A Dynamical Central Limit Theorem for Shallow Neural Networks

Zhengdao Chen[b], Grant M. Rotskoff[a,b], Joan Bruna[b,c], and Eric Vanden-Eijnden[b]

[a]Department of Chemistry, Stanford University, California
[b]Courant Institute of Mathematical Sciences, New York University, New York
[c]Center for Data Science, New York University, New York

August 25, 2020

Abstract

Recent theoretical work has characterized the dynamics of wide shallow neural networks trained via gradient descent in an asymptotic regime called the mean-field limit as the number of parameters tends towards infinity. At initialization, the randomly sampled parameters lead to a deviation from the mean-field limit that is dictated by the classical Central Limit Theorem (CLT). However, the dynamics of training introduces correlations among the parameters, raising the question of how the fluctuations evolve during training. Here, we analyze the mean-field dynamics as a Wasserstein gradient flow and prove that the deviations from the mean-field limit scaled by the width, in the width-asymptotic limit, remain bounded throughout training. In particular, they eventually vanish in the CLT scaling if the mean-field dynamics converges to a measure that interpolates the training data. This observation has implications for both the approximation rate and the generalization: the upper bound we obtain is given by a Monte-Carlo type resampling error, which does not depend explicitly on the dimension. This bound motivates a regularizaton term on the 2-norm of the underlying measure, which is also connected to generalization via the variation-norm function spaces.

1 Introduction

Theoretical analyses of neural networks aim to understand their computational and statistical advantages seen in practice. On the computation side, the training of neural networks can often succeed despite being inherently non-convex as an optimization problem and known to be hard in certain settings [20, 31, 41]. On the statistics side, neural networks can often generalize well despite having large numbers of parameters [8, 70]. In this context, the notion of *over-parametrization* has been useful, by providing insights into the optimization and generalization properties as the network widths tend to infinity [2, 4, 21, 36, 38, 63, 67]. In particular, under appropriate scaling, one can view shallow (a.k.a. single-hidden-layer or two-layer) networks as interacting particle systems that admit a mean-field limit. Their training dynamics can then be studied as Wasserstein Gradient Flows [12, 47, 51, 58], leading to global convergence guarantees in the mean-field limit under certain

Correspondence to: zc1216@nyu.edu, rotskoff@stanford.edu, bruna@cims.nyu.edu and eve2@cims.nyu.edu.

assumptions. On the statistics side, such an approach can lead to powerful generalization guarantees for learning high-dimensional functions with hidden low-dimensional structures, as compared to learning in Reproducing Kernel Hilbert Spaces (RKHS) [5, 30].

However, since ultimately we are concerned with neural networks of finite width, it is critical to study the deviation of finite-width networks from their infinite-width limits, especially how the discrepancy scales with the width n. At the random initial state, neurons do not interact and therefore a standard Monte-Carlo (MC) argument shows that the fluctuations in the underlying measure scale as $n^{-1/2}$ (or equivalently, n^{-1} if measured by the function variance), which we refer to as the Central Limit Theorem (CLT) scaling. As optimization introduces complex dependencies among the parameters, the key question is to understand how the fluctuation evolves during training. To make this investigation tractable, we aim to obtain insight on an asymptotic scale as the width grows, and focus on the evolution in time. An application of Grönwall's inequality can already show that this asymptotic deviation remains bounded at all finite time [46], but the dependence on time is exponential, making it difficult to assess the long-time behavior.

The main focus of this paper is to investigate this question in-depth, by analysing the interplay between the deviations from the mean-field limit and the gradient flow dynamics. First, we develop a dynamical CLT; we derive differential equations governing how the fluctuations away from the mean-field limit evolve as a function of training time and show that these fluctuations remain on the initial $n^{-1/2}$-scale for all finite times. Next, we examine the long-time behavior of the fluctuations, proving that, in several scenarios, the long-time fluctuations are controlled by the error of Monte-Carlo resampling from the limiting measure. We focus on two main setups relevant for supervised learning and scientific computing: the unregularized case with global convergence of mean-field gradient flows to minimizers that interpolate the data, and the regularized case where the limiting measure has atomic support and is nondegenerate. In the former setup, we prove particularly that the fluctuations eventually vanish in the CLT scaling. These asymptotic predictions are complemented by empirical results in a teacher-student model.

Related Works: This paper continues the line of work initiated by [12, 47, 51, 58] that studies optimization of over-parameterized shallow neural networks under the mean-field scaling. Global convergence for the unregularized setting is discussed in [46, 47, 51, 58]. In the regularized setting, [12] establishes global convergence in the mean-field limit under specific homogeneity conditions on the neuron activation. Other works that study asymptotic properties of wide neural networks include [1, 6, 23, 28, 29, 34, 35, 43, 69], notably investigating the transition between the so-called *lazy* and *active* regimes [14], corresponding respectively to linear versus nonlinear learning. Our focus is on the dynamics under the mean-field scaling, which encompasses the active, nonlinear regime.

A relevant work concerning the sparse optimization of measures is [11], where under a different metric for gradient flow and additional assumptions on the nature of the minimizer, it can be established that fluctuations vanish for sufficiently large n. Our results are only asymptotic in n but apply to broader settings in the context of shallow neural networks. Concerning the next-order deviations of finite neural networks from their mean-field limit, [51] show that the scale of fluctuations is below that of MC resampling for unregularized problems using non-rigorous arguments. [60] provides a CLT for the fluctuations at finite time under stochastic gradient descent (SGD) and proves that the fluctuations decay in time in the case where there is a single critical point in the parameter space. Our focus is on the long-time behavior of the fluctuations in more general settings. Another relevant topic is the propagation of chaos in McKean-Vlasov systems, which study the deviations of randomly-forced interacting particle systems from their infinite-particle limits [7, 10, 65, 66]. In particular, a line of work provides uniform-in-time bounds to the fluctuations in various settings [16, 19, 22, 55, 56], but the conditions are not applicable to shallow neural networks. Concurrently to our work, [17] studies quantitative propagation of chaos of shallow neural networks trained by SGD, but the bound grows exponentially in time, and therefore cannot address the long-time behavior of the fluctuations.

Learning with neural networks exhibits the phenomenon that generalization error can decrease with the level of overparameterization [8, 64]. [48] proposes a bias-variance decomposition that contains a variance term initialization in optimization. They show in experiments that this term decreases as the width of the network increases, and justifies this theoretically under the strong assumption that model parameters remain Gaussian-distributed in the components that are irrelevant for the task, which does not hold in the scenario we consider, for example. [27] provides scaling arguments for the dependence of this term on the width of the network. Our work provides a more rigorous analysis of the dependence of this term on the width of the network and training time.

2 Background

2.1 Shallow Neural Networks and the Integral Representation

On a data space $\Omega \subseteq \mathbb{R}^d$, we consider parameterized models of the following form

$$f^{(n)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{\theta}_i, \boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \Omega$ is the feature of the input data, $\{\boldsymbol{\theta}_i\}_{i=1}^n \subseteq D$ is the set of model parameters, and $\varphi : D \times \Omega \to \mathbb{R}$ is some activation function. Of particular interest are the shallow neural network models, which admit a more specific form:

Assumption 2.1 (The shallow neural networks setting)**.** *$D = \mathbb{R} \times \hat{D}$, and for $\boldsymbol{\theta} = (c, \boldsymbol{z}) \in D$, $\varphi(\boldsymbol{\theta}, \boldsymbol{x}) = c\hat{\varphi}(\boldsymbol{z}, \boldsymbol{x})$ with $\hat{\varphi} : \hat{D} \times \Omega \to \mathbb{R}$. Thus,* (1) *can be rewritten as*

$$f^{(n)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \hat{\varphi}(\boldsymbol{z}_i, \boldsymbol{x}) \tag{2}$$

As many of our results hold for general models of the form (1), we will invoke Assumption 2.1 only when needed. We shall also assume the following:

Assumption 2.2. *Ω is compact; D is an Euclidean space (or a subset thereof); $\varphi(\boldsymbol{\theta}, \boldsymbol{x})$ is twice differentiable in $\boldsymbol{\theta}$; $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}, \boldsymbol{x})$ is Lipschitz in $\boldsymbol{\theta}$, uniformly in $\boldsymbol{x}$.*

The regularity assumptions are standard in the literature [10, 11, 37]. We note that they are not satisfied by shallow ReLU neural networks (i.e., $\hat{\varphi}(\boldsymbol{z}, \boldsymbol{x}) = \max\{0, \langle \boldsymbol{a}, \boldsymbol{x}\rangle + b\}$, where $\boldsymbol{z} = [\boldsymbol{a}, b]^\intercal$, with $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$), though prior work [12, 13] has considered differentiable approximations of these models.

As observed in [12, 24, 47, 51, 58], a model of the form (1) can be expressed in integral form in terms of a probability measure over D as $f^{(n)} = f[\mu^{(n)}]$, where we define

$$f[\mu](\boldsymbol{x}) = \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x})\mu(d\boldsymbol{\theta}) \ , \tag{3}$$

and $\mu^{(n)}$ is the empirical measure of the parameters $\{\boldsymbol{\theta}_i\}_{i=1}^n$:

$$\mu^{(n)}(d\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}(d\boldsymbol{\theta}) \ . \tag{4}$$

Suppose we are given a dataset $\{(\boldsymbol{x}_l, y_l)\}_{l=1\dots L}$, which can be represented by an empirical data measure $\hat{\nu} = L^{-1}\sum_{l=1}^{L}\delta_{\boldsymbol{x}_l}$, and $y_l = f_*(\boldsymbol{x}_l)$ are generated by an target function f_* that we wish to estimate using least-squares regression. A canonical approach to this regression task is to consider an Empirical Risk Minimization (ERM) problem of the form

$$\min_{\mu\in\mathcal{P}(D)} \mathcal{L}(\mu) \quad \text{with} \quad \mathcal{L}(\mu) := \|f[\mu] - f_*\|_{\hat{\nu}}^2 + \lambda \int_D r(\boldsymbol{\theta})\mu(d\boldsymbol{\theta}) \ . \tag{5}$$

where $\mathcal{P}(D)$ is the space of probability measures on D, $\|f - f_*\|_{\hat{\nu}}^2 = \int_\Omega |f(\boldsymbol{x}) - f_*(\boldsymbol{x})|^2\hat{\nu}(d\boldsymbol{x})$ denotes the function reconstruction error averaged over the data, and $\lambda\int_D r(\boldsymbol{\theta})\mu(d\boldsymbol{\theta})$ is some optional regularization term. While we can allow r to be a general convex function, in Section 3.3 we will motivate a choice of r in the shallow neural networks setting based on the function variation norm.

Models like shallow neural networks allow efficient approximations to this regression problem, which we review next.

2.2 Approximation and Optimization with Finite Number of Neurons

Integral representations with a probability measure such as those defined in (3) are amenable to efficient approximation in high dimensions via Monte-Carlo sampling. Namely, if the parameters $\boldsymbol{\theta}_i$ in $f^{(n)}$ are drawn i.i.d. from an underlying measure μ on D, then by the Law of Large Numbers (LLN), the resulting empirical measure $\mu^{(n)}$ converges μ almost surely, and moreover,

$$\mathbb{E}_{\mu^{(n)}}\|f[\mu^{(n)}] - f[\mu]\|_{\hat{\nu}}^2 = n^{-1}\left(\int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2\mu(d\boldsymbol{\theta}) - \|f[\mu]\|_{\hat{\nu}}^2\right) \ , \tag{6}$$

Such a Monte-Carlo estimator showcases the benefit of normalized integral representations for high-dimensional approximation, as the ambient dimension appears in the rate of approximation only through the term $\int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2\mu(d\boldsymbol{\theta})$. In the case of shallow neural networks, this is given by the variation norm on the space $\mathcal{F}_1$ of the function we wish to approximate [5], which we will revisit in more detail in Section 3.3.

While the Monte-Carlo sampling strategy above can be seen as a 'static' approximation of a function representable as (3), it also gives rise to an efficient algorithm to optimize (5). Indeed, under (4), the loss $\mathcal{L}(\mu^{(n)})$ equals

$$L(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n) = \|f^{(n)} - f_*\|_{\hat{\nu}}^2 + \frac{\lambda}{n}\sum_{i=1}^{n} r(\boldsymbol{\theta}_i) \ , \tag{7}$$

as a function of the parameters $\{\boldsymbol{\theta}_i\}_{i=1}^n$. In the shallow neural network setting, with suitable choices of the function r, the regularization term corresponds to *weight decay* over the parameters.

Performing gradient descent (GD) on L with respect to $\{\boldsymbol{\theta}_i\}_{i=1}^n$ induces gradient dynamics on the functional $\mathcal{L}$ over $\mathcal{P}(D)$ with respect to a Wasserstein metric [12, 47, 51, 58], as we now review.

2.3 From Particle to Wasserstein Gradient Flows

Expanding (7), we get

$$L(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n) = C_{f_*} - \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{\theta}_i) + \frac{1}{2n^2}\sum_{i,j=1}^{n} K(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j), \tag{8}$$

where we have defined $C_f = \frac{1}{2}\|f\|^2_{\hat{\nu}}$, and

$$F(\boldsymbol{\theta}) = \int_\Omega f_*(\boldsymbol{x})\varphi(\boldsymbol{\theta},\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) - \lambda r(\boldsymbol{\theta}), \qquad K(\boldsymbol{\theta},\boldsymbol{\theta}') = \int_\Omega \varphi(\boldsymbol{\theta},\boldsymbol{x})\varphi(\boldsymbol{\theta}',\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) \,. \tag{9}$$

Performing GD on L amounts to discretizing in time the following ODE system that govern the gradient flow of $\{\boldsymbol{\theta}_i\}_{i=1}^n$:

$$\dot{\boldsymbol{\theta}}_i = -n\partial_{\boldsymbol{\theta}_i} L(\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_n) = \nabla F(\boldsymbol{\theta}_i) - \frac{1}{n}\sum_{j=1}^n \nabla K(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) =: -\nabla V(\boldsymbol{\theta}_i, \mu_t^{(n)}). \tag{10}$$

where we defined the potential

$$V(\boldsymbol{\theta},\mu) = -F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu(d\boldsymbol{\theta}) \,. \tag{11}$$

Heuristically, the 'particles' $\boldsymbol{\theta}_i$ perform GD according to the potential $V(\boldsymbol{\theta}, \mu_t^{(n)})$ which itself evolves, depending on the particles positions through their empirical measure. Such dynamics can also be expressed in terms of the empirical measure via the *continuity equation*:

$$\partial_t \mu_t^{(n)} = \nabla \cdot (\nabla V(\boldsymbol{\theta}, \mu_t^{(n)})\mu_t^{(n)}) \tag{12}$$

This equation should be understood in the weak sense of testing it against continuous functions $\chi : D \to \mathbb{R}$, and it can be interpreted as the gradient flow on the loss defined in (5) under the 2-Wasserstein metric induced by the metric on D where the particle gradient descent operates [12, 47, 51, 58]. This insight provides powerful analytical tools to understand convergence properties, by considering the mean-field limit when $n \to \infty$.

2.4 Law of Large Numbers and Mean-Field Gradient Flow

From now on, we assume that the particle gradient flow is initialized in the following way:

Assumption 2.3. *The gradient flow in (10) has the initial condition $\boldsymbol{\theta}_i(0) = \boldsymbol{\theta}_i^0$, with $\boldsymbol{\theta}_i^0$ drawn i.i.d. from a compactly supported measure $\mu_0 \in \mathcal{P}(D)$ for each $i = 1,\dots,n$. Hence, $\mu_0^{(n)}(d\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{\theta}_i^0}(d\boldsymbol{\theta})$.*

We use $\mathbb{P}_0$ to denote the probability measure associated with the set $\{\boldsymbol{\theta}_i^0\}_{i\in\mathbb{N}}$ with each $\boldsymbol{\theta}_i^0$ drawn i.i.d. from μ_0, and use $\mathbb{E}_0$ to denote the expectation under $\mathbb{P}_0$. The Law of Large Numbers (LLN) indicates that $\mathbb{P}_0$-almost surely, $\mu_t^{(n)} \rightharpoonup \mu_t$ as $n \to \infty$, where μ_t satisfies the mean-field gradient flow [12, 47, 52, 61]:

$$\partial_t \mu_t = \nabla \cdot (\nabla V(\boldsymbol{\theta},\mu_t)\mu_t) \,, \qquad \mu_{t=0} = \mu_0 \,. \tag{13}$$

The solution to this equation should be understood via the representation formula

$$\int_D \chi(\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\mu_0(d\boldsymbol{\theta}) \,, \tag{14}$$

where χ is a continuous test function $\chi : D \to \mathbb{R}$ and $\boldsymbol{\Theta}_t : D \to D$ is the *characteristic flow* associated with (12), which in direct analogy with (10) solves

$$\dot{\boldsymbol{\Theta}}_t(\boldsymbol{\theta}) = -\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t), \qquad \boldsymbol{\Theta}_0(\boldsymbol{\theta}) = \boldsymbol{\theta} \,. \tag{15}$$

Using expression (11) for V as well as (14), this equation can be written in closed form explicitly as

$$\dot{\boldsymbol{\Theta}}_t(\boldsymbol{\theta}) = \nabla F(\boldsymbol{\Theta}_t(\boldsymbol{\theta})) - \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\mu_0(d\boldsymbol{\theta}'), \qquad \boldsymbol{\Theta}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}\,. \tag{16}$$

It is easy to see that this equation is itself a gradient flow since it is the continuous-time limit of a proximal scheme (mirror descent), which we state as:

Proposition 2.4. *Given $\bar{\boldsymbol{\Theta}}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\tau > 0$, for $n \in \mathbb{N}$ let $\boldsymbol{\Theta}_{n\tau}$ be specified via*

$$\bar{\boldsymbol{\Theta}}_{n\tau} \in \operatorname{argmin}\left(\frac{1}{2\tau}\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}_{(n-1)\tau}\|_0^2 + \mathcal{E}(\boldsymbol{\Theta})\,,\right) \tag{17}$$

where we defined

$$\|\boldsymbol{\Theta}\|_0^2 = \int_D |\boldsymbol{\Theta}(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \tag{18}$$

and

$$\mathcal{E}(\boldsymbol{\Theta}) = -\int_D F(\boldsymbol{\Theta}(\boldsymbol{\theta}))\mu_0(d\boldsymbol{\theta}) + \frac{1}{2}\int_D K(\boldsymbol{\Theta}(\boldsymbol{\theta}), \boldsymbol{\Theta}(\boldsymbol{\theta}'))\mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}')\,. \tag{19}$$

Then

$$\lim_{\tau\to 0} \bar{\boldsymbol{\Theta}}_{\lfloor t/\tau \rfloor \tau} = \boldsymbol{\Theta}_t \qquad \mu_0\text{-almost surely}\,, \tag{20}$$

where $\boldsymbol{\Theta}_t$ solves (16).

2.5 Long-Time Properties of the Mean-Field Gradient Flow

In the shallow neural networks setting, a series of earlier work [12, 47, 51, 58] has established that under certain assumptions μ_t will converge to a global minimizer of the loss functional $\mathcal{L}$. In particular, [12] studies global convergence for the regularized loss $\mathcal{L}$ under homogeneity assumptions on $\hat{\varphi}$, and [50] considers modified dynamics using *double-lifting*. Here, to study the long time behavior of the fluctuations, we will often work with the following weaker assumptions:

Assumption 2.5. *The solution to (16) exists for all time, and has a limit:*

$$\boldsymbol{\Theta}_t \to \boldsymbol{\Theta}_\infty \quad \mu_0\text{-almost surely as } t \to \infty. \tag{21}$$

Assumption 2.6. *The limiting $\boldsymbol{\Theta}_\infty$ is a local minimizer of the energy (19).*

Note that while the convergence of gradient flows in finite-dimensional Euclidean space to local minimizers is guaranteed under mild assumptions [39, 62], its infinite-dimensional counterpart, Assumption 2.6, may require further technical assumptions that are left for future study. In addition, these assumptions impose conditions on the initial measure μ_0 [12, 47, 51].

With these assumptions, we have

Proposition 2.7. *Under Assumptions 2.3 and 2.5, we have*

$$\cup_{t\geq 0} \operatorname{supp} \mu_t = \cup_{t\geq 0}\{\boldsymbol{\Theta}_t(\boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \operatorname{supp} \mu_0\} \ \text{ is compact,} \tag{22}$$

and $\mu_t \rightharpoonup \mu_\infty$ weakly as $t \to \infty$, with μ_∞ satisfying

$$\int_D \chi(\boldsymbol{\theta})\mu_\infty(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\mu_0(d\boldsymbol{\theta}), \tag{23}$$

for all continuous test function $\chi : D \to \mathbb{R}$. Additionally, if Assumption 2.6 also holds, then

$$\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) \ \text{ is positive semidefinite for } \mu_0\text{-almost all } \boldsymbol{\theta} \tag{24}$$

We prove this proposition in Appendix B. Here, $\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$ denotes

$$\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) = -\nabla\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \int_D \nabla\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'))\mu_0(d\boldsymbol{\theta}') \tag{25}$$

and (24) will be useful later in Section 3.2 when we analyze the long time properties of the fluctuations around the mean-field limit. We note that Assumption 2.6 does not imply that μ_∞ is a minimizer of the energy $\mathcal{L}$ defined in (5), as it only needs to be a stationary point of (13).

3 Fluctuations from Mean-Field Gradient Flow

The main goal of this section is to characterize the deviations of finite-particle shallow networks from their mean-field evolution, by first deriving an estimate for $f_t^{(n)} - f_t$ for $t \geq 0$ (Section 3.1), and then analyzing its long-time properties (Section 3.2). In Section 3.3, we then motivate a choice of the regularization term in (5) that controls the bound on the long-time fluctuations derived in Section 3.2, and which is also connected to the variation-norm function spaces of Bach [5].

3.1 A Dynamical Central Limit Theorem

Let us start by defining

$$g_t^{(n)} := n^{1/2}\big(f_t^{(n)} - f_t\big) \, . \tag{26}$$

By the static Central Limit Theorem (CLT) we know that, if we draw the initial values of the parameters $\boldsymbol{\theta}_i$ independently from μ_0 as specified in Assumption 2.3, $g_{t=0}^{(n)}$ has a limit as $n \to \infty$, leading to estimates similar to (6) with $\mu^{(n)}$ and μ replaced by the initial $\mu_0^{(n)}$ and μ_0, respectively. For $t > 0$, however, this estimate is not preserved by the gradient flow: the static CLT no longer applies and needs to be replaced by a dynamical variant [10, 60, 65, 66]. Next, we derive this dynamical CLT in the context of neural network optimization.

To this end let us define the discrepancy measure $\omega_t^{(n)}$ such that

$$\int_D \chi(\boldsymbol{\theta})\omega_t^{(n)}(d\boldsymbol{\theta}) := n^{1/2}\int_D \chi(\boldsymbol{\theta})\left(\mu_t^{(n)}(d\boldsymbol{\theta}) - \mu_t(d\boldsymbol{\theta})\right) \, , \tag{27}$$

for any continuous test function $\chi : D \to \mathbb{R}$. We can then represent $g_t^{(n)}$ in terms of $\omega_t^{(n)}$ as

$$g_t^{(n)} = \int_D \varphi(\boldsymbol{\theta}, \cdot)\omega_t^{(n)}(d\boldsymbol{\theta}) \, . \tag{28}$$

Hence, we will first establish how the limit of $\omega_t^{(n)}$ as $n \to \infty$ evolves over time. This can be done upon noting that the representation formula (14) implies that

$$\int_D \chi(\boldsymbol{\theta})\omega_t^{(n)}(d\boldsymbol{\theta}) = n^{1/2}\int_D \left(\chi(\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}))\mu_0^{(n)}(d\boldsymbol{\theta}) - \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\mu_0(d\boldsymbol{\theta})\right) \, , \tag{29}$$

where $\boldsymbol{\Theta}_t^{(n)}$ solves (16) with μ_0 replaced by $\mu_0^{(n)}$. Defining

$$\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}) = n^{1/2}\big(\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}) - \boldsymbol{\Theta}_t(\boldsymbol{\theta})\big) \, , \tag{30}$$

we can write (29) as

$$\begin{aligned}\int_D \chi(\boldsymbol{\theta})\omega_t^{(n)}(d\boldsymbol{\theta}) &= \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\omega_0^{(n)}(d\boldsymbol{\theta}) \\ &\quad + \int_0^1 \int_D \nabla\chi\big(\boldsymbol{\Theta}_t(\boldsymbol{\theta}) + n^{-1/2}\eta\, \boldsymbol{T}_t^{(n)}(\boldsymbol{\theta})\big) \cdot \boldsymbol{T}_t^{(n)}(\boldsymbol{\theta})\mu_0^{(n)}(d\boldsymbol{\theta})d\eta\,. \end{aligned} \tag{31}$$

As shown in Appendix C.1, we can take the limit $n \to \infty$ of this formula to obtain:

Proposition 3.1 (Dynamical CLT - I)**.** *Under Assumptions 2.2 and 2.3, $\forall t \geq 0$, as $n \to \infty$ we have $\omega_t^{(n)} \rightharpoonup \omega_t$ weakly in law with respect to $\mathbb{P}_0$, where ω_t is such that given any test function $\chi : D \to \mathbb{R}$,*

$$\int_D \chi(\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\omega_0(d\boldsymbol{\theta}) + \int_D \nabla\chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta})) \cdot \boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\,. \tag{32}$$

Here ω_0 is the Gaussian measure with mean zero and covariance

$$\mathbb{E}_0\left[\omega_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}')\right] = \mu_0(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}') - \mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}')\,, \tag{33}$$

where $\mathbb{E}_0$ denotes expectation over $\mathbb{P}_0$, and $\boldsymbol{T}_t = \lim_{n\to\infty} n^{1/2}(\boldsymbol{\Theta}_t^{(n)} - \boldsymbol{\Theta}_t)$ is the flow solution to

$$\begin{aligned}\dot{\boldsymbol{T}}_t(\boldsymbol{\theta}) = &-\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{T}_t(\boldsymbol{\theta}) - \int_D \nabla\nabla' K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') \\ &- \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta}')\end{aligned} \tag{34}$$

with initial condition $\boldsymbol{T}_0 = 0$ and where $\boldsymbol{\Theta}_t$ solves (15) and $\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)$ is a shorthand for

$$\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t) = -\nabla\nabla F(\boldsymbol{\Theta}_t(\boldsymbol{\theta})) + \int_D \nabla\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\mu_0(d\boldsymbol{\theta}')\,. \tag{35}$$

A direct consequence of this proposition and formula (28) is:

Corollary 3.2. *Under Assumptions 2.2 and 2.3, $\forall t \geq 0$, as $n \to \infty$ we have $g_t^{(n)} \to g_t$ pointwise in law with respect to $\mathbb{P}_0$, where g_t can be expressed in terms of the limiting measure ω_t or the flow $\boldsymbol{T}_t$ as*

$$g_t = \int_D \varphi(\boldsymbol{\theta}, \cdot)\omega_t(d\boldsymbol{\theta}) = \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot)\omega_0(d\boldsymbol{\theta}) + \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot) \cdot \boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\,. \tag{36}$$

It is interesting to comment on the origin of both terms at the right hand side of (32) and, consequently, (36). The first term captures the deviations induced by fluctuations of $\mu_0^{(n)}$ around μ_0 assuming that the flow $\boldsymbol{\Theta}_t^{(n)}$ is unaffected by these fluctuations, and remains equal to $\boldsymbol{\Theta}_t$. In particular, this term is the one we would obtain if we were to resample $\mu_t^{(n)}$ from μ_t at every $t \geq 0$, i.e. use $\bar{\mu}_t^{(n)} = n^{-1}\sum_{i=1}^n \delta_{\bar{\boldsymbol{\theta}}_t^i}$ with $\{\bar{\boldsymbol{\theta}}_t^i\}_{i=1}^n$ sampled i.i.d. from μ_t, so that $\boldsymbol{\Theta}_t^{(n)}$ is identical to $\boldsymbol{\Theta}_t$ in (29). In this case, the limiting discrepancy measure $\bar{\omega}_t$ would simply be given by

$$\int_D \chi(\boldsymbol{\theta})\bar{\omega}_t(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\omega_0(d\boldsymbol{\theta})\,, \tag{37}$$

while the associated deviation in the represented function would read

$$\bar{g}_t = \int_D \varphi(\boldsymbol{\theta}, \cdot)\bar{\omega}_t(d\boldsymbol{\theta}) = \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot)\omega_0(d\boldsymbol{\theta})\,. \tag{38}$$

The second term at right hand side of (32) and (36) captures the deviations to the flow $\boldsymbol{\Theta}_t$ in (16) induced by the perturbation of μ_0, i.e. how much $\boldsymbol{\Theta}_t^{(n)}$ differs from $\boldsymbol{\Theta}_t$ in (29). In the limit as $n \to \infty$, these deviations are captured by the solution $\boldsymbol{T}_t$ to (34), as is apparent from (31).

The difference between g_t and $\bar{g}_t$ can also be quantified via the following Volterra equation, which can be derived from Proposition 3.1 and relates the evolution of g_t to that of $\bar{g}_t$.

Corollary 3.3 (Dynamical CLT - II)**.** *Under Assumptions 2.2 and 2.3, $\forall t \geq 0$, pointwise on Ω, we have $g_t^{(n)} \to g_t$ in law with respect to $\mathbb{P}_0$ as $n \to \infty$, where g_t solves the Volterra equation*

$$g_t(\boldsymbol{x}) + \int_0^t \int_\Omega \Gamma_{t,s}(\boldsymbol{x}, \boldsymbol{x}') g_s(\boldsymbol{x}') \hat{\nu}(d\boldsymbol{x}') ds = \bar{g}_t(\boldsymbol{x}) \ . \tag{39}$$

Here $\bar{g}_t$ is given in (38) and we defined

$$\Gamma_{t,s}(\boldsymbol{x}, \boldsymbol{x}') = \int_D \langle \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta})), J_{t,s}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta})) \rangle \mu_0(d\boldsymbol{\theta}) \ , \tag{40}$$

where $J_{t,s}$ is the solution to

$$\frac{d}{dt} J_{t,s}(\boldsymbol{\theta}) = -\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t) J_{t,s}(\boldsymbol{\theta}), \qquad J_{s,s}(\boldsymbol{\theta}) = Id \ . \tag{41}$$

This corollary is proven in Appendix C.2. In a nutshell, (39) can be established using Duhamel's principle on (34) by considering all terms at the right hand side except the first as the source term (hence the role of $J_{t,s}$) and inserting the result in (36).

3.2 Long-Time Behavior of the Fluctuations

Next, we study the long-time behavior of g_t and, in particular, evaluate

$$\lim_{t\to\infty} \mathbb{E}_0 \|g_t\|_{\hat{\nu}}^2 = \lim_{t\to\infty} \lim_{n\to\infty} n \mathbb{E}_0 \|f_t^{(n)} - f_t\|_{\hat{\nu}}^2. \tag{42}$$

This limit quantifies the asymptotic approximation error of $f_t^{(n)}$ around its mean field limit f_t after gradient flow, i.e. if we take $n \to \infty$ first, then $t \to \infty$ – taking these limits in opposite order is of practical interest too but is beyond the scope of the present paper. Our main result is to show that, under certain assumptions to be specified below, the limit in (42) is not only finite but it is then necessarily smaller than $\lim_{t\to\infty} \mathbb{E}_0 \|\bar{g}_t\|_{\hat{\nu}}^2$ with $\bar{g}_t$ given in (38). That is, the approximation error at the end of training by gradient flow is always lower than that obtained by resampling the mean-field measure μ_∞ defined in Proposition 2.7.

It is useful to start by considering an idealized case, namely when the initial conditions are sampled as in Assumption 2.3 with $\mu_0 = \mu_\infty$. In that case, there is no evolution at mean field level, i.e. $\boldsymbol{\Theta}_t(\boldsymbol{\theta}) = \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\mu_t = \mu_\infty$, and $f_t = f_\infty = \int_D \varphi_\infty(\boldsymbol{\theta}, \cdot) \mu_\infty(d\boldsymbol{\theta})$, but the CLT fluctuations still evolve. In particular, it is easy to see that the Volterra equation in (39) for g_t becomes

$$g_t(\boldsymbol{x}) + \int_0^t \int_\Omega \Gamma_{t-s}^\infty(\boldsymbol{x}, \boldsymbol{x}') g_s(\boldsymbol{x}') \hat{\nu}(d\boldsymbol{x}') ds = \bar{g}_\infty(\boldsymbol{x}) \ . \tag{43}$$

Here $\Gamma_{t-s}^\infty(\boldsymbol{x}, \boldsymbol{x}')$ is the Volterra kernel obtained by solving (41) with $\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)$ replaced by $\nabla\nabla V(\boldsymbol{\theta}, \mu_\infty)$ and inserting the result in (40) with $\boldsymbol{\Theta}_t(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\mu_0 = \mu_\infty$,

$$\Gamma_{t-s}^\infty(\boldsymbol{x}, \boldsymbol{x}') = \int_D \langle \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}, \boldsymbol{x}), e^{-(t-s)\nabla\nabla V(\boldsymbol{\theta}, \mu_\infty)} \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}, \boldsymbol{x}') \rangle \mu_\infty(d\boldsymbol{\theta}) \ , \tag{44}$$

and $\bar{g}_\infty$ is the Gaussian field with variance

$$\mathbb{E}_0\|\bar{g}_\infty\|_{\hat{\nu}}^2 = \int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2 \mu_\infty(d\boldsymbol{\theta}) - \|f_\infty\|_{\hat{\nu}}^2 \,. \tag{45}$$

From (24) in Proposition 2.7 we know that $\nabla\nabla V(\boldsymbol{\theta},\mu_\infty)$ is positive semidefinite for μ_∞-almost all $\boldsymbol{\theta}$. As a result, we prove in D.1 that the Volterra kernel (44) viewed as an operator on functions defined on $D\times[0,T]$ is positive semidefinite. Therefore, we have

$$\begin{aligned}\int_0^T \|g_t\|_{\hat{\nu}}^2 dt &\le \int_0^T \|g_t\|_{\hat{\nu}}^2 dt + \int_0^T\int_0^t\int_{\Omega\times\Omega} g_t(\boldsymbol{x})\Gamma_{t-s}^\infty(\boldsymbol{x},\boldsymbol{x}')g_s(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}')dsdt \\ &= \int_0^T \mathbb{E}_{\hat{\nu}}(g_t\bar{g}_\infty)dt \\ &\le T^{1/2}\|\bar{g}_\infty\|_{\hat{\nu}}\left(\int_0^T \|g_t\|_{\hat{\nu}}^2 dt\right)^{1/2} \,.\end{aligned} \tag{46}$$

Together with (45), this implies that

Theorem 3.4. *Under Assumptions 2.2, 2.3, 2.5 and 2.6, with $\mu_0=\mu_\infty$ and μ_∞ as specified in Proposition 2.7, we have*

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}_0\|g_t\|_{\hat{\nu}}^2 dt \le \int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2\mu_\infty(d\boldsymbol{\theta}) - \|f_\infty\|_{\hat{\nu}}^2 \,. \tag{47}$$

This theorem indicates that, if we knew μ_∞ and were able to sample initial conditions for the parameters from it, it would still be favorable to train these parameters by gradient descent as this would reduce the approximation error. Of course, in practice we have no *a priori* access to μ_∞, and so the relevant question is whether (47) also holds if we sample initial conditions from any μ_0 such that Proposition 2.7 holds.

In light of (36), one way to address this question is to study the long-time behavior of $\boldsymbol{T}_t$. In the setup without regularization ($\lambda=0$), we can do so by leveraging existing results that, under certain assumptions, the mean-field gradient flow converges to a global minimizer which interpolates the training data points exactly [12, 47, 52, 60]. In this case, the following theorem shows that we can obtain stronger controls on the fluctuations than (47), which we prove in Appendix D.2.

Theorem 3.5 (Long-time fluctuations in the unregularized case)**.** *Consider the ERM setting with $\lambda=0$ and under Assumptions 2.2, 2.3 and 2.5. Suppose that as $t\to\infty$, μ_t converges to a global minimizer μ_∞ that interpolates the data, i.e. the function $f_\infty=\int_D\varphi(\boldsymbol{\theta},\cdot)\mu_\infty(d\theta)$ satisfies*

$$\forall\boldsymbol{x}\in\operatorname{supp}\hat{\nu}\ :\ f_\infty(\boldsymbol{x}) = f_*(\boldsymbol{x})\,, \tag{48}$$

and, furthermore, the convergence satisfies

$$\int_0^\infty |\mathcal{L}(\mu_t)|^{1/2}\,dt<\infty \tag{49}$$

Then (47) holds. Additionally,

1. *if Assumption 2.1 also holds, i.e., in the shallow neural network setting, we further have*

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}_0\|g_t\|_{\hat{\nu}}^2 dt = 0\,; \tag{50}$$

2. *if $\mu_0 = \mu_\infty$, then $\|g_t\|_{\hat{\nu}}$ decreases monotonically in t.*

Remark 3.6. *This implies that in the shallow neural networks setting, under these assumptions, the fluctuations will eventually vanish in the $O(n^{-1/2})$ scale of CLT.*

Remark 3.7. *For (49) to hold, it is sufficient to have the loss $\mathcal{L}(\mu_t)$ decrease at an asymptotic rate of $O(t^{-\alpha})$ with $\alpha > 2$. We note that this is a stronger requirement than the $O(t^{-1})$ achieved by Mirror Descent in certain related settings [11, 50]. We leave the search for weaker sufficient conditions for future work.*

When the limiting measure μ_∞ does not necessarily interpolate the training data, such as in the regularized case, we have the following condition on $\boldsymbol{T}_t$ which guarantees that (47) holds:

Lemma 3.8. *If*

$$\lim_{T\to\infty} \mathbb{E}_0 \int_0^T \int_D \langle \boldsymbol{T}_t(\boldsymbol{\theta}), \nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t) \boldsymbol{T}_t(\boldsymbol{\theta})\rangle \mu_0(d\boldsymbol{\theta}) dt \geq 0 \,, \tag{51}$$

(including when this limit is $+\infty$) then (47) holds.

This lemma is proven in Appendix D.3. Note that condition (51) is natural since we know from Proposition 2.7 that $\lim_{t\to\infty} \nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t) = \nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$ exists and is positive semidefinite μ_0-almost surely. This lemma allow us to derive the following result:

Theorem 3.9 (Long-time fluctuations under assumptions on the curvature)**.** *Let $\Lambda_t(\boldsymbol{\theta})$ denotes to smallest eigenvalue of the tensor $\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)$ defined in (35) and assume that there exists a constant C (to be specified in Appendix D.4) such that*

$$-\int_D \min\{\Lambda_t(\boldsymbol{\theta}), 0\} \mu_0(d\boldsymbol{\theta}) = O(e^{-Ct}) \qquad \text{as } t \to \infty. \tag{52}$$

Then (51) and hence (47) hold.

This theorem is proven in Appendix D.4. To intuitively understand (52), note that we know from (24) in Proposition 2.7 that $\Lambda_t(\boldsymbol{\theta}) \to 0$ μ_0-almost surely as $t \to \infty$. Condition (52) can therefore be satisfied by having $\Lambda_t(\boldsymbol{\theta})$ converge to zero sufficiently fast in the regions of D where it is negative, or having the measure of these regions with respect to μ_0 converge to zero sufficiently fast, or both.

Finally, in the regularized ($\lambda > 0$) ERM setting, we can obtain the following result when the support of μ_∞ is atomic, as expected on general grounds [5, 9, 18, 26, 71]:

Theorem 3.10 (Long-time fluctuations in the regularized case)**.** *Consider the ERM setting under Assumptions 2.2, 2.3 and 2.5. Suppose further that as $t \to \infty$, μ_t converges to μ_∞ satisfying*

$$\exists \sigma > 0 \ \text{ s.t. } \ \forall \boldsymbol{\theta} \in \operatorname{supp} \mu_\infty \ : \ \nabla\nabla V(\boldsymbol{\theta}, \mu_\infty) \succ \sigma \boldsymbol{Id} \,, \text{ and} \tag{53}$$

$$\boldsymbol{\Theta}_t \text{ admits an asymptotic uniform convergence rate of } O(t^{-\alpha}) \text{ with } \alpha > 3/2. \tag{54}$$

Then (47) holds with the "lim" replaced by "lim sup" on its LHS.

Remark 3.11. *We prove in Appendix D.5 that the theorem holds with (54) replaced by a weaker condition:*

$$\int_0^\infty \int_D \left(|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})| + |U_t(\boldsymbol{\theta})|^2\right) e^{C_1(U_t(\boldsymbol{\theta}) + \bar{U}_t)} \mu_0(d\boldsymbol{\theta}) dt < \infty \,, \tag{55}$$

and show in Appendix D.5.2 that (54) is a sufficient condition for (55).

Remark 3.12. *[18] provides evidence for (55) being satisfied by the minimizers of the TV-norm-regularized loss function. In addition, as a comparison to our result, Chizat [11, Theorem 3.8] shows that under assumptions including (53) as well as the uniqueness and sparseness of the global minimizer, an alternative type of particle gradient descent (with a different homogeneity degree in the loss function and under the conic metric, which give rise to gradient flow in Wasserstein-Fisher-Rao metric instead of Wasserstein metric) converges to the global minimizer for large enough n (depending exponentially on d) with a uniform rate. This implies that in that setting, $\lim_{t\to\infty}\lim_{n\to\infty} n\|f_t^{(n)} - f_t\|_{\hat{\nu}}^2 = \lim_{n\to\infty}\lim_{t\to\infty} n\|f_t^{(n)} - f_t\|_{\hat{\nu}}^2 = 0$, $\mathbb{P}_0$-almost surely.*

Theorem 3.10 is proven in Appendix D.5 by analyzing directly the Voterra equation (39) and establishing that its solution coincides with that of (43) in the limit as $t \to \infty$, a property that we also expect to hold more generally than under the assumptions of Theorem 3.10.

3.3 The Monte-Carlo bound and variation norm

The bound (47) on the long-time fluctuations motivates us to control the term $\int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2 \mu_\infty(d\boldsymbol{\theta})$ using a suitable choice of regularization in (5). In the following, we restrict our attention to the shallow neural networks setting, and further assume that

Assumption 3.13. *$\hat{D}$ is compact.*

Under this assumption, there is

$$\int_D \|\varphi(\boldsymbol{\theta},\cdot)\|_{\hat{\nu}}^2 \mu(d\boldsymbol{\theta}) = \int_D \int_\Omega |\varphi(\boldsymbol{\theta},\boldsymbol{x})|^2 \hat{\nu}(d\boldsymbol{x})\mu(d\boldsymbol{\theta}) \le \hat{K}_M \int_D c^2 \mu(d\boldsymbol{\theta}) \,, \tag{56}$$

where $\hat{K}_M = \max_{\boldsymbol{z}\in\hat{D}} \|\hat{\varphi}(\boldsymbol{z},\cdot)\|_{\hat{\nu}}^2$. Thus, we consider regularization with $r(\boldsymbol{\theta}) = \frac{1}{2}c^2$, in which case (5) becomes

$$\min_{\mu\in\mathcal{P}(D)} \mathcal{L}(\mu) \quad \text{with} \quad \mathcal{L}(\mu) := \|f[\mu] - f_*\|_{\hat{\nu}}^2 + \frac{\lambda}{2}\int_D c^2 \mu(d\boldsymbol{\theta}) \,. \tag{57}$$

Interestingly, this choice of regularization leads to learning in the function space $\mathcal{F}_1$ [5] associated with $\hat{\varphi}$, which is equipped with the *variation norm* defined as

$$|\gamma_q(f)| := \inf_{\mu\in\mathcal{P}(D)} \left\{ \int_D |c|^q \mu(d\boldsymbol{\theta});\ f(\boldsymbol{x}) = \int_D c\hat{\varphi}(\boldsymbol{z},\boldsymbol{x})\mu(d\boldsymbol{\theta}) \right\} = |\gamma_1(f)|^q \,, \qquad q \ge 1 \,. \tag{58}$$

We call $\int_D |c|^q \mu(d\boldsymbol{\theta})$ the *q-norm* of μ. One can verify [44, Proposition 1] that indeed, using any $q \ge 1$ above yields the same norm because μ, the object defining the integral representation (3), is in fact a *lifted* version of a more 'fundamental' object $\gamma = \int_{\mathbb{R}} \mu(dc,\cdot) \in \mathcal{M}(\hat{D})$, the space of signed Radon measures over $\hat{D}$. They are related via the projection

$$\int_{\hat{D}} \chi(\boldsymbol{z})\gamma(d\boldsymbol{z}) = \int_D c\chi(\boldsymbol{z})\mu(d\boldsymbol{\theta}) \tag{59}$$

for all continuous test functions $\chi : \hat{D} \to \mathbb{R}$. One can also verify [11] that $\gamma_1(f) = \inf\{\|\gamma\|_{\mathrm{TV}};\ f(\boldsymbol{x}) = \int_{\hat{D}} \hat{\varphi}(\boldsymbol{z},\boldsymbol{x})\gamma(d\boldsymbol{z})\}$, where $\|\gamma\|_{\mathrm{TV}}$ is the *total variation* of γ [5].

The space $\mathcal{F}_1$ contains any RKHS whose kernel is generated as an expectation over features $k(\boldsymbol{x},\boldsymbol{x}') = \int_{\hat{D}} \hat{\varphi}(\boldsymbol{z},\boldsymbol{x})\hat{\varphi}(\boldsymbol{z},\boldsymbol{x}')\hat{\mu}_0(d\boldsymbol{z})$ with a base measure $\hat{\mu}_0 \in \mathcal{P}(\hat{D})$, but it provides crucial approximation advantages over such RKHS at approximating certain non-smooth, high-dimensional functions with hidden low-dimensional structure, giving rise to powerful generalization guarantees

[5]. This also motivates the study of overparametrized shallow networks with the scaling as in (1), as opposed to the NTK scaling of $1/\sqrt{n}$ [36].

To learn in $\mathcal{F}_1$, a canonical approach is to consider the ERM problem

$$\min_{f\in\mathcal{F}_1} \|f - f_*\|_{\hat{\nu}}^2 + \lambda\gamma_1(f), \tag{60}$$

By (58), this is equivalent to (57). In Appendix E, we prove the following proposition, which shows that the measure obtained from (57) indeed has its 2-norm controlled:

Proposition 3.14. *Under Assumption 2.2, the minimum value of $\mathcal{L}$ is unique and can only be attained at minimizers $\mu_\lambda \in \mathcal{P}(D)$ such that $f_\lambda = \int_D \varphi(\boldsymbol{\theta}, \cdot)\mu_\lambda(d\boldsymbol{\theta})$ is unique and satisfies*

$$\lambda^2|c_\lambda|^2\hat{K}_M^{-1} \leq \|f_\lambda - f_*\|_{\hat{\nu}}^2, \qquad \|f_\lambda - f_*\|_{\hat{\nu}}^2 + \lambda|c_\lambda|^2 \leq \lambda|\gamma_1(f_*)|^2. \tag{61}$$

where $c_\lambda = \int_D |c|\mu_\lambda(d\boldsymbol{\theta}) = \left(\int_D |c|^2\mu_\lambda(d\boldsymbol{\theta})\right)^{1/2} \leq \gamma_1(f_*)$ and $\hat{K}_M = \max_{\boldsymbol{z}\in\hat{D}} \|\hat{\varphi}(\boldsymbol{z},\cdot)\|_{\hat{\nu}}^2$.

4 Numerical Experiments

We perform numerical experiments under the student-teacher setting, with a shallow teacher network with 2 neurons in the hidden layer, and shallow student networks with varying widths (n) of the hidden layer. Both $\hat{D}$ and Ω are taken to be the unit sphere of $d = 16$ dimensions, and we take $\hat{\varphi}(\boldsymbol{z}, \boldsymbol{x}) = \max(0, \langle \boldsymbol{z}, \boldsymbol{x}\rangle)$. As a proof of concept, we optimize the student networks by gradient descent under the population loss where the data distribution is uniform on Ω. This allows an analytical formula for the gradient that greatly accelerates computations and provides an approximation for the empirical setting. For each choice of n, we run the experiment $\kappa = 20$ times with different random initializations of the student network. Additional experimental setup and results are provided in Appendix F. The mean squared fluctuation is defined as $\frac{1}{\kappa}\sum_{k=1}^{\kappa} \|f_k^{(n)} - \bar{f}^{(n)}\|_{\nu}^2$, with $\bar{f}^{(n)} = \frac{1}{\kappa}\sum_{k=1}^{\kappa} f_k^{(n)}$ being the averaged model, similar to the approach in [28]. The other plotted quantities – loss, TV-norm and 2-norm – are averaged across the κ number of runs.

We observe that the mean squared fluctuations indeed remain at a $1/n$ scaling with a general tendency to decay over time, except for in the zero-initialization case, where an increase occurs in the first few epochs of training. In the unregularized case with non-zero initialization, the mean squared fluctuation decays at the same rate for different n in roughly the first 10^3 epochs, after which it decays faster for smaller n. Interestingly, this coincides with the tendency for student neurons with $\boldsymbol{z}$ not aligned with the teacher neurons to slowly have their $|c|$ decrease to zero due to a finite-n effect, which is also reflected in the decrease in TV-norm. Aside from this, the mean squared fluctuations decay at similar rates for different choices of n, which is consistent with our theory, since their dynamics are governed by the same dynamical CLT. Moreover, the value of the fluctuations for these choices of n are indeed lower than the asymptotic Monte-Carlo bound given in (47) with μ_∞ and f_∞ replaced by the target measure and function, whose analytical expression and numerical value in this setup are provided in Appendix G. We also see that the average loss values remain similar over time for different choices of n, justifying the approximation by a mean-field dynamics. Also, we notice that with either regularization or zero-initialization, the student neurons are aligned with one of the teacher neurons in both $\boldsymbol{z}$ and c after training, which then results in lower TV-norms and 2-norms than using non-zero initialization and without regularization.

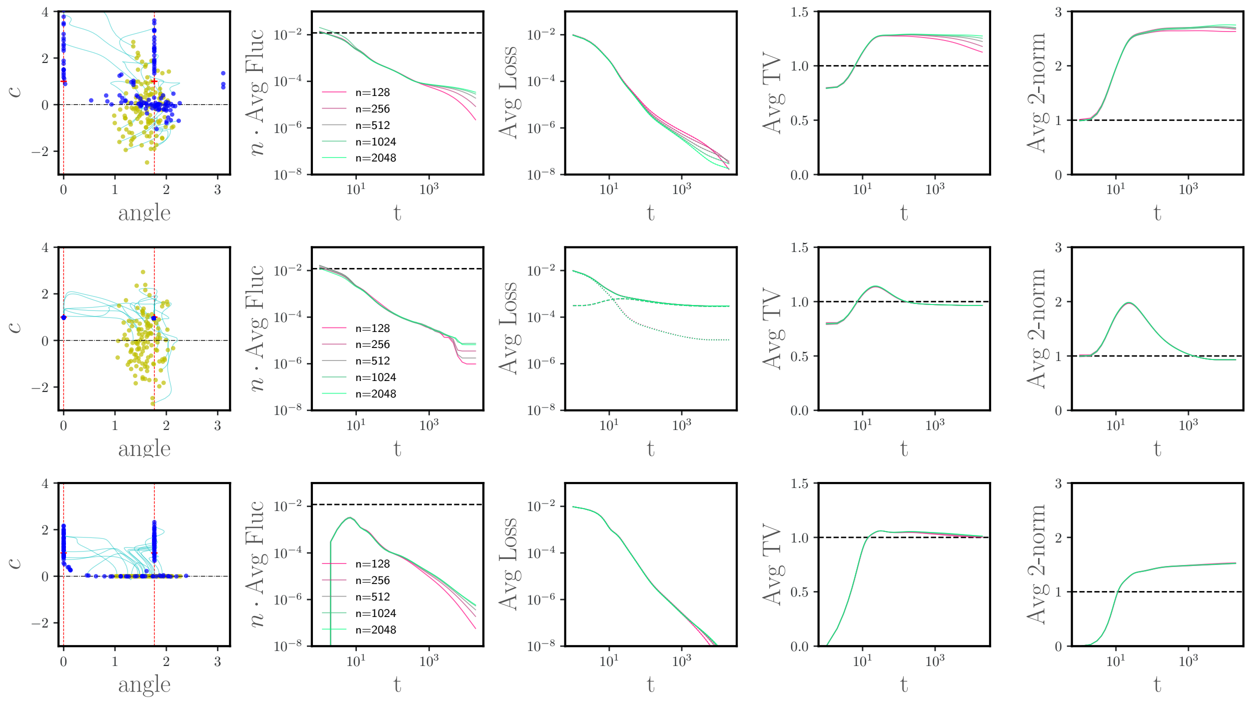

Figure 1: Results of numerical experiments described in Section 4, where student networks of different widths are trained to learn teacher networks of width 2 by performing gradient descent on the exact population loss. Each row corresponds to one setup. *Row 1*: Using unregularized loss and non-zero initialization; *Row 2*: Using regularized loss with $\lambda = 0.01$ and non-zero initialization; *Row 3*: Using regularized loss with $\lambda = 0.01$ and zero initialization. On each row, the *first* panel from the left plots the trajectory of the neurons, $\boldsymbol{\theta}_i = (c_i, \boldsymbol{z}_i)$, of a student network of width 128 during its training, with x-coordinate being the angle between $\boldsymbol{z}_i$ and that of a chosen teacher's neuron and y-coordinate being c_i, and where the yellow dots, blue dots and cyan curves marking their initial values, terminal values, and trajectory during training. The *second* to *fifth* panels plot the average fluctuations (scaled by n), average loss, average TV norm, and average 2-norm during training, respectively, under different choices of n, the width of the student network, and computed over 5 runs with different random initializations of the student network. The several quantities are defined in Section 4. In the *third* panel, the dotted curve indicates the reconstruction error, the dashed curve the regularization term, and the solid curve the sum of these two, i.e. total regularized loss. In the *fourth* and *fifth* panels, the horizontal dashed line gives the relevant norm of the teacher network.

5 Conclusions

We studied the deviation of shallow neural networks from their infinite-width limit. In the ERM setting, we establish that under different sets of conditions, finite-width networks approach the mean-field limit with an error that scales strictly better or no worse than that of MC resampling, giving width-asymptotic guarantees that do not depend on the data dimension explicitly. The MC resampling bound motivates a choice of regularization that is also connected to generalization via the variation-norm function spaces.

Our results thus seem to paint a favorable picture for high-dimensional learning, in which the optimization and generalization guarantees for the idealized mean-field limit could be transferred to their finite-particle counterparts. However, our results are still asymptotic, in that we take limits both in the number of particles and time. In the face of negative results for the computational efficiency of training shallow networks [20, 31, 41, 45, 54], an important challenge is to leverage additional structure in the problem (such as the empirical data distribution [32], or the structure of the minimisers [18]) to provide nonasymptotic versions of our results, along the lines of [11] or [40]. Finally, another clear direction for future research is to extend our techniques to deep neural architectures, in light of recent works that consider deep or residual models [3, 25, 42, 49, 59, 68].

Acknowledgements: We appreciate the helpful discussions with Lenaic Chizat and Carles Domingo-Enrich. Z.C. acknowledges support from the Henry MacCraken Fellowship. G.M.R. acknowledges support from the James S. McDonnell Foundation. J.B. acknowledges support from the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, and the Institute for Advanced Study. E. V.-E. acknowledges support by National Science Foundation (NSF) Materials Research Science and Engineering Center Program Grant No. DMR-1420073, and by NSF Grant No. DMS-1522767.

References

[1] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[3] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

[4] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[5] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[6] Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

[7] Javier Baladron, Diego Fasoli, Olivier Faugeras, and Jonathan Touboul. Mean field description of and propagation of chaos in recurrent multipopulation networks of hodgkin-huxley and fitzhugh-nagumo neurons. *arXiv preprint arXiv:1110.4294*, 2011.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv:1812.11118 [cs, stat]*, September 2019. arXiv: 1812.11118.

[9] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.

[10] Werner Braun and K Hepp. The vlasov dynamics and its fluctuations in the $1/n$ limit of interacting classical particles. *Communications in mathematical physics*, 56(2):101–113, 1977.

[11] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

[12] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

[13] Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[14] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[15] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[16] Roberto Cortez. Uniform propagation of chaos for kac's 1d particle system. *Journal of Statistical Physics*, 165(6):1102–1113, 2016.

[17] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. *arXiv preprint arXiv:2007.06352*, 2020.

[18] Jaume de Dios and Joan Bruna. On sparsity in overparametrised shallow relu networks. *arXiv preprint arXiv:2006.10225*, 2020.

[19] Pierre Del Moral and Laurent Miclo. Branching and interacting particle systems approximations of feynman-kac formulae with applications to non-linear filtering. In *Seminaire de probabilites XXXIV*, pages 1–145. Springer, 2000.

[20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539, 2020.

[21] 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.

[22] Alain Durmus, Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. An elementary approach to uniform in time propagation of chaos. *arXiv preprint arXiv:1805.11387*, 2018.

[23] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. *arXiv preprint arXiv:1909.11304*, 2019.

[24] Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint, 2019.

[25] Cong Fang, Jason D Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. *arXiv preprint arXiv:2007.01452*, 2020.

[26] SD Fisher and Joseph W Jerome. Spline solutions to l1 extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.

[27] 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. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[28] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy learning in deep neural networks: an empirical study. *arXiv preprint arXiv:1906.08034*, 2019.

[29] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9111–9121, 2019.

[30] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

[31] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. *arXiv preprint arXiv:2006.12011*, 2020.

[32] Sebastian Goldt, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv preprint arXiv:2006.14709*, 2020.

[33] G. Gripenberg, S. O. Londen, and O. Staffans. *Volterra Integral and Functional Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.

[34] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

[35] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

[36] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[37] Carlo Lancellotti. On the fluctuations about the vlasov limit for n-particle systems with mean-field interactions. *Journal of Statistical Physics*, 136(4):643–665, 2009.

[38] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[39] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[40] Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. volume 125 of *Proceedings of Machine Learning Research*, pages 2613–2682. PMLR, 09–12 Jul 2020.

[41] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[42] Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep resnet and beyond: Towards provable optimization via overparameterization from depth. *arXiv preprint arXiv:2003.05508*, 2020.

[43] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *arXiv preprint arXiv:2007.07497*, 2020.

[44] Chao Ma, Lei Wu, and Weinan E. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

[45] Pasin Manurangsi and Daniel Reichman. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

[46] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

[47] 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*, 115(33):E7665–E7671, 2018.

[48] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[49] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

[50] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*, 2019.

[51] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, pages 7146–7155, 2018.

[52] 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 preprint arXiv:1805.00915*, 2018.

[53] Nicolas Le Roux and Yoshua Bengio. Continuous neural networks. volume 2 of *Proceedings of Machine Learning Research*, pages 404–411, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[54] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, pages 4433–4441, 2018.

[55] Samir Salem. A gradient flow approach of uniform in time propagation of chaos for particles in double a well confinement. *arXiv preprint arXiv:1810.08946*, 2018.

[56] Jamil Salhi, James MacLaurin, and Salwa Toumi. On uniform propagation of chaos. *Stochastics*, 90(1):49–60, 2018.

[57] Sylvia Serfaty. Coulomb gases and ginzburg-landau vortices. *arXiv preprint arXiv:1403.6860*, 2014.

[58] Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

[59] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *arXiv preprint arXiv:1903.04440*, 2019.

[60] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[61] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[62] Stephen Smale. Stable manifolds for differential equations and diffeomorphisms. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 17(1-2):97–116, 1963.

[63] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[64] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

[65] Herbert Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

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

[67] Luca Venturi, Afonso S. Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

[68] Stephan Wojtowytsch et al. On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics. *arXiv preprint arXiv:2007.15623*, 2020.

[69] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

[70] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[71] S. Zuhovickii. Remarks on problems in approximation theory. *Mat. Zbirnik KDU*, 1948.

Appendix

Table of Contents

A Notations

We will use $\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})$ and $\nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})$ to denote $\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta}, \boldsymbol{x})$ and $\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta}, \boldsymbol{x})$, respectively. We will use $\nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to denote $\nabla_{\boldsymbol{\theta}} K(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\nabla\nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to denote $\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} K(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\nabla'\nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to denote $\nabla_{\boldsymbol{\theta}'}\nabla_{\boldsymbol{\theta}} K(\boldsymbol{\theta}, \boldsymbol{\theta}')$, and $\nabla'\nabla' K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to denote $\nabla_{\boldsymbol{\theta}'}\nabla_{\boldsymbol{\theta}'} K(\boldsymbol{\theta}, \boldsymbol{\theta}')$. We will write $V_t(\cdot)$ for $V(\cdot, \mu_t)$ and $V_\infty(\cdot)$ for $V(\cdot, \mu_\infty)$.

Let $D' = \cup_{t>0} \operatorname{supp} \mu_t$. Under Assumption 2.5 and Proposition 2.7, D' is bounded, and we denote its diameter by $|D'|$. We will use C_φ, $C_{\nabla\varphi}$ and $C_{\nabla\nabla\varphi}$ to denote the supremum of $|\varphi(\boldsymbol{\theta}, \boldsymbol{x})|$, $|\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})|$ and $|\nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})|$ over $\boldsymbol{\theta} \in D'$ and $\boldsymbol{x} \in \operatorname{supp} \hat{\nu}$, which are all finite under Assumptions 2.2 and the boundedness of D'. We will use $L_{\nabla\nabla\varphi}$ to denote the (uniform-in-$\boldsymbol{x}$) Lipschitz constant of $\nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})$ in $\boldsymbol{\theta}$, which is also finite under Assumption 2.2.

The following notations will be used in Appendix D.2: Assuming that D is Euclidean (under Assumption 2.2), let $\mathcal{V}(D)$ denote the space of random vector fields on D. It becomes a Hilbert space once equipped with the inner product

$$\langle \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \rangle_0 := \mathbb{E}_0 \int_D \boldsymbol{\xi}_1(\boldsymbol{\theta}) \cdot \boldsymbol{\xi}_2(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}), \tag{62}$$

where $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ denotes two random vector fields in $\mathcal{V}(D)$. This inner product gives rise to the norm

$$\|\boldsymbol{\xi}\|_0^2 := \mathbb{E}_0 \int_D |\boldsymbol{\xi}(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \,. \tag{63}$$

For each t, we define $\boldsymbol{b}_t \in \mathcal{V}(D)$ as

$$\boldsymbol{b}_t(\boldsymbol{\theta}) = \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta}') \tag{64}$$

which depends on the random measure ω_0. We define two linear operators, $\mathcal{A}_t^{(K)}$ and $\mathcal{A}_t^{(V)}$ on $\mathcal{V}(D)$, as

$$\begin{aligned}
(\mathcal{A}_t^{(K)}\boldsymbol{\xi})(\boldsymbol{\theta}) &= \int_D \nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{\xi}(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') &(65)\\
&= \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\Big(\int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal \boldsymbol{\xi}(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}')\Big)\hat{\nu}(d\boldsymbol{x}) \,, &(66)\\
(\mathcal{A}_t^{(V)}\boldsymbol{\xi})(\boldsymbol{\theta}) &= \nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{\xi}(\boldsymbol{\theta}) \,, &(67)
\end{aligned}$$

for $\boldsymbol{\xi} \in \mathcal{V}(D)$. Under Assumption 2.5, we also define $\boldsymbol{b}_\infty$, $\mathcal{A}_\infty^{(K)}$, and $\mathcal{A}_\infty^{(V)}$ similarly by replacing $\boldsymbol{\Theta}_t(\cdot)$ with $\boldsymbol{\Theta}_\infty(\cdot)$.

Let $\mathcal{W}_L(\Omega)$ denote the space of random functions on Ω. It becomes a Hilbert space once equipped with the inner product

$$\langle \eta_1, \eta_2 \rangle_{\hat{\nu},0} := \mathbb{E}_0 \int_\Omega \eta_1(\boldsymbol{x})\eta_2(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) = \frac{1}{L}\mathbb{E}_0 \sum_{l=1}^{L} \eta_1(\boldsymbol{x}_l)\eta_2(\boldsymbol{x}_l) \,, \tag{68}$$

which gives rise to the norm

$$\|\eta\|_{\hat{\nu},0}^2 := \langle \eta, \eta \rangle_{\hat{\nu},0} = \mathbb{E}_0 \|\eta\|_{\hat{\nu}}^2 \,. \tag{69}$$

With an abuse of notation, we will consider elements in $\mathcal{W}_L(\Omega)$ equivalently as random vectors on $\mathbb{R}^L$. Next, we can define $\mathcal{B}_t$ to be the operator that maps $\eta \in \mathcal{W}_L(\Omega)$ into the vector field

$$(\mathcal{B}_t\eta)(\boldsymbol{\theta}) = \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\eta(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) \tag{70}$$

in $\mathcal{V}(D)$. Its transpose is

$$(\mathcal{B}_t^\intercal\boldsymbol{\xi})(\boldsymbol{x}) = \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\boldsymbol{\xi}(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}), \tag{71}$$

which maps a vector field $\boldsymbol{\xi} \in \mathcal{V}(D)$ back into $\mathcal{W}_L(\Omega)$.

B Long-Time Properties of the Mean-Field Gradient Flow

Proof of Proposition 2.7. The compactness of $\cup_{t\geq 0} \operatorname{supp} \mu_t$ follows from (21) and the compactness of $\operatorname{supp} \mu_0$ assumed in Assumption 2.3. $\mu_t \rightharpoonup \mu_\infty$ follows from (14) and (21).

Under Assumption 2.5, $\boldsymbol{\Theta}_\infty$ is a local minimizer of the energy defined in (19). Consider a local perturbation $\epsilon\boldsymbol{\Theta}_\Delta$ to $\boldsymbol{\Theta}$. The energy value after the perturbation is

$$\begin{aligned}
\mathcal{E}(\boldsymbol{\Theta}_\infty + \epsilon\boldsymbol{\Theta}_\Delta) =& -\int_D F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) + \epsilon\boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}))\mu_0(d\boldsymbol{\theta}) \\
&+ \frac{1}{2}\int_D\int_D K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) + \epsilon\boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}') + \epsilon\boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}'))\mu_0(d\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') \,.
\end{aligned} \tag{72}$$

Under Assumptions 2.2, using Taylor expansion, we have

$$\begin{aligned} F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) + \epsilon \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta})) =& F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \epsilon \nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) \cdot \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \\ &+ \frac{1}{2}\epsilon^2 \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \rangle + O(\epsilon^3) \end{aligned} \tag{73}$$

$$\begin{aligned} & K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) + \epsilon \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}') + \epsilon \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}')) \\ =& K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) + \epsilon \nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \\ &+ \epsilon \nabla' K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') + \frac{1}{2}\epsilon^2 \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \rangle \\ &+ \frac{1}{2}\epsilon^2 \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}'), \nabla'\nabla' K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \\ &+ \epsilon^2 \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla'\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') \rangle + O(\epsilon^3) \; . \end{aligned} \tag{74}$$

Hence, there is

$$\begin{aligned} & \mathcal{E}(\boldsymbol{\Theta}_\infty + \epsilon \boldsymbol{\Theta}_\Delta) - \mathcal{E}(\boldsymbol{\Theta}_\infty) \\ =& \epsilon \int_D \left(-\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \int_D \nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \mu_0(d\boldsymbol{\theta}') \right) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \\ &+ \frac{1}{2}\epsilon^2 \Bigg(\int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \left(\nabla\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \int_D \nabla\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \mu_0(d\boldsymbol{\theta}') \right) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ &\quad + \int_D \int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla'\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \mu_0(d\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}') \Bigg) + O(\epsilon^3) \; . \end{aligned} \tag{75}$$

Since $\boldsymbol{\Theta}_\Delta$ is arbitrary can ϵ can be taken arbitrarily small, we see that for $\boldsymbol{\Theta}_\infty$ to be a local minimizer, the first-order condition is, $\forall \boldsymbol{\theta} \in \operatorname{supp} \mu_0$,

$$-\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \int_D \nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \mu_0(d\boldsymbol{\theta}') = 0 \; , \tag{76}$$

or

$$\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) = 0 \; , \tag{77}$$

and the second-order condition is, $\forall \boldsymbol{\Theta}_\Delta$,

$$\begin{aligned} & \int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \left(\nabla\nabla F(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) + \int_D \nabla\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \mu_0(d\boldsymbol{\theta}') \right) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ &\quad + \int_D \int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla'\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \mu_0(d\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}') \geq 0 \; , \end{aligned} \tag{78}$$

or

$$\begin{aligned} & \int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ &+ \int_D \int_D \langle \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}), \nabla'\nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')) \boldsymbol{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \mu_0(d\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}') \geq 0 \; . \end{aligned} \tag{79}$$

Suppose for contradiction that $\exists D^- \subseteq D$ with $\mu_0(D^-) > 0$ such that $\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$ is not positive semidefinite. Define $\Lambda_\infty(\boldsymbol{\theta})$ to be the least eigenvalue of $\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$. Then there is $\Lambda_\infty(\boldsymbol{\theta}) < 0$ on D^-. In addition, $\exists \zeta > 0$, $\exists D_0^- \subseteq D^-$ with $\mu_0(D_0^-) > 0$ such that $\Lambda_\infty(\boldsymbol{\theta}) < -\zeta$. For $\boldsymbol{\theta} \in D_0^-$, let $\boldsymbol{\Theta}_{\Delta,0}(\boldsymbol{\theta})$ be a normalized eigenvector to $\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$ associated with its least eigenvalue. Moreover, for $n \in \mathbb{N}^*$ that is large enough, we can select any subset $D_n^- \subset D_0^-$ such that $\mu_0(D_n^-) = \frac{1}{n} < \mu_0(D_0^-)$. Then, define

$$\mathbf{\Theta}_{\Delta,n}(\boldsymbol{\theta}) = n^{1/2} \mathbb{1}_{\boldsymbol{\theta} \in D_n^-} \mathbf{\Theta}_{\Delta,0}(\boldsymbol{\theta}) \; , \tag{80}$$

Then, there is

$$\begin{aligned} &\int_D \int_D \langle \mathbf{\Theta}_\Delta(\boldsymbol{\theta}), \nabla' \nabla K(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mathbf{\Theta}_\infty(\boldsymbol{\theta}')) \mathbf{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \mu_0(d\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}') \\ =& \int_\Omega \left| \int_D \nabla \varphi(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \mathbf{\Theta}_{\Delta,n} \mu_0(d\boldsymbol{\theta}) \right|^2 \hat{\nu}(d\boldsymbol{x}) \\ =& \int_\Omega \left| n^{1/2} \int_{D_n^-} \nabla \varphi(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \mathbf{\Theta}_{\Delta,0} \mu_0(d\boldsymbol{\theta}) \right|^2 \hat{\nu}(d\boldsymbol{x}) \\ \leq& C_{\nabla\varphi}^2 n^{-1} \; . \end{aligned} \tag{81}$$

On the other hand

$$\begin{aligned} &\int_D \langle \mathbf{\Theta}_{\Delta,n}(\boldsymbol{\theta}), \nabla\nabla V(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) \mathbf{\Theta}_{\Delta,n}(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ =& \int_{D_n^-} n^{-1} \langle \mathbf{\Theta}_{\Delta,0}(\boldsymbol{\theta}), \nabla\nabla V(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) \mathbf{\Theta}_{\Delta,0}(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ \leq& -\zeta \; . \end{aligned} \tag{82}$$

Therefore, for n large enough, we will have

$$\begin{aligned} &\int_D \int_D \langle \mathbf{\Theta}_\Delta(\boldsymbol{\theta}), \nabla' \nabla K(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mathbf{\Theta}_\infty(\boldsymbol{\theta}')) \mathbf{\Theta}_\Delta(\boldsymbol{\theta}') \rangle \mu_0(d\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}') \\ &+ \int_D \langle \mathbf{\Theta}_{\Delta,n}(\boldsymbol{\theta}), \nabla\nabla V(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty) \mathbf{\Theta}_{\Delta,n}(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) < 0 \; , \end{aligned} \tag{83}$$

which contradicts (79). Hence, we can conclude that μ_0-almost surely, $\nabla\nabla V(\mathbf{\Theta}_\infty(\boldsymbol{\theta}), \mu_\infty)$ is positive semidefinite. □

C Derivations of the Dynamical Central Limit Theorem

C.1 Proof of Proposition 3.1 (Dynamical CLT - I)

The following derivation is an adaptation of the approach in [10] for Vlasov interacting particle systems to our scenario. To start, $\mathbf{\Theta}_t$ and $\mathbf{\Theta}_t^{(n)}$ are governed by the following equations, respectively:

$$\begin{aligned} \dot{\mathbf{\Theta}}_t(\boldsymbol{\theta}) &= -\nabla V(\mathbf{\Theta}_t(\boldsymbol{\theta}), \mu_t), & \mathbf{\Theta}_0(\boldsymbol{\theta}) &= \boldsymbol{\theta} \\ \dot{\mathbf{\Theta}}_t^{(n)}(\boldsymbol{\theta}) &= -\nabla V(\mathbf{\Theta}_t^{(n)}(\boldsymbol{\theta}), \mu_t^{(n)}), & \mathbf{\Theta}_0^{(n)}(\boldsymbol{\theta}) &= \boldsymbol{\theta} \end{aligned} \tag{84}$$

Taking the difference between the two equations in (84) and using the mean value theorem, we get

$$
\begin{aligned}
&\dot{\boldsymbol{T}}_t^{(n)}(\boldsymbol{\theta}) \\
=&n^{1/2}\Big(\dot{\boldsymbol{\Theta}}_t^{(n)}(\boldsymbol{\theta}) - \dot{\boldsymbol{\Theta}}_t(\boldsymbol{\theta})\Big) \\
=&-n^{1/2}\Big(\nabla V(\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}), \mu_t^{(n)}) - \nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\Big) \\
=&-n^{1/2}\Big(\nabla V(\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}), \mu_t) - \nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\Big) - n^{1/2}\Big(\nabla V(\boldsymbol{\Theta}_t, \mu_t^{(n)}) - \nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\Big) \\
&\quad - n^{1/2}\Big[\Big(\nabla V(\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}), \mu_t^{(n)}) - \nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t^{(n)})\Big) - \Big(\nabla V(\boldsymbol{\Theta}_t^{(n)}, \mu_t) - \nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\Big)\Big] \\
=&-\nabla\nabla V(\tilde{\boldsymbol{\Theta}}_{t,1}^{(n)}(\boldsymbol{\theta}), \mu_t)\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}) - \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\theta}')\omega_t^{(n)}(d\boldsymbol{\theta}') \\
&\quad - n^{-1/2}\Big(\int_D \nabla\nabla K(\tilde{\boldsymbol{\Theta}}_{t,2}^{(n)}(\boldsymbol{\theta}), \boldsymbol{\theta}')\omega_t^{(n)}(d\boldsymbol{\theta}')\Big)\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}) \, ,
\end{aligned}
\tag{85}
$$

where $\tilde{\boldsymbol{\Theta}}_{t.1}^{(n)}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\Theta}}_{t,2}^{(n)}(\boldsymbol{\theta})$ denote points that lie on the line segment between $\boldsymbol{\Theta}_t(\boldsymbol{\theta})$ and $\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta})$. Using (31), we can substitute $\omega_t^{(n)}$ in the second term at the right hand side, for which we get

$$
\begin{aligned}
\int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\theta}')\omega_t^{(n)}(d\boldsymbol{\theta}') =& \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0^{(n)}(d\boldsymbol{\theta}') \\
&+ \int_D \nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \tilde{\boldsymbol{\Theta}}_{t,3}^{(n)}(\boldsymbol{\theta}'))\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') \\
&+ n^{-1/2}\int_D \nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \tilde{\boldsymbol{\Theta}}_{t,3}^{(n)}(\boldsymbol{\theta}'))\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}')\omega_0^{(n)}(d\boldsymbol{\theta}') \, .
\end{aligned}
\tag{86}
$$

Therefore, under Assumption 2.2, we have

$$
\begin{aligned}
\dot{\boldsymbol{T}}_t^{(n)}(\boldsymbol{\theta}) =& -\nabla\nabla V(\tilde{\boldsymbol{\Theta}}_{t,1}^{(n)}, \mu_t)\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}) \\
&- \int_D \nabla'\nabla K(\boldsymbol{\Theta}_{t,3}(\boldsymbol{\theta}), \tilde{\boldsymbol{\Theta}}_t^{(n)}(\boldsymbol{\theta}'))\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') \\
&- \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0^{(n)}(d\boldsymbol{\theta}') + O(n^{-1/2}) \, .
\end{aligned}
\tag{87}
$$

Now, we consider the limit as $n \to \infty$. By the standard CLT, we have that $\omega_0^{(n)}(d\boldsymbol{\theta}) \rightharpoonup \omega_0(d\boldsymbol{\theta})$ weakly with respect to $\mathbb{P}_0$, where $\omega_0(d\boldsymbol{\theta})$ is the Gaussian measure with mean zero and covariance defined in (33). On the other hand, by finite-time LLN, we have $\boldsymbol{\Theta}_t^{(n)}(\boldsymbol{\theta}) \to \boldsymbol{\Theta}_t(\boldsymbol{\theta})$ pointwise, $\mathbb{P}_0$-almost surely, and as a consequence $\tilde{\boldsymbol{\Theta}}_{t,1}^{(n)}(\boldsymbol{\theta}), \bar{\boldsymbol{\Theta}}_{t,3}^{(n)}(\boldsymbol{\theta}) \to \boldsymbol{\Theta}_t(\boldsymbol{\theta})$ as well. Therefore, $\boldsymbol{T}_t^{(n)}(\boldsymbol{\theta}) \to \boldsymbol{T}_t(\boldsymbol{\theta})$ pointwise, $\mathbb{P}_0$-almost surely, where the limiting $\boldsymbol{T}_t(\boldsymbol{\theta})$ solves the equation obtained by taking the limit $n \to \infty$ on both sides of (87), which becomes (34). (34) should be solved with initial condition $\boldsymbol{T}_0(\boldsymbol{\theta}) = 0$ since $\boldsymbol{T}_0^{(n)}(\boldsymbol{\theta}) = n^{1/2}(\boldsymbol{\Theta}_0^{(n)}(\boldsymbol{\theta}) - \boldsymbol{\Theta}_0(\boldsymbol{\theta})) = 0$.

Finally, taking the limit $n \to \infty$ on both sides of the equation (31), we deduce that $\omega_t^{(n)}(d\boldsymbol{\theta}) \rightharpoonup \omega_t(d\boldsymbol{\theta})$ weakly, in law with respect to $\mathbb{P}_0$, where the limiting $\omega_t(d\boldsymbol{\theta})$ satisfies

$$
\int_D \chi(\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\omega_0(d\boldsymbol{\theta}) + \int_D \nabla\chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta})) \cdot \boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \, .
\tag{88}
$$

This ends the proof of Proposition 3.1. □

C.2 Proof of Proposition 3.3 (Dynamical CLT - II)

Recall from (34) that

$$
\begin{aligned}
\dot{\boldsymbol{T}}_t(\boldsymbol{\theta}) &= -\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{T}_t(\boldsymbol{\theta}) - \int_D \nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}') \\
&\quad - \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta}') \\
&= -\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{T}_t(\boldsymbol{\theta}) - \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}') \,.
\end{aligned} \tag{89}
$$

Since $\boldsymbol{T}_0(\boldsymbol{\theta}) = 0$, we can use Duhamel's principle to deduce that

$$
\begin{aligned}
\boldsymbol{T}_t(\boldsymbol{\theta}) &= -\int_0^t J_{t,s}(\boldsymbol{\theta}) \int_D \nabla K(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{\theta}')\omega_s(d\boldsymbol{\theta}')ds \\
&= -\int_0^t \int_\Omega J_{t,s}(\boldsymbol{\theta})\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}) \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x})\omega_s(d\boldsymbol{\theta}')\hat{\nu}(d\boldsymbol{x})ds \\
&= -\int_0^t \int_\Omega J_{t,s}(\boldsymbol{\theta})\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x})g_s(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x})ds,
\end{aligned} \tag{90}
$$

where the tensor $J_{t,s}(\boldsymbol{\theta})$ is the Jacobian defined in Proposition 3.3. As a result

$$
\begin{aligned}
g_t(\boldsymbol{x}) &= \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x})\omega_t(d\boldsymbol{\theta}) \\
&= \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\omega_0(d\boldsymbol{\theta}) + \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \\
&= \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\omega_0(d\boldsymbol{\theta}) \\
&\quad - \int_D \int_0^t \int_\Omega \langle \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}), J_{t,s}(\boldsymbol{\theta})\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}')\rangle g_s(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')ds\mu_0(d\boldsymbol{\theta}) \\
&= \bar{g}_t(\boldsymbol{x}) - \int_0^t \int_\Omega \int_D \langle \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}), J_{t,s}(\boldsymbol{\theta})\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}')\rangle \mu_0(d\boldsymbol{\theta})g_s(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')ds \\
&= \bar{g}_t(\boldsymbol{x}) - \int_0^t \int_\Omega \Gamma_{t,s}(\boldsymbol{x}, \boldsymbol{x}')g_s(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')ds,
\end{aligned} \tag{91}
$$

with $\bar{g}_t(\boldsymbol{x})$ and $\Gamma_{t,s}(\boldsymbol{x}, \boldsymbol{x}')$ defined in (38) and (40), respectively. This is (39). □

D Long-Time Behavior of the Fluctuations

D.1 Proof of Theorem 3.4 ($\mu_0 = \mu_\infty$ case)

With the argument outlined in Section 3.2, what remains to be shown is that Γ^∞_{t-s} is positive-semidefinite as a Volterra kernel, according to the definition in [33]. We will utilize the following known result:

Proposition D.1 (Gripenberg et al. [33])**.** *Let $k : [0, \infty) \to \mathbb{R}^{L\times L}$ be a convolution-type kernel for a linear Volterra equation in $\mathbb{R}^L$. If $\forall \eta \in \mathbb{R}^L$, the function $t \mapsto \langle \eta, k(t)\eta\rangle$ is a nonnegative, nonincreasing and convex function on $(0, \infty)$, then k is nonnegative, meaning that $\forall \phi : [0, \infty) \to \mathbb{R}^L$ with compact support, there is*

$$
\int_0^\infty \int_0^t \langle \phi(t), k(t-s)\phi(s)\rangle ds dt \geq 0 \,. \tag{92}
$$

Thus, to take advantage of this proposition, we need to verify that $\forall \eta \in \mathbb{R}^L$, $\langle \eta, \Gamma_t^\infty \eta \rangle$ is

(1) nonnegative:

$$\begin{aligned} &\langle \eta, \Gamma_t^\infty \eta \rangle \\ &= \int_{\Omega\times\Omega}\int_D \big\langle \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}), e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big\rangle \eta(\boldsymbol{x})\eta(\boldsymbol{x}')\mu_0(d\boldsymbol{\theta})\hat{\nu}(d\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}') \\ &= \int_D \Big\langle \boldsymbol{b}(\boldsymbol{\theta}), e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\boldsymbol{b}(\boldsymbol{\theta})\Big\rangle \mu_0(d\boldsymbol{\theta}) \geq 0\,, \end{aligned} \tag{93}$$

where

$$\boldsymbol{b}(\boldsymbol{\theta}) = \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\eta(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) \tag{94}$$

because by assumption, $\forall \boldsymbol{\theta} \in D$, $\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))$ is positive semidefinite, and hence $e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}$ is a positive semidefinite operator;

(2) nonincreasing: Taking derivative with respect to time,

$$\frac{d}{dt}\langle \eta, \Gamma_t^\infty \eta \rangle = -\int_D \Big\langle \boldsymbol{b}(\boldsymbol{\theta}), \nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\boldsymbol{b}(\boldsymbol{\theta})\Big\rangle \mu_0(d\boldsymbol{\theta}) \leq 0, \tag{95}$$

because again, $\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))$ is positive semidefinite;

(3) convex: Taking one more derivative with respect to time,

$$\frac{d^2}{dt^2}\langle \eta, \Gamma_t^\infty \eta \rangle = \int_D \Big\langle \boldsymbol{b}(\boldsymbol{\theta}), (\nabla\nabla V(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})))^2 e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\boldsymbol{b}(\boldsymbol{\theta})\Big\rangle \mu_0(d\boldsymbol{\theta}) \geq 0, \tag{96}$$

Therefore, we can apply Proposition D.1 to conclude that Γ_{t-s}^∞ is PSD as a Volterra kernel, and so $\int_{t_0}^T \int_{t_0}^t \langle g_t, \Gamma_{t-s}^\infty g_s \rangle ds dt \geq 0$.

D.2 Proof of Theorem 3.5 (Unregularized case)

Recall that

$$\begin{aligned} \lim_{n\to\infty} n\mathbb{E}_0 \|f_t^{(n)} - f_t\|_{\hat{\nu}}^2 = \mathbb{E}_0\|g_t\|_{\hat{\nu}}^2 &= \mathbb{E}_0 \int_\Omega \Big| \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x})\omega_t(d\boldsymbol{\theta})\Big|^2 \hat{\nu}(d\boldsymbol{x}) \\ &= \mathbb{E}_0 \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}')\,, \end{aligned} \tag{97}$$

where, with a slight abuse of notation, in this equation $\mathbb{E}_0$ also denotes expectation over the randomness of the Gaussian distribution ω_0 defined in Proposition 3.1. From (32) in Proposition 3.1, this can be further expanded into

$$\begin{aligned} &\mathbb{E}_0 \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}') \\ &= \mathbb{E}_0 \int_{D\times D} \langle \boldsymbol{T}_t(\boldsymbol{\theta}), \nabla\nabla' K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta}')\rangle \mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}') \\ &+ 2\mathbb{E}_0 \int_{D\times D} \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}') \\ &+ \mathbb{E}_0 \int_{D\times D} K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}')\,. \end{aligned} \tag{98}$$

The last term at the RHS is equal to $\mathbb{E}_0\|\bar g_t\|^2_{\hat\nu}$ with $\bar g_t$ defined in (38). Using (33), it can be explicitly computed as

$$
\begin{aligned}
\mathbb{E}_0\|\bar g_t\|^2_{\hat\nu} =&\mathbb{E}_0\int_{D\times D} K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}')\\
&=\int_{D\times D} K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\left(\mu_0(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}')-\mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}')\right)\\
&=\int_D K(\boldsymbol{\theta},\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})-\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}')\\
&=\int_D K(\boldsymbol{\theta},\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})-\|f_t\|^2_{\hat\nu}\,.
\end{aligned}
\tag{99}
$$

Thus,

$$
\begin{aligned}
\lim_{t\to\infty}\mathbb{E}_0\|\bar g_t\|^2_{\hat\nu}=\lim_{t\to\infty}&\int_D K(\boldsymbol{\theta},\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})-\|f_t\|^2_{\hat\nu}\\
&=\int_D K(\boldsymbol{\theta},\boldsymbol{\theta})\mu_\infty(d\boldsymbol{\theta})-\|f_\infty\|^2_{\hat\nu}\\
&=\mathbb{E}_0\|\bar g_\infty\|^2_{\hat\nu}
\end{aligned}
\tag{100}
$$

and so

$$
\lim_{T\to\infty}⨍_0^T \mathbb{E}_0\|\bar g_t\|^2_{\hat\nu}dt=\mathbb{E}_0\|\bar g_\infty\|^2_{\hat\nu}\,,
\tag{101}
$$

where here and below we denote $⨍_0^t[\cdot]\,dt=\frac{1}{t}\int_0^t[\cdot]\,dt$. As a result, to prove (47) or (50) in Theorem 3.5, it suffices to establish that

$$
\lim_{T\to\infty}⨍_0^T \mathfrak{D}_t dt\le 0\,,
\tag{102}
$$

or

$$
\lim_{T\to\infty}⨍_0^T \mathfrak{D}_t dt\le -\mathbb{E}_0\|\bar g_\infty\|^2_{\hat\nu}\,,
\tag{103}
$$

respectively, where we defined

$$
\begin{aligned}
\mathfrak{D}_t :=& \mathbb{E}_0\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}')-\mathbb{E}_0\int_{D\times D} K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}')\\
=&\mathbb{E}_0\int_{D\times D}\langle \boldsymbol{T}_t(\boldsymbol{\theta}),\nabla\nabla' K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta}')\rangle\mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}')\\
&+2\mathbb{E}_0\int_{D\times D}\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\omega_0(d\boldsymbol{\theta}')\,.
\end{aligned}
\tag{104}
$$

To this end, we examine (34) as an infinite-dimensional ODE. With the Hilbert space $\mathcal{V}(D)$ defined in Appendix A and $\boldsymbol{b}_t$, $\mathcal{A}^{(K)}_t$ and $\mathcal{A}^{(V)}_t$ defined by (64), (66) and (67), respectively, we can rewrite (34) as the following ODE on $\mathcal{V}(D)$:

$$
\dot{\boldsymbol{T}}_t=-(\mathcal{A}^{(K)}_t+\mathcal{A}^{(V)}_t)\boldsymbol{T}_t-\boldsymbol{b}_t,
\tag{105}
$$

We can also rewrite (104) as

$$
\mathfrak{D}_t=\langle \boldsymbol{T}_t,\mathcal{A}^{(K)}_t\boldsymbol{T}_t\rangle_0+2\langle \boldsymbol{T}_t,\boldsymbol{b}_t\rangle_0\,.
\tag{106}
$$

From (105), we can deduce that

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{T}_t\|_0^2 = -\langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0 - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0 - \langle \boldsymbol{T}_t, \boldsymbol{b}_t\rangle_0, \tag{107}$$

or equivalently

$$\langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0 + \langle \boldsymbol{T}_t, \boldsymbol{b}_t\rangle_0 = -\frac{1}{2}\frac{d}{dt}\|\boldsymbol{T}_t\|_0^2 - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0\,. \tag{108}$$

Therefore, we can rewrite (104) as

$$\begin{aligned}\mathfrak{D}_t =& 2\left(\langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0 + \langle \boldsymbol{T}_t, \boldsymbol{b}_t\rangle_0\right) - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0\\ =& 2\left(-\frac{1}{2}\frac{d}{dt}\|\boldsymbol{T}_t\|_0^2 - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0\right) - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0\\ =& -\frac{d}{dt}\|\boldsymbol{T}_t\|_0^2 - 2\langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0 - \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0\end{aligned} \tag{109}$$

and as a result, since $\boldsymbol{T}_0 = 0$,

$$⨍_0^T \mathfrak{D}_t dt = -\frac{1}{T}\|\boldsymbol{T}_T\|_0^2 - 2⨍_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0 dt - ⨍_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0 dt\,. \tag{110}$$

Note that for all t, $\mathcal{A}_t^{(K)}$ is a positive semidefinite (PSD) operator on $\mathcal{V}(D)$, as $\forall \boldsymbol{\xi} \in \mathcal{V}(D)$,

$$\begin{aligned}\langle \mathcal{A}_t^{(K)}\boldsymbol{\xi}, \boldsymbol{\xi}\rangle_0 &= \mathbb{E}_0 \int_{D\times D} \langle \boldsymbol{\xi}(\boldsymbol{\theta}), \nabla\nabla' K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{\xi}(\boldsymbol{\theta}')\rangle \mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}')\\ &= \mathbb{E}_0 \int_\Omega \left|\int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\cdot\boldsymbol{\xi}(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\right|^2 \hat{\nu}(d\boldsymbol{x}) \geq 0\,.\end{aligned} \tag{111}$$

This implies that $⨍_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)}\boldsymbol{T}_t\rangle_0 dt \geq 0$. Hence, to establish (102), it is sufficient to show that

$$\lim_{T\to\infty} ⨍_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)}\boldsymbol{T}_t\rangle_0 dt = 0\,. \tag{112}$$

To this end, we need two lemmas that are proved below in Appendices D.2.1 and D.2.2, respectively:

Lemma D.2. *Assuming* (48) *and* (49) *together with Assumptions* 2.2, 2.3 *and* 2.5, *we have*

$$\int_0^\infty \|\mathcal{A}_t^{(V)}\|_0 dt < \infty \tag{113}$$

$$\int_0^\infty \|\mathcal{A}_\infty^{(K)} - \mathcal{A}_t^{(K)}\|_0 dt < \infty \tag{114}$$

$$\int_0^\infty \|\boldsymbol{b}_t - \boldsymbol{b}_\infty\|_0 dt < \infty \tag{115}$$

Lemma D.3. *Assuming* (48) *and* (49) *together with Assumptions* 2.2, 2.3 *and* 2.5, *we have*

$$\sup_{t<\infty} \|\boldsymbol{T}_t\|_0^2 < \infty\,. \tag{116}$$

With these two lemmas, we can show that

$$\left|\int_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)} \boldsymbol{T}_t\rangle_0 dt\right| \le \int_0^T \|\mathcal{A}_t^{(V)}\|_0 \|\boldsymbol{T}_t\|_0^2 dt \\ \le \left(\int_0^T \|\mathcal{A}_t^{(V)}\|_0 dt\right) \sup_{t<\infty} \|\boldsymbol{T}_t\|_0^2 \\ <\infty\ , \tag{117}$$

and so (112) is satisfied. This finishes the proof of (47) under (48) and (49) together with Assumptions 2.2, 2.3 and 2.5.

Next, we show (50) under the additional condition of Assumption 2.1. Thanks to (110) and (112), it is sufficient to establish that

$$\lim_{T\to\infty} \fint_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)} \boldsymbol{T}_t\rangle_0 dt = \mathbb{E}_0 \|\bar{g}_\infty\|_{\hat{\nu}}^2\ . \tag{118}$$

Heuristically, if $\boldsymbol{T}_\infty := \lim_{t\to\infty} \boldsymbol{T}_t$ exists, then from (105), it has to satisfy

$$-\boldsymbol{b}_\infty = \left(\mathcal{A}_\infty^{(V)} + \mathcal{A}_\infty^{(K)}\right) \boldsymbol{T}_\infty = \mathcal{A}_\infty^{(K)} \boldsymbol{T}_\infty\ , \tag{119}$$

as $\mathcal{A}_\infty^{(V)} = 0$ (because $\nabla\nabla V(\boldsymbol{\theta}, \mu_\infty) = \int_\Omega \varphi(\boldsymbol{\theta}, \boldsymbol{x})(f_\infty(\boldsymbol{x}) - f_*(\boldsymbol{x}))\hat{\nu}(d\boldsymbol{x}) = 0$ under the assumption of (48)). This equation implies that

$$(\boldsymbol{T}_\infty)^{||} = -\left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\ , \tag{120}$$

where $(\boldsymbol{T}_\infty)^{||}$ denotes the component of $\boldsymbol{T}_\infty$ in the range of $\mathcal{A}_\infty^{(K)}$, and $\left(\mathcal{A}_\infty^{(K)}\right)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathcal{A}_\infty^{(K)}$. As a result,

$$\langle \boldsymbol{T}_\infty, \mathcal{A}_\infty^{(K)} \boldsymbol{T}_\infty\rangle_0 = \langle (\boldsymbol{T}_\infty)^{||}, \mathcal{A}_\infty^{(K)} (\boldsymbol{T}_\infty)^{||}\rangle_0 \\ = \langle -\left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty, -\mathcal{A}_\infty^{(K)} \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\rangle_0 \\ = \langle \boldsymbol{b}_\infty, \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\rangle_0\ . \tag{121}$$

Rigorously, without assuming the existence of $\boldsymbol{T}_\infty$, we can establish that

Lemma D.4. *Assuming* (48) *and* (49) *together with Assumptions* 2.2, 2.3 *and* 2.5, *we have*

$$\lim_{t\to\infty} \fint_0^t \langle \boldsymbol{T}_s, \mathcal{A}_s^{(K)} \boldsymbol{T}_s\rangle_0 ds \ge \langle \boldsymbol{b}_\infty, \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\rangle_0\ . \tag{122}$$

As a consequence,

$$\lim_{t\to\infty} \fint_0^t \mathbb{E}_0 \|g_s\|_{\hat{\nu}}^2 dt \le \mathbb{E}_0 \|\bar{g}_\infty\|_{\hat{\nu}}^2 - \langle \boldsymbol{b}_\infty, \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\rangle_0\ . \tag{123}$$

This lemma is proved in D.2.3. It implies that we only need to show that

$$\langle \boldsymbol{b}_\infty, \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty\rangle_0 = \mathbb{E}_0 \|\bar{g}_\infty\|_{\hat{\nu}}^2\ . \tag{124}$$

This requires us to further exploit the relationship among $\mathcal{A}_\infty^{(K)}$, $\boldsymbol{b}_\infty$ and $\bar{g}_\infty$. With the Hilbert space $\mathcal{W}_L(\Omega)$ defined in Appendix A and $\mathcal{B}_t$ defined by (70), we can rewrite (66) as

$$\mathcal{A}_t^{(K)} = \mathcal{B}_t \mathcal{B}_t^\intercal \ . \tag{125}$$

Further, recall that

$$g_t = \int_D \varphi(\boldsymbol{\theta}, \cdot)\omega_t(d\boldsymbol{\theta}) = \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot)\omega_0(d\boldsymbol{\theta}) + \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot) \cdot \boldsymbol{T}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \tag{126}$$

$$\bar{g}_t = \int_D \varphi(\boldsymbol{\theta}, \cdot)\bar{\omega}_t(d\boldsymbol{\theta}) = \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \cdot)\omega_0(d\boldsymbol{\theta}) \ . \tag{127}$$

Therefore, we can write

$$g_t = \bar{g}_t + \mathcal{B}_t^\intercal \boldsymbol{T}_t \ , \tag{128}$$

and

$$\boldsymbol{b}_t = \mathcal{B}_t \bar{g}_t, \tag{129}$$

Similar formulas hold when we replace t by ∞. With these relations, we see that

$$\begin{aligned} \left\langle \boldsymbol{b}_\infty, \left(\mathcal{A}_\infty^{(K)}\right)^\dagger \boldsymbol{b}_\infty \right\rangle_0 &= \left\langle \mathcal{B}_\infty \bar{g}_\infty, \left(\mathcal{B}_\infty \mathcal{B}_\infty^\intercal\right)^\dagger \mathcal{B}_\infty \bar{g}_\infty \right\rangle_0 \\ &= \mathbb{E}_0 \| \left(\mathcal{B}_\infty\right)^\dagger \mathcal{B}_\infty \bar{g}_\infty \|_{\hat{\nu}}^2 \ , \end{aligned} \tag{130}$$

because $(\mathcal{B}_\infty \mathcal{B}_\infty^\intercal)^\dagger = (\mathcal{B}_\infty^\intercal)^\dagger (\mathcal{B}_\infty)^\dagger$. Since $(\mathcal{B}_\infty)^\dagger \mathcal{B}_\infty$ is the projection operator (matrix) onto the range of $\mathcal{B}_\infty^\intercal$ in $\mathbb{R}^L$, it is then sufficient to prove that

Lemma D.5. *Under Assumptions 2.1, 2.2, 2.3 and 2.5, $\mathbb{P}_0$-almost surely, $\bar{g}_\infty \in \mathrm{Ran}(\mathcal{B}_\infty^\intercal)$.*

Lemma D.5 is proven in Appendix D.2.4 and it concludes the proof of (50) in Theorem 3.5.

To show that $\|g_t\|_{\hat{\nu}}$ decreases monotonically when $\mu_0 = \mu_\infty$, note that in this case $\mu_t = \mu_\infty$, $\forall t \geq 0$, and so $\mathcal{A}_t^{(V)} = \mathcal{A}_\infty^{(V)} = 0$, $\mathcal{A}_t^{(K)} = \mathcal{A}_\infty^{(K)}$ and $\boldsymbol{b}_t = \boldsymbol{b}_\infty$, $\forall t \geq 0$. Thus, (105) becomes

$$\dot{\boldsymbol{T}}_t = -\mathcal{A}_\infty^{(K)} \boldsymbol{T}_t - \boldsymbol{b}_\infty, \tag{131}$$

As will be shown in Lemma D.6, $\boldsymbol{b}_\infty$ is in the range of $\mathcal{A}_\infty^{(K)}$. Therefore, defining

$$\boldsymbol{u}_\infty = (\mathcal{A}_\infty^{(K)})^\dagger \boldsymbol{b}_\infty \ , \tag{132}$$

and

$$\boldsymbol{z}_t = \boldsymbol{T}_t + \boldsymbol{u}_\infty \ , \tag{133}$$

there is

$$\dot{\boldsymbol{z}}_t = -\mathcal{A}_\infty^{(K)} \boldsymbol{z}_t \ , \tag{134}$$

whose solution can be written analytically as

$$\boldsymbol{z}_t = e^{-\mathcal{A}_\infty^{(K)} t} \boldsymbol{z}_0 = e^{-\mathcal{A}_\infty^{(K)} t} \boldsymbol{u}_\infty \ . \tag{135}$$

Thus,

$$\boldsymbol{T}_t = \boldsymbol{z}_t - \boldsymbol{u}_\infty = -(I - e^{-\mathcal{A}_\infty^{(K)} t}) \boldsymbol{u}_\infty \tag{136}$$

Therefore, as $\boldsymbol{b}_\infty = \mathcal{B}_\infty \bar{g}_\infty$, there is

$$\begin{aligned} g_t =& \bar{g}_\infty + \mathcal{B}_\infty^\intercal \boldsymbol{T}_t \\ =& \bar{g}_\infty - \mathcal{B}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) \boldsymbol{u}_\infty \\ =& \bar{g}_\infty - \mathcal{B}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) (\mathcal{A}_\infty^{(K)})^\dagger \mathcal{B}_\infty \bar{g}_\infty \ . \end{aligned} \tag{137}$$

Hence,

$$|g_\infty|^2 = |\bar{g}_\infty|^2 - 2(*) + (**) \ , \tag{138}$$

where

$$\begin{aligned} (*) =& (\mathcal{B}_\infty \bar{g}_\infty)^\intercal \, (I - e^{-\mathcal{A}_\infty^{(K)} t}) (\mathcal{A}_\infty^{(K)})^\dagger \mathcal{B}_\infty \bar{g}_\infty \\ =& \boldsymbol{b}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) (\mathcal{A}_\infty^{(K)})^\dagger \boldsymbol{b}_\infty \end{aligned} \tag{139}$$

and

$$\begin{aligned} (**) =& (\mathcal{B}_\infty \bar{g}_\infty)^\intercal (\mathcal{A}_\infty^{(K)})^\dagger (I - e^{-\mathcal{A}_\infty^{(K)} t}) \mathcal{B}_\infty \mathcal{B}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) (\mathcal{A}_\infty^{(K)})^\dagger \mathcal{B}_\infty \bar{g}_\infty \\ =& \boldsymbol{b}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) \mathcal{B}_\infty \mathcal{B}_\infty^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) \boldsymbol{b}_\infty \ . \end{aligned} \tag{140}$$

In the ERM setting, $\mathcal{A}_\infty^{(K)}$ is PSD with a finite number of nonzero eigenspaces. Consider a set of orthonormal eigenfunctions that span those nonzero eigenspaces, $v_1, ..., v_k$, corresponding to eigenvalues $\lambda_1, ..., \lambda_k > 0$, respectively. As $\boldsymbol{b}_\infty$ is in the range of $\mathcal{A}_\infty^{(K)}$ by Lemma D.6, we can decompose it as

$$\boldsymbol{b}_\infty = \sum_{i=1}^k c_i v_i \tag{141}$$

for some real numbers c_i's. Thus, we can write

$$\begin{aligned} (*) =& \left(\sum_{i=1}^k c_i v_i \right)^\intercal (I - e^{-\mathcal{A}_\infty^{(K)} t}) (\mathcal{A}_\infty^{(K)})^\dagger \left(\sum_{j=1}^k c_j v_j \right) \\ =& \left(\sum_{i=1}^k c_i v_i \right)^\intercal \left(\sum_{j=1}^k c_j \lambda_j^{-1} (1 - e^{-\lambda_j t}) v_j \right) \\ =& \sum_{i=1}^k \lambda_j^{-1} (1 - e^{-\lambda_j t}) c_i^2 \ , \end{aligned} \tag{142}$$

$$\begin{aligned} (**) =& \left(\sum_{i=1}^k c_i v_i \right)^\intercal (\mathcal{A}_\infty^{(K)})^\dagger (I - e^{-\mathcal{A}_\infty^{(K)} t}) \mathcal{B}_\infty \mathcal{B}_\infty^\intercal (I - e^{-\mathcal{A}_\infty t}) (\mathcal{A}_\infty^{(K)})^\dagger \left(\sum_{j=1}^k c_j v_j \right) \\ =& \left(\sum_{i=1}^k c_i v_i \right)^\intercal (\mathcal{A}_\infty^{(K)})^\dagger (I - e^{-\mathcal{A}_\infty^{(K)} t}) \mathcal{A}_\infty^{(K)} (I - e^{-\mathcal{A}_\infty t}) (\mathcal{A}_\infty^{(K)})^\dagger \left(\sum_{j=1}^k c_j v_j \right) \\ =& \left(\sum_{i=1}^k c_i v_i \right)^\intercal \left(\sum_{j=1}^k \lambda_j^{-1} \left(1 - e^{-\lambda_j t}\right)^2 c_j v_j \right) \\ =& \sum_{i=1}^k \lambda_j^{-1} \left(1 - e^{-\lambda_j t}\right)^2 c_i^2 \ . \end{aligned} \tag{143}$$

Therefore,

$$\begin{aligned}
|g_\infty|^2 =&|\bar{g}_\infty|^2 - 2\sum_{i=1}^{k} \lambda_j^{-1}\left(1-e^{-\lambda_j t}\right) c_i^2 + \sum_{i=1}^{k} \lambda_j^{-1}\left(1-e^{-\lambda_j t}\right)^2 c_i^2 \\
=&|\bar{g}_\infty|^2 + \sum_{i=1}^{k} \lambda_j^{-1}\left(1-e^{-\lambda_j t}\right)\left(-1-e^{-\lambda_j t}\right) c_i^2 \\
=&|\bar{g}_\infty|^2 - \sum_{i=1}^{k} \lambda_j^{-1}\left(1-e^{-2\lambda_j t}\right) c_i^2 \ ,
\end{aligned} \tag{144}$$

which is decreasing in time. This completes the proof of Theorem 3.5. □

D.2.1 Proof of Lemma D.2

Proof of (113): $\int_0^\infty \|\mathcal{A}_t^{(V)}\|_0 dt < \infty$.
By the definition of the operator norm induced by $\|\cdot\|_0$ on $\mathcal{V}(D)$, $\|\mathcal{A}_t^{(V)}\|_0$ is the smallest number C_t such that $\forall \boldsymbol{\xi}$, there is

$$\|\mathcal{A}_t^{(V)}\|_0 = \sup_{\boldsymbol{\xi}\in\mathcal{V}(D), \|\boldsymbol{\xi}\|_0\neq 0} \frac{\left|\langle \boldsymbol{\xi}, \mathcal{A}_t^{(V)}\boldsymbol{\xi}\rangle_0\right|}{\|\boldsymbol{\xi}\|_0^2} \ . \tag{145}$$

In the unregularized case, a straightforward bound of $\left|\langle \boldsymbol{\xi}, \mathcal{A}_t^{(V)}\boldsymbol{\xi}\rangle_0\right|$ is

$$\begin{aligned}
\left|\langle \boldsymbol{\xi}, \mathcal{A}_t^{(V)}\boldsymbol{\xi}\rangle_0\right| =& \left|\mathbb{E}_0 \int_D \langle \boldsymbol{\xi}(\boldsymbol{\theta}), \nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{\xi}(\boldsymbol{\theta})\rangle \mu_0(d\boldsymbol{\theta})\right| \\
=& \left|\mathbb{E}_0 \int_D \int_\Omega \langle \boldsymbol{\xi}(\boldsymbol{\theta}), \nabla\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\boldsymbol{\xi}(\boldsymbol{\theta})\rangle (f_t(\boldsymbol{x}) - f_*(\boldsymbol{x}))\hat{\nu}(d\boldsymbol{x})\, \mu_0(d\boldsymbol{\theta})\right| \\
\leq& \mathbb{E}_0 \int_D \int_\Omega C_{\nabla\nabla\varphi} \left|\boldsymbol{\xi}(\boldsymbol{\theta})\right|^2 \left|f_t(\boldsymbol{x}) - f_*(\boldsymbol{x})\right| \hat{\nu}(d\boldsymbol{x})\mu_0(d\boldsymbol{\theta}) \\
=& C_{\nabla\nabla\varphi}\|\boldsymbol{\xi}\|_0^2 \int_\Omega \left|f_t(\boldsymbol{x}) - f_*(\boldsymbol{x})\right| \hat{\nu}(d\boldsymbol{x}) \\
\leq& L^{1/2} C_{\nabla\nabla\varphi}\|\boldsymbol{\xi}\|_0^2 \|f_t - f_*\|_{\hat{\nu}} \\
=& L^{1/2} C_{\nabla\nabla\varphi}\|\boldsymbol{\xi}\|_0^2 \left(\mathcal{L}(\mu_t)\right)^{1/2} \ .
\end{aligned} \tag{146}$$

Thus, we have

$$\|\mathcal{A}_t^{(V)}\|_0 \leq L^{1/2} C_{\nabla\nabla\varphi}\|\boldsymbol{\xi}\|_0^2 \left(\mathcal{L}(\mu_t)\right)^{1/2} \ . \tag{147}$$

By the assumption (49), we thus have

$$\int_0^\infty \|\mathcal{A}_t^{(V)}\|_0 dt \leq L^{1/2} C_{\nabla\nabla\varphi} \int_0^\infty \left(\mathcal{L}(\mu_t)\right)^{1/2} dt < \infty \tag{148}$$

which gives us the desired bound. □

Proof of (114): $\int_0^\infty \|\mathcal{A}_\infty^{(K)} - \mathcal{A}_t^{(K)}\|_0 dt < \infty$.

We have

$$
\begin{aligned}
&\langle \boldsymbol{\xi}, (\mathcal{A}_t^{(K)} - \mathcal{A}_\infty^{(K)})\boldsymbol{\xi}\rangle_0 \\
=&\mathbb{E}_0 \int_\Omega \Big(\Big(\int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \Big)^2 - \Big(\int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \Big)^2 \Big) \hat{\nu}(d\boldsymbol{x}) \\
=&\mathbb{E}_0 \int_\Omega \Big(\int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) + \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \Big) \\
&\times \Big(\int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \Big) \hat{\nu}(d\boldsymbol{x}) \,.
\end{aligned}
\tag{149}
$$

Hence, the absolute value of the expression above is upper-bounded by

$$
\begin{aligned}
&\mathbb{E}_0 \Big(\int_D |\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) + \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})| |\boldsymbol{\xi}(\boldsymbol{\theta})| \mu_0(d\boldsymbol{\theta}) \\
&\quad \times \int_D |\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})| |\boldsymbol{\xi}(\boldsymbol{\theta})| \mu_0(d\boldsymbol{\theta}) \Big) \\
\leq& 2C_{\nabla\varphi} C_{\nabla\nabla\varphi} \|\boldsymbol{\xi}\|_0^2 \Big(\int_D |\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \Big)^{1/2} \,.
\end{aligned}
\tag{150}
$$

Thus, by the assumption (49), we have

$$
\begin{aligned}
\int_0^\infty \|\mathcal{A}_\infty^{(K)} - \mathcal{A}_t^{(K)}\|_0 dt \leq& 2C_{\nabla\varphi} C_{\nabla\nabla\varphi} \int_0^\infty \left(\int_D |\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \right)^{1/2} dt \\
\leq& 2C_{\nabla\varphi} C_{\nabla\nabla\varphi} \int_0^\infty (\mathcal{L}(\mu_t))^{\frac{1}{2}} \, dt \\
<& \infty
\end{aligned}
\tag{151}
$$

□

Proof of (115): $\int_0^\infty \|\boldsymbol{b}_t - \boldsymbol{b}_\infty\|_0 dt < \infty$.
There is

$$
\begin{aligned}
\boldsymbol{b}_t(\boldsymbol{\theta}) - \boldsymbol{b}_\infty(\boldsymbol{\theta}) =& \int_D \big(\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}')) - \nabla K(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'))\big) \omega_0(d\boldsymbol{\theta}') \\
=& \int_D \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal \hat{\nu}(d\boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \\
=& \int_D \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal - \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal \hat{\nu}(d\boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \\
&+ \int_D \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})^\intercal \omega_0(d\boldsymbol{\theta}') \\
=& \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \left(\int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\big) \omega_0(d\boldsymbol{\theta}') \right)^\intercal \hat{\nu}(d\boldsymbol{x}) \\
&+ \int_\Omega \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big) \left(\int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \right)^\intercal \hat{\nu}(d\boldsymbol{x}) \,.
\end{aligned}
\tag{152}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}_0 \left|\boldsymbol{b}_t(\boldsymbol{\theta}) - \boldsymbol{b}_\infty(\boldsymbol{\theta})\right|^2 \leq & \mathbb{E}_0 \left| \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \left(\int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\big)\, \omega_0(d\boldsymbol{\theta}') \right)^\intercal \hat{\nu}(d\boldsymbol{x}) \right|^2 \\
& + \mathbb{E}_0 \left| \int_\Omega \left(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\right) \left(\int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \right)^\intercal \hat{\nu}(d\boldsymbol{x}) \right|^2 \\
\leq & \int_\Omega \left|\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\right|^2 \mathbb{E}_0 \left| \int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\big)\, \omega_0(d\boldsymbol{\theta}') \right|^2 \hat{\nu}(d\boldsymbol{x}) \\
& + \int_\Omega \left|\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\right|^2 \mathbb{E}_0 \left| \int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \right|^2 \hat{\nu}(d\boldsymbol{x}) \\
\leq & C_{\nabla\varphi}^2 \int_\Omega \mathbb{E}_0 \left| \int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\big)\, \omega_0(d\boldsymbol{\theta}') \right|^2 \hat{\nu}(d\boldsymbol{x}) \\
& + C_{\nabla\nabla\varphi}^2 \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})\right|^2 \int_\Omega \mathbb{E}_0 \left| \int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x}) \omega_0(d\boldsymbol{\theta}') \right|^2 \hat{\nu}(d\boldsymbol{x}) \,.
\end{aligned}
\tag{153}
$$

By the property of ω_0, there is

$$
\begin{aligned}
\mathbb{E}_0 \left| \int_D \chi(\boldsymbol{\theta}) \omega_0(d\boldsymbol{\theta}) \right|^2 &= \int_D \left| \chi(\boldsymbol{\theta}) - \int_D \chi(\boldsymbol{\theta}') \mu_0(d\boldsymbol{\theta}') \right|^2 \mu_0(d\boldsymbol{\theta}) \\
&\leq \int_D \left|\chi(\boldsymbol{\theta})\right|^2 \mu_0(d\boldsymbol{\theta})
\end{aligned}
\tag{154}
$$

for a test function χ on D. Thus,

$$
\begin{aligned}
\mathbb{E}_0 \left|\boldsymbol{b}_t(\boldsymbol{\theta}) - \boldsymbol{b}_\infty(\boldsymbol{\theta})\right|^2 \leq & C_{\nabla\varphi}^2 \int_\Omega \int_D \left|\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\right|^2 \mu_0(d\boldsymbol{\theta}') \\
& + C_{\nabla\nabla\varphi}^2 \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})\right|^2 \int_\Omega \int_D \left|\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\right|^2 \mu_0(d\boldsymbol{\theta}') \hat{\nu}(d\boldsymbol{x}) \\
\leq & C_{\nabla\varphi}^2 C_{\nabla\nabla\varphi}^2 \int_D \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}') - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta}')\right|^2 \mu_0(d\boldsymbol{\theta}') \\
& + C_{\nabla\nabla\varphi}^2 C_{\nabla\varphi}^2 \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})\right|^2 \,.
\end{aligned}
\tag{155}
$$

Therefore,

$$
\begin{aligned}
\|\boldsymbol{b}_t - \boldsymbol{b}_\infty\|_0^2 = & \mathbb{E}_0 \int_D \left|\boldsymbol{b}_t(\boldsymbol{\theta}) - \boldsymbol{b}_\infty(\boldsymbol{\theta})\right|^2 \mu_0(d\boldsymbol{\theta}) \\
\leq & 2 C_{\nabla\varphi}^2 C_{\nabla\nabla\varphi}^2 \int_D \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})\right|^2 \mu_0(d\boldsymbol{\theta}) \,.
\end{aligned}
\tag{156}
$$

Since

$$
\begin{aligned}
\int_D \left|\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})\right|^2 \mu_0(d\boldsymbol{\theta}) \leq & \int_D \int_t^\infty \left|\dot{\boldsymbol{\Theta}}_s(\boldsymbol{\theta})\right|^2 ds \mu_0(d\boldsymbol{\theta}) \\
= & \int_D \int_t^\infty \left|\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\right|^2 ds \mu_0(d\boldsymbol{\theta}) \\
= & -\int_t^\infty \frac{d}{ds} \mathcal{L}(\mu_s) ds \\
= & \mathcal{L}(\mu_t) - \mathcal{L}(\mu_\infty) \\
= & \mathcal{L}(\mu_t)
\end{aligned}
\tag{157}
$$

we can conclude that

$$\int_0^\infty \|\boldsymbol{b}_t - \boldsymbol{b}_\infty\|_0 dt \leq \int_0^\infty |\mathcal{L}(\mu_t)|^{\frac{1}{2}}\, dt < \infty\,. \tag{158}$$

□

D.2.2 Proof of Lemma D.3

Our goal is to show that $\|\boldsymbol{T}_t\|_0$ remains bounded for all time. First note that, for all t, $\mathcal{A}_t^{(K)}$ is a positive semidefinite (PSD) operator on $\mathcal{V}(D)$ since

$$\begin{aligned}\langle \mathcal{A}_t^{(K)}\boldsymbol{\xi}, \boldsymbol{\xi}\rangle_0 &= \mathbb{E}_0 \int_{D\times D} \langle \boldsymbol{\xi}(\boldsymbol{\theta}), \nabla\nabla' K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{\xi}(\boldsymbol{\theta}')\rangle \mu_0(d\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}') \\ &= \mathbb{E}_0 \int_\Omega \Big| \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\cdot\boldsymbol{\xi}(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\Big|^2 \hat{\nu}(d\boldsymbol{x}) \geq 0\,.\end{aligned} \tag{159}$$

Second, by Assumption 2.5, for μ_0-almost-every $\boldsymbol{\theta} \in D$, $\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}) = \lim_{t\to\infty} \boldsymbol{\Theta}_t(\boldsymbol{\theta})$ exists, which allows us to define $\boldsymbol{b}_\infty$, $\mathcal{A}_\infty^{(K)}$, and $\mathcal{A}_\infty^{(V)}$ similarly to (64), (66) and (67) by replacing $\boldsymbol{\Theta}_t(\cdot)$ with $\boldsymbol{\Theta}_\infty(\cdot)$. Since we assume that

$$\forall \boldsymbol{x}_k \in \operatorname{supp}\hat{\nu} \quad : \quad f_\infty(\boldsymbol{x}_k) = \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_k)\mu_\infty(d\boldsymbol{\theta}) = f_*(\boldsymbol{x}_k) \tag{160}$$

we have

$$\forall \boldsymbol{\theta} \in D \quad : \quad \nabla\nabla V(\boldsymbol{\theta}, \mu_\infty) = \int_\Omega \nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})(f_\infty(\boldsymbol{x}) - f_*(\boldsymbol{x}))d\boldsymbol{x} = 0\,. \tag{161}$$

This implies that $\mathcal{A}_\infty^{(V)}$ is the zero operator on $\mathcal{V}(D)$.

Third, we have the following observation:

Lemma D.6. *Under Assumptions 2.2, 2.3 and 2.5, $\boldsymbol{b}_t \in \operatorname{Ran}(\mathcal{A}_t^{(K)})$ for all t, and $\boldsymbol{b}_\infty \in \operatorname{Ran}(\mathcal{A}_\infty^{(K)})$. Specifically, $\exists \tilde{\boldsymbol{u}}_\infty \in \mathcal{V}(D)$ such that $\|\boldsymbol{u}_\infty\|_0 < \infty$ and $\mathcal{A}_\infty^{(K)}\tilde{\boldsymbol{u}}_\infty = \boldsymbol{b}_\infty$.*

Proof of Lemma D.6: Recall from (129) that $\boldsymbol{b}_\infty = \mathcal{B}_\infty \bar{g}_\infty$. Define $\tilde{\boldsymbol{u}}_\infty = \mathcal{B}_\infty\left(\mathcal{B}_\infty^\intercal \mathcal{B}_\infty\right)^\dagger \bar{g}_\infty$. We claim that $\mathcal{A}_\infty^{(K)}\tilde{\boldsymbol{u}}_\infty = \boldsymbol{b}_\infty$, because

$$\begin{aligned}\mathcal{A}_\infty^{(K)}\tilde{\boldsymbol{u}}_\infty &= \left(\mathcal{B}_\infty\mathcal{B}_\infty^\intercal\right)\mathcal{B}_\infty\left(\mathcal{B}_\infty^\intercal\mathcal{B}_\infty\right)^\dagger \bar{g}_\infty \\ &= \mathcal{B}_\infty\mathcal{B}_\infty^\intercal\left(\mathcal{B}_\infty\left(\mathcal{B}_\infty\right)^\dagger\right)\left(\mathcal{B}_\infty^\intercal\right)^\dagger \bar{g}_\infty \\ &= \mathcal{B}_\infty\left(\mathcal{B}_\infty^\intercal\left(\mathcal{B}_\infty^\intercal\right)^\dagger\right)\bar{g}_\infty \\ &= \mathcal{B}_\infty\bar{g}_\infty \\ &= \boldsymbol{b}_\infty\,,\end{aligned} \tag{162}$$

where the third equality is because $\mathcal{B}_\infty\left(\mathcal{B}_\infty\right)^\dagger$ is the projection operator onto $\operatorname{Ran}(\mathcal{B}_\infty) = \operatorname{Nul}^\perp(\mathcal{B}_\infty^\intercal)$, and the fourth equality is because $\mathcal{B}_\infty^\intercal\left(\mathcal{B}_\infty^\intercal\right)^\dagger$ is the projection operator onto $\operatorname{Ran}(\mathcal{B}_\infty^\intercal) = \operatorname{Nul}^\perp(\mathcal{B}_\infty)$.

It remains to establish that $\|\tilde{\boldsymbol{u}}_\infty\|_0 < \infty$. To show this, we see that

$$\begin{aligned}
&\int_D |\tilde{\boldsymbol{u}}_\infty(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \\
&= \int_D \int_{\Omega\times\Omega} \Big(\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big(\mathcal{M}_\infty^\dagger \bar{g}_\infty\big)(\boldsymbol{x})\Big) \\
&\qquad\qquad \cdot \Big(\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}')\big((\mathcal{M}_\infty)_\infty^\dagger \bar{g}_\infty\big)(\boldsymbol{x}')\Big)\hat{\nu}(d\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}')\mu_0(d\boldsymbol{\theta}') \\
&= \int_\Omega \int_\Omega M(\boldsymbol{x}, \boldsymbol{x}', \mu_\infty)\big(\mathcal{M}_\infty^\dagger \bar{g}_\infty\big)(\boldsymbol{x})\big(\mathcal{M}_\infty^\dagger \bar{g}_\infty\big)(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}') \\
&= \int_\Omega \big(\mathcal{M}_\infty^\dagger \bar{g}_\infty\big)(\boldsymbol{x}) \cdot \bar{g}_\infty(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) \\
&\leq \lambda_{\min}^{-1} \int_\Omega |\bar{g}_\infty(\boldsymbol{x})|^2 \hat{\nu}(d\boldsymbol{x})\ ,
\end{aligned} \tag{163}$$

where $\lambda_{\min}$ is the least nonzero eigenvalue of the matrix $\mathcal{M}_\infty$ (and hence $\lambda_{\min}^{-1}$ is the largest eigenvalue of $\mathcal{M}_\infty^\dagger$). Since

$$\begin{aligned}
\mathbb{E}_0|\bar{g}_\infty(\boldsymbol{x})|^2 &= \mathbb{E}_0\Big|\int_D \varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\omega_0(d\boldsymbol{\theta})\Big|^2 \\
&= \int_D \Big(\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) - \int_D \varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'), \boldsymbol{x})\mu_0(d\boldsymbol{\theta}')\Big)^2 \mu_0(d\boldsymbol{\theta}) \\
&\leq \int_D \big|\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big|^2 \mu_0(d\boldsymbol{\theta})\ ,
\end{aligned} \tag{164}$$

there is

$$\begin{aligned}
\|\tilde{\boldsymbol{u}}_\infty\|_0^2 &\leq \mathbb{E}_0 \int_D |\tilde{\boldsymbol{u}}_\infty(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \\
&\leq \lambda_{\min}^{-1} \int_\Omega \int_D \big(\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big)^2 \mu_0(d\boldsymbol{\theta})\nu(d\boldsymbol{x}) \\
&\leq \lambda_{\min}^{-1} C_\varphi^2 < \infty\ ,
\end{aligned} \tag{165}$$

(*End of the proof of Lemma D.6*) □

Coming back to the prof of Lemma D.3, we have shown that, as $t \to \infty$, (105) approaches the asymptotic dynamics

$$\dot{\boldsymbol{T}}_t = -\mathcal{A}_\infty^{(K)} \boldsymbol{T}_t - \boldsymbol{b}_\infty, \tag{166}$$

with $\mathcal{A}_\infty^{(K)}$ positive semidefinite and $\boldsymbol{b}_\infty$ in the range of $\mathcal{A}_\infty^{(K)}$. This is a stable system. Hence, the rest of the task is to examine what happens at finite time. To do so, we perform a change-of-variable with

$$\boldsymbol{z}_t = \boldsymbol{T}_t + \tilde{\boldsymbol{u}}_\infty, \tag{167}$$

with

$$\boldsymbol{u}_\infty = \mathcal{B}_\infty(\mathcal{B}_\infty^\intercal \mathcal{B}_\infty)^\dagger \bar{g}_\infty \tag{168}$$

as is defined in the proof of Lemma D.6. The dynamics of $\boldsymbol{z}_t$ is governed by

$$\begin{aligned}
\dot{\boldsymbol{z}}_t = \dot{\boldsymbol{T}}_t &= -\big(\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)}\big)\boldsymbol{T}_t - \boldsymbol{b}_t \\
&= -\mathcal{A}_t^{(K)}\boldsymbol{z}_t - \mathcal{A}_t^{(V)}\boldsymbol{z}_t - \big(\boldsymbol{b}_t - (\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)})\tilde{\boldsymbol{u}}_\infty\big)\ .
\end{aligned} \tag{169}$$

Thus, in integral form,

$$\boldsymbol{z}_t = \Pi(t,0)\boldsymbol{z}_0 + \int_0^t \Pi(t,s)\big(- \mathcal{A}_s^{(V)}\boldsymbol{z}_s - (\boldsymbol{b}_s - (\mathcal{A}_s^{(K)} + \mathcal{A}_s^{(V)})\tilde{\boldsymbol{u}}_\infty)\big)ds, \tag{170}$$

where $\Pi(t,s)$ is the fundamental solution (a.k.a. Green's function) associated with the time-variant homogeneous system

$$\dot{\boldsymbol{z}}_t = -\mathcal{A}_t^{(K)}\boldsymbol{z}_t\ . \tag{171}$$

Since $\mathcal{A}_t^{(K)}$ is positive semidefinite for all t, there is $\|\Pi(t,s)\|_0 \leq 1$ for $t > s$, where with a slight abuse of notation we also use $\|\cdot\|_0$ for the operator norm. Hence,

$$\begin{aligned}\|\boldsymbol{z}_t\|_0 \leq & \|\Pi(t,0)\|_0\|\boldsymbol{z}_0\|_0 + \int_0^t \|\Pi(t,s)\|_0\Big(\|\mathcal{A}_s^{(V)}\|_0\|\boldsymbol{z}_s\|_0 + \|\boldsymbol{b}_s - (\mathcal{A}_s^{(K)} + \mathcal{A}_s^{(V)})\tilde{\boldsymbol{u}}_\infty\|_0\Big)ds \\ \leq & \|\boldsymbol{z}_0\|_0 + \int_0^t \Big(\|\mathcal{A}_s^{(V)}\|_0\|\boldsymbol{z}_s\|_0 + \|\boldsymbol{b}_s - (\mathcal{A}_s^{(K)} + \mathcal{A}_s^{(V)})\tilde{\boldsymbol{u}}_\infty\|_0\Big)ds\ .\end{aligned} \tag{172}$$

By Grönwall's inequality, we thus have

$$\|\boldsymbol{z}_t\|_0 \leq \Big(\|\boldsymbol{z}_0\|_0 + \int_0^t \|\boldsymbol{b}_s - (\mathcal{A}_s^{(K)} + \mathcal{A}_s^{(V)})\tilde{\boldsymbol{u}}_\infty\|_0 ds\Big)e^{\int_0^t \|\mathcal{A}_s^{(V)}\|_0 ds}\ . \tag{173}$$

Therefore, $\|\boldsymbol{z}_t\|_0$ remains bounded for all time if we can show that

$$\int_0^\infty \|\boldsymbol{b}_t - (\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)})\tilde{\boldsymbol{u}}_\infty\|_0 dt < \infty, \qquad \int_0^\infty \|\mathcal{A}_t^{(V)}\|_0 dt < \infty\ . \tag{174}$$

Since

$$\|\boldsymbol{b}_t - (\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)})\tilde{\boldsymbol{u}}_\infty\|_0 \leq \|\boldsymbol{b}_t - \boldsymbol{b}_\infty\|_0 + \|(\mathcal{A}_t^{(K)} - \mathcal{A}_\infty^{(K)})\tilde{\boldsymbol{u}}_\infty\|_0 + \|\mathcal{A}_\infty^{(V)}\tilde{\boldsymbol{u}}_\infty\|_0 \tag{175}$$

we see that (174) is guaranteed by Lemmas D.2 and D.6.

This completes the proof of Lemma D.3. □

D.2.3 Proof of Lemma D.4

From D.3, we have that

$$\lim_{t\to\infty}\left\|\fint_0^t \dot{\boldsymbol{T}}_s ds\right\|_0 = \lim_{t\to\infty}\left\|\frac{1}{t}\left(\boldsymbol{T}_t - \boldsymbol{T}_0\right)ds\right\|_0 = 0\ . \tag{176}$$

By (105), we then obtain that

$$\lim_{t\to\infty}\left\|\fint_0^t \left(\mathcal{A}_s^{(K)}\boldsymbol{T}_s + \boldsymbol{b}_s\right)ds + \fint_0^t \mathcal{A}_s^{(V)}\boldsymbol{T}_s ds\right\|_0 = 0\ . \tag{177}$$

By (113) in Lemma D.2 as well as Lemma D.3, we know that

$$\lim_{t\to\infty}\left\|\fint_0^t \mathcal{A}_s^{(V)}\boldsymbol{T}_s ds\right\|_0 = 0\ . \tag{178}$$

Therefore,

$$\lim_{t\to\infty}\left\|\fint_0^t \left(\mathcal{A}_s^{(K)}\boldsymbol{T}_s + \boldsymbol{b}_s\right)ds\right\|_0 = 0\ . \tag{179}$$

Next, by (114) and (115) in Lemma D.2 as well as Lemma D.3, we know that

$$\lim_{t\to\infty}\left\|\fint_0^t \left(\mathcal{A}_s^{(K)}\boldsymbol{T}_s+\boldsymbol{b}_s\right)ds-\fint_0^t \left(\mathcal{A}_\infty^{(K)}\boldsymbol{T}_s+\boldsymbol{b}_\infty\right)ds\right\|_0=0\,. \tag{180}$$

Therefore,

$$\lim_{t\to\infty}\left\|\fint_0^t \left(\mathcal{A}_\infty^{(K)}\boldsymbol{T}_s+\boldsymbol{b}_\infty\right)ds\right\|_0=0\,. \tag{181}$$

With $\tilde{\boldsymbol{u}}_\infty$ defined in (168), as $\boldsymbol{b}_\infty=\mathcal{A}_\infty^{(K)}\boldsymbol{u}_\infty$, there is

$$\lim_{t\to\infty}\left\|\mathcal{A}_\infty^{(K)}\left(\fint_0^t \boldsymbol{T}_s ds-\boldsymbol{u}_\infty\right)\right\|_0=0\,. \tag{182}$$

Let $\boldsymbol{\xi}^{||}$ denote the component of a vector field $\boldsymbol{\xi}\in\mathcal{V}(D)$ that is in the range of $\mathcal{A}_\infty^{(K)}$. In the ERM setting, $\mathcal{A}_\infty^{(K)}$ has a least nonzero eigenvalue that is positive, and hence the above implies that

$$\lim_{t\to\infty}\left\|\left(\fint_0^t \boldsymbol{T}_s ds-\tilde{\boldsymbol{u}}_\infty\right)^{||}\right\|_0=0 \tag{183}$$

or

$$\lim_{t\to\infty}\left\|\left(\fint_0^t \boldsymbol{T}_s ds\right)^{||}-\tilde{\boldsymbol{u}}_\infty\right\|_0=0 \tag{184}$$

and therefore, as $\mathrm{Nul}(\mathcal{A}_\infty^{(K)})=\mathrm{Nul}(\mathcal{B}_\infty\mathcal{B}_\infty^\intercal)=\mathrm{Nul}(\mathcal{B}_\infty^\intercal)$, it follows that

$$\lim_{t\to\infty}\left\|\mathcal{B}_\infty^\intercal\left(\fint_0^t \boldsymbol{T}_s ds\right)-\mathcal{B}_\infty^\intercal\tilde{\boldsymbol{u}}_\infty\right\|_0=0\,. \tag{185}$$

Similar to (114), it can be shown that $\int_0^\infty\|\mathcal{B}_t-\mathcal{B}_\infty\|_0 dt<\infty$. Therefore, we have

$$\lim_{t\to\infty}\left\|\left(\fint_0^t \mathcal{B}_s^\intercal\boldsymbol{T}_s ds\right)-\mathcal{B}_\infty^\intercal\tilde{\boldsymbol{u}}_\infty\right\|_0=0\,. \tag{186}$$

Now,

$$\begin{aligned}\fint_0^t\langle\boldsymbol{T}_s,\mathcal{A}_s^{(K)}\boldsymbol{T}_s\rangle_0 ds&=\fint_0^t\langle\mathcal{B}_s^\intercal\boldsymbol{T}_s,\mathcal{B}_s^\intercal\boldsymbol{T}_s\rangle_{\hat{\nu},0}ds\\&\geq\left\langle\left(\fint_0^t\mathcal{B}_s^\intercal\boldsymbol{T}_s ds\right),\left(\fint_0^t\mathcal{B}_s^\intercal\boldsymbol{T}_s ds\right)\right\rangle_{\hat{\nu},0}\,.\end{aligned} \tag{187}$$

Hence,

$$\begin{aligned}\lim_{t\to\infty}\fint_0^t\langle\boldsymbol{T}_s,\mathcal{A}_s^{(K)}\boldsymbol{T}_s\rangle_0 ds&\geq\lim_{t\to\infty}\left\langle\left(\fint_0^t\mathcal{B}_s^\intercal\boldsymbol{T}_s ds\right),\left(\fint_0^t\mathcal{B}_s^\intercal\boldsymbol{T}_s ds\right)\right\rangle_{\hat{\nu},0}\\&=\langle\mathcal{B}_\infty^\intercal\tilde{\boldsymbol{u}}_\infty,\mathcal{B}_\infty^\intercal\tilde{\boldsymbol{u}}_\infty\rangle_{\hat{\nu},0}\\&=\left\langle\mathcal{B}_\infty^\intercal\left(\mathcal{A}_\infty^{(K)}\right)^\dagger\boldsymbol{b}_\infty,\mathcal{B}_\infty^\intercal\left(\mathcal{A}_\infty^{(K)}\right)^\dagger\boldsymbol{b}_\infty\tilde{\boldsymbol{u}}_\infty\right\rangle_{\hat{\nu},0}\\&=\left\langle\left(\mathcal{A}_\infty^{(K)}\right)^\dagger\boldsymbol{b}_\infty,\left(\mathcal{A}_\infty^{(K)}\right)\left(\mathcal{A}_\infty^{(K)}\right)^\dagger\boldsymbol{b}_\infty\right\rangle_0\\&=\left\langle\boldsymbol{b}_\infty,\left(\mathcal{A}_\infty^{(K)}\right)^\dagger\boldsymbol{b}_\infty\right\rangle_0\,.\end{aligned} \tag{188}$$

□

D.2.4 Proof of Lemma D.5

Since

$$\bar{g}_\infty(\boldsymbol{x}) = \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x})\omega_0(d\boldsymbol{\theta}) \;, \tag{189}$$

we know that when viewed as a L-dimensional random vector, $\bar{g}_\infty$ has the distribution

$$\bar{g}_\infty \sim \mathcal{N}(0, \bar{C}_\infty) \;, \tag{190}$$

where

$$\begin{aligned}\left(\bar{C}_\infty\right)_{ij} :=& \mathbb{E}_0\left[\bar{g}_\infty(\boldsymbol{x}_i)\bar{g}_\infty(\boldsymbol{x}_j)\right] \\ =& \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)\varphi(\boldsymbol{\theta}, \boldsymbol{x}_j)\mu_\infty(d\boldsymbol{\theta}) - \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)\mu_\infty(d\boldsymbol{\theta}) \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x}_j)\mu_\infty(d\boldsymbol{\theta}') \;,\end{aligned} \tag{191}$$

by the covariance of ω_0, (33). Thus, we decompose $\bar{C}_\infty$ as $\bar{C}_\infty = \bar{C}_\infty^{(1)} - \bar{C}_\infty^{(2)}$, with

$$\left(\bar{C}_\infty^{(1)}\right)_{ij} = \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)\varphi(\boldsymbol{\theta}, \boldsymbol{x}_j)\mu_\infty(d\boldsymbol{\theta}) \;, \tag{192}$$

$$\left(\bar{C}_\infty^{(2)}\right)_{ij} = \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)\mu_\infty(d\boldsymbol{\theta}) \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x}_j)\mu_\infty(d\boldsymbol{\theta}') \;. \tag{193}$$

Since $\bar{C}_\infty$ is PSD, its square root $\left(\bar{C}_\infty\right)^{\frac{1}{2}}$ is well-defined. By the property of multivariate Gaussian, we can write

$$\bar{g}_\infty \overset{\mathrm{d}}{=} \left(\bar{C}_\infty\right)^{\frac{1}{2}} w \;, \tag{194}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution, and $w \in \mathbb{R}^L$ follows the distribution

$$w \sim \mathcal{N}(0, \mathrm{Id}_L) \;. \tag{195}$$

This means that almost surely, $\bar{g}_\infty \in \mathrm{Ran}\left(\left(\bar{C}_\infty\right)^{\frac{1}{2}}\right)$, and which would imply that $\bar{g}_\infty \in \mathrm{Ran}\left(\bar{C}_\infty\right)$. This means that almost surely, we can write

$$\bar{g}_\infty = \bar{C}_\infty^{(1)} w^{(1)} - \bar{C}_\infty^{(2)} w^{(2)} \tag{196}$$

for some pair of $w^{(1)}, w^{(2)} \in \mathbb{R}^L$. Our goal is then to show that both $\bar{C}_\infty^{(1)} w^{(1)}$ and $\bar{C}_\infty^{(2)} w^{(2)}$ belong to $\mathrm{Ran}(\mathcal{B}_\infty^\intercal)$. Here, under Assumption 2.1, since $\varphi(\boldsymbol{\theta}, \boldsymbol{x}) = c\hat{\varphi}(\boldsymbol{z}, \boldsymbol{x})$ when $\boldsymbol{\theta} = \begin{bmatrix} c & \boldsymbol{z} \end{bmatrix}^\intercal$, there is

$$\nabla \varphi(\boldsymbol{\theta}, \boldsymbol{x}) = \begin{bmatrix} \hat{\varphi}(\boldsymbol{z}, \boldsymbol{x}) \\ c\nabla_{\boldsymbol{z}} \hat{\varphi}(\boldsymbol{z}, \boldsymbol{x}) \end{bmatrix} \;. \tag{197}$$

Therefore, first, we have

$$\begin{aligned}\left(\bar{C}_\infty^{(1)} w^{(1)}\right)_i =& \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i) \left(\sum_{j=1}^L \varphi(\boldsymbol{\theta}, \boldsymbol{x}_j) w_j^{(1)}\right) \mu_\infty(d\boldsymbol{\theta}) \\ =& \int_D \nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)^\intercal \cdot \begin{bmatrix} c(\boldsymbol{\theta}) \left(\sum_{j=1}^L \varphi(\boldsymbol{\theta}, \boldsymbol{x}_j) w_j^{(1)}\right) \\ 0 \end{bmatrix} \mu_\infty(d\boldsymbol{\theta}) \\ =& \mathcal{B}_\infty^\intercal \boldsymbol{\xi}^{(1)} \;,\end{aligned} \tag{198}$$

with

$$\boldsymbol{\xi}(\boldsymbol{\theta})^{(1)} = \begin{bmatrix} c(\boldsymbol{\theta}) \left(\sum_{j=1}^{L} \varphi(\boldsymbol{\theta}, \boldsymbol{x}_j) w_j^{(1)} \right) \\ 0 \end{bmatrix} . \tag{199}$$

This means that $\left(\bar{C}_{\infty}^{(1)} w^{(1)}\right) \in \mathrm{Ran}(\mathcal{B}_{\infty}^{\intercal})$.

Second, there is

$$\begin{aligned} \left(\bar{C}_{\infty}^{(2)} w^{(2)}\right)_i &= \left(\int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i) \mu_{\infty}(d\boldsymbol{\theta}) \right) \left(\sum_{j=1}^{L} w_j^{(2)} \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x}_j) \mu_{\infty}(d\boldsymbol{\theta}') \right) \\ &= \int_D \nabla \varphi(\boldsymbol{\theta}, \boldsymbol{x}_i)^{\intercal} \cdot \begin{bmatrix} c(\boldsymbol{\theta}) \left(\sum_{j=1}^{L} w_j^{(2)} \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x}_j) \mu_{\infty}(d\boldsymbol{\theta}') \right) \\ 0 \end{bmatrix} \mu_{\infty}(d\boldsymbol{\theta}) \\ &= \mathcal{B}_{\infty}^{\intercal} \boldsymbol{\xi}^{(2)} , \end{aligned} \tag{200}$$

with

$$\boldsymbol{\xi}(\boldsymbol{\theta})^{(2)} = \begin{bmatrix} c(\boldsymbol{\theta}) \left(\sum_{j=1}^{L} w_j^{(2)} \int_D \varphi(\boldsymbol{\theta}', \boldsymbol{x}_j) \mu_{\infty}(d\boldsymbol{\theta}') \right) \\ 0 \end{bmatrix} \tag{201}$$

This means that $\left(\bar{C}_{\infty}^{(2)} w^{(2)}\right) \in \mathrm{Ran}(\mathcal{B}_{\infty}^{\intercal})$. Hence the lemma is proved. □

D.3 Proof of Lemma 3.8

With $\mathfrak{D}_t$ defined in (104), for (47) to hold, it is sufficient to show that

$$\lim_{T \to \infty} \fint_0^T \mathfrak{D}_t dt \le 0 . \tag{202}$$

Recall from (110) that

$$\fint_0^T \mathfrak{D}_t dt = -\frac{1}{T} \|\boldsymbol{T}_T\|_0^2 - 2 \fint_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(V)} \boldsymbol{T}_t \rangle_0 dt - \fint_0^T \langle \boldsymbol{T}_t, \mathcal{A}_t^{(K)} \boldsymbol{T}_t \rangle_0 dt . \tag{203}$$

Since $\boldsymbol{T}_0 = 0$ and $\mathcal{A}_t^{(K)}$ is PSD, we see that the assumption (51) is sufficient.

D.4 Proof of Theorem 3.9 (The case with curvature assumptions)

Our goal is to verify (51) in order to apply Lemma 3.8. We first see that

$$\begin{aligned} &\mathbb{E}_0 \int_D \langle \boldsymbol{T}_t(\boldsymbol{\theta}), \nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t) \boldsymbol{T}_t(\boldsymbol{\theta}) \rangle \mu_0(d\boldsymbol{\theta}) \\ \ge &\mathbb{E}_0 \int_D \lambda_{\min}(\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)) |\boldsymbol{T}_t(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \\ \ge &\mathbb{E}_0 \int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)), 0\right\} |\boldsymbol{T}_t(\boldsymbol{\theta})|^2 \mu_0(d\boldsymbol{\theta}) \\ = &\int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)), 0\right\} \left(\mathbb{E}_0 |\boldsymbol{T}_t(\boldsymbol{\theta})|^2\right) \mu_0(d\boldsymbol{\theta}) \\ \ge &\int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)), 0\right\} \left(\sup_{\boldsymbol{\theta} \in \operatorname{supp} \mu_0} \mathbb{E}_0 |\boldsymbol{T}_t(\boldsymbol{\theta})|^2 \right) \mu_0(d\boldsymbol{\theta}) \\ \ge &\|\boldsymbol{T}_t\|_{\sup}^2 \left(\int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)), 0\right\} \mu_0(d\boldsymbol{\theta}) \right) , \end{aligned} \tag{204}$$

where we define, for $\boldsymbol{\xi} \in \mathcal{V}(D)$,

$$\|\boldsymbol{\xi}\|_{\sup} := \sup_{\boldsymbol{\theta} \in \operatorname{supp} \mu_0} \left(\mathbb{E}_0 |\boldsymbol{\xi}(\boldsymbol{\theta})|^2\right)^{\frac{1}{2}}, \tag{205}$$

which is a norm on $\mathcal{V}(D)$.

Hence, if we assume that $\left|\int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\theta}, \mu_t)), 0\right\} \mu_0(d\boldsymbol{\theta})\right|$ is small asymptotically, then what remains is to upper-bound $\|\boldsymbol{T}_t\|_{\sup}$. Recall from (105) that the dynamics of $\boldsymbol{T}_t$ is governed by

$$\dot{\boldsymbol{T}}_t = -(\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)})\boldsymbol{T}_t - \boldsymbol{b}_t, \tag{206}$$

Thus, in the $\|\cdot\|_{\sup}$ norm defined above, we have

$$\begin{aligned} \frac{d}{dt}\|\boldsymbol{T}_t\|_{\sup} \leq & \| - (\mathcal{A}_t^{(K)} + \mathcal{A}_t^{(V)})\boldsymbol{T}_t - \boldsymbol{b}_t\|_{\sup} \\ \leq & \|\mathcal{A}_t^{(K)}\boldsymbol{T}_t\|_{\sup} + \|\mathcal{A}_t^{(V)}\boldsymbol{T}_t\|_{\sup} + \|\boldsymbol{b}_t\|_{\sup} . \end{aligned} \tag{207}$$

We then want to bound the growth of $\|\boldsymbol{T}_t\|_{\sup}$ by upper-bounding the RHS. Note that for $\boldsymbol{\xi} \in \mathcal{V}(D)$,

$$\begin{aligned} \|\mathcal{A}_t^{(V)}\boldsymbol{\xi}\|_{\sup}^2 =& \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 |(\mathcal{A}_t^{(V)}\boldsymbol{\xi})(\boldsymbol{\theta})|^2 \\ =& \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 |\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)\boldsymbol{\xi}(\boldsymbol{\theta})|^2 \\ \leq & \sup_{\boldsymbol{\theta}\in D} |\nabla\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \mu_t)|^2 \mathbb{E}_0 |\boldsymbol{\xi}(\boldsymbol{\theta})|^2 \\ \leq & (C_{\nabla\nabla\varphi} C_\varphi + \lambda)^2 \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 |\boldsymbol{\xi}(\boldsymbol{\theta})|^2 \\ =& (C_{\nabla\nabla\varphi} C_\varphi + \lambda)^2 \|\boldsymbol{\xi}\|_{\sup}^2 , \end{aligned} \tag{208}$$

$$\begin{aligned} \|\mathcal{A}_t^{(K)}\boldsymbol{\xi}\|_{\sup}^2 =& \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 |(\mathcal{A}_t^{(K)}\boldsymbol{\xi})(\boldsymbol{\theta})|^2 \\ =& \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 \left| \int_D \nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\boldsymbol{\xi}(\boldsymbol{\theta}')\mu_0(d\boldsymbol{\theta}')\right|^2 \\ \leq & \sup_{\boldsymbol{\theta}\in D} \mathbb{E}_0 \int_D |\nabla'\nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))|^2 |\boldsymbol{\xi}(\boldsymbol{\theta}')|^2 \mu_0(d\boldsymbol{\theta}') \\ \leq & \sup_{\boldsymbol{\theta}\in D} C_{\nabla\varphi}^4 \int_D \mathbb{E}_0 |\boldsymbol{\xi}(\boldsymbol{\theta}')|^2 \mu_0(d\boldsymbol{\theta}') \\ \leq & C_{\nabla\varphi}^4 \sup_{\boldsymbol{\theta}'\in D} \mathbb{E}_0 |\boldsymbol{\xi}(\boldsymbol{\theta}')|^2 \\ =& C_{\nabla\varphi}^4 \|\boldsymbol{\xi}\|_{\sup}^2 . \end{aligned} \tag{209}$$

Thus,

$$\|\mathcal{A}_t^{(K)}\boldsymbol{T}_t\|_{\sup} + \|\mathcal{A}_t^{(V)}\boldsymbol{T}_t\|_{\sup} \leq (C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi} C_\varphi + \lambda)\|\boldsymbol{T}_t\|_{\sup} . \tag{210}$$

To bound $\|\boldsymbol{b}_t\|_{\sup}$, we recall that

$$\begin{aligned} \boldsymbol{b}_t(\boldsymbol{\theta}) =& \int_D \nabla K(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{\Theta}_t(\boldsymbol{\theta}'))\omega_0(d\boldsymbol{\theta}') \\ =& \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\bar{g}_t(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) , \end{aligned} \tag{211}$$

with

$$\bar{g}_t(\boldsymbol{x}) = \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\omega_0(d\boldsymbol{\theta}) \ . \tag{212}$$

This implies that $\forall \boldsymbol{\theta} \in \operatorname{supp} \mu_0$,

$$|\boldsymbol{b}_t(\boldsymbol{\theta})| \le \frac{1}{L} C_{\nabla\varphi} \sum_{l=1}^{L} |\bar{g}_t(\boldsymbol{x}_l)| \tag{213}$$

and so

$$\begin{aligned} \mathbb{E}_0|\boldsymbol{b}_t(\boldsymbol{\theta})|^2 \le& C_{\nabla\varphi}^2 \mathbb{E}_0 \left(\frac{1}{L} \sum_{l=1}^{L} |\bar{g}_t(\boldsymbol{x}_l)| \right)^2 \\ \le& C_{\nabla\varphi}^2 \mathbb{E}_0 \left(\frac{1}{L} \sum_{l=1}^{L} |\bar{g}_t(\boldsymbol{x}_l)|^2 \right) \\ \le& C_{\nabla\varphi}^2 \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_0 |\bar{g}_t(\boldsymbol{x}_l)|^2 \ . \end{aligned} \tag{214}$$

On the other hand, similar to (164), we have

$$\begin{aligned} \mathbb{E}_0|\bar{g}_t(\boldsymbol{x})|^2 =& \mathbb{E}_0 \Big| \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\omega_0(d\boldsymbol{\theta}) \Big|^2 \\ &= \int_D \Big(\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \int_D \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}'), \boldsymbol{x})\mu_0(d\boldsymbol{\theta}') \Big)^2 \mu_0(d\boldsymbol{\theta}) \\ &\le \int_D \big| \varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \big|^2 \mu_0(d\boldsymbol{\theta}) \\ &\le C_\varphi^2 \ , \end{aligned} \tag{215}$$

Thus, there is $\forall \boldsymbol{\theta} \in \operatorname{supp} \mu_0$,

$$\mathbb{E}_0|\boldsymbol{b}_t(\boldsymbol{\theta})|^2 \le C_{\nabla\varphi}^2 C_\varphi^2 \tag{216}$$

and so

$$\|\boldsymbol{b}_t\|_{\sup} \le C_{\nabla\varphi} C_\varphi \ . \tag{217}$$

Therefore, based on (207), we have

$$\frac{d}{dt}\|\boldsymbol{T}_t\|_{\sup} \le (C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)\|\boldsymbol{T}_t\|_{\sup} + C_{\nabla\varphi}C_\varphi \ . \tag{218}$$

Since $\boldsymbol{T}_0 = 0$, we thus have

$$\begin{aligned} \|\boldsymbol{T}_t\|_{\sup} \le& C_{\nabla\varphi}C_\varphi \int_0^t e^{(C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)(t-s)} ds \\ =& C_{\nabla\varphi}C_\varphi e^{(C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)t} \int_0^t e^{-(C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)s} ds \\ \le& \frac{C_{\nabla\varphi}C_\varphi}{C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda} e^{(C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)t} \end{aligned} \tag{219}$$

Now, using (204), we see that in order for (51) to hold, it is sufficient to have

$$\lim_{t\to\infty} e^{(C_{\nabla\varphi}^2 + C_{\nabla\nabla\varphi}C_\varphi + \lambda)t} \left(\int_D \min\left\{ \lambda_{\min}(\nabla\nabla V(\boldsymbol{\theta}, \mu_t)), 0 \right\} \mu_0(d\boldsymbol{\theta}) \right) = 0 \tag{220}$$

and therefore sufficient to have

$$-\int_D \min\left\{\lambda_{\min}(\nabla\nabla V(\boldsymbol{\theta},\mu_t)),0\right\}\mu_0(d\boldsymbol{\theta}) \sim O\left(e^{-(C^2_{\nabla\varphi}+C_{\nabla\nabla\varphi}C_\varphi+\lambda)t}\right) \tag{221}$$

□

D.5 Proof of Theorem 3.10 (Regularized case)

Recall from Proposition 3.3 that the dynamics of g_t is governed by

$$g_t(\boldsymbol{x}) + \int_0^t \int_\Omega \Gamma_{t,s}(\boldsymbol{x},\boldsymbol{x}')g_s(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')ds = \bar{g}_t(\boldsymbol{x}) \,, \tag{222}$$

with

$$\Gamma_{t,s}(\boldsymbol{x},\boldsymbol{x}') = \int_D \langle \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}),\boldsymbol{x}), J_{t,s}(\boldsymbol{\theta})\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}),\boldsymbol{x}')\rangle\mu_0(d\boldsymbol{\theta}) \,, \tag{223}$$

with $J_{t,s}$ being the Jacobian of the flow $\boldsymbol{\Theta}_t$.

In the ERM setting, $\operatorname{supp}\hat{\nu}$ is singular, thus we have $\hat{\nu}(d\boldsymbol{x}) = L^{-1}\sum_{k=1}^L \delta_{\boldsymbol{x}_k}(d\boldsymbol{x})$, where L is the total number of training data points. We define $\mathcal{W}_L(\Omega)$ together with the inner product $\langle\cdot,\cdot\rangle_{\hat{\nu},0}$ and the norm $\|\cdot\|_{\hat{\nu},0}$ as in Appendix A. We will also continue to consider g_t and $\bar{g}_t$ equivalently as L-dimensional vectors,

$$\begin{pmatrix} g_t(\boldsymbol{x}_1) & \cdots & g_t(\boldsymbol{x}_L)\end{pmatrix}^T, \qquad \begin{pmatrix} \bar{g}_t(\boldsymbol{x}_1) & \cdots & \bar{g}_t(\boldsymbol{x}_L)\end{pmatrix}^T \,, \tag{224}$$

respectively. Thus, $\Gamma_{t,s}$ can also be represented by the $L\times L$ matrix

$$\begin{pmatrix} \Gamma_{t,s}(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & \Gamma_{t,s}(\boldsymbol{x}_1,\boldsymbol{x}_L) \\ \vdots & & \vdots \\ \Gamma_{t,s}(\boldsymbol{x}_L,\boldsymbol{x}_1) & \cdots & \Gamma_{t,s}(\boldsymbol{x}_L,\boldsymbol{x}_L)\end{pmatrix} . \tag{225}$$

Under such an abuse of notations, we can simplify (222) into

$$g_t + \int_0^t \Gamma_{t,s} g_s ds = \bar{g}_t \,. \tag{226}$$

Thus, the goal is to prove that

$$\lim_{t\to\infty}\sup \fint_0^t \mathbb{E}_0\|g_t\|^2_{\hat{\nu}} dt \le \mathbb{E}_0\|\bar{g}_\infty\|^2_{\hat{\nu}} \,. \tag{227}$$

As in (44), we also define

$$\Gamma^\infty_{t-s}(\boldsymbol{x},\boldsymbol{x}') = \int_D \langle \nabla\varphi(\boldsymbol{\theta},\boldsymbol{x}), e^{-(t-s)\nabla\nabla V_\infty(\boldsymbol{\theta})}\nabla\varphi(\boldsymbol{\theta},\boldsymbol{x}')\rangle\mu_\infty(d\boldsymbol{\theta}) \,, \tag{228}$$

where for simplicity, we write $V_t(\cdot)$ for $V(\cdot,\mu_t)$ and $V_\infty(\cdot)$ for $V(\cdot,\mu_\infty)$. Then the heuristic argument outlined in Section 3.2 before Theorem 3.4 amounts to rewriting (226) as

$$g_t + \int_0^t \Gamma^\infty_{t-s} g_s ds = \bar{g}_t + \int_0^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds \tag{229}$$

and then arguing that 1) Γ^∞ is a nonnegative convolution-type Volterra kernel, and 2) the second term on the RHS is small. Rigorously, we need to introduce an extra level of complication: for every $t_0 > 0$, we can rewrite (226) into

$$\begin{aligned} g_t &= \bar{g}_t - \int_{t_0}^t \Gamma_{t,s} g_s ds - \int_0^{t_0} \Gamma_{t,s} g_s ds \\ &= \bar{g}_t - \int_{t_0}^t \Gamma^\infty_{t-s} g_s ds + \int_{t_0}^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds - \int_0^{t_0} \Gamma_{t,s} g_s ds \;. \end{aligned} \tag{230}$$

Then, for any $T > t_0$, by multiplying g_t and integrating from t_0 to T, we get

$$\begin{aligned} &\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt + \int_{t_0}^T \int_{t_0}^t \langle g_t, \Gamma^\infty_{t-s} g_s \rangle_{\hat{\nu}} ds dt \\ \leq &\int_{t_0}^T \langle g_t, \bar{g}_t \rangle_{\hat{\nu}} dt + \int_{t_0}^T \langle g_t, \int_{t_0}^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds \rangle_{\hat{\nu}} dt + \int_{t_0}^T \langle g_t, \int_0^{t_0} \Gamma_{t,s} g_s ds \rangle_{\hat{\nu}} dt \;. \end{aligned} \tag{231}$$

Then firstly, the second term on the LHS is nonnegative because of the nonnegativity of Γ^∞_t as a convolution-type Volterra kernel, as proven in Appendix D.1.

Hence, we have

$$\begin{aligned} \int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \leq &\int_{t_0}^T \langle g_t, \bar{g}_t \rangle_{\hat{\nu}} dt + \int_{t_0}^T \langle g_t, \int_{t_0}^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds \rangle_{\hat{\nu}} dt \\ &+ \int_{t_0}^T \left\langle g_t, \int_0^{t_0} \Gamma_{t,s} g_s ds \right\rangle_{\hat{\nu}} dt \;. \end{aligned} \tag{232}$$

By Cauchy-Schwartz,

$$\int_{t_0}^T \langle g_t, \bar{g}_t \rangle_{\hat{\nu}} dt \leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \|\bar{g}_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} , \tag{233}$$

$$\begin{aligned} \int_{t_0}^T \left\langle g_t, \int_{t_0}^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds \right\rangle_{\hat{\nu}} dt &\leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \left\| \int_{t_0}^t (\Gamma^\infty_{t-s} - \Gamma_{t,s}) g_s ds \right\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \\ &\leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right) \left(\int_{t_0}^T \int_{t_0}^t \|\Gamma^\infty_{t-s} - \Gamma_{t,s}\|^2_{\hat{\nu}} ds dt \right)^{\frac{1}{2}} , \text{ and} \end{aligned} \tag{234}$$

$$\begin{aligned} \int_{t_0}^T \left\langle g_t, \int_0^{t_0} \Gamma_{t,s} g_s ds \right\rangle_{\hat{\nu}} dt &\leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \left\| \int_0^{t_0} \Gamma_{t,s} g_s ds \right\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \\ &\leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \left(\int_0^{t_0} \|\Gamma_{t,s}\|^2_{\hat{\nu}} ds \right) \left(\int_0^{t_0} \|g_s\|^2_{\hat{\nu}} ds \right) dt \right)^{\frac{1}{2}} \\ &\leq \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_0^{t_0} \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \int_0^{t_0} \|\Gamma_{t,s}\|^2_{\hat{\nu}} ds dt \right)^{\frac{1}{2}} . \end{aligned} \tag{235}$$

Therefore, putting everything together, we have

$$\begin{aligned} \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \leq &\left(\int_{t_0}^T \|\bar{g}_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} + \left(\int_{t_0}^T \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \int_{t_0}^t \|\Gamma^\infty_{t-s} - \Gamma_{t,s}\|^2_{\hat{\nu}} ds dt \right)^{\frac{1}{2}} \\ &+ \left(\int_0^{t_0} \|g_t\|^2_{\hat{\nu}} dt \right)^{\frac{1}{2}} \left(\int_{t_0}^T \int_0^{t_0} \|\Gamma_{t,s}\|^2_{\hat{\nu}} ds dt \right)^{\frac{1}{2}} , \end{aligned} \tag{236}$$

and hence, using $\fint_a^b \cdot\, dt$ to denote the averaged integral $\frac{1}{b-a}\int_a^b \cdot\, dt$,

$$\begin{aligned} \Big(\fint_{t_0}^T \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \leq \Big(\fint_{t_0}^T \|\bar{g}_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} + \Big(\fint_{t_0}^T \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \Big(\int_{t_0}^T \int_{t_0}^t \|\Gamma_{t-s}^\infty - \Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt\Big)^{\frac{1}{2}} \\ + \Big(\int_0^{t_0} \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \Big(\fint_{t_0}^T \int_0^{t_0} \|\Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt\Big)^{\frac{1}{2}}, \end{aligned} \tag{237}$$

or

$$\begin{aligned} \left(1 - \Big[\int_{t_0}^T \int_{t_0}^t \|\Gamma_{t-s}^\infty - \Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt\Big]^{\frac{1}{2}}\right) \Big(\fint_{t_0}^T \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \\ \leq \Big(\fint_{t_0}^T \|\bar{g}_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} + \Big(\int_0^{t_0} \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \Big(\fint_{t_0}^T \int_0^{t_0} \|\Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt\Big)^{\frac{1}{2}}. \end{aligned} \tag{238}$$

Lemma D.7. *Under the assumptions in Theorem 3.10, we have*

$$\lim_{t_0 \to \infty} \int_{t_0}^\infty \int_{t_0}^t \|\Gamma_{t-s}^\infty - \Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt = 0 \tag{239}$$

and $\forall t_0 > 0$,

$$\lim_{T \to \infty} \fint_{t_0}^T \int_0^{t_0} \|\Gamma_{t,s}\|^2 ds dt = 0\,. \tag{240}$$

The lemma is proved in Appendix D.5.1, and let us first proceed with the proof of the theorem assuming this lemma. Suppose for contradiction that (227) does not hold, meaning that

$$\lim_{T\to\infty} \sup \big(\fint_0^T \|g_t\|_{\hat{\nu}}^2 dt\big)^{\frac{1}{2}} = \|\bar{g}_\infty\|_{\hat{\nu}} + \epsilon \tag{241}$$

for some $\epsilon > 0$. We will select a pair of t_0 and T for which the inequality (238) cannot be satisfied. Firstly, by the convergence of $\bar{g}_t$ to $\bar{g}_\infty$, $\exists t_a > 0$ such that $\forall t_1, t_2 > t_a$,

$$\Big(\fint_{t_1}^{t_2} \|\bar{g}_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \leq \|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{6}\epsilon\,. \tag{242}$$

Secondly, by our assumption (241) and the first part of Lemma D.7, $\exists t_0 > t_a$ such that both

$$\Big(\fint_0^{t_0} \|g_t\|_{\hat{\nu}}^2 dt\Big)^{\frac{1}{2}} \leq \|\bar{g}_\infty\|_{\hat{\nu}} + 2\epsilon \tag{243}$$

and

$$\int_{t_0}^\infty \int_{t_0}^t \|\Gamma_{t-s}^\infty - \Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt < \frac{\epsilon}{6\|\bar{g}_\infty\|_{\hat{\nu}} + 3\epsilon} \tag{244}$$

are satisfied. In particular, (243) implies

$$\Big(\int_0^{t_0} |g_t|^2 dt\Big)^{\frac{1}{2}} \leq t_0^{\frac{1}{2}} \cdot (|\bar{g}_\infty| + 2\epsilon) \tag{245}$$

Let

$$\delta = \left(\frac{\epsilon}{6t_0^{\frac{1}{2}} \cdot (\|\bar{g}_\infty\|_{\hat{\nu}} + 2\epsilon)} \right)^2 > 0 \,. \tag{246}$$

By the second part of Lemma D.7, $\exists t_b > t_0$ such that $\forall T > t_b$,

$$\fint_{t_0}^{T} \int_0^{t_0} \|\Gamma_{t,s}\|^2 ds dt < \delta \tag{247}$$

so that the last term in (238) satisfies

$$\left(\int_0^{t_0} \|g_t\|_{\hat{\nu}}^2 dt \right)^{\frac{1}{2}} \left(\fint_{t_0}^{T} \int_0^{t_0} \|\Gamma_{t,s}\|^2 ds dt \right)^{\frac{1}{2}} < \tfrac{1}{6}\epsilon \tag{248}$$

By our assumption (241), we can choose a $T > t_b$ such that

$$\Big(\fint_0^{T} \|g_t\|_{\hat{\nu}}^2 dt \Big)^{\frac{1}{2}} \geq \|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{2}{3}\epsilon \,. \tag{249}$$

Since

$$\left(\fint_0^{t_0} \|g_t\|_{\hat{\nu}}^2 dt \right)^{\frac{1}{2}} \leq \|\bar{g}_\infty\|_{\hat{\nu}} + 2\epsilon, \tag{250}$$

we can assume without loss of generality that $\frac{T}{t_0}$ is large enough so that

$$\Big(\fint_{t_0}^{T} \|g_t\|_{\hat{\nu}}^2 dt \Big)^{\frac{1}{2}} \geq \|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{2}\epsilon \,. \tag{251}$$

Thus, back to the inequality (238), the LHS is strictly lower-bounded by

$$\big(\|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{2}\epsilon\big)\big(1 - \frac{\epsilon}{6\|\bar{g}_\infty\|_{\hat{\nu}} + 3\epsilon}\big) = \|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{3}\epsilon, \tag{252}$$

whereas the RHS is strictly upper-bounded by

$$\|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{6}\epsilon + \tfrac{1}{6}\epsilon = \|\bar{g}_\infty\|_{\hat{\nu}} + \tfrac{1}{3}\epsilon \,. \tag{253}$$

This gives contradiction and we are done with the proof of Theorem 3.10. □

D.5.1 Proof of Lemma D.7

It remains to prove Lemma D.7. To do so we will need an auxiliary result, that we state and prove first:

Lemma D.8. *Let $\Delta\Gamma_{t,s} := \Gamma_{t,s} - \Gamma_{t-s}^{\infty}$. If $\nabla\nabla V$ is uniformly positive definite with eigenvalues lower-bounded by λ, then there exists constants C and C' whose values depend on $|D'|$, C_φ, $C_{\nabla\varphi}$, $C_{\nabla\nabla\varphi}$, and $L_{\nabla\nabla\varphi}$ such that*

$$\|\Delta\Gamma_{t,s}\|_{\hat{\nu}} \leq C e^{-\lambda(t-s)} \int_D \Big(|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| + \big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})\big)\, e^{C'(U_s(\boldsymbol{\theta}) + \bar{U}_s)} \Big)\, \mu_0(d\boldsymbol{\theta}) \tag{254}$$

where $\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta}) = \boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})$.

Proof of Lemma D.8: To bound $\|\Delta\Gamma_{t,s}\|_{\hat{\nu}}$, we bound $\|\Delta\Gamma_{t,s}\eta\|_{\hat{\nu}}$ for $\eta \in \mathbb{R}^L$. Note that $\Delta\Gamma_{t,s}\eta$ can be obtained in the following way. Consider the two systems

$$\begin{cases} \frac{d}{dt}\boldsymbol{\xi}_t(\boldsymbol{\theta}) = -\nabla\nabla V_t(\boldsymbol{\Theta}_t(\boldsymbol{\theta}))\boldsymbol{\xi}_t(\boldsymbol{\theta}) \\ \boldsymbol{\xi}_s(\boldsymbol{\theta}) = \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}')\eta(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}') \end{cases} \tag{255}$$

$$\begin{cases} \frac{d}{dt}\boldsymbol{\xi}'_t(\boldsymbol{\theta}) = -\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\boldsymbol{\xi}'_t(\boldsymbol{\theta}) \\ \boldsymbol{\xi}'_s(\boldsymbol{\theta}) = \int_\Omega \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}')\eta(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}') \end{cases} \tag{256}$$

Then there is

$$\begin{aligned} (\Gamma_{t,s}\eta)(\boldsymbol{x}) &= \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \boldsymbol{\xi}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \\ (\Gamma^\infty_{t-s}\eta)(\boldsymbol{x}) &= \int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \boldsymbol{\xi}'_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \end{aligned} \tag{257}$$

and hence

$$\begin{aligned} (\Delta\Gamma_{t,s}\eta)(\boldsymbol{x}) &= \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})\boldsymbol{\xi}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) - \int_D \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\boldsymbol{\xi}'_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \\ &= \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) \cdot \big(\boldsymbol{\xi}_t(\boldsymbol{\theta}) - \boldsymbol{\xi}'_t(\boldsymbol{\theta})\big)\mu_0(d\boldsymbol{\theta}) \\ &\quad + \int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big) \cdot \boldsymbol{\xi}'_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \ . \end{aligned} \tag{258}$$

We will first try to bound $\boldsymbol{\xi}_t(\boldsymbol{\theta}) - \boldsymbol{\xi}'_t(\boldsymbol{\theta})$ as a function of η. Define $\Delta\boldsymbol{\xi}_t(\boldsymbol{\theta}) = \boldsymbol{\xi}_t(\boldsymbol{\theta}) - \boldsymbol{\xi}'_t(\boldsymbol{\theta})$. Then

$$\begin{aligned} \frac{d}{dr}\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta}) =& -\big(\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\big)\boldsymbol{\xi}_r - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta}) \\ =& -\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta}) \\ & -\big(\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\big)\boldsymbol{\xi}'_r \\ & -\big(\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\big)\Delta\boldsymbol{\xi}_r \ . \end{aligned} \tag{259}$$

Thus,

$$\begin{aligned} \Delta\boldsymbol{\xi}_t(\boldsymbol{\theta}) &= e^{-(t-s)\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta}) \\ &\quad + \int_s^t e^{-(t-r)\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\big(\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\big)\boldsymbol{\xi}'_r(\boldsymbol{\theta})dr \\ &\quad + \int_s^t e^{-(t-r)\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\big(\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\big)\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta})dr \ . \end{aligned} \tag{260}$$

Since $\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) - \lambda I_d$ is positive semidefinite, we first have

$$|\boldsymbol{\xi}'_r(\boldsymbol{\theta})| \leq e^{-\lambda(r-s)}|\boldsymbol{\xi}'_s(\boldsymbol{\theta})| \tag{261}$$

as well as

$$
\begin{aligned}
|\Delta\boldsymbol{\xi}_t(\boldsymbol{\theta})| \leq & \, e^{-\lambda(t-s)}|\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| \\
& + \int_s^t e^{-\lambda(t-r)}\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_r(\boldsymbol{\theta})|dr \\
& + \int_s^t e^{-\lambda(t-r)}\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta})|dr \\
\leq & \, e^{-\lambda(t-s)}|\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| \\
& + \int_s^t e^{-\lambda(t-s)}\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr \\
& + \int_s^t e^{-\lambda(t-r)}\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta})|dr \; .
\end{aligned}
\tag{262}
$$

To prepare for an application of Gronwall's inequality, we introduce a change-of-variable by defining, for $r \in [s, t]$,

$$
\overline{\Delta\boldsymbol{\xi}}_r(\boldsymbol{\theta}) = e^{\lambda(t-s)}\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta}) \; . \tag{263}
$$

Then we can rewrite the equation above as

$$
\begin{aligned}
|\overline{\Delta\boldsymbol{\xi}}_t(\boldsymbol{\theta})| = & \, e^{\lambda(r-s)}|\Delta\boldsymbol{\xi}_t(\boldsymbol{\theta})| \\
\leq & \, |\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr \\
& + \int_s^t e^{\lambda(r-s)}\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\Delta\boldsymbol{\xi}_r(\boldsymbol{\theta})|dr \\
\leq & \, |\overline{\Delta\boldsymbol{\xi}}_s(\boldsymbol{\theta})| + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr \\
& + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\overline{\Delta\boldsymbol{\xi}}_r(\boldsymbol{\theta})|dr \; .
\end{aligned}
\tag{264}
$$

Thus, by Gronwall's inequality,

$$
\begin{aligned}
|\overline{\Delta\boldsymbol{\xi}}_t(\boldsymbol{\theta})| \leq & \Big(|\overline{\Delta\boldsymbol{\xi}}_s(\boldsymbol{\theta})| + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr\Big) \\
& \times e^{\int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\|dr} \; ,
\end{aligned}
\tag{265}
$$

or, back in the original variable that we are interested in,

$$
\begin{aligned}
|\Delta\boldsymbol{\xi}_t(\boldsymbol{\theta})| \leq & \Big(|\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr\Big) \\
& \times e^{-\lambda(t-s) + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\|dr} \; .
\end{aligned}
\tag{266}
$$

Now, we have

$$
\begin{aligned}
& |\Delta\Gamma_{t,s}\eta(\boldsymbol{x})| \\
\leq & \| \int_D \nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x})^\intercal \cdot \Delta\boldsymbol{\xi}_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\|_{\hat{\nu}} \\
& + \| \int_D \big(\nabla\varphi(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big)^\intercal \boldsymbol{\xi}'_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta})\|_{\hat{\nu}} \\
\leq & C_{\nabla\varphi}\int_D |\Delta\boldsymbol{\xi}_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta}) + C_{\nabla\nabla\varphi}\int_D |\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})||\boldsymbol{\xi}'_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta}) \\
\leq & C_{\nabla\varphi}e^{-\lambda(t-s)}\int_D \Big(|\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| + \int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|dr\Big) \\
& \qquad e^{\int_s^t \|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\|dr}\mu_0(d\boldsymbol{\theta}) \\
& + C_{\nabla\nabla\varphi}e^{-\lambda(t-s)}\int_D |\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})||\boldsymbol{\xi}'_s(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta})\ .
\end{aligned} \tag{267}
$$

Note that we have,

$$
|\boldsymbol{\xi}'_s(\boldsymbol{\theta})| = |\int_\Omega \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}')\eta(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')| \leq C_{\nabla\varphi}\sup_{1\leq p\leq P}|\eta(\boldsymbol{x}_p)| \leq P^{\frac{1}{2}}C_{\nabla\varphi}\|\eta\|_{\hat{\nu}}, \tag{268}
$$

$$
\begin{aligned}
|\Delta\boldsymbol{\xi}_s(\boldsymbol{\theta})| &= |\int_\Omega \big(\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}) - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x})\big)\eta(\boldsymbol{x}')\hat{\nu}(d\boldsymbol{x}')| \\
&\leq \int_\Omega |\nabla\varphi(\boldsymbol{\Theta}_s(\boldsymbol{\theta}), \boldsymbol{x}') - \nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}), \boldsymbol{x}')||\eta(\boldsymbol{x}')|\hat{\nu}(d\boldsymbol{x}') \\
&\leq L^{\frac{1}{2}}C_{\nabla\nabla\varphi}|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})|\|\eta\|_{\hat{\nu}}
\end{aligned} \tag{269}
$$

and, since $\nabla\nabla V_r(\boldsymbol{\theta}) = \int_\Omega \nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})(f_r(\boldsymbol{x}) - f_*(\boldsymbol{x}))\hat{\nu}(d\boldsymbol{x})$ and $\nabla\nabla V_\infty(\boldsymbol{\theta}) = \int_\Omega \nabla\nabla\varphi(\boldsymbol{\theta}, \boldsymbol{x})(f_\infty(\boldsymbol{x}) - f_*(\boldsymbol{x}))\hat{\nu}(d\boldsymbol{x})$,

$$
\begin{aligned}
&\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\| \\
\leq &\|\nabla\nabla V_r(\boldsymbol{\Theta}_r(\boldsymbol{\theta})) - \nabla\nabla V_r(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\| \\
&+ \|\nabla\nabla V_r(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta})) - \nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))\| \\
\leq & L_{\nabla\nabla\varphi}C_\varphi|\Delta\boldsymbol{\Theta}_r(\boldsymbol{\theta})| + C_{\nabla\nabla\varphi}\|\Delta f_r\|_{\hat{\nu},\infty},
\end{aligned} \tag{270}
$$

where we use $\|f\|_{\hat{\nu},\infty}$ to denote $\sup_{\boldsymbol{x}\in\operatorname{supp}\hat{\nu}}|f(\boldsymbol{x})|$ and we defined $\Delta f_t = f_t - f_\infty$.

As a result, we have

$$
\begin{aligned}
& \|\Delta\Gamma_{t,s}\|_{\hat{\nu}} \\
\leq & \|\eta\|_{\hat{\nu}}^{-1}\|\Delta\Gamma_{t,s}\eta\|_{\hat{\nu}} \\
\leq & C_{\nabla\varphi}e^{-\lambda(t-s)}\int_D \Big(C_{\nabla\nabla\varphi}|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + \int_s^t \big(L_{\nabla\nabla\varphi}C_\varphi|\Delta\boldsymbol{\Theta}_r(\boldsymbol{\theta})| + C_{\nabla\nabla\varphi}\|\Delta f_r\|_{\hat{\nu},\infty}\big)C_{\nabla\varphi}dr\Big) \\
& \qquad \times e^{\int_s^t L_{\nabla\nabla\varphi}C_\varphi|\Delta\boldsymbol{\Theta}_r(\boldsymbol{\theta})| + C_{\nabla\nabla\varphi}\|\Delta f_r\|_{\hat{\nu},\infty}dr}\mu_0(d\boldsymbol{\theta}) \\
& + C_{\nabla\nabla\varphi}e^{-\lambda(t-s)}\int_D C_{\nabla\varphi}|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta})\ .
\end{aligned} \tag{271}
$$

Therefore, using C_0, C_1, etc. to represent constants that depend on C_φ, $C_{\nabla\varphi}$, $C_{\nabla\nabla\varphi}$, $C_{\nabla\nabla\varphi}$ and

$L_{\nabla\nabla\varphi}$, we have

$$
\begin{aligned}
&\|\Delta\Gamma_{t,s}\|_{\hat{\nu}} \\
&\le C_0 e^{-\lambda(t-s)}\Big(\int_D |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta}) + \int_D |\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| e^{C_1\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \|\Delta f_r\|_{\hat{\nu},\infty} dr}\mu_0(d\boldsymbol{\theta}) \\
&\qquad + \int_D \big(\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \|f_r - f_\infty\|_{\hat{\nu},\infty} dr\big) e^{C_1\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \|\Delta f_r\|_{\hat{\nu},\infty} dr}\mu_0(d\boldsymbol{\theta})\Big) \,.
\end{aligned}
\tag{272}
$$

Note that $\|\Delta f_r\|_{\hat{\nu},\infty}$ can be further upper-bounded by $C_\varphi \int_D |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta})$. Furthermore, defining

$$
\overline{\Delta\mathbf{\Theta}_t} = \int_D |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta}) \tag{273}
$$

we can write the bound above as

$$
\begin{aligned}
\|\Delta\Gamma_{t,s}\|_{\hat{\nu}} \le{}& C_0 e^{-\lambda(t-s)}\Big(\int_D |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})|\mu_0(d\boldsymbol{\theta}) + \int_D |\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| e^{C_1\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \overline{\Delta\mathbf{\Theta}_r} dr}\mu_0(d\boldsymbol{\theta}) \\
&+ \int_D \big(\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \overline{\Delta\mathbf{\Theta}_r} dr\big) e^{C_1\int_s^t |\Delta\mathbf{\Theta}_r(\boldsymbol{\theta})| + \overline{\Delta\mathbf{\Theta}_r} dr}\mu_0(d\boldsymbol{\theta})\Big) \,.
\end{aligned}
\tag{274}
$$

Finally, let

$$
U_t(\boldsymbol{\theta}) = \int_t^\infty |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})| dt \tag{275}
$$

and

$$
\bar{U}_t = \int_D U_t(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) = \int_t^\infty \overline{\Delta\mathbf{\Theta}_t} dt \,. \tag{276}
$$

Then there is

$$
\begin{aligned}
\|\Delta\Gamma_{t,s}\|_{\hat{\nu}} &\le C_0 e^{-\lambda(t-s)} \int_D \Big(|\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})| + \big(|\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta}) + \bar{U}_s\big) e^{C_1(U_s(\boldsymbol{\theta}) + \bar{U}_s)}\Big) \mu_0(d\boldsymbol{\theta}) \\
&\le 2C_0 e^{-\lambda(t-s)} \int_D \Big(|\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})| + (|\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})) e^{C_1(U_s(\boldsymbol{\theta}) + \bar{U}_s)}\Big) \mu_0(d\boldsymbol{\theta}) \,.
\end{aligned}
\tag{277}
$$

(End of the proof of Lemma D.8.) □

Proof of Lemma D.7: Lemma D.8 entails that, $\exists C, C' > 0$ such that

$$
\begin{aligned}
\|\Delta\Gamma_{t,s}\|_{\hat{\nu}}^2 &\le C e^{-2\lambda(t-s)} \bigg(\int_D \Big(|\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})| + \Big(|\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})\Big) e^{C'(U_s(\boldsymbol{\theta}) + \bar{U}_s)}\Big)\mu_0(d\boldsymbol{\theta})\bigg)^2 \\
&\le 4C e^{-2\lambda(t-s)} \int_D |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})|^2 + \Big(|\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})|^2 + U_s(\boldsymbol{\theta})^2\Big) e^{2C'(U_s(\boldsymbol{\theta}) + \bar{U}_s)}\mu_0(d\boldsymbol{\theta}) \\
&\le 4C|D'| e^{-2\lambda(t-s)} \int_D |\Delta\mathbf{\Theta}_t(\boldsymbol{\theta})| + \Big(|\Delta\mathbf{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big) e^{2C'(U_s(\boldsymbol{\theta}) + \bar{U}_s)}\mu_0(d\boldsymbol{\theta}),
\end{aligned}
\tag{278}
$$

where for the last inequality, we assume that $|D'| \ge 1$ (or, to accommodate the more general case, just replace $|D'|$ by $\max\{|D'|, 1\}$).

To prove Lemma D.7, the first goal is to show

$$
\lim_{t_0 \to \infty} \int_{t_0}^\infty \int_{t_0}^t \|\Delta\Gamma_{t,s}\|_{\hat{\nu}}^2 ds dt = 0 \,. \tag{279}
$$

There is

$$
\begin{aligned}
&\int_{t_0}^{\infty}\int_{t_0}^{t}\|\Delta\Gamma_{t,s}\|_{\hat\nu}^2 dsdt \\
\leq& 4C|D'|\int_D\int_{t_0}^{\infty}\int_{t_0}^{t} e^{-2\lambda(t-s)}\Big(|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| + \Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}\Big) dsdt\mu_0(d\boldsymbol{\theta}) \\
\leq& 4C|D'|\int_D\Big(\int_{t_0}^{\infty}\Big(\int_{t_0}^{t} e^{-2\lambda(t-s)}ds\Big)|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})|dt \\
&\qquad + \int_{t_0}^{\infty}\Big(\int_{s}^{\infty} e^{-2\lambda(t-s)}dt\Big)\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\Big)\mu_0(d\boldsymbol{\theta}) \\
\leq& 2C|D'|\lambda^{-1}\int_D\Big(\int_{t_0}^{\infty}|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})|dt + \int_{t_0}^{\infty}\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\Big)\mu_0(d\boldsymbol{\theta}) \\
\leq& 4C|D'|\lambda^{-1}\int_D\int_{t_0}^{\infty}\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\mu_0(d\boldsymbol{\theta})\ .
\end{aligned}
\tag{280}
$$

By our assumption, the RHS is finite for $t_0 > 0$. Hence, by taking t_0 large enough, the value of $\int_{t_0}^{\infty}\int_{t_0}^{t}\|\Delta\Gamma_{t,s}\|_{\hat\nu}^2 dsdt$ can be made arbitrarily close to zero.

The second goal is to show that $\forall t_0 > 0$,

$$
\lim_{T\to\infty}\fint_{t_0}^{T}\int_0^{t_0}\|\Gamma_{t,s}\|^2 dsdt = 0\ . \tag{281}
$$

As a first step, we show that

$$
\lim_{T\to\infty}\fint_{t_0}^{T}\int_0^{t_0}\|\Gamma_{t-s}^{\infty}\|_{\hat\nu}^2 dsdt = 0 \tag{282}
$$

because $\forall\eta\in\mathcal{W}_L(\Omega)$, there is

$$
\begin{aligned}
|\langle\eta,\Gamma_{t-s}^{\infty}\eta\rangle_{\hat\nu}| =& \int_D\Big\langle\boldsymbol{b}(\boldsymbol{\theta}), e^{-t\nabla\nabla V_\infty(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}))}\boldsymbol{b}(\boldsymbol{\theta})\Big)\Big\rangle\mu_0(d\boldsymbol{\theta}) \\
\leq& e^{-\lambda(t-s)}\int_D|\boldsymbol{b}(\boldsymbol{\theta})|^2\mu_0(d\boldsymbol{\theta}) \\
\leq& e^{-\lambda(t-s)}\|\mathcal{M}_\infty\|_{\hat\nu}\|\eta\|_{\hat\nu}^2\ ,
\end{aligned}
\tag{283}
$$

where

$$
\boldsymbol{b}(\boldsymbol{\theta}) = \int_\Omega\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}),\boldsymbol{x})\eta(\boldsymbol{x})\hat\nu(d\boldsymbol{x}). \tag{284}
$$

and $\mathcal{M}_\infty$ is defined as $\mathcal{M}_\infty := \mathcal{B}_\infty^\intercal\mathcal{B}_\infty$, or concretely, for $\eta\in\mathcal{W}_L(\omega)$,

$$
\begin{aligned}
(\mathcal{M}_\infty\eta)(\boldsymbol{x}) :=& \int_\Omega\Big(\int_D\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'),\boldsymbol{x})^\intercal\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'),\boldsymbol{x}')\mu_0(d\boldsymbol{\theta}')\Big)\eta(\boldsymbol{x}')\hat\nu(d\boldsymbol{x}') \\
=& \int_\Omega M(\boldsymbol{x},\boldsymbol{x}',\mu_\infty)\eta(\boldsymbol{x}')\hat\nu(d\boldsymbol{x}')\ ,
\end{aligned}
\tag{285}
$$

where

$$
M(\boldsymbol{x},\boldsymbol{x}',\mu_\infty) := \int_D\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'),\boldsymbol{x})\cdot\nabla\varphi(\boldsymbol{\Theta}_\infty(\boldsymbol{\theta}'),\boldsymbol{x}')\mu_0(d\boldsymbol{\theta}')\ . \tag{286}
$$

In the ERM setting, $\mathcal{M}_\infty$ is effectively an $L \times L$ matrix. Thus,

$$\begin{aligned}\fint_{t_0}^{T}\int_0^{t_0}\|\Gamma^\infty_{t-s}\|^2_{\hat\nu}dsdt \leq &\fint_{t_0}^{T}\int_0^{t_0} e^{-2\lambda(t-s)}\|\mathcal{M}_\infty\|^2_{\hat\nu}dsdt \\ \leq &\|\mathcal{M}_\infty\|^2_{\hat\nu}\fint_{t_0}^{T} e^{-2\lambda(t-t_0)}dt \to 0 \qquad \text{as} \quad T\to\infty\end{aligned} \tag{287}$$

Hence, it is sufficient to show that

$$\lim_{T\to\infty}\fint_{t_0}^{T}\int_0^{t_0}\|\Delta\Gamma_{t,s}\|^2 dsdt = 0\,. \tag{288}$$

We have

$$\begin{aligned}&\int_{t_0}^{T}\int_0^{t_0}\|\Delta\Gamma_{t,s}\|^2 dsdt \\ \leq &4C|D'|\int_D\Big(\int_{t_0}^{T}\Big(\int_0^{t_0} e^{-2\lambda(t-s)}ds\Big)|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})|dt \\ &\quad + \int_0^{t_0}\Big(\int_{t_0}^{T} e^{-2\lambda(t-s)}dt\Big)\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\Big)\mu_0(d\boldsymbol{\theta}) \\ \leq &2C|D'|\lambda^{-1}\int_D\Big(\int_{t_0}^{T} e^{-2\lambda(t-t_0)}|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})|dt \\ &\quad + \int_0^{t_0} e^{-2\lambda(t_0-s)}\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\Big)\mu_0(d\boldsymbol{\theta}) \\ \leq &4C|D'|\lambda^{-1}\int_D\int_0^{\infty}\Big(|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| + U_s(\boldsymbol{\theta})^2\Big)e^{2C'(U_s(\boldsymbol{\theta})+\bar U_s)}ds\mu_0(d\boldsymbol{\theta}) \\ < &\infty\end{aligned} \tag{289}$$

by assumption (55). Therefore,

$$\fint_{t_0}^{T}\int_0^{t_0}\|\Delta\Gamma_{t,s}\|^2_{\hat\nu}dsdt = \frac{1}{T-t_0}\int_{t_0}^{T}\int_0^{t_0}\|\Delta\Gamma_{t,s}\|^2_{\hat\nu}dsdt \xrightarrow[T\to\infty]{} 0\,. \tag{290}$$

This concludes the proof of Lemma D.7. □

D.5.2 Interpretation of the Assumption (55)

Below, we will illustrate the assumption (55)

$$Q := \int_D\int_0^{\infty}\big(|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| + U_t(\boldsymbol{\theta})^2\big)e^{C_1(U_t(\boldsymbol{\theta})+\bar U_t)}dt\mu_0(d\boldsymbol{\theta}) < \infty, \tag{291}$$

in Theorem 3.10 by giving examples that satisfy this condition.

First, consider an example where $\exists\kappa > 0, \alpha > 1$ such that $\forall\boldsymbol{\theta}\in\operatorname{supp}\mu_0$ and $\forall\, t > 0$,

$$|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| < \kappa(t+1)^{-\alpha}, \tag{292}$$

that is, all characteristic flows share a uniform asymptotic convergence rate on the order of $t^{-\alpha}$. Then $\forall\boldsymbol{\theta}\in\operatorname{supp}\mu_0$,

$$U_t(\boldsymbol{\theta}) = \int_t^{\infty}|\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})|ds \leq \frac{\kappa}{\alpha-1}(t+1)^{-(\alpha-1)} \tag{293}$$

and thus

$$\bar{U}_t \leq \frac{\kappa}{\alpha-1}(t+1)^{-(\alpha-1)} . \tag{294}$$

Therefore,

$$\begin{aligned} Q &\leq \int_D \int_0^\infty \left(|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| + U_t(\boldsymbol{\theta})^2\right) e^{C_1(U_0(\boldsymbol{\theta})+\bar{U}_0)} dt \mu_0(d\boldsymbol{\theta}) \\ &\leq \int_0^\infty \Big(\kappa(t+1)^{-\alpha} + \big(\frac{\kappa}{\alpha-1}\big)^2 (t+1)^{-2(\alpha-1)}\Big) e^{\frac{2C_1\kappa}{\alpha-1}} dt, \end{aligned} \tag{295}$$

which is finite as long as $\alpha > \frac{3}{2}$. Thus,

Proposition D.9. *If* $\exists \kappa > 0, \alpha > \frac{3}{2}$ *such that* $\forall \boldsymbol{\theta} \in \operatorname{supp} \mu_0$ *and* $\forall t \geq 0$,

$$|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| = |\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})| < \kappa(t+1)^{-\alpha}, \tag{296}$$

then the condition (55) *is satisfied.*

Moreover, the assumption allows flexibility in having non-uniform convergence rate for different characteristic flows, $\boldsymbol{\Theta}_t(\boldsymbol{\theta})$. Suppose that $\exists \kappa : \operatorname{supp} \mu_0 \to \mathbb{R}_+$ and $\alpha > \frac{3}{2}$ such that $\forall \theta \in \operatorname{supp} \mu_0$,

$$|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| < \kappa(\boldsymbol{\theta})(t+1)^{-\alpha} . \tag{297}$$

Then

$$U_t(\boldsymbol{\theta}) = \int_t^\infty |\Delta\boldsymbol{\Theta}_s(\boldsymbol{\theta})| ds \leq \frac{\kappa}{\alpha-1}(t+1)^{-(\alpha-1)} \tag{298}$$

and so

$$\begin{aligned} Q &\leq \int_D \int_0^\infty \left(|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| + U_t(\boldsymbol{\theta})^2\right) e^{2C_1(U_0(\boldsymbol{\theta}))} dt \mu_0(d\boldsymbol{\theta}) \\ &\leq \int_D \int_0^\infty \Big(\kappa(\boldsymbol{\theta})(t+1)^{-\alpha} + \big(\frac{\kappa(\boldsymbol{\theta})}{\alpha-1}\big)^2 (t+1)^{-2(\alpha-1)}\Big) e^{\frac{2C_1\kappa(\boldsymbol{\theta})}{\alpha-1}} dt \\ &\leq C_2 \int_D \left(\kappa(\boldsymbol{\theta}) + \kappa(\boldsymbol{\theta})^2\right) e^{\frac{2C_1\kappa(\boldsymbol{\theta})}{\alpha-1}} \mu_0(d\boldsymbol{\theta}) . \end{aligned} \tag{299}$$

Therefore,

Proposition D.10. *Suppose* $\exists \alpha > \frac{3}{2}$ *and a function* $\kappa : \operatorname{supp} \mu_0 \to \mathbb{R}_+$, *which satisfies*

$$\int_D \Big(\kappa(\boldsymbol{\theta}) + \kappa(\boldsymbol{\theta})^2\Big) e^{\frac{2C_1\kappa(\boldsymbol{\theta})}{\alpha-1}} \mu_0(d\boldsymbol{\theta}) < \infty, \tag{300}$$

such that $\forall \boldsymbol{\theta} \in \operatorname{supp} \mu_0$,

$$|\Delta\boldsymbol{\Theta}_t(\boldsymbol{\theta})| = |\boldsymbol{\Theta}_t(\boldsymbol{\theta}) - \boldsymbol{\Theta}_\infty(\boldsymbol{\theta})| \leq \kappa(\boldsymbol{\theta})(t+1)^{-\alpha} . \tag{301}$$

Then the condition (55) *is satisfied.*

E Properties of the Minimizers of the Regularized Loss

We prove Proposition 3.14, which we extend into:

Proposition E.1. *Under Assumption 2.2, the minimum of the loss $\mathcal{L}(\mu)$ defined in (5) is unique and can only be attained at minimizers $\mu_\lambda \in \mathcal{P}(D)$ that are each of the form*

$$\mu_\lambda(dc, d\boldsymbol{z}) = \delta_{c_\lambda}(dc)\hat{\mu}_+(d\boldsymbol{z}) + \delta_{-c_\lambda}(dc)\hat{\mu}_-(d\boldsymbol{z}) \tag{302}$$

where $c_\lambda \geq 0$ and $\hat{\mu}_\pm \in \mathcal{P}(\hat{D})$ satisfy

$$\begin{aligned}
\forall \boldsymbol{z} \in \operatorname{supp}\hat{\mu}_- \quad &: \quad -\hat{F}(\boldsymbol{z}) + c_\lambda \int_{\hat{D}} \hat{K}(\boldsymbol{z}, \boldsymbol{z}') \left(\hat{\mu}_+(d\boldsymbol{z}') - \hat{\mu}_-(d\boldsymbol{z}')\right) = \lambda c_\lambda, \\
\forall \boldsymbol{z} \in \operatorname{supp}\hat{\mu}_+ \quad &: \quad -\hat{F}(\boldsymbol{z}) + c_\lambda \int_{\hat{D}} \hat{K}(\boldsymbol{z}, \boldsymbol{z}') \left(\hat{\mu}_+(d\boldsymbol{z}') - \hat{\mu}_-(d\boldsymbol{z}')\right) = -\lambda c_\lambda, \\
\forall \boldsymbol{z} \in \hat{D} \quad &: \quad \left| -\hat{F}(\boldsymbol{z}) + c_\lambda \int_{\hat{D}} \hat{K}(\boldsymbol{z}, \boldsymbol{z}') \left(\hat{\mu}_+(d\boldsymbol{z}') - \hat{\mu}_-(d\boldsymbol{z}')\right) \right| \leq \lambda c_\lambda.
\end{aligned} \tag{303}$$

In addition, the constant c_λ is unique and positive if $F(\boldsymbol{z})$ is not identically zero on $\hat{D}$, the closure of the supports of $\hat{\mu}_\pm$ are disjoint, $\overline{\operatorname{supp}\hat{\mu}_+} \cap \overline{\operatorname{supp}\hat{\mu}_-} = \emptyset$, and the function

$$f_\lambda = \int_D c\hat{\varphi}(\boldsymbol{z}, \cdot)\mu_\lambda(dc, d\boldsymbol{z}) = c_\lambda \int_{\hat{D}} \hat{\varphi}(\boldsymbol{z}, \cdot)\left(\hat{\mu}_+(d\boldsymbol{z}) - \hat{\mu}_-(d\boldsymbol{z})\right) \tag{304}$$

is unique and satisfies

$$\tfrac{1}{4}\lambda^2 |c_\lambda|^2 \hat{K}_M^{-1} \leq \|f_* - f_\lambda\|_{\hat{\nu}}^2, \qquad \|f_* - f_\lambda\|_{\hat{\nu}}^2 + \lambda |c_\lambda|^2 \leq \lambda |\gamma_1(f_*)|^2. \tag{305}$$

where $\hat{K}_M = \max_{\boldsymbol{z}\in\hat{D}} \|\hat{\varphi}(\boldsymbol{z}, \cdot)\|_{\hat{\nu}}^2 = \max_{\boldsymbol{z}\in\hat{D}} \hat{K}(\boldsymbol{z}, \boldsymbol{z})$.

Remark E.2. *Note that the proposition automatically implies that $\gamma_1(f_\lambda) \leq \gamma_1(f_*) < \infty$. It also implies that*

$$\int_D |c|^q \mu_\lambda(dc, d\boldsymbol{z}) = |c_\lambda|^q = |\gamma_\lambda|_{TV}^q \leq |\gamma_1(f_*)|^q \qquad \forall q \in \mathbb{R}_+ \tag{306}$$

where $\gamma_\lambda = \int_{\mathbb{R}} c\mu_\lambda(dc, \cdot)$. Finally note that the proposition holds if we replace the empirical loss by the population loss.

Proof: The uniqueness of the minimum of the regularized energy (5) follows from its convexity in the linear topology; the fact that this energy can only be minimized by minimizers follows from the compactness of the sets $\{\mu \in \mathcal{P}(D) \ : \ \mathcal{E}_\lambda[\mu] \leq u, u \in \mathbb{R}\}$. The minimizers of $\mathcal{E}_\lambda[\mu]$ must satisfy the following Euler-Lagrange equations [57]:

$$\forall (c, \boldsymbol{z}) \in D \quad : \quad -c\hat{F}(\boldsymbol{z}) + c\int_D c'\hat{K}(\boldsymbol{z}, \boldsymbol{z}')\mu(dc', d\boldsymbol{z}') + \tfrac{1}{2}\lambda|c|^2 \equiv c\hat{V}(\boldsymbol{z}) + \tfrac{1}{2}\lambda|c|^2 \geq \bar{V}\ , \tag{307}$$

with equality on the support of μ and where $\bar{V}$ is the expectation of the left hand side with respect to $\mu(dc, d\boldsymbol{z})$. Minimizing the left hand side of (307) over c at fixed $\boldsymbol{z}$, we deduce that

$$\forall \boldsymbol{z} \in \hat{D} \quad : \quad \min_c \left(c\hat{V}(\boldsymbol{z}) + \tfrac{1}{2}\lambda|c|^2\right) \geq \bar{V}\ , \tag{308}$$

with equality for $\boldsymbol{z}$ in the support of $\hat{\mu} = \int_{\mathbb{R}} \mu(dc, \cdot)$. This means that for any $\boldsymbol{z} \in \operatorname{supp}\hat{\mu}$, there can only be one $c = c(\boldsymbol{z})$ in $\operatorname{supp}\mu$, with $c(\boldsymbol{z})$ satisfying the Euler-Lagrange equation associated with (308)

$$\hat{V}(\boldsymbol{z}) + \lambda c(\boldsymbol{z}) = 0 \quad \Leftrightarrow \quad \hat{V}(\boldsymbol{z}) = -\lambda c(\boldsymbol{z}) \tag{309}$$

If we insert this equality back in $c(\boldsymbol{z})\hat{V}(\boldsymbol{z}) + \frac{1}{2}\lambda|c(\boldsymbol{z})|^2 = \bar{V}$, we deduce that $|c(\boldsymbol{z})| = c_\lambda$, with the constant c_λ related to $\bar{V}$ as

$$-\frac{1}{2}\lambda|c_\lambda|^2 = \bar{V}\ , \tag{310}$$

and furthermore, $\forall \boldsymbol{z} \in \operatorname{supp}\hat{\mu}$,

$$\hat{V}(\boldsymbol{z}) = \begin{cases} -\lambda c_\lambda & \text{if } c(\boldsymbol{z}) = c_\lambda \\ \lambda c_\lambda & \text{if } c(\boldsymbol{z}) = -c_\lambda \end{cases} . \tag{311}$$

These considerations imply that the minimizer must be of the form (302), and if we combine (308) and (310) and evaluate the minimum on c explicitly we deduce that $\hat{\mu}_\pm$ and c_λ must satisfy the equations in (303). It is also clear from (303) that we must have $\overline{\operatorname{supp}\hat{\mu}_+} \cap \overline{\operatorname{supp}\hat{\mu}_-} = \emptyset$: indeed if there was a point $\boldsymbol{z} \in \overline{\operatorname{supp}\hat{\mu}_+} \cap \overline{\operatorname{supp}\hat{\mu}_-}$, then at that point $\hat{V}(\boldsymbol{z})$ would be discontinuous, which is not possible since this function is continuously differentiable for any μ by our assumptions on $\hat{\varphi}$. Finally, to show that we must have that $c_\lambda > 0$ if $F(\boldsymbol{z})$ is not identically zero on $\hat{D}$, note that if $c_\lambda = 0$, (307) reduces to

$$\forall (c, \boldsymbol{z}) \in D \quad : \quad -c\hat{F}(\boldsymbol{z}) + \tfrac{1}{2}\lambda|c|^2 \geq 0 \tag{312}$$

which can only be satisfied if $\hat{F}(\boldsymbol{z}) = 0$.

To show that c_λ and the function in (304) are unique, let μ_λ and μ'_λ be two different minimizers and consider

$$f_\lambda = \int_D c\hat{\varphi}(\boldsymbol{z}, \cdot)\mu_\lambda(dc, d\boldsymbol{z}) \quad \text{and} \quad f'_\lambda = \int_D c\hat{\varphi}(\boldsymbol{z}, \cdot)\mu'_\lambda(dc, d\boldsymbol{z}) \tag{313}$$

Let us evaluate the energy on $a\mu_\lambda + (1-a)\mu'_\lambda \in \mathcal{P}(D)$ with $a \in [0,1]$. By convexity of $\mathcal{E}_\lambda$ we have

$$\mathcal{L}(a\mu_\lambda + (1-a)\mu'_\lambda) \leq a\mathcal{L}(\mu_\lambda) + (1-a)\mathcal{L}(\mu'_\lambda) = \mathcal{L}(\mu_\lambda) = \mathcal{L}(\mu'_\lambda) \tag{314}$$

Since $a\mu_\lambda + (1-a)\mu'_\lambda$ cannot have a lower energy than this minimum, we must have equality in (314), which reduces to

$$\begin{aligned} &\left\| f_* - af_\lambda - (1-a)f'_\lambda \right\|^2_{\hat{\nu}} + a\lambda|c_\lambda|^2 + (1-a)\lambda|c'_\lambda|^2 \\ &= \|f_* - f_\lambda\|^2_{\hat{\nu}} + \lambda|c_\lambda|^2 \\ &= \left\|f_* - f'_\lambda\right\|^2_{\hat{\nu}} + \lambda|c'_\lambda|^2\ , \end{aligned} \tag{315}$$

where c_λ and c'_λ are associated with μ_λ and μ'_λ, respectively. Clearly these equations can only be fulfilled for all $a \in [0,1]$ if $c_\lambda = c'_\lambda$ and $f_\lambda = f'_\lambda$ $\hat{\nu}$-a.e. on Ω.

To establish (305), notice that if μ_λ is a minimizer and f_λ is given by (304), then we can derive from (311) that

$$-\int_\Omega f_\lambda(\boldsymbol{x})f_*(\boldsymbol{x})\hat{\nu}(d\boldsymbol{x}) + \|f_\lambda\|^2_{\hat{\nu}} + \lambda|c_\lambda|^2 = 0. \tag{316}$$

This gives, using Cauchy-Schwartz,

$$\lambda|c_\lambda|^2 = \int_\Omega f_\lambda(\boldsymbol{x})(f_*(\boldsymbol{x}) - f_\lambda(\boldsymbol{x}))\hat{\nu}(d\boldsymbol{x}) \leq \|f_\lambda\|_{\hat{\nu}}\, \|f_* - f_\lambda\|_{\hat{\nu}}\ . \tag{317}$$

Now notice that

$$\|f_\lambda\|^2_{\hat{\nu}} = c^2_\lambda \int_{\hat{D}\times\hat{D}} \hat{K}(\boldsymbol{z}, \boldsymbol{z}')\left(\hat{\mu}_+(d\boldsymbol{z}) - \hat{\mu}_-(d\boldsymbol{z})\right)\left(\hat{\mu}_+(d\boldsymbol{z}') - \hat{\mu}_-(d\boldsymbol{z}')\right) \leq 4c^2_\lambda \hat{K}_M\ . \tag{318}$$

Using (318) in (317) and reorganizing gives the first inequality in (305). To establish the second, let $\mu_* \in \mathcal{M}_+(D)$ be the measure that minimizes $\int_D |c|\mu(dc, d\boldsymbol{z})$ under the constraint that $f_* =$

$\int_D c\hat{\varphi}(\boldsymbol{z},\cdot)\mu_*(dc,d\boldsymbol{z})$, so that $\int_D |c|\mu_*(dc,d\boldsymbol{z}) = \gamma_1(f_*)$—the measure μ_* exists since we assumed that $f_* \in \mathcal{F}_1$. Evaluated on μ_*, the loss is

$$\mathcal{L}(\mu_*) = \lambda|\gamma_1(f_*)|^2. \tag{319}$$

Any minimizer μ_λ of $\mathcal{L}(\mu)$ must do at least as well, i.e we must have

$$\|f_* - f_\lambda\|^2_{\hat{\nu}} + \lambda \int_D |c|^2 \mu_\lambda(dc,d\boldsymbol{z}) = \|f_* - f_\lambda\|^2_{\hat{\nu}} + \lambda|c_\lambda|^2 \le \lambda|\gamma(f_*)|^2. \tag{320}$$

This establishes the second inequality in (305). □

F Experimental Setup

The code is implemented in C++ and run on a cluster with single CPU. We take both Ω and $\hat{D}$ to be unit spheres of dimension $d = 16$. Both the student network and the teacher network are of the form (1), with $\hat{\varphi}(\boldsymbol{z},\boldsymbol{x}) = \max(0,\langle\boldsymbol{z},\boldsymbol{x}\rangle)$ being the ReLU nonlinearity without bias. The teacher network has two neurons, $(c_1,\boldsymbol{z}_1)$ and $(c_2,\boldsymbol{z}_2)$, in the hidden layer, with $c_1 = c_2 = 1$ and $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ sampled i.i.d. from the uniform distribution on $\hat{D}$ and then fixed across the experiments. We vary the width of the student network in the range of $n = 128, 256, 512, 1024$ and 2048, with their $\boldsymbol{z}$'s sampled i.i.d. from the uniform distribution on $\hat{D}$ and their c's sampled i.i.d. from $\mathcal{N}(0,1)$. The models are trained by performing SGD on the exact population loss, where the loss and its gradients are computed analytically, both of which we rescale by d in order to adjust to the $\frac{1}{d}$ factor resulting from spherical integrals. The models are trained for 20000 epochs with learning rate (which is multiplied to the RHS of (10)) set to be 1.

For $n = 128, 256, 512, 1024$ and 2048, the approximate running times of each run of 20000 epochs are 1 min, 2 min, 8 min, 30 min and 150 min, respectively.

The hyperparameter in regularization, λ, is manually selected from $0.01, 0.05$ and 0.1. We select 0.01 because it yields the best loss value (unsurprisingly) while also controlling the TV norm and the 2-norm of the solution. Figure 2 shows the same plots as the second row of Figure 1 but with $\lambda = 0.05$ and $\kappa = 10$ runs.

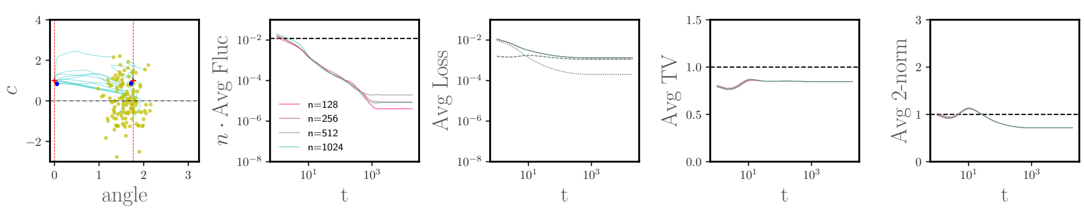

Figure 2: Same setup as the second row of Figure 1, except for $\lambda = 0.05$ instead of 0.01.

G Analytical Calculation of the Resampling Error

Derivations to similar to the following can be found in [5, 15, 53]. In the setting of ReLU without bias on unit sphere, we take $\hat{D} = \Omega = \mathbb{S}^d \subseteq \mathbb{R}^{d+1}$, $\hat{\varphi}(\boldsymbol{z},\boldsymbol{x}) = \max(\langle\boldsymbol{z},\boldsymbol{x}\rangle, 0)$, and ν is equal to the uniform measure on $\mathbb{S}^d$. In this case,

$$\hat{K}(\boldsymbol{z},\boldsymbol{z}') = \int_\Omega \hat{\varphi}(\boldsymbol{z},\boldsymbol{x})\hat{\varphi}(\boldsymbol{z}',\boldsymbol{x})\nu(d\boldsymbol{x}) = \frac{1}{2(d+1)\pi}(\sin\alpha + (\pi-\alpha)\cos\alpha), \tag{321}$$

with α being the angle between $\boldsymbol{z}$ and $\boldsymbol{z}'$, and

$$\int_{\Omega} |\hat{\varphi}(\boldsymbol{z}, \boldsymbol{x})|^2 \nu(d\boldsymbol{x}) = \frac{1}{2} \int_{\Omega} (\langle \boldsymbol{x}, \boldsymbol{z} \rangle)^2 \nu(d\boldsymbol{x}) = \frac{1}{2(d+1)} \tag{322}$$

Thus, taking μ_* to be the measure representing the teacher network, $\mu_* = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta_{\boldsymbol{z}_i}(d\boldsymbol{z}) \delta_1(dc)$, we have

$$\begin{aligned} \int_D \|\varphi(\boldsymbol{\theta}, \cdot)\|_\nu^2 \mu_*(d\boldsymbol{\theta}) &= \int_D \int_{\Omega} |\varphi(\boldsymbol{\theta}, \boldsymbol{x})|^2 \nu(d\boldsymbol{x}) \mu_*(d\boldsymbol{\theta}) \\ &= \int_D \frac{c^2}{2(d+1)} \mu_*(d\boldsymbol{\theta}) \\ &= \frac{1}{2(d+1)} \end{aligned} \tag{323}$$

On the other hand,

$$\begin{aligned} \|f_*\|_\nu^2 &= \int_{\Omega} \Big| \int_D \varphi(\boldsymbol{\theta}, \boldsymbol{x}) \mu_*(d\boldsymbol{\theta}) \Big|^2 \nu(d\boldsymbol{x}) \\ &= \int_D \int_D cc' \hat{K}(\boldsymbol{z}, \boldsymbol{z}') \mu_*(d\boldsymbol{\theta}) \mu_*(d\boldsymbol{\theta}') \\ &= \frac{1}{n_t^2} \sum_{i,j=1}^{n_t} \hat{K}(\boldsymbol{z}_i, \boldsymbol{z}_j) \end{aligned} \tag{324}$$

In the experiments described in the main text, we take $n_t = 2$, and $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are initialized with a fixed random seed such that their angle, α_{12}, equal to 1.766. Thus,

$$\|f_*\|_\nu^2 = \frac{1}{4(d+1)\pi}(0+\pi) + \frac{1}{4(d+1)\pi}(\sin\alpha_{12} + (\pi - \alpha_{12})\cos\alpha_{12}) \approx 0.012 \tag{325}$$

Together, we get a numerical value of the RHS of (47) if we replace μ_∞, f_∞ and $\hat{\nu}$ by μ_*, f_* and ν, respectively.