boldsymbol{\theta} \in \mathbb{R}^D$ depending on the context. Further, we will denote for $\boldsymbol{\theta} = (a, \boldsymbol{w})$ and $\boldsymbol{\theta}' = (a', \boldsymbol{w}')$:

$$
V(\boldsymbol{\theta}) = av(\boldsymbol{w}), \qquad U(\boldsymbol{\theta},\boldsymbol{\theta}') = aa'u(\boldsymbol{w},\boldsymbol{w}').
$$

  In particular,

$$
\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = (v(\boldsymbol{w}), a\nabla_{\boldsymbol{w}} v(\boldsymbol{w})), \qquad \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta},\boldsymbol{\theta}') = (a'u(\boldsymbol{w},\boldsymbol{w}'), aa'\nabla_{\boldsymbol{w}} u(\boldsymbol{w},\boldsymbol{w}')).
$$

  In the case of fixed coefficients, without loss of generality, we will fix in the proof $a_i = 1$ for notational simplicity and freely denote $(\boldsymbol{\theta}_i)_{i=1}^N = (\boldsymbol{w}_i)_{i=1}^N$,

$$
\begin{aligned}
V(\boldsymbol{\theta}) &= v(\boldsymbol{w}), & U(\boldsymbol{\theta},\boldsymbol{\theta}') &= u(\boldsymbol{w},\boldsymbol{w}'),\\
\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{w}} v(\boldsymbol{w}), & \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta},\boldsymbol{\theta}') &= \nabla_{\boldsymbol{w}} u(\boldsymbol{w},\boldsymbol{w}').
\end{aligned}
$$

- $W_2(\cdot,\cdot)$ is the Wasserstein distance between probability measures

$$
W_2(\mu,\nu) = \Big(\inf\Big\{\int_{\mathbb{R}^D\times\mathbb{R}^D}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2):\ \gamma \text{ is a coupling of } \mu,\nu\Big\}\Big)^{1/2}.
$$

- For $N \in \mathbb{N}$, we will denote $[N] = \{1, 2, \dots, N\}$. With a little abuse of notation, for $s \in \mathbb{R}$, we will denote $[s] = \varepsilon\lfloor s/\varepsilon\rfloor$, with ε the time discretization parameter.

- K will denote a generic constant depending on K_i for $i = 1, 2, 3, 4, 5, 6$, where the K_i's are constants that will be specified from the context.

- In the proof and the statements of the theorems, we will only consider the leading order in T. In particular, we freely use that $KT^k\log^l Te^{KT} \le K'e^{K'T}$ for a constant $K' \ge K$.

- For readers convenience, we copy here the two simplified versions of Gronwall's lemma that will be used extensively in the proof.

  (i) Consider an interval $I = [0, t]$ and ϕ a real-valued function defined on I, assume there exists positive constants α, β such that ϕ satisfies the integral inequality

$$
\phi(t) \le \alpha + \beta\int_0^t \phi(s)\mathrm{d}s, \qquad \forall t \in I,
$$

  then $\phi(t) \le \alpha e^{\beta t}$ for all $t \in I$.

  (ii) Consider a non-negative sequence $\{\phi_k\}_{k=0}^n$ and assume there exists positive constants α, β such that $\{\phi_k\}_{k=0}^n$ satisfies the summation inequality

$$
\phi_k \le \alpha + \beta\sum_{0\le l<k}\phi_l, \qquad \forall k \in \{0, 1, \dots, n\},
$$

  then $\phi_k \le \alpha + \alpha\beta k e^{\beta k}$ for all $k \in \{0, 1, \dots, n\}$.

B Proof of Theorem 1 part (A)

Throughout this section, the assumptions of Theorem 1 (A) are understood to hold. These are assumptions A1-A4 in Section 3. In writing the proofs, for notational simplicity, we consider the following special setting:

R1. The coefficients $a_i \equiv 1$.

R2. The step size function $\xi(t) \equiv 1/2$.

The proof can be easily generalized to the case of general bounded coefficient $|a_i| \le K$, and non-constant function $\xi(t)$.

In the proof of this theorem, we have $(\boldsymbol{\theta}_i)_{i=1}^N = (\boldsymbol{w}_i)_{i=1}^N$, and

$$
\begin{aligned}
V(\boldsymbol{\theta}_i) =& a_i v(\boldsymbol{w}_i) = v(\boldsymbol{w}_i),\\
U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) =& a_i a_j u(\boldsymbol{w}_i, \boldsymbol{w}_j) = u(\boldsymbol{w}_i, \boldsymbol{w}_j).
\end{aligned}
$$

We will consider four dynamics (note we choose $\xi(t) = 1/2$ in these equations):

- The *nonlinear dynamics (ND)*: we introduce $(\bar{\boldsymbol{\theta}}_i^t)_{i\in[N], t\ge 0}$ with initialization $\bar{\boldsymbol{\theta}}_i^0 \sim \rho_0$ i.i.d.:

  $$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{\theta}}_i^t = -2\xi(t)\Big[\nabla V(\bar{\boldsymbol{\theta}}_i^t) + \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big].$$

  Equivalently, we have the integral equation

  $$\bar{\boldsymbol{\theta}}_i^t = \bar{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\mathrm{d}s, \tag{24}$$

  where we denoted $\boldsymbol{G}(\boldsymbol{\theta}; \rho) = -\nabla\Psi(\boldsymbol{\theta}; \rho) = -\nabla V(\boldsymbol{\theta}) - \int \nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')$. Note that $\bar{\boldsymbol{\theta}}_i^t$ is random because of its random initialization, and its law is ρ_t.

- The *particle dynamics (PD)*: we introduce $(\underline{\boldsymbol{\theta}}_i^t)_{i\in[N], t\ge 0}$ with initialization $\underline{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:

  $$\frac{\mathrm{d}}{\mathrm{d}t}\underline{\boldsymbol{\theta}}_i^t = -2\xi(t)\Big[\nabla V(\underline{\boldsymbol{\theta}}_i^t) + \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\underline{\boldsymbol{\theta}}_i^t, \underline{\boldsymbol{\theta}}_j^t)\Big].$$

  We introduce the particle distribution $\underline{\rho}_t^{(N)} = (1/N)\sum_{i=1}^N \delta_{\underline{\boldsymbol{\theta}}_i^t}$. In integration form, we get:

  $$\underline{\boldsymbol{\theta}}_i^t = \underline{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)})\mathrm{d}s. \tag{25}$$

- The *gradient descent (GD)*: we introduce $(\tilde{\boldsymbol{\theta}}_i^k)_{i\in[N], k\in\mathbb{N}}$ with initialization $\tilde{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:

  $$\tilde{\boldsymbol{\theta}}_i^{k+1} = \tilde{\boldsymbol{\theta}}_i^k - 2s_k\Big[\nabla V(\tilde{\boldsymbol{\theta}}_i^k) + \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k, \tilde{\boldsymbol{\theta}}_j^k)\Big],$$

  where $s_k = \varepsilon\xi(k\varepsilon)$. We introduce the particle distribution $\tilde{\rho}_k^{(N)} = (1/N)\sum_{i=1}^N \delta_{\tilde{\boldsymbol{\theta}}_i^k}$. In summation form, we get:

  $$\tilde{\boldsymbol{\theta}}_i^k = \tilde{\boldsymbol{\theta}}_i^0 + 2\varepsilon\sum_{l=0}^{k-1} \xi(l\varepsilon)\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)}). \tag{26}$$

  The GD dynamic corresponds to the discretized particle dynamic (25).

- The *stochastic gradient descent (SGD)*: we introduce $(\boldsymbol{\theta}_i^k)_{i\in[N],k\in\mathbb{N}}$ with initialization $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:
$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k - 2s_k \boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}),$$
where $\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) = (y_{k+1} - \hat{y}_{k+1})\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}_i^k)$, with $\boldsymbol{z}_k \equiv (\boldsymbol{x}_k, y_k)$ and $\hat{y}_{k+1} = (1/N)\sum_{j=1}^N \sigma_\star(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}_j^k)$. In summation form, we have
$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon \sum_{l=0}^{k-1} \xi(l\varepsilon)\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}). \tag{27}$$

Denote $\boldsymbol{\theta}^t = (\boldsymbol{\theta}_1^t, \ldots, \boldsymbol{\theta}_N^t)$, $\bar{\boldsymbol{\theta}}^t = (\bar{\boldsymbol{\theta}}_1^t, \ldots, \bar{\boldsymbol{\theta}}_N^t)$, $\tilde{\boldsymbol{\theta}}^t = (\tilde{\boldsymbol{\theta}}_1^t, \ldots, \tilde{\boldsymbol{\theta}}_N^t)$, and $\underline{\boldsymbol{\theta}}^t = (\underline{\boldsymbol{\theta}}_1^t, \ldots, \underline{\boldsymbol{\theta}}_N^t)$. For $t \in \mathbb{R}_{\geq 0}$, define $[t] = \varepsilon\lfloor t/\varepsilon \rfloor$. We will use the nonlinear dynamics, particle dynamics, gradient descent dynamics as interpolation dynamics

$$\begin{aligned}
&\Big|R(\rho_{k\varepsilon}) - R_N(\boldsymbol{\theta}^k)\Big| \\
\leq &\underbrace{\Big|R(\rho_{k\varepsilon}) - R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon})\Big|}_{\text{PDE}-\text{ND}} + \underbrace{\Big|R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon})\Big|}_{\text{ND}-\text{PD}} + \underbrace{\Big|R_N(\boldsymbol{\theta}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)\Big|}_{\text{PD}-\text{GD}} + \underbrace{\Big|R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)\Big|}_{\text{GD}-\text{SGD}}.
\end{aligned}$$

By Proposition 1, 2, 3, 4 proved below, we have with probability at least $1 - e^{-z^2}$,

$$\begin{aligned}
\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \leq& K\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z], \\
\sup_{t\in[0,T]} |R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \leq& Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z], \\
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon})| \leq& Ke^{KT}\varepsilon, \\
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\boldsymbol{\theta}^k) - R_N(\tilde{\boldsymbol{\theta}}^k)| \leq& Ke^{KT}\sqrt{T\varepsilon}[\sqrt{D + \log N} + z].
\end{aligned}$$

Combining these inequalities gives the conclusion of Theorem 1 (A). In the following subsections, we prove all the above interpolation bounds, under the setting of Theorem 1 (A).

B.1 Technical lemmas

Assumptions A1 - A3 immediately implies that

Lemma 2. *There exists a constant K depending on K_1, K_2, K_3, such that*
$$|V|, |U|, \|\nabla V\|_2, \|\nabla U\|_2, \|\nabla^2 V\|_{\text{op}}, \|\nabla^2 U\|_{\text{op}} \leq K.$$
For any $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i=1}^N$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}_i')_{i=1}^N$, we have
$$|R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}')| \leq K \max_{i\leq N} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2. \tag{28}$$

Proof of Lemma 2. Note we have
$$\begin{aligned}
V(\boldsymbol{\theta}) =& -\mathbb{E}_{y,\boldsymbol{x}}[y\sigma(\boldsymbol{x}; \boldsymbol{\theta})], \\
U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) =& \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}; \boldsymbol{\theta}_1)\sigma(\boldsymbol{x}; \boldsymbol{\theta}_2)].
\end{aligned}$$
The boundedness of V and U are implied by the boundedness of $\|\sigma\|_\infty$ and $|y|$ in Assumption A1. The boundedness of $\|\nabla V\|_2, \|\nabla U\|_2, \|\nabla^2 V\|_{\text{op}}, \|\nabla^2 U\|_{\text{op}}$ are implied by Assumption A3.

Finally, Eq. (28) holds by noting that
$$|R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}')| \leq \frac{1}{N}\sum_{i=1}^N |V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}_i')| + \frac{1}{N^2}\sum_{i,j=1}^N |U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_j')|,$$
and by the Lipschitz property of V and U. □

Using Eq. (24) and (25), we immediately have

Lemma 3. *There exists a constant K such that for any time s,t*

$$\begin{aligned}\|\underline{\boldsymbol{\theta}}_i^t - \underline{\boldsymbol{\theta}}_i^s\|_2 \leq & K|t-s|,\\ \|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^s\|_2 \leq & K|t-s|,\\ W_2(\rho_t, \rho_s) \leq & K|t-s|.\end{aligned}$$

Proof of Lemma 3. The first two inequalities are simply implied by the boundedness of ∇V and $\nabla_1 U$, and Eq. (24) and (25). The third inequality is simply implied by

$$W_2(\rho_t, \rho_s) \leq (\mathbb{E}[\|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^s\|_2^2])^{1/2}.$$

□

B.2 Bound between PDE and nonlinear dynamics

Proposition 1 (PDE-ND)**.** *There exists a constant K depending only on the K_i, $i = 1, 2, 3$, such that with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \leq K\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z].$$

Proof of Proposition 1. We decompose the difference into the following two terms

$$|R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \leq \underbrace{|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|}_{\text{I}} + \underbrace{|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)|}_{\text{II}}.$$

where the expectation is taken with respect to $\bar{\boldsymbol{\theta}}_i^0 \sim \rho_0$. The result holds simply by combining Lemma 4 and Lemma 5. □

Lemma 4 (Term II bound)**.** *We have*

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \leq K/N.$$

Proof of Lemma 4. The bound holds simply by observing that

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| = \frac{1}{N}\Big|\int U(\boldsymbol{\theta}, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta}) - \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\rho_t(\mathrm{d}\boldsymbol{\theta}_1)\rho_t(\mathrm{d}\boldsymbol{\theta}_2)\Big| \leq K/N.$$

□

Lemma 5 (Term I bound)**.** *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \leq K[\sqrt{\log(NT)} + z]/\sqrt{N}\Big) \geq 1 - e^{-z^2}.$$

Proof of Lemma 5. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_i, \dots, \boldsymbol{\theta}_N)$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_i', \dots \boldsymbol{\theta}_N)$ be two configurations that differ only in the i'th variable. Then

$$\begin{aligned}&|R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}')|\\ \leq&\frac{2}{N}|V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}_i')| + \frac{1}{N^2}|U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_i')| + \frac{2}{N^2}\sum_{j\in[N], j\neq i}|U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_j)|\\ \leq&\frac{K}{N}.\end{aligned} \tag{29}$$

Applying McDiarmid's inequality, we have

$$\mathbb{P}\Big(|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \geq \delta\Big) \leq \exp\{-N\delta^2/K\}.$$

By Lemma 3 and 2, we have

$$\Big| |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| - |R_N(\bar{\boldsymbol{\theta}}^s) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^s)| \Big| \le K|s-t|.$$

Hence taking the union bound over $s \in \eta\{0, 1, \dots, \lfloor T/\eta \rfloor\}$ and bounding the difference between time in the interval and grid, we have

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \ge \delta + K\eta \Big) \le (T/\eta)\exp\{-N\delta^2/K\}.$$

Now taking $\eta = 1/\sqrt{N}$ and $\delta = K[\sqrt{\log(NT)} + z]/\sqrt{N}$, we get the desired result. □

B.3 Bound between nonlinear dynamics and particle dynamics

Proposition 2 (ND-PD)**.** *There exists a constant K, such that with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{t\in[0,T]} \max_{i\in[N]} \|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \le K e^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z], \tag{30}$$

$$\sup_{t\in[0,T]} |R_N(\boldsymbol{\theta}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \le K e^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z]. \tag{31}$$

Proof of Proposition 2. Note we have

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2^2 =& \langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \nabla V(\bar{\boldsymbol{\theta}}_i^t) - \nabla V(\underline{\boldsymbol{\theta}}_i^t)\rangle + \Big\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \nabla_1 U(\underline{\boldsymbol{\theta}}_i^t, \underline{\boldsymbol{\theta}}_j^t)\Big\rangle \\
&- \frac{1}{N}\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\rangle \\
&- \Big\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \frac{1}{N}\sum_{j\ne i} \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big\rangle \\
\le& K\|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \cdot \max_{j\in[N]}\|\underline{\boldsymbol{\theta}}_j^t - \bar{\boldsymbol{\theta}}_j^t\|_2 + \|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 (K/N + I_i^t),
\end{aligned} \tag{32}$$

where

$$I_i^t \equiv \Big\| \frac{1}{N}\sum_{j\ne i}\Big[\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big]\Big\|_2.$$

We would like to prove a uniform bound for I_i^t for $i \in [N]$ and $t \in [0, T]$.

Lemma 6. *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} \max_{i\in[N]} I_i^t \le K[\sqrt{\log(NT)} + z]/\sqrt{N} \Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 6. Define $\boldsymbol{X}_i^t = \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})$. Note we have $\mathbb{E}[\boldsymbol{X}_i^t | \bar{\boldsymbol{\theta}}_i^t] = 0$ (where expectation is taken with respect to $\bar{\boldsymbol{\theta}}_j^0 \sim \rho_0$ for $j \ne i$), and $\|\boldsymbol{X}_i^t\|_2 \le 2K$ (by assumption that $\|\nabla U\|_2 \le K$). By Lemma 30, we have for any fixed $i \in [N]$ and $t \in [0, T]$,

$$\mathbb{P}\Big(I_i^t \ge K(\sqrt{1/N} + \delta)\Big) = \mathbb{E}\Big[\mathbb{P}\Big(I_i^t \ge K(\sqrt{1/N} + \delta) | \bar{\boldsymbol{\theta}}_i^t\Big)\Big] \le \exp\{-N\delta^2\}.$$

By Lemma 3, there exists K such that, for any $0 \le t, s \le T$ and $i \in [N]$, we have

$$|I_i^t - I_i^s| \le K|t - s|.$$

Taking the union bound over $i \in [N]$ and $s \in \eta\{0, 1, \dots, \lfloor T/\eta \rfloor\}$ and bounding time in the interval and the grid, we have

$$\mathbb{P}\Big(\sup_{t\in[0,T]} \max_{i\in[N]} I_i^t \ge K(\sqrt{1/N} + \delta) + K\eta \Big) \le (NT/\eta)\exp\{-N\delta^2\}.$$

Taking $\eta = \sqrt{1/N}$, and $\delta = K[\sqrt{\log(NT)} + z]/\sqrt{N}$, we get the desired result. □

Let $\delta(N,T,z) = K[\sqrt{\log(NT)} + z]/\sqrt{N}$, and define

$$\Delta(t) = \sup_{s \in [t]} \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^s - \bar{\boldsymbol{\theta}}_i^s\|_2.$$

We condition on the good event in Lemma 6 to happen. By Eq. (32), we have

$$\frac{\mathrm{d}\Delta}{\mathrm{d}t}(t) \le K \cdot \Delta(t) + \delta(N,T,z),$$

and by Gronwall's inequality, we obtain

$$\Delta(T) \le K e^{KT} \delta(N,T,z).$$

By Eq. (28), this proves Eq. (30) and (31) hold with probability at least $1 - e^{-z^2}$. □

B.4 Bound between particle dynamics and GD

Proposition 3 (PD-GD)**.** *There exists a constant K such that:*

$$\begin{aligned}\sup_{k \in [0, t/\varepsilon] \cap \mathbb{N}} \max_{i \le N} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le& K e^{KT} \varepsilon, \\ \sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} |R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le& K e^{KT} \varepsilon.\end{aligned}$$

Proof of Proposition 3. By Lemma 3, we have

$$\begin{aligned}\|\underline{\boldsymbol{\theta}}_i^t - \underline{\boldsymbol{\theta}}_i^s\|_2 \le& K|t-s|, \\ W_2(\underline{\rho}_t^{(N)}, \underline{\rho}_s^{(N)}) \le& K|t-s|.\end{aligned}$$

For $k \in \mathbb{N}$ and $t = k\varepsilon$, we have

$$\begin{aligned}\|\underline{\boldsymbol{\theta}}_i^t - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le& \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s \\ \le& \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \underline{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s + \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s \\ \le& K t\varepsilon + K \int_0^t \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^{[s]} - \tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}\|_2 \mathrm{d}s.\end{aligned}$$

Denoting $\Delta(t) \equiv \sup_{k \in [0, t/\varepsilon] \cap \mathbb{N}} \max_{i \le N} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2$. We get the equation

$$\Delta(t) \le K \int_0^t \Delta(s) \mathrm{d}s + K t \varepsilon = K \int_0^t [\Delta(s) + \varepsilon] \mathrm{d}s.$$

Applying Gronwall's lemma, we get:

$$\Delta(T) \le K e^{KT} \varepsilon.$$

Using Eq. (28) concludes the proof. □

B.5 Bound between GD and SGD

Proposition 4 (GD-SGD)**.** *There exists a constant K, such that with probability at least $1 - e^{-z^2}$, we have*

$$\begin{aligned}\sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \max_{i \in [N]} \|\tilde{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_i^k\|_2 \le& K e^{KT} \sqrt{T\varepsilon}[\sqrt{D + \log N} + z], \qquad (33) \\ \sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)| \le& K e^{KT} \sqrt{T\varepsilon}[\sqrt{D + \log N} + z]. \qquad (34)\end{aligned}$$

Proof of Proposition 4. Denoting $\mathcal{F}_k = \sigma((\boldsymbol{\theta}_i^0)_{i\in[N]}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k)$ the σ-algebra generated by observations $\boldsymbol{z}_\ell = (y_\ell, \boldsymbol{x}_\ell)$ up to step k, we get:

$$\mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})|\mathcal{F}_k] = -\nabla V(\boldsymbol{\theta}_i^k) - \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k) = \boldsymbol{G}(\boldsymbol{\theta}_i^k, \rho_k^{(N)}),$$

where $\rho_k^{(N)} \equiv (1/N)\sum_{i\in[N]} \delta_{\boldsymbol{\theta}_i^k}$ is the empirical distribution of the SGD iterates. Hence we get:

$$\begin{aligned}\|\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^k\|_2 =& \Big\|\varepsilon\sum_{l=0}^{k-1}\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) - \varepsilon\sum_{l=0}^{k-1}\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)})\Big\|_2\\ \le& \Big\|\varepsilon\sum_{l=0}^{k-1}\boldsymbol{Z}_i^l\Big\|_2 + \varepsilon\sum_{l=0}^{k-1}\Big\|\boldsymbol{G}(\boldsymbol{\theta}_i^l; \rho_l^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)})\Big\|_2\\ \equiv& A_i^k + B_i^k,\end{aligned}$$

where we denoted $\boldsymbol{Z}_i^l \equiv \boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) - \mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1})|\mathcal{F}_l]$ and $A_i^k = \|\varepsilon\sum_{l=0}^{k-1}\boldsymbol{Z}_i^l\|_2$.

Note $\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) = (y_{l+1} - \hat{y}_{l+1})\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}_{l+1}; \boldsymbol{w}_i^l)$ for $\boldsymbol{z}_{l+1} = (y_{l+1}, \boldsymbol{x}_{l+1})$. Since we assumed in A2 that $\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}; \boldsymbol{w})$ is K-sub-Gaussian, and since y_{l+1} and $\hat{y}_{l+1}$ are K bounded, we have that $\boldsymbol{Z}_i^l$ is K-sub-Gaussian (the product of a bounded random variable and a sub-Gaussian random variable is sub-Gaussian). We can therefore apply Azuma-Hoeffding inequality (Lemma 31) and get:

$$\mathbb{P}\Big(\max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} A_i^k \ge K\sqrt{T\varepsilon}(\sqrt{D} + z)\Big) \le e^{-z^2}.$$

Taking the union bound over $i \in [N]$, we get:

$$\mathbb{P}\Big(\max_{i\in[N]}\max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} A_i^k \ge K\sqrt{T\varepsilon}(\sqrt{D + \log N} + z)\Big) \le e^{-z^2}. \tag{35}$$

Introducing $\Delta(t) \equiv \sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}\|\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^k\|_2$, the B_i^k terms can be bounded by:

$$B_i^k \le K\int_0^{k\varepsilon}\|\boldsymbol{G}(\boldsymbol{\theta}_i^{[s]/\varepsilon}; \rho_{[s]/\varepsilon}^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s \le K\int_0^{k\varepsilon}\Delta(s)\mathrm{d}s.$$

Assuming the bad events in Eq. (35) does not happen, we have

$$\Delta(t) \le K\int_0^t \Delta(s)\mathrm{d}s + K\sqrt{T\varepsilon}(\sqrt{D + \log N} + z).$$

Applying Gronwall's inequality and applying Eq. (28) concludes the proof. □

C Proof of Theorem 1 part (B)

The difference in the proof of part (B) with the proof of part (A) comes from the fact that the functions V and U are not bounded and Lipschitz anymore, and that $\hat{f}(\boldsymbol{x}; \boldsymbol{\theta})$ is not bounded by a constant. However, we show that when starting from an initial distribution ρ_0 with compact support in the variable a, the support of ρ_t in the variable a remains bounded uniformly on the interval $[0, T]$ by a constant that only depends on the K_i, $i = 1, 2, 3, 4$, and T.

For $\boldsymbol{\theta} = (a, \boldsymbol{w})$ and $\boldsymbol{\theta}' = (a', \boldsymbol{w}')$, remember we have

$$\begin{aligned}\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}) =& a\sigma(\boldsymbol{x}; \boldsymbol{w}),\\ v(\boldsymbol{w}) =& -\mathbb{E}_{y,\boldsymbol{x}}[y\sigma(\boldsymbol{x}; \boldsymbol{w})],\\ u(\boldsymbol{w}, \boldsymbol{w}') =& \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}; \boldsymbol{w})\sigma(\boldsymbol{x}; \boldsymbol{w}')],\\ V(\boldsymbol{\theta}) =& a \cdot v(\boldsymbol{w}),\\ U(\boldsymbol{\theta}, \boldsymbol{\theta}') =& aa' \cdot u(\boldsymbol{w}, \boldsymbol{w}'),\end{aligned}$$

hence we have

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) =&(v(\boldsymbol{w}), a\nabla_{\boldsymbol{w}} v(\boldsymbol{w})),\\ \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}, \boldsymbol{\theta}') =&(a' \cdot u(\boldsymbol{w}, \boldsymbol{w}'), aa' \cdot \nabla_{\boldsymbol{w}} u(\boldsymbol{w}, \boldsymbol{w}')).\end{aligned}$$

Throughout this section, the assumptions A1 - A4 are understood to hold. For the sake of simplicity we will write the proof under the following restriction:

R1. The step size function $\xi(t) \equiv 1/2$.

The proof for a general function $\xi(t)$ is obtained by a straightforward adaptation.

We define the four dynamics with the same definitions as at the beginning of Section B. We copy them here for reader's convenience.

- The *nonlinear dynamics (ND)*: $(\bar{\boldsymbol{\theta}}_i^t)_{i\in[N],t\geq 0}$ with initialization $\bar{\boldsymbol{\theta}}_i^0 \sim \rho_0$ i.i.d.:
$$\bar{\boldsymbol{\theta}}_i^t = \bar{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\mathrm{d}s, \tag{36}$$
where we denoted $\boldsymbol{G}(\boldsymbol{\theta}; \rho) = -\nabla\Psi(\boldsymbol{\theta}; \rho) = -\nabla V(\boldsymbol{\theta}) - \int \nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')$.
- The *particle dynamics (PD)*: $(\underline{\boldsymbol{\theta}}_i^t)_{i\in[N],t\geq 0}$ with initialization $\underline{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:
$$\underline{\boldsymbol{\theta}}_i^t = \underline{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)})\mathrm{d}s, \tag{37}$$
where $\underline{\rho}_t^{(N)} = (1/N)\sum_{i=1}^N \delta_{\underline{\boldsymbol{\theta}}_i^t}$.
- The *gradient descent (GD)*: $(\tilde{\boldsymbol{\theta}}_i^k)_{i\in[N],k\in\mathbb{N}}$ with initialization $\tilde{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:
$$\tilde{\boldsymbol{\theta}}_i^k = \tilde{\boldsymbol{\theta}}_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)}). \tag{38}$$
where $s_k = \varepsilon\xi(k\varepsilon)$ and $\tilde{\rho}_k^{(N)} = (1/N)\sum_{i=1}^N \delta_{\tilde{\boldsymbol{\theta}}_i^k}$.
- The *stochastic gradient descent (SGD)*: $(\boldsymbol{\theta}_i^k)_{i\in[N],k\in\mathbb{N}}$ with initialization $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}_i^0$:
$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}), \tag{39}$$
where $\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) = (y_{k+1} - \hat{y}_{k+1})\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}_i^k)$, with $\boldsymbol{z}_k \equiv (\boldsymbol{x}_k, y_k)$ and $\hat{y}_{k+1} = (1/N)\sum_{j=1}^N a_j^k \sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_j^k)$.

We have the decomposition

$$\begin{aligned}&\Big|R(\rho_{k\varepsilon}) - R_N(\boldsymbol{\theta}^k)\Big|\\ \leq&\underbrace{\Big|R(\rho_{k\varepsilon}) - R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon})\Big|}_{\text{PDE}-\text{ND}} + \underbrace{\Big|R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon})\Big|}_{\text{ND}-\text{PD}} + \underbrace{\Big|R_N(\boldsymbol{\theta}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)\Big|}_{\text{PD}-\text{GD}} + \underbrace{\Big|R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)\Big|}_{\text{GD}-\text{SGD}}.\end{aligned}$$

By Proposition 5, 6, 7, 8, there exists constants K and K_0, such that if we take $\varepsilon \leq 1/[K_0(D + \log N + z^2)e^{K_0(1+T)^3}]$, with probability at least $1 - e^{-z^2}$, we have

$$\begin{aligned}\sup_{t\in[0,T]}|R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \leq& K(1+T)^4\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z],\\ \sup_{t\in[0,T]}|R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \leq& Ke^{K(1+T)^3}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z],\\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}|R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon})| \leq& Ke^{K(1+T)^3}\varepsilon,\\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}|R_N(\boldsymbol{\theta}^k) - R_N(\tilde{\boldsymbol{\theta}}^k)| \leq& Ke^{K(1+T)^3}\sqrt{\varepsilon}[\sqrt{D + \log N} + z].\end{aligned}$$

Combining these inequalities, and noting that $Ke^{K(1+T)^3} \le K'e^{K'T^3}$ for some $K' \ge K$, give the conclusion of Theorem 1 (B). In the following subsections, we prove all the above interpolation bounds, under the setting of Theorem 1 (B).

C.1 Technical lemmas

Lemma 7. *There exists a constant K depending only on the K_i, $i = 1, 2, 3, 4$, such that*

$$\begin{aligned}\mathrm{supp}(\rho_t) \subseteq& [-K(1+t), K(1+t)] \times \mathbb{R}^{D-1},\\ |\bar{a}_i^t| \le& K(1+t),\\ |\underline{a}_i^t| \le& K(1+t).\end{aligned}$$

Proof of Lemma 7.
Step 1. Let $\bar{\boldsymbol{\theta}}_i^t = (\bar{a}_i^t, \bar{\boldsymbol{w}}_i^t)$, and $\hat{y}(\boldsymbol{x}; \rho_t) = \int a\sigma(\boldsymbol{x}; \boldsymbol{w})\rho_t(\mathrm{d}\boldsymbol{\theta})$. Note that along the PDE, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} R(\rho_t) = -\int \|\nabla\Psi(\boldsymbol{\theta}; \rho_t)\|_2^2 \rho_t(\mathrm{d}\boldsymbol{\theta}) \le 0.$$

Hence we have (note $|y| \le K$, $|\sigma| \le K$, and $\mathrm{supp}(\rho_0) \subseteq [-K, K] \times \mathbb{R}^{D-1}$)

$$R(\rho_t) = \mathbb{E}_{y,\boldsymbol{x}}[(y - \hat{y}(\boldsymbol{x}; \rho_t))^2] \le R(\rho_0) = \mathbb{E}_{y,\boldsymbol{x}}\Big[\Big(y - \int a\sigma(\boldsymbol{x}; \boldsymbol{w})\rho_0(\mathrm{d}\boldsymbol{\theta})\Big)^2\Big] \le K.$$

The nonlinear dynamics for $\bar{a}_i^t$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{a}_i^t = \mathbb{E}_{y,\boldsymbol{x}}[(y - \hat{y}(\boldsymbol{x}; \rho_t))\sigma(\boldsymbol{x}; \bar{\boldsymbol{w}}_i^t)],$$

which gives

$$\Big|\frac{\mathrm{d}}{\mathrm{d}t}\bar{a}_i^t\Big| \le \{\mathbb{E}_{y,\boldsymbol{x}}[(y - \hat{y}(\boldsymbol{x}; \rho_t))^2]\mathbb{E}_{y,\boldsymbol{x}}[\sigma(\boldsymbol{x}; \bar{\boldsymbol{w}}_i^t)^2]\}^{1/2} \le K.$$

Hence, we have

$$|\bar{a}_i^t| \le |\bar{a}_i^0| + Kt \le K(1+t).$$

Note $(\bar{a}_i^t, \bar{\boldsymbol{w}}_i^t) \sim \rho_t$, hence we have $\mathrm{supp}(\rho_t) \subseteq [-K(1+t), K(1+t)] \times \mathbb{R}^{D-1}$.
Step 2. Denote $\underline{\boldsymbol{\theta}}_i^t = (\underline{a}_i^t, \underline{\boldsymbol{w}}_i^t)$, $\underline{\rho}_t^{(N)} = (1/N)\sum_{i=1}^N \delta_{\underline{\boldsymbol{\theta}}_i^t}$, and denote $\underline{y}(\boldsymbol{x}; \underline{\boldsymbol{\theta}}^t) = (1/N)\sum_{i\in[N]} \underline{a}_i^t\sigma(\boldsymbol{x}; \underline{\boldsymbol{w}}_i^t)$. Note along the PDE, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} R_N(\underline{\boldsymbol{\theta}}^t) = -\int \|\nabla\Psi(\boldsymbol{\theta}; \underline{\rho}_t^{(N)})\|_2^2 \underline{\rho}_t^{(N)}(\mathrm{d}\boldsymbol{\theta}) \le 0.$$

Hence we have (note $|y| \le K$, $|\sigma| \le K$, and $|\underline{a}_i^0| \le K$)

$$R_N(\underline{\boldsymbol{\theta}}^t) = \mathbb{E}_{y,\boldsymbol{x}}[(y - \underline{y}(\boldsymbol{x}; \underline{\rho}_t^{(N)}))^2] \le R_N(\underline{\boldsymbol{\theta}}^0) = \mathbb{E}_{y,\boldsymbol{x}}\Big[\Big(y - \int a\sigma(\boldsymbol{x}; \boldsymbol{w})\underline{\rho}_0^{(N)}(\mathrm{d}\boldsymbol{\theta})\Big)^2\Big] \le K.$$

The nonlinear dynamics for $\underline{a}_i^t$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{a}_i^t = \mathbb{E}_{y,\boldsymbol{x}}[(y - \underline{y}(\boldsymbol{x}; \underline{\boldsymbol{\theta}}^t))\sigma(\boldsymbol{x}; \underline{\boldsymbol{w}}_i^t)],$$

which gives

$$\Big|\frac{\mathrm{d}}{\mathrm{d}t}\underline{a}_i^t\Big| \le \{\mathbb{E}_{y,\boldsymbol{x}}[(y - \underline{y}(\boldsymbol{x}; \underline{\boldsymbol{\theta}}^t))^2]\mathbb{E}_{y,\boldsymbol{x}}[\sigma(\boldsymbol{x}; \underline{\boldsymbol{w}}_i^t)^2]\}^{1/2} \le K.$$

Hence, we have

$$|\underline{a}_i^t| \le |\underline{a}_i^0| + Kt \le K(1+t).$$

This proves the lemma. □

Lemma 8 (Boundness and Lipschitzness)**.** *Denoting* $\boldsymbol{\theta} = (a, \boldsymbol{w})$*,* $\boldsymbol{\theta}_1 = (a_1, \boldsymbol{w}_1)$ *and* $\boldsymbol{\theta}_2 = (a_2, \boldsymbol{w}_2)$*. We have*

$$\begin{aligned}
|V(\boldsymbol{\theta})|, \|\nabla V(\boldsymbol{\theta})\|_2 \leq& K(1+|a|),\\
|V(\boldsymbol{\theta}_1) - V(\boldsymbol{\theta}_2)|, \|\nabla V(\boldsymbol{\theta}_1) - \nabla V(\boldsymbol{\theta}_2)\|_2 \leq& K \cdot [1 + |a_1| \wedge |a_2|] \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,\\
|U(\boldsymbol{\theta}, \boldsymbol{\theta}')|, \|\nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 \leq& K(1+|a|)(1+|a'|),\\
|U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - U(\boldsymbol{\theta}_2, \boldsymbol{\theta})|, \|\nabla_{(1,2)} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - \nabla_{(1,2)} U(\boldsymbol{\theta}_2, \boldsymbol{\theta})\|_2 \leq& K(1+|a|) \cdot [1 + |a_1| \wedge |a_2|] \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,\\
|R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}')| \leq& K \max_{i \in [N]} (1 + |a_i| \vee |a_i'|)^2 \cdot \max_{j \in [N]} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j'\|_2.
\end{aligned}$$

Proof of Lemma 8. We have

$$\begin{aligned}
|V(\boldsymbol{\theta})| =& |a v(\boldsymbol{w})| \leq K|a|,\\
\|\nabla V(\boldsymbol{\theta})\|_2 =& \|(v(\boldsymbol{w}), a \nabla_{\boldsymbol{w}} v(\boldsymbol{w}))\|_2 \leq K(1+|a|),
\end{aligned}$$

and (assuming $|a_1| \geq |a_2|$)

$$|V(\boldsymbol{\theta}_1) - V(\boldsymbol{\theta}_2)| = |a_1 v(\boldsymbol{w}_1) - a_2 v(\boldsymbol{w}_2)| \leq K[|a_1 - a_2| + |a_2| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2] \leq K[1 + |a_2|] \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,$$

and

$$\begin{aligned}
\|\nabla V(\boldsymbol{\theta}_1) - \nabla V(\boldsymbol{\theta}_2)\|_2 =& \|(v(\boldsymbol{w}_1) - v(\boldsymbol{w}_2), a_1 \nabla v(\boldsymbol{w}_1) - a_2 \nabla v(\boldsymbol{w}_2))\|_2\\
\leq& K\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + K\|a_1 \nabla v(\boldsymbol{w}_1) - a_2 \nabla v(\boldsymbol{w}_1)\|_2 + \|a_2[\nabla v(\boldsymbol{w}_1) - \nabla v(\boldsymbol{w}_2)]\|_2\\
\leq& K[\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + |a_1 - a_2|] + K|a_2| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2\\
\leq& K(1 + |a_2|) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,
\end{aligned}$$

and

$$|U(\boldsymbol{\theta}, \boldsymbol{\theta}')| = |a a' u(\boldsymbol{w}, \boldsymbol{w}')| \leq K|a||a'|,$$

and

$$\|\nabla U(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 = \|(a' u(\boldsymbol{w}, \boldsymbol{w}'), a a' \cdot \nabla_1 u(\boldsymbol{w}, \boldsymbol{w}'))\|_2 \leq K|a'|(1+|a|),$$

and (assuming $|a_1| \geq |a_2|$)

$$\begin{aligned}
|U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - U(\boldsymbol{\theta}_2, \boldsymbol{\theta})| =& |a_1 a u(\boldsymbol{w}_1, \boldsymbol{w}) - a_2 a u(\boldsymbol{w}_2, \boldsymbol{w})|\\
\leq& K[|a_1 - a_2||a| + |a_2||a| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2]\\
\leq& K(1 + |a_2|)|a| \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - \nabla_1 U(\boldsymbol{\theta}_2, \boldsymbol{\theta})\|_2 =& \|(a u(\boldsymbol{w}_1, \boldsymbol{w}) - a u(\boldsymbol{w}_2, \boldsymbol{w}), a_1 a \nabla_1 u(\boldsymbol{w}_1, \boldsymbol{w}) - a_2 a \nabla_1 u(\boldsymbol{w}_2, \boldsymbol{w}))\|_2\\
\leq& |a| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + K|a||a_1 - a_2| + K|a||a_2| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2\\
\leq& K|a|(1 + |a_2|) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla_2 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - \nabla_2 U(\boldsymbol{\theta}_2, \boldsymbol{\theta})\|_2 =& \|(a_1 u(\boldsymbol{w}_1, \boldsymbol{w}) - a_2 u(\boldsymbol{w}_2, \boldsymbol{w}), a_1 a \nabla_2 u(\boldsymbol{w}_1, \boldsymbol{w}) - a_2 a \nabla_2 u(\boldsymbol{w}_2, \boldsymbol{w}))\|_2\\
\leq& K|a_1 - a_2| + K|a_2| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + K|a||a_1 - a_2| + K|a||a_2| \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2\\
\leq& K(1 + |a|)(1 + |a_2|) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&|R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}')|\\
\leq& 2 \max_{i \in [N]} |V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}_i')| + \max_{i,j \in [N]} |U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_j')|\\
\leq& K\Big[\max_{i \in [N]} (1 + |a_i| \wedge |a_i'|) \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2 + \max_{i,j \in [N]} (1 + |a_i| \wedge |a_i'|)(|a_j| \vee |a_j'|) \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2\Big]\\
\leq& K \max_{i \in [N]} (1 + |a_i| \vee |a_i'|)^2 \cdot \max_{j \in [N]} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j'\|_2.
\end{aligned}$$

This concludes the proof. □

Lemma 9. *There exists a constant K such that for any time $0 \le s < t$*

$$\begin{aligned}
\|\underline{\boldsymbol{\theta}}_i^t - \underline{\boldsymbol{\theta}}_i^s\|_2 \le& K(1+s)^2|t-s|,\\
\|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^s\|_2 \le& K(1+s)^2|t-s|,\\
W_2(\rho_t,\rho_s) \le& K(1+s)^2|t-s|.
\end{aligned}$$

Proof of Lemma 9. This lemma holds by the bounds of ∇V and $\nabla_1 U$ in Lemma 8 and the bounds for $|\bar{a}_i^t|, |\underline{a}_i^t|$ in Lemma 7, and by the inequality

$$W_2(\rho_t,\rho_s) \le (\mathbb{E}[\|\bar{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^s\|_2^2])^{1/2}.$$

$\square$

C.2 Bound between PDE and nonlinear dynamics

Proposition 5 (PDE-ND)**.** *There exists a constant K, such that with probability at least $1-e^{-z^2}$, we have*

$$\sup_{t\in[0,T]\cap\mathbb{N}} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le K(1+T)^4 \frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z]$$

Proof of Proposition 5. We decompose the difference into the following two terms

$$|R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le \underbrace{|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|}_{\mathrm{I}} + \underbrace{|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)|}_{\mathrm{II}}.$$

where the expectation is taken with respect to $\bar{\boldsymbol{\theta}}_i^0 \sim \rho_0$. The result holds simply by combining Lemma 10 and Lemma 11. $\square$

Lemma 10 (Term II bound)**.** *We have*

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le K(1+t)^2/N.$$

Proof of Lemma 10. The bound hold simply by observing that

$$\begin{aligned}
|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| =& \frac{1}{N}\Big|\int a^2 u(\boldsymbol{w},\boldsymbol{w})\rho_t(\mathrm{d}\boldsymbol{\theta}) - \int a_1 a_2 u(\boldsymbol{w}_1,\boldsymbol{w}_2)\rho_t(\mathrm{d}\boldsymbol{\theta}_1)\rho_t(\mathrm{d}\boldsymbol{\theta}_2)\Big|\\
\le& (K/N)\int a^2\rho_t(\mathrm{d}\boldsymbol{\theta}) \le K(1+t)^2/N.
\end{aligned}$$

$\square$

Lemma 11 (Term I bound)**.** *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \le K(1+T)^4[\sqrt{\log(NT)} + z]/\sqrt{N}\Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 11. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_N)$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}'_i,\ldots\boldsymbol{\theta}_N)$ be two configurations that differ only in the i'th variable. Assuming $a, a' \in [-K(1+t), K(1+t)]$, then

$$\begin{aligned}
&|R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}')|\\
\le& \frac{2}{N}|V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}'_i)| + \frac{1}{N^2}|U(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i) - U(\boldsymbol{\theta}'_i,\boldsymbol{\theta}'_i)| + \frac{2}{N^2}\sum_{j\in[N],j\neq i}|U(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) - U(\boldsymbol{\theta}'_i,\boldsymbol{\theta}_j)| \qquad (40)\\
\le& \frac{K}{N}(1+t)^2.
\end{aligned}$$

Note we have $\bar{a}_i^t \in [-K(1+t), K(1+t)]$, applying McDiarmid's inequality, we have

$$\mathbb{P}\Big(|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \ge \delta\Big) \le \exp\{-N\delta^2/[K(1+t)^4]\}.$$

By Lemma 9, 8 and 7, for $0 \le s < t$, we have

$$\begin{aligned}&|R_N(\bar{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^s)|\\ \le&K \max_{i\in[N]}(1 + |\bar{a}_i^s| \vee |\bar{a}_i^t|)^2 \cdot \max_{j\in[N]} \|\bar{\boldsymbol{\theta}}_j^t - \bar{\boldsymbol{\theta}}_j^s\|_2 \le K(1+t)^4|t-s|,\end{aligned}$$

which gives

$$\Big||R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| - |R_N(\bar{\boldsymbol{\theta}}^s) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^s)|\Big| \le K(1+t)^4|t-s|.$$

Hence taking union bound over $s \in \eta\{0, 1, \dots, \lfloor T/\eta \rfloor\}$ and bounding difference between time in the interval and grid, we have

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \ge \delta + K(1+T)^4\eta\Big) \le (T/\eta)\exp\{-N\delta^2/[K(1+T)^4]\}.$$

Now taking $\eta = 1/\sqrt{N}$ and $\delta = K(1+T)^4[\sqrt{\log(NT)} + z]/\sqrt{N}$, we get the desired inequality. □

C.3 Bound between nonlinear dynamics and particle dynamics

Proposition 6 (ND-PD)**.** *There exists a constant K, such that with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{t\in[0,T]} \max_{i\in[N]} \|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \le K e^{K(1+T)^3} \frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z], \tag{41}$$

$$\sup_{t\in[0,T]} |R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \le K e^{K(1+T)^3} \frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z]. \tag{42}$$

Proof of Proposition 6. Note we have

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2^2 =&\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \nabla V(\bar{\boldsymbol{\theta}}_i^t) - \nabla V(\underline{\boldsymbol{\theta}}_i^t)\rangle + \Big\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \underline{\boldsymbol{\theta}}_j^t)\Big\rangle\\ &+ \Big\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \underline{\boldsymbol{\theta}}_j^t) - \nabla_1 U(\underline{\boldsymbol{\theta}}_i^t, \underline{\boldsymbol{\theta}}_j^t)\Big\rangle\\ &- \frac{1}{N}\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\rangle\\ &- \Big\langle \underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t, \frac{1}{N}\sum_{j\ne i} \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big\rangle\\ \le& K(1+t)^2\|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \cdot \max_{j\in[N]} \|\underline{\boldsymbol{\theta}}_j^t - \bar{\boldsymbol{\theta}}_j^t\|_2 + \|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2(K(1+t)^2/N + I_i^t),\end{aligned} \tag{43}$$

where

$$I_i^t = \Big\|\frac{1}{N}\sum_{j\ne i}\Big[\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big]\Big\|_2.$$

The last inequality follows by Lemma 8 and 7. Now we would like to prove a uniform bound for I_i^t for $i \in [N]$ and $t \in [0, T]$.

Lemma 12. *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} \max_{i\in[N]} I_i^t \le K(1+T)^2[\sqrt{\log(NT)} + z]/\sqrt{N}\Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 12. Denote $\boldsymbol{X}_i^t = \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})$. Note we have $\mathbb{E}[\boldsymbol{X}_i^t|\bar{\boldsymbol{\theta}}_i^t] = 0$ (where expectation is taken with respect to $\bar{\boldsymbol{\theta}}_j^0 \sim \rho_0$ for $j \ne i$), and $\|\boldsymbol{X}_i^t\|_2 \le 2(1+t)^2K$ (by Lemma 8 and 7). By Lemma 30, we have for any fixed $i \in [N]$ and $t \in [0, T]$,

$$\mathbb{P}\Big(I_i^t \ge K(1+t)^2(\sqrt{1/N} + \delta)\Big) = \mathbb{E}\Big[\mathbb{P}\Big(I_i^t \ge K(1+t)^2(\sqrt{1/N} + \delta)|\bar{\boldsymbol{\theta}}_i^t\Big)\Big] \le \exp\{-N\delta^2\}.$$

By Lemma 9, there exists K such that, for any $0 \le s < t \le T$ and $i \in [N]$, we have

$$|I_i^t - I_i^s| \le K(1+t)^2 |t-s|.$$

Taking the union bound over $i \in [N]$ and $s \in \eta[T/\eta]$ and bounding time in the interval and the grid, we have

$$\mathbb{P}\Big(\sup_{t \in [0,T]} \max_{i \in [N]} I_i^t \ge K(1+T)^2(\sqrt{1/N} + \delta) + K(1+T)^2 \eta\Big) \le (NT/\eta) \exp\{-N\delta^2\}.$$

Taking $\eta = \sqrt{1/N}$, and $\delta = K[\sqrt{\log(NT)} + z]/\sqrt{N}$, we get the desired result. □

Denote $\delta(N,T,z) = K(1+T)^2[\sqrt{\log(NT)} + z]/\sqrt{N}$ and

$$\Delta(t) = \sup_{s \in [t]} \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^s - \bar{\boldsymbol{\theta}}_i^s\|_2.$$

We condition on the good event in Lemma 12 to happen. By Eq. (43), we have

$$\Delta'(t) \le K(1+T)^2 \cdot \Delta(t) + \delta(N,T,z),$$

By Gronwall's inequality, we have

$$\Delta(T) \le K e^{K(1+T)^3} \delta(N,T,z).$$

This happens with probability $1 - e^{-z^2}$. This proves Eq. (41). Finally, Eq. (42) holds by Lemma 8. □

C.4 Bound between particle dynamics and GD

Proposition 7 (PD-GD)**.** *There exists constants K and K_0 such that, letting $\varepsilon \le 1/(K_0 e^{K_0(1+T)^3})$, we have for any $t \le T$,*

$$\begin{aligned}
\sup_{k \in [0,t/\varepsilon] \cap \mathbb{N}} |\tilde{a}_i^k| \le& K(1+t),\\
\sup_{k \in [0,t/\varepsilon] \cap \mathbb{N}} \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le& K e^{K(1+T)^2 t} \varepsilon,\\
\sup_{k \in [0,t/\varepsilon] \cap \mathbb{N}} |R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le& K e^{K(1+T)^2 t} \varepsilon.
\end{aligned}$$

Proof of Proposition 7. Let $\underline{\rho}_s^{(N)} = (1/N)\sum_{i=1}^N \delta_{\underline{\boldsymbol{\theta}}_i^s}$, and $\tilde{\rho}_k^{(N)} = (1/N)\sum_{i=1}^N \delta_{\tilde{\boldsymbol{\theta}}_i^k}$. For $k \in \mathbb{N}$ and $t = k\varepsilon$, we have

$$\begin{aligned}
\|\underline{\boldsymbol{\theta}}_i^t - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le& 2\int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s\\
\le& 2\int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2 \mathrm{d}s + 2\int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s.
\end{aligned}$$

By Lemma 9 and 8, for $0 \le s \le t$, we have

$$\begin{aligned}
&\|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2\\
\le& \|\nabla V(\underline{\boldsymbol{\theta}}_i^s) - \nabla V(\underline{\boldsymbol{\theta}}_i^{[s]})\|_2 + \sup_{j \in [N]} \|\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s, \underline{\boldsymbol{\theta}}_j^s) - \nabla_1 U(\underline{\boldsymbol{\theta}}_i^s, \underline{\boldsymbol{\theta}}_j^{[s]})\|_2\\
&+ \sup_{j \in [N]} \|\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s, \underline{\boldsymbol{\theta}}_j^{[s]}) - \nabla_1 U(\underline{\boldsymbol{\theta}}_i^{[s]}, \underline{\boldsymbol{\theta}}_j^{[s]})\|_2\\
\le& K[1 + |\underline{a}_i^s|]\|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2 + \sup_{j \in [N]} K(1 + |\underline{a}_i^s|)(1 + |\underline{a}_j^{[s]}|)[\|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2 + \|\underline{\boldsymbol{\theta}}_j^s - \underline{\boldsymbol{\theta}}_j^{[s]}\|_2]\\
\le& K(1+t)^4(s - [s]) \le K(1+t)^4 \varepsilon,
\end{aligned}$$

and for $u = k\varepsilon \leq t$,

$$\begin{aligned}
&\|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^u; \underline{\rho}_u^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^k; \tilde{\rho}_k^{(N)})\|_2 \\
\leq &\|\nabla V(\underline{\boldsymbol{\theta}}_i^u) - \nabla V(\tilde{\boldsymbol{\theta}}_i^k)\|_2 + \sup_{j\in[N]} \|\nabla_1 U(\underline{\boldsymbol{\theta}}_i^u, \underline{\boldsymbol{\theta}}_j^u) - \nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k, \underline{\boldsymbol{\theta}}_j^u)\|_2 \\
&+ \sup_{j\in[N]} \|\nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k, \underline{\boldsymbol{\theta}}_j^u) - \nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k, \tilde{\boldsymbol{\theta}}_j^k)\|_2 \\
\leq &K(1 + |\underline{a}_i^u|)\|\underline{\boldsymbol{\theta}}_i^u - \tilde{\boldsymbol{\theta}}_i^k\|_2 + \sup_{j\in[N]} K(1 + |\underline{a}_i^u|)(1 + |\underline{a}_j^u|)\|\underline{\boldsymbol{\theta}}_i^u - \tilde{\boldsymbol{\theta}}_i^k\|_2 \\
&+ \sup_{j\in[N]} K(1 + |\tilde{a}_i^k|)(1 + |\underline{a}_j^u|)\|\underline{\boldsymbol{\theta}}_j^u - \tilde{\boldsymbol{\theta}}_j^k\|_2 \\
\leq &\max_{j\in[N]} K(1 + t + |\tilde{a}_j^k - \underline{a}_j^u|)(1 + t)\|\underline{\boldsymbol{\theta}}_i^u - \tilde{\boldsymbol{\theta}}_j^k\|_2 \leq K(1+t)^2 \cdot \max_{j\in[N]}\{\|\underline{\boldsymbol{\theta}}_j^u - \tilde{\boldsymbol{\theta}}_j^k\|_2, \|\underline{\boldsymbol{\theta}}_j^u - \tilde{\boldsymbol{\theta}}_j^k\|_2^2\}.
\end{aligned}$$

Denoting $\Delta(t) \equiv \sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}} \max_{i\leq N} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2$, we get the equation

$$\begin{aligned}
\Delta(t) \leq &K(1+t)^2 \int_0^t \max\{\Delta(s), \Delta(s)^2\} \mathrm{d}s + K(1+t)^4 t\varepsilon \\
\leq &K(1+T)^2 \int_0^t [\max\{\Delta(s), \Delta(s)^2\} + (1+T)^2\varepsilon] \mathrm{d}s.
\end{aligned}$$

Let $T_\Delta = \inf\{t : \Delta(t) \geq 1\}$. For $t \leq T_\Delta$, we have $\Delta(s)^2 \leq \Delta(s)$. Applying Gronwall's lemma, we get for any $t \leq T_\Delta$,

$$\Delta(t) \leq K e^{K(1+T)^2 t}\varepsilon.$$

Note we assumed $\varepsilon \leq 1/(K_0 e^{K_0(1+T)^3})$, which gives $K e^{K(1+T)^2 T}\varepsilon \leq 1/2$. This shows that $T_\Delta \geq T$. Hence we get

$$\Delta(T) \leq K e^{K(1+T)^2 T}\varepsilon.$$

Moreover, we immediately have,

$$\max_{i\in[N]} \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |\tilde{a}_i^k| \leq \max_{i\in[N]} \sup_{t\in[0,T]} |\underline{a}_i^t| + \max_{i\in[N]} \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \|\tilde{\boldsymbol{\theta}}_i^t - \underline{\boldsymbol{\theta}}_i^t\|_2 \leq K(1+t) + K e^{K(1+T)^2 T}\varepsilon \leq 2K(1+t).$$

Finally, applying the last inequality in Lemma 8 concludes the proof. □

C.5 Bound between GD and SGD

Proposition 8 (GD-SGD)**.** *There exists constants K and K_0, such that if we take $\varepsilon \leq 1/[K_0(D + \log N + z^2)e^{K_0(1+T)^3}]$, the following holds with probability at least $1 - e^{-z^2}$: for any $t \leq T$, we have*

$$\begin{aligned}
\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}} \max_{i\in[N]} |a_i^k| \leq &K(1+t), \\
\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}} \max_{i\in[N]} \|\tilde{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_i^k\|_2 \leq &K e^{K(1+T)^2 t}\sqrt{\varepsilon}[\sqrt{D + \log N} + z], \\
\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)| \leq &K e^{K(1+T)^2 t}\sqrt{\varepsilon}[\sqrt{D + \log N} + z].
\end{aligned}$$

Proof of Proposition 8. Denoting $\mathcal{F}_k = \sigma((\boldsymbol{\theta}_i^0)_{i\in[N]}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k)$ the σ-algebra generated by the data sample $\boldsymbol{z}_\ell = (y_\ell, \boldsymbol{x}_\ell)$ for $\ell \leq k$, we get:

$$\mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})|\mathcal{F}_k] = -\nabla V(\boldsymbol{\theta}_i^k) - \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k) = \boldsymbol{G}(\boldsymbol{\theta}_i^k, \rho_k^{(N)}),$$

where $\rho_k^{(N)} \equiv (1/N)\sum_{i\in[N]}\delta_{\boldsymbol{\theta}_i^k}$ denotes the empirical distribution of the iterates of SGD. Hence we get:

$$
\begin{aligned}
\|\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^k\|_2 =& \Big\|\varepsilon \sum_{l=0}^{k-1} \boldsymbol{F}_i(\boldsymbol{\theta}_i^l; \boldsymbol{z}_{l+1}) - \varepsilon \sum_{l=0}^{k-1} \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)})\Big\|_2 \\
\le& \Big\|\varepsilon \sum_{l=0}^{k-1} \boldsymbol{Z}_i^l\Big\|_2 + \varepsilon \sum_{l=0}^{k-1} \Big\|\boldsymbol{G}(\boldsymbol{\theta}_i^l; \rho_l^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)})\Big\|_2,
\end{aligned}
$$

where $\boldsymbol{Z}_i^l \equiv \boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) - \mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1})|\mathcal{F}_l]$.

Denote $\boldsymbol{A}_i^k = \sum_{l=0}^{k-1} \varepsilon \boldsymbol{Z}_i^l$. Hence $\{\boldsymbol{A}_i^k\}_{k\in\mathbb{N}}$ is a martingale adapted to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. Note we have

$$
\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) =((y_{k+1} - \hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k))\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k), (y_{k+1} - \hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k))a_i^k \nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)),
$$

where $\hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k) = (1/N)\sum_{j=1}^n a_j^k \sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_j^k)$.

The following discussion is under the conditional law $\mathcal{L}(\,\cdot\,|\mathcal{F}_k)$. Note that $|\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)| \le K$, and $|y_{k+1} - \hat{y}_{k+1}(\boldsymbol{\theta}^k)| \le K(1 + \max_j |a_j^k|)$, hence $(y_{k+1} - \hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k))\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)$ is $K(1 + \max_i |a_i^k|)$-sub-Gaussian. Furthermore, $\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)$ is a K-sub-Gaussian random vector, and $|(y_{k+1} - \hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k))a_i^k| \le K(1+\max_i |a_i^k|)^2$, hence $(y_{k+1} - \hat{y}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}^k))a_i^k \nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)$ is a $K(1+\max_j |a_j^k|)^2$-sub-Gaussian random vector. As a result, we have $\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})$ under the conditional law $\mathcal{L}(\,\cdot\,|\mathcal{F}_k)$ is a $K(1 + \max_j |a_j^k|)^2$-sub-Gaussian random vector (concatenation of two possibly dependent sub-Gaussian random vectors is sub-Gaussian).

Let $T_a = \min\{l : \max_{i\in[N]} |a_i^l| \ge M_T\}$ where $M_T \equiv 2K(1 + T)$. Then we have

$$
\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \varepsilon \boldsymbol{Z}_i^k\rangle}|\mathcal{F}_k]\mathbf{1}\{\max_{i\in[N]} |a_i^k| \le M_T\} \le e^{\varepsilon^2 K^2 M_T^4 \|\boldsymbol{\lambda}\|_2^2/2}.
$$

Now let $\bar{\boldsymbol{A}}_i^k = \boldsymbol{A}_i^{k\wedge T_a}$. Then $\bar{\boldsymbol{A}}_i^k$ is also a martingale. Furthermore, we have

$$
\begin{aligned}
&\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{A}}_i^k - \bar{\boldsymbol{A}}_i^{k-1}\rangle}|\mathcal{F}_{k-1}] \\
=&\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \boldsymbol{A}_i^k - \boldsymbol{A}_i^{k-1}\rangle}\mathbf{1}\{T_a \ge k\}|\mathcal{F}_{k-1}] + \mathbb{E}[e^{\langle \boldsymbol{\lambda}, \boldsymbol{A}_i^{T_a} - \boldsymbol{A}_i^{T_a}\rangle}\mathbf{1}\{T_a \le k-1\}|\mathcal{F}_{k-1}] \\
=&\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \varepsilon \boldsymbol{Z}_i^{k-1}\rangle}|\mathcal{F}_{k-1}]\mathbf{1}\{T_a \ge k\} + \mathbf{1}\{T_a \le k-1\} \\
=&\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \varepsilon \boldsymbol{Z}_i^{k-1}\rangle}|\mathcal{F}_{k-1}]\mathbf{1}\{\max_{i\in[N]} |a_i^{k-1}| < M_T\} + \mathbf{1}\{T_a \le k-1\} \\
\le& e^{\varepsilon^2 K^2 M_T^4 \|\boldsymbol{\lambda}\|_2^2/2}.
\end{aligned}
$$

Hence we can apply Azuma-Hoeffding's concentration bound (Lemma 31) to $\|\bar{\boldsymbol{A}}_i^l\|_2$,

$$
\mathbb{P}\Big(\max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \|\bar{\boldsymbol{A}}_i^k\|_2 \ge K M_T^2 \sqrt{T\varepsilon}(\sqrt{D} + z)\Big) \le e^{-z^2},
$$

and taking the union bound over $i \in [N]$, we get:

$$
\mathbb{P}\Big(\max_{i\in[N]} \max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \|\boldsymbol{A}_i^{k\wedge T_a}\|_2 \le K M_T^2 \sqrt{T\varepsilon}(\sqrt{D + \log N} + z)\Big) \ge 1 - e^{-z^2}. \tag{44}
$$

Denote the above event to be a good event $E_{\rm good}$,

$$
E_{\rm good} = \Big\{\max_{i\in[N]} \max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \|\boldsymbol{A}_i^{k\wedge T_a}\|_2 \le K M_T^2 \sqrt{T\varepsilon}(\sqrt{D + \log N} + z)\Big\}.
$$

We consider the case in which $E_{\rm good}$ happens. We have (note we assumed $\varepsilon \le 1/(K_0 e^{K_0(1+T)^3})$, by Proposition 7, we have $\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}|\tilde{a}_i^k|\le K(1+t)$)

$$\begin{aligned}
&\|\boldsymbol{G}(\boldsymbol{\theta}_i^k;\rho_k^{(N)})-\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^k;\tilde{\rho}_k^{(N)})\|_2\\
\le&\|\nabla V(\boldsymbol{\theta}_i^k)-\nabla V(\tilde{\boldsymbol{\theta}}_i^k)\|_2+\sup_{j\in[N]}\|\nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k,\boldsymbol{\theta}_j^k)-\nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k,\tilde{\boldsymbol{\theta}}_j^k)\|_2\\
&+\sup_{j\in[N]}\|\nabla_1 U(\boldsymbol{\theta}_i^k,\boldsymbol{\theta}_j^k)-\nabla_1 U(\tilde{\boldsymbol{\theta}}_i^k,\boldsymbol{\theta}_j^k)\|_2\\
\le&K(1+|\tilde{a}_i^k|)\|\boldsymbol{\theta}_i^k-\tilde{\boldsymbol{\theta}}_i^k\|_2+\sup_{j\in[N]}K(1+|\tilde{a}_i^k|)(1+|\tilde{a}_j^k|)\|\boldsymbol{\theta}_j^k-\tilde{\boldsymbol{\theta}}_j^k\|_2\\
&+\sup_{j\in[N]}K(1+|\tilde{a}_i^k|)(1+|a_j^k|)\|\boldsymbol{\theta}_i^k-\tilde{\boldsymbol{\theta}}_i^k\|_2\\
\le&K(1+T+|\tilde{a}_i^k-a_i^k|)(1+T)\max_{j\in[N]}\|\boldsymbol{\theta}_j^k-\tilde{\boldsymbol{\theta}}_j^k\|_2\\
\le&K(1+T)^2\cdot\max_{j\in[N]}\{\|\boldsymbol{\theta}_j^k-\tilde{\boldsymbol{\theta}}_j^k\|_2,\|\boldsymbol{\theta}_j^k-\tilde{\boldsymbol{\theta}}_j^k\|_2^2\}.
\end{aligned}$$

Denoting $\Delta(t)\equiv\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}\|\boldsymbol{\theta}_i^k-\tilde{\boldsymbol{\theta}}_i^k\|_2$. Denote $T_\Delta=\inf\{u:\Delta(u)\ge 1\}$. For $t\le T_a\wedge T_\Delta\wedge T$, we get the equation

$$\Delta(t)\le KM_T^2\int_0^t\Delta(s)\mathrm{d}s+KM_T^2\sqrt{\varepsilon T}(\sqrt{D+\log N}+z),$$

which gives

$$\Delta(t)\le KM_T^2\sqrt{\varepsilon T}(\sqrt{D+\log N}+z)e^{KM_T^2 t}.$$

Since we choose $\varepsilon\le 1/[K_0(D+\log N+z^2)e^{K_0(1+T)^3}]$, we have

$$\Delta(T_a\wedge T_\Delta\wedge T)\le M_T^2\sqrt{T\varepsilon}(\sqrt{D+\log N}+z)e^{KM_T^2 T}\le 1/2.$$

Moreover, for $t\le T_a\wedge T_\Delta\wedge T$, we have

$$\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}|a_i^k|\le\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}|\tilde{a}_i^k|+\Delta(t)\le K(1+T)+1/2<2K(1+T).$$

This means that the stopping times $T_a,T_\Delta\ge T$. Hence, for any $t\le T$, we have

$$\begin{aligned}
\Delta(t)\le&M_T^2\sqrt{\varepsilon T}(\sqrt{D+\log N}+z)e^{KM_T^2 t},\\
\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}|a_i^k|\le&2K(1+t).
\end{aligned}$$

Note all these happens when event E_{good} happens. Hence, the probability such that the events above happens is at least $1-e^{-z^2}$. Finally, by Lemma 8, we have the desired bound on R_N. This concludes the proof. □

D Proof of Theorem 2 part (A)

The proof follows the same scheme as for Theorem 1 (A) and we will limit ourselves to describing the differences.

Throughout this section, the assumptions A1-A6 of Theorem 2 are understood to hold. For the sake of simplicity we will write the proof under the following restriction:

R1. The coefficients $a_i\equiv 1$.

R2. The step size function $\xi(t)\equiv 1/2$.

The proof for a general function $\xi(t)$ is obtained by a straightforward adaptation.

For the reader's convenience, we copy here the limiting PDE:

$$\begin{aligned}
\partial_t\rho_t=&2\xi(t)\nabla\cdot[\rho(\boldsymbol{\theta})\nabla\Psi_\lambda(\boldsymbol{\theta};\rho_t)]+2\xi(t)\tau D^{-1}\Delta_{\boldsymbol{\theta}}\rho_t,\\
\Psi_\lambda(\boldsymbol{\theta};\rho)=&V(\boldsymbol{\theta})+\int U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')+\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\,.
\end{aligned}$$

We will consider four different coupled dynamics with same initialization $(\bar{\boldsymbol{\theta}}_i^0)_{i\le N} \sim_{iid} \rho_0$ and stochastic term. We will denote $\{\boldsymbol{W}_i(s)\}_{s\ge 0}$ for $i \in [N]$ independent D-dimensional Brownian motions. The integral equations and summation forms of the four dynamics are as follows:

- The *nonlinear dynamics (ND)*:
$$\bar{\boldsymbol{\theta}}_i^t = \bar{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s;\rho_s)\mathrm{d}s + \int_0^t \sqrt{2\xi(s)\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s), \tag{45}$$
where we denoted $\boldsymbol{G}(\boldsymbol{\theta};\rho) = -\nabla\Psi_\lambda(\boldsymbol{\theta};\rho) = -\lambda\boldsymbol{\theta} - \nabla V(\boldsymbol{\theta}) - \int \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')$, and $\bar{\boldsymbol{\theta}} \sim \rho_0$ i.i.d.
- The *particle dynamics (PD)*:
$$\underline{\boldsymbol{\theta}}_i^t = \underline{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s;\hat{\rho}_s^{(N)})\mathrm{d}s + \int_0^t \sqrt{2\xi(s)\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s), \tag{46}$$
where $\underline{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.
- The *gradient descent (GD)*:
$$\tilde{\boldsymbol{\theta}}_i^k = \tilde{\boldsymbol{\theta}}_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l;\tilde{\rho}_l^{(N)}) + \int_0^{k\varepsilon} \sqrt{2\xi([s])\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s),$$
where $\tilde{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.
- The *stochastic gradient descent (SGD)*:
$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon\sum_{l=0}^{k-1}\xi(l\varepsilon)\boldsymbol{F}_i(\boldsymbol{\theta}^l;\boldsymbol{z}_{l+1}) + \int_0^{k\varepsilon} \sqrt{2\xi([s])\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s),$$
where we denoted $\boldsymbol{F}_i(\boldsymbol{\theta}^k;\boldsymbol{z}_{k+1}) = -\lambda\boldsymbol{\theta}_i^k + (y_{k+1} - \hat{y}_{k+1})\nabla_{\boldsymbol{\theta}_i}\sigma_\star(\boldsymbol{x}_{k+1};\boldsymbol{\theta}_i^k)$, and $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.

By Proposition 9, 10, 11, 12, there exists constants K and K_0, such that with probability at least $1 - e^{-z^2}$, we have
$$\begin{aligned}
\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le& Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z],\\
\sup_{t\in[0,T]} |R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \le& Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z],\\
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le& Ke^{KT}[\sqrt{\log(N(T/\varepsilon \vee 1))} + z]\sqrt{\varepsilon},\\
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)| \le& Ke^{KT}\sqrt{T\varepsilon}[\sqrt{D + \log N} + z].
\end{aligned}$$
Combining these inequalities gives the conclusion of Theorem 2 (A). In the following subsections, we prove all the above interpolation bounds, under the setting of Theorem 2 (A).

D.1 Technical lemmas

Define the maximum and the average of the norm of the initialization:
$$\Theta_\infty \equiv \max_{i\le N}\|\boldsymbol{\theta}_i^0\|_2, \qquad \Theta_1 \equiv \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{\theta}_i^0\|_2.$$
Similarly define the following bounds on the Brownian noise:
$$\overline{\boldsymbol{W}}_i(t) \equiv \sqrt{\frac{\tau}{D}}\boldsymbol{W}_i(t) = \int_0^t \sqrt{\frac{\tau}{D}}\mathrm{d}\boldsymbol{W}_i(s), \qquad W_\infty \equiv \max_{i\le N}\sup_{t\le T}\|\overline{\boldsymbol{W}}_i(t)\|_2, \qquad W_1 \equiv \sup_{t\le T}\frac{1}{N}\sum_{i=1}^N\|\overline{\boldsymbol{W}}_i(t)\|_2.$$

Lemma 13. *There exists a constant K such that:*

$$\mathbb{P}\Big(\max(\Theta_\infty, W_\infty) \le K(1+T)\big[\sqrt{\log N} + z\big]\Big) \ge 1 - e^{-z^2},$$
$$\mathbb{P}\Big(\max(\Theta_1, W_1) \le K(1+T)\big[1 + z\big]\Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 13. Let us first consider a generic D-dimensional K^2-sub-Gaussian random vector $\boldsymbol{X}$, we have:

$$\mathbb{E}_{\boldsymbol{X}}[\exp\{\mu\|\boldsymbol{X}\|_2^2/2\}] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{G}}[\exp\{\sqrt{\mu}\langle\boldsymbol{G},\boldsymbol{X}\rangle\}] \le \mathbb{E}_{\boldsymbol{G}}[\exp\{\mu K^2\|\boldsymbol{G}\|_2^2/2\}] = (1-\mu K^2/2)^{-D/2},$$

where $\boldsymbol{G} \sim N(0, \boldsymbol{I}_D)$. Recall that $\{(a_i^0, \boldsymbol{w}_i^0)\}_{i\in[N]} \sim_{iid} \rho_0$ with $\boldsymbol{w}_i^0$ being a K^2/D-sub-Gaussian vector in $\mathbb{R}^{D-1}$ independent of a_i^0. Using the above inequality, we get

$$\mathbb{P}\Big(\|\boldsymbol{w}_i^0\|_2 \ge u\Big) \le \mathbb{E}[\exp\{\mu\|\boldsymbol{w}_i^0\|_2^2/2\}]/\exp\{\mu u^2/2\} \le (1-\mu K^2/D)^{-(D-1)/2}\exp\{-\mu u^2/2\}.$$

Taking the union bound over $i \in [N]$, and noting that $|a_i^0| \le K$, we get:

$$\mathbb{P}\Big(\max_{i\in[N]}\|\boldsymbol{\theta}_i^0\|_2 \ge u + K\Big) \le (1-\mu K^2/D)^{-(D-1)/2}\exp\{-\mu u^2/2 + \log N\}.$$

Taking $\mu = D/(2K^2)$ and $u = 2K[\sqrt{D + \log N} + z]/\sqrt{D}$, we get:

$$\mathbb{P}\Big(\Theta_\infty \ge 2K\left[\sqrt{D+\log N} + z\right]/\sqrt{D}\Big) \le e^{-z^2}.$$

Let us now consider the average over $i \in [N]$ of the $\|\boldsymbol{w}_i^0\|_2$, which are independent, we get:

$$\mathbb{P}\Big(N^{-1}\sum_{i=1}^N \|\boldsymbol{w}_i^0\|_2 \ge u\Big) \le \mathbb{P}\Big(\sum_{i=1}^N \|\boldsymbol{w}_i^0\|_2^2 \ge Nu^2\Big) \le (1-\mu K^2/D)^{-N(D-1)/2}\exp\{-\mu N u^2/2\}.$$

Taking $\mu = D/(2K^2)$ and $u = 2K\,[1+z]$, noting $(1/N)\sum_{i=1}^N |a_i^0| \le K$, we get:

$$\mathbb{P}(\Theta_1 \ge 2K\,[1+z]) \le e^{-z^2}.$$

Similarly, we consider $\overline{\boldsymbol{W}}_i(t) \equiv \sqrt{\tau/D}\boldsymbol{W}_i(t)$ which is a D-dimensional Gaussian random variable with variance $\mathrm{Var}(\overline{W}_i^j(t)) = \int_0^t(\tau/D)\mathrm{d}s = \tau t/D$. We note that $\exp\{\mu\|\boldsymbol{W}_i(t)\|_2^2\}$ is a sub-martingale and by Doob's martingale inequality, we have:

$$\mathbb{P}\Big(\sup_{t\in[0,T]}\|\boldsymbol{W}_i(t)\|_2 \ge u\Big) \le \mathbb{E}[\exp\{\mu\|\boldsymbol{W}_i(T)\|_2^2/2\}]/\exp\{\mu u^2/2\} \le (1-2\mu(\tau T/D))^{-D/2}\exp\{-\mu u^2/2\}.$$

Taking the union bound over $i \in [N]$ gives:

$$\mathbb{P}\Big(\max_{i\in[N]}\sup_{t\in[0,T]}\|\boldsymbol{W}_i(t)\|_2 \ge u\Big) \le (1-2\mu\tau T/D)^{-D/2}\exp\{-\mu u^2/2 + \log N\}.$$

Taking $\mu = D/(4\tau T)$ and $u = 4\sqrt{T\tau}[\sqrt{D+\log N} + z]/\sqrt{D}$, we get:

$$\mathbb{P}\Big(W_\infty \ge 4\sqrt{T\tau}[\sqrt{D+\log N} + z]/\sqrt{D}\Big) \le e^{-z^2}.$$

We can consider the average over $i \in [N]$ of the preceding bound, by noticing that:

$$\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{W}_i(t)\|_2 \le \Big(\frac{1}{N}\sum_{i=1}^N\|\boldsymbol{W}_i(t)\|_2^2\Big)^{1/2} \equiv \|\boldsymbol{W}(t)\|_2/\sqrt{N},$$

where $\boldsymbol{W}(t)$ is a ND-dimensional Brownian motion. We can therefore apply Doob's martingale inequality to the sub-martingale $\exp\{\mu\|\boldsymbol{W}(t)\|_2^2\}$. We have

$$\begin{aligned}\mathbb{P}\Big(\sup_{t\in[0,T]}\|\boldsymbol{W}(t)\|_2 \geq \sqrt{N}u\Big) &\leq \mathbb{E}[\exp\{\mu\|\boldsymbol{W}(t)\|_2^2/2\}]/\exp\{N\mu u^2/2\}\\ &\leq (1-2\mu T\tau/D)^{-ND/2}\exp\{-N\mu u^2/2\}.\end{aligned}$$

Taking $\mu = D/(4\tau T)$ and $u = 4\sqrt{T\tau}[1+z]$, we get:

$$\mathbb{P}\Big(W_1 \geq 4\sqrt{T\tau}[1+z]\Big) \leq e^{-z^2}.$$

This proves the lemma. □

The two following lemmas are modified from [MMN18, Section 7.2, Lemma 7.5].

Lemma 14. *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{i\leq N}\sup_{k\in[0,T/\eta]\cap\mathbb{N}}\sup_{u\in[0,\eta]}\|\bar{\boldsymbol{\theta}}_i^{k\eta+u}-\bar{\boldsymbol{\theta}}_i^{k\eta}\|_2 \leq Ke^{KT}\Big[\sqrt{\log\left(N(T/\eta\vee 1)\right)}+z\Big]\sqrt{\eta}\Big) \leq 1-e^{-z^2},$$

and for any $t,h\geq 0, t+h\leq T$,

$$W_2(\rho_t,\rho_{t+h}) \leq (\mathbb{E}[\|\bar{\boldsymbol{\theta}}_i^t-\bar{\boldsymbol{\theta}}_i^{t+h}\|_2^2])^{1/2} \leq Ke^{KT}\sqrt{h}.$$

Proof of Lemma 14. Define $\Delta_i(t)\equiv\sup_{s\leq t}\|\bar{\boldsymbol{\theta}}_i^t\|_2$. From Eq. (45),

$$\|\bar{\boldsymbol{\theta}}_i^t\|_2 \leq K\int_0^t \|\bar{\boldsymbol{\theta}}_i^s\|_2\mathrm{d}s + \Theta_\infty + W_\infty,$$

which gives, after applying Gronwall's inequality with the bounds of Lemma 13:

$$\mathbb{P}\Big(\Delta_i(t) \leq Ke^{KT}\big[\sqrt{\log N}+z\big]\Big) \geq 1-e^{-z^2}. \tag{47}$$

Consider $\Delta_i(h;k,\varepsilon) = \sup_{0\leq u\leq\varepsilon}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2$. We have

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 &\leq \Big\|\int_{k\varepsilon}^{k\varepsilon+u}\xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s;\rho_s)\mathrm{d}s\Big\|_2 + \|\overline{\boldsymbol{W}}_{i,k}(u)\|_2\\ &\leq Kh\sup_{s\leq T}\big[\lambda\|\bar{\boldsymbol{\theta}}_i^s\|_2+1\big] + \|\overline{\boldsymbol{W}}_{i,k}(u)\|_2,\end{aligned}$$

where we defined $\overline{\boldsymbol{W}}_{i,k}(u)\equiv\int_{k\varepsilon}^{k\varepsilon+u}\sqrt{\tau/D}\mathrm{d}\boldsymbol{W}_i(s)$. By a similar computation as in Lemma 13, we have

$$\mathbb{P}\Big(\max_{i\leq N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\leq u\leq\varepsilon}\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2 \geq 4\sqrt{K\varepsilon}\Big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\Big]\Big) \leq e^{-z^2}.$$

Combining this bound and Eq. (47) yields:

$$\mathbb{P}\Big(\max_{i\leq N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\Delta_i(h;k,\varepsilon) \leq Ke^{KT}\big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\big]\sqrt{\varepsilon}\Big) \geq 1-e^{-z^2}. \tag{48}$$

We now bound $W_2(\rho_t,\rho_{t+h})$:

$$W_2(\rho_t,\rho_{t+h})^2 \leq \mathbb{E}[\|\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t+h}\|_2^2] = \int_0^\infty \mathbb{P}(\|\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t+h}\|_2^2 \geq u)\mathrm{d}u.$$

Using Eq. (48), we have (where we removed the union bound on $i\in[N]$ and $k\in[0,T/\varepsilon]\cap\mathbb{N}$)

$$\mathbb{P}\Big(\|\bar{\boldsymbol{\theta}}_i^{t+h}-\bar{\boldsymbol{\theta}}_i^t\|_2 \geq Ke^{KT}[1+z]\sqrt{h}\Big) \leq e^{-z^2}.$$

Integrating this upper bound on the probability yields the desired inequality. □

The exact same proof shows a similar lemma for the particle dynamics.

Lemma 15. *There exists a constant K, such that*

$$\mathbb{P}\Big(\max_{i\leq N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{u\in[0,\varepsilon]}\|\underline{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\underline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 \leq Ke^{KT}\Big[\sqrt{\log\left(N(T/\varepsilon\vee 1)\right)}+z\Big]\sqrt{\varepsilon}\Big) \leq 1-e^{-z^2}.$$

D.2 Bound between PDE and nonlinear dynamics

Proposition 9 (PDE-ND)**.** *There exists a constant K such that with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z].$$

We will follow the same decomposition as in the proof of Proposition 1. The proof of term II only depend on the upper bound on the potential U and still apply. The term I bound follow from a similar proof as lemma 5.

Lemma 16 (Term I bound)**.** *There exists K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \le Ke^{KT}[\sqrt{\log(NT)} + z]/\sqrt{N}\Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 16. Applying McDiarmid's inequality, we have

$$\mathbb{P}\Big(|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \ge \delta\Big) \le \exp\{-N\delta^2/K\}.$$

Furthermore we have the following increment bound for $t, h \ge 0$:

$$\begin{aligned}
&\Big||R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h})| - |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|\Big| \\
\le&\Big|R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)\Big| + \Big|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)\Big| \\
\le&K\Big[\sup_{i\in[N]} \|\bar{\boldsymbol{\theta}}_i^{t+h} - \bar{\boldsymbol{\theta}}_i^t\|_2 + \mathbb{E}[\|\bar{\boldsymbol{\theta}}_j^{t+h} - \bar{\boldsymbol{\theta}}_j^t\|_2]\Big].
\end{aligned}$$

Using Lemma 14, we get

$$\sup_{k\in[0,T/\eta]\cap\mathbb{N}}\ \sup_{u\in[0,\eta]} \Big||R_N(\bar{\boldsymbol{\theta}}^{k\eta+u}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta+u})| - |R_N(\bar{\boldsymbol{\theta}}^{k\eta}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta})|\Big| \le Ke^{KT}\Big[\sqrt{\log N(T/\eta \vee 1)} + z\Big]\sqrt{\eta},$$

with probability at least $1 - e^{-z^2}$. Hence taking an union bound over $s \in \eta\{0, 1, \dots, \lfloor T/\eta\rfloor\}$ and bounding the variation inside the grid intervals, we have

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \ge \delta + Ke^{KT}\Big[\sqrt{\log N(T/\eta \vee 1)} + z\Big]\sqrt{\eta}\Big) \le (T/\eta)\exp\{-N\delta^2/K\} + e^{-z^2}.$$

Taking $\eta = 1/N$ and $\delta = K[\sqrt{\log(NT)} + z]/\sqrt{N}$ concludes the proof. □

D.3 Bound between nonlinear dynamics and particle dynamics

Proposition 10 (ND-PD)**.** *There exists a constant K, such that with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{t\in[0,T]} \max_{i\in[N]} \|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \le Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z], \tag{49}$$

$$\sup_{t\in[0,T]} |R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \le Ke^{KT}\frac{1}{\sqrt{N}}[\sqrt{\log(NT)} + z]. \tag{50}$$

Proof of Proposition 10. The nonlinear dynamics and the particle dynamics are coupled by using the same Brownian motion, and the noise term cancel out. By the same calculation as in Proposition 2, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\underline{\boldsymbol{\theta}}_i^t - \bar{\boldsymbol{\theta}}_i^t\|_2 \le K \cdot \max_{j\in[N]} \|\underline{\boldsymbol{\theta}}_j^t - \bar{\boldsymbol{\theta}}_j^t\|_2 + K/N + I_i^t, \tag{51}$$

where

$$I_i^t = \Big\|\frac{1}{N}\sum_{j\neq i}\Big[\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big]\Big\|_2.$$

Now we would like to prove a uniform bound for I_i^t for $i \in [N]$ and $t \in [0, T]$.

Lemma 17. *There exists a constant K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]}\max_{i\in[N]} I_i^t \le Ke^{KT}[\sqrt{\log(NT)}+z]/\sqrt{N}\Big) \ge 1-e^{-z^2}.$$

Proof of Lemma 17. Denoting $\boldsymbol{X}_i^t = \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})$, we have $\mathbb{E}[\boldsymbol{X}_i^t|\bar{\boldsymbol{\theta}}_i^t] = 0$, (where the expectation is taken with respect to $\bar{\boldsymbol{\theta}}_j^0 \sim \rho_0$ and $\{\boldsymbol{W}_j(s)\}_{s\ge 0}$ for $j \neq i$), and $\|\boldsymbol{X}_i^t\|_2 \le 2K$ (by assumption that $\|\nabla U\|_2 \le K$). By Lemma 30, we have for any fixed $i \in [N]$ and $t \in [0,T]$,

$$\mathbb{P}\Big(I_i^t \ge K(\sqrt{1/N}+\delta)\Big) = \mathbb{E}\Big[\mathbb{P}\Big(I_i^t \ge K(\sqrt{1/N}+\delta)|\bar{\boldsymbol{\theta}}_i^t\Big)\Big] \le \exp\{-N\delta^2\}.$$

We then bound the variation of I_i^s over an interval $[t, t+h]$, with $t, h \ge 0$:

$$\begin{aligned}|I_i^{t+h} - I_i^t| \le& \frac{1}{N}\sum_{j\le i}\Big\|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\Big\|_2 \\ &+ \Big\|\int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \boldsymbol{\theta})\rho_{t+h}(\mathrm{d}\boldsymbol{\theta}) - \int \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta})\Big\|_2 \\ \le& K\Big[\sup_{i\le N}\|\bar{\boldsymbol{\theta}}_i^{t+h} - \bar{\boldsymbol{\theta}}_i^t\|_2 + \mathbb{E}[\|\bar{\boldsymbol{\theta}}_j^{t+h} - \bar{\boldsymbol{\theta}}_j^t\|_2]\Big].\end{aligned}$$

By Lemma 14, there exists K such that, we have

$$\mathbb{P}\Big(\sup_{k\in[0,T/\eta]\cap\mathbb{N}}\sup_{u\in[0,\eta]}|I_i^{k\eta+u} - I_i^{k\eta}| \le Ke^{KT}\Big[\sqrt{\log\left(N(T/\eta\vee 1)\right)}+z\Big]\sqrt{\eta}\Big) \ge 1-e^{-z^2}.$$

Taking an union bound for $i \in [N]$ and $s \in \eta\{0, 1, \dots, \lfloor T/\eta\rfloor\}$ and bounding the variation inside the grid intervals, we have

$$\mathbb{P}\Big(\sup_{t\in[0,T]}\max_{i\in[N]} I_i^t \ge K(\sqrt{1/N}+\delta) + Ke^{KT}\left[\sqrt{\log N(T/\eta\vee 1)}\right]\sqrt{\eta}\Big) \le (NT/\eta)\exp\{-N\delta^2\} + e^{-z^2}.$$

Taking $\eta = 1/N$, and $\delta = K[\sqrt{\log(NT)}+z]/\sqrt{N}$, we get the desired result. □

Denote $\delta_N(T,z) = Ke^{KT}[\sqrt{\log(NT)}+z]/\sqrt{N}$ and

$$\Delta(t) = \sup_{s\in[t]}\max_{i\in[N]}\|\underline{\boldsymbol{\theta}}_i^s - \bar{\boldsymbol{\theta}}_i^s\|_2.$$

With probability at least $1-e^{-z^2}$, we have

$$\Delta'(t) \le K\cdot\Delta(t) + \delta_N(T,z),$$

which, after applying Gronwall's inequality, concludes the proof. □

D.4 Bound between particle dynamic and GD

Proposition 11 (PD-GD)**.** *There exists a constant K such that with probability at least $1-e^{-z^2}$, we have*

$$\begin{aligned}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\max_{i\le N}\|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le& Ke^{KT}\Big[\sqrt{\log\left(N(T/\varepsilon\vee 1)\right)}+z\Big]\sqrt{\varepsilon}, \\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}|R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le& Ke^{KT}[\sqrt{\log(N(T/\varepsilon\vee 1))}+z]\sqrt{\varepsilon}.\end{aligned}$$

Proof of Proposition 11. For $k \in \mathbb{N}$ and $t = k\varepsilon$,

$$\|\underline{\boldsymbol{\theta}}_i^t - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2\mathrm{d}s + \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2\mathrm{d}s.$$

We have by Lemma 15

$$\int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2 \mathrm{d}s \leq KT \sup_{s\in[0,T]} \max_{i\in[N]} \|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2$$

$$\leq TKe^{KT}\left[\sqrt{\log\left(N(T/\varepsilon \vee 1)\right)} + z\right]\sqrt{\varepsilon},$$

with probability at least $1 - e^{-z^2}$. Denote $\delta(N, T, z) = TKe^{KT}\left[\sqrt{\log\left(N(T/\varepsilon \vee 1)\right)} + z\right]\sqrt{\varepsilon}$ and

$$\Delta(t) \equiv \sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}} \max_{i\leq N} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2.$$

With probability at least $1 - e^{-z^2}$, we get

$$\Delta(t) \leq K\int_0^t \Delta(s)\mathrm{d}s + \delta(N, T, z).$$

Applying Gronwall's inequality concludes the proof. □

D.5 Bound between GD and SGD

Proposition 12 (GD-SGD)**.** *There exists a constant K such that, with probability at least $1 - e^{-z^2}$, we have*

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \max_{i\in[N]} \|\tilde{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_i^k\|_2 \leq Ke^{KT}\sqrt{T\varepsilon}[\sqrt{D + \log N} + z],$$

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)| \leq Ke^{KT}\sqrt{T\varepsilon}[\sqrt{D + \log N} + z].$$

Proof of Proposition 12. We coupled the noise between the GD and SGD such that the noise cancels out. Noticing furthermore that the regularization term does not depend on $\boldsymbol{z}_k$ and vanishes in the martingale difference $\boldsymbol{Z}_i^l \equiv \boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) - \mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1})|\mathcal{F}_l]$, where $\mathcal{F}_k = \sigma((\boldsymbol{\theta}_i^0)_{i\in[N]}, (\boldsymbol{z}_l)_{l=0}^k, (\boldsymbol{W}_i(s))_{s\leq k\varepsilon})$. Therefore the same proof as Proposition 4 applies here. □

E Proof of Theorem 2 part (B)

We remind the notations used in the proof of Theorem 1 (B): for $\boldsymbol{\theta} = (a, \boldsymbol{w})$ and $\boldsymbol{\theta}' = (a', \boldsymbol{w}')$,

$$\begin{aligned}
v(\boldsymbol{w}) =& -\mathbb{E}_{y,\boldsymbol{x}}[y\sigma(\boldsymbol{x}; \boldsymbol{w})],\\
u(\boldsymbol{w}, \boldsymbol{w}') =& \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}; \boldsymbol{w})\sigma(\boldsymbol{x}; \boldsymbol{w}')],\\
V(\boldsymbol{\theta}) =& a \cdot v(\boldsymbol{w}),\\
U(\boldsymbol{\theta}, \boldsymbol{\theta}') =& aa' \cdot u(\boldsymbol{w}, \boldsymbol{w}')\\
\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) =& (v(\boldsymbol{w}), a\nabla_{\boldsymbol{w}} v(\boldsymbol{w})),\\
\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}, \boldsymbol{\theta}') =& (a' \cdot u(\boldsymbol{w}, \boldsymbol{w}'), aa' \cdot \nabla_{\boldsymbol{w}} u(\boldsymbol{w}, \boldsymbol{w}')).
\end{aligned}$$

For convenience, we copy here the properties of the potentials $V(\boldsymbol{\theta})$ and $U(\boldsymbol{\theta}, \boldsymbol{\theta}')$ listed in Lemma 8. Denoting $\boldsymbol{\theta} = (a, \boldsymbol{w})$, $\boldsymbol{\theta}_1 = (a_1, \boldsymbol{w}_1)$ and $\boldsymbol{\theta}_2 = (a_2, \boldsymbol{w}_2)$. We have

$$\begin{aligned}
|V(\boldsymbol{\theta})|, \|\nabla V(\boldsymbol{\theta})\|_2 \leq& K(1 + |a|),\\
\|\nabla V(\boldsymbol{\theta}_1) - \nabla V(\boldsymbol{\theta}_2)\|_2 \leq& K \cdot [1 + \min\{|a_1|, |a_2|\}] \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,\\
|U(\boldsymbol{\theta}, \boldsymbol{\theta}')|, \|\nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 \leq& K(1 + |a|)(1 + |a'|),\\
\|\nabla_{(1,2)} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}) - \nabla_{(1,2)} U(\boldsymbol{\theta}_2, \boldsymbol{\theta})\|_2 \leq& K(1 + |a|) \cdot [1 + \min\{|a_1|, |a_2|\}] \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.
\end{aligned}$$

Throughout this section, the assumptions A1 - A6 are understood to hold. For the sake of simplicity we will write the proof under the following restriction:

R1. The step size function $\xi(t) \equiv 1/2$.

The proof for a general function $\xi(t)$ is obtained by a straightforward adaptation.

We recall the form of the limiting PDE:

$$\begin{aligned}\partial_t \rho_t =& 2\xi(t)\nabla \cdot [\rho(\boldsymbol{\theta})\nabla\Psi_\lambda(\boldsymbol{\theta};\rho_t)] + 2\xi(t)\tau D^{-1}\Delta_{\boldsymbol{\theta}}\rho_t,\\ \Psi_\lambda(\boldsymbol{\theta};\rho) =& V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}') + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 .\end{aligned}$$

We will consider four different coupled dynamics with same initialization $(\bar{\boldsymbol{\theta}}_i^0)_{i\le N} \sim_{iid} \rho_0$. The integral equations and summation form are as follows:

- The *nonlinear dynamics (ND)*:

$$\bar{\boldsymbol{\theta}}_i^t = \bar{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s;\rho_s)\mathrm{d}s + \int_0^t \sqrt{2\xi(s)\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s), \tag{52}$$

  where we denoted $\boldsymbol{G}(\boldsymbol{\theta};\rho) = -\nabla\Psi_\lambda(\boldsymbol{\theta};\rho) = -\lambda\boldsymbol{\theta} - \nabla V(\boldsymbol{\theta}) - \int \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')$, and $\bar{\boldsymbol{\theta}} \sim \rho_0$ iid.

- The *particle dynamics (PD)*:

$$\underline{\boldsymbol{\theta}}_i^t = \underline{\boldsymbol{\theta}}_i^0 + 2\int_0^t \xi(s)\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s;\hat{\rho}_s^{(N)})\mathrm{d}s + \int_0^t \sqrt{2\xi(s)\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s), \tag{53}$$

  where $\underline{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.

- The *gradient descent (GD)*:

$$\tilde{\boldsymbol{\theta}}_i^k = \tilde{\boldsymbol{\theta}}_i^0 + 2\varepsilon\sum_{l=0}^{k-1} \xi(l\varepsilon)\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l;\tilde{\rho}_l^{(N)}) + \int_0^{k\varepsilon} \sqrt{2\xi([s])\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s),$$

  where $\tilde{\boldsymbol{\theta}}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.

- The *stochastic gradient descent (SGD)*:

$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon\sum_{l=0}^{k-1} \xi(l\varepsilon)\boldsymbol{F}_i(\boldsymbol{\theta}^l;\boldsymbol{z}_{l+1}) + \int_0^{k\varepsilon} \sqrt{2\xi([s])\tau D^{-1}}\mathrm{d}\boldsymbol{W}_i(s),$$

  where we defined $\boldsymbol{F}_i(\boldsymbol{\theta}^k;\boldsymbol{z}_{k+1}) = -\lambda\boldsymbol{\theta}_i^k + (y_{k+1} - \hat{y}_{k+1})\nabla_{\boldsymbol{\theta}_i}\sigma_\star(\boldsymbol{x}_{k+1};\boldsymbol{\theta}_i^k)$, and $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}_i^0$.

By Proposition 13, 14, 15, 16, there exists constants K, such that with probability at least $1 - e^{-z^2}$, we have

$$\begin{aligned}\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - R(\rho_t)| \le& Ke^{KT}[\log^{3/2}(NT) + z^3]/\sqrt{N},\\ \sup_{t\in[0,T]} |R_N(\underline{\boldsymbol{\theta}}^t) - R_N(\bar{\boldsymbol{\theta}}^t)| \le& Ke^{e^{KT}[\sqrt{\log N}+z^2]}[\sqrt{D\log N} + \log^{3/2}(NT) + z^5]/\sqrt{N},\\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le& Ke^{e^{KT}[\sqrt{\log N}+z^2]}[\log(N(T/\varepsilon\vee 1)) + z^6]\sqrt{\varepsilon},\\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k)| \le& Ke^{e^{KT}[\sqrt{\log N}+z^2]}[\sqrt{D}\log N + \log^{3/2} N + z^5]\sqrt{\varepsilon}.\end{aligned}$$

Combining these inequalities gives the conclusion of Theorem 2 (B). In the following subsections, we prove all the above interpolation bounds, under the setting of Theorem 2 (B).

E.1 Technical lemmas

The bounds on the potentials U, V, and their derivatives scales with the coefficients a, which can be arbitrarily large with non-zero probability due to the Brownian noise. In our analysis we will need to keep track of the maximum and the first moment of $|a|$ for each of the different dynamics. In this section we will show that there exists high probability bounds along the trajectories.

We recall the following notations introduced in Appendix Section D.1,

$$\Theta_\infty \equiv \max_{i\le N} \|\boldsymbol{\theta}_i^0\|_2, \qquad \Theta_1 \equiv \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{\theta}_i^0\|_2.$$

and on the Brownian motion,

$$\overline{\boldsymbol{W}}_i(t) \equiv \sqrt{\frac{\tau}{D}}\boldsymbol{W}_i(t) = \int_0^t \sqrt{\frac{\tau}{D}}\mathrm{d}\boldsymbol{W}_i(s), \qquad W_\infty \equiv \max_{i\le N}\sup_{t\le T}\|\overline{\boldsymbol{W}}_i(t)\|_2, \qquad W_1 \equiv \sup_{t\le T}\frac{1}{N}\sum_{i=1}^N \|\overline{\boldsymbol{W}}_i(t)\|_2.$$

For convenience, we recall here the bounds derived in Lemma 13:

$$\mathbb{P}\Big(\max(\Theta_\infty, W_\infty) \le K(1+T)\big[\sqrt{\log N} + z\big]\Big) \ge 1 - e^{-z^2},$$
$$\mathbb{P}\Big(\max(\Theta_1, W_1) \le K(1+T)\big[1 + z\big]\Big) \ge 1 - e^{-z^2}.$$

In the following lemma, and throughout the proof, we will denote $\bar{\boldsymbol{a}}^t \equiv (\bar{a}_1^t, \ldots, \bar{a}_N^t) \in \mathbb{R}^N$ the vector of the $\bar{a}_i^t$ variables of the nonlinear dynamics. Similarly we will denote $\underline{\boldsymbol{a}}^t$, $\tilde{\boldsymbol{a}}^k$ and $\boldsymbol{a}^k$ the vectors of variable a associated to the particle dynamics, gradient descent and stochastic gradient descent. We will furthermore use $\|\boldsymbol{a}\|_1$, $\|\boldsymbol{a}\|_\infty$ to denote the ℓ_1 and ℓ_∞ norms of the coefficients vector.

Lemma 18. *There exists a constant K, such that denoting $M_2(t) = Ke^{Kt}$, we have*

$$\sup_{s\in[0,t]} \int a^2 \rho_s(\mathrm{d}a) \le M_2(t).$$

Furthermore, letting $(\bar{a}^t, \bar{\boldsymbol{w}}^t) \sim \rho_t$, then $\bar{a}^t$ is $M_2(t)$-sub-Gaussian.

Proof of Lemma 18. Denote $A(t) = \int a^2 \rho_t(\mathrm{d}a)/2$. For simplicity, we will directly take the derivative of this function. This computation can be made rigorous by considering smooth approximation of a truncated squared function, with bounded second derivative, and using the definition of weak solution. We get:

$$\frac{\mathrm{d}}{\mathrm{d}t}A(t) = \tau/D - \int \Big[\lambda a^2 + a\cdot v(\boldsymbol{w}) + a\cdot \int a' u(\boldsymbol{w}, \boldsymbol{w}')\rho_t(\mathrm{d}\boldsymbol{\theta}')\Big]\rho_t(\mathrm{d}\boldsymbol{\theta}) \le K + KA(t),$$

which implies by applying Gronwall's lemma we have

$$\sup_{s\in[0,t]} \int a^2 \rho_s(\mathrm{d}a) \le Ke^{Kt}.$$

Let us consider the nonlinear dynamics for the variable $\bar{a}^t \sim \rho_t$:

$$\mathrm{d}\bar{a}^t = -\lambda \bar{a}^t \mathrm{d}t + \Big[-v(\bar{\boldsymbol{w}}^t) - \int a' u(\bar{\boldsymbol{w}}^t, \boldsymbol{w}')\rho_t(\mathrm{d}\boldsymbol{\theta}')\Big]\mathrm{d}t + \sqrt{\frac{\tau}{D}}\mathrm{d}W^a(t).$$

Denote $u_\lambda(t) = \bar{a}^t e^{\lambda t}$ and

$$K(\bar{\boldsymbol{w}}^t, \rho_t) = -v(\bar{\boldsymbol{w}}^t) - \int a' u(\bar{\boldsymbol{w}}^t, \boldsymbol{w}')\rho_t(\mathrm{d}\boldsymbol{\theta}'),$$

we get

$$\mathrm{d}u_\lambda(t) = e^{\lambda t}K(\bar{\boldsymbol{w}}^s, \rho_t)\mathrm{d}t + e^{\lambda t}\sqrt{\frac{\tau}{D}}\mathrm{d}W^a(t),$$

and in integration form we have

$$u_\lambda(t) = u_\lambda(0) + \int_0^t e^{\lambda s} K(\bar{\boldsymbol{w}}^s, \rho_s)\mathrm{d}s + \int_0^t e^{\lambda s}\sqrt{\frac{\tau}{D}}\mathrm{d}W^a(s).$$

We deduce that we can rewrite $\bar{a}^t \sim \rho_t$ as the sum of three random variables:

$$\bar{a}^t = \underbrace{e^{-\lambda t}a^0}_{\Gamma_1} + \underbrace{\int_0^t e^{-\lambda(t-s)}K(\bar{\boldsymbol{w}}^s, \rho_s)\mathrm{d}s}_{\Gamma_2} + \underbrace{\int_0^t e^{-\lambda(t-s)}\sqrt{\frac{\tau}{D}}\mathrm{d}W^a(s)}_{\Gamma_3}.$$

By assumption a_0 is K-bounded, and thus Γ_1 is K^2-sub-Gaussian. By the boundedness of u and v, Cauchy Schwartz inequality, and by $A(t) \leq M_2(t)$, then for $s \leq t$, we have $|K(\bar{\boldsymbol{w}}^s, \rho_s)| \leq Ke^{Kt}$, hence the random variable Γ_2 is Ke^{Kt}-bounded and thus Ke^{Kt}-sub-Gaussian. The random variable Γ_3 is a Gaussian random variable with variance

$$\mathrm{Var}(\Gamma_3) = \int_0^t e^{-2\lambda(t-s)}\frac{\tau}{D}\mathrm{d}s \leq Kt.$$

We deduce that $\bar{a}^t$ is the sum of three (dependent) sub-Gaussian random variables with parameters K^2, Ke^{Kt}, Kt respectively, and therefore the sum $\bar{a}^t$ is Ke^{Kt}-sub-Gaussian. □

Lemma 19. *There exists a constant K such that with probability at least $1 - e^{-z^2}$, we have*

$$\max\Big(\sup_{t\in[0,T]}\|\bar{\boldsymbol{a}}^t\|_1, \sup_{t\in[0,T]}\|\underline{\boldsymbol{a}}^t\|_1, \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\|\tilde{\boldsymbol{a}}^k\|_1, \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\|\boldsymbol{a}^k\|_1\Big) \leq N\cdot Ke^{KT}[1+z] \equiv N\cdot M_1,$$

$$\max\Big(\sup_{t\in[0,T]}\|\bar{\boldsymbol{a}}^t\|_\infty, \sup_{t\in[0,T]}\|\underline{\boldsymbol{a}}^t\|_\infty, \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\|\tilde{\boldsymbol{a}}^k\|_\infty, \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\|\boldsymbol{a}^k\|_\infty\Big) \leq Ke^{KT}[\sqrt{\log N}+z] \equiv M_\infty.$$

Proof of Lemma 19. Let us start with the non-linear dynamics trajectories. We have in integral form:

$$\begin{aligned}\big|\bar{a}_i^t\big| &= \Big|\bar{a}_i^0 + \int_0^t\Big[-\lambda\bar{a}_i^s - v(\bar{\boldsymbol{w}}_i^s) - \int au(\bar{\boldsymbol{w}}_i^s, \boldsymbol{w})\rho_s(\mathrm{d}\boldsymbol{\theta})\Big]\mathrm{d}s + \int_0^t\sqrt{\frac{\tau}{D}}\mathrm{d}W_i^a(s)\Big| \\ &\leq |\bar{a}_i^0| + K\int_0^t|\bar{a}_i^s|ds + KT\sqrt{M_2} + |\overline{W}_i^a(t)| \qquad (54)\\ &\leq K\int_0^t|\bar{a}_i^s|\,ds + \Theta_\infty + KT\sqrt{M_2} + W_\infty,\end{aligned}$$

where we recall that $\overline{W}_i^a(t) = \sqrt{\tau/D}W_i^a(t)$. Applying Gronwall's lemma to $\Delta(t) = \sup_{s\in[0,t]}|\bar{a}_i^s|$ with Lemma 13 gives:

$$\Delta(T) \leq Ke^{KT}[\sqrt{\log N}+z],$$

while summing (54) over i yields:

$$(\|\bar{\boldsymbol{a}}^t\|_1/N) \leq \Theta_1 + K\int_0^t(\|\bar{\boldsymbol{a}}^s\|_1/N)\mathrm{d}s + Ke^{KT} + W_1,$$

and by Gronwall's lemma: $\sup_{t\in[0,T]}\|\bar{\boldsymbol{a}}^s\|_1/N \leq Ke^{KT}[1+z]$. The same proof applies to the other trajectories and we will only write down the corresponding inequality on the integral or summation form:

$$\begin{aligned}|\underline{a}_i^t| &\leq |a_i^0| + KT + K\int_0^t|\underline{a}_i^s|\mathrm{d}s + K\int_0^t(\|\underline{\boldsymbol{a}}^s\|_1/N)\mathrm{d}s + |\overline{W}_i^a(t)|,\\ |\tilde{a}_i^k| &\leq |a_i^0| + KT + K\varepsilon\sum_{l=1}^{k-1}|\tilde{a}_i^l| + K\varepsilon\sum_{l=1}^{k-1}(\|\tilde{\boldsymbol{a}}^l\|_1/N) + |\overline{W}_i^a(t)|,\\ |a_i^k| &\leq |a_i^0| + KT + K\varepsilon\sum_{l=1}^{k-1}|a_i^l| + K\varepsilon\sum_{l=1}^{k-1}(\|\boldsymbol{a}^l\|_1/N) + |\overline{W}_i^a(t)|.\end{aligned}$$

□

Lemma 20. *There exists a constant K such that:*

$$\mathbb{P}\Big(\sup_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{u\in[0,\varepsilon]}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}\big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\big]\sqrt{\varepsilon}\Big)\le 1-e^{-z^2},$$

$$\mathbb{P}\Big(\sup_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{u\in[0,\varepsilon]}\|\underline{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\underline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}\big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\big]\sqrt{\varepsilon}\Big)\le 1-e^{-z^2}.$$

Furthermore, we have for $t,h\ge 0, t+h\le T$,

$$W_2(\rho_t,\rho_{t+h})\le\Big(\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t+h}\|_2^2]\Big)^{1/2}\le Ke^{KT}\sqrt{h}.$$

Proof of Lemma 20. We will only show the result for the non-linear dynamic. The proof for the particle dynamic will be exactly the same, upon replacing $\sqrt{M_2}$ by M_1.
Step 1. Let us consider $\Delta_i(t)\equiv\sup_{s\le t}\|\bar{\boldsymbol{\theta}}_i^t\|_2$ and $\Delta_0(t)\equiv\sup_{s\le t}\frac{1}{N}\sum_{i\le N}\|\bar{\boldsymbol{\theta}}_i^t\|_2$:

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}_i^t\|_2&\le\|\boldsymbol{\theta}_i^0\|_2+2K\int_0^t\Big(\lambda\|\bar{\boldsymbol{\theta}}_i^s\|_2+K(1+|\bar{a}_i^s|)+K\sqrt{M_2}(1+|\bar{a}_i^s|)\Big)\mathrm{d}s+\|\overline{\boldsymbol{W}}_i\|_2\\&\le K\int_0^t\|\bar{\boldsymbol{\theta}}_i^s\|_2\mathrm{d}s+Ke^{KT}T\sup_{s\in[0,t]}|\bar{a}_i^s|+\Theta_\infty+W_\infty,\end{aligned}$$

which gives, after applying Gronwall's inequality with the bounds of Lemma 13 and 19:

$$\mathbb{P}\Big(\Delta_i(t)\le Ke^{KT}\big[\sqrt{\log N}+z\big]\Big)\ge 1-e^{-z^2}.$$

Similarly:

$$\Delta_0(t)\le K\int_0^t\Delta_0(s)\mathrm{d}s+Ke^{KT}\sup_{s\in[0,t]}(\|\bar{\boldsymbol{a}}^s\|_1/N)+\Theta_1+W_1,$$

and thus:

$$\mathbb{P}\Big(\Delta_0(t)\le Ke^{KT}[1+z]\Big)\ge 1-e^{-z^2}. \tag{55}$$

Step 2. Let us bound $\sup_{0\le u\le\varepsilon}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2$:

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2&\le\Big\|\int_{k\varepsilon}^{k\varepsilon+u}\xi(s)\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s;\rho_s)\mathrm{d}s\Big\|_2+\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2\\&\le Kh\sup_{s\le T}\Big[\lambda\|\bar{\boldsymbol{\theta}}_i^s\|_2+(1+\sqrt{M_2})(1+|\bar{a}_i^s|)\Big]+\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2,\end{aligned}$$

where we defined $\overline{\boldsymbol{W}}_{i,k}(u)\equiv\int_{k\varepsilon}^{k\varepsilon+u}\sqrt{\tau/D}\mathrm{d}\boldsymbol{W}_i(s)$. By a similar computation as in Lemma 13, we have

$$\mathbb{P}\Big(\max_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\le u\le\varepsilon}\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2\ge 4\sqrt{K\varepsilon}\Big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\Big]\Big)\le e^{-z^2}.$$

Injecting this bound in the above inequality yields:

$$\mathbb{P}\Big(\max_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\le u\le\varepsilon}\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}\big[\sqrt{\log(N(T/\varepsilon\vee 1))}+z\big]\sqrt{\varepsilon}\Big)\ge 1-e^{-z^2}.$$

Another useful bound can be obtained by taking the average over $i\in[N]$:

$$\frac{1}{N}\sum_{i=1}^N\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le K\Delta_0(t)+Ke^{KT}\sup_{s\le T}\|\bar{\boldsymbol{a}}^s\|_1+\frac{1}{N}\sum_{i=1}^N\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2.$$

We get by a similar computation as in Lemma 13, we have

$$\mathbb{P}\Big(\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\le u\le\varepsilon}\frac{1}{N}\sum_{i=1}^N\|\overline{\boldsymbol{W}}_{i,k}(u)\|_2\ge 4\sqrt{K\varepsilon}[\sqrt{\log(T/\varepsilon\vee 1)}+z]\Big)\le e^{-z^2}.$$

We get the following bound:

$$\mathbb{P}\Big(\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\le u\le\varepsilon}\frac{1}{N}\sum_{i=1}^N\|\bar{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\bar{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}[\sqrt{\log(T/\varepsilon\vee 1)}+z]\sqrt{\varepsilon}\Big)\ge 1-e^{-z^2}. \tag{56}$$

Step 3. We now bound $W_2(\rho_t,\rho_{t+h})$:

$$W_2(\rho_t,\rho_{t+h})^2\le\mathbb{E}[\|\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t+h}\|_2^2]=\int_0^\infty\mathbb{P}(\|\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t+h}\|_2^2\ge u)\mathrm{d}u.$$

Using step 2, we have (where we removed the union bound over $i\in[N]$ and $k\in[0,T/\varepsilon]\cap\mathbb{N}$):

$$\mathbb{P}\Big(\|\bar{\boldsymbol{\theta}}_i^{t+h}-\bar{\boldsymbol{\theta}}_i^t\|_2\ge Ke^{KT}[1+z]\sqrt{h}\Big)\le e^{-z^2}.$$

Integrating this upper bound on the probability yields the desired inequality. □

Lemma 21. *There exists a constant K, such that for $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^{ND}$*

$$|R_N(\boldsymbol{\theta})-R_N(\boldsymbol{\theta}')|\le K(1+\|\boldsymbol{a}\|_1/N+\|\boldsymbol{a}'\|_1/N+\|\boldsymbol{a}'\|_1^2/N^2)\max_{i\in[N]}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_i'\|_2.$$

Proof of Lemma 21. We have

$$\begin{aligned}&|R_N(\boldsymbol{\theta})-R_N(\boldsymbol{\theta})|\\\le&\frac{2}{N}\sum_{i=1}^N\big|a_iv(\boldsymbol{w}_i)-a_i'v(\boldsymbol{w}_i')\big|+\frac{1}{N^2}\sum_{i,j=1}^N|a_ia_ju(\boldsymbol{w}_i,\boldsymbol{w}_j)-a_i'a_j'u(\boldsymbol{w}_i',\boldsymbol{w}_j')|\\\le&\frac{2}{N}\sum_{i=1}^N K(|a_i'-a_i|+|a_i'|\|\boldsymbol{w}_i-\boldsymbol{w}_i'\|_2)\\&+\frac{1}{N^2}\sum_{i,j=1}^N K\Big[|a_i||a_j-a_j'|+|a_j'||a_i-a_i'|+|a_i'a_j'|(\|\boldsymbol{w}_i-\boldsymbol{w}_i'\|_2+\|\boldsymbol{w}_j-\boldsymbol{w}_j'\|_2)\Big]\\\le&K(1+\|\boldsymbol{a}\|_1/N+\|\boldsymbol{a}'\|_1/N+\|\boldsymbol{a}'\|_1^2/N^2)\max_{i\in[N]}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_i'\|_2.\end{aligned}$$

□

E.2 Bound between PDE and nonlinear dynamics

Proposition 13 (PDE-ND)**.** *There exists a constant K such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]}|R_N(\bar{\boldsymbol{\theta}}^t)-R(\rho_t)|\le Ke^{KT}\Big[\log^{3/2}(NT)+z^3\Big]/\sqrt{N}\Big)\ge 1-e^{-z^2}.$$

The proof will use the same decomposition in two terms as in the proof of proposition 1.

Lemma 22 (Term II bound)**.** *We have*

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)-R(\rho_t)|\le Ke^{KT}/N.$$

Proof of Lemma 22. The bound hold simply by observing that

$$\begin{aligned}|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)-R(\rho_t)|=&\frac{1}{N}\Big|\int a^2u(\boldsymbol{w},\boldsymbol{w})\rho_t(\mathrm{d}\boldsymbol{\theta})-\int a_1a_2u(\boldsymbol{w}_1,\boldsymbol{w}_2)\rho_t(\mathrm{d}\boldsymbol{\theta}_1)\rho_t(\mathrm{d}\boldsymbol{\theta}_2)\Big|\\\le&K/N\int a^2\rho_t(\mathrm{d}a)\le Ke^{KT}/N\end{aligned}$$

where we used the upper bound on the second moment of variable a in Lemma 18. □

Lemma 23 (Term I bound)**.** *There exists K, such that*

$$\mathbb{P}\Big(\sup_{t\in[0,T]} |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \le Ke^{KT}\left[\log(NT) + z^3\right]/\sqrt{N}\Big) \ge 1 - e^{-z^2}.$$

Proof of Lemma 23. We have:

$$\begin{aligned}\big|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)\big| \le& 2\Big|\frac{1}{N}\sum_{i=1}^N \left[V(\bar{\boldsymbol{\theta}}_i^t) - \mathbb{E}V(\bar{\boldsymbol{\theta}}_i^t)\right]\Big| + \frac{1}{N^2}\sum_{i=1}^N \Big|U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t)\Big| \\ &+ \frac{1}{N}\sum_{i=1}^N \Big|\frac{1}{N}\sum_{j=1, j\neq i}^N \Big[U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\Big]\Big| \\ &+ \frac{1}{N}\sum_{i=1}^N \Big|\frac{1}{N}\sum_{j=1, j\neq i}^N \Big[\mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\Big]\Big|.\end{aligned}$$

We will bound each of these terms separately. For any fixed t, we have $(\bar{\boldsymbol{\theta}}_i^t)_{i\in[N]} \sim \rho_t$ independently. Define:

$$Q_1(t) = \Big|\frac{1}{N}\sum_{i=1}^N \left[V(\bar{\boldsymbol{\theta}}_i^t) - \mathbb{E}V(\bar{\boldsymbol{\theta}}_i^t)\right]\Big|,$$

which is the absolute value of the sum of martingale differences. Furthermore, we can rewrite $V(\bar{\boldsymbol{\theta}}_i^t) = \bar{a}_i^t v(\bar{\boldsymbol{w}}_i^t)$ which is Ke^{KT}-sub-Gaussian (product of a sub-Gaussian random variable, by Lemma 18, and a bounded random variable). We can therefore apply Azuma-Hoeffding's inequality (Lemma 31),

$$\mathbb{P}\Big(Q_1(t) \le Ke^{KT}\left[1+z\right]/\sqrt{N}\Big) \ge 1 - e^{-z^2}.$$

The second term is bounded as follow:

$$\begin{aligned}E_2(t) \equiv \frac{1}{N^2}\sum_{i=1}^N \Big|U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_i^t)\Big| \le& \frac{1}{N^2}\sum_{i=1}^N |(\bar{a}_i^t)^2 u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_i^t)| + \frac{1}{N^2}\sum_{i=1}^N \Big|\mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t}\Big[(\bar{a}_i^t)^2 u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_i^t)\Big]\Big| \\ \le& \frac{K}{N^2}\cdot\|\bar{\boldsymbol{a}}^t\|_\infty \cdot \|\bar{\boldsymbol{a}}^t\|_1 + \frac{Ke^{KT}}{N},\end{aligned}$$

where we used that $\int a^2 \rho_t(\mathrm{d}a) \le Ke^{KT}$. Using Lemma 19, we get:

$$\mathbb{P}\Big(E_2(t) \le Ke^{KT}\left[\sqrt{\log N} + z^2\right]/N\Big) \ge 1 - e^{-z^2}.$$

Define:

$$Q_2^i(t) = \Big|\frac{1}{N}\sum_{j=1, j\neq i}^N \Big[U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\Big]\Big|.$$

Because $\bar{\boldsymbol{\theta}}_i^t$ is independent of the $(\bar{\boldsymbol{\theta}}_j^t)_{j\in[N], j\neq i}$, we can condition on $\bar{\boldsymbol{\theta}}_i^t$, and restrict ourselves to the event where $\bar{\boldsymbol{\theta}}_i^t \le M_\infty$. $Q_2^i(t)$ is the absolute value of a sum of martingale difference, with $U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) = \bar{a}_i^t \bar{a}_j^t u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_j^t)$ which is $Ke^{KT}|\bar{a}_i^t|^2$-sub-Gaussian (product of a sub-Gaussian random variable and a bounded random variable). We apply Azuma-Hoeffding's inequality (Lemma 31),

$$\begin{aligned}&\mathbb{P}\Big(Q_2^i(t) \ge Ke^{KT}M_\infty\left[1+z\right]/\sqrt{N}\Big) \\ \le& \mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t}\Big[\mathbb{P}\Big(Q_2^i(t) \ge Ke^{KT}M_\infty\left[1+z\right]/\sqrt{N}\Big|\bar{\boldsymbol{\theta}}_i^t\Big)\mathbf{1}(|\bar{a}_i^t| \le M_\infty)\Big] + \mathbb{P}(|\bar{a}_i^t| > M_\infty) \\ \le& 2e^{-z^2}\end{aligned}$$

We take the union bound over $i \in [N]$ and get:

$$\mathbb{P}\Big(\max_{i\in[N]} Q_2^i(t) \geq Ke^{KT}\left[\log N + z^2\right]/\sqrt{N}\Big) \leq e^{-z^2}.$$

Define:

$$Q_3^i(t) = \Big|\frac{1}{N}\sum_{j=1,j\neq i}^{N}\left[\mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) - \mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\right]\Big|.$$

We have:

$$\mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t) = \bar{a}_i^t \cdot \int a u(\bar{\boldsymbol{w}}_i^t, \boldsymbol{w})\rho(\mathrm{d}\boldsymbol{\theta}),$$

with $\Big|\int au(\bar{\boldsymbol{w}}_i^t, \boldsymbol{w})\rho(\mathrm{d}\boldsymbol{\theta})\Big| \leq K\Big(\int a^2 \rho_t(\mathrm{d}a)\Big)^{1/2} \leq Ke^{KT}$. Thus, $\mathbb{E}_{\bar{\boldsymbol{\theta}}_j^t} U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)$ is Ke^{KT}-sub-Gaussian (product of a sub-Gaussian random variables and of a bounded random variable). Applying Azuma-Hoeffding's inequality Lemma 31, followed by an union bound over $i \in [N]$, we get

$$\mathbb{P}\Big(\max_{i\in[N]} Q_3^i(t) \geq Ke^{KT}\left[\sqrt{\log N} + z\right]/\sqrt{N}\Big) \leq e^{-z^2}.$$

Combining the above bounds with the bound on $\sup_{s\in[0,T]}\{\|\bar{\boldsymbol{a}}^s\|_1, \|\bar{\boldsymbol{a}}^s\|_\infty\}$ of Lemma 19 yields:

$$\mathbb{P}\Big(\big|R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)\big| \geq Ke^{KT}\left[\log N + z^2\right]/\sqrt{N}\Big) \leq e^{-z^2}. \tag{57}$$

In order to extend this concentration uniformly on the interval $[0, T]$, we use the following result:

Lemma 24. *There exists K, such that*

$$\sup_{k\in[0,T/\eta]\cap\mathbb{N}} \sup_{u\in[0,\eta]} \Big||R_N(\bar{\boldsymbol{\theta}}^{k\eta+u}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta+u})| - |R_N(\bar{\boldsymbol{\theta}}^{k\eta}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta})|\Big|$$
$$\leq Ke^{KT}\left[\sqrt{\log\left(N(T/\eta \vee 1)\right)} + z^3\right]\sqrt{\eta},$$

with probability at least $1 - e^{-z^2}$.

Proof of Lemma 24. Consider $t, h \geq 0, t + h \leq T$. From Lemma 21,

$$|R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)| \leq K(1 + \|\bar{\boldsymbol{a}}^{t+h}\|_1/N + \|\bar{\boldsymbol{a}}^t\|_1/N + \|\bar{\boldsymbol{a}}^{t+h}\|_1^2/N^2)\max_{i\in[N]}\|\bar{\boldsymbol{\theta}}_i^{t+h} - \bar{\boldsymbol{\theta}}_i^t\|_2.$$

Using Lemma 20 without the union bound over $s \in \eta\{0, 1, \ldots, \lfloor T/\eta\rfloor\}$ and the bounds on $\sup_{t\in[0,T]}\{\|\bar{\boldsymbol{a}}^t\|_1\}$ of Lemma 19, we get

$$\mathbb{P}\Big(|R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)| \geq Ke^{KT}\left[\sqrt{\log N} + z^3\right]\sqrt{h}\Big) \leq e^{-z^2}.$$

The difference in expectation, where the expectation is taken over $(\bar{\boldsymbol{\theta}}_i)_{i\in[N]}$, is therefore bounded by

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \leq \mathbb{E}|R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)| \leq \int_0^\infty \mathbb{P}\Big(|R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)| \geq u\Big)\mathrm{d}u.$$

Doing a change of variable, we get:

$$|\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)| \leq Ke^{KT}\sqrt{h\log N} + \int_0^\infty e^{-z^2}Ke^{KT}\sqrt{h}z^2\mathrm{d}z$$
$$\leq Ke^{KT}(\sqrt{\log N} + 1)\sqrt{h}.$$

Hence using that

$$\Big||R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h})| - |R_N(\bar{\boldsymbol{\theta}}^t) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|\Big| \leq |R_N(\bar{\boldsymbol{\theta}}^{t+h}) - R_N(\bar{\boldsymbol{\theta}}^t)| + |\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{t+h}) - \mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|,$$

with Lemma 20, we get

$$\sup_{k\in[0,T/\eta]\cap\mathbb{N}}\sup_{u\in[0,\eta]}\Big||R_N(\bar{\boldsymbol{\theta}}^{k\eta+u})-\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta+u})|-|R_N(\bar{\boldsymbol{\theta}}^{k\eta})-\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^{k\eta})|\Big|$$
$$\leq Ke^{KT}\left[\sqrt{\log\left(N(T/\eta\vee 1)\right)}+z^3\right]\sqrt{\eta},$$

with probability at least $1-e^{-z^2}$. □

Taking an union bound over $s\in\eta\{0,\dots,\lfloor T/\eta\rfloor\}$ in Eq. (57) and bounding the variation inside the grid intervals, we get

$$\begin{aligned}\mathbb{P}\Big(\sup_{t\in[0,T]}|R_N(\bar{\boldsymbol{\theta}}^t)-\mathbb{E}R_N(\bar{\boldsymbol{\theta}}^t)|\geq& Ke^{KT}\left[\log N+z^2\right]/\sqrt{N}+Ke^{KT}\left[\sqrt{\log\left(N(T/\eta\vee 1)\right)}+z^3\right]\sqrt{\eta}\Big)\\ &\leq (T/\eta)\exp\{-z^2\}.\end{aligned}$$

Taking $\eta=1/(N\log N)$ and $z=[\sqrt{\log(NT\log N)}+z']$ concludes the proof. □

E.3 Bound between nonlinear dynamics and particle dynamics

Proposition 14 (ND-PD)**.** *There exists a constant K, such that with probability at least $1-e^{-z^2}$, we have*

$$\begin{aligned}\sup_{t\in[0,T]}\max_{i\in[N]}\|\bar{\boldsymbol{\theta}}_i^t-\underline{\boldsymbol{\theta}}_i^t\|_2\leq& Ke^{e^{KT}[\sqrt{\log N}+z^2]}\left[\sqrt{D\log N}+\log^{3/2}(NT)+z^3\right]/\sqrt{N},\\ \sup_{t\in[0,T]}|R_N(\underline{\boldsymbol{\theta}}^t)-R_N(\bar{\boldsymbol{\theta}}^t)|\leq& Ke^{e^{KT}[\sqrt{\log N}+z^2]}\left[\sqrt{D\log N}+\log^{3/2}(NT)+z^5\right]/\sqrt{N}.\end{aligned}$$

Proof of Proposition 14. Define $\Delta(t)\equiv\sup_{s\leq t}\max_{i\in[N]}\|\bar{\boldsymbol{\theta}}_i^s-\underline{\boldsymbol{\theta}}_i^s\|_2$. We have

$$\begin{aligned}\|\underline{\boldsymbol{\theta}}_i^t-\bar{\boldsymbol{\theta}}_i^t\|_2\leq&\int_0^t\|\boldsymbol{G}(\bar{\boldsymbol{\theta}}_i^s;\rho_s)-\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s;\underline{\rho}_s^{(N)})\|_2\mathrm{d}s\\ \leq&\int_0^t\lambda\|\bar{\boldsymbol{\theta}}_i^s-\underline{\boldsymbol{\theta}}_i^s\|_2\mathrm{d}s+\int_0^t\|\nabla V(\bar{\boldsymbol{\theta}}_i^s)-\nabla V(\underline{\boldsymbol{\theta}}_i^s)\|_2\mathrm{d}s\\ &+\int_0^t\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2\mathrm{d}s\\ &+\int_0^t\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\int\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\boldsymbol{\theta})\rho_s(\mathrm{d}\boldsymbol{\theta})\Big\|_2\mathrm{d}s.\end{aligned}\tag{58}$$

Let us bound each term separately. We have

$$\begin{aligned}\|\nabla V(\bar{\boldsymbol{\theta}}_i^s)-\nabla V(\underline{\boldsymbol{\theta}}_i^s)\|_2\leq&|v(\bar{\boldsymbol{w}}_i^s)-v(\underline{\boldsymbol{w}}_i^s)|+\|\bar{a}_i^s\nabla v(\bar{\boldsymbol{w}}_i^s)-\underline{a}_i^s\nabla v(\underline{\boldsymbol{w}}_i^s)\|_2\\ \leq&K(\|\bar{\boldsymbol{w}}_i^s-\underline{\boldsymbol{w}}_i^s\|_2+|\bar{a}_i^s-\underline{a}_i^s|+|\bar{a}_i^s|\|\bar{\boldsymbol{w}}_i^s-\underline{\boldsymbol{w}}_i^s\|_2)\\ \leq&K(1+\|\bar{\boldsymbol{a}}^s\|_\infty)\|\bar{\boldsymbol{\theta}}_i^s-\underline{\boldsymbol{\theta}}_i^s\|_2.\end{aligned}$$

We decompose the second term into two terms

$$\begin{aligned}\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2\leq&\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2\\ &+\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2,\end{aligned}$$

where

$$
\begin{aligned}
&\Big\|\frac{1}{N}\sum_{j=1}^{N}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2\\
\le&\Big|\frac{1}{N}\sum_{j=1}^{N}\bar{a}_j^s u(\bar{\boldsymbol{w}}_i^s,\bar{\boldsymbol{w}}_j^s)-\underline{a}_j^s u(\bar{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)\Big|+\Big\|\frac{1}{N}\sum_{j=1}^{N}\bar{a}_i^s\bar{a}_j^s\nabla_1 u(\bar{\boldsymbol{w}}_i^s,\bar{\boldsymbol{w}}_j^s)-\bar{a}_i^s\underline{a}_j^s\nabla_1 u(\bar{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)\Big\|_2\\
\le&K(1+|\bar{a}_i^s|)\Big[\max_{j\in[N]}|\bar{a}_j^s-\underline{a}_j^s|+\Big[\frac{1}{N}\sum_{j=1}^{N}|\bar{a}_j^s|\Big]\max_{j\in[N]}\|\bar{\boldsymbol{w}}_j^s-\underline{\boldsymbol{w}}_j^s\|_2\Big]\\
\le&K(1+\|\bar{\boldsymbol{a}}^s\|_\infty)\cdot(1+\|\bar{\boldsymbol{a}}^s\|_1/N)\cdot\max_{j\in[N]}\|\bar{\boldsymbol{\theta}}_j^s-\underline{\boldsymbol{\theta}}_j^s\|_2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\Big\|\frac{1}{N}\sum_{j=1}^{N}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)-\nabla_1 U(\underline{\boldsymbol{\theta}}_i^s,\underline{\boldsymbol{\theta}}_j^s)\Big\|_2\\
\le&\Big|\frac{1}{N}\sum_{j=1}^{N}\underline{a}_j^s u(\bar{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)-\underline{a}_j^s u(\underline{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)\Big|+\Big\|\frac{1}{N}\sum_{j=1}^{N}\bar{a}_i^s\underline{a}_j^s\nabla_1 u(\bar{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)-\underline{a}_i^s\underline{a}_j^s\nabla_1 u(\underline{\boldsymbol{w}}_i^s,\underline{\boldsymbol{w}}_j^s)\Big\|_2\\
\le&\Big[\frac{K}{N}\sum_{j=1}^{N}|\underline{a}_j^s|\Big]\sup_{j\in[N]}\|\bar{\boldsymbol{w}}_j^s-\underline{\boldsymbol{w}}_j^s\|_2+K|\bar{a}_i^s-\underline{a}_i^s|\Big[\frac{1}{N}\sum_{j=1}^{N}|\underline{a}_j^s|\Big]+K|\underline{a}_i^s|\Big[\frac{1}{N}\sum_{j=1}^{N}|\underline{a}_j^s|\Big]\|\bar{\boldsymbol{w}}_i^s-\underline{\boldsymbol{w}}_i^s\|_2\\
\le&K(1+\|\underline{\boldsymbol{a}}^s\|_\infty)\cdot(1+\|\underline{\boldsymbol{a}}^s\|_1/N)\cdot\max_{j\in[N]}\|\bar{\boldsymbol{\theta}}_j^s-\underline{\boldsymbol{\theta}}_j^s\|_2.
\end{aligned}
$$

The last term in Eq. (58) can be decomposed into two terms. Consider $j=i$:

$$
\begin{aligned}
&\frac{1}{N}\|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_i^s)-\int\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\boldsymbol{\theta})\rho_s(\mathrm{d}\boldsymbol{\theta})\|_2\\
\le&\frac{1}{N}\|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_i^s)\|_2+\frac{1}{N}\int\|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\boldsymbol{\theta})\|_2\rho_s(\mathrm{d}\boldsymbol{\theta})\\
\le&\frac{1}{N}\left[|\bar{a}_i^s u(\bar{\boldsymbol{w}}_i^s,\bar{\boldsymbol{w}}_i^s)|+\|(\bar{a}_i^s)^2\nabla_1 u(\bar{\boldsymbol{w}}_i^s,\bar{\boldsymbol{w}}_i^s)\|_2\right]+\int\left[|au(\bar{\boldsymbol{w}}_i^s,\boldsymbol{w})|+\|\bar{a}_i^s a\nabla_1 u(\bar{\boldsymbol{w}}_i^s,\boldsymbol{w})\|_2\right]\rho_s(\mathrm{d}\boldsymbol{\theta})\\
\le&\frac{1}{N}K\|\bar{\boldsymbol{a}}^s\|_\infty\cdot(1+\|\bar{\boldsymbol{a}}^s\|_\infty)+Ke^{KT}(1+\|\bar{\boldsymbol{a}}^s\|_\infty),
\end{aligned}
$$

where we used that $\int|a|\rho_s(\mathrm{d}\boldsymbol{\theta})\le\Big(\int a^2\rho_s(\mathrm{d}\boldsymbol{\theta})\Big)^{1/2}$ and Lemma 18. We consider $j\neq i$ and denote:

$$
Q^i(s)=\Big\|\frac{1}{N}\sum_{j=1,j\neq i}^{N}\left[\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)-\int\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}})\rho_s(\mathrm{d}\boldsymbol{\theta})\right]\Big\|_2,
$$

which is bounded in the following lemma:

Lemma 25. *There exists a constant K, such that:*

$$
\mathbb{P}\Big(\sup_{s\in[0,T]}\max_{i\le N}Q^i(s)\ge Ke^{KT}\left[\sqrt{D\log N}+\log^{3/2}(NT)+z^3\right]/\sqrt{N}\Big)\le e^{-z^2}.
$$

Proof of Lemma 25. The concentration of $Q^i(s)$ follows from a similar method as in the proof of Lemma 23. For any fixed s, we have $(\bar{\boldsymbol{\theta}}_i^s)_{i\in[N]}\sim\rho_s$ independently. In particular, we have

$$
\int\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}})\rho_s(\mathrm{d}\boldsymbol{\theta})=\mathbb{E}\left[\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s,\bar{\boldsymbol{\theta}}_j^s)\Big|\bar{\boldsymbol{\theta}}_i^s\right],
$$

and $Q^i(s)$ conditioned on $\bar{\boldsymbol{\theta}}_i^s$ is the norm of a martingale difference sum. We furthermore restrict ourselves to the event where $\bar{a}_i^s \leq M_\infty$. We have $\nabla_1 U(\bar{\boldsymbol{\theta}}_i^s, \bar{\boldsymbol{\theta}}_j^s) = \bar{a}_j^t \cdot (u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_j^t), \bar{a}_i^s \nabla_1 u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_j^t))$ which is $Ke^{KT} M_\infty^2$-sub-Gaussian (the product of a sub-Gaussian random variable and a bounded random variable is sub-Gaussian). We can therefore apply Azuma-Hoeffding 's inequality (Lemma 31),

$$\begin{aligned}
&\mathbb{P}\Big(Q^i(s) \geq Ke^{KT} M_\infty \left[\sqrt{D} + z\right] / \sqrt{N}\Big) \\
\leq &\mathbb{E}_{\bar{\boldsymbol{\theta}}_i^t} \left[\mathbb{P}\Big(Q^i(s) \geq Ke^{KT} M_\infty \left[\sqrt{D} + z\right] / \sqrt{N} \Big| \bar{\boldsymbol{\theta}}_i^s\Big) \mathbf{1}(|\bar{a}_i^s| \leq M_\infty)\right] + \mathbb{P}(\|\bar{\boldsymbol{a}}^s\|_\infty \geq M_\infty) \\
\leq &2e^{-z^2}.
\end{aligned}$$

Taking the union bound over the $i \in [N]$

$$\mathbb{P}\left(\max_{i \leq N} Q^i(s) \geq Ke^{KT} \left[\sqrt{D \log N} + \log(N) + z^2\right] / \sqrt{N}\right) \leq e^{-z^2}.$$

Furthermore, let us consider $t, h \geq 0, t + h \leq T$:

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1, j \neq i}^{N} \|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 \\
\leq &\frac{1}{N} \sum_{j=1, j \neq i}^{N} \left[|\bar{a}_j^{t+h} u(\bar{\boldsymbol{w}}_i^{t+h}, \bar{\boldsymbol{w}}_j^{t+h}) - \bar{a}_j^t u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_j^t)| + \|\bar{a}_i^{t+h} \bar{a}_j^{t+h} \nabla_1 u(\bar{\boldsymbol{w}}_i^{t+h}, \bar{\boldsymbol{w}}_j^{t+h}) - \bar{a}_i^t \bar{a}_j^t \nabla_1 u(\bar{\boldsymbol{w}}_i^t, \bar{\boldsymbol{w}}_j^t)\|_2\right] \\
\leq &K(1 + \left\|\bar{\boldsymbol{a}}^t\right\|_\infty) \cdot (1 + \left\|\bar{\boldsymbol{a}}^t\right\|_1 / N) \cdot \sup_{i \leq N} \|\bar{\boldsymbol{\theta}}_i^{t+h} - \bar{\boldsymbol{\theta}}_i^t\|_2.
\end{aligned}$$

Considering Lemma 20 without the union bound over $s \in \eta\{0, 1, \ldots, \lfloor T/\eta \rfloor\}$ and the high probability bounds on $\sup_{t \in [0,T]} \{\|\bar{\boldsymbol{a}}^t\|_\infty, \|\bar{\boldsymbol{a}}^t\|_1\}$ of Lemma 19, we get:

$$\mathbb{P}\Big(\frac{1}{N} \sum_{j=1, j \neq i}^{N} \|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 \geq Ke^{KT}(1 + z) \left[\sqrt{\log N} + z\right]^2 \sqrt{h}\Big) \leq e^{-z^2}.$$

The difference in expectation, where the expectation is taken over $\bar{\boldsymbol{\theta}}_j$, is bounded by

$$\begin{aligned}
\|\mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 &\leq \mathbb{E}\Big[\frac{1}{N} \sum_{j=1, j \neq i}^{N} \|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2\Big] \\
&\leq \int_0^\infty \mathbb{P}\Big(\frac{1}{N} \sum_{j=1, j \neq i}^{N} \|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 \geq u\Big) \mathrm{d}u.
\end{aligned}$$

Noticing that $(1 + z) \left[\sqrt{\log N} + z\right]^2 \leq (\sqrt{\log N} + z)^3$ and doing a change of variable, we get:

$$\begin{aligned}
\|\mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 &\leq Ke^{KT} \log N \sqrt{h} + \int_{-\sqrt{\log N}}^{\infty} e^{-z^2} Ke^{KT} z^2 \sqrt{h} \mathrm{d}z \\
&\leq Ke^{KT} (\log N + 1) \sqrt{h}.
\end{aligned}$$

Hence using that

$$\begin{aligned}
|Q^i(t + h) - Q^i(t)| \leq &\frac{1}{N} \sum_{j=1, j \neq i}^{N} \|\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2 \\
&+ \|\mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^{t+h}, \bar{\boldsymbol{\theta}}_j^{t+h}) - \mathbb{E}\nabla_1 U(\bar{\boldsymbol{\theta}}_i^t, \bar{\boldsymbol{\theta}}_j^t)\|_2,
\end{aligned}$$

and the bounds derived above, with an union bound over $t \in \eta\{0, 1, \dots, \lfloor T/\eta \rfloor\}$, we get

$$\mathbb{P}\Big(\sup_{k \in [0,T/\eta] \cap \mathbb{N}} \sup_{u \in [0,\eta]} \max_{i \in [N]} |Q^i(k\eta + u) - Q^i(k\eta)| \le K e^{KT} \left[\log\left(N(T/\eta \vee 1)\right) + z^3 \right] \sqrt{\eta} \Big) \ge 1 - e^{-z^2}.$$

We can therefore take the supremum over the interval $[0, T]$:

$$\mathbb{P}\Big(\max_{i \le N} \sup_{s \in [0,T]} Q^i(s) \ge K e^{KT} \left[\sqrt{D \log N} + \log(N) + z^2 \right] / \sqrt{N} + K e^{KT} \left[\log\left(N(T/\eta \vee 1)\right) + z^3 \right] \sqrt{\eta} \Big)$$
$$\le (T/\eta) \exp\{-z^2\}.$$

Taking $\eta = 1/N$ and $z = [\sqrt{\log(NT)} + z']$:

$$\mathbb{P}\Big(\max_{i \le N} \sup_{s \in [0,T]} Q^i(s) \ge K e^{KT} \left[\sqrt{D \log N} + \log^{3/2}(NT) + z^3 \right] / \sqrt{N} \Big) \le e^{-z^2}.$$

□

Using the high probability bound on $\sup_{s \in [0,T]} \{ \|\bar{\boldsymbol{a}}^s\|_1/N, \|\bar{\boldsymbol{a}}^s\|_\infty \}$ of Lemma 19, we get with probability at least $1 - e^{-z^2}$ that for all $t \in [0, T]$

$$\begin{aligned} \Delta(t) \le & K e^{KT}(1 + z) \left[\sqrt{\log N} + z \right] \int_0^t \Delta(s) \mathrm{d}s + T K e^{KT} \left[\sqrt{\log N} + z \right]^2 / N \\ & + T K e^{KT} \left[\sqrt{D \log N} + \log^{3/2}(NT) + z^3 \right] / \sqrt{N}. \end{aligned}$$

Applying Gronwall's inequality, we get:

$$\mathbb{P}\Big(\Delta(T) \le K e^{e^{KT}[\sqrt{\log N} + z^2]} \left[\sqrt{D \log N} + \log^{3/2}(NT) + z^3 \right] / \sqrt{N} \Big) \ge 1 - e^{-z^2}.$$

Using Lemma 21 and the high probability bounds on $\sup_{t \in [0,T]} \{ \|\bar{\boldsymbol{a}}^t\|_1/N, \|\bar{\boldsymbol{a}}^t\|_\infty, \|\tilde{\boldsymbol{a}}^t\|_1/N, \|\tilde{\boldsymbol{a}}^t\|_\infty \}$ of Lemma 19 concludes the proof. □

E.4 Bound between particle dynamics and GD

Proposition 15 (PD-GD)**.** *There exists constant K, such that with probability at least $1 - e^{-z^2}$, we have*

$$\begin{aligned} \sup_{k \in [0,T/\varepsilon] \cap \mathbb{N}} \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le & K e^{e^{KT}[\sqrt{\log N} + z^2]} \left[\log(N(T/\varepsilon \vee 1)) + z^4 \right] \sqrt{\varepsilon}, \\ \sup_{k \in [0,T/\varepsilon] \cap \mathbb{N}} |R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k)| \le & K e^{e^{KT}[\sqrt{\log N} + z^2]} \left[\log(N(T/\varepsilon \vee 1)) + z^6 \right] \sqrt{\varepsilon}. \end{aligned}$$

Proof of Proposition 15. Denote $\Delta(t) \equiv \sup_{k \in [0,t/\varepsilon] \cap \mathbb{N}} \max_{i \in [N]} \|\underline{\boldsymbol{\theta}}_i^{k\varepsilon} - \tilde{\boldsymbol{\theta}}_i^k\|_2$. For $k \in \mathbb{N}$ and $t = k\varepsilon$,

$$\begin{aligned} \|\underline{\boldsymbol{\theta}}_i^t - \tilde{\boldsymbol{\theta}}_i^k\|_2 \le & \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s \\ \le & \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s; \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2 \mathrm{d}s \\ & + \int_0^t \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}; \tilde{\rho}_{[s]/\varepsilon}^{(N)})\|_2 \mathrm{d}s. \end{aligned}$$

Let us consider each terms separately:

$$\begin{aligned} & \|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s, \underline{\rho}_s^{(N)}) - \boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]}; \underline{\rho}_{[s]}^{(N)})\|_2 \\ \le & \lambda \|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2 + \|\nabla V(\underline{\boldsymbol{\theta}}_i^s) - \nabla V(\underline{\boldsymbol{\theta}}_i^{[s]})\|_2 + \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_1 U(\underline{\boldsymbol{\theta}}_i^s, \underline{\boldsymbol{\theta}}_j^s) - \nabla_1 U(\underline{\boldsymbol{\theta}}_i^{[s]}, \underline{\boldsymbol{\theta}}_j^{[s]}) \Big\|_2 \\ \le & K(1 + \|\underline{\boldsymbol{a}}^s\|_\infty) \cdot \|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2 + K(1 + \|\underline{\boldsymbol{a}}^s\|_\infty) \cdot (1 + \|\underline{\boldsymbol{a}}^s\|_1/N) \cdot \max_{j \in [N]} \|\underline{\boldsymbol{\theta}}_i^s - \underline{\boldsymbol{\theta}}_i^{[s]}\|_2 \\ & + K(1 + \|\underline{\boldsymbol{a}}^{[s]}\|_\infty) \cdot (1 + \|\underline{\boldsymbol{a}}^{[s]}\|_1/N) \cdot \max_{j \in [N]} \|\underline{\boldsymbol{\theta}}_j^s - \underline{\boldsymbol{\theta}}_j^{[s]}\|_2. \end{aligned}$$

From Lemma 20, we know that

$$\mathbb{P}\Big(\sup_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{u\in[0,\varepsilon]}\|\underline{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\underline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}\left[\sqrt{\log(N(T/\varepsilon\vee 1))}+z^2\right]\sqrt{\varepsilon}\Big)\le 1-e^{-z^2},$$

which combined with the upper bound on $\sup_{s\in[0,T]}\{\|\underline{\boldsymbol{a}}^s\|_1/N,\|\underline{\boldsymbol{a}}^s\|_\infty\}$ of Lemma 19, shows that with probability at least $1-e^{-z^2}$, we have

$$\int_0^{k\varepsilon}\|\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^s;\underline{\rho}_s^{(N)})-\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]};\underline{\rho}_{[s]}^{(N)})\|_2\mathrm{d}s\le KTe^{KT}\left[\log(N(T/\varepsilon\vee 1))+z^4\right]\sqrt{\varepsilon}.$$

Consider the second term:

$$\begin{aligned}
&\|\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon},\tilde{\rho}_{[s]/\varepsilon}^{(N)})-\boldsymbol{G}(\underline{\boldsymbol{\theta}}_i^{[s]};\underline{\rho}_{[s]}^{(N)})\|_2\\
\le&\lambda\|\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}-\underline{\boldsymbol{\theta}}_i^{[s]}\|_2+\|\nabla V(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon})-\nabla V(\underline{\boldsymbol{\theta}}_i^{[s]})\|_2+\Big\|\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon},\tilde{\boldsymbol{\theta}}_j^{[s]/\varepsilon})-\nabla_1 U(\underline{\boldsymbol{\theta}}_i^{[s]},\underline{\boldsymbol{\theta}}_j^{[s]})\Big\|_2\\
\le&K(1+\|\tilde{\boldsymbol{a}}^{[s]}\|_\infty)\cdot\|\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}-\underline{\boldsymbol{\theta}}_i^{[s]}\|_2+K(1+\|\tilde{\boldsymbol{a}}^{[s]/\varepsilon}\|_\infty)\cdot(1+\|\tilde{\boldsymbol{a}}^{[s]/\varepsilon}\|_1/N)\cdot\max_{j\in[N]}\|\tilde{\boldsymbol{\theta}}_i^{[s]/\varepsilon}-\underline{\boldsymbol{\theta}}_i^{[s]}\|_2\\
&+K(1+\|\underline{\boldsymbol{a}}^{[s]}\|_\infty)\cdot(1+\|\underline{\boldsymbol{a}}^{[s]}\|_1/N)\cdot\max_{j\in[N]}\|\tilde{\boldsymbol{\theta}}_j^{[s]/\varepsilon}-\underline{\boldsymbol{\theta}}_j^{[s]}\|_2.
\end{aligned}$$

Using the high probability bound on $\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\{\|\underline{\boldsymbol{a}}^{k\varepsilon}\|_1/N,\|\underline{\boldsymbol{a}}^{k\varepsilon}\|_\infty,\|\tilde{\boldsymbol{a}}^k\|_1/N,\|\tilde{\boldsymbol{a}}^k\|_\infty\}$ of Lemma 19, we get with probability at least $1-e^{-z^2}$ that for all $t\in[0,T]$

$$\Delta(t)\le Ke^{KT}(1+z)\left[\sqrt{\log N}+z\right]\int_0^t\Delta(s)\mathrm{d}s+Ke^{KT}\left[\log(N(T/\varepsilon\vee 1))+z^4\right]\sqrt{\varepsilon}.$$

Applying Gronwall's inequality, we get with probability at least $1-e^{-z^2}$,

$$\mathbb{P}\Big(\Delta(T)\le Ke^{e^{KT}[\sqrt{\log N}+z^2]}\left[\log(N(T/\varepsilon\vee 1))+z^4\right]\sqrt{\varepsilon}\Big)\ge 1-e^{-z^2}.$$

This bound combined with Lemma 21 concludes the proof. □

E.5 Bound between GD and SGD

Proposition 16 (GD-SGD)**.** *There exists K, such that with probability at least $1-e^{-z^2}$, we have*

$$\begin{aligned}
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}\|\tilde{\boldsymbol{\theta}}_i^k-\boldsymbol{\theta}_i^k\|_2\le&Ke^{e^{KT}[\sqrt{\log N}+z^2]}\left[\sqrt{D}\log N+\log^{3/2}N+z^3\right]\sqrt{\varepsilon},\\
\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}|R_N(\tilde{\boldsymbol{\theta}}^k)-R_N(\boldsymbol{\theta}^k)|\le&Ke^{e^{KT}[\sqrt{\log N}+z^2]}\left[\sqrt{D}\log N+\log^{3/2}N+z^5\right]\sqrt{\varepsilon}.
\end{aligned}$$

Proof of Proposition 16. Define $\Delta(t)\equiv\sup_{k\in[0,t/\varepsilon]\cap\mathbb{N}}\max_{i\in[N]}\|\tilde{\boldsymbol{\theta}}_i^k-\boldsymbol{\theta}_i^k\|_2$. Denote the generated σ-algebra:

$$\mathcal{F}_k=\sigma((\boldsymbol{\theta}_i^0)_{i\in[N]},\{\boldsymbol{W}_i(s)\}_{i\in[N],s\le k\varepsilon},\boldsymbol{z}_1,\dots,\boldsymbol{z}_k).$$

We get:

$$\mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^k;\boldsymbol{z}_{k+1})|\mathcal{F}_k]=-\lambda\boldsymbol{\theta}_i^k-\nabla V(\boldsymbol{\theta}_i^k)-\frac{1}{N}\sum_{j=1}^N\nabla_1 U(\boldsymbol{\theta}_i^k,\boldsymbol{\theta}_j^k)=\boldsymbol{G}(\boldsymbol{\theta}_i^k,\rho_k^{(N)}),$$

where we denoted $\rho_k^{(N)}\equiv(1/N)\sum_{i\in[N]}\delta_{\boldsymbol{\theta}_i^k}$ the particle distribution of SGD. Hence we get

$$\begin{aligned}
\|\boldsymbol{\theta}_i^k-\tilde{\boldsymbol{\theta}}_i^k\|_2=&\Big\|\varepsilon\sum_{l=0}^{k-1}\boldsymbol{F}_i(\boldsymbol{\theta}_i^l;\boldsymbol{z}_{l+1})-\varepsilon\sum_{l=0}^{k-1}\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l;\tilde{\rho}_l^{(N)})\Big\|_2\\
\le&\Big\|\varepsilon\sum_{l=0}^{k-1}\boldsymbol{Z}_i^l\Big\|_2+\varepsilon\sum_{l=0}^{k-1}\Big\|\boldsymbol{G}(\boldsymbol{\theta}_i^l;\rho_l^{(N)})-\boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l;\tilde{\rho}_l^{(N)})\Big\|_2\\
\le&A_i^k+B_i^k,
\end{aligned}$$

where we denoted $\boldsymbol{Z}_i^l \equiv \boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1}) - \mathbb{E}[\boldsymbol{F}_i(\boldsymbol{\theta}^l; \boldsymbol{z}_{l+1})|\mathcal{F}_l]$ and $A_i^k = \|\varepsilon \sum_{l=0}^{k-1} \boldsymbol{Z}_i^l\|_2$.

Denote $\boldsymbol{A}_i^k = \sum_{l=0}^{k-1} \varepsilon \boldsymbol{Z}_i^l$. Hence $\{\boldsymbol{A}_i^k\}_{k \in \mathbb{N}}$ is a martingale adapted to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. Note the regularization term cancels out. We have component-wise

$$\begin{aligned}\boldsymbol{Z}_i^k =& \Big((y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))\sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k) - \mathbb{E}\left[(y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))\sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)|\mathcal{F}_k\right], \\ & (y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))a_i^k \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)) - \mathbb{E}\left[(y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))a_i^k \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)|\mathcal{F}_k\right]\Big).\end{aligned}$$

The following discussion is under the conditional law $\mathcal{L}(\cdot|\mathcal{F}_k)$. Note $|\sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)| \le K$, and $|y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k)| \le K(1 + \|\boldsymbol{a}^k\|_1/N)$, hence $(y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))\sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)$ is $K(1 + \|\boldsymbol{a}^k\|_1/N)^2$-sub-Gaussian. Note that by assumption, $\nabla_{\boldsymbol{w}} \sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)$ is K-sub-Gaussian (random vector), and $|(y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))a_i^k| \le K(1 + \|\boldsymbol{a}^k\|_1/N)\|\boldsymbol{a}^k\|_\infty$, hence $(y^{k+1} - \hat{y}(\boldsymbol{x}^{k+1}; \boldsymbol{\theta}^k))a_i^k \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{x}^{k+1}; \boldsymbol{w}_i^k)$ is a $K(1 + \|\boldsymbol{a}^k\|_1/N)^2 \|\boldsymbol{a}^k\|_\infty^2$-sub-Gaussian random vector. As a result, we have $\boldsymbol{F}_k(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})$ under the conditional law $\mathcal{L}(\cdot|\mathcal{F}_k)$ is a $K(1 + \|\boldsymbol{a}^k\|_1/N)^2\|\boldsymbol{a}^k\|_\infty^2$-sub-Gaussian random vector..

Let $\tau \equiv \inf\{k | \|\boldsymbol{a}^k\|_\infty \ge M_\infty \text{ or } \|\boldsymbol{a}^k\|_1 \ge N \cdot M_1\}$. Notice that $\boldsymbol{A}_i^{t\wedge\tau} - \boldsymbol{A}_i^{t\wedge\tau-1} = \boldsymbol{Z}_i^{k\wedge\tau-1}$. Following the same argument as in the proof of Proposition 8, we deduce that for $\overline{\boldsymbol{A}}_i^k \equiv \boldsymbol{A}_i^{k\wedge\tau}$, the martingale difference $\overline{\boldsymbol{A}}_i^k - \overline{\boldsymbol{A}}_i^{k-1}$ is $\varepsilon^2 K^2 M_1^2 M_\infty^2$-sub-Gaussian under the conditional law $\mathcal{L}(\cdot|\mathcal{F}_k)$. We apply Azuma-Hoeffding's inequality (Lemma 31)

$$\mathbb{P}\Big(\max_{k \in [0, T/\varepsilon]\cap\mathbb{N}} \|\overline{\boldsymbol{A}}_i^k\|_2 \ge K M_1 M_\infty \sqrt{\varepsilon}\left[\sqrt{D} + z\right]\Big) \le e^{-z^2}.$$

We get:

$$\begin{aligned}&\mathbb{P}\Big(\max_{k \in [0, T/\varepsilon]\cap\mathbb{N}} \|\boldsymbol{A}_i^k\|_2 \ge K M_1 M_\infty \sqrt{\varepsilon}\left[\sqrt{D} + z\right]\Big) \\ \le & \mathbb{P}\Big(\max_{k \in [0, T/\varepsilon]\cap\mathbb{N}} \|\overline{\boldsymbol{A}}_i^k\|_2 \ge K M_1 M_\infty \sqrt{\varepsilon}\left[\sqrt{D} + z\right]\Big) + \mathbb{P}(\tau \le T/\varepsilon) \\ \le & 2e^{-z^2},\end{aligned}$$

where we used the high probability bound of $\sup_{k \in [0,T/\varepsilon]\cap\mathbb{N}}\{\|\boldsymbol{a}^k\|_1, \|\boldsymbol{a}^k\|_1\}$ in Lemma 19. Taking the union bound over $i \in [N]$ yields

$$\mathbb{P}\Big(\max_{i \le N} \max_{k \in [0, T/\varepsilon]\cap\mathbb{N}} A_i^k \ge K e^{KT}\left[\sqrt{D}\log N + \log^{3/2} N + z^3\right]\sqrt{\varepsilon}\Big) \le e^{-z^2}.$$

For the second term, we get:

$$\begin{aligned}&\|\boldsymbol{G}(\boldsymbol{\theta}_i^l, \rho_l^{(N)}) - \boldsymbol{G}(\tilde{\boldsymbol{\theta}}_i^l; \tilde{\rho}_l^{(N)})\|_2 \\ \le & \lambda\|\boldsymbol{\theta}_i^l - \tilde{\boldsymbol{\theta}}_i^l\|_2 + \|\nabla V(\boldsymbol{\theta}_i^l) - \nabla V(\tilde{\boldsymbol{\theta}}_i^l)\|_2 + \Big\|\frac{1}{N}\sum_{j=1}^N \nabla_1 U(\boldsymbol{\theta}_i^l, \boldsymbol{\theta}_j^l) - \nabla_1 U(\tilde{\boldsymbol{\theta}}_i^l, \tilde{\boldsymbol{\theta}}_j^l)\Big\|_2 \\ \le & K(1 + \|\boldsymbol{a}^l\|_\infty) \cdot \|\boldsymbol{\theta}_i^l - \tilde{\boldsymbol{\theta}}_i^l\|_2 + K(1 + \|\bar{\boldsymbol{a}}^l\|_\infty) \cdot (1 + \|\bar{\boldsymbol{a}}^l\|_1/N) \cdot \max_{j \in [N]} \|\boldsymbol{\theta}_j^l - \tilde{\boldsymbol{\theta}}_j^l\|_2 \\ & + K(1 + \|\tilde{\boldsymbol{a}}^l\|_\infty) \cdot (1 + \|\tilde{\boldsymbol{a}}^l\|_1/N) \cdot \max_{j \in [N]} \|\boldsymbol{\theta}_j^l - \tilde{\boldsymbol{\theta}}_j^l\|_2.\end{aligned}$$

Using the high probability bound on $\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\{\|\boldsymbol{a}^k\|_1/N, \|\boldsymbol{a}^k\|_\infty, \|\tilde{\boldsymbol{a}}^k\|_1/N, \|\tilde{\boldsymbol{a}}^k\|_\infty\}$ of Lemma 19, we get with probability at least $1 - e^{-z^2}$ that for all $t \in [0, T]$

$$\Delta(t) \le K e^{KT}(1 + z)\left[\sqrt{\log N} + z\right]\int_0^t \Delta(s)\mathrm{d}s + K e^{KT}\left[\sqrt{D}\log N + \log^{3/2} N + z^3\right]\sqrt{\varepsilon}.$$

Applying Gronwall's inequality, we get:

$$\mathbb{P}\Big(\Delta(T) \le K e^{e^{KT}[\sqrt{\log N} + z^2]}\left[\sqrt{D}\log N + \log^{3/2} N + z^3\right]\sqrt{\varepsilon}\Big) \ge 1 - e^{-z^2}.$$

This bound combined with Lemma 21 concludes the proof. □

F Existence and uniqueness of PDEs solutions

F.1 Equation (DD) (noiseless SGD)

For the readers convenience, we reproduce here the form of the limiting PDE

$$\partial_t \rho_t = 2\xi(t)\nabla \cdot \big(\rho_t \nabla \Psi(\boldsymbol{\theta}; \rho_t)\big), \tag{59}$$

$$\Psi(\boldsymbol{\theta}; \rho_t) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, \rho_t(\mathrm{d}\tilde{\boldsymbol{\theta}}). \tag{60}$$

This PDE describes an evolution in the space of probability distribution on $\mathbb{R}^D$ and has to be interpreted in the weak sense. Namely ρ_t is a solution of Eq. (59), if for any bounded function $h : \mathbb{R}^D \mapsto \mathbb{R}$ differentiable with bounded gradient:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int h(\boldsymbol{\theta})\rho_t(\mathrm{d}\boldsymbol{\theta}) = -2\xi(t)\int \langle \nabla h(\boldsymbol{\theta}), \nabla\Psi(\boldsymbol{\theta}; \rho_t\rangle \rho_t(\mathrm{d}\boldsymbol{\theta}). \tag{61}$$

For fixed coefficient, under assumptions A1, A2, A3, A4, we have $\nabla V(\boldsymbol{\theta})$ and $\nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')$ bounded Lipschitz. By [Szn91, Theorem 1.1], these assumptions are sufficient to guarantee the existence and uniqueness of solution of PDE (59).

For general coefficients, the potentials are not bounded and Lipschitz anymore. The existence and uniqueness under assumptions A1, A2, A3, A4, can be derived by a similar argument as in [SS18, Section 4], which uses an adaptation of the argument of [Szn91, Theorem 1.1].

F.2 Equation (diffusion-DD) (noisy SGD)

For the readers convenience, we reproduce here the form of the limiting PDE

$$\partial_t \rho_t = 2\xi(t)\nabla \cdot \big(\rho_t \nabla \Psi_\lambda(\boldsymbol{\theta}; \rho_t)\big) + 2\xi(t)/\beta \Delta_{\boldsymbol{\theta}} \rho_t, \tag{62}$$

$$\Psi_\lambda(\boldsymbol{\theta}; \rho_t) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, \rho_t(\mathrm{d}\tilde{\boldsymbol{\theta}}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2. \tag{63}$$

We say that ρ_t is a weak solution of Eq. (62) if for any $\zeta \in C_0^\infty(\mathbb{R} \times \mathbb{R}^D)$ (the space of smooth functions decaying to 0 at infinity), we have for any $T > 0$

$$\begin{aligned}&\int_{\mathbb{R}^D} \zeta_0(\boldsymbol{\theta})\rho_0(\mathrm{d}\boldsymbol{\theta}) - \int_{\mathbb{R}^D} \zeta_0(\boldsymbol{\theta})\rho_T(\mathrm{d}\boldsymbol{\theta})\\ &= -\int_{(0,T)\times\mathbb{R}^D} [\partial_t \zeta_t(\boldsymbol{\theta}) - 2\xi(t)\langle \nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta}; \rho_t), \nabla_{\boldsymbol{\theta}}\zeta_t(\boldsymbol{\theta})\rangle + 2\xi(t)\Delta_{\boldsymbol{\theta}}\zeta_t(\boldsymbol{\theta})]\rho_t(\mathrm{d}\boldsymbol{\theta})\mathrm{d}t.\end{aligned} \tag{64}$$

Note that this notion of weak solution is equivalent to the one introduced earlier in Eq. (61), see for instance [San15, Proposition 4.2].

For fixed coefficients, the existence and uniqueness of solution of Eq. (62) was proven in [MMN18, Section 10.2], under the assumptions A1, A2, A3, A6. The proof follows from an adaptation of the proof of [JKO98, Theorem 5.1].

For general coefficients, we can follow a similar contraction argument as in [SS18, Section 4] and [Szn91, Theorem 1.1], by bounding more carefully each term.

Proposition 17. *Assume conditions* A1-A5. *Then PDE* (62) *admits a weak solution* $(\rho_t)_{t\ge 0}$ *which is unique.*

Proof of Lemma 17. Without loss of generality, we assume $\xi(t) = 1/2$, which corresponds to a reparametrization of variable time t. Denote by $\mathscr{P}(\mathbb{R}^D)$ the set of probability measures on $\mathbb{R}^D$, endowed with the topology of weak convergence. Note that Eq. (64) immediately implies that $t \mapsto \rho_t$ is continuous in $\mathscr{P}(\mathbb{R}^D)$.

Denote by $D([0,T]; \mathscr{P}(\mathbb{R}^D))$ the set of maps from $[0,T]$ into $\mathscr{P}(\mathbb{R}^D)$ and by $C([0,T]; \mathscr{P}(\mathbb{R}^D))$ the set of continuous maps in this class. We introduce the map $\Phi_T : C([0,T]; \mathscr{P}(\mathbb{R}^D)) \to D([0,T]; \mathscr{P}(\mathbb{R}^D))$, which associates $m \in D([0,T]; \mathscr{P}(\mathbb{R}^D))$ to the law of the solution

$$\bar{\boldsymbol{\theta}}^t = \bar{\boldsymbol{\theta}}^0 + \int_0^t \boldsymbol{G}(\bar{\boldsymbol{\theta}}^s; m_s)\mathrm{d}s + \overline{\boldsymbol{W}}(t), \qquad \text{for } t \le T,\ \bar{\boldsymbol{\theta}}_0 \sim \rho_0.$$

Observe that if m is a weak solution of PDE (62) defined on interval $[0,T]$, then m is a fixed point of Φ_T. Further, for any such fixed point m, Lemma 18 and Lemma 20 both apply. In particular, $t \mapsto m_t$ is continuous in $\mathscr{P}(\mathbb{R}^D)$ and therefore Φ_T maps $C([0,T];\mathscr{P}(\mathbb{R}^D))$ to $C([0,T];\mathscr{P}(\mathbb{R}^D))$. Further, again by the same derivation, there exists a constant C, such that

$$\int a^2 m_t(\mathrm{d}a) \le Ce^{Ct}, \qquad \text{for all } t \in [0,T].$$

Let us define $\mathscr{P}_{C_0,T_0}(\mathbb{R}^D)$ the space of probability measures such that $\int a^2\mu(\mathrm{d}a) \le C_0 e^{C_0 T_0}$. We consider $m \in C([0,T_0];\mathscr{P}_{C_0,T_0}(\mathbb{R}^D))$, the set of continuous mapping from $[0,T_0]$ on $\mathscr{P}_{C_0,T_0}(\mathbb{R}^D)$. Using the same computation as in the proof of Lemma 18, we have:

$$\bar{a}^t = e^{-\lambda t}\bar{a}^0 + \int_0^t e^{-\lambda(t-s)} K(\bar{\boldsymbol{w}}^s, m_s)\mathrm{d}s + \int_0^t e^{-\lambda(t-s)}\sqrt{\tau/D}\,\mathrm{d}W^a(s),$$

where $|K(\bar{\boldsymbol{w}}^s, m_s)| = \Big| - v(\bar{\boldsymbol{w}}^s) - \int a u(\bar{\boldsymbol{w}}^s, \boldsymbol{w}) m_s(\mathrm{d}a, \mathrm{d}\boldsymbol{w})\Big| \le K + K\sqrt{C_0}e^{C_0 s/2}$. We get:

$$(\bar{a}^t)^2 \le 9K^2 + 18K^2 t + 18K^2 t C_0 e^{C_0 t} + B_t^2,$$

where B_t is a normal random variable with variance bounded by $9t\tau/D$. Taking the expectation with respect to $\Phi_T(m)$, we get:

$$\int a^2 \Phi_T(m)_t(\mathrm{d}a) \le (9K^2 + 18K^2 t + 18K^2 t C_0 + 9t\tau/D)e^{C_0 t}.$$

Hence we deduce that for C_0 sufficiently big and T_0 sufficiently small, we have for every $T \in [0,T_0]$, $\Phi_T(m) \in C([0,T];\mathscr{P}_{C_0,T}(\mathbb{R}^D))$. We can therefore restrict our mapping Φ to the subsets $C([0,T];\mathscr{P}_{C_0,T}(\mathbb{R}^D))$ for $T \le T_0$, which must contains all the fixed points by the above discussion.

We introduce the following metric on $C([0,T];\mathscr{P}_{C_0,T}(\mathbb{R}^D))$:

$$\mathscr{D}_T(m^1, m^2) = \Big(\inf\Big\{\int \sup_{t\le T}\|\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t\|_2^2 \gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2) : \gamma \text{ is a coupling of } m^1, m^2\Big\}\Big)^{1/2}.$$

We show that for $T_1 \le T_0$ sufficiently small, the mapping Φ_{T_1} is a contraction with respect to this distance.

Lemma 26. *There exists a constant K such that, for all $T \le T_0$, and for all $m^1, m^2 \in C([0,T];\mathscr{P}_{C_0,T}(\mathbb{R}^D))$, we have*

$$\mathscr{D}_T(\Phi_T(m^1), \Phi_T(m^2)) \le TK\mathscr{D}_T(m^1, m^2).$$

Proof of Lemma 26. Fix $T \le T_0$, and consider a coupling γ between $m^1, m^2 \in C([0,T];\mathscr{P}_{C_0,T}(\mathbb{R}^D))$. We consider the following coupling between $\Phi_T(m^1)$ and $\Phi_T(m^2)$:

$$\begin{aligned}\bar{\boldsymbol{\theta}}_1^t &= \bar{\boldsymbol{\theta}}^0 + \int \boldsymbol{G}_1(\bar{\boldsymbol{\theta}}_1^s;\gamma_s)\mathrm{d}s + \overline{\boldsymbol{W}}(t),\\ \bar{\boldsymbol{\theta}}_2^t &= \bar{\boldsymbol{\theta}}^0 + \int \boldsymbol{G}_2(\bar{\boldsymbol{\theta}}_2^s;\gamma_s)\mathrm{d}s + \overline{\boldsymbol{W}}(t),\end{aligned}$$

where $\boldsymbol{G}_1(\bar{\boldsymbol{\theta}}_1^s;\gamma_s) = -\lambda\bar{\boldsymbol{\theta}}_1^s - \nabla V(\bar{\boldsymbol{\theta}}_1^s) - \int_{\mathbb{R}^D\times\mathbb{R}^D}\nabla_1 U(\bar{\boldsymbol{\theta}}_1^s, \boldsymbol{\theta}_1)\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)$ (and similarly for $\boldsymbol{G}_2$). We have:

$$\begin{aligned}\|\boldsymbol{G}_1(\bar{\boldsymbol{\theta}}_1^s;\gamma_s) - \boldsymbol{G}_2(\bar{\boldsymbol{\theta}}_2^s;\gamma_s)\|_2 \le & K(1+|\bar{a}_1^s|)\|\bar{\boldsymbol{\theta}}_1^s - \bar{\boldsymbol{\theta}}_2^s\|_2 + K\int |a_1|(1+|\bar{a}_1^s|)\|\bar{\boldsymbol{\theta}}_1^s - \bar{\boldsymbol{\theta}}_2^s\|_2\gamma_s(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\ &+ \int (1+|a_1|)(1+|\bar{a}_2^s|)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2\gamma_s(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2).\end{aligned}$$

Hence, we get (using that $m_s^1 \in \mathcal{P}_{C_0,T}(\mathbb{R}^D)$)

$$\|\bar{\boldsymbol{\theta}}_1^t - \bar{\boldsymbol{\theta}}_2^t\|_2 \le Ke^{KT_0}\int_0^t (1+|\bar{a}_1^s|)\|\bar{\boldsymbol{\theta}}_1^s - \bar{\boldsymbol{\theta}}_2^s\|_2\mathrm{d}s + K\int_0^t (1+|\bar{a}_2^s|)\int (1+|a_1|)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2\gamma_s(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\mathrm{d}s,$$

where K is a constant depending on the constants of the assumptions and C_0. Taking the square and using Cauchy-Schwartz inequality

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}_1^t-\bar{\boldsymbol{\theta}}_2^t\|_2^2 \le& Ke^{KT_0}\int_0^t(1+|\bar{a}_1^s|)^2\mathrm{d}s\int_0^t\|\bar{\boldsymbol{\theta}}_1^s-\bar{\boldsymbol{\theta}}_2^s\|_2^2\mathrm{d}s\\ &+K\int_0^t(1+|\bar{a}_2^s|)^2\mathrm{d}s\int_0^t\Big(\int(1+|a_1|)^2\gamma_s(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2)\int\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2\gamma_s(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2)\Big)\mathrm{d}s\\ \le& Ke^{KT_0}T_0M_{T_0}\int_0^t\|\bar{\boldsymbol{\theta}}_1^s-\bar{\boldsymbol{\theta}}_2^s\|_2^2\mathrm{d}s+Ke^{KT_0}M_{T_0}t^2\int\sup_{t\le T}\|\boldsymbol{\theta}_1^t-\boldsymbol{\theta}_2^t\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2),\end{aligned}$$

where $M_{T_0}=(1+\sup_{t\le T_0}(|\bar{a}_1^t|\vee|\bar{a}_2^t|))^2$. Applying Gronwall's lemma, we get, for any $T<T_0$,

$$\sup_{t\le T}\|\bar{\boldsymbol{\theta}}_1^t-\bar{\boldsymbol{\theta}}_2^t\|_2^2\le KT^2e^{KT_0^2e^{KT_0}M_{T_0}}\int\sup_{t\le T}\|\boldsymbol{\theta}_1^t-\boldsymbol{\theta}_2^t\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2).$$

Taking the expectation:

$$\mathbb{E}[\sup_{t\le T}\|\bar{\boldsymbol{\theta}}_1^t-\bar{\boldsymbol{\theta}}_2^t\|_2^2]\le KT^2\mathbb{E}\{\exp(KT_0^2e^{KT_0}M_{T_0})\}\int\sup_{t\le T}\|\boldsymbol{\theta}_1^t-\boldsymbol{\theta}_2^t\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2).$$

By a similar argument as in Lemma 19, we have $\mathbb{P}(M_{T_0}\ge Ke^{KT_0}(1+z^2))\le e^{-z^2}$, i.e.

$$\mathbb{P}(\exp\{KT_0^2e^{KT_0}M_{T_0}\}\ge\exp\{KT_0^2e^{KT_0}(1+z^2)\})\le e^{-z^2}.$$

Doing a change of variable, we get:

$$\begin{aligned}\mathbb{E}\{\exp(KT_0^2e^{KT_0}M_{T_0})\}=&\int\mathbb{P}(\exp(KT_0^2e^{KT_0}M_{T_0})\ge u)\mathrm{d}u\\ \le& KT_0^2e^{KT_0}+KT_0^2e^{KT_0^2e^{KT_0}}\int_0^\infty z\exp\{-(1-KT_0^2e^{KT_0})z^2\}\mathrm{d}z<\infty,\end{aligned}$$

for T_0 small enough. We conclude that there exists a constant $K<\infty$ such that

$$\mathscr{D}_T(\Phi_T(m^1),\Phi_T(m^2))\le(\inf_\gamma\mathbb{E}[\sup_{t\le T}\|\bar{\boldsymbol{\theta}}_1^t-\bar{\boldsymbol{\theta}}_2^t\|_2^2])^{1/2}\le TK\mathscr{D}_T(m^1,m^2),$$

where we used that the coupling γ was chosen arbitrarily. □

We can therefore consider $T_1<1/K$. The mapping Φ_{T_1} is a contraction on the space $C([0,T_1];\mathscr{P}_{C_0,T_1}(\mathbb{R}^D))$. By the Banach fixed-point theorem, there exists a fixed point for Φ_{T_1} on the interval $[0,T_1]$, which is unique. We can further iterate the same argument. Assume that the fixed point of Φ_T is unique, for some $T>0$. Then $\Phi_{[0,T+T_1]}$ has a unique fixed point, which is a map $m:[0,T+T_1]\to\mathscr{P}(\mathbb{R}^D)$. This suffices to conclude that PDE (62) admits a weak solution on $[0,\infty)$, and this solution is unique. □

Further, Duhamel's principle for PDE (62) holds. Denote $\mathscr{G}(\boldsymbol{\theta},\boldsymbol{\theta}';t)$ the heat kernel:

$$\mathscr{G}(\boldsymbol{\theta},\boldsymbol{\theta}';t)\equiv\frac{1}{(2\pi t)^{d/2}}\exp\{-\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2/(2t)\}.$$

Lemma 27. *Assume conditions* A1-A5. *Let ρ be a weak solution of PDE* (62). *Then, for any $t>0$, $\rho_t(\mathrm{d}\boldsymbol{\theta})$ has a density, denoted $\rho(t,\cdot)$, which satisfies*

$$\rho(t,\boldsymbol{\theta})=\int\mathscr{G}(\boldsymbol{\theta},\boldsymbol{\theta}_1;\tau t/D)\rho_0(\mathrm{d}\boldsymbol{\theta}_1)-\int_0^t\int\langle\nabla_{\boldsymbol{\theta}_1}\mathscr{G}(\boldsymbol{\theta},\boldsymbol{\theta}_1;\tau(t-s)/D),\nabla_{\boldsymbol{\theta}_1}\Psi(\boldsymbol{\theta}_1;\rho_s)\rangle\rho(s,\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}s.$$

Proof of Lemma 27. For ease of notation, let us set $\tau/D = 1$ and $\xi(t) = 1/2$, which amounts to rescaling time. Consider $\eta \in C^\infty(\mathbb{R}^D)$ (space of smooth real-valued functions) with bounded support, and define:

$$\mathscr{G}_\eta(\boldsymbol{\theta}; t) = \int \mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\theta}_1; t)\eta(\boldsymbol{\theta}_1)\, \mathrm{d}\boldsymbol{\theta}_1.$$

By property of the heat kernel, we have

$$(\partial_t - \Delta)\mathscr{G}_\eta(\boldsymbol{\theta}; t) = 0, \qquad \forall t > 0, \forall \boldsymbol{\theta} \in \mathbb{R}^D.$$

Take $\zeta(\boldsymbol{\theta}, s) = \mathscr{G}_\eta(\boldsymbol{\theta}; t - s)$ (which indeed decays to 0 at infinity) as a test function in Eq. (64) for $T = t$. We get:

$$\int \eta(\boldsymbol{\theta}_1)\rho_t(\mathrm{d}\boldsymbol{\theta}_1) = \int \mathscr{G}_\eta(\boldsymbol{\theta}; t)\rho_0(d\boldsymbol{\theta}) - \int_{(0,t)\times\mathbb{R}^D} \langle \mathscr{G}_\eta(\boldsymbol{\theta}; t - s), \nabla\Psi(\boldsymbol{\theta}; \rho_s)\rangle \rho_s(\mathrm{d}\boldsymbol{\theta})\mathrm{d}s.$$

By applying Fubini's theorem, we get

$$\begin{aligned}&\int \eta(\boldsymbol{\theta}_1)\rho_t(\mathrm{d}\boldsymbol{\theta}_1)\\ =&\int \mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\theta}_1; t)\rho_0(d\boldsymbol{\theta})\eta(\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1 - \int_{(0,t)\times\mathbb{R}^D\times\mathbb{R}^D} \langle \mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\theta}_1; t - s), \nabla\Psi(\boldsymbol{\theta}, \boldsymbol{\theta}_1; \rho_s)\rangle \rho_s(\mathrm{d}\boldsymbol{\theta})\mathrm{d}s\ \eta(\boldsymbol{\theta}_1)\mathrm{d}\boldsymbol{\theta}_1,\end{aligned}$$

where η is an arbitrary function with bounded support, which concludes the proof. □

Lemma 28. *Assume conditions* A1- A6*. Assume further that ρ_0 has a density. Denote $(\rho_t)_{t\geq 0}$ the solution of PDE* (62)*, with density $(\rho(t, \cdot))_{t\geq 0}$. Then $(t, \boldsymbol{\theta}) \mapsto \rho(t, \boldsymbol{\theta})$ is in $C^{1,2}((0, \infty)\times\mathbb{R}^D)$, where $C^{1,2}((0, \infty)\times\mathbb{R}^D)$ is the function space of continuous function with continuous derivative in time, and second order continuous derivative in space.*

Proof of Lemma 28. The proof follows exactly from the proof of Lemma [MMN18, Lemma 10.7]. □

F.3 The noisy PDE as a gradient flow in the space of probability distributions

We include a second independent proof of the existence of a weak solution, which is interesting in itself. It relies on a deep connection pioneered by [JKO98], between Fokker-Planck PDEs and gradient flow in probability space. The proof follows closely the steps detailed in [JKO98]. The arguments are similar to [MMN18, Section 10.2], and we will only detail the differences.

We will consider the set $\mathcal{K}$ of admissible probability densities,

$$\mathcal{K} = \Big\{\rho : \mathbb{R}^D \mapsto [0, +\infty) \text{ measurable } : \int_{\mathbb{R}^D} \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1, M(\rho) < \infty\Big\},$$

where

$$M(\rho) \equiv \int_{\mathbb{R}^D} \|\boldsymbol{\theta}\|_2^2 \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Recall

$$R(\rho) = \mathbb{E}(y^2) + 2\int_{\mathbb{R}^D} V(\boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \int_{\mathbb{R}^D\times\mathbb{R}^D} U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\boldsymbol{\theta})\rho(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}'.$$

We will define

$$\mathrm{Ent}(\rho) = -\int_{\mathbb{R}^D} \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$
$$F(\rho) = 1/2 \cdot [\lambda M(\rho) + R(\rho)] - 1/\beta \cdot \mathrm{Ent}(\rho).$$

The PDE (62) can be interpreted as a gradient flow on the free energy functional $F(\rho)$ in the space of probability measures on $\mathbb{R}^D$ endowed with the $W_2(\cdot, \cdot)$ Wassertein distance [MMN18, Section 10.2]. Recall that for μ, ν probability distributions over $\mathbb{R}^D$, we have:

$$W_2^2(\mu, \nu) = \inf\Big\{\int_{\mathbb{R}^D\times\mathbb{R}^D} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2) : \ \gamma \text{ is a coupling of } \mu, \nu\Big\}$$

Proposition 18. *Assume conditions* A1, A2, A3, A6. *Let initialization $\rho_0 \in \mathcal{K}$ so that $F(\rho_0) < \infty$. Then the PDE* (62) *admits a weak solution $(\rho_t)_{t \geq 0}$ which is unique. Moreover, for any fixed t, $\rho_t \in \mathcal{K}$ is absolutely continuous with respect to the Lebesgue measure, and $M(\rho_t)$ and $\mathrm{Ent}(\rho_t)$ are uniformly bounded in t.*

Proof of Proposition 18. Without loss of generality, we assume $\xi(t) = 1/2$, which corresponds to a reparametrization of variable time t. To prove the existence of the solution, we consider the limit of the following discretized scheme when the step-size h goes to zero: we define recursively a sequence of distributions $\{\overline{\rho}_k^h\}_{k \in \mathbb{N}}$, with $\overline{\rho}_0^h = \rho_0$ and

$$\overline{\rho}_{k+1}^h \in \arg\min_{\rho \in \mathcal{K}} \Big\{ hF(\rho) + \frac{1}{2} W_2^2(\rho, \overline{\rho}_k^h) \Big\}. \tag{65}$$

Lemma 29. *Given an initialization $\rho_0 \in \mathcal{K}$, there exists a unique solution of the scheme* (65).

Proof of Lemma 29. Clearly it is sufficient to analyze a single step of the scheme (65). The proof follows from the same arguments as in [JKO98, Proposition 4.1], which shows that there exists a sequence of measures $\{\rho_\nu\}_{\nu \in \mathbb{N}} \in \mathcal{K}$ that converges weakly to $\rho^* \in \mathcal{K}$ such that

$$\lim_{\nu \to \infty} \Big\{ F(\rho_\nu) + \frac{1}{2} W_2^2(\rho_\nu, \rho_0) \Big\} = \inf_{\rho \in \mathcal{K}} \Big\{ F(\rho) + \frac{1}{2} W_2^2(\rho_\nu, \rho_0) \Big\} > -\infty.$$

Moreover, there exists a constant C such that $M(\rho_\nu) \leq C$ and $M(\rho^*) \leq C$ by lower semi-continuity of $M(\rho)$. We only need to check lower semi-continuity of $R(\rho)$ to conclude that ρ^* is indeed a minimizer. Uniqueness comes from convexity of the functional and strict convexity of $-\mathrm{Ent}(\rho)$.

Denote for $x \in \mathbb{R}$, the functions $\overline{\phi}_m(x) = \mathrm{sign}(x) \cdot \max\{|x| - m, 0\}$ and $\underline{\phi}_m(x) = x - \overline{\phi}_m(x)$, and $\mathsf{B}(r) = \mathsf{B}(0, r) \subset \mathbb{R}^D$:

$$\begin{aligned} |R(\rho_\nu) - R(\rho^*)| \leq &\Big| \int \underline{\phi}_m(V(\boldsymbol{\theta}))[\rho_\nu(\boldsymbol{\theta}) - \rho^*(\boldsymbol{\theta})]\mathrm{d}\boldsymbol{\theta} \Big| + \Big| \int \underline{\phi}_m(U(\boldsymbol{\theta}, \boldsymbol{\theta}'))[\rho_\nu(\boldsymbol{\theta})\rho_\nu(\boldsymbol{\theta}') - \rho^*(\boldsymbol{\theta})\rho^*(\boldsymbol{\theta}')]\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}' \Big| \\ &+ \Big| \int \overline{\phi}_m(V(\boldsymbol{\theta}))[\rho_\nu(\boldsymbol{\theta}) - \rho^*(\boldsymbol{\theta})]\mathrm{d}\boldsymbol{\theta} \Big| + \Big| \int \overline{\phi}_m(U(\boldsymbol{\theta}, \boldsymbol{\theta}'))[\rho_\nu(\boldsymbol{\theta})\rho_\nu(\boldsymbol{\theta}') - \rho^*(\boldsymbol{\theta})\rho^*(\boldsymbol{\theta}')]\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}' \Big|. \end{aligned}$$

By weak convergence in $L^1(\mathbb{R}^D)$, the first two terms converge to zero. Recalling that $V(\boldsymbol{\theta}) = av(\boldsymbol{w})$ and $U(\boldsymbol{\theta}, \boldsymbol{\theta}') = aa'u(\boldsymbol{w}, \boldsymbol{w}')$, with $|v(\boldsymbol{w})| \leq K$ and $|u(\boldsymbol{w}, \boldsymbol{w}')| \leq K$, we deduce

$$\Big| \int \overline{\phi}_m(V(\boldsymbol{\theta}))[\rho_\nu(\boldsymbol{\theta}) - \rho^*(\boldsymbol{\theta})]\mathrm{d}\boldsymbol{\theta} \Big| \leq \Big| \int_{\mathsf{B}(m/K)} \overline{\phi}_m(V(\boldsymbol{\theta}))[\rho_\nu(\boldsymbol{\theta}) - \rho^*(\boldsymbol{\theta})]\mathrm{d}\boldsymbol{\theta} \Big| \leq 2KC/m,$$

and

$$\begin{aligned} &\Big| \int \overline{\phi}_m(U(\boldsymbol{\theta}, \boldsymbol{\theta}'))[\rho_\nu(\boldsymbol{\theta})\rho_\nu(\boldsymbol{\theta}') - \rho^*(\boldsymbol{\theta})\rho^*(\boldsymbol{\theta}')]\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}' \Big| \\ \leq &\Big| \int_{\mathsf{B}(\sqrt{m}/K) \times \mathsf{B}(\sqrt{m}/K)} \overline{\phi}_m(U(\boldsymbol{\theta}, \boldsymbol{\theta}'))[\rho_\nu(\boldsymbol{\theta})\rho_\nu(\boldsymbol{\theta}') - \rho^*(\boldsymbol{\theta})\rho^*(\boldsymbol{\theta}')]\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}' \Big| \leq 2KC^2/m, \end{aligned}$$

where we used that $\int_{\mathsf{B}(r)} |a| \rho_\nu(\mathrm{d}a) \leq \int a^2/r \rho_\nu(\mathrm{d}a) \leq C/r$. Because m is arbitrarily large, we conclude that

$$\lim_{\nu \to \infty} \big| R(\rho_\nu) - R(\rho^*) \big| = 0.$$

□

The rest of the proof follows the proof of [JKO98, Theorem 5.1], which shows that for a given $T < \infty$, there exists C such that for any h and k with $hk \leq T$, we have $M(\overline{\rho}_k^h) \leq C$. If we denote $\rho^h(t, .)$ the piece wise constant distribution trajectory, we deduce that it converges weakly to ρ in $L^1((0, T) \times \mathbb{R}^D)$. Furthermore, the weak convergence applies for each given time $t \in [0, +\infty)$, i.e. $\rho^h(t) \mapsto \rho(t)$ weakly.

We still need to show that this limiting distribution is a weak solution (61) of PDE (62). Let $\boldsymbol{\xi} \in C_0^\infty(\mathbb{R}^D, \mathbb{R}^D)$ be a smooth vector field with bounded support, and define $\{\Phi_\tau\}_{\tau \in \mathbb{R}}$ the corresponding flux:

$$\partial_\tau \Phi_\tau = \boldsymbol{\xi} \circ \Phi_\tau \text{ for all } \tau \in \mathbb{R} \text{ and } \Phi_0 = \mathrm{id}. \tag{66}$$

Further, for $\tau \in \mathbb{R}$, define ν_τ to be the push forward measure of $\overline{\rho}_k^h$ under Φ_τ. Namely,

$$\int_{\mathbb{R}^D} \nu_\tau(\boldsymbol{\theta})\zeta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\mathbb{R}^D} \overline{\rho}_k^h(\boldsymbol{\theta})\zeta(\Phi_\tau(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}, \qquad \forall \zeta \in C(\mathbb{R}^D),$$

or equivalently $\nu_\tau = \frac{1}{\det \nabla \Phi_\tau}\overline{\rho}_k^h \circ \Phi_\tau^{-1}$. We only need to consider the term $R(\rho)$. See the proof of [MMN18, Lemma 10.6] for more details.

From the assumption of bounded support, we must have $\sup_{\boldsymbol{\theta} \in \mathbb{R}^D} \|\boldsymbol{\xi}(\boldsymbol{\theta})\|_2 \leq K$. From Eq. (66), we have

$$\Phi_\tau(\boldsymbol{\theta}) = \boldsymbol{\theta} + \int_0^\tau \Phi_s(\boldsymbol{\xi}(\boldsymbol{\theta}))\mathrm{d}s. \tag{67}$$

Hence applying Gronwall's inequality to $u(\tau) = \sup_{\boldsymbol{\theta} \in \mathsf{B}(r)} \|\Phi_\tau(\boldsymbol{\theta})\|_2$, and considering $\tau \leq 1$, we get $u(\tau) \leq K$. Therefore, for $\tau \leq 1$, we get $|(\partial^2/\partial\tau^2)\Phi_\tau(\boldsymbol{\theta})| = |\Phi_\tau(\boldsymbol{\xi}(\boldsymbol{\xi}(\boldsymbol{\theta})))| \leq K$. We deduce that

$$\|\Phi_\tau(\boldsymbol{\theta}) - \boldsymbol{\theta} - \tau\boldsymbol{\xi}(\boldsymbol{\theta})\|_2 \leq K\tau^2. \tag{68}$$

Let us consider the derivative of $R(v_\tau)$ with respect to τ. Recall that U is symmetric.

$$\begin{aligned}
&\int [U(\Phi_\tau(\boldsymbol{\theta}_1), \Phi_\tau(\boldsymbol{\theta}_2)) - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - 2\tau\langle \nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle]\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2\\
=&\int [U(\Phi_\tau(\boldsymbol{\theta}_1), \Phi_\tau(\boldsymbol{\theta}_2)) - U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - \tau\langle \nabla_2 U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle]\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2\\
&+ \int [U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - \tau\langle \nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle]\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2\\
&+ \int [\tau\langle \nabla_2 U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle - \tau\langle \nabla_2 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle]\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2.
\end{aligned}$$

Denote $(a_1^\tau, \boldsymbol{w}_1^\tau) = \Phi_\tau(\boldsymbol{\theta}_1)$ and $(a_2^\tau, \boldsymbol{w}_2^\tau) = \Phi_\tau(\boldsymbol{\theta}_2)$, and $\boldsymbol{\xi}(\boldsymbol{\theta}) = (\xi_a(\boldsymbol{\theta}), \boldsymbol{\xi}_w(\boldsymbol{\theta}))$. Consider the first term

$$\begin{aligned}
&U(\Phi_\tau(\boldsymbol{\theta}_1), \Phi_\tau(\boldsymbol{\theta}_2)) - U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - \tau\langle \nabla_2 U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle\\
=&a_1^\tau\{[a_2^\tau - a_2]u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2^\tau) + a_2[u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2) - u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2^\tau)] - \tau\xi_a(\boldsymbol{\theta}_2)u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2) - \tau a_2\langle \nabla_{\boldsymbol{w}_2} u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2), \boldsymbol{\xi}_w(\boldsymbol{\theta}_2)\rangle\}\\
=&a_1^\tau\{[a_2^\tau - a_2 - \tau\xi_a(\boldsymbol{\theta}_2)]u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2^\tau) + a_2[u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2) - u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2^\tau) - \tau\langle \nabla_{\boldsymbol{w}_2} u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2), \boldsymbol{\xi}_w(\boldsymbol{\theta}_2)\rangle]\}\\
&+ \tau a_1^\tau\xi_a[u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2^\tau) - u(\boldsymbol{w}_1^\tau, \boldsymbol{w}_2)].
\end{aligned}$$

Using that $\|\nabla u\|_{\mathrm{op}}, \|\nabla^2 u\|_{\mathrm{op}} \leq K$, and Eq. (67) and Eq. (68), we get for $\tau \leq 1$

$$\begin{aligned}
&\Big|\int [U(\Phi_\tau(\boldsymbol{\theta}_1), \Phi_\tau(\boldsymbol{\theta}_2)) - U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - \tau\langle \nabla_2 U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle]\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2\Big|\\
\leq&K\tau^2 \int |a_1^\tau(K + a_2)|\overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2 \leq K\tau^2 C(K + C),
\end{aligned}$$

where we used that $|a^\tau| \leq |a| + K\tau$ from Eq. (67), and $M(\overline{\rho}_k^h) \leq C$. The same computation shows that the second and third terms, as well as the term depending on $V(\boldsymbol{\theta})$ are $O(\tau^2)$.

Taking $\tau \to 0$, we conclude that:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}[R(\nu_\tau)]_{\tau=0} = \int_{\mathbb{R}^d} \langle \nabla\Psi(\boldsymbol{\theta}, \overline{\rho}_k^h), \boldsymbol{\xi}(\boldsymbol{\theta})\rangle\overline{\rho}_k^h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

This equality combined with the analysis of [JKO98, Theorem 5.1] shows that $\rho(t)$ is indeed a weak solution of PDE (62). The proof of uniqueness follows from the regularity Lemma 28 and a standard method from elliptic-parabolic equations (see [JKO98, Theorem 5.1] for details). □

G Proof of Theorem 4

Proof of Theorem 4. Let $L^2(\mathbb{R}^d, \mathbb{P})$ be the space of functions on $\mathbb{R}^d$ that is square integrable with respect to the measure $\mathbb{P}$. For any functions $u, v \in L^2(\mathbb{R}^d, \mathbb{P})$, we denote by $\langle u, v\rangle_{L^2} = \int_{\mathbb{R}^d} u(\boldsymbol{x})v(\boldsymbol{x})\mathbb{P}(\mathrm{d}\boldsymbol{x})$ the scalar product of u, v and $\|u\|_{L^2} = (\langle u, u\rangle_{L^2})^{1/2}$ the norm of u in $L^2(\mathbb{R}^d, \mathbb{P})$.

We prove the case for general coefficients. The proof of fixed coefficient is the same but simpler.

Step 1. Bound the support of $\bar{a}^{t,\alpha}$.

Let $\bar{\boldsymbol{\theta}}^{t,\alpha} = (\bar{a}^{t,\alpha}, \bar{\boldsymbol{w}}^{t,\alpha})$ satisfying the non-linear dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{\theta}}^{t,\alpha} = -\frac{1}{\alpha}\nabla_{\boldsymbol{\theta}}\Psi_\alpha(\bar{\boldsymbol{\theta}}^{t,\alpha}; \rho_t^\alpha)$$

with initialization $\bar{\boldsymbol{\theta}}^{0,\alpha} \sim \rho_0$, and ρ_t^α given by Eq. (Rescaled-DD). Then we have

$$\begin{aligned}\Big|\frac{\mathrm{d}}{\mathrm{d}t}\bar{a}^{t,\alpha}\Big| =& \Big|(1/\alpha)\mathbb{E}[(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \rho_t^\alpha))\sigma(\boldsymbol{x}; \bar{\boldsymbol{w}}^{t,\alpha})]\Big|\\ \leq& (1/\alpha)\mathbb{E}[(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}; \rho_t^\alpha))^2]^{1/2}\mathbb{E}[\sigma(\boldsymbol{x}; \bar{\boldsymbol{w}}^{t,\alpha})^2]^{1/2}\\ \leq& (1/\alpha)KR_\alpha(\rho_t^\alpha)^{1/2}.\end{aligned}$$

The last inequality follows from the assumption that $\|\sigma\|_\infty \leq K$. Note $R_\alpha(\rho_t^\alpha)$ will always decrease along the trajectory, i.e., we have $R_\alpha(\rho_t^\alpha) \leq R_\alpha(\rho_0) \leq B$. As a result, we have $|\mathrm{d}\bar{a}^{t,\alpha}/\mathrm{d}t| \leq KB^{1/2}/\alpha$, so that

$$|\bar{a}^{t,\alpha}| \leq K(1 + B^{1/2}t/\alpha) \equiv M_{t,\alpha}.$$

Denoting $A(\rho) = \sup_{(a,\boldsymbol{w})\in\mathrm{supp}(\rho)} |a|$. Since $(\bar{a}^{t,\alpha}, \bar{\boldsymbol{w}}^{t,\alpha}) \sim \rho_t^\alpha$, we have

$$A(\rho_t^\alpha) \leq M_{t,\alpha} = K(1 + B^{1/2}t/\alpha).$$

Step 2. Bound $W_2(\rho_t^\alpha, \rho_0)$.

For $\boldsymbol{\theta} = (a, \boldsymbol{w})$, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}}\Psi_\alpha(\boldsymbol{\theta}, \rho_t^\alpha)\| &= \|\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})[f(\boldsymbol{x}) - \hat{f}_\alpha(\boldsymbol{x}; \rho_t^\alpha)]\}\| &(69)\\ &\leq \mathbb{E}\{\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})\|^2\}^{1/2}\mathbb{E}\{[f(\boldsymbol{x}) - \hat{f}_\alpha(\boldsymbol{x}; \rho_t^\alpha)]^2\}^{1/2} &(70)\\ &= \{\mathbb{E}\{\sigma(\boldsymbol{x}; \boldsymbol{w})^2\} + a^2\mathbb{E}\{\|\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}; \boldsymbol{w})\|_2^2\}\}^{1/2}R_\alpha(\rho_t^\alpha)^{1/2} &(71)\\ &\leq K(1 + |a|\sqrt{D})B^{1/2}\,. &(72)\end{aligned}$$

The last inequality follows from $\|\sigma\|_\infty \leq K$ and

$$\mathbb{E}\{\|\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}; \boldsymbol{w})\|_2^2\} = \mathrm{tr}(\nabla_1\nabla_2 u(\boldsymbol{w}, \boldsymbol{w})) \leq D\|\nabla_1\nabla_2 u(\boldsymbol{w}, \boldsymbol{w})\|_{\mathrm{op}} \leq KD.$$

Hence, for $s \leq t$,

$$\|\bar{\boldsymbol{\theta}}^{t,\alpha} - \bar{\boldsymbol{\theta}}^{s,\alpha}\|_2 = \frac{1}{\alpha}\Big\|\int_s^t \nabla_{\boldsymbol{\theta}}\Psi_\alpha(\bar{\boldsymbol{\theta}}^{u,\alpha}; \rho_u^\alpha)\mathrm{d}u\Big\|_2 \leq \frac{K}{\alpha}|t - s|M_{t,\alpha}B^{1/2}\sqrt{D}.$$

Note that, by the coupling in terms of nonlinear dynamics, for any $s \leq t$, we have

$$W_2(\rho_s^\alpha, \rho_t^\alpha) \leq \mathbb{E}\{\|\bar{\boldsymbol{\theta}}^{s,\alpha} - \bar{\boldsymbol{\theta}}^{t,\alpha}\|^2\}^{1/2} \leq \frac{K}{\alpha}|t - s|M_{t,\alpha}B^{1/2}\sqrt{D}. \tag{73}$$

Step 2. Bound $\|\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}\|_{\mathrm{op}}$.

Note that, for $v \in L^2(\mathbb{R}^d, \mathbb{P})$,

$$\langle v, \mathcal{H}_\rho v\rangle_{L^2} = \int \big\|\mathbb{E}_{\boldsymbol{x}}\{\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})v(\boldsymbol{x})\}\big\|_2^2\,\rho(\mathrm{d}\boldsymbol{\theta})\,. \tag{74}$$

Letting γ denote the coupling that achieves the W_2 distance between ρ_1 and ρ_2, we have

$$\begin{aligned}\langle v, [\mathcal{H}_{\rho_1} - \mathcal{H}_{\rho_2}]v\rangle_{L^2} =& \int \Big\{ \big\|\mathbb{E}_{\boldsymbol{x}}\{\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1)v(\boldsymbol{x})\}\big\|_2^2 - \big\|\mathbb{E}_{\boldsymbol{x}}\{\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)v(\boldsymbol{x})\}\big\|_2^2 \Big\}\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\\ \le& \Big[\int A_-(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\cdot \int A_+(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\Big]^{1/2}.\end{aligned}$$

where

$$\begin{aligned}A_-(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \equiv& \big\|\mathbb{E}_{\boldsymbol{x}}\{[\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)]v(\boldsymbol{x})\}\big\|_2^2\\ \le& \mathbb{E}_{\boldsymbol{x}}\{\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\|_2^2\}\|v\|_{L^2}^2\,,\\ A_+(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \equiv& (\mathbb{E}_{\boldsymbol{x}}\{[\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1)\|_2 + \|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\|_2]\|v(\boldsymbol{x})\|_2\})^2\\ \le& \mathbb{E}_{\boldsymbol{x}}\{(\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1)\|_2 + \|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\|_2)^2\}\|v\|_{L^2}^2\,.\end{aligned}$$

Note we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}}\{(\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1)\|_2 + \|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\|_2)^2\}\\ =&\mathrm{tr}[\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)] + \mathrm{tr}[\nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2)] + 2\{\mathrm{tr}[\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)]\cdot \mathrm{tr}[\nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2)]\}^{1/2}\\ \le& D(\|\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)\|_{\mathrm{op}} + \|\nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2)\|_{\mathrm{op}} + 2\{\|\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)\|_{\mathrm{op}}\|\nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2)\|_{\mathrm{op}}\}^{1/2})\\ \le& KD(1+|a_1|\vee|a_2|)^2,\end{aligned}$$

where the last inequality is by

$$\nabla_1\nabla_2 U(\boldsymbol{\theta},\boldsymbol{\theta}') = \begin{bmatrix} u(\boldsymbol{w},\boldsymbol{w}') & a'\nabla_1 u(\boldsymbol{w},\boldsymbol{w}')\\ a\nabla_2 u(\boldsymbol{w},\boldsymbol{w}') & aa'\nabla_1\nabla_2 u(\boldsymbol{w},\boldsymbol{w}')\end{bmatrix},$$

and the assumption that $|u|, \|\nabla u\|_2, \|\nabla^2 u\|_{\mathrm{op}} \le K$. This gives

$$A_+(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \le KD(1+|a_1|\vee|a_2|)^2\|v\|_{L^2}^2.$$

Moreover, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}}[\|\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_2)\|_2^2] = \mathrm{tr}[\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) + \nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2) - 2\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)]\\ \le& D\|\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) + \nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2) - 2\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\|_{\mathrm{op}} \le K\kappa D(1+|a_1|\vee|a_2|)^2\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2,\end{aligned}$$

where the last inequality follows from

$$\|\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) + \nabla_1\nabla_2 U(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2) - 2\nabla_1\nabla_2 U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\|_{\mathrm{op}} \le \|\nabla_1^2\nabla_2^2 U(\tilde{\boldsymbol{\theta}}_1,\tilde{\boldsymbol{\theta}}_2)\|_{\mathrm{op}}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2,$$

and $\|\nabla^3 u\|_{\mathrm{op}}, \|\nabla^4 u\|_{\mathrm{op}} \le \kappa$. This gives

$$A_-(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \le K\kappa D(1+|a_1|\vee|a_2|)^2\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2\|v\|_{L^2}^2.$$

Remember the notation $A(\rho) = \sup_{(a,\boldsymbol{w})\in\mathrm{supp}(\rho)}|a|$ and we have shown $A(\rho_t^\alpha) \le M_{t,\alpha} = K(1+B^{1/2}t/\alpha)$ in step 1, we have

$$\begin{aligned}&\langle v, [\mathcal{H}_{\rho_1} - \mathcal{H}_{\rho_2}]v\rangle_{L^2}\\ =&\Big[KD[1+A(\rho_1)\vee A(\rho_2)]^2\cdot\|v\|_{L^2}^2\cdot K\kappa D[1+A(\rho_1)\vee A(\rho_2)]^2\cdot\int\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2\gamma(\mathrm{d}\boldsymbol{\theta}_1,\mathrm{d}\boldsymbol{\theta}_2)\|v\|_{L^2}^2\Big]^{1/2}\\ \le& K\kappa^{1/2}D[1+A(\rho_1)\vee A(\rho_2)]^2 W_2(\rho_1,\rho_2)\cdot\|v\|_{L^2}^2.\end{aligned}$$

Substituting above, we get

$$\|\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}\|_{\mathrm{op}} \le K\kappa^{1/2}DW_2(\rho_0,\rho_t^\alpha)(1+M_{t,\alpha})^2 \le K\kappa^{1/2}D^{3/2}(1+B^{1/2}t/\alpha)^3B^{1/2}t/\alpha. \tag{75}$$

Step 3. Bound the difference of mean field and linearized residue dynamics $v_t = u_t^\alpha - u_t^*$.

We now consider the mean field residual dynamics (RD) and the linearized residual dynamics (17). Defining $v_t = u_t^\alpha - u_t^*$, we have

$$\partial_t v_t = -\mathcal{H}_{\rho_t^\alpha} v_t + (\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}) u_t^* \, . \tag{76}$$

Since $\mathcal{H}_{\rho_t^\alpha} \succeq \mathbf{0}$, this implies

$$\frac{\mathrm{d}}{\mathrm{d}t} \|v_t\|_{L^2}^2 \le 2\langle v_t, (\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}) u_t^* \rangle_{L^2} \le 2\|v_t\|_{L^2} \|\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}\|_{\mathrm{op}} \|u_t^*\|_{L^2} . \tag{77}$$

Using the bound (75), and $\|u_t^*\|_{L^2}^2 \le \|u_0^*\|_{L^2}^2 = R_\alpha(\rho_0) \le B_\alpha$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \|v_t\|_{L^2} \le \|\mathcal{H}_{\rho_0} - \mathcal{H}_{\rho_t^\alpha}\|_{\mathrm{op}} \|u_t^*\|_{L^2} \le K\kappa^{1/2} D^{3/2} (1 + B^{1/2} t/\alpha)^3 B t/\alpha \, , \tag{78}$$

Integrating this inequality yields Eq. (20). Eq. (21) follows by triangle inequality.

Step 4. Proving Eq. (22).

For $\rho_0 = \rho_0^a \times \rho_0^w$ with $|\mathbb{E}(a)| \le K/\alpha$, we have

$$\|\hat{f}(\boldsymbol{x}; \rho_0)\| = \alpha \Big\| \int a \rho_0^a(\mathrm{d}a) \cdot \int \sigma(\boldsymbol{x}; \boldsymbol{w}) \rho_0^w(\mathrm{d}\boldsymbol{w}) \Big\| \le K$$

Then we have

$$R_\alpha(\rho_0) = 2\mathbb{E}[f(\boldsymbol{x})^2] + 2\mathbb{E}[\hat{f}(\boldsymbol{x}; \rho_0)^2] \le K,$$

which is independent of α. Hence we have in both cases

$$\lim_{\alpha \to \infty} R_\alpha(\rho_t^\alpha) \le \|u_t^*\|_{L^2}^2 .$$

Equation (22) holds by Lemma 1. □

H The mean field limit and kernel limit

This section is a self-contained note comparing the *mean field limit* and *kernel limit*. We introduce the *distributional dynamics* and *residual dynamics*, which we consider in the pre-limit and in the limit of infinite number of neurons.

Let us emphasize that the material presented here is not new and appears in the literature, possibly in a slightly different formulations.

H.1 Two layers neural networks with a scale parameter α

Let $f : \mathbb{R}^d \to \mathbb{R}$. We use a two layer's neural network to fit this function f over data $\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}$. We denote $\hat{f}_{\alpha,N}(\boldsymbol{x}; \boldsymbol{\theta})$ the N-neurons prediction function at point $\boldsymbol{x} \in \mathbb{R}^d$ with weights $\boldsymbol{\theta} \in \mathbb{R}^{D \times N}$,

$$\hat{f}_{\alpha,N}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\alpha}{N} \sum_{j=1}^N \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_j) \, .$$

Here α serves as a scale parameter, which can be used to explore different regimes of the learning dynamics. We minimize the population risk over $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$:

$$R_{\alpha,N}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}} \Big[\Big(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x}; \boldsymbol{\theta}) \Big)^2 \Big].$$

In the rest of this appendix, we will first consider the gradient flow dynamics of the finite neuron risk function. This can be described via a *distributional dynamics*, which is a flow in the space of probability measures. The distributional dynamics induces an evolution of the residuals at the data points, which we call *residual dynamics*. We then consider the limit $N \to \infty$, which we refer to as the *mean field limit*.

Finally, we consider the limit of both $\alpha \to \infty$ *after* $N \to \infty$, that we call the *kernel limit*. Of course, it is also possible (and interesting) to study joint limits $\alpha, N \to \infty$ [JGH18]. Our rationale for the focusing on $\alpha \to \infty$ *after* $N \to \infty$ (following [CB18a]) is that it allows to explore the crossover between mean field and kernel behaviors.

H.2 The residual dynamics in the pre-limit

Calculating the gradient $\nabla_{\boldsymbol{\theta}_j} R_{\alpha,N}(\boldsymbol{\theta})$ using chain rule, we get

$$\nabla_{\boldsymbol{\theta}_j} R_{\alpha,N}(\boldsymbol{\theta}) = -\frac{\alpha}{N}\hat{\mathbb{E}}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_j)].$$

We consider the gradient flow ODE with time reparameterization given by $N/(2\alpha^2)$,

$$\frac{\mathrm{d}\boldsymbol{\theta}_j^t}{\mathrm{d}t} = -\frac{N}{2\alpha^2}\nabla_{\boldsymbol{\theta}_j} R_{\alpha,N}(\boldsymbol{\theta}^t) = \frac{1}{\alpha}\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}^t))\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_j^t)].$$

The time derivative of $\hat{f}_{\alpha,N}(\boldsymbol{z};\boldsymbol{\theta}^t)$ can be calculated using the chain rule. We have

$$\begin{aligned}\partial_t \hat{f}_{\alpha,N}(\boldsymbol{z};\boldsymbol{\theta}^t) =& \frac{\alpha}{N}\sum_{j=1}^N \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z};\boldsymbol{\theta}_j^t), \frac{\mathrm{d}\boldsymbol{\theta}_j^t}{\mathrm{d}t}\rangle \\ =& \mathbb{E}_{\boldsymbol{x}}\Big[\Big(\frac{1}{N}\sum_{j=1}^N \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_j^t), \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z};\boldsymbol{\theta}_j^t)\rangle\Big)\Big(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}^t)\Big)\Big].\end{aligned}$$

Define the kernel function $\mathcal{H}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ with weights $\boldsymbol{\theta} \in \mathbb{R}^{D\times N}$ to be

$$\mathcal{H}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=1}^N \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_j), \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z};\boldsymbol{\theta}_j)\rangle,$$

then we have

$$\partial_t \hat{f}_{\alpha,N}(\boldsymbol{z};\boldsymbol{\theta}^t) = \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}^t)\Big)\mathcal{H}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^t)\Big].$$

Taking the residue function to be $u_t^{\alpha,N}(\boldsymbol{z}) = f(\boldsymbol{z}) - \hat{f}_{\alpha,N}(\boldsymbol{z};\boldsymbol{\theta}^t)$, we have

$$\partial_t u_t^{\alpha,N}(\boldsymbol{z}) = -\,\mathbb{E}_{\boldsymbol{x}}[\mathcal{H}(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^t)u_t^{\alpha,N}(\boldsymbol{x})], \tag{79}$$

with initialization $u_0^{\alpha,N}(\boldsymbol{z}) = f(\boldsymbol{z}) - f_{\alpha,N}(\boldsymbol{z};\boldsymbol{\theta}^0)$ and $\boldsymbol{\theta}_i^0 \sim \rho_0$ independently. We call Eq. (79) the *residual dynamics*. The residual dynamics is not a self-contained equation and depends on $\boldsymbol{\theta}^t$.

H.3 The distributional dynamics in the pre-limit

Define

$$\rho_t^{\alpha,N}(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=1}^N \delta_{\boldsymbol{\theta}_j^t}.$$

Define the prediction function with distribution ρ and scaled parameter α to be

$$\hat{f}_\alpha(\boldsymbol{x};\rho) = \alpha\int \sigma_\star(\boldsymbol{x};\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}).$$

Consider again the gradient flow dynamics

$$\frac{\mathrm{d}\boldsymbol{\theta}_j^t}{\mathrm{d}t} = \frac{1}{\alpha}\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}^t))\nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}_j)] = -\frac{1}{\alpha}\nabla_{\boldsymbol{\theta}}\Psi_\alpha(\boldsymbol{\theta}_j^t;\rho_t^{\alpha,N}).$$

where we defined

$$\Psi_\alpha(\boldsymbol{\theta};\rho) = -\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_\alpha(\boldsymbol{x};\rho))\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})].$$

Then we have

$$\begin{aligned}\partial_t \rho_t^{\alpha,N} =& (1/\alpha)\nabla_{\boldsymbol{\theta}}\cdot(\rho_t^{\alpha,N}[\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta};\rho_t^{\alpha,N})]), \\ \rho_0^{\alpha,N} =& \frac{1}{N}\sum_{j=1}^N \delta_{\boldsymbol{\theta}_i^0},\end{aligned} \tag{80}$$

with $\boldsymbol{\theta}_i^0 \sim \rho_0$ independently. We call dynamics (80) the *distributional dynamics*. The distributional dynamics is equivalent to the gradient flow.

H.4 The coupled dynamics

Writing the *distributional dynamics* and *residual dynamics* together (in the pre-limit), we have

$$\begin{aligned}\partial_t \rho_t^{\alpha,N} =& (1/\alpha)\nabla_{\boldsymbol{\theta}} \cdot (\rho_t^{\alpha,N}[\nabla_{\boldsymbol{\theta}}\Psi_\alpha(\boldsymbol{\theta};\rho_t^{\alpha,N})]),\\ \partial_t u_t^{\alpha,N}(\boldsymbol{z}) =& -\mathbb{E}_{\boldsymbol{x}}[u_t^{\alpha,N}(\boldsymbol{x})\mathcal{H}_{\rho_t^{\alpha,N}}(\boldsymbol{x},\boldsymbol{z})],\end{aligned}$$

where

$$\begin{aligned}\mathcal{H}_\rho(\boldsymbol{x},\boldsymbol{z}) \equiv& \int \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x};\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z};\boldsymbol{\theta})\rangle \rho(\mathrm{d}\boldsymbol{\theta}),\\ \Psi_\alpha(\boldsymbol{\theta};\rho^{\alpha,N}) =& -\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \hat{f}_\alpha(\boldsymbol{x};\rho^{\alpha,N}))\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})] = -\mathbb{E}_{\boldsymbol{x}}[u_t^{\alpha,N}(\boldsymbol{x})\sigma_\star(\boldsymbol{x};\boldsymbol{\theta})],\end{aligned}$$

with initialization conditions $\rho_0^N = (1/N)\sum_{i=1}^N \delta_{\boldsymbol{\theta}_i^0}$, $u_N(0,\boldsymbol{x}) = f(\boldsymbol{x}) - \hat{f}_{\alpha,N}(\boldsymbol{x};\boldsymbol{\theta}^0)$, and $(\boldsymbol{\theta}_i^0)_{i\le N} \sim_{i.i.d.} \rho_0$.

Note these coupled dynamics are random, where the randomness comes from the random initialization $(\boldsymbol{\theta}_i^0)_{i\le N} \sim_{i.i.d.} \rho_0$.

H.5 The mean field limit

In the mean field limit, we fix α and take $N \to \infty$. Under some conditions, it can be shown that there exists $(\rho_t)_{t\ge 0}$ satisfying the *mean field distributional dynamics*

$$\partial_t \rho_t^\alpha = (1/\alpha)\nabla \cdot (\rho_t^\alpha[\nabla_{\boldsymbol{\theta}}\Psi_\alpha(\boldsymbol{\theta};\rho_t^\alpha)]), \tag{81}$$

with initialization condition $\rho_0^\alpha = \rho_0$. Moreover, we have almost surely (over $\boldsymbol{\theta}_i^0 \sim \rho_0$ independently)

$$\lim_{N\to\infty} W_2(\rho_t^{\alpha,N},\rho_t^\alpha) \to 0.$$

The mean field distributional dynamics was proposed and studied in [MMN18, SS18, RVE18, CB18b] under various conditions.

Now define the *mean field residual function* $u_t^\alpha(\boldsymbol{z})$ to be

$$u_t^\alpha(\boldsymbol{z}) \equiv f(\boldsymbol{z}) - \hat{f}_\alpha(\boldsymbol{z};\rho_t^\alpha).$$

For any fixed $\boldsymbol{z}$, we have almost surely

$$\lim_{N\to\infty} u_t^{\alpha,N}(\boldsymbol{z}) = u_t^\alpha(\boldsymbol{z}).$$

Under some regularity conditions, it is not hard to show that this mean field residual function satisfies *mean field residual dynamics*

$$\partial_t u_t^\alpha(\boldsymbol{z}) = -\mathbb{E}_{\boldsymbol{x}}[u_t^\alpha(\boldsymbol{x})\mathcal{H}_{\rho_t^\alpha}(\boldsymbol{x},\boldsymbol{z})].$$

The *mean field residual dynamics* is not a self-contained equation. It depends on the distribution through the kernel $\mathcal{H}_{\rho_t^\alpha}$. The mean field residual dynamics was first explicitly given in [RVE18, Proposition 2.5].

H.6 The kernel limit

Theorem 4 shows that, as α becomes large, for any fixed t, we have

$$\lim_{\alpha\to\infty} W_2(\rho_t^\alpha,\rho_0) = 0,$$

and hence

$$\lim_{\alpha\to\infty} \|\mathcal{H}_{\rho_t^\alpha} - \mathcal{H}_{\rho_0}\|_{\mathrm{op}} = 0.$$

In this limit, the mean field residual dynamics converges to the *linearized residual dynamics*,

$$\partial_t u_t^*(\boldsymbol{z}) = -\mathbb{E}_{\boldsymbol{x}}[u_t^*(\boldsymbol{x})\mathcal{H}_{\rho_0}(\boldsymbol{x},\boldsymbol{z})]. \tag{82}$$

The linearized residual dynamics is exactly the same as the continuous time kernel boosting dynamics with kernel $\mathcal{H}_{\rho_0}$, whose solution can be written down explicitly

$$u_t^* = e^{-\mathcal{H}_{\rho_0} t} u_0^*. \tag{83}$$

When the kernel is strictly positive definite, one can show that the L^2-norm of the residual function converges to 0 as time goes to infinity.

The kernel limit is studied in [JGH18, GJS+19] in the joint limit $\alpha = N^{1/2} \to \infty$, and in a multi-layer neural network settings. The specific limit considered here ($N \to \infty$ followed by $\alpha \to \infty$) is discussed in [CB18a].

An interesting line of research [LL18, DZPS18, DLL+18, AZLS18] also studies the kernel limit, but focusing on dynamics on empirical risk. Note that all the equations discussed above also holds for $\mathbb{P}_{\boldsymbol{x}} = (1/n)\sum_{k=1}^n \delta_{\boldsymbol{x}_k}$. The benefit of working with the empirical risk is that, under mild assumptions, the kernel matrix $\{\mathcal{H}_{\rho_0}(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j\in[n]}$ is strictly positive definite with least eigenvalue $\lambda_{\min} > 0$. As a result, it is possible to upper bound the convergence time of the empirical risk to 0 using Eq. (82). Hence, it is possible to choose the number of neurons large enough that the residual dynamics (79) is well approximated by the linearized residual dynamics (82) along the whole trajectory.

H.7 Kernel limit as kernel ridge regression

Consider the case when $\mathbb{P}_{\boldsymbol{x}} = (1/n)\sum_{k=1}^n \delta_{\boldsymbol{x}_k}$ is the empirical data distribution. We make an additional assumption on the initialization weight distribution ρ_0:

(I) The initialization distribution $(a, \boldsymbol{w}) \sim \rho_0$ verifies: a is independent of $\boldsymbol{w}$ and $\mathbb{E}(a) = 0$. In other words, $\rho_0 = \rho_0^a \times \rho_0^{\boldsymbol{w}}$ with $\int a \rho_0^a(\mathrm{d}a) = 0$.

Under this assumption, we have $\hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha) \equiv 0$ for any $\boldsymbol{z} \in \mathbb{R}^d$, so that $u_0^\alpha(\boldsymbol{x}_k) = f(\boldsymbol{x}_k)$ for $k \in [n]$.

Denote

$$\begin{aligned}\boldsymbol{u}_t^\alpha =& [u_t^\alpha(\boldsymbol{x}_1), \dots, u_t^\alpha(\boldsymbol{x}_n)]^\mathsf{T},\\ \boldsymbol{u}_t^* =& [u_t^*(\boldsymbol{x}_1), \dots, u_t^*(\boldsymbol{x}_n)]^\mathsf{T},\\ \boldsymbol{y} =& [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^\mathsf{T}.\end{aligned}$$

Further we denote the data kernel matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{H}_{ij} = \mathcal{H}_{\rho_0}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then Eq. (17) can be rewritten as

$$\boldsymbol{u}_t^* = e^{-\boldsymbol{H}t/n} \boldsymbol{u}_0^* = e^{-\boldsymbol{H}t/n} \boldsymbol{y}.$$

Note Theorem 4 holds also in the case when $\mathbb{P}_{\boldsymbol{x}}$ is an empirical data distribution. Hence we have

$$\lim_{\alpha\to\infty} \sup_{t\in[0,T]} \frac{1}{\sqrt{n}} \|\boldsymbol{u}_t^\alpha - \boldsymbol{u}_t^*\|_2 = \lim_{\alpha\to\infty} \sup_{t\in[0,T]} \|u_t^\alpha - u_t^*\|_{L^2} = 0.$$

The following proposition considers the scaling limit (kernel limit) of the prediction function at time t,

$$\hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha) = \alpha \int \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}) \rho_t^\alpha(\mathrm{d}\boldsymbol{\theta}),$$

where ρ_t^α is the solution of the rescaled distributional dynamics (Rescaled-DD).

This fact already appears (implicitly or explicitly) in several of the papers mentioned above. We state and prove it here for the sake of completeness.

Proposition 19. *Assume conditions A1 - A4 hold, and $\mathbb{P}_{\boldsymbol{x}} = (1/n)\sum_{k=1}^n \delta_{\boldsymbol{x}_k}$ to be the empirical data distribution. Additionally assume the finite data kernel matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is invertible, and ρ_0 verifies property* (I)*. Then for any fixed $\boldsymbol{z} \in \mathbb{R}^d$, we have*

$$\lim_{t\to\infty} \lim_{\alpha\to\infty} \hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha) = \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} \boldsymbol{H}^{-1} \boldsymbol{y},$$

where

$$\boldsymbol{h}(\boldsymbol{z}) = [\mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_1), \dots, \mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_n)]^\mathsf{T}.$$

Remark H.1. Given a data set $\{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$, kernel ridge regression is a function estimator $\hat{f}_\lambda$ that solves the following minimization problem

$$\min_f \; \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}_{\rho_0}}.$$

The norm $\|f\|_{\mathcal{H}_{\rho_0}}$ is the *reproducible kernel Hilbert space* (RKHS) norm of function f, where the RKHS is associated to the kernel $\mathcal{H}_{\rho_0}$. The solution of the minimization problem above gives

$$\hat{f}_\lambda(\boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{z})^\mathsf{T}(\boldsymbol{H} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}.$$

Proposition 19 shows that, the mean field prediction function in the kernel limit is performing a kernel ridge regression with regularization parameter $\lambda = 0$.

Proof of Proposition 19. Recall that

$$\begin{aligned}
\boldsymbol{u}_t^\alpha =& [u_t^\alpha(\boldsymbol{x}_1), \ldots, u_t^\alpha(\boldsymbol{x}_n)]^\mathsf{T},\\
\boldsymbol{u}_t^* =& [u_t^*(\boldsymbol{x}_1), \ldots, u_t^*(\boldsymbol{x}_n)]^\mathsf{T},\\
\boldsymbol{y} =& [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^\mathsf{T}.
\end{aligned}$$

The data kernel matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is given by $\boldsymbol{H}_{in} = \mathcal{H}_{\rho_0}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. By Eq. (17) and the assumption on ρ_0, we have

$$\boldsymbol{u}_t^* = e^{-\boldsymbol{H}t/n}\boldsymbol{u}_0^* = e^{-\boldsymbol{H}t/n}\boldsymbol{y}.$$

For any fixed $\boldsymbol{z} \in \mathbb{R}^d$, denote

$$\begin{aligned}
\boldsymbol{h}_t^\alpha(\boldsymbol{z}) =& [\mathcal{H}_{\rho_t^\alpha}(\boldsymbol{z}, \boldsymbol{x}_1), \ldots, \mathcal{H}_{\rho_t^\alpha}(\boldsymbol{z}, \boldsymbol{x}_n)]^\mathsf{T},\\
\boldsymbol{h}(\boldsymbol{z}) =& [\mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_1), \ldots, \mathcal{H}_{\rho_0}(\boldsymbol{z}, \boldsymbol{x}_n)]^\mathsf{T}.
\end{aligned}$$

Using chain rule, the time derivative of the prediction function $\hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha) = \alpha \int \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})\rho_t^\alpha(\mathrm{d}\boldsymbol{\theta})$ gives

$$\begin{aligned}
\partial_t \hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha) =& \alpha\partial_t \int \sigma_\star(\boldsymbol{z}; \boldsymbol{\theta})\rho_t^\alpha(\mathrm{d}\boldsymbol{\theta}) = \int \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z}; \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}}\Psi_\alpha(\boldsymbol{\theta}; \rho_t^\alpha)\rangle \rho_t^\alpha(\mathrm{d}\boldsymbol{\theta})\\
=& \mathbb{E}_{\boldsymbol{x}}\Big[u_t^\alpha(\boldsymbol{x}) \int \langle \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{z}; \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}}\sigma_\star(\boldsymbol{x}; \boldsymbol{\theta})\rangle \rho_t^\alpha(\mathrm{d}\boldsymbol{\theta})\Big] = \boldsymbol{h}_t^\alpha(\boldsymbol{z})\boldsymbol{u}_t^\alpha/n.
\end{aligned} \tag{84}$$

By the same argument as Step 2 of Theorem 4, we have

$$\sup_{t\in[0,T]} \|\boldsymbol{h}(\boldsymbol{z}) - \boldsymbol{h}_t^\alpha(\boldsymbol{z})\|_2 = O(1/\alpha). \tag{85}$$

By Theorem 4, we have

$$\sup_{t\in[0,T]} \|\boldsymbol{u}_t^\alpha - \boldsymbol{u}_t^*\|_2 = \sup_{t\in[0,T]} \|u_t^\alpha - u_t^*\|_{L^2} = O(1/\alpha). \tag{86}$$

Now, we denote $\hat{f}_t(\boldsymbol{z})$ be the solution of the following *linearized prediction dynamics*,

$$\begin{aligned}
\partial_t \hat{f}_t(\boldsymbol{z}) =& \boldsymbol{h}(\boldsymbol{z})^\mathsf{T}\boldsymbol{u}_t^*/n,\\
\hat{f}_0(\boldsymbol{z}) =& 0.
\end{aligned} \tag{87}$$

By Eq (84), (85), (86) and (87), we have

$$\sup_{t\in[0,T]} |\partial_t \hat{f}_t(\boldsymbol{z}) - \partial_t \hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha)| = O(1/\alpha),$$

together with $\hat{f}_0(\boldsymbol{z}) = \hat{f}_\alpha(\boldsymbol{z}; \rho_0^\alpha) = 0$ we get

$$\hat{f}_t(\boldsymbol{z}) = \lim_{\alpha\to\infty} \hat{f}_\alpha(\boldsymbol{z}; \rho_t^\alpha).$$

Note the solution of Eq. (87) gives

$$\hat{f}_t(\boldsymbol{z}) = n^{-1} \int_0^t \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} \boldsymbol{u}_s^* \mathrm{d}s = n^{-1} \int_0^t \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} e^{-\boldsymbol{H}s/n} \boldsymbol{y} \mathrm{d}s = \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} \boldsymbol{H}^{-1} (\boldsymbol{I} - e^{-\boldsymbol{H}t/n}) \boldsymbol{y},$$

so that

$$\hat{f}_\infty(\boldsymbol{z}) = \lim_{t \to \infty} \hat{f}_t(\boldsymbol{z}) = \lim_{t \to \infty} \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} \boldsymbol{H}^{-1} (\boldsymbol{I} - e^{-\boldsymbol{H}t/n}) \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{z})^\mathsf{T} \boldsymbol{H}^{-1} \boldsymbol{y}.$$

This proves the proposition. □

I Technical lemmas

Lemma 30. *Let $\boldsymbol{X}_i \in \mathbb{R}^D$ with $\{\boldsymbol{X}_i\}_{i \in [N]}$ to be i.i.d. random variables, with $\|\boldsymbol{X}_i\|_2 \le K$ and $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{0}$. Then we have (the constant K in the result is up to some universal constant)*

$$\mathbb{P}\Big(\Big\| \frac{1}{N} \sum_{i=1}^N \boldsymbol{X}_i \Big\|_2 \ge K(\sqrt{1/N} + \delta) \Big) \le e^{-N\delta^2}.$$

Proof. Denote $f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) = \|(1/N) \sum_{i=1}^N \boldsymbol{X}_i\|_2$. Then we have

$$\begin{aligned} |\mathbb{E}[f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)]| \le& \mathbb{E}[f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)^2]^{1/2} = \mathbb{E}\Big[\Big\langle \frac{1}{N} \sum_{i=1}^N \boldsymbol{X}_i, \frac{1}{N} \sum_{j=1}^N \boldsymbol{X}_j \Big\rangle \Big]^{1/2} \\ =& \Big\{ \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}[\|\boldsymbol{X}_i\|_2^2] \Big\}^{1/2} \le K \sqrt{\frac{1}{N}}. \end{aligned}$$

Note by triangle inequality, we have

$$|f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_N) - f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i', \ldots, \boldsymbol{X}_N)| \le \frac{1}{N} \|\boldsymbol{X}_i - \boldsymbol{X}_i'\|_2 \le \frac{2K}{N}.$$

By McDiarmid's inequality, we have

$$\mathbb{P}\Big(|f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) - \mathbb{E}[f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)]| \ge \delta \Big) \le \exp\{-N\delta^2/K\},$$

which gives the desired result. □

Lemma 31 (Azuma-Hoeffding bound)**.** *Let $(\boldsymbol{X}_k)_{k \ge 0}$ be a martingale taking values in $\mathbb{R}^D$ with respect to the filtration $(\mathcal{F}_k)_{k \ge 0}$, with $\boldsymbol{X}_0 = 0$. Assume that the following holds almost surely for all $k \ge 1$:*

$$\mathbb{E}\{e^{\langle \lambda, \boldsymbol{X}_k - \boldsymbol{X}_{k-1} \rangle} | \mathcal{F}_{k-1}\} \le e^{L^2 \|\lambda\|_2^2 / 2}$$

Then we have

$$\mathbb{P}\Big(\max_{k \le n} \|\boldsymbol{X}_k\|_2 \ge 2L\sqrt{n} \Big[\sqrt{D} + \delta \Big] \Big) \le e^{-\delta^2}.$$

Proof. This lemma is proven in [MMN18, Section A, Lemma A.1]. □