

Overparameterization of Deep ResNet: Zero Loss and Mean-field Analysis

Zhiyan Ding

ZDING49@MATH.WISC.EDU

*Mathematics Department
University of Wisconsin-Madison
Madison, WI 53706 USA.*

Shi Chen

SCHEN636@WISC.EDU

*Mathematics Department
University of Wisconsin-Madison
Madison, WI 53706 USA.*

Qin Li

QINLI@MATH.WISC.EDU

*Mathematics Department
University of Wisconsin-Madison
Madison, WI 53706 USA.*

Stephen J. Wright

SWRIGHT@CS.WISC.EDU

*Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706 USA.*

Editor: Ohad Shamir

Abstract

Finding parameters in a deep neural network (NN) that fit training data is a nonconvex optimization problem, but a basic first-order optimization method (gradient descent) finds a global optimizer with perfect fit (zero-loss) in many practical situations. We examine this phenomenon for the case of Residual Neural Networks (ResNet) with smooth activation functions in a limiting regime in which both the number of layers (depth) and the number of weights in each layer (width) go to infinity. First, we use a mean-field-limit argument to prove that the gradient descent for parameter training becomes a gradient flow for a probability distribution that is characterized by a partial differential equation (PDE) in the large-NN limit. Next, we show that under certain assumptions, the solution to the PDE converges in the training time to a zero-loss solution. Together, these results suggest that the training of the ResNet gives a near-zero loss if the ResNet is large enough. We give estimates of the depth and width needed to reduce the loss below a given threshold, with high probability.

Keywords: Residual neural network, overparameterization, mean-field analysis, zero loss, gradient flow

1. Introduction

In the training of deep neural networks (NN), the loss function is nonconvex, but algorithms based on gradient descent (GD) appear to converge to a zero-loss solution that interpolates the training data. As the number of layers and the width of the NN increase to infinity,

the training problem becomes vastly overparameterized, potentially yielding a large set of interpolating solutions. Still, it remains surprising that GD approaches find the global minimizer reliably.

What is the mechanism that allows gradient descent to find the minimizer?

This article proposes a partial answer to this question for residual neural networks (ResNet), relying on three main toolboxes: a continuous limit argument, a mean-field limit argument, and gradient flow analysis. These toolboxes are used to translate gradient descent for parameter training into a partial differential equation (PDE), where PDE analysis (specifically, steady-state equilibrium analysis) is employed to trace the convergence to the minimizer.

We denote by L and M the number of layers, and the number of weights per layer, respectively. We study highly overparameterized NN in which both L and M go to infinity. As $L \rightarrow \infty$, the NN approaches infinite depth, and ResNet translates into an ordinary differential equation (ODE) with a force term encoding the parameter configuration. We refer to this limit as the *continuous limit*. As $M \rightarrow \infty$, the NN approaches infinite width, and the ODE that characterizes ResNet further translates into an ordinary-integral equation (OIE), with the integrand encoding the configuration of neuron parameters. We refer to this limit as the *mean-field limit*. In the limiting regime, the training of the parameters becomes the training of the integrand, and gradient descent on the parameters translates into a gradient flow on the integrand. The gradient flow process turns out to be characterized by a PDE. We observe that under certain assumptions, the long “time” limit of the PDE solution has zero cost, suggesting that gradient descent applied to a (finite) parameter configuration yields almost-zero-cost solutions too. The smallness of this cost depends on the largeness of L and M ; we give a precise quantification.

We describe the problem setup in Section 2, where we also formally derive the continuous and mean-field limits. We unify notations in Section 3 and present our main results in Section 4, including rigorous justifications for both limiting procedures, the result about convergence of the PDE solution to the global minimum, and a result about convergence to near-zero losses of gradient descent on large (but finite) networks. As a byproduct of our analysis, well-posedness for both the ODE formulation and the PDE formulation is obtained in Section 5.

There is a vast of literature addressing the overparameterization of DNN, its associated nonconvex optimization problem, and the application of the first order (gradient descent) optimization methods. Three major perspectives have been taken. One approach is to analyze the landscape of the objective functions, identifying properties that a nonconvex optimization problem needs to satisfy for a first order method to converge to an optimizer. Different NN structures are then analyzed to fit these properties (Jin et al., 2017; Ge et al., 2015; Du et al., 2017; Ge et al., 2018; Nguyen and Hein, 2018; Du and Lee, 2018; Soltanolkotabi et al., 2019; Nguyen and Hein, 2017; Kawaguchi, 2016; Yun et al., 2018; Safran et al., 2020; Bartlett et al., 2018b,a). The second approach largely depends on the insensibility of the Gram matrix as the number of weights goes to infinity (Allen-Zhu et al., 2019; Du et al., 2019a; Zhang et al., 2019; Chatterji et al., 2021; Zou et al., 2020; Du et al., 2019b; Frei et al., 2019; Chen et al., 2020; Li and Liang, 2018), as a way to analyze Neural Tangent Kernel (Jacot et al., 2018; Liu et al., 2020). The analysis crucially depends on the fact that GD confines the iteration in a small neighborhood if the initial conditions are random. However, as point out in (Allen-Zhu

and Li, 2019; Ba et al., 2020; Wei et al., 2019), NTK, like many other kernel models Arora et al. (2019), is rather a limited description of NN. The success of NTK highly depends on the fact that a nonlinear NN can be approximated by a linear model of the infinite dimensional random features associated with the NTK. The third approach, which we adopt in this paper, involves the mean-field limit. The idea is to let the number of weights M and/or the number of layers L go to infinity, and translate the gradient descent for the parameters in a finite dimensional space to a gradient flow in a (infinite-dimensional) function space (Araújo et al., 2019; Fang et al., 2019; Nguyen, 2019; Allen-Zhu et al., 2019; Du et al., 2019a; Zhang et al., 2019; Chatterji et al., 2021; Chizat and Bach, 2018; Mei et al., 2018; Wojtowytsch, 2020; Lu et al., 2020; Sirignano and Spiliopoulos, 2021, 2020; E et al., 2020b; Nguyen, 2019; Pham and Nguyen, 2021; Nguyen and Pham, 2021; Fang et al., 2021; Lu et al., 2020; E et al., 2020a; Jabir et al., 2021). In particular, in (E et al., 2020a), the authors have formally derived the limiting gradient flow equation for deep and overparameterized ResNet. Some PDE analysis techniques are borrowed to show the convergence of the limiting gradient flow equation. In (Chizat and Bach, 2018; Mei et al., 2018; Wojtowytsch, 2020), the authors study the two-layer (one hidden layer) NN in the regime where M is infinite. In these works, both the mean-field limit and the convergence of the gradient flow are proved, giving a complete answer about the convergence to the zero-loss when NN has two layers. The problem is significantly more complicated for NNs that have multiple layers, since the convexity of the objective function is lost. To the best of our knowledge, the current paper is the first that both rigorously proves the mean-field limit that translates the optimization problem into a gradient flow, and demonstrates why the gradient flow leads to a global minimizer.

2. ResNet and Gradient Descent

General ResNet uses the following model:

$$z_{l+1}(x) = z_l(x) + \frac{1}{ML} \sum_{m=1}^M f(z_l(x), \theta_{l,m}), \quad l = 0, 1, \dots, L-1 \quad (1)$$

$$z_0(x) = x,$$

where $x \in \mathbb{R}^d$ is the input data. The configuration of the NN is encoded in $\Theta_{L,M} = \{\theta_{l,m}\}_{l=0, m=1}^{L-1, M}$, where $\theta_{l,m} \in \mathbb{R}^k$, and $f : \mathbb{R}^d \times \mathbb{R}^k \rightarrow \mathbb{R}^d$ is a given activation function. (E et al., 2020b, Section 3.3) verifies that “conventional” ResNets are included in this framework. Define

$$z_{l+1}(x) = z_l(x) + \frac{1}{ML} \sum_{m=1}^M U_{l,m} \sigma(W_{l,m}^\top z_l(x) + b_{l,m}), \quad l = 0, 1, \dots, L-1$$

$$z_0(x) = x,$$

where $W_{l,m}, U_{l,m} \in \mathbb{R}^d$, $b_{l,m} \in \mathbb{R}$, and σ is the ReLU activation function. In this example, we have $\theta_{l,m} = (W_{l,m}, U_{l,m}, b_{l,m}) \in \mathbb{R}^k$, with $k = 2d + 1$.

Denote by $Z_{\Theta_{L,M}}(l; x)$ the solution to (1). The goal of training ResNet is to seek parameters $\Theta_{L,M}$ such that the following cost is minimized:

$$E(\Theta_{L,M}) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_{\Theta_{L,M}}(L; x)) - y(x))^2 \right], \quad (2)$$

where $g(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a given measuring function, $y(x) \in \mathbb{R}$ is the label corresponding to x , and μ is the probability distribution from which x is drawn.

Classical gradient descent updates the parameters according to the formula

$$\Theta_{L,M}^{n+1} = \Theta_{L,M}^n - h \nabla_{\Theta_{L,M}} E(\Theta_{L,M}^n),$$

where h is the step length. In the $h \rightarrow 0$ limit, these parameters satisfy the following ODE (Chizat and Bach, 2018, Def 2.2), with s denoting the rescaled pseudo-time:

$$\frac{d\Theta_{L,M}(s)}{ds} = -ML \nabla_{\Theta_{L,M}} E(\Theta_{L,M}), \quad \text{for } s \geq 0. \quad (3)$$

2.1 The Continuous Limit and Mean-field Limit

We obtain the *continuous limit* of (1) by making the ResNet infinitely deep, that is, $L \rightarrow \infty$. By reparameterizing the indices $l = [0, 1, \dots, L-1]$ with the continuous variable $t \in [0, 1]$, we view z in (1) as a function in t that satisfies a coupled ODE, with $1/L$ being the stepsize in time. Accordingly, $\theta_{l,m}$ can be viewed as $\theta_m(t = l/L)$. Denoting $\Theta(t) = \{\theta_m(t)\}_{m=1}^M$, the continuous limit of (1) can be written as

$$\frac{dz(t; x)}{dt} = \frac{1}{M} \sum_{m=1}^M f(z(t; x), \theta_m(t)), \quad t \in [0, 1], \quad \text{with } z(0; x) = x. \quad (4)$$

Following (2), we define the cost functional E for this formulation as

$$E(\Theta) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_\Theta(1; x)) - y(x))^2 \right], \quad (5)$$

where $Z_\Theta(t; x)$ solves (4) for a given collection $\Theta(t)$ of the M functions $\{\theta_m(t), m = 1, 2, \dots, M\}$. Similar to (3), we can use GD to find the configuration of $\Theta(t)$ that minimizes (5) by making $\Theta(t)$ flow in the descending direction of $E(\Theta)$, that is,

$$\frac{\partial \Theta(s; t)}{\partial s} = -M \left. \frac{\delta E}{\delta \Theta} \right|_{\Theta(s; \cdot)}, \quad s > 0, \quad t \in [0, 1], \quad (6)$$

where $\frac{\delta E}{\delta \Theta}$ is the functional derivative of E with respect to $\Theta(t)$.

The *mean-field limit* is obtained by making the ResNet infinitely wide, that is, $M \rightarrow \infty$. Considering that the right hand side of (4) has the form of an expectation, it naturally becomes an integral in the limit with respect to a certain probability density. Denoting this density by $\rho(\theta, t)$, and assuming that the θ_m are drawn from it, the ODE (4) for z translates to the following OIE:

$$\frac{dz(t; x)}{dt} = \int_{\mathbb{R}^k} f(z(t; x), \theta) d\rho(\theta, t), \quad t \in [0, 1] \quad \text{with } z(0; x) = x. \quad (7)$$

By mimicking (5), we define the cost function in the mean-field setting as

$$E(\rho) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_\rho(1; x)) - y(x))^2 \right], \quad (8)$$

where $Z_\rho(t; x)$ is the solution to (7) for a given ρ . Similar to the gradient flow for $\Theta_{L,M}$ and $\Theta(t)$, the probability distribution ρ that encodes the configuration of θ in the mean-field setting, flows in the descending direction of $E(\rho)$ in pseudo-time s . Considering that $\rho(\theta, t, s)$ should be a probability density for all s and t , this gradient flow is conducted in the Wasserstein metric. Following (Chizat and Bach, 2018; Lu et al., 2020), we have

$$\frac{\partial \rho(\theta, t, s)}{\partial s} = \nabla_\theta \cdot \left(\rho(\theta, t, s) \nabla_\theta \frac{\delta E(\rho)}{\delta \rho} \Big|_{\rho(\cdot, \cdot, s)} \right), \quad s > 0, t \in [0, 1], \quad (9)$$

where $\frac{\delta E}{\delta \rho}$ is the functional derivative with respect to ρ . With the classical variational method, one can explicitly compute this derivative as follows (see (Lu et al., 2020, Section 3.1)):

$$\frac{\delta E}{\delta \rho} \Big|_\rho (\theta, t) = \mathbb{E}_{x \sim \mu} \left(p_\rho^\top(t; x) f(Z_\rho(t; x), \theta) \right), \quad (10)$$

where $p_\rho(t; x)$, parameterized by x , maps $[0, 1] \rightarrow \mathbb{R}^d$, and is a vector solution to the following ODE:

$$\begin{cases} \frac{dp_\rho^\top}{dt} = -p_\rho^\top \int_{\mathbb{R}^k} \partial_z f(Z_\rho, \theta) \rho(\theta, t) d\theta, \\ p_\rho(t = 1; x) = (g(Z_\rho(1; x)) - y(x)) \nabla g(Z_\rho(1; x)). \end{cases} \quad (11)$$

In the later sections, to emphasize the s dependence, we use $\frac{\delta E(\Theta(s))}{\delta \Theta}$ and $\frac{\delta E(\rho(s))}{\delta \rho}$ to denote $\frac{\delta E}{\delta \Theta} \Big|_{\Theta(s; \cdot)}$ and $\frac{\delta E}{\delta \rho} \Big|_{\rho(\cdot, \cdot, s)}$ respectively. As a summary, to update $\rho(\theta, t, s)$ to $\rho(\theta, t, s + \delta s)$ with an infinitesimal δs , we solve (7) for $Z_\rho(t; x)$ using the given $\rho(\theta, t, s)$ and compute p_ρ using (11). This allows us to compute $\frac{\delta E(\rho(s))}{\delta \rho}(\theta, t)$ which, in turn, yields $\rho(\theta, t, s + \delta s)$ from (9). In (11), $\partial_z f$ is a $d \times d$ matrix that stands for the Jacobian of f with respect to its z argument.

Remark 1 *We stress that equation (9) is not a classical Wasserstein-2 gradient flow in the probability space. The classical gradient flow looks for one probability function of θ in the pseudo-time s (Ambrosio et al., 2008). Instead, we now have $\rho(\theta, t)$, where ρ is a probability density of θ for every t , and their flows are all coupled through the definition of E . Note that this feature emerges exactly because we have infinitely many layers $L \rightarrow \infty$, reformulated using a continuous variable t . When there are only 2 layers, the dependence on t can be dropped, and the problem reduces to the classical gradient flow (Chizat and Bach, 2018). A core obstacle to be overcome in this paper is the need for new technicalities to handle the non-standard gradient flow.*

2.2 Contributions

With the roadmap above, it is clear that we need to justify that the gradient flow in the discrete setting (3) is tracked closely by the PDE, so that $E(\Theta_{L,M}(s)) \approx E(\rho(\cdot, \cdot, s))$, and that the PDE (9) achieves the global minimum, for which $E(\rho(\cdot, \cdot, s = \infty)) = 0$.

In this paper we fully address the first question, and we provide some conditions under which the global minimum can be achieved. More specifically, the two main contributions of the paper are as follows.

- Contribution 1: We give a rigorous proof of the continuous and mean-field limit: Theorem 6. This is to justify that when $M, L \rightarrow \infty$, $E(\Theta_{L,M}(s)) \approx E(\Theta(s; \cdot)) \approx E(\rho(\cdot, \cdot, s))$. The dependence of these approximations on L and M are made precise in the result.
- Contribution 2: We show that the global minimum can be achieved in the continuous setting under certain conditions: Theorem 7. That is, the gradient flow equation (20), a modified version of (9), sends its distribution $\rho(\theta, t, s)$ to a global minimizer in the long-time regime (as $s \rightarrow \infty$). The conditions required may be stringent in real applications and we discuss their validity in certain scenarios in Appendix F.

Theorems 6 and 7 are combined in Theorem 9, which shows that $E(\Theta_{L,M}(s)) \rightarrow 0$ for large s .

By comparison with existing mean-field limit results, Theorem 6 gives rigorous justifications for both $L \rightarrow \infty$ and $M \rightarrow \infty$. By contrast, (Chizat and Bach, 2018; Araújo et al., 2019; Fang et al., 2019; Nguyen, 2019) consider $M \rightarrow \infty$ with L fixed. In most results for NTK (Allen-Zhu et al., 2019; Du et al., 2019a; Zhang et al., 2019; Chatterji et al., 2021), the size of M grows as L increases, while our results have these two parameters independent of each other. Furthermore, we have no requirement on the number of training samples.

For the a priori estimates, we also show the well-posedness of all six equations (three for z and three for θ) and that these equations have unique solutions. These results are mostly collected in Section 5, as follows.

- Well-posedness of Z : Theorem 10. For a given ρ , equation (7) has a unique solution Z_ρ , and the solution is stable with respect to ρ . With Remark 13, the uniqueness and stability result extend to treating (4) for a given $\Theta(t)$.
- Well-posedness of $\Theta_{L,M}, \Theta(t), \rho$: Theorems 11 and 12. These quantities satisfy the modified gradient flow equations (18), (19), and (20) respectively, and these equations have unique solutions.

3. Notations, Assumptions, and Definitions

Throughout the paper, we denote the collection of probability distribution that has a bounded second moment by $\mathcal{P}^2(\mathbb{R}^k)$, that is, $\mathcal{P}^2(\mathbb{R}^k) = \{\rho : \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta) < \infty\}$. We assume certain regularity properties for the activation function f , the measuring function g , the data y , and the input measure μ , as follows

Assumption 3.1 *Let $f : \mathbb{R}^d \times \mathbb{R}^k \rightarrow \mathbb{R}^d$ and $g, y : \mathbb{R}^d \rightarrow \mathbb{R}$ be C^2 functions. We assume that these functions and μ satisfy the following properties:*

1. *For all $x \in \mathbb{R}^d, \theta \in \mathbb{R}^k$, we have*

$$\left| \partial_x^i \partial_\theta^j f \right| \leq C_1 |\theta|^i (|x| + 1)^j, \quad 0 < i + j \leq 2, \quad i \geq 0, \quad j \geq 0, \quad (12)$$

where C_1 is a constant, $\partial_x^i \partial_\theta^j f$ is (i, j) derivative of f , and $|\cdot|$ is the Frobenius norm.

2. *The function set $\{h \mid h = \int_{\mathbb{R}^k} f(x, \theta) d\rho(\theta), \rho \in \mathcal{P}^2(\mathbb{R}^k)\}$ is dense in $\mathcal{C}(|x| < R; \mathbb{R}^d)$ for any $R > 0$.*

3. $g(x)$ and $\nabla g(x)$ are Lipschitz continuous. Moreover, $|\nabla g(x)|$ has a positive lower bound, that is, $\inf_{x \in \mathbb{R}^d} |\nabla g(x)| > 0$.
4. The associated probability density function of $\mu(x)$ is in L^∞ , and is compactly supported:

$$\mu(\{|x| \geq R\}) = 0, \quad (13)$$

for some fixed $R > 0$.

5. $y(x) \in L^\infty_{\text{loc}}(\mathbb{R}^d)$, that is, $\sup_{|x| \leq R} |y(x)| < \infty$.

The assumption (12) is satisfied by most commonly used activation functions (E et al., 2020b). Consider, for example,

$$f(x, \theta) = f(x, \theta_1, \theta_2, \theta_3, \theta_4) = \sigma(\theta_1 x + \theta_2) - \sigma(\theta_3 x + \theta_4),$$

where $\theta_1, \theta_3 \in \mathbb{R}^{d \times d}$, and $\theta_2, \theta_4 \in \mathbb{R}^d$, and σ is a component-wise regularized ReLU activation function. The form of the assumption controls the growth of its derivatives, and is used widely (Chizat and Bach, 2018; Lu et al., 2020).

For every fixed s , the solution to (9) is in the space $\mathcal{C}([0, 1]; \mathcal{P}^2)$. This is a non-standard probability space due to the introduction of t , so we need to build a new metric.

Definition 2 ¹ $\mathcal{C}([0, 1]; \mathcal{P}^2)$ is a collection of continuous paths of probability distribution $\rho(\theta, t)$ ($\theta \in \mathbb{R}^k, t \in [0, 1]$) where

1. For every fixed $t \in [0, 1]$, $\rho(\cdot, t) \in \mathcal{P}^2(\mathbb{R}^k)$.
2. For any $t_0 \in [0, 1]$, $\lim_{t \rightarrow t_0} W_2(\rho(\cdot, t), \rho(\cdot, t_0)) = 0$, where W_2 is the classical Wasserstein-2 distance.

The space $\mathcal{C}([0, 1]; \mathcal{P}^2)$ is equipped with the metric: $d_1(\rho_1, \rho_2) = \sup_t W_2(\rho_1(\cdot, t), \rho_2(\cdot, t))$.

We further define the space involving the pseudo-time s .

Definition 3 ² $\mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ is a collection of continuous paths of probability distribution $\rho(\theta, t, s)$ ($\theta \in \mathbb{R}^k, t \in [0, 1], s \in [0, \infty)$) where

1. For every fixed $s \in [0, \infty)$, $\rho(\cdot, \cdot, s) \in \mathcal{C}([0, 1]; \mathcal{P}^2)$.
2. For any $s_0 \in [0, \infty)$, $\lim_{s \rightarrow s_0} d_1(\rho(\cdot, \cdot, s), \rho(\cdot, \cdot, s_0)) = 0$ where d_1 is given in Definition 2.

The metric in $\mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ is defined by $d_2(\rho_1, \rho_2) = \sup_{t,s} W_2(\rho_1(\cdot, t, s), \rho_2(\cdot, t, s))$.

Since \mathcal{P}^2 is complete in W_2 distance, $\mathcal{C}([0, 1]; \mathcal{P}^2)$ and $\mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ are complete metric spaces under d_1 and d_2 respectively also.

For carrying out the mean-field limit, we use the classical approach and follow the particle presentation of ρ , this means we require the probability distribution ρ to be admissible.

-
1. $\mathcal{C}([0, T]; \mathcal{A})$ is a collection of functions $f(\theta, t)$ that are continuous in time $t \in [0, T]$, and for each fixed time t_0 , $f(t_0, \theta) \in \mathcal{A}$. In this case, $T = 1$ and $\mathcal{A} = \mathcal{P}^2$, equipped with W_2 metric.
 2. Similar to Definition 3, $\mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ is a collection of functions $f(\theta, t, s)$ that are continuous in s and t , and for each fixed $(s_0, t_0) \in [0, \infty) \times [0, 1]$, $f(\theta, t_0, s_0) \in \mathcal{P}^2$.

Definition 4 We call $\rho(\theta, t)$ admissible if it has a particle presentation, namely there exists a continuous stochastic process $\theta(t) : [0, 1] \rightarrow \mathbb{R}^k$ such that

$$\theta(t_0) \sim \rho(\theta, t_0), \quad \lim_{t \rightarrow t_0} \mathbb{E}(|\theta(t) - \theta(t_0)|^2) = 0. \quad (14)$$

Furthermore, $\rho(\theta, t)$ is called limit-admissible if its M averaged trajectory is bounded and Lipschitz with high probability. See the rigorous definition in Definition 29.

Remark 5 We have two comments on the admissible condition.

- Without the dependence on t , all probability distributions are admissible: One can define a particle by drawing it from this distribution. With the dependence on t , the admissible requirement is stricter: There needs to be a continuous stochastic process $\theta(t)$ whose underlying distribution is the given $\rho(\theta, t)$.
- According to (Ambrosio et al., 2008, Theorem 8.2.1), if $\rho(\cdot, t)$ is absolutely continuous in t using the Wasserstein metric, it can be regarded as a probability measure over continuous function $C([0, 1]; \mathbb{R}^k)$, and can be represented by a continuous stochastic process $\theta(t)$.

4. Mean-field and Convergence of the Gradient Flow

We present our main result in this section. To ensure the well-posedness of the gradient flow equation (9), we need to add a regularizer to the cost function, as discussed in Section 4.1. Sections 4.2 and 4.3 discuss validity of the continuous and mean-field limits and convergence to the global minimizer. The convergence to the global minimizer requires a fully-supported assumption; we discuss the validity of this assumption in Appendix F.

4.1 Modified Cost Function

The analysis is conducted with a modified cost function to ensure the well-posedness of (9), the gradient flow equation. To do so, we add a regularizer to the cost function (2), as follows:

$$E_s(\Theta_{L,M}) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_{\Theta_{L,M}}(L; x)) - y(x))^2 \right] + e^{-s} \frac{1}{ML} \sum_{l=0}^{L-1} \sum_{m=1}^M |\theta_{l,m}|^2. \quad (15)$$

Recalling our notation $\Theta(t) = \{\theta(t)\}_{m=1}^M$, the corresponding regularized cost function in the continuous limit ($L \rightarrow \infty$) is

$$E_s(\Theta(\cdot)) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_{\Theta}(1; x)) - y(x))^2 \right] + \frac{e^{-s}}{M} \sum_{m=1}^M \int_0^1 |\theta_m(t)|^2 dt. \quad (16)$$

Finally, in the mean-field limit ($M \rightarrow \infty$), we assume θ_m is i.i.d. drawn from ρ , the corresponding regularized cost function becomes

$$E_s(\rho) = \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} (g(Z_{\rho}(1; x)) - y(x))^2 \right] + e^{-s} \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t). \quad (17)$$

In these definitions, $Z_{\Theta_{L,M}}$, Z_{Θ} , and Z_{ρ} are solutions to the difference equation (1), ODE (4) and OIE (7), respectively. It is straightforward to see that E_s differs from the (non-regularized) cost function E by only an exponentially small term, which disappears as $s \rightarrow \infty$. As a consequence, the global minimum of E_s coincides with the global minimum of the true cost E in the large- s limit, so it makes sense to analyze gradient descent applied to E_s .

4.2 Continuous and Mean-field Limit of the Gradient Flow

Mimicking the derivation in Section 2.1, the gradient flows for $\Theta_{L,M}$, $\Theta(t)$, and ρ can be changed according to the regularized cost functions involving pseudo-time $s \in [0, \infty)$ defined above. $\Theta_{L,M}$ solves

$$\frac{d\Theta_{L,M}(s)}{ds} = -ML\nabla_{\Theta_{L,M}}E(\Theta_{L,M}(s)) - 2e^{-s}\Theta_{L,M}(s), \quad \text{for } s > 0. \quad (18)$$

For the continuous limit ($L \rightarrow \infty$), we have

$$\frac{d\theta_m(s; t)}{ds} = -M\frac{\delta E(\Theta(s))}{\delta \theta_m} - 2e^{-s}\theta_m(s; t), \quad m = 1, 2, \dots, M. \quad (19)$$

For the mean-field limit ($M \rightarrow \infty$), we have

$$\begin{cases} \frac{\partial \rho(\theta, t, s)}{\partial s} = \nabla_{\theta} \cdot \left(\rho(\theta, t, s) \nabla_{\theta} \frac{\delta E_s(\rho(s))}{\delta \rho} \right), & \text{for } s > 0, \quad t \in [0, 1], \\ \rho(\theta, t, 0) = \rho_{\text{ini}}(\theta, t). \end{cases} \quad (20)$$

Noting the definition of E_s in (17), we have the explicit expression for the Fréchet derivative:

$$\frac{\delta E_s(\rho)}{\delta \rho} = \frac{\delta E(\rho)}{\delta \rho} + e^{-s}|\theta|^2, \quad (21)$$

where $\frac{\delta E(\rho)}{\delta \rho}$ is defined in (10).

The connection between these three equations is made rigorous in the following result.

Theorem 6 *Suppose that Assumption 3.1 holds. Let $\rho_{\text{ini}}(\theta, t)$ be limit-admissible, and $\{\theta_m(0; t)\}_{m=1}^M$ in (19) be i.i.d. drawn from $\rho_{\text{ini}}(\theta, t)$. Let*

- $\Theta_{L,M}(s) = \{\theta_{l,m}(s)\}$ be the solution to (18) with initial condition $\theta_{l,m}(s=0) = \theta_m(0; \frac{t}{L})$, and
- $\rho(\theta, t, s)$ be the solution to (20) with initial condition $\rho_{\text{ini}}(\theta, t)$.

Then for any positive ϵ , η , and S , there exists a constant $C > 0$ that depends on $\rho_{\text{ini}}(\theta, t)$ and S such that when

$$M > \frac{C(\rho_{\text{ini}}(\theta, t), S)}{\epsilon^2 \eta}, \quad L > \frac{C(\rho_{\text{ini}}(\theta, t), S)}{\epsilon},$$

we have

$$\mathbb{P}(|E(\Theta_{L,M}(s)) - E(\rho(\cdot, \cdot, s))| \leq \epsilon) \geq 1 - \eta, \quad \forall s < S,$$

where $E(\Theta_{L,M}(s))$ is defined in (2) and $E(\rho(\cdot, \cdot, s))$ is defined in (8).

The proof of this result appears in Appendix G. This theorem suggests that for every fixed $S > 0$, arbitrarily large, the cost generated by $\Theta_{L,M}$ that follows its gradient flow coded in GD is comparable (with high probability) to the cost generated by ρ , the continuous mean-field limit of $\Theta_{L,M}$, provided that L and M are sufficiently large. The size of the ResNet depends negative-algebraically on ϵ (the desired accuracy) and η (the confidence of success). The essence of this theorem is that it matches the zero-loss property of the parameter configuration of a finite sized ResNet to its limiting PDE, whose analysis can be performed with some standard PDE tools.

The proof of Theorem 6 divides naturally into two components: first, showing $E(\rho(\cdot, \cdot, s)) \approx E(\Theta(s; \cdot))$ and second, showing $E(\Theta(s; \cdot)) \approx E(\Theta_{L,M}(s))$ for all $s < S$. The former amounts to utilizing the mean-field limit theory and justifying that the particle trajectory $\theta_m(t, s)$ follows $\rho(\theta, t, s)$ for all t in pseudo-time $s \in [0, S]$. The latter is to adopt the continuity in t and trace the differences between $\theta_m(\frac{t}{L})$ and $\theta_{l,m}$. These two components of the proof are summarized in Theorems 31 and 32, respectively. According to the formula of the Fréchet derivatives (10) and (11), the estimates in these theorems naturally route through the boundedness of p_ρ , p_Θ , $p_{\Theta_{L,M}}$, and similarly Z_ρ , Z_Θ and $Z_{\Theta_{L,M}}$. These bounds are collected in Theorem 10 and Lemmas 24 and 37.

To intuitively see the equivalence between (19) and (20), we test them on the same smooth function $h(\theta)$. To test (20), we multiply it on both sides and perform integration by parts to obtain:

$$\frac{d}{ds} \int_{\mathbb{R}^k} h d\rho(\theta) = - \int_{\mathbb{R}^k} \nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho} d\rho.$$

This is to say $\frac{d}{ds} \mathbb{E}(h) = -\mathbb{E} \left(\nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho} \right)$. Testing h on (19) gives the same equivalence. Suppose $\rho = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m}$, then:

$$\frac{d}{ds} \mathbb{E}(h) = \frac{1}{M} \sum_{m=1}^M \nabla_\theta h(\theta_m) \frac{d}{ds} \theta_m = - \sum_{m=1}^M \nabla_\theta h(\theta_m) \frac{\delta E_s(\Theta)}{\delta \theta_m}.$$

The right hand side is also $-\mathbb{E} \left(\nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho} \right)$ if and only if $M \frac{\delta E_s(\Theta(s))}{\delta \theta_m} = \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho}(\theta_m, t)$. This holds true, as presented in Lemma 34. The proof in Theorem 31 gives the rigorous quantification in the convergence in M .

4.3 Descending to the Global Minimum

We investigate the convergence in s of the ρ system by studying the large- s behavior of the limiting gradient flow equation (20). Our main theorem is the following.

Theorem 7 *Assume that Assumption 3.1 holds and that $\rho_{\text{ini}}(\theta, t)$ is admissible. Suppose that $\rho(\theta, t, s) \in \mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ solves (20). If the following two conditions hold:*

- *there exists a long-time limit $\rho_\infty(\theta, t) \in \mathcal{C}([0, 1]; \mathcal{P}^2)$ such that $\rho(\theta, t, s)$ converges to $\rho_\infty(\theta, t)$ in $\mathcal{C}([0, 1]; \mathcal{P}^2)$ as $s \rightarrow \infty$;*
- *the support of $\rho_\infty(\theta, t_0)$ is the full space \mathbb{R}^k for some $t_0 \in [0, 1]$,*

then $E(\rho_\infty) = 0$, so that ρ_∞ is a global minimizer. More specifically, for any $\epsilon > 0$, there is a constant $C_0 > 0$ depending only on $\rho_{\text{ini}}(\theta, t)$ and ϵ such that

$$E(\rho(s)) \leq \epsilon, \quad \forall s > C_0(\rho_{\text{ini}}(\theta, t), \epsilon).$$

This theorem gives the conditions under which the long-time limit ρ_∞ , if achievable, becomes a global minimizer with zero value of the objective E . Its proof, which uses PDE analysis techniques, appears in Appendix E. The approach follows other similar results on the global minimum: For every ρ with $E(\rho) > 0$, we can find a measure ν with $\int d\nu = 0$ such that $\frac{\delta E(\rho)}{\delta \rho}$ in the direction of ν is negative, suggesting that a better probability measure than ρ can be obtained by descending along the ν direction. This essentially equates a steady state with the global minimum; see Proposition 26. In addition, under the assumption that ρ_∞ is supported in the entire space, we show that the global minimum is achievable with the flow suggested by the equation.

Remark 8 (Assumptions in Theorem 7.) *There are two notable assumptions in Theorem 7: First, that $\rho(\theta, t, s)$ converges to $\rho_\infty(\theta, t)$ as $s \rightarrow \infty$, and second, that ρ_∞ is supported on the whole domain \mathbb{R}^k for some time $t_0 \in [0, 1]$. These two assumptions are difficult to justify, but in Appendix F we describe some scenarios under which both conditions hold, though these scenarios are themselves stringent. We leave an investigation of weaker conditions under which the limit exists to future research.*

When the neural network has two layers, with $M \rightarrow \infty$, (Chizat and Bach, 2018) shows that the limiting gradient flow equation can achieve the global minimum without the full-support condition. However, they require f to be 2-homogeneous or partially 1-homogeneous (in θ). They also assume the convergence of $\rho(\theta, s)$ to ρ_∞ in the $s \rightarrow \infty$ limit. Without additional structure in the activation function, proving global convergence is a challenging task that we leave to future work.

4.4 Main Result: Global Convergence of Finite-sized ResNet

We are now ready to present the main result of the paper. Combining Theorems 6 and 7 we naturally obtain the main theorem that justifies the overparameterizing deep ResNet architecture.

Theorem 9 *Suppose that the conditions in Theorem 6 and 7 hold. Then for any positive ϵ and η , there exist positive constants C_0 depending on $\rho_{\text{ini}}(\theta, t), \epsilon$ and C depending on $\rho_{\text{ini}}(\theta, t), s$ such that when*

$$s > C_0(\rho_{\text{ini}}(\theta, t), \epsilon), \quad M > \frac{C(\rho_{\text{ini}}(\theta, t), s)}{\epsilon^2 \eta}, \quad L > \frac{C(\rho_{\text{ini}}(\theta, t), s)}{\epsilon},$$

we have

$$\mathbb{P}(|E(\Theta_{L,M}(s))| \leq \epsilon) \geq 1 - \eta.$$

This theorem states the final result of the paper: If ResNet is big enough, the gradient descent, after running long enough time, finds nearly-zero loss solution with high probability. The size of the ResNet algebraically depends on the accuracy and confidence level.

5. Well-posedness and Mean-field Result

Some a priori estimates on the well-posedness of the dynamical system are necessary in the course of establishing the long-time convergence result. We list well-posedness results in this section. We show that the dynamical system for z has a unique and stable solution, and the gradient flow of the parameter configuration is also well-posed.

5.1 Well-posedness of the OIE (7)

As L and M approach ∞ , z satisfies the ordinary-integral equation presented in (7). We justify that this differential equation is well-posed, in the sense that the solution is unique and stable.

Theorem 10 *Suppose that Assumption 3.1 holds, that x is in the support of μ , and that $\rho_1, \rho_2 \in \mathcal{C}([0, 1]; \mathcal{P}^2)$. Then (7) has a unique C^1 solution. Further, this solution is stable, in the sense that*

$$|Z_{\rho_1}(t; x)| \leq C(\mathcal{L}_1) \quad (22a)$$

$$|Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)| \leq C(\mathcal{L}_1, \mathcal{L}_2) d_1(\rho_1, \rho_2), \quad \text{for all } t \in [0, 1]. \quad (22b)$$

Here $C(\cdot)$ are constants depending only on the given arguments, and \mathcal{L}_i are the second moments of ρ_i ($i = 1, 2$), defined by $\mathcal{L}_i = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_i(\theta, t) dt$.

The proof of this result appears in Appendix B. The theorem suggests that a small perturbation to ρ is linearly reflected in Z_ρ , the solution to (7). Thus, a small perturbation in the parameterization of the ResNet leads only to a small perturbation to the ResNet output.

5.2 Well-posedness of the Adjusted Gradient Flow

The gradient flow of the parametrization is also well-posed, in both the discrete setting and the continuous mean-field limit.

In the discrete setting, we have the following.

Theorem 11 *Suppose that Assumption 3.1 holds, then (18) has a unique solution.*

Next, we show that equation (20), which characterizes the dynamics of the continuous mean-field limit of the parameter configuration, is also well-posed.

Theorem 12 *Suppose that Assumption 3.1 holds and that $\rho_{\text{ini}}(\theta, t)$ is admissible. Then (20) has a unique solution $\rho(\theta, t, s)$ in $\mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ with initial condition $\rho_{\text{ini}}(\theta, t)$. Furthermore, $\rho(\theta, t, s)$ is admissible for any s and*

$$\frac{dE_s(\rho(\cdot, \cdot, s))}{ds} \leq 0. \quad (23)$$

The proofs of both theorems can be found in Appendix D.

Remark 13 *Although we do not directly show the well-posedness of (4) and (19), it is effectively a corollary of Theorem 10 and 12 above. One way to connect them is to reformulate the discrete probability distribution as*

$$\rho_{\Theta}^{\text{dis}}(\theta, t, s) = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m(s; t)}(\theta), \quad (24)$$

where $\Theta(s; t) = \{\theta_m(s; t)\}_{m=1}^M$ is the list of trajectories. Since $\theta_m(0; t)$ is continuous in t , we have $\rho_{\Theta}^{\text{dis}}(\theta, 0; t) \in \mathcal{C}([0, 1]; \mathcal{P}^2)$. Because

$$\frac{1}{M} \sum_{m=1}^M f(z(t; x), \theta_m(0; t)) = \int_{\mathbb{R}^k} f(z(t; x), \theta) d\rho_{\Theta}^{\text{dis}}(\theta, t, 0),$$

using Theorem 10, (4) has a unique \mathcal{C}^1 solution $Z_{\Theta}(t; x)$ when $\Theta(t)$ is continuous. Furthermore, according to the definition (16) and (17), we have

$$E_s \left(\rho_{\Theta}^{\text{dis}}(\cdot, \cdot, s) \right) = E_s \left(\Theta(s; \cdot) \right).$$

As a consequence, if $\Theta(s; t)$ satisfies (19), then $\rho_{\Theta}^{\text{dis}}$ satisfies (20), and vice versa. The well-posedness result in Theorem 12 for (20) then can be extended to justify the well-posedness of (19). The rigorous proof for this comment can be found in Remark 35.

6. Conclusion

First-order methods such as gradient descent can find a global optimizer that provides zero-loss for fitting deep neural network. We explain this mechanism for ResNet in the limiting regime when both the number of layers and the number of weights per layer approach infinity. We show that GD of parameter configuration of ResNet can be translated to a gradient flow of the limiting probability distribution. Furthermore, we give an estimate of the size of such ResNets. Moreover, we show that under some conditions, the limiting gradient flow captures the zero-loss as its global minimum.

In future work, we would like to relax the assumptions for the activation function and the fully-supported condition required in Theorem 7. Furthermore, the regularized version E_s is used as a replacement of E for technical reasons, but we believe such strategy can be relaxed as well.

Acknowledgments

Q.L. acknowledges support from Vilas Early Career award. The research of Z.D., S.C., and Q.L. is supported in part by NSF-CAREER-1750488 and Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin Madison with funding from the Wisconsin Alumni Research Foundation. The work of Z.D., S.C., Q.L., and S.W. is supported in part by NSF via grant DMS-2023239. S.W. also acknowledges support from NSF Awards 1628384, 1740707, 1839338, and 1934612, and Subcontract 8F-30039 from Argonne National Laboratory. The authors also acknowledge two anonymous reviewers for insightful suggestions.

Appendix A. Introduction of the Appendix

This appendix contains proofs and supporting analysis for the theorems in the main text. We start by proving well-posedness of the ResNet ODE and the gradient flow. This is followed by the global convergence of the gradient-flow PDE. Finally, we prove the validity of continuous and mean-field limit.

The specific organization of the appendix is as follows.

Appendix B: Proof of Theorem 10: Well-posedness of the ResNet ODE in its continuous limit (4), and the mean-field limit (7).

Appendix C: Preparation of a-priori estimates, subsequently used to show well-posedness of the gradient flow equation.

Appendix D: Proof of Theorem 11 and 12: Well-posedness of gradient flows (18) and (20).

Appendix E, F: Proof of Theorem 7: Global convergence of the gradient flow. (Section F demonstrates validity of the assumption we made in the statement of Theorem 7.)

Appendix G-I: Proof of Theorem 6. Section G lays out the structure of the proof, Section H shows the continuous limit, and Section I shows the mean-field limit.

The analytical core of the paper lies in Appendices E and G, which describe properties of the gradient flow PDEs and explain why the gradient descent method for ResNet can be explained by these gradient flow equations. The technical results of Appendix B-D can be skipped by readers who are interested to proofs of the main results.

Throughout, we denote by $C(\cdot)$ a generic constant that depends on its arguments (\cdot) . The precise value of this constant may change at each instance.

Appendix B. Proof of Theorem 10

First, we study well-posedness of the dynamical system by proving Theorem 10. Recall the equation:

$$\frac{dZ_\rho(t; x)}{dt} = F(Z_\rho, t), \quad \forall t \in [0, 1] \text{ with } z(0; x) = x, \quad (25)$$

where for a given $\rho \in \mathcal{C}([0, 1]; \mathcal{P}^2)$ we use the notation

$$F(z, t) = \int_{\mathbb{R}^k} f(z, \theta) d\rho(\theta, t). \quad (26)$$

The proof of Theorem 10 relies on the classical Lipschitz condition for the well-posedness of an ODE.

Proof [Proof of Theorem 10] Since $\rho \in \mathcal{C}([0, 1]; \mathcal{P}^2)$, we obtain that

$$\sup_{0 \leq t \leq 1} \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t) < C < \infty.$$

This implies that $\int_{\mathbb{R}^k} |\theta| d\rho(\theta, t) < \sqrt{C}$ for all t .

For any $t \in [0, 1]$, using (12) from Assumption 3.1, we have

$$|f(z, \theta)| \leq C(|\theta| + 1)(|z| + 1), \quad |f(z_1, \theta) - f(z_2, \theta)| \leq C_1|\theta||z_1 - z_2|. \quad (27)$$

Then, with the boundedness of the first moment of θ , we have

$$\begin{aligned} |F(z_1, t) - F(z_2, t)| &\leq \left| \int_{\mathbb{R}^k} (f(z_1, \theta) - f(z_2, \theta)) \, d\rho(\theta, t) \right| \\ &\leq C_1 \left| \int_{\mathbb{R}^k} |\theta| \, d\rho(\theta, t) \right| |z_1 - z_2| < C|z_1 - z_2|, \end{aligned} \quad (28)$$

meaning that $F(z, t)$ is uniformly Lipschitz in z for all $t \in [0, 1]$. Therefore (7) has a unique \mathcal{C}^1 solution.

To show the boundedness in (22a), we have from (27) and (26) that

$$|F(z, t)| \leq C(|z| + 1) \int_{\mathbb{R}^k} (|\theta| + 1) \, d\rho(\theta, t). \quad (29)$$

Multiplying (25) and using (29), we obtain

$$\begin{aligned} &\frac{d|Z_{\rho_1}(t; x)|^2}{dt} \\ &\leq 2C(|Z_{\rho_1}|^2 + |Z_{\rho_1}|) \int_{\mathbb{R}^k} (|\theta| + 1) \, d\rho_1(\theta, t) \\ &\leq 4C \int_{\mathbb{R}^k} (|\theta| + 1) \, d\rho_1(\theta, t) (|Z_{\rho_1}(t; x)|^2 + 1), \end{aligned}$$

Using Grönwall's inequality, we have

$$\begin{aligned} |Z_{\rho_1}(t; x)| &\leq \exp \left(2C \left(\int_0^1 \int_{\mathbb{R}^k} |\theta| \, d\rho_1 \, dt + 1 \right) \right) (|x| + 1) \\ &\leq \exp \left(2C \left(\left(\int_0^1 \int_{\mathbb{R}^k} |\theta|^2 \, d\rho_1 \, dt \right)^{1/2} + 1 \right) \right) (|x| + 1) \\ &\leq \exp \left(2C \left(\sqrt{\mathcal{L}_1} + 1 \right) \right) (|x| + 1), \end{aligned}$$

where $\mathcal{L}_1 = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 \, d\rho_1 \, dt$. Finally, to prove the stability result (22b), we define for $\rho_1, \rho_2 \in \mathcal{C}([0, 1]; \mathcal{P}^2)$ the notation

$$\Delta(t; x) = Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x).$$

Then we have (using (28)) that

$$\begin{aligned}
 & \frac{d|\Delta(t; x)|^2}{dt} \\
 &= 2 \left\langle \Delta(t; x), \int_{\mathbb{R}^k} f(Z_{\rho_1}(t; x), \theta) d\rho_1(\theta, t) - \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_2(\theta, t) \right\rangle \\
 &= 2 \left\langle \Delta(t; x), \int_{\mathbb{R}^k} f(Z_{\rho_1}(t; x), \theta) d\rho_1(\theta, t) - \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_1(\theta, t) \right\rangle \\
 &\quad + 2 \left\langle \Delta(t; x), \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_1(\theta, t) - \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_2(\theta, t) \right\rangle \\
 &\leq 2C_1 |\Delta(t; x)|^2 \int_{\mathbb{R}^k} |\theta| d\rho_1(\theta, t) + 2C_1 |\Delta(t; x)| (|Z_{\rho_2}(t; x)| + 1) d_1(\rho_1, \rho_2) \\
 &\leq 2C_1 \left(\int_{\mathbb{R}^k} |\theta| d\rho_1(\theta, t) + 1 \right) |\Delta(t; x)|^2 + 2C_1 (|Z_{\rho_2}(t; x)| + 1)^2 d_1^2(\rho_1, \rho_2),
 \end{aligned} \tag{30}$$

where in the first inequality we first use the mean-value theorem and Assumption 3.1 (12) to obtain

$$|f(Z_{\rho_1}(t; x), \theta) - f(Z_{\rho_2}(t; x), \theta)| \leq C_1 |\theta| |Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)| = C_1 |\theta| |\Delta(t; x)|.$$

Next, let $\theta_1 \sim \rho_1(\theta, t)$ and $\theta_2 \sim \rho_2(\theta, t)$ such that $(\mathbb{E}(|\theta_1 - \theta_2|^2))^{1/2} = W_2(\rho_1(\cdot, t), \rho_2(\cdot, t))$. Similarly, we also have

$$\begin{aligned}
 & \left| \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_1(\theta, t) - \int_{\mathbb{R}^k} f(Z_{\rho_2}(t; x), \theta) d\rho_2(\theta, t) \right| \\
 &\leq \mathbb{E}(|f(Z_{\rho_2}(t; x), \theta_1) - f(Z_{\rho_2}(t; x), \theta_2)|) \\
 &\leq C_1 (|Z_{\rho_2}(t; x)| + 1) \mathbb{E}(|\theta_1 - \theta_2|) \\
 &\leq C_1 (|Z_{\rho_2}(t; x)| + 1) (\mathbb{E}(|\theta_1 - \theta_2|^2))^{1/2} \\
 &\leq C_1 (|Z_{\rho_2}(t; x)| + 1) d_1(\rho_1, \rho_2)
 \end{aligned}$$

Finally, since $|\Delta(0; x)| = 0$, we have

$$|\Delta(t; x)| \leq 2C_1 d_1^2(\rho_1, \rho_2) \exp \left(2C_1 \left(\sqrt{\mathcal{L}_1} + t \right) \right) \int_0^t (|Z_{\rho_2}(s; x)| + 1)^2 ds$$

where we use Grönwall's inequality and $\int_0^1 \int_{\mathbb{R}^k} |\theta| d\rho_1(\theta, t) dt \leq \sqrt{\mathcal{L}_1}$ by Hölder's inequality. By substituting $|Z_{\rho_2}(t; x)|$ from (22a), we complete the proof of (22b). \blacksquare

Appendix C. A Priori Estimates of the Cost Function

Some a priori estimates are necessary in the proof for the main theorems. We prove two lemmas here.

For E_s , standard techniques (see (Lu et al., 2020)) show that

$$\frac{\delta E_s(\rho)}{\delta \rho}(\theta, t) = \mathbb{E}_{x \sim \mu} \left(p_\rho^\top(t; x) f(Z_\rho(t; x), \theta) \right) + e^{-s} |\theta|^2, \tag{31}$$

where $p_\rho(t; x)$ is the solution to the following ODE:

$$\begin{cases} \frac{\partial p_\rho^\top}{\partial t} = -p_\rho^\top \int_{\mathbb{R}^k} \partial_z f(Z_\rho(t; x), \theta) d\rho(\theta, t), \\ p_\rho(1; x) = (g(Z_\rho(1; x)) - y(x)) \nabla g(Z_\rho(1; x)). \end{cases} \quad (32)$$

According to assumption (13), $\mu(x)$ has a compact support with $|x| < R$.

Our first lemma shows that p_ρ is Lipschitz continuous with respect to ρ .

Lemma 14 *Suppose that Assumption 3.1 holds and that x is in the support of μ . Suppose that $\rho_1, \rho_2 \in \mathcal{C}([0, 1]; \mathcal{P}^2)$ with corresponding solutions p_{ρ_1}, p_{ρ_2} of (32), and denote*

$$\mathcal{L}_i = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_i(\theta, t) dt.$$

Then the following two bounds are satisfied:

$$|p_{\rho_1}(t; x)| \leq C(\mathcal{L}_1), \quad (33a)$$

$$|p_{\rho_1}(t; x) - p_{\rho_2}(t; x)| \leq C(\mathcal{L}_1, \mathcal{L}_2) d_1(\rho_1, \rho_2). \quad (33b)$$

Proof From (12) in Assumption 3.1, with $i = 1$ and $j = 0$, we have

$$\left| \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_1}(t; x), \theta) d\rho_1(\theta, t) \right| \leq C_1 \int_{\mathbb{R}^k} |\theta| d\rho_1(\theta, t). \quad (34)$$

it follows from the initial conditions of (32) that

$$|p_{\rho_1}(1; x)| \leq C(|Z_{\rho_1}(1; x)| + 1) \leq C(\mathcal{L}_1). \quad (35)$$

where we use Assumption 3.1 in the first inequality and (22a) in the second inequality. Noting that (32) is a linear equation, (33a) follows naturally when we combine (34) with (35) and use $\int_0^1 \int_{\mathbb{R}^k} |\theta| d\rho_1(\theta, t) dt \leq \mathcal{L}_1^{1/2}$.

To prove (33b), we define

$$\Delta(t; x) = p_{\rho_1}(t; x) - p_{\rho_2}(t; x).$$

First, when $t = 1$ and $|x| < R$, we have from the initial conditions that

$$\begin{aligned} |\Delta(1; x)| &= |p_{\rho_1}(1; x) - p_{\rho_2}(1; x)| \\ &= |(g(Z_{\rho_1}(1; x)) - y(x)) \nabla g(Z_{\rho_1}(1; x)) - (g(Z_{\rho_2}(1; x)) - y(x)) \nabla g(Z_{\rho_2}(1; x))| \\ &\leq C(\mathcal{L}_1, \mathcal{L}_2) |Z_{\rho_1}(1; x) - Z_{\rho_2}(1; x)| \leq C(\mathcal{L}_1, \mathcal{L}_2) d_1(\rho_1, \rho_2), \end{aligned} \quad (36)$$

where we use Assumption 3.1, (22a), and $|x| < R$ in the first inequality and (22b) in the second inequality. The following ODE is satisfied by Δ :

$$\frac{\partial \Delta^\top(t; x)}{\partial t} = -\Delta^\top(t; x) \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_1}(t; x), \theta) d\rho_1(\theta, t) + p_{\rho_2}^\top(t; x) D_{\rho_1, \rho_2}(t; x), \quad (37)$$

where

$$D_{\rho_1, \rho_2}(t; x) = \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_2}(t; x), \theta) d\rho_2(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_1}(t; x), \theta) d\rho_1(\theta, t).$$

The boundedness of $D_{\rho_1, \rho_2}(t; x)$ can be shown similarly to (30), by first splitting into two terms:

$$\begin{aligned} |D_{\rho_1, \rho_2}(t; x)| &\leq \underbrace{\left| \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_2}(t; x), \theta) d\rho_2(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\rho_2}(t; x), \theta) d\rho_1(\theta, t) \right|}_{\text{(I)}} \\ &\quad + \underbrace{\left| \int_{\mathbb{R}^k} [\partial_z f(Z_{\rho_2}(t; x), \theta) - \partial_z f(Z_{\rho_1}(t; x), \theta)] d\rho_1(\theta, t) \right|}_{\text{(II)}}. \end{aligned} \quad (38)$$

The bound of (II) relies on Assumption 3.1 and (22b):

$$\begin{aligned} \text{(II)} &\leq C_1 |Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)| \int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) \\ &\leq C(\mathcal{L}_1, \mathcal{L}_2) \left(\int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) \right) d_1(\rho_1, \rho_2). \end{aligned} \quad (39)$$

To bound (I), noticing that ρ_i are admissible, we let $\theta_1 \sim \rho_1(\theta, t)$ and $\theta_2 \sim \rho_2(\theta, t)$ such that $(\mathbb{E}(|\theta_1 - \theta_2|^2))^{1/2} = W_2(\rho_1(\cdot, t), \rho_2(\cdot, t))$. We then have

$$\begin{aligned} \text{(I)} &\leq \mathbb{E}(|\partial_z f(Z_{\rho_2}(t; x), \theta_1) - \partial_z f(Z_{\rho_2}(t; x), \theta_2)|) \\ &\leq C_1 (|Z_{\rho_2}(t; x)| + 1) \mathbb{E}((|\theta_1| + |\theta_2|)|\theta_1 - \theta_2|) \\ &\leq C(\mathcal{L}_2) \mathbb{E}((|\theta_1| + |\theta_2|)|\theta_1 - \theta_2|) \\ &\leq C(\mathcal{L}_2) (\mathbb{E}(|\theta_1|^2 + |\theta_2|^2))^{1/2} (\mathbb{E}(|\theta_1 - \theta_2|^2))^{1/2} \\ &\leq C(\mathcal{L}_2) (\mathbb{E}(|\theta_1|^2 + |\theta_2|^2))^{1/2} d_1(\rho_1, \rho_2) \\ &= C(\mathcal{L}_2) \left(\int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) + \int_{\mathbb{R}^k} |\theta|^2 d\rho_2(\theta, t) \right)^{1/2} d_1(\rho_1, \rho_2), \end{aligned} \quad (40)$$

where we use the mean-value theorem and Assumption 3.1 (12) in the second inequality: For some $\lambda \in (0, 1)$, we have

$$\begin{aligned} |\partial_z f(Z_{\rho_2}(t; x), \theta_1) - \partial_z f(Z_{\rho_2}(t; x), \theta_2)| &\leq |\partial_z \partial_\theta f(Z_{\rho_2}(t; x), \theta_1 + \lambda \theta_2)| |\theta_1 - \theta_2| \\ &\leq C_1 (|Z_{\rho_2}(t; x)| + 1) (|\theta_1| + |\theta_2|) |\theta_1 - \theta_2|. \end{aligned} \quad (41)$$

We use (22a) in the third inequality.

By substituting (40) and (39) into (38) and using Hölder's inequality, we obtain

$$|D_{\rho_1, \rho_2}(t; x)| \leq C(\mathcal{L}_1, \mathcal{L}_2) \left(\int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) + \int_{\mathbb{R}^k} |\theta|^2 d\rho_2(\theta, t) + 1 \right) d_1(\rho_1, \rho_2).$$

By using this bound in (37), using (34) and Hölder's inequality, we have

$$\begin{aligned} \frac{d|\Delta(t; x)|^2}{dt} &\leq C(\mathcal{L}_1, \mathcal{L}_2) \left(\int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) + \int_{\mathbb{R}^k} |\theta|^2 d\rho_2(\theta, t) + 1 \right) |\Delta(t; x)|^2 \\ &\quad + C(\mathcal{L}_1, \mathcal{L}_2) \left(\int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) + \int_{\mathbb{R}^k} |\theta|^2 d\rho_2(\theta, t) + 1 \right) d_1^2(\rho_1, \rho_2). \end{aligned} \quad (42)$$

Finally, we use Grönwall's inequality with the initial condition (36) to obtain (33b). \blacksquare

Our second lemma concerns the continuity of $\nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}$.

Lemma 15 *Assume $\rho, \rho_1, \rho_2 \in \mathcal{C}([0, 1]; \mathcal{P}^2)$. Defining*

$$\mathcal{L} = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t) dt, \quad \tilde{\mathcal{L}} = \sup \left\{ \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_1(\theta, t) dt, \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_2(\theta, t) dt \right\},$$

then for any $(\theta_1, t_1), (\theta_2, t_2) \in \mathbb{R}^k \times [0, 1]$ and $s > 0$, we have

- *Boundedness:*

$$\left| \nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}(\theta_1, t_1) \right| \leq 2|\theta_1| + C(\mathcal{L}), \quad (43)$$

- *Lipschitz continuity in θ and t :*

$$\begin{aligned} &\left| \nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}(\theta_1, t_1) - \nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}(\theta_2, t_2) \right| \\ &\leq C(\mathcal{L}) \left(|\theta_1 - \theta_2| + (|\theta_2| + 1)|t_1 - t_2|^{1/2} \right), \end{aligned} \quad (44)$$

- *Lipschitz continuity in ρ :*

$$\left| \nabla_\theta \frac{\delta E_s(\rho_1)}{\delta \rho}(\theta, t) - \nabla_\theta \frac{\delta E_s(\rho_2)}{\delta \rho}(\theta, t) \right| \leq C(\tilde{\mathcal{L}}) d_1(\rho_1, \rho_2) (|\theta| + 1), \quad (45)$$

where d_1 is defined in (2).

Proof From (31), we have

$$\nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}(\theta, t) = \mathbb{E}_{x \sim \mu} \left(p_\rho^\top(t; x) \partial_\theta f(Z_\rho(t; x), \theta) \right) + 2e^{-s}\theta, \quad (46)$$

which gives

$$\left| \nabla_\theta \frac{\delta E_s(\rho)}{\delta \rho}(\theta, t) \right| \leq \mathbb{E}_{x \sim \mu} (|\partial_\theta f(Z_\rho(t; x), \theta)| |p_\rho(t; x)|) + 2|\theta| \leq 2|\theta| + C(\mathcal{L}),$$

where we use (13), (22a), and (33a) in the second inequality. Thus, (43) is proved.

To prove (44), we use assume $t_1 > t_2$ without loss of generality, use (46), and reformulate bound the left-hand side of (44) as follows:

$$\begin{aligned}
 & \left| \nabla_{\theta} \frac{\delta E_s(\rho)}{\delta \rho}(\theta_1, t_1) - \nabla_{\theta} \frac{\delta E_s(\rho)}{\delta \rho}(\theta_2, t_2) \right| \\
 & \leq \underbrace{\left| \mathbb{E}_{x \sim \mu} \left(p_{\rho}^{\top}(t_1; x) \partial_{\theta} f(Z_{\rho}(t_1; x), \theta_1) - p_{\rho}^{\top}(t_1; x) \partial_{\theta} f(Z_{\rho}(t_1; x), \theta_2) \right) \right|}_{(I)} \\
 & \quad + \underbrace{\left| \mathbb{E}_{x \sim \mu} \left(p_{\rho}^{\top}(t_1; x) \partial_{\theta} f(Z_{\rho}(t_1; x), \theta_2) - p_{\rho}^{\top}(t_1; x) \partial_{\theta} f(Z_{\rho}(t_2; x), \theta_2) \right) \right|}_{(II)} \\
 & \quad + \underbrace{\left| \mathbb{E}_{x \sim \mu} \left(p_{\rho}^{\top}(t_1; x) \partial_{\theta} f(Z_{\rho}(t_2; x), \theta_2) - p^{\top}(t_2; x) \partial_{\theta} f(Z_{\rho}(t_2; x), \theta_2) \right) \right|}_{(III)} + 2e^{-s} |\theta_1 - \theta_2|.
 \end{aligned} \tag{47}$$

The bounds on all three terms in this expression rely on (12) with the mean-value theorem in Assumption 3.1. For (I), we have

$$\begin{aligned}
 |\partial_{\theta} f(Z_{\rho}(t_1; x), \theta_1) - \partial_{\theta} f(Z_{\rho}(t_1; x), \theta_2)| & \leq |\partial_{\theta}^2 f(Z_{\rho}(t_1; x), (1 - \lambda)\theta_1 + \lambda\theta_2)| |\theta_1 - \theta_2| \\
 & \leq C_1 (|Z_{\rho}(t_1; x)| + 1)^2 |\theta_1 - \theta_2| \\
 & \leq C(\mathcal{L}) |\theta_1 - \theta_2|,
 \end{aligned}$$

where we use (12) with mean-value theorem ($\lambda \in [0, 1]$) in the first two inequalities, and (22a) in the last inequality.

To bound term (II), we have from (25) and (27) that

$$\begin{aligned}
 |Z_{\rho}(t_1; x) - Z_{\rho}(t_2; x)| & \leq \left| \int_{t_2}^{t_1} \int_{\mathbb{R}^k} f(Z_{\rho}(t; x), \theta) d\rho(\theta, t) dt \right| \\
 & \leq \int_{t_2}^{t_1} \int_{\mathbb{R}^k} |f(Z_{\rho}(t; x), \theta)| d\rho(\theta, t) dt \\
 & \leq C \int_{t_2}^{t_1} \int_{\mathbb{R}^k} (|Z_{\rho}(t; x)| + 1)(|\theta| + 1) d\rho(\theta, t) dt \\
 & \stackrel{(a)}{\leq} C(\mathcal{L}) \left(\int_{t_2}^{t_1} \int_{\mathbb{R}^k} (|\theta| + 1) d\rho(\theta, t) dt \right) \\
 & \stackrel{(b)}{\leq} C(\mathcal{L}) \left(\int_{t_2}^{t_1} \int_{\mathbb{R}^k} (|\theta| + 1)^2 d\rho(\theta, t) dt \right)^{1/2} \left(\int_{t_2}^{t_1} \int_{\mathbb{R}^k} 1 d\rho(\theta, t) dt \right)^{1/2} \\
 & \stackrel{(c)}{\leq} C(\mathcal{L}) \sqrt{t_1 - t_2},
 \end{aligned}$$

where we use (22a) in (a), Hölder's inequality in (b), and $\int_{\mathbb{R}^k} d\rho(\theta, t) = 1$ for all $t \in [0, 1]$ in (c). This implies

$$\begin{aligned}
 (II) & \leq \mathbb{E}_{x \sim \mu} \left(|p_{\rho}^{\top}(t_1; x)| |\partial_{\theta} f(Z_{\rho}(t_1; x), \theta_2) - \partial_{\theta} f(Z_{\rho}(t_2; x), \theta_2)| \right) \\
 & \leq C(\mathcal{L}) \mathbb{E}_{x \sim \mu} (|\partial_{\theta} f(Z_{\rho}(t_1; x), \theta_2) - \partial_{\theta} f(Z_{\rho}(t_2; x), \theta_2)|) \\
 & \leq C(\mathcal{L}) \mathbb{E}_{x \sim \mu} (|\theta_2| |Z_{\rho}(t_1; x) - Z_{\rho}(t_2; x)|) \leq C(\mathcal{L}) |\theta_2| \sqrt{t_1 - t_2},
 \end{aligned}$$

where we use (33a) in the first inequality, $|\partial_x \partial_\theta f(x, \theta)| \leq C_1 |\theta|(|x| + 1)$ with mean-value theorem and (22a) in the second inequality.

To bound term (III), we also use (32) to obtain

$$\begin{aligned} |p_\rho(t_1; x) - p_\rho(t_2; x)| &\leq \left| \int_{t_2}^{t_1} \int_{\mathbb{R}^k} p_\rho^\top(t; x) \partial_z f(Z_\rho(t; x), \theta) d\rho(\theta, t) dt \right| \\ &\leq C_1 |p_\rho^\top(t; x)| \int_{t_2}^{t_1} \int_{\mathbb{R}^k} |\theta| d\rho(\theta, t) dt \\ &\leq C(\mathcal{L}) \left| \int_{t_2}^{t_1} \int_{\mathbb{R}^k} |\theta| d\rho(\theta, t) dt \right| \\ &\leq C(\mathcal{L}) \sqrt{t_1 - t_2}, \end{aligned}$$

where we use $|\partial_z f(Z_\rho(t; x), \theta)| \leq C_1 |\theta|$ in the second inequality, (33a) in the third inequality, and Hölder's inequality with $\int_{\mathbb{R}^k} d\rho(\theta, t) = 1$ in the last inequality. This implies

$$(III) \leq \mathbb{E}_{x \sim \mu} (|\partial_\theta f(Z_\rho(t_2; x), \theta_2)| |p_\rho(t_1; x) - p_\rho(t_2; x)|) \leq C(\mathcal{L}) \sqrt{t_1 - t_2},$$

where we also use $|\partial_\theta f(Z_\rho(t_2; x), \theta_2)| \leq C_1 (|Z_\rho(t_2; x)| + 1) \leq C(\mathcal{L})$ by (22a). These estimates together with (47) lead to (44).

Finally, to prove (45), we have from (31) that

$$\begin{aligned} &\left| \nabla_\theta \frac{\delta E_s(\rho_1)}{\delta \rho}(\theta, t) - \nabla_\theta \frac{\delta E_s(\rho_2)}{\delta \rho}(\theta, t) \right| \\ &= \mathbb{E}_{x \sim \mu} (|\partial_\theta f(Z_{\rho_1}(t; x), \theta) p_{\rho_1}(t; x) - \partial_\theta f(Z_{\rho_2}(t; x), \theta) p_{\rho_2}(t; x)|) \\ &\leq \mathbb{E}_{x \sim \mu} (|\partial_\theta f(Z_{\rho_1}(t; x), \theta) - \partial_\theta f(Z_{\rho_2}(t; x), \theta)| |p_{\rho_1}(t; x)|) \\ &\quad + \mathbb{E}_{x \sim \mu} (|\partial_\theta f(Z_{\rho_2}(t; x), \theta)| |p_{\rho_1}(t; x) - p_{\rho_2}(t; x)|) \\ &\leq C(\tilde{\mathcal{L}}) (|\theta| \mathbb{E}_{x \sim \mu} (|Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)|) + \mathbb{E}_{x \sim \mu} (|p_{\rho_1}(t; x) - p_{\rho_2}(t; x)|)) \\ &\leq C(\tilde{\mathcal{L}}) d_1(\rho_1, \rho_2) (|\theta| + 1). \end{aligned}$$

In the second inequality we use $|\partial_z \partial_\theta f(z, \theta)| \leq C_1 |\theta|(|z| + 1)$ with mean-value theorem and (22a) to obtain that

$$\begin{aligned} &|\partial_\theta f(Z_{\rho_1}(t; x), \theta) - \partial_\theta f(Z_{\rho_2}(t; x), \theta)| \\ &\leq |\partial_z \partial_\theta f((1 - \lambda)Z_{\rho_1}(t; x) + \lambda Z_{\rho_2}(t; x), \theta)| |Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)| \\ &\leq C(\tilde{\mathcal{L}}) |\theta| |Z_{\rho_1}(t; x) - Z_{\rho_2}(t; x)|, \end{aligned}$$

where $\lambda \in [0, 1]$. ■

Appendix D. Proof of Theorem 12 and Theorem 11

Theorem 12 states the well-posedness of equation (20), the continuous mean-field limit of the gradient flow. We verify that the solution is unique and the energy is nonincreasing.

The proof uses a fixed-point argument: We build a map and show that this map allows only a single fixed point. Fixing $S > 0$, for any $\phi(\theta, t, s) \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ and $\phi(\theta, t, 0) = \rho_{\text{ini}}(\theta, t)$, we define a map

$$\varphi = \mathcal{T}_S(\phi) : \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2)) \rightarrow \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2)) \quad (48)$$

where φ solves:

$$\begin{cases} \frac{\partial \varphi(\theta, t, s)}{\partial s} = \nabla_\theta \cdot \left(\varphi(\theta, t, s) \nabla_\theta \frac{\delta E_s(\phi(s))}{\delta \rho}(\theta, t) \right), \\ \varphi(\theta, t, 0) = \rho_{\text{ini}}(\theta, t). \end{cases} \quad (49)$$

The proof is divided into three steps:

Step 1: We show this map is well-defined and maps a $\mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ measure to a $\mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ measure. Furthermore, if the second moment of ϕ is bounded, then $\mathcal{T}(\phi)$ also has a bounded second moment, and we can specify such boundedness.

Step 2: We give a bound of $d_2(\mathcal{T}_S(\phi_1), \mathcal{T}_S(\phi_2))$ using $d_2(\phi_1, \phi_2)$. Furthermore, the associated energy is non-increasing.

By combining steps 1 and 2, we can find a small enough S so that \mathcal{T}_S is a contraction map in a complete subset of $\mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$. Then, according to the contraction map theorem, there is a unique distribution in the function space so that $\phi^* = \mathcal{T}_S(\phi^*)$, meaning that ϕ^* solves (49) itself, and is thus the unique solution to (20) for $s < S$.

Step 3: We extend the local solution to a global solution.

D.1 Step 1

For fixed $\phi(\theta, t, s)$, we define the gradient flow:

$$\begin{cases} \frac{d\theta_\phi(s; t)}{ds} = -\nabla_\theta \frac{\delta E_s(\phi(s))}{\delta \rho}(\theta, t) (\theta_\phi(s; t), t), \\ \theta_\phi(0; t) \sim \rho_{\text{ini}}(\theta, t). \end{cases} \quad (50)$$

Then $\varphi = \mathcal{T}_S(\phi)$ is the probability measure of θ_ϕ for any $s \in [0, S]$. Well-posedness of φ translates to the well-posedness of θ_ϕ .

According to (44) in Lemma 15, the force $\nabla_\theta \frac{\delta E_s(\phi(s))}{\delta \rho}(\theta, t)(\cdot, t)$ is a Lipschitz function with any $t \in [0, 1]$. The classical ODE theory then suggests there is a unique solution for $s \in [0, S]$, which depends continuously on the initial $\theta(0; t)$.

Denoting

$$\mathcal{L}_{S, \phi} = \sup_{0 \leq s \leq S} \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\phi(\theta, t, s) dt, \quad \mathcal{L}_{\text{ini}}^{\text{sup}} = \sup_{0 \leq t \leq 1} \int_{\mathbb{R}^k} |\theta|^2 d\rho_{\text{ini}}(\theta, t), \quad (51)$$

we have the following proposition.

Proposition 16 *Suppose that $\theta_\phi(0; t)$ is a random measurable path drawn from $\rho_{\text{ini}}(\theta, t)$ and that $\theta_\phi(s; t)$ is the solution to (50). Then for any $(t_1, s_1), (t_2, s_2) \in [0, 1] \times [0, S]$, we have*

$$\mathbb{E} \left(|\theta_\phi(s_1; t_1)|^2 \right) \leq \exp(5S) \left(\mathcal{L}_{\text{ini}}^{\text{sup}} + SC(\mathcal{L}_{S,\phi}) \right), \quad (52)$$

and

$$\begin{aligned} & \mathbb{E} \left(|\theta_\phi(s_1; t_1) - \theta_\phi(s_2; t_2)|^2 \right) \\ & \leq C \left(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S \right) \left(\mathbb{E} \left(|\theta_\phi(0; t_1) - \theta_\phi(0; t_2)|^2 \right) + |t_1 - t_2| + |s_1 - s_2|^2 \right), \end{aligned} \quad (53)$$

Proof To prove (52), we use (43) and (50) to obtain that

$$\frac{d |\theta_\phi(s; t_1)|^2}{ds} \leq 4 |\theta_\phi(s; t_1)|^2 + 2 |\theta_\phi(s; t_1)| C(\mathcal{L}_{S,\phi}) \leq 5 |\theta_\phi(s; t_1)|^2 + C(\mathcal{L}_{S,\phi}).$$

Using Grönwall's inequality, we obtain, for all $s_1 < S$

$$|\theta_\phi(s_1; t_1)|^2 \leq \exp(5S) |\theta_\phi(0; t_1)|^2 + \exp(5S) SC(\mathcal{L}_{S,\phi}).$$

By taking the expectation with respect to $\rho_{\text{ini}}(\theta, t)$, we have

$$\mathbb{E} \left(|\theta_\phi(s_1; t_1)|^2 \right) \leq \exp(5S) \mathbb{E} \left(|\theta_\phi(0; t_1)|^2 \right) + \exp(5S) SC(\mathcal{L}_{S,\phi}) \leq \exp(5S) (\mathcal{L}_{\text{ini}}^{\text{sup}} + SC(\mathcal{L}_{S,\phi})),$$

completing the proof of (52).

For the left-hand side of (53), we have

$$\begin{aligned} & \mathbb{E} \left(|\theta_\phi(s_1; t_1) - \theta_\phi(s_2; t_2)|^2 \right) \\ & \leq 2 \underbrace{\mathbb{E} \left(|\theta_\phi(s_1; t_1) - \theta_\phi(s_1; t_2)|^2 \right)}_{\text{(I)}} + 2 \underbrace{\mathbb{E} \left(|\theta_\phi(s_1; t_2) - \theta_\phi(s_2; t_2)|^2 \right)}_{\text{(II)}}, \end{aligned} \quad (54)$$

and bound the two terms (I) and (II) separately.

(I): From (50) and (44) in Lemma 15, we have

$$\frac{d |\theta_\phi(s; t_1) - \theta_\phi(s; t_2)|^2}{ds} \leq C(\mathcal{L}_{S,\phi}) |\theta_\phi(s; t_1) - \theta_\phi(s; t_2)|^2 + |\theta_\phi(s; t_1)|^2 |t_1 - t_2|,$$

which implies

$$\begin{aligned} & \mathbb{E} \left(|\theta_\phi(s_1; t_1) - \theta_\phi(s_1; t_2)|^2 \right) \\ & \leq C \left(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S \right) \left(\mathbb{E} \left(|\theta_\phi(0; t_1) - \theta_\phi(0; t_2)|^2 \right) + \left(\int_0^{s_1} \mathbb{E} |\theta_\phi(s; t_1)|^2 ds \right) |t_1 - t_2| \right) \\ & \leq C \left(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S \right) \left(\mathbb{E} \left(|\theta_\phi(0; t_1) - \theta_\phi(0; t_2)|^2 \right) + |t_1 - t_2| \right), \end{aligned} \quad (55)$$

where we use the bound (52) in the second inequality.

(II): This is estimated by integrating (50) from s_1 to s_2 and using the bound of $\nabla_{\theta} \frac{\delta E_s(\rho)}{\delta \rho}$ from (43). From the Grönwall inequality and (52), we have

$$|\theta_{\phi}(s_1; t_2) - \theta_{\phi}(s_2; t_2)| \leq C(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S) (|s_1 - s_2| + |e^{s_1} - e^{-s_2}|) \leq C(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S) |s_1 - s_2|.$$

The second inequality comes from the fact that $s_i < S$. This leads to

$$\mathbb{E} |\theta_{\phi}(s_1; t_2) - \theta_{\phi}(s_2; t_2)|^2 \leq C(\mathcal{L}_{S,\phi}, \mathcal{L}_{\text{ini}}^{\text{sup}}, S) |s_1 - s_2|^2. \quad (56)$$

By substituting (55) and (56) into (54), we complete the proof of (53). \blacksquare

An immediate corollary of Proposition 16 is that the map $\mathcal{T}_S(\cdot)$ is well-defined.

Corollary D.1 *For every $S > 0$, the map \mathcal{T}_S is well-defined. That is, for any $\phi \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$, one can find $\varphi = \mathcal{T}_S(\phi) \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ as the unique solution of (49). In particular, for any $(t, s) \in [0, 1] \times [0, S]$, we have*

$$\int_{\mathbb{R}^k} |\theta|^2 d\varphi(\theta, t, s) \leq \exp(5S) (\mathcal{L}_{\text{ini}}^{\text{sup}} + SC(\mathcal{L}_{S,\phi})), \quad (57)$$

where $C(\mathcal{L}_{S,\phi})$ is a quantity depending only on $\mathcal{L}_{S,\phi}$.

Proof For fixed $(t, s) \in [0, 1] \times [0, S]$, define $\varphi(\theta, t, s)$ as the distribution of $\theta_{\phi}(s; t)$. Using the classical stochastic theory (Ambrosio et al., 2008, Prop 8.1.8), $\varphi(\theta, t, s)$ is a solution to (49). For fixed $(t, s) \in [0, 1] \times [0, S]$, we prove (57) due to (52). Finally, using (14) and (53), we obtain that

$$\lim_{(t,s) \rightarrow (t_0,s_0)} W_2(\varphi(\cdot, t, s), \varphi(\cdot, t_0, s_0)) \leq \lim_{(t,s)} \left(\mathbb{E} \left(|\theta_{\phi}(s; t) - \theta_{\phi}(s_0; t_0)|^2 \right) \right)^{1/2} = 0,$$

which proves that $\varphi \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$. \blacksquare

D.2 Step 2.

We show first the contraction property.

Proposition 17 *For any $\phi_1, \phi_2 \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$, we have*

$$d_2(\mathcal{T}_S(\phi_1), \mathcal{T}_S(\phi_2)) \leq Q(\mathcal{L}_S) (S \exp(5S) (\mathcal{L}_{\text{ini}}^{\text{sup}} + SQ(\mathcal{L}_S)) + S)^{1/2} d_2(\phi_1, \phi_2), \quad (58)$$

where $Q : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is an increasing function and $\mathcal{L}_S = \max\{\mathcal{L}_{S,\phi_1}, \mathcal{L}_{S,\phi_2}\}$ with $\mathcal{L}_{S,\phi}$ defined in (51).

Proof Denote by $\theta_{\phi_i}(s; t)$ the solutions to (50) with $\phi = \phi_i$, and suppose that the initial conditions match, that is,

$$\theta_{\phi_1}(0; t) = \theta_{\phi_2}(0; t).$$

Similar to the previous subsection, we translate the study of φ_i to the study of θ_{ϕ_i} . Define

$$\Delta_t(s) = |\theta_{\phi_1}(s; t) - \theta_{\phi_2}(s; t)|,$$

we have

$$d_2(\mathcal{T}_S(\phi_1), \mathcal{T}_S(\phi_2)) \leq \sup_{(t,s) \in [0,1] \times [0,S]} \mathbb{E}(\Delta_t^2(s)).$$

Using (50), we have

$$\begin{aligned} \frac{d(\Delta_t(s))^2}{ds} &\leq 4(\Delta_t(s))^2 + 4 \left| \nabla_{\theta} \frac{\delta E_s(\phi_1(s))}{\delta \rho}(\theta_{\phi_1}, t) - \nabla_{\theta} \frac{\delta E_s(\phi_2(s))}{\delta \rho}(\theta_{\phi_2}, t) \right|^2 \\ &\leq 4(\Delta_t(s))^2 + 8 \left| \nabla_{\theta} \frac{\delta E_s(\phi_1(s))}{\delta \rho}(\theta_{\phi_1}, t) - \nabla_{\theta} \frac{\delta E_s(\phi_1(s))}{\delta \rho}(\theta_{\phi_2}, t) \right|^2 \\ &\quad + 8 \left| \nabla_{\theta} \frac{\delta E_s(\phi_1(s))}{\delta \rho}(\theta_{\phi_2}, t) - \nabla_{\theta} \frac{\delta E_s(\phi_2(s))}{\delta \rho}(\theta_{\phi_2}, t) \right|^2. \end{aligned} \quad (59)$$

The second term on the right-hand side concerns the continuity addressed in Lemma 15 (44). To bound the last term, we use (45) to obtain

$$\begin{aligned} &\left| \nabla_{\theta} \frac{\delta E_s(\phi_1(s))}{\delta \rho}(\theta_{\phi_2}, t) - \nabla_{\theta} \frac{\delta E_s(\phi_2(s))}{\delta \rho}(\theta_{\phi_2}, t) \right| \\ &\leq C(\mathcal{L}_S) d_1(\phi_1(s), \phi_2(s)) (|\theta_{\phi_2}| + 1) \leq C(\mathcal{L}_S) d_2(\phi_1, \phi_2) (|\theta_{\phi_2}| + 1), \end{aligned}$$

where we use (12), (22a), (33a) in the first inequality and (22b), (33b) in the second inequality, with d_2 defined in Definition 3. By substituting into (59), we obtain

$$\frac{d(\Delta_t(s))^2}{ds} \leq C(\mathcal{L}_S) [(\Delta_t(s))^2 + d_2^2(\phi_1, \phi_2) (|\theta_{\phi_2}|^2 + 1)],$$

Using the Grönwall inequality, we then have

$$\begin{aligned} \mathbb{E}((\Delta_t(s))^2) &\leq C(\mathcal{L}_S) \left(\int_0^s \mathbb{E}|\theta_{\phi_2}(u; t)|^2 du + s \right) d_2^2(\phi_1, \phi_2) \\ &\leq C(\mathcal{L}_S) d_2^2(\phi_1, \phi_2) (S \exp(5S) (\mathcal{L}_{\text{ini}}^{\text{sup}} + SC(\mathcal{L}_S)) + s), \end{aligned}$$

completing the proof. ■

We are now ready to run the contraction-map argument that justifies the existence and uniqueness of the local solution. According to the contraction mapping theorem, we need to verify two conditions in order to show that there is a $\phi^* = \mathcal{T}_S(\phi^*)$:

- There is a closed subset in $C([0, S]; C([0, 1], W_2))$ so that \mathcal{T}_S maps it to itself; and
- \mathcal{T}_S is a contraction map in this subset.

We define the closed subset next.

Definition 18

$$\mathcal{B}_{\rho_0} = \left\{ \phi \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2)) \left| \sup_{t \in [0, 1], 0 \leq s \leq S} \int_{\mathbb{R}^k} |\theta|^2 d\phi(\theta, t, s) \leq 4\mathcal{L}_{\text{ini}}^{\text{sup}} \right. \right\}.$$

We now claim that for small enough S , \mathcal{T}_S is a contraction map in \mathcal{B}_{ρ_0} .

Proposition 19 *Suppose that S is small enough that*

$$\begin{aligned} \exp(5S)(\mathcal{L}_{\text{ini}}^{\text{sup}} + SQ(4\mathcal{L}_{\text{ini}}^{\text{sup}})) &< 4\mathcal{L}_{\text{ini}}^{\text{sup}}, \\ Q(4\mathcal{L}_{\text{ini}}^{\text{sup}})(S \exp(5S)(\mathcal{L}_{\text{ini}}^{\text{sup}} + SQ(4\mathcal{L}_{\text{ini}}^{\text{sup}})) + S)^{1/2} &< \frac{1}{2}, \end{aligned} \quad (60)$$

where Q comes from Proposition 17. Then the following are true.

- If $\phi \in \mathcal{B}_{\rho_0}$, then $\mathcal{T}_S(\phi) \in \mathcal{B}_{\rho_0}$, that is,

$$\sup_{t \in [0, 1], 0 \leq s \leq S} \int_{\mathbb{R}^k} |\theta|^2 d\mathcal{T}(\phi)(\theta, t, s) \leq 4\mathcal{L}_{\text{ini}}^{\text{sup}} \quad (61)$$

- \mathcal{T}_S is a contraction map in this subset, meaning that for any $\phi_1, \phi_2 \in \mathcal{B}_{\rho_0}$, we have

$$d_2(\mathcal{T}_S(\phi_1), \mathcal{T}_S(\phi_2)) \leq \frac{1}{2} d_2(\phi_1, \phi_2). \quad (62)$$

Proof First, using Corollary D.1 (57), we have

$$\int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\mathcal{T}(\phi_1)(\theta, t, s) dt \leq \exp(5S)(\mathcal{L}_{\text{ini}}^{\text{sup}} + SQ(4\mathcal{L}_{\text{ini}}^{\text{sup}})) \leq 4\mathcal{L}_{\text{ini}}^{\text{sup}},$$

which proves (61). Then, using (58), we have

$$d_2(\mathcal{T}_S(\phi_1), \mathcal{T}_S(\phi_2)) \leq Q(4\mathcal{L}_{\text{ini}}^{\text{sup}})(S \exp(5S)(\mathcal{L}_{\text{ini}}^{\text{sup}} + SQ(4\mathcal{L}_{\text{ini}}^{\text{sup}})) + S)^{1/2} d_2(\phi_1, \phi_2) < \frac{1}{2} d_2(\phi_1, \phi_2),$$

which proves (62). ■

Using the contraction mapping theorem, we can obtain directly that $\mathcal{T}_S(\phi)$ has a fixed point in \mathcal{B}_{ρ_0} when S is small enough.

Corollary D.2 *If S satisfies (60), then there exists $\phi^*(\theta, t, s) \in \mathcal{B}_{\rho_0} \subset \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ such that $\phi^*(\theta, t, s)$ is a solution to (20) with initial condition $\rho_{\text{ini}}(\theta, t)$ and admissible for each s .*

Proof According to Proposition 19, \mathcal{T}_S is a contraction map in \mathcal{B}_{ρ_0} . The result then follows immediately from the contraction mapping theorem and the admissibility of $\mathcal{T}_S(\phi)$ by its definition. ■

Finally, we prove that the cost function decays along the flow.

Lemma 20 Fix $S > 0$. If $\phi^*(\theta, t, s) \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ such that $\phi^*(\theta, t, s)$ is a solution to (20) with initial condition $\rho_{\text{ini}}(\theta, t)$, then for $0 < s < S$, we have

$$\frac{dE_s(\phi^*(\cdot, \cdot, s))}{ds} \leq 0.$$

Proof Denote by $\theta^*(s; t)$ the solution to (50) with $\phi = \phi^*$. Then since ϕ^* is the unique solution up to $s \leq S$, the probability distribution of θ^* is ϕ^* itself.

According to (20), we obtain formally that

$$\begin{aligned} & \frac{dE_s(\phi^*(\cdot, \cdot, s))}{ds} \\ &= - \int_0^1 \int_{\mathbb{R}^k} \left| \nabla_{\theta} \frac{\delta E_s(\phi^*(s))}{\delta \rho}(\theta, t) \right|^2 d\phi^*(\theta, t, s) dt - \exp^{-s} \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\phi^*(\theta, t, s) dt \\ &= - \int_0^1 \mathbb{E} \left(\left| \nabla_{\theta} \frac{\delta E_s(\phi^*(s))}{\delta \rho}(\theta^*(s; t), t) \right|^2 \right) dt - \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt \\ &\leq 0, \end{aligned} \tag{63}$$

which proves the result. We note that the rigorous proof of (63) is in Appendix J. ■

D.3 Step 3:

In this final step of the proof, we extend the local solution of Corollary D.2 to a global solution. Since Lemma 20 shows that E_s is decaying with s , we can then improve the bound for the second moment of the solution (57). We have the following corollary.

Corollary D.3 For fixed $S > 0$, denote by $\phi^*(\theta, t, s) \in \mathcal{C}([0, S]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ the solution to (20) with initial condition $\rho_{\text{ini}}(\theta, t)$. Then for any $(t, s) \in [0, 1] \times [0, S]$, we have

$$\int_{\mathbb{R}^k} |\theta|^2 d\phi^*(\theta, t, s) \leq \exp(5s)(\mathcal{L}_{\text{ini}}^{\text{sup}} + C(s)), \tag{64}$$

where the quantity C depends only on s .

Proof Using $E_s(\phi^*(\theta, t, s)) < E_0(\phi^*(\theta, t, 0))$, we first have

$$\sup_{0 \leq u \leq s} \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\phi^*(\theta, t, u) dt \leq \exp(s) E_0(\rho_{\text{ini}}(\theta, t)).$$

Then, using (57), we obtain that

$$\int_{\mathbb{R}^k} |\theta|^2 d\mathcal{T}(\phi^*)(\theta, t, s) \leq \exp(5s)(\mathcal{L}_{\text{ini}}^{\text{sup}} + sC(\exp(s)E_0(\rho_{\text{ini}}(\theta, t)))),$$

proving (64). ■

This corollary takes an important step. By contrast with (57), it removes the dependence of the bound on $\mathcal{L}_{S,\phi}$, so the fixed-point argument need no longer depend on the initial guess ϕ .

We are now ready to prove the Theorem 12.

Proof [Proof of Theorem 12]

From Corollary D.2, if S_1 satisfies (60), we obtain a local solution $\phi^* \in C([0, S_1]; \mathcal{C}([0, 1]; \mathcal{P}^2))$ to (20), which is admissible for any s .

Denote by S^* the supremum of such constants for which the solution exists. We hope to show that $S^* = \infty$. If not, then denote the solution by $\phi^* \in C([0, S^*]; \mathcal{C}([0, 1]; \mathcal{P}^2))$. According to (64), for any $s < S^*$ and $t \in [0, 1]$, we have

$$\int_{\mathbb{R}^k} |\theta|^2 d\phi^*(\theta, t, s) \leq \exp(5S^*)(\mathcal{L}_{\text{ini}}^{\text{sup}} + Q_2(S^*)) = \mathcal{L}^{*,\text{sup}}.$$

Let us choose Δ_{S^*} small enough to satisfy

$$\begin{aligned} \exp(5\Delta_{S^*})(\mathcal{L}^{*,\text{sup}} + \Delta_{S^*}Q(4\mathcal{L}^{*,\text{sup}})) &< 4\mathcal{L}^{*,\text{sup}}, \\ Q(4\mathcal{L}^{*,\text{sup}})(\Delta_{S^*} \exp(5\Delta_{S^*})(\mathcal{L}^{*,\text{sup}} + \Delta_{S^*}Q(4\mathcal{L}^{*,\text{sup}})) + \Delta_{S^*})^{1/2} &< \frac{1}{2}. \end{aligned}$$

Then, using Proposition 19 and Corollary D.2, we can further extend ϕ^* to be supported on $C([0, S^* + \Delta_{S^*}]; \mathcal{C}([0, 1]; \mathcal{P}^2))$. This contradicts the assumption that we have to stop at a finite value of S^* . Thus $S^* = \infty$.

Finally, (23) is a direct result of Lemma 20. ■

D.4 Proof of Theorem 11

In this section, we consider the finite layer case (1) and prove Theorem 11. In (2), $\nabla_{\Theta_{L,M}} E(\Theta_{L,M})$ can also be written as

$$\frac{\partial E(\Theta_{L,M})}{\partial \theta_{l,m}} = \frac{1}{ML} \mathbb{E}_{x \sim \mu} (\partial_{\theta} f(Z_{\Theta_{L,M}}(l; x), \theta_{l,m}) p_{\Theta_{L,M}}(l; x)), \quad (65)$$

where $p_{\Theta_{L,M}}(l; x)$ can be solved by the following iteration formula:

$$\begin{cases} p_{\Theta_{L,M}}^{\top}(l; x) = p_{\Theta_{L,M}}^{\top}(l+1; x) \left(I + \frac{1}{ML} \sum_{m=1}^M \partial_z f(Z_{\Theta_{L,M}}(l+1; x), \theta_{l+1,i}) \right), \\ p_{\Theta_{L,M}}(L-1; x) = (g(Z_{\Theta_{L,M}}(L; x)) - y(x)) \nabla g(Z_{\Theta_{L,M}}(L; x)) \end{cases} \quad (66)$$

for $0 \leq l \leq L-2$. Similar to Definitions 2 and 3, we make the following definition.

Definition 21 $\Theta_{L,M} = \{\theta_{l,m}\}_{l=0,m=1}^{L-1,M} \in L_{L,M}^{\infty}$ if and only if

$$\sup_{l,m} |\theta_{l,m}| < \infty.$$

The metric in $L_{L,M}^{\infty}$ is defined as

$$d_{1,L,M}(\Theta_{L,M}, \tilde{\Theta}_{L,M}) = \max_l \left(\frac{1}{M} \sum_{m=1}^M |\theta_{l,m} - \tilde{\theta}_{l,m}|^2 \right)^{1/2}.$$

Definition 22 For $s \geq 0$, we have $\Theta_{L,M}(s) = \{\theta_{l,m}(s)\}_{l=0,m=1}^{L-1,M} \in \mathcal{C}([0, \infty); L_{L,M}^\infty)$ if and only if

1. For fixed $s \in [0, \infty)$, $\Theta_{L,M}(s) \in L_{L,M}^\infty$, and
2. For any $s_0 \in [0, \infty)$, we have

$$\lim_{s \rightarrow s_0} d_{1,L,M}(\Theta_{L,M}(s), \Theta_{L,M}(s_0)) = 0,$$

where $d_{1,L,M}$ is defined in Definition 21.

The metric in $\mathcal{C}([0, \infty); L_{L,M}^\infty)$ is defined by

$$d_{2,L,M}(\Theta_{L,M}, \tilde{\Theta}_{L,M}) = \sup_s d_{1,L,M}(\Theta_{L,M}(s), \tilde{\Theta}_{L,M}(s)).$$

Theorem 11 is a corollary of the following result.

Theorem 23 Suppose that Assumption 3.1 holds. If $\Theta_{L,M}(0) \in L_{L,M}^\infty$ then (18) has a unique solution in $\mathcal{C}([0, \infty); L_{L,M}^\infty)$.

Before proving the Theorem 23, we first prove boundedness and stability of $Z_{\Theta_{L,M}}$ and $p_{\Theta_{L,M}}$, similarly to Theorem 10 and Lemma 14.

Lemma 24 Suppose that Assumption 3.1 holds and that x in the support of μ . Let

$$\tilde{\Theta}_{L,M} = \left\{ \tilde{\theta}_{l,m} \right\}_{l=0,m=1}^{L-1,M}$$

be a set of parameters, in addition to $\Theta_{L,M}$, and denote

$$\mathcal{L}_{\Theta_{L,M}} = \frac{1}{LM} \sum_{l=0}^{L-1} \sum_{m=1}^M |\theta_{l,m}|^2, \quad \mathcal{L}_{\tilde{\Theta}_{L,M}} = \frac{1}{LM} \sum_{l=0}^{L-1} \sum_{m=1}^M |\tilde{\theta}_{l,m}|^2.$$

Then there exist constants $C(\mathcal{L}_{\Theta_{L,M}})$ depending on $\mathcal{L}_{\Theta_{L,M}}$ and $C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}})$ depending on $\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}}$ such that for all $0 \leq l \leq L-1$, we have the following bounds:

$$|Z_{\Theta_{L,M}}(l+1; x)| \leq C(\mathcal{L}_{\Theta_{L,M}}), \quad (67a)$$

$$\left| Z_{\Theta_{L,M}}(l+1; x) - Z_{\tilde{\Theta}_{L,M}}(l+1; x) \right| \leq C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}}) d_{1,L,M}(\Theta_{L,M}, \tilde{\Theta}_{L,M}), \quad (67b)$$

$$|p_{\Theta_{L,M}}(l; x)| \leq C(\mathcal{L}_{\Theta_{L,M}}), \quad (68a)$$

$$\left| p_{\Theta_{L,M}}(l; x) - p_{\tilde{\Theta}_{L,M}}(l; x) \right| \leq C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}}) d_{1,L,M}(\Theta_{L,M}, \tilde{\Theta}_{L,M}). \quad (68b)$$

Proof Using (1) and (27), we obtain that

$$\begin{aligned} (|Z_{\Theta_{L,M}}(l+1; x)| + 1) &\leq C_1 \left(1 + \frac{1}{LM} \sum_{m=1}^M |\theta_{l,m}| \right) (|Z_{\Theta_{L,M}}(l; x)| + 1) \\ &\leq C_1 \exp \left(\frac{1}{LM} \sum_{m=1}^M |\theta_{l,m}| \right) (|Z_{\Theta_{L,M}}(l; x)| + 1), \end{aligned}$$

which proves (67a) by iteration.

Similarly, using (12), we obtain that

$$\frac{1}{ML} \sum_{m=1}^M |\partial_z f(Z_{\Theta_{L,M}}(l+1; x), \theta_{l,m})| \leq \frac{C_1}{ML} \sum_{m=1}^M |\theta_{l,m}|,$$

which by (66) implies that

$$|p_{\Theta_{L,M}}(l; x)| \leq \left(1 + \frac{C_1}{LM} \sum_{m=1}^M |\theta_{l,m}| \right) |p_{\Theta_{L,M}}(l+1; x)|.$$

From this bound, together with $|p_{\Theta_{L,M}}(x, L-1)| \leq C|Z_{\Theta_{L,M}}(L; x)| \leq C(\mathcal{L}_{\Theta_{L,M}})$, we prove (68a) by iteration.

To prove (67b), similarly to (30), we use (1), (12), and (67a) to obtain

$$\begin{aligned} &\left| Z_{\Theta_{L,M}}(l+1; x) - Z_{\tilde{\Theta}_{L,M}}(l+1; x) \right| \\ &\leq \left(1 + \frac{C_1}{ML} \sum_{m=1}^M |\theta_{l,m}| \right) \left| Z_{\Theta_{L,M}}(l; x) - Z_{\tilde{\Theta}_{L,M}}(l; x) \right| + \frac{C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}})}{ML} \sum_{m=1}^M |\theta_{l,m} - \tilde{\theta}_{l,m}|. \end{aligned}$$

By using this bound iteratively in conjunction with the boundary condition $|Z_{\Theta_{L,M}}(0; x) - Z_{\tilde{\Theta}_{L,M}}(0; x)| = 0$, we prove (67b).

To prove (68b), similarly to the proof of Lemma 14 (37)-(42), we use (12) and (66)-(68a) to obtain

$$\begin{aligned} &\left| p_{\Theta_{L,M}}(l; x) - p_{\tilde{\Theta}_{L,M}}(l; x) \right| \\ &\leq \left| (p_{\Theta_{L,M}}(l+1; x) - p_{\tilde{\Theta}_{L,M}}(l+1; x))^{\top} \left(I + \frac{1}{ML} \sum_{m=1}^M \partial_z f(Z_{\Theta_{L,M}}(l+1; x), \theta_{l+1,m}) \right) \right| \\ &\quad + \left| p_{\tilde{\Theta}_{L,M}}^{\top}(l+1; x) \left(\frac{1}{ML} \sum_{m=1}^M \left(\partial_z f(Z_{\Theta_{L,M}}(l+1; x), \theta_{l+1,i}) - \partial_z f(Z_{\tilde{\Theta}_{L,M}}(l+1; x), \tilde{\theta}_{l+1,m}) \right) \right) \right| \\ &\leq C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}}) \left(\frac{1}{ML} \sum_{m=1}^M |\theta_{l,m}| + 1 \right) |p_{\Theta_{L,M}}(l+1; x) - p_{\tilde{\Theta}_{L,M}}(l+1; x)| \\ &\quad + C(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}}) \left(\frac{1}{ML} \sum_{m=1}^M |\theta_{l,m}|^2 + |\tilde{\theta}_{l,m}|^2 + \frac{1}{L} \right) d_{1,L,M}(\Theta_{L,M}, \tilde{\Theta}_{L,M}). \end{aligned} \tag{69}$$

Similarly to (36), we have

$$\begin{aligned} |p_{\Theta_{L,M}}(L-1; x) - p_{\tilde{\Theta}_{L,M}}(L-1; x)| &\leq C |Z_{\Theta_{L,M}}(L; x) - Z_{\tilde{\Theta}_{L,M}}(L; x)| \\ &\leq C \left(\mathcal{L}_{\Theta_{L,M}}, \mathcal{L}_{\tilde{\Theta}_{L,M}} \right) d_{1,L,M} \left(\Theta_{L,M}, \tilde{\Theta}_{L,M} \right). \end{aligned}$$

By combining with (69), we prove (68b) by iteration. \blacksquare

We are now ready to prove Theorem 23, and thus also its corollary, Theorem 11.

Proof [Proof of Theorem 23] The proof is quite similar to that of Theorem 12, so we omit some details.

Recall (18):

$$\frac{d\Theta_{L,M}(s)}{ds} = -ML\nabla_{\Theta_{L,M}} E(\Theta_{L,M}(s)) - 2e^{-s}\Theta_{L,M}(s) = -ML\nabla_{\Theta_{L,M}} E_s(\Theta_{L,M}(s)).$$

Similarly to Lemma 15, using (65) and Lemma 24, we have the following bound and stability inequalities for $ML\nabla_{\Theta_{L,M}} E_s(\Theta_{L,M})$:

$$|ML\nabla_{\Theta_{L,M}} E_s(\Theta_{L,M})| \leq 2|\Theta_{L,M}| + C(\mathcal{L}) \quad (70a)$$

$$\left| ML\nabla_{\Theta_{L,M}} E_s(\Theta_{L,M}) - ML\nabla_{\tilde{\Theta}_{L,M}} E_s(\tilde{\Theta}_{L,M}) \right| \leq C(\mathcal{L}) |\Theta_{L,M} - \tilde{\Theta}_{L,M}|, \quad (70b)$$

where

$$\mathcal{L} = \max \left\{ \frac{1}{LM} \sum_{l=0}^{L-1} \sum_{m=1}^M |\theta_{l,m}|^2, \frac{1}{LM} \sum_{l=0}^{L-1} \sum_{m=1}^M |\tilde{\theta}_{l,m}|^2 \right\}.$$

Similarly to Step 1 of the proof in Section D, we can first construct a map induced by (18). Using (70a) and (70b), similar to Step 1 in the earlier proof, we can show that this map is well-defined. Further, similar to Step 2 in Section D, we can show that this map is locally contracted in $\mathcal{C}([0, \infty); L_{L,M}^\infty)$. Finally, for a local solution to (18), $\frac{dE_s}{ds} \leq 0$ is clearly true, which implies that the proof in Step 3 of Section D can also be applied here. Thus, we obtain a global solution. \blacksquare

Appendix E. Long-time Equilibrium and the Convergence to the Global Minimizer

We prove Theorem 7 in this section. The strategy is as follows. First, we show that for any ρ for which $E(\rho) > 0$, we can find a direction ν so that the Fréchet derivative of $E(\cdot)$ along this direction is negative, meaning that an infinitesimal change to ρ in direction ν will decrease E . This fact suggests that any stable point ρ , which has $\frac{\delta E(\rho)}{\delta \rho} = 0$, must be a global minimizer of E , with $E(\rho) = 0$. This part of the discussion presented in Section E.1, Proposition 26. Second, in Section E.2, we show that such a minimizer of E can be obtained as the long-time equilibrium of (20).

Note that we assume throughout this section that Assumption 3.1 holds (it being one of the assumptions in Theorem 7).

E.1 $\frac{\delta E(\rho)}{\delta \rho} = 0$ vs. $E(\rho) = 0$

We prove here that if ρ is a stable point for E , then it is actually a global minimizer. We start by proving a lower bound for p_ρ .

Lemma 25 *Suppose that $\rho \in \mathcal{C}([0, 1]; \mathcal{P}^2)$ and that p_ρ is a solution to (11). Denoting*

$$\mathcal{L}_\rho = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t),$$

then for any $t \in [0, 1]$ we have

$$\mathbb{E}_{x \sim \mu} (|p_\rho(t; x)|^2) \geq Q_2(\mathcal{L}_\rho) E(\rho), \quad (71)$$

where $Q_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a decreasing function.

Proof Recall the initial condition for p_ρ in (11):

$$p_\rho(1; x) = (g(Z_\rho(1; x)) - y(x)) \nabla g(Z_\rho(1; x)).$$

According to Assumption 3.1 (part 3) and the definition (8) we therefore have

$$\mathbb{E}_{x \sim \mu} (|p_\rho(1; x)|^2) \geq \delta_1^2 E(\rho).$$

Further, since the equation (11) is linear, that is,

$$\frac{\partial p_\rho^\top}{\partial t} = -p_\rho^\top \int_{\mathbb{R}^k} \partial_z f(Z_\rho, \theta) d\rho(\theta, t), \quad (72)$$

we have

$$\frac{d|p_\rho(t; x)|^2}{dt} \leq \left(2C_1 \int_{\mathbb{R}^k} |\theta| d\rho(\theta, t) \right) |p_\rho(t; x)|^2,$$

where we used (34) to bound $|\int_{\mathbb{R}^k} \partial_z f(Z_\rho(t; x), \theta) d\rho(\theta, t)|$. By solving this equation, we obtain

$$|p_\rho(t; x)|^2 \geq |p_\rho(1; x)|^2 \exp \left(-2C_1 \int_t^1 \int_{\mathbb{R}^k} |\theta| d\rho(\theta, t) \right) \geq C(\mathcal{L}_\rho) |p_\rho(1; x)|^2.$$

We finalize the proof by taking expectation of both sides. (Monotonicity is a consequence of the format of the exponential term.) ■

The following proposition shows existence of a descent direction from any ρ for which $E(\rho) > 0$.

Proposition 26 *Suppose that $\rho \in \mathcal{C}([0, 1]; \mathcal{P}^2)$. If $E(\rho) > 0$, then for any $t_0 \in [0, 1]$, there exists a measure $\nu(\theta)$ of \mathbb{R}^k such that $\int_{\mathbb{R}^k} d\nu(\theta) = 0$ and*

$$\int_{\mathbb{R}^k} \frac{\delta E(\rho)}{\delta \rho}(\theta, t_0) d\nu(\theta) < 0. \quad (73)$$

Proof Denote

$$\mathcal{L}_\rho = \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t) dt.$$

According to the existence and uniqueness of solution to (7), for any $t \in [0, 1]$, we can construct a map \mathcal{Z}_t such that

$$\mathcal{Z}_t(x) = Z_\rho(t; x).$$

Since the trajectory can be computed backwards in time, \mathcal{Z}_t^{-1} is well-defined. Furthermore, we denote $\mu_t^* = (\mathcal{Z}_t)_\# \mu$ the pushforward of μ under map \mathcal{Z}_t and let:

$$p^*(t; x) = p_\rho(t; \mathcal{Z}_t^{-1}(x)).$$

By Assumption 3.1 and classical ODE theory, \mathcal{Z}_t and \mathcal{Z}_t^{-1} are both continuous maps in x , and so are $p_\rho(t; x)$ and $p^*(t; x)$. With the change of variables, for all $t \in [0, 1]$

$$\frac{\delta E(\rho)}{\delta \rho}(\theta, t) = \int_{\mathbb{R}^d} p_\rho^\top(t; x) f(\mathcal{Z}_t(x), \theta) d\mu = \int_{\mathbb{R}^d} (p^*(t; x))^\top f(x, \theta) d\mu_t^*. \quad (74)$$

For a fixed $t_0 \in [0, 1]$, calling Lemma 14, we have the boundedness of the Jacobian, meaning $\sup_{x \in \text{supp}(\mu)} \left\| \frac{d\mu_{t_0}^*(\mathcal{Z}_{t_0}^{-1}(x))}{d\mu(x)} \right\|_2 \leq C(\mathcal{L}_\rho)$. As a consequence, $\mu_{t_0}^*(x)$ has a compact support since $\mu(x)$ has one. We denote the size of the support R^* , meaning $\text{supp}(\mu_{t_0}^*(x)) \subset \{x : |x| < R^*\}$.

We now derive a general formula for $\int \frac{\delta E(\rho)}{\delta \rho}(\theta, t) d\nu$. Recall (74), we have:

$$\begin{aligned} & \int_{\mathbb{R}^k} \frac{\delta E(\rho)}{\delta \rho}(\theta, t_0) d\nu(\theta) \\ &= \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(\int_{\mathbb{R}^k} f(x, \theta) d\nu(\theta) \right) d\mu_{t_0}^*(x) \\ &= \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(\int_{\mathbb{R}^k} f(x, \theta) d\nu(\theta) + p^*(t_0; x) \right) d\mu_{t_0}(x) \\ &\quad - \int_{\mathbb{R}^d} (p^*(t_0; x))^\top p^*(t_0; x) d\mu_{t_0}^*(x) \\ &= \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(\int_{\mathbb{R}^k} f(x, \theta) d\rho(\theta) + p^*(t_0; x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \nu)(\theta) \right) d\mu_{t_0}(x) \\ &\quad - \int_{\mathbb{R}^d} (p^*(t_0; x))^\top p^*(t_0; x) d\mu_{t_0}^*(x). \end{aligned} \quad (75)$$

Noticing that according to Lemma 25, if $E(\rho) \neq 0$, the second term above is strictly negative ($< -Q_2(\mathcal{L}_\rho)E(\rho)$), the goal then is to find ν so that $\int d\nu = 0$ and that the first term is rather small. This makes the full term $\int_{\mathbb{R}^k} \frac{\delta E(\rho)}{\delta \rho}(\theta, t_0) d\nu(\theta)$ negative.

Denote a continuous function

$$h(x) = \int_{\mathbb{R}^k} f(x, \theta) d\rho(\theta, t_0) + p^*(t_0; x),$$

then according to Assumption 3.1, for arbitrarily small ϵ , there is a $\hat{\nu}$ so that $\int \hat{\nu} = 0$ and

$$\left\| h(x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right\|_{L_{|x| < R^*}^\infty} \leq \epsilon.$$

This implies

$$\begin{aligned}
& \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(\int_{\mathbb{R}^k} f(x, \theta) d\rho(\theta) + p^*(t_0; x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right) d\mu_{t_0}(x) \\
&= \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(h(x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right) d\mu_{t_0}(x) \\
&\leq \int_{\mathbb{R}^d} |p^*(t_0; x)| \left| h(x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right| d\mu_{t_0}^*(x) \\
&\leq \|p^*(t_0; x)\|_{L_{|x| < R^*}^\infty} \left\| h(x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right\|_{L_{|x| < R^*}^\infty} \leq \epsilon \|p^*(t_0; x)\|_{L_{|x| < R^*}^\infty}
\end{aligned} \tag{76}$$

Choose ϵ small enough so that

$$\begin{aligned}
& \int_{\mathbb{R}^d} (p^*(t_0; x))^\top \left(\int_{\mathbb{R}^k} f(x, \theta) d\rho(\theta) + p^*(t_0; x) - \int_{\mathbb{R}^k} f(x, \theta) d(\rho - \hat{\nu})(\theta) \right) d\mu_{t_0}(x) \\
&\leq \epsilon \|p^*(t_0; x)\|_{L_{|x| < R^*}^\infty} \leq \frac{1}{2} Q(\mathcal{L}_\rho) E(\rho).
\end{aligned} \tag{77}$$

Let $\nu = \hat{\nu}$. Plugging (77) into (75) and using Lemma 25 (71), we have

$$\int \frac{\delta E(\rho)}{\delta \rho}(\theta, t_0) d\nu(\theta) < 0,$$

which finishes the proof. ■

E.2 Proof of Theorem 7

Proof [Proof of Theorem 7] If $\rho(\theta, t, s)$ converges to $\rho_\infty(\theta, t)$ in $\mathcal{C}([0, 1]; \mathcal{P}^2)$, then we have

$$\sup_{s \geq 0, t \in [0, 1]} \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t, s) < \infty.$$

Since $\rho(\theta, t, s)$ is a weak solution to (20) and it converges to $\rho_\infty(\theta, t)$, we obtain that $\partial_s \rho|_{\rho_\infty} = 0$, so that³

$$\nabla \cdot \left(\rho_\infty \nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) \right) = 0, \quad a.e. \tag{78}$$

From (12), (27), and (33a),

$$\left| \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) \right| \leq C(|\theta| + 1), \quad \left| \nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) \right| \leq C.$$

3. This statement can be made rigorous. Since $\lim_{s \rightarrow \infty} d_1(\rho(\theta, t, s), \rho_\infty(\theta, t)) = 0$, we have from (45) that $\lim_{s \rightarrow \infty} \nabla \frac{\delta E_s(\rho)}{\delta \rho(s)}(\theta, t) = \nabla \frac{\delta E(\rho)}{\delta \rho_\infty}(\theta, t)$. Due to the bound (45), the convergence is uniform in $(\theta, t) \in \mathcal{B}_r \times [0, 1]$ for any $r > 0$. Following the protocol for weak solution, we multiply smooth compact support test functions on both sides of (20) and integrate. As $s \rightarrow \infty$, the integrand converges uniformly, leading to (78).

Multiplying $\frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t)$ on the left hand-side of (78) and integrating by parts, since $\rho_\infty(\theta, t)$ in $\mathcal{C}([0, 1]; \mathcal{P}^2)$, the boundary term vanishes. And we obtain

$$\int_0^1 \int_{\mathbb{R}^k} \left| \nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) \right|^2 d\rho_\infty dt = 0,$$

From continuity of $\nabla \frac{\delta E(\rho_\infty)}{\delta \rho}$ and $\rho_\infty(\theta, t)$ in $\mathcal{C}([0, 1]; \mathcal{P}^2)$, $\int_{\mathbb{R}^k} \left| \nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) \right|^2 d\rho_\infty$ is continuous in t . Thus, we finally have

$$\rho_\infty(\theta, t) \nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t) = 0, \quad \forall t \in [0, 1]$$

When the support of $\rho_\infty(\theta, t_0)$ is \mathbb{R}^k , we have from continuity of $\nabla \frac{\delta E(\rho_\infty)}{\delta \rho}$ in (44) that

$$\nabla \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t_0) = 0, \quad \forall \theta \in \mathbb{R}^k,$$

implying that there is a constant c such that

$$\frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t_0) = c, \quad \forall \theta \in \mathbb{R}^k.$$

When $E(\rho) > 0$, then according to Proposition 26, we can find a measure $\nu(\theta)$ so that

$$\int_{\mathbb{R}^k} d\nu(\theta) = 0 \quad \text{and} \quad \int_{\mathbb{R}^k} \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t_0) d\nu(\theta) < 0.$$

We thus have a contradiction, since

$$0 > \int_{\mathbb{R}^k} \frac{\delta E(\rho_\infty)}{\delta \rho}(\theta, t_0) d\nu(\theta) = c \int_{\mathbb{R}^k} d\nu(\theta) = 0.$$

We conclude that $E(\rho_\infty) = 0$, proving the theorem. ■

Appendix F. Discussion of the Fully-supported Condition in Theorem 7

In this section we give two examples where the fully-supported condition holds true.

Proposition 27 *Suppose that Assumption 3.1 holds and that $\rho_{\text{ini}}(\theta, t)$ is admissible. Let $\rho(\theta, t, s) \in \mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ be the solution to (20). If the support of $\rho_{\text{ini}}(\theta, t_0)$ is \mathbb{R}^k for some $t_0 \in [0, 1]$, then for any finite s , the support of $\rho(\theta, t_0, s)$ is \mathbb{R}^k .*

This proposition suggests that if the initial probability distribution ρ_{ini} is supported on the whole domain, this property is preserved for all finite s . A second example is as follows.

Proposition 28 *Suppose that Assumption 3.1 holds, that $\rho_{\text{ini}}(\theta, t)$ is admissible, and that the support of $\rho_{\text{ini}}(\theta, t_0)$ is \mathbb{R}^k for some $t_0 \in [0, 1]$. Let $\rho(\theta, t, s) \in \mathcal{C}([0, \infty); \mathcal{C}([0, 1]; \mathcal{P}^2))$ be the solution to (20), and $\theta_\rho(s; t, \theta_0)$ be the solution to corresponding particle presentation:*

$$\begin{cases} \frac{\partial \theta_\rho(s; t, \theta_0)}{\partial s} = -\nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho}(\theta_\rho(s; t, \theta_0), t), & \forall s > 0, t \in [0, 1] \\ \theta_\rho(0; t, \theta_0) = \theta_0. \end{cases} \quad (79)$$

If $\lim_{s \rightarrow \infty} \theta_\rho(s; t, \theta_0) = \theta_\rho(\infty; t, \theta_0)$ uniformly in θ_0, t , then there exists $\rho_\infty(\theta, t)$ so that $\lim_{s \rightarrow \infty} \rho(\theta, t, s) = \rho_\infty(\theta, t)$ in $\mathcal{C}([0, 1]; \mathcal{P}^2)$ and the support of $\rho_\infty(\theta, t_0)$ is the full space \mathbb{R}^k .

This proposition validates the condition required in Theorem 6, thus allowing $E(\rho_\infty) = 0$ in the prescribed situation.

We prove the two propositions here.

Proof [Proof of Proposition 27] According to Section D.1, $\rho(\theta, t, s)$ has a particle representation from the solution of the ODE (50). Let $S > 0$ be fixed. According to Lemma 15, we have for any $s \in [0, S]$ that $\nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho}$ is Lipschitz in θ . Therefore, the following ODE also has a unique solution for $s \in [0, S]$ for any $t \in [0, 1]$ and $\theta_S \in \mathbb{R}^k$:

$$\begin{cases} \frac{\partial \theta(s; t)}{\partial s} = -\nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho}(\theta(s; t), t) \\ \theta(S; t) = \theta_S \end{cases}.$$

This fact suggests that (50) produces a bijection from $\theta(0; t) \in \mathbb{R}^k$ to $\theta(S; t) \in \mathbb{R}^k$. Because the solution to (50) also depends continuously on initial data and the support of $\rho_{\text{ini}}(\theta, t_0)$ is \mathbb{R}^k , we obtain that the support of $\rho(\theta, t_0, s)$ is also \mathbb{R}^k for any $s < \infty$, proving the result. ■

Proof [Proof of Proposition 28] According to Section D.1, $\rho(\theta, t, s)$ has a particle representation from the solution of the ODE (79), where θ_0 is replaced by $\theta_0(t)$ and $\theta_0(t) \sim \rho_{\text{ini}}(\theta, t)$. We define $\rho_\infty(\theta, t)$ as the distribution of $\theta_\rho(\infty; t, \theta_0(t))$. Since

$$\lim_{s \rightarrow \infty} \theta_\rho(s; t, \theta) = \theta_\rho(\infty; t, \theta)$$

uniformly in θ, t , we obtain that

$$\lim_{s \rightarrow \infty} \sup_{t \in [0, 1]} W_2(\rho_\infty(\theta, t), \rho(\theta, t, s)) \leq \lim_{s \rightarrow \infty} \sup_{t \in [0, 1]} \mathbb{E}(|\theta_\rho(s; t, \theta_0(t)) - \theta_\rho(\infty; t, \theta_0(t))|^2) = 0,$$

which proves the convergence condition of Theorem 7.

Next, consider any small ball $B_r(\tilde{\theta}) \subset \mathbb{R}^k$, where $\tilde{\theta} \in \mathbb{R}^k$ and $r > 0$. According to Proposition 27, we obtain that

$$\int_{B_{r/2}(\tilde{\theta})} d\rho(\theta, t_0, s) > 0, \quad \text{for all } s > 0,$$

which implies that

$$\mathbb{P}(\theta_\rho(s; t, \theta_0(t)) \in B_{r/2}(\tilde{\theta})) > 0, \quad \text{for all } s > 0.$$

Because of the uniformly convergence, we can choose S large enough so that

$$\sup_{t \in [0,1], \theta \in \mathbb{R}^k} |\theta_\rho(S; t, \theta) - \theta_\rho(\infty; t, \theta)| \leq \frac{r}{2}.$$

We then obtain

$$\int_{B_r(\tilde{\theta})} d\rho_\infty(\theta, t_0) = \mathbb{P}\left(\theta_\rho(\infty; t, \theta_0(t)) \in B_r(\tilde{\theta})\right) \geq \mathbb{P}\left(\theta_\rho(S; t, \theta_0(t)) \in B_{r/2}(\tilde{\theta})\right) > 0.$$

Since $B_r(\tilde{\theta})$ is a arbitrary ball in \mathbb{R}^k , this proves that the support of $\rho_\infty(\theta, t_0)$ is \mathbb{R}^k . \blacksquare

Appendix G. Proof of Theorem 6

We start by giving the full definition of limit-admissible ρ .

Definition 29 For an admissible $\rho(\theta, t)$, we say $\rho(\theta, t)$ is limit-admissible if the average of a large number of particle presentations is bounded and Lipschitz with high probability. That is, for an admissible $\rho(\theta, t)$, there are two constants C_3 and C_4 , both greater than $\sup_{t \in [0,1]} \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t)$ such that, for any M stochastic process presentation $\{\theta_m(t)\}_{m=1}^M$ that are i.i.d. drawn from $\rho(\theta, t)$, the following properties are satisfied for any $\eta > 0$ and $M > \frac{C_3}{\eta}$:

1. Second moment boundedness in time:

$$\mathbb{P}\left(\sup_{t \in [0,1]} \frac{1}{M} \sum_{m=1}^M |\theta_m(t)|^2 \leq C_4\right) \geq 1 - \eta. \quad (80)$$

2. For all $L > 0$, we have

$$\mathbb{P}\left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left|\theta_m(t) - \theta_m\left(\frac{l}{L}\right)\right|^2 dt \leq \frac{C_4}{L^2}\right) \geq 1 - \eta. \quad (81)$$

Remark 30 We note that it is relatively easy to have $\rho(\theta, t)$ be limit-admissible and satisfy (80) and (81). For example, if $\rho(\theta, t) = \rho(\theta)$ and $\theta(t) = \theta$, then (81) is directly satisfied. Furthermore, if $\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta)$ is finite, then using Chebyshev's inequality, we have

$$\begin{aligned} & \mathbb{P}\left(\left|\sup_{t \in [0,1]} \frac{1}{M} \sum_{m=1}^M |\theta_m(t)|^2 - \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta)\right| \leq \left(\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta)\right)^{1/2}\right) \\ &= \mathbb{P}\left(\left|\frac{1}{M} \sum_{m=1}^M \left(|\theta_m|^2 - \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta)\right)\right| \leq \left(\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta)\right)^{1/2}\right) \\ &\geq 1 - \frac{\mathbb{E}\left(\left|\frac{1}{M} \sum_{m=1}^M (|\theta_m|^2 - \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta))\right|^2\right)}{\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta)} \\ &\geq 1 - \frac{1}{M}. \end{aligned}$$

Since $\int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta) \leq (\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta))^{1/2}$, this implies (80) with

$$C_3 = 1, \quad C_4 = 2 \left(\int_{\mathbb{R}^k} |\theta|^4 d\rho(\theta) \right)^{1/2}.$$

Theorem 6 addresses the convergence in both M and L . We organize the discussion into the following two theorems, which are subsequently proved in the latter two sections.

The first theorem addresses the limit in M , specifically, the setup in which ResNet has infinite layers $L = \infty$ and finite width $M < \infty$, but we are interested in the limiting loss for larger M . We refer this part of analysis as “mean-field analysis.”

Theorem 31 *Suppose that Assumption 3.1 holds, $\rho_{\text{ini}}(\theta, t)$ is limit-admissible, and $\{\theta_m(0; t)\}_{m=1}^M$ are i.i.d drawn from $\rho_{\text{ini}}(\theta, t)$. Suppose too that*

- $\rho(\theta, t, s)$ solves (20) with the initial condition $\rho_{\text{ini}}(\theta, t)$, and
- $\theta_m(s; t)$ solves (19) with the initial condition $\theta_m(0; t)$.

Then for any positive values of ϵ , η , and S , there exists a constant $C(\rho_{\text{ini}}(\theta, t), S) > 0$ depending on $\rho_{\text{ini}}(\theta, t)$ and S such that when

$$M > \frac{C(\rho_{\text{ini}}(\theta, t), S)}{\epsilon^2 \eta},$$

we have

$$\mathbb{P}(|E(\Theta(s; \cdot)) - E(\rho(\cdot, \cdot, s))| \leq \epsilon) \geq 1 - \eta, \quad \forall s < S.$$

Proof See Appendix H. ■

The conclusion of this result suggests that for a $1 - \eta$ confidence of an ϵ accuracy, M grows polynomially with respect to $1/\epsilon$ and $1/\eta$.

The second result considers the convergence of discrete ResNet (1) to continuous ResNet (4) as $L \rightarrow \infty$. This part of analysis is called “continuous limiting analysis.”

Theorem 32 *Suppose that Assumption 3.1 holds, $\rho_{\text{ini}}(\theta, t)$ is limit-admissible, and $\{\theta_m(0; t)\}_{m=1}^M$ are i.i.d drawn from $\rho_{\text{ini}}(\theta, t)$. Suppose too that*

- $\theta_m(s; t)$ solves (19) with initial condition $\theta_m(0; t)$, and
- $\theta_{l,m}(s)$ solves (18) with initial condition $\theta_m(0; \frac{l}{L})$.

Then for any positive ϵ , η , and S , there exists a constant $C(\rho_{\text{ini}}(\theta, t), S) > 0$ depending on $\rho_{\text{ini}}(\theta, t)$ and S such that when

$$M \geq \frac{C(\rho_{\text{ini}}(\theta, t), S)}{\eta}, \quad L \geq \frac{C(\rho_{\text{ini}}(\theta, t), S)}{\epsilon}$$

we have for all $s < S$ that

$$\mathbb{P}(|E(\Theta(s; \cdot)) - E(\Theta_{L,M}(s))| \leq \epsilon) \geq 1 - \eta.$$

Proof See Appendix I. ■

This theorem shows that when the width M is large enough, then with high probability, the difference of loss functions between the discrete ResNet and its continuous limit decreases to 0 as $L \rightarrow \infty$.

Remark 33 *According to the proof in Appendix I, the result of Theorem 32 can be strengthened: In particular, the lower bound of M could be relaxed. If one incorporates η dependence in C_4 (in (80)) and let it be large enough, the lower bound in M in Definition 29 can be removed, relaxing the requirement on M in the previous theorem.*

Appendix H. Convergence to the Mean-field PDE

This section is dedicated to mean-field analysis and the proof of Theorem 31. We first present some intuition.

H.1 An Intuitive Argument for the Equivalence

To intuitively see the equivalence between (19) and (20), we test them on the same smooth function $h(\theta)$. Testing (20) amounts to multiplying h on both sides and performing integration by parts:

$$\frac{d}{ds} \int_{\mathbb{R}^k} h d\rho(\theta) = - \int_{\mathbb{R}^k} \nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho} d\rho.$$

This is to say $\frac{d}{ds} \mathbb{E}(h) = \mathbb{E}(\nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho})$. Testing h on (19), we should expect the same equivalence.

Setting $\rho = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m}$, we have from (19) that

$$\frac{d}{ds} \mathbb{E}(h) = \frac{1}{M} \sum_{m=1}^M \nabla_\theta h(\theta_m) \frac{d}{ds} \theta_m = - \sum_{m=1}^M \nabla_\theta h(\theta_m) \frac{\delta E_s}{\delta \theta_m}.$$

The right hand side is also $\mathbb{E}(\nabla_\theta h \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho})$ if and only if

$$M \frac{\delta E_s(\Theta(s))}{\delta \theta_m} = \nabla_\theta \frac{\delta E_s(\rho(s))}{\delta \rho}(\theta_m, t),$$

a fact that we show rigorously below.

Lemma 34 *For a list of continuous trajectories $\Theta(t) = \{\theta_m(t)\}_{m=1}^M$, denote*

$$\rho_\Theta = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m(t)}(\theta)$$

then for all $1 \leq m \leq M$ and $t \in [0, 1]$, we have

$$\frac{\delta E_s(\Theta)}{\delta \theta_m}(t) = \frac{1}{M} \nabla_\theta \frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\theta_m(t), t). \quad (82)$$

Proof For every fixed m with $1 \leq m \leq M$, we denote by $\tilde{\theta}_m(t)$ the direction of perturbation, and define the new system

$$\tilde{\Theta}_\epsilon(t) = \left\{ \tilde{\theta}_{i,\epsilon}(t) \right\}_{i=1}^M,$$

by setting $\tilde{\theta}_{i,\epsilon}(t) = \theta_i(t)$ for $i \neq m$ and $\tilde{\theta}_{m,\epsilon}(t) = \theta_m(t) + \epsilon \tilde{\theta}_m(t)$ for small $\epsilon > 0$. Using the definition of the Fréchet derivative, we have

$$\begin{aligned} \int_0^1 \frac{\delta E_s(\Theta)}{\delta \theta_m}(t) \tilde{\theta}_m(t) dt &= \lim_{\epsilon \rightarrow 0} \frac{E_s(\tilde{\Theta}_\epsilon) - E_s(\Theta)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{E_s(\rho_{\tilde{\Theta},\epsilon}) - E_s(\rho_\Theta)}{\epsilon} \\ &\stackrel{(I)}{=} \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left[\int_0^1 \int_{\mathbb{R}^k} \frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\theta, t) (d\rho_{\tilde{\Theta},\epsilon} - d\rho_\Theta) dt \right] \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{M\epsilon} \left[\int_0^1 \left(\frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\tilde{\theta}_{m,\epsilon}(t), t) - \frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\theta_m(t), t) \right) dt \right] \\ &= \frac{1}{M} \int_0^1 \nabla_\theta \frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\theta_m(t), t) \tilde{\theta}_m(t) dt. \end{aligned}$$

In (I), we use $d_1(\rho_{\tilde{\Theta},\epsilon}, \rho_{\Theta,\epsilon}) = O(\epsilon)$ and a property of Fréchet derivative ((Lu et al., 2020, Theorem 2)) to obtain that

$$\left| E_s(\rho_{\tilde{\Theta},\epsilon}) - E_s(\rho_\Theta) - \left[\int_0^1 \int_{\mathbb{R}^k} \frac{\delta E_s(\rho_\Theta)}{\delta \rho}(\theta, t) (d\rho_{\tilde{\Theta},\epsilon} - d\rho_\Theta) dt \right] \right| \leq o(d_1(\rho_{\tilde{\Theta},\epsilon}, \rho_{\Theta,\epsilon})) = o(\epsilon).$$

Because $\tilde{\theta}_m(t)$ is an arbitrary perturbation, the result (82) is proved. \blacksquare

Remark 35 *As a direct consequence of this result, we can verify Remark 13 rigorously as well.*

Proof [Proof of Remark 13] From Lemma 34, we see that (19) can be written as

$$\frac{d\theta_m(s; t)}{ds} = -\nabla_\theta \frac{\delta E_s(\rho_\Theta^{\text{dis}})}{\delta \rho}(\theta_m(s; t), t), \quad m = 1, 2, \dots, M. \quad (83)$$

For any smooth test function h and $t \in [0, 1]$, we have

$$\begin{aligned} \frac{d}{ds} \left(\int_{\mathbb{R}^k} h(\theta) d\rho_\Theta^{\text{dis}}(\theta, t, s) \right) &= \frac{d}{ds} \left(\frac{1}{M} \sum_{m=1}^M h(\theta_m(s; t)) \right) \\ &= \frac{1}{M} \sum_{m=1}^M -\nabla h(\theta_m(s; t)) \cdot \nabla_\theta \frac{\delta E_s(\rho_\Theta^{\text{dis}})}{\delta \rho}(\theta_m(s; t), t) \\ &= -\int_{\mathbb{R}^k} \nabla h(\theta) \cdot \nabla_\theta \frac{\delta E_s(\rho_\Theta^{\text{dis}})}{\delta \rho}(\theta, t) d\rho_\Theta^{\text{dis}}(\theta, t, s), \end{aligned}$$

which implies from integration by parts that $\rho_\Theta^{\text{dis}}(\theta, t, s)$ is a weak solution to (20). \blacksquare

H.2 Discussion of Stability in the Mean-field Regime

We first provide a priori estimates to show the stability of Z and p . Let $\rho(\theta, t) \in \mathcal{C}([0, 1]; \mathcal{P}^2)$, and $\{\tilde{\theta}_m(t)\}_{m=1}^M$ and $\{\theta_m(t)\}_{m=1}^M$ be two distinct paths. Defining

$$\rho^{\text{dis}}(\theta, t) = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m(t)}(\theta), \quad \tilde{\rho}^{\text{dis}}(\theta, t) = \frac{1}{M} \sum_{m=1}^M \delta_{\tilde{\theta}_m(t)}(\theta), \quad (84)$$

we have the following lemma.

Lemma 36 *Let ρ^{dis} and $\tilde{\rho}^{\text{dis}}$ be as defined in (84). Suppose that Z_ρ and $Z_{\rho^{\text{dis}}}$ are the solutions of (7) using ρ and ρ^{dis} , respectively, and p_ρ and $p_{\rho^{\text{dis}}}$ solve (11) using ρ and ρ^{dis} , respectively. Denote*

$$\mathcal{L}^{\text{sup}} = \sup_{t \in [0, 1]} \left\{ \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, u) du, \frac{1}{M} \sum_{m=1}^M |\theta_m(t)|^2, \frac{1}{M} \sum_{m=1}^M |\tilde{\theta}_m(t)|^2 \right\}. \quad (85)$$

There exists a constant $C(\mathcal{L}^{\text{sup}})$ that depends only on \mathcal{L}^{sup} such that for all $t \in [0, 1]$, we have

$$\begin{aligned} & |Z_\rho(t; x) - Z_{\rho^{\text{dis}}}(t; x)| \\ & \leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M \int_0^1 |\theta_m(\tau) - \tilde{\theta}_m(\tau)|^2 d\tau \right)^{1/2} \\ & \quad + C(\mathcal{L}^{\text{sup}}) \left(\int_0^1 \left| \int_{\mathbb{R}^k} f(Z_\rho(\tau; x), \theta) d(\rho(\theta, \tau) - \tilde{\rho}^{\text{dis}}(\theta, \tau)) \right|^2 d\tau \right)^{1/2}, \end{aligned} \quad (86)$$

and

$$\begin{aligned} & |p_\rho(t; x) - p_{\rho^{\text{dis}}}(t; x)| \\ & \leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M \int_0^1 |\theta_m(\tau) - \tilde{\theta}_m(\tau)|^2 d\tau \right)^{1/2} \\ & \quad + C(\mathcal{L}^{\text{sup}}) \left(\int_0^1 \left| \int_{\mathbb{R}^k} f(Z_\rho(\tau; x), \theta) d(\rho(\theta, \tau) - \tilde{\rho}^{\text{dis}}(\theta, \tau)) \right|^2 d\tau \right)^{1/2} \\ & \quad + C(\mathcal{L}^{\text{sup}}) \left(\int_0^1 \left| \int_{\mathbb{R}^k} \partial_z f(Z_\rho(\tau; x), \theta) d(\rho(\theta, \tau) - \tilde{\rho}^{\text{dis}}(\theta, \tau)) \right|^2 d\tau \right)^{1/2}. \end{aligned} \quad (87)$$

Proof For ease of notation, we define

$$\begin{aligned} Z(t; x) &= Z_\rho(t; x), \quad Z_{\text{dis}}(t; x) = Z_{\rho^{\text{dis}}}(t; x), \quad \tilde{Z}_{\text{dis}}(t; x) = Z_{\tilde{\rho}^{\text{dis}}}(t; x), \\ p(t; x) &= p_\rho(t; x), \quad p_{\text{dis}}(t; x) = p_{\rho^{\text{dis}}}(t; x), \quad \tilde{p}_{\text{dis}}(t; x) = p_{\tilde{\rho}^{\text{dis}}}(t; x), \end{aligned}$$

and denote

$$\tilde{\Delta}(t; x) = Z(t; x) - \tilde{Z}_{\text{dis}}(t; x), \quad \Delta(t; x) = \tilde{Z}_{\text{dis}}(t; x) - Z_{\text{dis}}(t; x), \quad \Delta_p(t; x) = p(t; x) - p_{\text{dis}}(t; x).$$

To prove (86), we have by similar reasoning to (30) that

$$\frac{d|\tilde{\Delta}(t; x)|^2}{dt} \leq \left(\frac{2C_1}{M} \sum_{m=1}^M |\tilde{\theta}_m(t)| + 1 \right) |\tilde{\Delta}(t; x)|^2 + \left| \int_{\mathbb{R}^k} f(Z(t; x), \theta) d(\rho(\theta, t) - \tilde{\rho}^{\text{dis}}(\theta, t)) \right|^2.$$

Using the Grönwall inequality and the fact that $|\tilde{\Delta}(0; x)| = 0$, we obtain the following, for all $t \in [0, 1]$:

$$|\tilde{\Delta}(t; x)| \leq C(\mathcal{L}^{\text{sup}}) \left(\int_0^1 \left| \int_{\mathbb{R}^k} f(Z(\tau; x), \theta) d(\rho(\theta, \tau) - \tilde{\rho}^{\text{dis}}(\theta, \tau)) \right|^2 d\tau \right)^{1/2}. \quad (88)$$

Similarly, we have

$$\begin{aligned} \frac{d|\Delta(t; x)|^2}{dt} &\leq \left(\frac{2C_1}{M} \sum_{m=1}^M |\theta_m(t)|^2 + 1 \right) |\Delta(t; x)|^2 \\ &\quad + \left| \int_{\mathbb{R}^k} f(\tilde{Z}_{\text{dis}}(t; x), \theta) d(\rho^{\text{dis}}(\theta, t) - \tilde{\rho}^{\text{dis}}(\theta, t)) \right|^2 \\ &\leq C(\mathcal{L}^{\text{sup}}) |\Delta(t; x)|^2 + C_1^2 \left(|\tilde{Z}_{\text{dis}}(t; x)| + 1 \right)^2 \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(t) - \tilde{\theta}_m(t)| \right)^2 \\ &\leq C(\mathcal{L}^{\text{sup}}) \left[|\Delta(t; x)|^2 + \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(t) - \tilde{\theta}_m(t)|^2 \right) \right], \end{aligned}$$

where we use Assumption 3.1 (12) in the second inequality and (22a) in the last inequality. Since $\Delta(0; x) = 0$, we use the Grönwall's inequality to obtain the following, for all $t \in [0, 1]$:

$$|\Delta(t; x)| \leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M \int_0^1 |\theta_m(\tau) - \tilde{\theta}_m(\tau)|^2 d\tau \right)^{1/2}. \quad (89)$$

We obtain (86) by adding (88) and (89).

To prove (87), we recall (11) to obtain

$$\begin{aligned}
& \frac{d|\Delta_p(t; x)|^2}{dt} \\
&= 2 \left\langle \Delta_p(t; x), p_\rho^\top \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - p_{\rho^{\text{dis}}}^\top \int_{\mathbb{R}^k} \partial_z f(Z_{\text{dis}}(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right\rangle \\
&= 2 \left\langle \Delta_p(t; x), (p_\rho^\top - p_{\rho^{\text{dis}}}^\top) \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) \right\rangle \\
&\quad + 2 \left\langle \Delta_p(t; x), p_{\rho^{\text{dis}}}^\top \left(\int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\text{dis}}(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right) \right\rangle \\
&\leq 2 \left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) \right| |\Delta_p(t; x)|^2 \\
&\quad + 2 |\Delta_p(t; x)| |p_{\rho^{\text{dis}}}^\top| \left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\text{dis}}(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right| \\
&\leq C(\mathcal{L}^{\text{sup}}) |\Delta_p(t; x)|^2 \\
&\quad + 2 \left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\text{dis}}(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right|^2 \\
&\leq C(\mathcal{L}^{\text{sup}}) |\Delta_p(t; x)|^2 \\
&\quad + 6 \left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\tilde{\rho}^{\text{dis}}(\theta, t) \right|^2 \\
&\quad + 6 \underbrace{\left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\tilde{\rho}^{\text{dis}}(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right|^2}_{\text{(I)}} \\
&\quad + 6 \underbrace{\left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho^{\text{dis}}(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z_{\text{dis}}(t; x), \theta) d\rho^{\text{dis}}(\theta, t) \right|^2}_{\text{(II)}},
\end{aligned} \tag{90}$$

where we use (12) from Assumption 3.1 along with (22a) and (33a) in the second inequality. We then bound the last two terms. To bound (I), we use (12) from Assumption 3.1 along with (22a), and an analysis similar to (41) to obtain

$$\begin{aligned}
\text{(I)} &\leq \left(\frac{1}{M} \sum_{m=1}^M \left| \partial_z f(Z(t; x), \theta_m(t)) - \partial_z f(Z(t; x), \tilde{\theta}_m(t)) \right| \right)^2 \\
&\leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M (|\theta_m(t)| + |\tilde{\theta}_m(t)|) |\theta_m(t) - \tilde{\theta}_m(t)| \right)^2 \\
&\leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M (|\theta_m(t)| + |\tilde{\theta}_m(t)|)^2 \right) \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(t) - \tilde{\theta}_m(t)|^2 \right) \\
&\leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(t) - \tilde{\theta}_m(t)|^2 \right),
\end{aligned} \tag{91}$$

where we use Hölder's inequality in the second inequality. For (III), we use (12) from Assumption 3.1 along with (85), to obtain

$$\begin{aligned}
(\text{II}) &\leq \left(\frac{1}{M} \sum_{m=1}^M |\partial_z f(Z(t; x), \theta_m(t)) - \partial_z f(Z_{\text{dis}}(t; x), \theta_m(t))| \right)^2 \\
&\leq \left(\frac{C_1}{M} \sum_{m=1}^M |\theta_m(t)|^2 |Z(t; x) - Z_{\text{dis}}(t; x)| \right)^2 \\
&\leq C(\mathcal{L}^{\text{sup}}) |Z(t; x) - Z_{\text{dis}}(t; x)|^2.
\end{aligned} \tag{92}$$

By substituting (91) and (92) into (90), we obtain

$$\begin{aligned}
\frac{d|\Delta_p(t; x)|^2}{dt} &= C(\mathcal{L}^{\text{sup}}) |\Delta_p(t; x)|^2 \\
&\quad + C(\mathcal{L}^{\text{sup}}) \left(\left(\frac{1}{M} \sum_{m=1}^M |\theta_m(t) - \tilde{\theta}_m(t)|^2 \right) + |Z(t; x) - Z_{\text{dis}}(t; x)|^2 \right) \\
&\quad + 6 \left| \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\rho(\theta, t) - \int_{\mathbb{R}^k} \partial_z f(Z(t; x), \theta) d\tilde{\rho}^{\text{dis}}(\theta, t) \right|^2.
\end{aligned} \tag{93}$$

Considering the initial condition, we obtain in a similar fashion to (36) that

$$|\Delta_p(1; x)| \leq C(\mathcal{L}^{\text{sup}}) |Z(1; x) - Z_{\text{dis}}(1; x)|,$$

By substituting (86) into (93) and using Grönwall's inequality, we arrive at (87). \blacksquare

H.3 Proof of Theorem 31

According to the statement of the theorem, $\rho(\theta, t, s)$ solves (20) with admissible initial condition $\rho_{\text{ini}}(\theta, t)$, and $\theta_m(s; t)$ solves (19) with initial conditions $\{\theta_m(t, 0)\}_{m=1}^M$ that are *i.i.d.* drawn from $\rho_{\text{ini}}(\theta, t)$. We now define

$$\rho_{\Theta_s}^{\text{dis}}(\theta, t) = \frac{1}{M} \sum_{m=1}^M \delta_{\theta_m(s; t)}(\theta),$$

with

$$\frac{d\theta_m(s; t)}{ds} = -\nabla_{\theta} \frac{\delta E_s(\rho_{\Theta_s}^{\text{dis}})}{\delta \rho}(\theta_m(s; t)) - 2e^{-s} \theta_m(s; t), \quad s > 0, \quad t \in [0, 1], \tag{94}$$

where $\frac{\delta E_s(\rho_{\Theta_s}^{\text{dis}})}{\delta \rho}$ is defined in (10).

Throughout this section, we denote

$$\mathcal{L}_{\text{ini}}^{\text{dis, sup}} = \sup_{t \in [0, 1]} \frac{1}{M} \sum_{m=1}^M |\theta_m(0; t)|^2, \quad \mathcal{L}_{\text{ini}}^{\text{sup}} = \sup_{t \in [0, 1]} \int_{\mathbb{R}^k} |\theta|^2 d\rho_{\text{ini}}(\theta, t).$$

We note that when M is large, $\mathcal{L}_{\text{ini}}^{\text{dis, sup}}$ is close to $\mathcal{L}_{\text{ini}}^{\text{sup}}$ (which has no randomness) with high probability. We have the following lemma.

Lemma 37 *For fixed $S > 0$, any $s \in [0, S]$, $|x| < R$, and $t \in [0, 1]$, there exists a constant $C(S, \mathcal{L}_{\text{ini}}^{\text{dis, sup}})$ such that*

$$\int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_{\Theta_s}^{\text{dis}}(\theta, t) dt < C(S, \mathcal{L}_{\text{ini}}^{\text{dis, sup}}). \quad (95)$$

Further, for any x with $|x| < R$, the ODE solution is bounded as follows:

$$\left| Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \leq C(S, \mathcal{L}_{\text{ini}}^{\text{dis, sup}}), \quad (96)$$

while the following bound holds on $p_{\rho_{\Theta_s}^{\text{dis}}}$:

$$\left| p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \leq C(S, \mathcal{L}_{\text{ini}}^{\text{dis, sup}}). \quad (97)$$

Proof First, using (16) and (22a), we have that

$$E_0(\rho_{\Theta_0}^{\text{dis}}) < C(\mathcal{L}_{\text{ini}}^{\text{dis, sup}}).$$

According to the definition of $\rho_{\Theta_s}^{\text{dis}}$ and $E_s(\rho)$ in (24) and (17), and recalling decay of energy, we have as shown in (23) that

$$\begin{aligned} \int_0^1 \int_{\mathbb{R}^k} |\theta|^2 d\rho_{\Theta_s}^{\text{dis}}(\theta, t) dt &= \int_0^1 \frac{1}{M} \sum_{m=1}^M |\theta_m(s; t)|^2 dt \leq \exp(s) E_s(\rho_{\Theta_s}^{\text{dis}}) \\ &\leq \exp(S) E_0(\rho_{\Theta_0}^{\text{dis}}) < C(S, \mathcal{L}_{\text{ini}}^{\text{dis, sup}}), \end{aligned}$$

for $s \in [0, S]$, proving (95).

We obtain (96) and (97) from the same argument that shows (22a) and (33a), with the \mathcal{L}_1 term replaced by the bound (95). \blacksquare

We are now ready to prove Theorem 31.

Proof [Proof of Theorem 31] To start, we define a new system $\tilde{\Theta}_s = \{\tilde{\theta}_m(s; t)\}_{m=1}^M$ where each $\tilde{\theta}_m$ solves

$$\frac{\partial \tilde{\theta}_m(s; t)}{\partial s} = -\nabla_{\theta} \frac{\delta E_s(\rho(s))}{\delta \rho}(\tilde{\theta}_m(s; t)), \quad \forall (t, s) \in [0, 1] \times [0, \infty), \quad (98)$$

with initial conditions $\tilde{\theta}_m(t, 0) = \theta_m(t, 0)$. As a consequence, we have

$$\tilde{\theta}_m(s; t) \sim \rho(\theta, t, s), \quad \forall (t, s) \in [0, 1] \times [0, \infty).$$

We further denote

$$\tilde{\rho}^{\text{dis}}(\theta, t, s) = \frac{1}{M} \sum_{m=1}^M \delta_{\tilde{\theta}_m(s; t)}(\theta), \quad \mathcal{L}_0^{\text{sup}} = \max \left\{ \mathcal{L}_{\text{ini}}^{\text{sup}}, \mathcal{L}_{\text{ini}}^{\text{dis, sup}} \right\} < \infty, \quad (99)$$

We first bound the second moment of $\tilde{\rho}^{\text{dis}}(\theta, t, s)$ for all $t \in [0, 1]$ and $s \in [0, S]$ with initial condition $\tilde{\rho}^{\text{dis}}(\theta, t, 0) = \frac{1}{M} \sum_{m=1}^M \delta_{\tilde{\theta}_m(0; t)}(\theta)$. Similar to the proof of Proposition 16, we multiply (98) by $\tilde{\theta}_m(s; t)$ on both sides, and utilize the bound (43) from Lemma 15, where \mathcal{L} in (43) is replaced by $C(\mathcal{L}_{\text{ini}}^{\text{sup}}, S)$ according to (64) in Corollary D.3. We thus obtain

$$|\tilde{\theta}_m(s; t)| \leq C(\mathcal{L}_{\text{ini}}^{\text{sup}}, S) \left(|\tilde{\theta}_m(0; t)| + 1 \right),$$

which implies that

$$\sup_{t \in [0, 1], 0 \leq s \leq S} \frac{1}{M} \sum_{m=1}^M \left| \tilde{\theta}_m(s; t) \right|^2 \leq C(\mathcal{L}_0^{\text{sup}}, S).$$

By combining this bound with (64) and (95), we obtain that there is a constant $C(\mathcal{L}_0^{\text{sup}}, S)$ depending only on $\mathcal{L}_0^{\text{sup}}$ and S such that

$$\sup_{t \in [0, 1], 0 \leq s \leq S} \left\{ \int_{\mathbb{R}^k} |\theta|^2 d\rho(\theta, t, s), \frac{1}{M} \sum_{m=1}^M |\theta_m(s; t)|^2, \frac{1}{M} \sum_{m=1}^M \left| \tilde{\theta}_m(s; t) \right|^2 \right\} \leq C(\mathcal{L}_0^{\text{sup}}, S).$$

To prove the theorem, we have from the definition (8) of E that

$$\begin{aligned} & \left| E(\rho(s)) - E(\rho_{\Theta_s}^{\text{dis}}) \right| \\ &= \mathbb{E}_{x \sim \mu} \left[\frac{1}{2} \left(g(Z_{\rho(s)}(1; x)) - y(x) \right)^2 - \frac{1}{2} \left(g(Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x)) - y(x) \right)^2 \right] \\ &\leq \mathbb{E}_{x \sim \mu} \left[\left| g(Z_{\rho(s)}(1; x)) - g(Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x)) \right| \left(\left| g(Z_{\rho(s)}(1; x)) + g(Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x)) \right| + |y(x)| \right) \right] \quad (100) \\ &\leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| g(Z_{\rho(s)}(1; x)) - g(Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x)) \right| \right) \\ &\leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\rho(s)}(1; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x) \right| \right), \end{aligned}$$

where the second inequality arises from boundedness of Z in (22a) and local boundedness of g and y , by Assumption 3.1.

To estimate $\left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right|$, we recall (86) to obtain

$$\begin{aligned} & \left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \\ &\leq C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M \int_0^1 \left| \theta_m(s; \tau) - \tilde{\theta}_m(s; \tau) \right|^2 d\tau \right)^{1/2} \quad (101) \\ &\quad + C(\mathcal{L}_0^{\text{sup}}, S) \left(\int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau \right)^{1/2}. \end{aligned}$$

Since $\tilde{\theta}_m(s; t) \sim \rho(\theta, t, s)$, the second term of (101) can be bounded by law of large numbers, with high probability. Thus we need only to control the first term. In the analysis below we control this term in Step 1 and utilize the law of large numbers in Step 2.

Step 1: Estimating $\frac{1}{M} \sum_{m=1}^M \int_0^1 \left| \theta_m(s; t) - \tilde{\theta}_m(s; t) \right|^2 dt$. Defining

$$\Delta_{t,m}(s) = \theta_m(s; t) - \tilde{\theta}_m(s; t),$$

we note that $|\Delta_{t,m}(0)| = 0$. By taking the difference of (94) and (98), we obtain

$$\begin{aligned} & \frac{d|\Delta_{t,m}(s)|^2}{ds} \\ &= -2 \exp(-s) |\Delta_{t,m}(s)|^2 \\ & \quad - 2 \left\langle \Delta_{t,m}(s), \mathbb{E}_{x \sim \mu} \left(\partial_{\theta} f(Z_{\rho(s)}(t; x), \tilde{\theta}_m) p_{\rho(s)}(t; x) - \partial_{\theta} f(Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x), \theta_m) p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right) \right\rangle \\ &= -2 \exp(-s) |\Delta_{t,m}(s)|^2 \\ & \quad - 2 \left\langle \Delta_{t,m}(s), \underbrace{\mathbb{E}_{x \sim \mu} \left(\partial_{\theta} f(Z_{\rho(s)}(t; x), \tilde{\theta}_m) p_{\rho(s)}(t; x) - \partial_{\theta} f(Z_{\rho(s)}(t; x), \tilde{\theta}_m) p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right)}_{\text{(I)}} \right\rangle \\ & \quad - 2 \left\langle \Delta_{t,m}(s), \underbrace{\mathbb{E}_{x \sim \mu} \left(\partial_{\theta} f(Z_{\rho(s)}(t; x), \tilde{\theta}_m) p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) - \partial_{\theta} f(Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x), \theta_m) p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right)}_{\text{(II)}} \right\rangle. \end{aligned} \tag{102}$$

For term (I), we have from the bound on $Z_{\rho(s)}$ in (22a) that

$$|\text{(I)}| \leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| p_{\rho(s)}(t; x) - p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \right).$$

For (II), we have

$$\begin{aligned} |\text{(II)}| &\leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| \partial_{\theta} f(Z_{\rho(s)}(t; x), \tilde{\theta}_m) - \partial_{\theta} f(Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x), \theta_m) \right| \right) \\ &\leq C(\mathcal{L}_0^{\text{sup}}, S) \left[\left(\left| \tilde{\theta}_m \right| + \left| \theta_m \right| \right) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \right) + \left| \tilde{\theta}_m - \theta_m \right| \right]. \end{aligned}$$

In both estimates we used the property of f in (12), and the bound on $Z_{\rho, \rho_{\Theta}^{\text{dis}}}$. By substituting these estimates into (102), we obtain

$$\begin{aligned} & \frac{d|\Delta_{t,m}(s)|^2}{ds} \\ &\leq C(\mathcal{L}_0^{\text{sup}}, S) |\Delta_{t,m}(s)|^2 \\ & \quad + C(\mathcal{L}_0^{\text{sup}}, S) |\Delta_{t,m}(s)| \mathbb{E}_{x \sim \mu} \left(\left(\left| \tilde{\theta}_m \right| + \left| \theta_m \right| \right) \left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| + \left| p_{\rho(s)}(t; x) - p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \right), \end{aligned}$$

which implies

$$\begin{aligned}
& \frac{1}{M} \sum_{m=1}^M \frac{d|\Delta_{t,m}(s)|^2}{ds} \\
& \leq C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) \\
& \quad + C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| \left(|\tilde{\theta}_m| + |\theta_m| \right) \right) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \right) \\
& \quad + C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| \right) \mathbb{E}_{x \sim \mu} \left(\left| p_{\rho(s)}(t; x) - p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right| \right) \\
& \leq C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) + C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| p_{\rho(s)}(t; x) - p_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right|^2 \right) \\
& \quad + C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x) \right|^2 \right),
\end{aligned}$$

where we used the Hölder's inequality and

$$\begin{aligned}
\left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| \left(|\tilde{\theta}_m| + |\theta_m| \right) \right)^2 & \leq \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) \left(\frac{1}{M} \sum_{m=1}^M \left(|\tilde{\theta}_m| + |\theta_m| \right)^2 \right) \\
& \leq C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right).
\end{aligned}$$

Noting the estimate in Lemma 36, we obtain that

$$\begin{aligned}
& \frac{d \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right)}{ds} \\
& \leq C(\mathcal{L}_0^{\text{sup}}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) \\
& \quad + C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau \right) \\
& \quad + C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\int_0^1 \left| \int_{\mathbb{R}^k} \partial_z f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau \right),
\end{aligned}$$

which implies, using Grönwall's inequality, that

$$\begin{aligned}
 & \frac{1}{M} \sum_{m=1}^M \int_0^1 |\Delta_{t,m}(s)|^2 dt \\
 & \leq C(\mathcal{L}_0^{\sup}, S) \mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau ds \right) \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \left| \int_{\mathbb{R}^k} \partial_z f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau ds \right). \tag{103}
 \end{aligned}$$

Step 2: Complete the proof. Since ρ_{ini} satisfies the limit-admissible condition, according to (80), there is a constant $C_4 \geq \mathcal{L}_{\text{ini}}^{\sup}$ such that $\mathbb{P}(\mathcal{L}_{\text{ini}}^{\text{dis}, \sup} \leq C_4) \geq 1 - \eta/2$ for M sufficiently large. Using the definition of \mathcal{L}_0^{\sup} in (99), and for some constant $C_3 \geq \mathcal{L}_{\text{ini}}^{\sup}$, this translates to:

$$\mathbb{P}(\mathcal{L}_0^{\sup} \leq C_4) \geq 1 - \eta/2, \quad \text{for } M > \frac{2C_3}{\eta}, \tag{104}$$

so that \mathcal{L}_0^{\sup} is bounded with high probability.

By substituting (103) into (86), we obtain

$$\begin{aligned}
 & |Z_{\rho(s)}(t; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(t; x)| \\
 & \leq C(\mathcal{L}_0^{\sup}, S) \left(\underbrace{\mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau ds \right)}_{\text{(I)}} \right)^{1/2} \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \left(\underbrace{\mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \left| \int_{\mathbb{R}^k} \partial_z f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau ds \right)}_{\text{(II)}} \right)^{1/2} \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \left(\underbrace{\int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau}_{\text{(III)}} \right)^{1/2}. \tag{105}
 \end{aligned}$$

All terms (I), (II), and (III) can be controlled. We have

$$\begin{aligned}
\mathbb{E}(\text{I}) &= \mathbb{E}_{x \sim \mu} \left(\mathbb{E} \left(\int_0^S \int_0^1 \left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 d\tau ds \right) \right) \\
&= \mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \mathbb{E} \left(\left| \int_{\mathbb{R}^k} f(Z_{\rho(s)}(\tau; x), \theta) d(\rho(\theta, \tau, s) - \tilde{\rho}^{\text{dis}}(\theta, \tau, s)) \right|^2 \right) d\tau ds \right) \\
&\leq \frac{C(\mathcal{L}_{\text{ini}}^{\text{sup}}, S)}{M} \mathbb{E}_{x \sim \mu} \left(\int_0^S \int_0^1 \int_{\mathbb{R}^k} |f(Z_{\rho(s)}(\tau; x), \theta)|^2 d\rho(\theta, \tau, s) d\tau ds \right) \\
&\leq \frac{C(\mathcal{L}_{\text{ini}}^{\text{sup}}, S)}{M} \int_0^S \int_0^1 \int_{\mathbb{R}^k} (|\theta|^2 + 1) d\rho(\theta, \tau, s) d\tau ds \leq \frac{C(C_4, S)}{M},
\end{aligned}$$

where we use $\tilde{\theta}_m(s; t) \sim \rho(\theta, t, s)$ in the first inequality, (27) with $|Z_{\rho(s)}| \leq C(\mathcal{L}_{\text{ini}}^{\text{sup}}, S)$ in the second inequality, and $\mathcal{L}_{\text{ini}}^{\text{sup}} \leq C_4$ in the third inequality. By similar reasoning, we obtain

$$\mathbb{E}(\text{II}) \leq \frac{C(C_4, S)}{M}, \quad \mathbb{E}(\text{III}) \leq \frac{C(C_4, S)}{M}.$$

From Markov's inequality, these bounds imply that when $M > \frac{C(C_4, S)}{\epsilon^2 \eta}$, we have

$$\mathbb{P}(\{(I) < \epsilon^2\} \cap \{(II) < \epsilon^2\} \cap \{(III) < \epsilon^2\}) > 1 - \eta/2. \quad (106)$$

By substituting (104) and (106) into (105), we see that there exists a constant $C(C_4, S)$ such that for any $\epsilon, \eta > 0$, when $M > \frac{C(C_4, S)}{\epsilon^2 \eta}$ we obtain that

$$\mathbb{P}\left(\left|Z_{\rho(s)}(1; x) - Z_{\rho_{\Theta_s}^{\text{dis}}}(1; x)\right| < \epsilon\right) > 1 - \eta.$$

The proof is completed by substituting this bound into (100). ■

Appendix I. Convergence to the Continuous Limit

This section is dedicated to the continuous limit and proof of Theorem 32.

I.1 Stability with Discretization

Before proving Theorem 32, and similarly to Appendix H.2, we first consider the stability Z and p under discretization. Defining the path of parameters $\Theta(t) = \{\theta_m(t)\}_{m=1}^M$ and the set of parameters $\Theta_{L, M} = \{\theta_{l, m}\}_{l=0, m=1}^{L-1, M}$, we have the following lemma.

Lemma 38 *Suppose that Assumption 3.1 holds and that x is in the support of μ . Denoting*

$$\mathcal{L}^{\text{sup}} = \sup_{t \in [0, 1], l} \left\{ \frac{1}{M} \sum_{m=1}^M |\theta_m(t)|^2, \frac{1}{M} \sum_{m=1}^M |\theta_{l, m}|^2 \right\}, \quad (107)$$

there exists a constant $C(\mathcal{L}^{\text{sup}})$ depending only on \mathcal{L}^{sup} such that for any $0 \leq l \leq L-1$, we have

$$\begin{aligned} & \sup_{\frac{l}{L} \leq t \leq \frac{l+1}{L}} \{ |Z_{\Theta}(t; x) - Z_{\Theta_{L,M}}(l; x)|, |Z_{\Theta}(t; x) - Z_{\Theta_{L,M}}(l+1; x)| \} \\ & \leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 d\tau \right)^{1/2} + \frac{C(\mathcal{L}^{\text{sup}})}{L}, \end{aligned} \quad (108)$$

and

$$\begin{aligned} & \sup_{\frac{l}{L} \leq t \leq \frac{l+1}{L}} |p_{\Theta}(t; x) - p_{\Theta_{L,M}}(l; x)| \\ & \leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 d\tau \right)^{1/2} + \frac{C(\mathcal{L}^{\text{sup}})}{L}. \end{aligned} \quad (109)$$

Proof Define

$$Z(t; x) = Z_{\Theta}(t; x), \quad p(t; x) = p_{\Theta}(t; x)$$

and

$$\tilde{Z}(t; x) = \sum_{l=0}^{L-1} Z_{\Theta_{L,M}}(l; x) \mathbf{1}_{\frac{l}{L} \leq t < \frac{l+1}{L}}, \quad \tilde{p}(t; x) = \sum_{l=0}^{L-1} p_{\Theta_{L,M}}(l; x) \mathbf{1}_{\frac{l}{L} < t \leq \frac{l+1}{L}},$$

with

$$\tilde{Z}(1; x) = Z_{\Theta_{L,M}}(L; x), \quad \tilde{p}(0; x) = p_{\Theta_{L,M}}(0; x).$$

Using (1), (66), and Lemma 24, we obtain

$$\begin{aligned} |Z_{\Theta_{L,M}}(l+1; x) - Z_{\Theta_{L,M}}(l; x)| & < \frac{C(\mathcal{L}^{\text{sup}})}{L}, \\ |p_{\Theta_{L,M}}(l+1; x) - p_{\Theta_{L,M}}(l; x)| & < \frac{C(\mathcal{L}^{\text{sup}})}{L}, \end{aligned} \quad (110)$$

for $0 \leq l \leq L-1$. Now define Δ_t by

$$\Delta_t = Z(t; x) - \tilde{Z}(t; x).$$

For $t \in [\frac{l}{L}, \frac{l+1}{L}]$, we have from (4) that

$$\begin{aligned} |\Delta_t| & \leq \left| \Delta_{\frac{l}{L}} \right| + \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^t |f(Z(\tau; x), \theta_m(\tau))| d\tau \\ & \leq \left| \Delta_{\frac{l}{L}} \right| + \frac{C(\mathcal{L}^{\text{sup}})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} (|\theta_m(\tau)| + 1) d\tau \\ & \leq \left| \Delta_{\frac{l}{L}} \right| + \frac{C(\mathcal{L}^{\text{sup}})}{L}, \end{aligned} \quad (111)$$

where we use (27) and (96) in the second inequality and (107) in the third inequality. From (1) and (4), we obtain further that

$$\begin{aligned}
\left| \Delta_{\frac{l+1}{L}} \right| &= \left| \Delta_{\frac{l}{L}} \right| + \left| \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} f(Z(\tau; x), \theta_m(\tau)) - f(\tilde{Z}(\tau; x), \theta_{l,m}) \, d\tau \right| \\
&\leq \left| \Delta_{\frac{l}{L}} \right| + \left| \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} f(Z(\tau; x), \theta_m(\tau)) - f(Z(\tau; x), \theta_{l,m}) \, d\tau \right| \\
&\quad + \left| \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} f(Z(\tau; x), \theta_{l,m}) - f(\tilde{Z}(\tau; x), \theta_{l,m}) \, d\tau \right| \\
&\stackrel{(I)}{\leq} \left| \Delta_{\frac{l}{L}} \right| + C(\mathcal{L}^{\sup}) \left(\frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_m(\tau) - \theta_{l,m}| \, d\tau \right) + C_1 |\Delta_\xi| \left(\frac{1}{ML} \sum_{m=1}^M |\theta_{l,m}| \right) \\
&\stackrel{(II)}{\leq} \left(1 + \frac{C(\mathcal{L}^{\sup})}{L} \right) \left| \Delta_{\frac{l}{L}} \right| + \frac{C(\mathcal{L}^{\sup})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_m(\tau) - \theta_{l,m}| \, d\tau + \frac{C(\mathcal{L}^{\sup})}{L^2}
\end{aligned}$$

where $\xi \in [\frac{l}{L}, \frac{l+1}{L}]$, and using (12), (96) in (I), and (107) and (111) in (II). By applying this bound iteratively, we obtain

$$\left| \Delta_{\frac{l}{L}} \right| \leq C(\mathcal{L}^{\sup}) |\Delta_0| + \frac{C(\mathcal{L}^{\sup})}{M} \sum_{j=0}^{l-1} \sum_{m=1}^M \int_{\frac{j}{L}}^{\frac{j+1}{L}} |\theta_{l,m} - \theta_m(\tau)| \, d\tau + \frac{C(\mathcal{L}^{\sup})}{L},$$

where $|\Delta_0| = 0$. Combining this with (111) and using Hölder's inequality, we obtain that

$$|\Delta_l| \leq C(\mathcal{L}^{\sup}) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 \, d\tau \right)^{1/2} + \frac{C(\mathcal{L}^{\sup})}{L}.$$

By combining this bound with (110), we prove (108).

To prove (109), we define

$$\Delta_p(t; x) = p(t; x) - \tilde{p}(t; x).$$

Similarly to (36), we obtain

$$\begin{aligned}
|\Delta_p(1; x)| &\leq C(\mathcal{L}^{\sup}) \left| \tilde{Z}(1; x) - Z(1; x) \right| \\
&\leq C(\mathcal{L}^{\sup}) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 \, d\tau \right)^{1/2} + \frac{C(\mathcal{L}^{\sup})}{L}.
\end{aligned} \tag{112}$$

For $t \in (\frac{l}{L}, \frac{l+1}{L}]$, using (11), we obtain that

$$\begin{aligned}
|\Delta_p(t; x)| &\leq \left| \Delta_p\left(\frac{l+1}{L}; x\right) \right| + \frac{1}{M} \sum_{m=1}^M \int_t^{\frac{l+1}{L}} |\partial_z f(Z(\tau; x), \theta_m(\tau))| |p(\tau; x)| \, d\tau \\
&\leq \left| \Delta_p\left(\frac{l+1}{L}; x\right) \right| + \frac{C(\mathcal{L}^{\sup})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_m(\tau)| \, d\tau \\
&\leq \left| \Delta_p\left(\frac{l+1}{L}; x\right) \right| + \frac{C(\mathcal{L}^{\sup})}{L},
\end{aligned} \tag{113}$$

where we use (12) and (97) in the second inequality, and Hölder's inequality together with the definition of \mathcal{L} from (107) in the third inequality.

From (11), we obtain that

$$p^\top \left(\frac{l}{L}; x \right) = p^\top \left(\frac{l+1}{L}; x \right) + \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} p^\top(\tau; x) \partial_z f(Z(\tau; x), \theta_m(\tau)) d\tau,$$

and from (66) that

$$\tilde{p}^\top \left(\frac{l}{L}; x \right) = \tilde{p}^\top \left(\frac{l+1}{L}; x \right) + \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z_{\Theta_{L,M}}(l; x), \theta_{l,m}) d\tau.$$

By bounding differences of these two expressions, we have

$$\begin{aligned} & \left| \Delta_p \left(\frac{l}{L}; x \right) \right| \\ & \leq \underbrace{\left| \Delta_p \left(\frac{l+1}{L}; x \right) \right| + \frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left| p^\top(\tau; x) \partial_z f(Z(\tau; x), \theta_m(\tau)) d\tau - \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z(\tau; x), \theta_m(\tau)) \right| d\tau}_{(I)} \\ & \quad + \underbrace{\frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left| \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z(\tau; x), \theta_m(\tau)) - \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z_{\Theta_{L,M}}(l; x), \theta_m(\tau)) \right| d\tau}_{(II)} \\ & \quad + \underbrace{\frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left| \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z_{\Theta_{L,M}}(l; x), \theta_m(\tau)) - \tilde{p}^\top \left(\frac{l+1}{L}; x \right) \partial_z f(Z_{\Theta_{L,M}}(l; x), \theta_{l,m}) \right| d\tau}_{(III)}. \end{aligned} \tag{114}$$

We bound (I), (II), and (III) as follows.

(I): Using (12) and (107), we obtain that

$$(I) \leq |\Delta_p(t; x)| \frac{C(\mathcal{L}^{\sup})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_m(\tau)| d\tau \leq \frac{C(\mathcal{L}^{\sup})}{L} |\Delta_p(t; x)|.$$

(II): Using (12), (68a), and (107), we obtain

$$\begin{aligned} (II) & \leq \frac{C(\mathcal{L}^{\sup})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |Z(\tau; x) - Z_{\Theta_{L,M}}(l; x)| |\theta_m(\tau)|^2 d\tau \\ & \leq \frac{C(\mathcal{L}^{\sup})}{L} \left(\left(\frac{1}{M} \sum_{l'=0}^{L-1} \sum_{m=1}^M \int_{\frac{l'}{L}}^{\frac{l'+1}{L}} |\theta_{l',m} - \theta_m(\tau)|^2 d\tau \right)^{1/2} + \frac{1}{L} \right), \end{aligned}$$

where we make use of (108) in the final inequality.

(III): Using (12), (67a), (68a), and (107), we obtain that

$$\begin{aligned}
\text{(III)} &\leq \frac{C(\mathcal{L}^{\text{sup}})}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} (|\theta_m(\tau)| + |\theta_{l,m}|) |\theta_m(\tau) - \theta_{l,m}| d\tau \\
&\leq C(\mathcal{L}^{\text{sup}}) \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left(\frac{1}{M} \sum_{m=1}^M (|\theta_m(\tau)| + |\theta_{l,m}|)^2 \right)^{1/2} \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(\tau) - \theta_{l,m}|^2 \right)^{1/2} d\tau \\
&\leq C(\mathcal{L}^{\text{sup}}) \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(\tau) - \theta_{l,m}|^2 \right)^{1/2} d\tau.
\end{aligned}$$

By substituting these three inequalities and (113) into (114), we obtain

$$\begin{aligned}
\left| \Delta_p \left(\frac{l}{L}; x \right) \right| &\leq \left(1 + \frac{C(\mathcal{L}^{\text{sup}})}{L} \right) \left| \Delta_p \left(\frac{l+1}{L}; x \right) \right| \\
&\quad + \frac{C(\mathcal{L}^{\text{sup}})}{L} \left(\left(\frac{1}{M} \sum_{l'=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l',m} - \theta_m(\tau)|^2 d\tau \right)^{1/2} + \frac{1}{L} \right) \\
&\quad + C(\mathcal{L}^{\text{sup}}) \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(\tau) - \theta_{l,m}|^2 \right)^{1/2} d\tau.
\end{aligned}$$

By applying this bound iteratively, and using (112) and (113), we obtain

$$|\Delta_p(t; x)| \leq C(\mathcal{L}^{\text{sup}}) \left(\left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 d\tau \right)^{1/2} + \frac{1}{L} \right), \quad (115)$$

where we also use Hölder's inequality to write

$$\begin{aligned}
&C(\mathcal{L}^{\text{sup}}) \sum_{l=0}^{L-1} \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left(\frac{1}{M} \sum_{m=1}^M |\theta_m(\tau) - \theta_{l,m}|^2 \right)^{1/2} d\tau \\
&\leq C(\mathcal{L}^{\text{sup}}) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\theta_{l,m} - \theta_m(\tau)|^2 d\tau \right)^{1/2}.
\end{aligned}$$

We obtain (109) by combining (115) with (113). ■

1.2 Proof of Theorem 32

We denote by $\theta_m(s; t)$ the solution to (19) with initial $\{\theta_m(0; t)\}_{m=1}^M$ *i.i.d.* drawn from $\rho_{\text{ini}}(\theta, t)$. Further, $\theta_{l,m}(s)$ is a solution to (18) with initial $\theta_{l,m}(0) = \theta_m(0; \frac{l}{L})$ for $0 \leq l \leq L-1$ and $1 \leq i \leq M$. Denoting

$$\mathcal{L}_{\text{ini}}^{\text{dis, sup}} = \sup_{t \in [0,1]} \frac{1}{M} \sum_{m=1}^M |\theta_m(0; t)|^2, \quad \Theta_{L,M}(s) = \{\theta_{l,m}(s)\}_{l=0, i=0}^{L-1, M},$$

we have the following lemma.

Lemma 39 *For fixed $S > 0$, any $s \in [0, S]$, x in the support of μ , and integer l with $0 \leq l \leq L - 1$, there exists a constant $C(S, \mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}})$ depending only on S and $\mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}}$ such that*

$$\sup_{s \in [0, S], l} \frac{1}{M} \sum_{m=1}^M |\theta_{l,m}(s)|^2 < C(S, \mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}}). \quad (116)$$

Further, the ODE solution and $p_{\Theta_{L,M}(s)}$ are bounded as follows:

$$|Z_{\Theta_{L,M}(s)}(l+1; x)| \leq C(S, \mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}}), \quad (117)$$

$$|p_{\Theta_{L,M}(s)}(l; x)| \leq C(S, \mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}}). \quad (118)$$

Proof Since the proof is quite similar to Lemma 37, we omit it. ■

We are now ready to prove Theorem 32.

Proof [Proof of Theorem 32] Define

$$\mathcal{L}_0^{\text{sup}} = \max \left\{ \sup_{t \in [0, 1]} \int_{\mathbb{R}^k} |\theta|^2 d\rho_{\text{ini}}(\theta, t), \mathcal{L}_{\text{ini}}^{\text{dis}, \text{sup}} \right\} < \infty.$$

From (95) and (116), we obtain

$$\sup_{(t,s) \in [0, 1] \times [0, S], l} \left\{ \frac{1}{M} \sum_{m=1}^M |\theta_m(s; t)|^2, \frac{1}{M} \sum_{m=1}^M |\theta_{l,m}(s)|^2 \right\} \leq C(\mathcal{L}_0^{\text{sup}}, S), \quad (119)$$

where $C(\mathcal{L}_0^{\text{sup}}, S)$ is a constant depends on $\mathcal{L}_0^{\text{sup}}, S$.

Using a similar derivation to (100), we have

$$|E(\Theta(s; \cdot)) - E(\Theta_{L,M}(s))| \leq C(\mathcal{L}_0^{\text{sup}}, S) |Z_{\Theta(s)}(1; x) - Z_{\Theta_{L,M}(s)}(L; x)|. \quad (120)$$

Thus, to prove the theorem, it suffices to prove that $|Z_{\Theta_{L,M}(s)}(L; x) - Z_{\Theta(s)}(1; x)|$ is small. According to (108), this requires us to bound the quantity

$$\frac{1}{M} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(s)|^2 dt, \quad \text{where } \Delta_{t,m}(s) = \theta_{l,m}(s) - \theta_m(s; t). \quad (121)$$

The next part of the proof obtains the required bound.

First, using (18) and (19), we obtain that

$$\begin{aligned}
& \frac{d|\Delta_{t,m}(s)|^2}{ds} \\
&= -2 \exp(-s) |\Delta_{t,m}(s)|^2 \\
&\quad - 2 \left\langle \Delta_{t,m}(s), \mathbb{E}_{x \sim \mu} \left(\partial_\theta f(Z_{\Theta_{L,M}(s)}(l; x), \theta_{l,m}) p_{\Theta_{L,M}(s)}(l; x) - \partial_\theta f(Z_{\Theta(s)}(t; x), \theta_m) p_{\Theta(s)}(t; x) \right) \right\rangle \\
&= -2 \exp(-s) |\Delta_{t,m}(s)|^2 \\
&\quad - 2 \left\langle \Delta_{t,m}(s), \underbrace{\mathbb{E}_{x \sim \mu} \left(\partial_\theta f(Z_{\Theta_{L,M}(s)}(l; x), \theta_{l,m}) p_{\Theta_{L,M}(s)}(l; x) - \partial_\theta f(Z_{\Theta(s)}(t; x), \theta_m) p_{\Theta_{L,M}(s)}(l; x) \right)}_{(I)} \right\rangle \\
&\quad - 2 \left\langle \Delta_{t,m}(s), \underbrace{\mathbb{E}_{x \sim \mu} \left(\partial_\theta f(Z_{\Theta(s)}(t; x), \theta_m) p_{\Theta_{L,M}(s)}(l; x) - \partial_\theta f(Z_{\Theta(s)}(t; x), \theta_m) p_{\Theta(s)}(t; x) \right)}_{(II)} \right\rangle
\end{aligned} \tag{122}$$

To bound (I), we use (118) to obtain

$$\begin{aligned}
|(\text{I})| &\leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| \partial_\theta f(Z_{\Theta_{L,M}(s)}(l; x), \theta_{l,m}) - \partial_\theta f(Z_{\Theta(s)}(t; x), \theta_m) \right| \right) \\
&\leq C(\mathcal{L}_0^{\text{sup}}, S) \left[(|\theta_{l,m}| + |\theta_m|) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\Theta_{L,M}(s)}(l; x) - Z_{\Theta(s)}(t; x) \right| \right) + |\theta_{l,m} - \theta_m| \right],
\end{aligned} \tag{123}$$

where we use (12), (96), and (117) in the second inequality. To bound (II), we use (12) and (96) to obtain

$$|(\text{II})| \leq C(\mathcal{L}_0^{\text{sup}}, S) \mathbb{E}_{x \sim \mu} \left(\left| p_{\Theta_{L,M}(s)}(l; x) - p_{\Theta(s)}(t; x) \right| \right). \tag{124}$$

By substituting (124) and (123) into (122), we obtain

$$\begin{aligned}
& \frac{d|\Delta_{t,m}(s)|^2}{ds} \\
&\leq C(\mathcal{L}_0^{\text{sup}}, S) |\Delta_{t,m}(s)|^2 \\
&\quad + C(\mathcal{L}_0^{\text{sup}}, S) |\Delta_{t,m}(s)| \mathbb{E}_{x \sim \mu} \left((|\theta_{l,m}| + |\theta_m|) \left| Z_{\Theta_{L,M}(s)}(l; x) - Z_{\Theta(s)}(t; x) \right| + \left| p_{\Theta_{L,M}(s)}(l; x) - p_{\Theta(s)}(t; x) \right| \right),
\end{aligned}$$

which implies that

$$\begin{aligned}
 & \frac{d \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right)}{ds} \\
 & \leq C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| (|\theta_{l,m}| + |\theta_m|) \right) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\Theta_{L,M}(s)}(l; x) - Z_{\Theta(s)}(t; x) \right| \right) \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| \right) \mathbb{E}_{x \sim \mu} \left(\left| p_{\Theta_{L,M}(s)}(l; x) - p_{\Theta(s)}(t; x) \right| \right) \\
 & \leq C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) + C(\mathcal{L}_0^{\sup}, S) \mathbb{E}_{x \sim \mu} \left(\left| p_{\Theta_{L,M}(s)}(t; x) - p_{\Theta(s)}(t; x) \right|^2 \right) \\
 & \quad + C(\mathcal{L}_0^{\sup}, S) \mathbb{E}_{x \sim \mu} \left(\left| Z_{\Theta_{L,M}(s)}(t; x) - Z_{\Theta(s)}(t; x) \right|^2 \right). \tag{125}
 \end{aligned}$$

The second inequality above uses Hölder's inequality together with the bound $2ab \leq a^2 + b^2$ and

$$\begin{aligned}
 \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)| (|\theta_{l,m}| + |\theta_m|) \right)^2 & \leq \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right) \left(\frac{1}{M} \sum_{m=1}^M (|\theta_{l,m}| + |\theta_m|)^2 \right) \\
 & \leq C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{m=1}^M |\Delta_{t,m}(s)|^2 \right).
 \end{aligned}$$

By substituting (108) and (109) into (125), we obtain

$$\frac{d \frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(s)|^2 dt}{ds} \leq C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(s)|^2 dt + \frac{1}{L^2} \right),$$

which implies that

$$\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(s)|^2 dt \leq C(\mathcal{L}_0^{\sup}, S) \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(0)|^2 dt + \frac{1}{L^2} \right) \tag{126}$$

by Grönwall's inequality. This is the bound we were seeking on (121). Here, we also have

$$\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(0)|^2 dt = \frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} \left| \theta_m \left(0; \frac{l}{L} \right) - \theta_m(0; t) \right|^2 dt. \tag{127}$$

To complete the proof, we first use (80) and (81), and take $M \geq \frac{2C_3}{\eta}$, to obtain

$$\mathbb{P} \left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(0)|^2 dt \leq \frac{C_4}{L^2} \right) \geq 1 - \eta/2. \tag{128}$$

Similarly to (104), we also have

$$\mathbb{P}(\mathcal{L}_0^{\sup} \leq C_4) \geq 1 - \eta/2. \quad (129)$$

Using (128), (129) to substitute $\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(0)|^2$, \mathcal{L}_0^{\sup} in (126), we find that there exists a constant $C'(C_4, S)$ depending on C_4 and S such that if

$$M \geq \frac{2C_3}{\eta}, \quad L \geq \frac{C'(C_4, S)}{\epsilon},$$

then we have

$$\mathbb{P}\left(\frac{1}{M} \sum_{l=0}^{L-1} \sum_{m=1}^M \int_{\frac{l}{L}}^{\frac{l+1}{L}} |\Delta_{t,m}(s)|^2 dt \leq \epsilon\right) \geq 1 - \eta. \quad (130)$$

Using (130), (119) and (129) to bound the right hand side of (108), we find that there exists another constant $C''(C_4, S)$ depending on C_4 and S such that if

$$M \geq \frac{2C_3}{\eta}, \quad L \geq \frac{C''(C_4, S)}{\epsilon},$$

then we have

$$\mathbb{P}\left(\left|Z_{\Theta(s)}(1; x) - Z_{\Theta_{L,M}(s)}(L; x)\right| \leq \epsilon\right) \geq 1 - \eta,$$

By using this result in conjunction with (120), we complete the proof. \blacksquare

Appendix J. Rigorous Proof of Lemma 20

Proof [Proof of Lemma 20] According to the brief proof in Appendix D.3, it suffices to prove

$$\frac{dE_s(\phi^*(s))}{ds} = - \int_0^1 \mathbb{E} \left(\left| \nabla_{\theta} \frac{\delta E_s(\phi^*(s))}{\delta \rho} (\theta^*(s; t), t) \right|^2 \right) dt - \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt. \quad (131)$$

To show (131), we essentially need to show:

$$|E_s(\phi^*(s)) - E_{s_0}(\phi^*(s_0)) + D(s_0)(s - s_0)| = o(|s - s_0|), \quad (132)$$

where

$$D(s_0) = \int_0^1 \mathbb{E} \left(\left| \nabla_{\theta} \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) \right|^2 \right) dt + \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt \quad (133)$$

Noticing that according to (43),

$$\mathbb{E} \left(\left| \nabla_{\theta} \frac{\delta E_s(\phi^*(s))}{\delta \rho} (\theta^*(s; t), t) \right|^2 \right) \leq 8\mathbb{E}|\theta^*(s; t)|^2 + 2C(\mathcal{L}_{S, \phi^*}) < \infty,$$

the term (133) is finite and well-defined. To prove (132), recall (17), we first write

$$\begin{aligned}
 & E_s(\phi^*(s)) - E_{s_0}(\phi^*(s_0)) \\
 &= \underbrace{\mathbb{E}_{x \sim \mu} \left(\frac{1}{2} (\langle w, Z_{\phi^*(s)}(1; x) \rangle - y(x))^2 - \frac{1}{2} (\langle w, Z_{\phi^*(s_0)}(1; x) \rangle - y(x))^2 \right)}_{(I)} \\
 & \quad + \underbrace{\exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt}_{(II)} .
 \end{aligned}$$

Then we consider (I) and (II) separately:

(I): First, (I) can be written as

$$\begin{aligned}
 (I) &= \mathbb{E}_{x \sim \mu} \left((\langle w, Z_{\phi^*(s_0)}(1; x) \rangle - y(x)) \langle w, Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x) \rangle \right) \\
 & \quad + o(|Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x)|) \\
 &= \mathbb{E}_{x \sim \mu} \left((\langle w, Z_{\phi^*(s_0)}(1; x) \rangle - y(x)) \langle w, Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x) \rangle \right) \\
 & \quad + o(s - s_0)
 \end{aligned}$$

where we use $|Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x)| \leq C(\mathcal{L}_{S, \phi^*})|s - s_0|$ according to (22b) and (53).

Similar to (Lu et al., 2020), using (12), (22b), and (53), we have

$$\begin{aligned}
& \frac{d(Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x))}{dt} \\
&= \mathbb{E} (f(Z_{\phi^*(s)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&= \mathbb{E} (f(Z_{\phi^*(s)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t))) \\
&\quad + \mathbb{E} (f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&= \{ \mathbb{E} (f(Z_{\phi^*(s)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t))) \\
&\quad - \mathbb{E} (f(Z_{\phi^*(s)}(t; x), \theta^*(s_0; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \} \\
&\quad + \mathbb{E} (f(Z_{\phi^*(s)}(t; x), \theta^*(s_0; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&\quad + \mathbb{E} (f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&\stackrel{(a)}{=} \{ [\mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t)))] (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \\
&\quad - [\mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)))] (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \} \\
&\quad + [\mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)))] (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \\
&\quad + \mathbb{E} (f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&\quad + o(|s - s_0|) \\
&\stackrel{(b)}{=} [\mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)))] (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \\
&\quad + \mathbb{E} (f(Z_{\phi^*(s_0)}(t; x), \theta^*(s; t)) - f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \\
&\quad + o(|s - s_0|) \\
&\stackrel{(c)}{=} \mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \\
&\quad + \mathbb{E} (\partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) (\theta^*(s; t) - \theta^*(s_0; t))) \\
&\quad + o(|s - s_0|) \\
&\stackrel{(d)}{=} \mathbb{E} (\partial_z f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) (Z_{\phi^*(s)}(t; x) - Z_{\phi^*(s_0)}(t; x)) \\
&\quad - \mathbb{E} \left(\partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) (s - s_0) \right) \\
&\quad + o(|s - s_0|),
\end{aligned} \tag{134}$$

where we use $|\partial_x^2 f| \leq C_1 |\theta|^2$ by (12) and $\mathbb{E}(|\theta^*(s; t)|^2) < C(\mathcal{L}_{S, \phi^*})$ by (57) in (a). In (b), we use $|\partial_z \partial_\theta f| \leq C_1 |\theta|(|x| + 1)$ by (12), $|Z_{\phi^*(s_0)}| < C(\mathcal{L}_{S, \phi^*})$ by (22a), and $\mathbb{E}(|\theta^*(s; t) - \theta^*(s_0; t)|) < C(\mathcal{L}_{S, \phi^*})(s - s_0)$ by (53) to obtain that first two terms are $o(s - s_0)$. In (c), we also use $|\partial_\theta f| \leq C_1(|x| + 1)$ by (12), $|Z_{\phi^*(s_0)}| < C(\mathcal{L}_{S, \phi^*})$ by (22a), and $\mathbb{E}(|\theta^*(s; t) - \theta^*(s_0; t)|) < C(\mathcal{L}_{S, \phi^*})(s - s_0)$ by (53). In (d), we first use (44), (59), and (45) to obtain

$$\begin{aligned}
& \left| \nabla_\theta \frac{\delta E_s(\phi^*(s))}{\delta \rho} (\theta^*(s; t), t) - \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) \right| \\
& \leq C (|\theta^*(s; t) - \theta^*(s_0; t)| + d_2(\phi^*(s), \phi^*(s_0))(|\theta^*(s_0; t)| + 1))
\end{aligned}$$

which implies

$$\mathbb{E} \left| \theta^*(s; t) - \theta^*(s_0; t) - \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) (s - s_0) \right| = o(|s - s_0|) \tag{135}$$

because $\mathbb{E}(|\theta^*(s; t)|^2) < C(\mathcal{L}_{S, \phi^*})$ by (57). Combining (135) with $|\partial_\theta f| \leq C_1(|x| + 1)$ by (12) and $|Z_{\phi^*(s_0)}| < C(\mathcal{L}_{S, \phi^*})$, we obtain (d). Since $Z_{\phi^*(s)}(0; x) = Z_{\phi^*(s_0)}(0; x)$, using the last equality, we obtain that

$$\begin{aligned} & Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x) \\ = & - \int_0^1 e^{\int_t^1 \mathbb{E}(\partial_z f(Z_{\phi^*(s_0)}(\tau; x), \theta^*(s_0; \tau))) d\tau} \\ & \cdot \mathbb{E} \left(\partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho}(\theta^*(s_0; t), t) \right) dt (s - s_0) \end{aligned} \quad (136)$$

Using (11), we also have

$$\begin{aligned} p_{\phi^*}(t; x) &= e^{\int_t^1 \mathbb{E}(\partial_z f(Z_{\phi^*(s_0)}(\tau; x), \theta^*(s_0; \tau))) d\tau} p_{\phi^*}(1; x) \\ &= e^{\int_t^1 \mathbb{E}(\partial_z f(Z_{\phi^*(s_0)}(\tau; x), \theta^*(s_0; \tau))) d\tau} (\langle w, Z_{\phi^*(s_0)}(1; x) \rangle - y(x)) w \end{aligned} \quad (137)$$

Combine (136) and (137), we obtain that

$$\begin{aligned} & \mathbb{E}_{x \sim \mu} ((\langle w, Z_{\phi^*(s_0)}(1; x) \rangle - y(x)) \langle w, Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x) \rangle) \\ = & \mathbb{E}_{x \sim \mu} (\langle p_{\phi^*}(1; x), Z_{\phi^*(s)}(1; x) - Z_{\phi^*(s_0)}(1; x) \rangle) \\ = & - \mathbb{E}_{x \sim \mu} \left(\int_0^1 p_{\phi^*}(t; x) \mathbb{E} \left(\partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho}(\theta^*(s_0; t), t) \right) dt \right) (s - s_0) \\ & + o(|s - s_0|), \end{aligned}$$

which implies

$$\begin{aligned} \text{(I)} &= - \int_0^1 \mathbb{E} \left(\mathbb{E}_{x \sim \mu} (p_{\phi^*}(t; x) \partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t))) \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho}(\theta^*(s_0; t), t) \right) dt (s - s_0) \\ &+ o(|s - s_0|) \end{aligned} \quad (138)$$

(II): Similar to before, using (53) and (135), we obtain that

$$\begin{aligned}
(\text{II}) &= \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt - \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt \\
&\quad + \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt \\
&= \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt - \exp^{-s} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt \\
&\quad - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt (s - s_0) + o(|s - s_0|) \\
&\stackrel{(a)}{=} \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s; t)|^2) dt - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt \\
&\quad - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt (s - s_0) + o(|s - s_0|) \\
&\stackrel{(b)}{=} -2 \exp^{-s_0} \int_0^1 \mathbb{E} \left(\theta^*(s_0; t) \nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) (s - s_0) \right) dt \\
&\quad - \exp^{-s_0} \int_0^1 \mathbb{E}(|\theta^*(s_0; t)|^2) dt (s - s_0) + o(|s - s_0|)
\end{aligned} \tag{139}$$

In (a), we use similar calculations as the third to the fifth equality in (134). In (b), we use similar calculations as the last equality in (134).

Noticing that

$$\nabla_\theta \frac{\delta E_s(\phi^*(s_0))}{\delta \rho} (\theta^*(s_0; t), t) = \mathbb{E}_{x \sim \mu} (\partial_\theta f(Z_{\phi^*(s_0)}(t; x), \theta^*(s_0; t)) p_{\phi^*}(t; x)) + 2 \exp^{-s_0} \theta^*(s_0; t),$$

we finally have

$$(\text{I}) + (\text{II}) = -D(s_0) + o(|s - s_0|)$$

by (133), (138), and (139). This proves (132) and the lemma. ■

References

- Z. Allen-Zhu and Y. Li. What can ResNet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems*, 2019.
- Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 242–252, 2019.
- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science and Business Media. Birkhäuser Basel, 2008.

- D. Araújo, R. Oliveira, and D. Yukimura. A mean-field limit for certain deep neural networks. *arXiv/1906.00193*, 2019.
- S. Arora, S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.
- J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International Conference on Learning Representations*, 2020.
- P. Bartlett, S. Evans, and P. Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv/1804.05012*, 2018a.
- P. Bartlett, D. Helmbold, and P. Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 521–530, 2018b.
- N. Chatterji, P. Long, and P. Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *Journal of Machine Learning Research*, 22(159):1–48, 2021.
- Z. Chen, Y. Cao, Q. Gu, and T. Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13363–13373, 2020.
- L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.
- S. Du and J. Lee. On the power of over-parametrization in neural networks with quadratic activation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1329–1338, 2018.
- S. Du, C. Jin, J. Lee, M. Jordan, A. Singh, and B. Póczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, 2017.
- S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 2019a.
- S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.
- W. E, C. Ma, and L. Wu. Machine learning from a continuous viewpoint, I. *Science China Mathematics*, 63(11):2233–2266, Sep 2020a.

- W. E, C. Ma, L. Wu, and S. Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *CSIAM Transactions on Applied Mathematics*, 1(4):561–615, 2020b.
- C. Fang, Y. Gu, W. Zhang, and T. Zhang. Convex formulation of overparameterized deep neural networks. *arXiv/1911/07626*, 2019.
- C. Fang, J. Lee, P. Yang, and T. Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on Learning Theory*, 2021.
- S. Frei, Y. Cao, and Q. Gu. Algorithm-dependent generalization bounds for overparameterized deep residual networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.
- R. Ge, J. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.
- J.-F. Jabir, D. Šiška, and Ł. Szpruch. Mean-field neural ODEs via relaxed optimal control. *arxiv/1912.05475*, 2021.
- A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.
- C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732, 2017.
- K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, 2016.
- Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.
- C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems*, pages 15954–15964, 2020.
- Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying. A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6426–6436, 2020.
- S. Mei, A. Montanari, and P. 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.
- P.-M. Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv/1902.02880*, 2019.

- P.-M. Nguyen and H. Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arxiv/2001.11443*, 2021.
- Q. Nguyen and M. Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2603–2612, 2017.
- Q. Nguyen and M. Hein. Optimization landscape and expressivity of deep CNNs. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- H. Pham and P.-M. Nguyen. Global convergence of three-layer neural networks in the mean field regime. In *International Conference on Learning Representations*, 2021.
- I. Safran, G. Yehudai, and O. Shamir. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Proceedings of The 34th Conference on Learning Theory*, 2020.
- J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.
- J. Sirignano and K. Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 2021.
- M. Soltanolkotabi, A. Javanmard, and J. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inf. Theor.*, 65(2): 742–769, February 2019.
- C. Wei, J. Lee, Q. Liu, and T. Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems*, 2019.
- S. Wojtowytsch. On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime. *arXiv/2005/13530*, 2020.
- C. Yun, S. Sra, and A. Jadbabaie. Global optimality conditions for deep neural networks. In *International Conference on Learning Representations*, 2018.
- H. Zhang, D. Yu, M. Yi, W. Chen, and T. Liu. Convergence theory of learning over-parameterized ResNet: A full characterization. *arXiv/1903.07120*, 2019.
- D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine learning*, 109:467–492, 2020.