Journal of Machine Learning Research 27 (2026) 1-62 Submitted 4/24; Revised 10/25; Published 4/26

A Mean-Field Analysis of Neural Stochastic Gradient Descent-Ascent for Functional Minimax Optimization

Yuchen Zhu YZHU738@GATECH.EDU
Department of Mathematics
Georgia Institute of Technology
Atlanta, GA 300332, USA

Yufeng Zhang YUFENGZHANG2023@U.NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Zhuoran Yang ZHUORAN.YANG@YALE.EDU
Department of Statistics and Data Science
Yale University
New Haven, CT 06520, USA

Xiaohong Chen XIAOHONG.CHEN@YALE.EDU
Department of Economics
Yale University
New Haven, CT 06511, USA

Editor: Ambuj Tewari

Abstract

This paper studies minimax optimization problems defined over infinite-dimensional function classes of over-parameterized two-layer neural networks. In particular, we consider the minimax optimization problem stemming from estimating linear functional equations defined by conditional expectations, where the objective functions are quadratic in the functional spaces. We address (i) the convergence of the stochastic gradient descent-ascent algorithm and (ii) the representation learning of the neural networks. We establish convergence in the mean-field regime by considering the continuous-time, infinite-width limit of the optimization dynamics. Under this regime, stochastic gradient descent-ascent corresponds to a Wasserstein gradient flow over the space of probability measures defined over the space of neural network parameters. We prove that the Wasserstein gradient flow converges globally to a stationary point of the minimax objective at a $\mathcal{O}(T^{-1} + \alpha^{-1})$ sublinear rate, and additionally finds the solution to the functional equation when the regularizer of the minimax objective is strongly convex. Here T denotes the time and α is a scaling parameter of the neural networks. In terms of representation learning, our results show that the feature representation induced by the neural networks may deviate from the initial representation by a factor of $\mathcal{O}(\alpha^{-1})$, measured by the Wasserstein distance. Finally, we

apply our general results to concrete examples, including policy evaluation, nonparametric instrumental variable regression, and asset pricing.

Keywords: Minimax optimization, neural networks, functional gradient descent-ascent, policy evaluation, asset pricing, instrumental variable regression.

1. Introduction

Minimax optimization problems are ubiquitous in machine learning, statistics, economics, and other fields. Examples include generative adversarial networks (GANs) (Goodfellow et al., 2020; Salimans et al., 2016), adversarial training (Ganin et al., 2016; Madry et al., 2017), robust optimization (Ben-Tal et al., 2009; Levy et al., 2020), and zero-sum games (Xie et al., 2020b; Zhao et al., 2022). The goal in minimax optimization is to find a solution (f^*, g^*) to the problem $\min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \mathcal{L}(f,g)$, where $\mathcal{L}$ is a bivariate objective function, and $\mathcal{F}$ and $\mathcal{G}$ are the feasible sets of the decision variables f and g. In modern machine learning applications, $\mathcal{F}$ and $\mathcal{G}$ are often function classes flexibly parameterized by neural networks, and the objective $\mathcal{L}(f,g)$ can be approximated using data. The minimax optimization problem is often solved using first-order optimization algorithms. Despite their widespread success across diverse applications, there is currently no global convergence theory for popular first-order algorithms that solve general minimax optimization problems using neural networks.

In this work, we study the convergence of first-order algorithms for solving minimax optimization problems where $\mathcal{F}$ and $\mathcal{G}$ are both flexibly parameterized by two-layer neural networks, and the objective functional is quadratic in f and g up to regularization:

$$\min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \mathcal{L}(f,g), \quad \mathcal{L}(f,g) = \mathbb{E}\big[g(Z)\cdot\Phi(X,Z;f) - 1/2\cdot g(Z)^2 + \texttt{Reg}(f)\big], \tag{1}$$

where $\texttt{Reg}(f)$ is a convex regularizer that penalizes the complexity of $f\in\mathcal{F}$. A popular choice is to choose a $\texttt{Reg}(f)$ that measures the norm $\|f\|_2^2$. Here, the expectation is taken with respect to the joint distribution of random variables (X,Z), g is a function of Z, and Φ takes (X,Z) and a function f as its input and is linear in f. The objective function (1) arises from solving a linear functional conditional moment equation of the form $\mathbb{E}[\Phi(X,Z;f)|Z=\cdot]=0$ if and only if $f=f^*\in\mathcal{F}$. Here, X is a vector containing all the endogenous variables, and Z contains all the exogenous/pre-determined variables. This problem has ample applications, including policy evaluation (Cai et al., 2019; Duan et al., 2020; Jin et al., 2021; Chen and Qi, 2022; Ramprasad et al., 2022), nonparametric instrumental variable regression (NPIV) (Ai and Chen, 2003; Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Darolles et al., 2011; Chen and Reiss, 2011; Chen and Pouzo, 2012), and asset pricing (Chen and Ludvigson, 2009; Chen et al., 2014, 2024). For example, when $\Phi(X,Z;f)=Y-f(X)$, the problem in (1) recovers the setting of NPIV. The minimax objective in (1) arises when we solve the conditional moment equation via adversarial estimation (Uehara et al., 2020; Duan et al., 2021; Chernozhukov et al., 2020; Liao et al., 2020; Wai et al., 2020; Bennett et al., 2019), which introduces a dual function and transforms equation solving into a minimax optimization.

We study the infinite-dimensional minimax optimization in (1) over the space of over-parameterized two-layer neural networks. Specifically, a neural network is represented by

$f_{\mathtt{NN}}(\cdot;\boldsymbol{\theta}) = \alpha/N \sum_{i=1}^{N} \phi(\cdot;\theta^i)$, where N is the number of neurons, $\phi(\cdot;\theta^i)$ denotes the i-th neuron, $\{\theta^i\}_{i\in[N]}$ are the network parameters, and α is a scaling parameter. We aim to solve the minimax optimization in (1) with both f and g represented by over-parameterized two-layer neural networks, which is favorable especially when Z is a high-dimensional vector. To solve this problem, we consider the arguably simplest first-order algorithm, stochastic gradient descent-ascent (SGDA), where the parameters of f and g are simultaneously updated using stochastic gradients of the objective functional. Specifically, we aim to address the following two questions:

- Does SGDA with over-parameterized neural networks converge to some solution?
- Does SGDA learn data-dependent features that yield a statistically accurate solution?

Answering these questions involves two intricate challenges in optimization and representation learning with neural networks. First, the minimax objective is nonconvex-nonconcave with respect to the neural network parameters of f and g, it is unclear whether first-order algorithms converge. Second, the neural network's representation evolves during optimization, and it is unclear how to track and assess the data-dependent features it learns. While there are some existing works on neural network optimization using the technique of neural tangent kernel (NTK) (Jacot et al., 2018; Du et al., 2018; Cai et al., 2019; Xu and Gu, 2020; Wang et al., 2022), such an approach suggests that the feature representation of the neural networks is fixed throughout training and is only determined by the initialization of the network parameters. Despite being an elegant theoretical framework, the NTK approach is limited in its ability to capture the representation learning aspect of neural network optimization. To show that the neural network optimization algorithms learn useful data-dependent features, in addition to establishing convergence, more importantly, we need to show that (i) the algorithm approximately finds a proper solution concept, e.g., a stationary point or a local or global optimizer of the minimax objective function, and (ii) the representation of the neural networks moves from the initialization by a considerable amount.

In this paper, we tackle both challenges by leveraging the framework of mean-field analysis of over-parameterized neural networks (Chizat and Bach, 2018; Mei et al., 2018, 2019; Zhang et al., 2020; Lu et al., 2020b; Zhang et al., 2021b; Sirignano and Spiliopoulos, 2020a,b, 2022; Chen et al., 2020b; Fang et al., 2021b). In particular, we focus on the continuous-time and infinite-width limit of the SGDA algorithm, where the step size approaches zero and the width N approaches infinity. From the mean-field lens, a neural network $f(\cdot;\boldsymbol{\theta})$ can be identified with a probability measure μ by writing $f(\cdot;\boldsymbol{\theta}) = \alpha \cdot \int_\theta \phi(\cdot;\theta)\ \mu(\mathrm{d}\theta)$, where μ is the empirical distribution of $\{\theta^i\}_{i\in[N]}$ and α is the scaling parameter of the neural network. Thus, parameter updates of SGDA can be regarded as updates to the probability measure μ. From this perspective, we prove that in the continuous-time and infinite-width limit, SGDA corresponds to a gradient flow of the minimax objective $\mathcal{L}$ in the Wasserstein space, i.e., the space of probability measures over the parameter space equipped with the Wasserstein-2 distance. Besides, by defining a proper potential function that characterizes the stationary point of the minimax objective, we prove that the Wasserstein gradient flow converges to a stationary point at a sublinear rate of $\mathcal{O}(1/T + 1/\alpha)$, where T is the time horizon and α is a scaling parameter of the neural network. Moreover, we prove that the Wasserstein distance between the parameter distribution found by SGDA and its

initialization is $\mathcal{O}(\alpha^{-1})$, indicating that the neural network's representation can deviate considerably from its initialization. Such behavior is not captured by the NTK analysis, which shows that the representation remains fixed at initialization. Furthermore, when the regularization on f satisfies a version of strong convexity, we prove that the Wasserstein gradient flow converges to the global optimizer f^* at a sublinear $\mathcal{O}(1/T + 1/\alpha)$ rate.

Furthermore, our setting presents unique challenges for a few reasons. Firstly, unlike many existing analyses of SGDA that focus on finite-dimensional Euclidean spaces (Beznosikov et al., 2023; Jin et al., 2019; Lin et al., 2020a), our work studies an infinite-dimensional functional minimax optimization problem, for which tools do not transfer directly. Moreover, while the targeted functional optimization problem is amenable to analysis (as it is convex-concave with respect to the input functions), we do not directly approach it from a functional gradient descent-ascent perspective. Instead, we adopt a realistic approach by representing both the primal and dual functions using neural networks. In terms of network parameters, while the problem is reduced to a finite-dimensional space, the convex-concave structure is destroyed, rendering existing SGDA analysis techniques hardly applicable. To circumvent the technical difficulties, we lift the problem onto the Wasserstein space of probability measures by taking a continuous-time, mean-field limit. Again, unlike the existing literature on mean-field analysis and minimax optimization on the Wasserstein space (Nitanda et al., 2022; Yamamoto et al., 2024; Kim et al., 2023; Cai et al., 2024) that considered only convex or convex-concave problems with respect to the geometry of Wasserstein space, we studied the problem that originates from (1), which again is nonconvex-nonconcave despite its convex-concave nature on the function space. This distinction sets our paper apart from the aforementioned line of work and highlights the challenges in our analysis once again.

To overcome these challenges, in the convergence analysis, we leverage the hidden structures of the nonconvex-nonconcave Wasserstein minimax optimization problem and build connections to convex-concave functional optimization. In particular, in the proof of our main results, Theorem 7, we utilize tools and techniques from optimal transport to establish a bound for the time decay rate of the Wasserstein distance between the trajectory and the optimal solution, as measured by another distance that quantifies the discrepancy between them. This relation can be understood as a significantly weak version of Gronwall's inequality, which paves the way to our convergence result. Building on top of such inequality, we also derive a precise characterization of the trajectories of the Wasserstein gradient flow, showing that its distance to the target solution is bound to decay at a constant rate before deviating too far from it, thereby establishing a sublinear convergence rate for the objective.

To the best of our knowledge, our work provides the first theoretical analysis of an optimization algorithm solving functional conditional moment equations using neural networks with representation learning. We apply our general theory to three important examples: policy evaluation, instrumental variables regression, and asset pricing. In these examples, we prove that the SGDA algorithm finds the global solution with over-parameterized neural networks. Moreover, SGDA learns data-dependent features that enable these statistically accurate estimators.

1.1 Related Works

Minimax Optimization. Our work is closely related to the literature on first-order methods for solving minimax optimization problems. These works establish the convergence rate or iteration complexity of first-order methods under various assumptions on the objective function. In particular, most of the existing works focus on finite-dimensional parameter spaces and one of the following objective functions: (i) convex-concave (Lin et al., 2020b; Ibrahim et al., 2019; Ouyang and Xu, 2021; Alkousa et al., 2019; Luo et al., 2021; Xie et al., 2020a; Han et al., 2024; Li et al., 2023; Jin et al., 2022; Beznosikov et al., 2023), (ii) nonconvex-concave (Jin et al., 2019; Lin et al., 2020a; Lu et al., 2020a; Ostrovskii et al., 2021b; Zhao, 2023; Huang et al., 2022; Luo et al., 2020; Zhang et al., 2021a; Nouiehed et al., 2019; Thekumparampil et al., 2019), and (iii) nonconvex-nonconcave (Li et al., 2022; Diakonikolas et al., 2021; Ostrovskii et al., 2021a; Yang et al., 2022; Grimmer et al., 2022; Hajizadeh et al., 2024; Grimmer et al., 2023; Yang et al., 2020).

Our work can be viewed as an extension of convex-concave minimax optimization to the infinite-dimensional functional space. In particular, our objective is a regularized quadratic functional over the input functions, which is then restricted to the class of over-parameterized neural networks. Note that the objective of interest is in fact nonconvex-nonconcave in the neural network parameter space. Compared with work on general nonconvex-nonconcave minimax optimization problems, our setting has a more favorable underlying functional-space structure with respect to convexity. This structure enables us to lift the neural network parameter updates to the Wasserstein space and analyze the gradient flow in the space of distributions. Our approach leverages the hidden convexity and concavity of the seemingly nonconvex and nonconcave objective function, thereby improving algorithm convergence and complexity.

Mean-field Analysis in Deep Learning. Our work is closely related to recent studies on neural network training using gradient-based methods. One line of research establishes the convergence of gradient-based algorithms for training over-parameterized neural networks under the "lazy training" regime, where the neural networks behave similarly to random kernel functions. Such a regime is also known as the neural tangent kernel regime (Jacot et al., 2018; Allen-Zhu et al., 2019a,b; Chen et al., 2020a; Frei and Gu, 2021; Zou and Gu, 2019; Du et al., 2018, 2019; Arora et al., 2019a,b; Huang and Yau, 2020). Our work is, however, closer to another line of research based on the perspective of mean-field approximation (Mei et al., 2018, 2019; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020a,b, 2022; Chen et al., 2020b; Fang et al., 2021b; Chen et al., 2019). Under the mean-field view, the neural network parameters are identified as a distribution over the parameter space. As a result, the evolution of parameters via gradient-based updates is captured by a differential equation governing the evolution of the corresponding distribution. By elevating the training dynamics to the Wasserstein space, the optimization objective often enjoys a benign landscape, which admits a more tractable analysis and global convergence. See, e.g, Zhang et al. (2020, 2021b); Fang et al. (2021b); Lu et al. (2020b); Fang et al. (2019); Chizat (2022); Hu et al. (2021); Nitanda et al. (2022) and the references therein. Also, see Fang et al. (2021a) for a recent survey.

Our work is particularly related to the mean-field analysis of the Neural Temporal Difference (TD) (Zhang et al., 2020) and the Neural Actor-Critic (AC) (Zhang et al., 2021b) in reinforcement learning. These previous works have analyzed the global convergence of the TD and AC algorithms for two-layer over-parameterized neural networks. The optimization problem for these two tasks is to minimize an objective function involving only a single neural network. Unlike these works, we focus on minimax optimization, which requires neural-network parameterizations of both the primal and dual functions. This presents new challenges to the analysis, as the gradient dynamics of the primal and dual neural networks give rise to a coupled system of PDEs. To the best of our knowledge, our paper is the first to apply the mean-field limit in studying the convergence of algorithms for solving the general form of functional conditional moment equations using over-parameterized neural networks.

Adversarial Estimation. Our work is also related to the literature on adversarial estimation, a method that solves a functional conditional moment equation by introducing a dual function and reformulating the original problem into a minimax optimization. Our work studies this type of minimax optimization with over-parameterized neural networks. Thus, our work is more related to the study of adversarial estimation within neural network function classes (Dikkala et al., 2020; Chernozhukov et al., 2020; Bennett et al., 2019; Xu et al., 2021). Compared with our work, these studies focus on statistical errors pertinent to neural networks, assuming the optimization problem is solved perfectly. Instead, we study the optimization algorithm and establish the convergence of stochastic gradient-descent-ascent for over-parameterized neural networks.

Several previous works have also explored the convergence of optimization dynamics in adversarial estimation with neural networks. In particular, Neural GTD (Wai et al., 2020) and Neural SEM (Liao et al., 2020) analyze respectively the convergence for off-policy evaluation and structural equation models estimation with an over-parameterized two-layered neural network. However, their analyses are based on the idea of neural tangent kernel (NTK), where the employed neural network has a fixed representation during training, and the initialization completely determines the representation. In contrast, our work adopts the mean-field approach, which enables learning a data-dependent representation.

2. Preliminaries

The functional conditional moment equations cover many important examples in statistics, machine learning, economics, and causal inference. In this section, we first introduce the general formulation of the functional conditional moment equations and then reformulate them into a minimax optimization problem. Then, we present a few concrete examples of function conditional moment equations, such as policy evaluation, nonparametric instrumental variables regression, and asset pricing. Finally, we introduce the background of mean-field neural networks and Wasserstein space, which are essential for the convergence analysis of the SGDA algorithm.

2.1 Functional Conditional Moment Equations

In this section, we introduce the general formulation of functional conditional moment equations. Let $X \in \mathcal{X}$ be a vector that includes all the endogenous variables, let $Z \in \mathcal{Z}$

denote all the exogenous variables, and let $\mathcal{D} \in \mathscr{P}(\mathcal{X} \times \mathcal{Z})$ denote the joint distribution of (X, Z). We let $\mathbb{E}_{\mathcal{D}}[\cdot]$ denote the expectation taken with respect to the joint distribution of (X, Z) and $\mathbb{E}_{X|Z}[\cdot]$ denote the conditional expectation using the conditional distribution of X given Z. Let $W \in \mathcal{W} \subseteq \mathcal{X} \times \mathcal{Z}$ be a subset of variables that may contain both the endogenous and exogenous variables, and let $L^2(\mathcal{W})$ denote a Hilbert space of measurable functions of W with finite second moment. Let $\mathcal{F} := \{f : \mathcal{W} \to \mathbb{R}\} \subset L^2(\mathcal{W})$ denote a class of functions defined on $\mathcal{W}$. In a *functional conditional moment equation* problem, we aim to find a function $f_0 \in \mathcal{F}$ that solves the following functional equation involving the conditional distribution of X given Z over $\mathcal{F}$:

$$\mathbb{E}_{X|Z}\big[\Phi(X, Z; f_0) \,\big|\, Z = z\big] = 0, \qquad \forall z \in \mathcal{Z}, \tag{2}$$

where $\Phi \colon \mathcal{X} \times \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ is a known functional.

For any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, we define a functional $\bar{\delta} \colon \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ as

$$\bar{\delta}(z; f) := \mathbb{E}_{X|Z}\big[\Phi(X, Z; f) \,\big|\, Z = z\big], \qquad \forall f \in \mathcal{F}, z \in \mathcal{Z}. \tag{3}$$

In other words, the conditional moment equation problem in (2) boils down to finding a function $f_0 \in \mathcal{F}$ such that $\bar{\delta}(\cdot; f_0)$ is a zero function on $\mathcal{Z}$. Therefore an equivalent way to solve $f_0 \in \mathcal{F}$ in (2) is by solving $\inf_{f \in \mathcal{F}} \mathbb{E}[(\bar{\delta}(Z; f))^2]$ (Ai and Chen, 2003; Chen and Pouzo, 2012). To control the complexity of the function class $\mathcal{F}$, Ai and Chen (2003) propose to use flexible sieve spaces $\mathcal{F}_{k(n)}$ that becomes dense in $\mathcal{F}$ as the sieve dimension $k(n)$ grows to infinity with data sample size n, and proposed the so-called sieve minimum distance criterion $\min_{f \in \mathcal{F}_{k(n)}} \mathbb{E}\big[\bar{\delta}(Z; f)^2\big]/2$. In particular, Ai and Chen (2003) allows for two-layer NNs, splines, wavelets, Fourier series, and all kinds of polynomial sieves $\mathcal{F}_{k(n)}$ to approximate functions in $\mathcal{F} \subseteq L^2(\mathcal{W})$. Alternatively, Chen and Pouzo (2012) proposes the following penalized (or regularized) minimum distance criterion:

$$\min_{f \in \mathcal{F}} J(f), \quad J(f) := \mathbb{E}\big[\bar{\delta}(Z; f)^2\big]/2 + \lambda \cdot \mathcal{R}(f), \tag{4}$$

where $\lambda \geq 0$ is a regularization parameter, $\mathcal{R}(f)$ is a regularizer on function $f \in \mathcal{F}$. They allow that $\mathcal{R}(f)$ to be any convex or lower-semicompact regularizer. In the minimum distance approach, for any fixed f, the authors first estimate $\bar{\delta}(z; f)$ by the following least squares criterion:

$$\operatorname*{argmin}_{\delta \in L^2(\mathcal{Z})} \mathbb{E}\Big[1/2 \cdot \big(\Phi(X, Z; f) - \delta(Z)\big)^2\Big] = \operatorname*{argmax}_{\delta \in L^2(\mathcal{Z})} \mathbb{E}\Big[\Phi(X, Z; f)\delta(Z) - 1/2 \cdot \delta(Z)^2\Big]$$

Furthermore, we assume that the functional Φ is affine in f, which captures several important applications in machine learning and causal inference, as listed in Section 2.2. Specifically, we define $\tilde{\Phi}(x, z, f) = \Phi(x, z, f) - \Phi(x, z, 0)$, where 0 stands for the zero function on $\mathcal{W}$. Then for any two functions $f_1, f_2 \in \mathcal{F}$ and any $a, b \in \mathbb{R}$, we have

$$\tilde{\Phi}(x, z; af_1 + bf_2) = a \cdot \tilde{\Phi}(x, z; f_1) + b \cdot \tilde{\Phi}(x, z; f_2), \qquad \forall (x, z) \in \mathcal{X} \times \mathcal{Z}. \tag{5}$$

Solving (2) with Overparameterized Neural Networks. In the sequel, we aim to solve the problem in (2) based on i.i.d. data points sampled from $\mathcal{D}$, with $\mathcal{F}$ being a class

of overparameterized neural networks. In this case, it is possible that (2) does not have a solution within $\mathcal{F}$. Furthermore, for the choice of regularizer, we consider the following specific form of $\mathcal{R}(f)$:

$$\mathcal{R}(f) = \mathbb{E}_{\mathcal{D}}[\Psi(X, Z; f)] \tag{6}$$

where for any given $(x, z) \in \mathcal{X} \times \mathcal{Z}$, $\Psi(x, z; f) : \mathcal{F} \to \mathbb{R}_+$ is a convex functional of f that maps each function f to a scalar. Moreover, Ψ satisfies

$$\Psi(x, z; 0) = 0, \qquad \Psi(x, z; f) \geq 0, \qquad \forall f \in \mathcal{F}, \tag{7}$$

$$\frac{\delta\Psi(x, z; af_1 + bf_2)}{\delta f} = a \cdot \frac{\delta\Psi(x, z; f_1)}{\delta f} + b \cdot \frac{\delta\Psi(x, z; f_2)}{\delta f}, \qquad \forall f_1, f_2 \in \mathcal{F},\ a, b \in \mathbb{R}. \tag{8}$$

Equation (7) requires that $\Psi(X, Z; f)$ is a non-negative functional of f that is equal to 0 if and only $f = 0$. Equation (8) requires that the functional derivative of $\Psi(X, Z; f)$ with respect to f, is linear in f. One example of Ψ is the L_2-regularizer of the following type, $\Psi(x, z; f) = f(w)^2$. Here $w \in \mathcal{W}$ is a subset of variables that contain values from both the endogenous variables x and exogenous variables z.

Minimax Estimation. To solve the optimization problem in (4), we first transform it into an *unconditional* moment formulation by introducing a dual function. Moreover, we assume that the primal problem in (4) has a unique solution. Note that such an assumption is not restrictive at all, as it is often a natural consequence when the used regularization $\mathcal{R}(f)$ is strongly convex in f. By Fenchel duality, we can rewrite the objective function $J(f)$ as

$$\begin{aligned} J(f) &= \mathbb{E}_{\mathcal{D}}\Big[1/2 \cdot \bar{\delta}(z; f)^2 + \lambda\Psi(X, Z; f)\Big] \\ &= \mathbb{E}_{\mathcal{D}}\Big[\max_{g:\mathcal{Z}\to\mathbb{R}} \big(g(z) \cdot \mathbb{E}\big[\Phi(X, z; f) \,\big|\, z\big] - 1/2 \cdot g(z)^2\big) + \lambda\Psi(X, Z; f)\Big] \\ &= \max_{g:\mathcal{Z}\to\mathbb{R}} \mathbb{E}_{\mathcal{D}}\Big[g(Z) \cdot \Phi(X, Z; f) - 1/2 \cdot g(Z)^2 + \lambda\Psi(X, Z; f)\Big]. \end{aligned} \tag{9}$$

Then the problem $\min_f J(f)$ in (4) becomes the following minimax optimization problem:

$$\min_f \max_g \mathcal{L}(f, g), \quad \mathcal{L}(f, g) := \mathbb{E}_{\mathcal{D}}\Big[g(Z) \cdot \Phi(X, Z; f) - 1/2 \cdot g(Z)^2 + \lambda\Psi(X, Z; f)\Big]. \tag{10}$$

We note that $\mathcal{L}$ is a convex-concave functional with respect to functions f and g. We denote by (f^*, g^*) the unique saddle point of (10). Here, the uniqueness of f^* comes from the assumed uniqueness of the solution to the primal problem, and $g^*(z) = \mathbb{E}[\Phi(X, Z; f^*)|Z = z]$ implies the uniqueness of g^*. Without the regularization, i.e., $\lambda = 0$, the saddle point of (10) is $f^* = f_0$ and $g^* = 0$.

2.2 Examples of Functional Conditional Moment Equation

In this section, we discuss several important applications of the functional conditional moment equation that serve as running examples for this paper.

Policy Evaluation. We consider a Markov decision process given by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S} \subseteq \mathbb{R}^d$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel,

$r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\gamma \in (0, 1)$ is the discount factor. Given a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$, an agent interacts with the environment in the following manner. At a state s_t, the agent takes an action $a_t \sim \pi(\cdot \,|\, s_t)$ and receives a reward $r_t = r(s_t, a_t)$. Then, the agent transits to the next state $s_{t+1} \sim \mathcal{P}(\cdot \,|\, s_t, a_t)$. We denote the transition kernel induced by policy π by $\mathcal{P}^\pi(s' \,|\, s) = \int_{\mathcal{A}} \mathcal{P}(s' \,|\, s, a)\pi(a \,|\, s)\mathrm{d}a$ for any $s, s' \in \mathcal{S}$. In policy evaluation, we aim to estimate the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ defined as follows,

$$V^\pi(s) = \mathbb{E}_\pi\Big[\sum_{i=0}^{\infty} \gamma^i r(s_i, a_i) \,\Big|\, s_0 = s\Big],$$

where the expectation $\mathbb{E}_\pi$ is taken with respect to $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot \,|\, s_t, a_t)$ for $t \geq 0$. By the Bellman equation (Sutton and Barto, 2018), it holds for any $s \in \mathcal{S}$ that

$$V^\pi(s) - \mathcal{T}^\pi V^\pi(s) = 0, \quad \mathcal{T}^\pi f(s) = \mathbb{E}_{a \sim \pi(\cdot \,|\, s)}\big[r(s, a)\big] + \gamma \mathbb{E}_{s' \sim \mathcal{P}^\pi(\cdot \,|\, s)}\big[f(s')\big]. \tag{11}$$

Corresponding to the Bellman equation in (11), let $\mathcal{D}$ denotes the joint distribution of the state-action tuple (s, a, s') under policy π, the value function V^π satisfies the following functional conditional moment equation,

$$\mathbb{E}_{s'|s}\Big[r(s, a) - V^\pi(s) + \gamma \cdot V^\pi(s') \,\Big|\, s\Big] = 0. \tag{12}$$

We notice that (12) is a special case of the functional conditional moment equation in (2) by setting the exogenous variable Z to be the current state s, the endogenous variable X to be the next state s' and the function to be estimated $f : \mathcal{S} \to \mathbb{R}$ to be defined on the state space $\mathcal{S}$. In this case, the functional is $\Phi(X, Z; f) = r + \gamma \cdot f(X) - f(Z)$, where r is the reward function. We remark that the reason the function f can be evaluated simultaneously on X and Z is that both X and Z are variables defined on $\mathcal{S}$. Following the same derivation of (2.1), policy evaluation can be formulated as the following minimax optimization problem,

$$\min_f \max_g \Big\{\mathcal{L}(f, g) = \mathbb{E}_{\mathcal{D}}\Big[g(Z) \cdot \big(r + \gamma \cdot f(X) - f(Z)\big) - 1/2 \cdot g(Z)^2 + \lambda \Psi(X, Z; f)\Big]\Big\}.$$

Nonparametric Instrumental Variables Regression. The nonparametric instrumental variables model is common and useful in statistics and economics. The model can be described simply by one line of equation,

$$Y = f_0(X) + \varepsilon, \quad \mathbb{E}\big[\varepsilon \,\big|\, Z\big] = 0.$$

where Y in an observed outcome, X is the endogenous variable, Z is the exogenous variable, f_0 is the true model that characterize the relationship between Y and X and is also the function we want to estimate. In this model, ε is a noise possibly correlated with the endogenous X but uncorrelated with the exogenous Z. It's straightforward to see that the NPIV model fits into the framework of the functional conditional moment equation by plugging the model equation into the equation about ε,

$$\mathbb{E}_{\mathcal{D}}\Big[Y - f_0(X) \,\Big|\, Z\Big] = 0. \tag{13}$$

We notice that (13) is a special case of the functional conditional moment equation in (5) by identifying X, Z with the endogenous and exogenous variable, respectively, and setting the functional as $\Phi(X, Z; f) = Y - f(X)$. Following the same derivation of (2.1), the problem of NPIV is equivalent to the following minimax optimization problem,

$$\min_f \max_g \Big\{ \mathcal{L}(f, g) = \mathbb{E}_{\mathcal{D}} \Big[g(Z) \cdot \big(Y - f(X) \big) - 1/2 \cdot g(Z)^2 + \lambda \Psi(X, Z; f) \Big] \Big\}.$$

Asset Pricing. Asset pricing refers to the process of determining the fair value of financial assets. This field is fundamental in finance and underpins much of the work in investment, portfolio management, and risk assessment. The Semiparametric Consumption Capital Asset Pricing Model (CCAPM) is a foundational asset pricing model that describes the relationship between systematic risk and expected asset returns, incorporating the influence of investors' consumption preferences over time. Moreover, CCAPM can be characterized through a functional conditional moment equation (Chen et al., 2014; Chen and Ludvigson, 2009). To describe the model, let C_t denote the consumption level at time t, $c_t \equiv C_t / C_{t-1}$ the consumption growth. The marginal utility of consumption at time t is given by $\mathrm{MU}_t = C_t^{-\gamma_0} f_0(c_t)$, where $\gamma_0 > 0$ is the discount factor, $f_0 : \mathcal{C} \to \mathbb{R}$ is the nonparametric structural demand function, which is an unknown positive function of our interest and is defined on $\mathcal{C}$, the space of consumption growth. The unknown function f_0 can be understood as a taste shifter that describes how the marginal utility of consumption changes with the state of the economy in terms of consumption growth.

Now, consider the growth-return tuple $(c_t, \tilde{r}_{t+1}, c_{t+1})$ for $t \in \mathbb{N}^+$ with joint distribution $\mathcal{D}$, where c_t is the consumption growth at the current time t, and c_{t+1} is the consumption growth at the next time $t+1$. $\tilde{r}_{t+1}$ is a modified return observed in this period, which is a known function of the actual return r_{t+1} and the consumption growth c_{t+1} at time $t+1$. We consider the scenario where the time series of consumption growth $\{c_t\}_{t \geq 0}$ follows a time-homogeneous Markov chain with a smooth transition kernel. That being said, both conditional transition probabilities $c_{t+1}|c_t$ and $c_t|c_{t+1}$ admit a smooth density function. The CCAPM model captures the behavior of f_0 through the following equation:

$$\mathbb{E}_{c_{t+1}|c_t} \big[\tilde{r}_{t+1} \cdot f_0(c_{t+1}) - f_0(c_t) \,\big|\, c_t \big] = 0, \tag{14}$$

where the modified return can be further expressed as $\tilde{r}_{t+1} = \delta_0 \cdot r_{t+1} \cdot c_{t+1}^{-\gamma_0}$, $\delta_0 \in (0, 1]$ is the rate of time preference. We focus on a setting where $\mathcal{C} \subseteq \mathbb{R}$ is a compact set, and the modified return $\tilde{r}_{t+1}$ is bounded for all $t \geq 0$. We notice that (14) is a special case of the functional conditional moment equation in (5). We can identify the exogenous variable Z with c_t, the consumption growth at the current time t, and the endogenous variable X with c_{t+1}, the consumption growth at the next time $t+1$. In this scenario, we identify the space $\mathcal{W}$ with $\mathcal{C}$, the space of consumption growth, and the function to be estimated $f : \mathcal{C} \to \mathbb{R}$ is defined on $\mathcal{C}$. The functional is $\Phi(X, Z; f) = \tilde{r}_{t+1} \cdot f(X) - f(Z)$, where $\tilde{r}_{t+1}$ again denotes the modified return. Similar to the scenario of policy evaluation, the reason function f can be evaluated simultaneously on X and Z is that both X and Z are variables defined on $\mathcal{C}$. Following the same derivation of (2.1), the problem of asset pricing through CCAPM is equivalent to the following minimax optimization problem,

$$\min_f \max_g \Big\{ \mathcal{L}(f, g) = \mathbb{E}_{\mathcal{D}} \Big[g(Z) \cdot (\tilde{r}_{t+1} \cdot f(X) - f(Z)) - 1/2 \cdot g(Z)^2 + \lambda \Psi(X, Z; f) \Big] \Big\}$$

2.3 Mean-Field Neural Network and Wasserstein Space

In the sequel, we will consider functions that can be represented through a class of neural networks. Consider a neural function defined on a given state space Ω, $\sigma : \Omega \times \mathbb{R}^D \to \mathbb{R}$ that takes an input $x \in \Omega$ and parameter $\theta \in \mathbb{R}^D$ and outputs a value in $\mathbb{R}$. For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ where $\theta_i \in \mathbb{R}^D$, we can define an over-parameterized two-layered neural network function h using neuron function σ,

$$h(x, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; \theta_i), \quad \forall x \in \Omega.$$

For such a form, we can further consider the infinite width limit when $N \to \infty$. When taking such a limit, the neural network function h becomes a mean-field neural network and can be parameterized with a probability measure over the parameter space, $\mu \in \mathscr{P}(\mathbb{R}^D)$.

$$h(x; \mu) = \int_{\mathbb{R}^D} \sigma(x; \theta) \mathrm{d}\mu(\theta), \quad \forall x \in \Omega.$$

When considering such a limit, the optimization problem over the neural network function class shifts from a finite-dimensional problem over the parameter space to an infinite-dimensional problem over the space of probability measures. Therefore, we will need to track the convergence of probability measures on the Wasserstein space when analyzing algorithm convergence.

We now introduce the background knowledge of the Wasserstein space for the reader's information. Let $\mathscr{P}_p(\mathbb{R}^D)$ be the space of all the probability measures over the D-dimensional Euclidean space $\mathbb{R}^D$ with finite p-th order moments. The Wasserstein-p distance between two probability measures $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^D)$ is defined as follows,

$$\mathcal{W}_p(\mu, \nu) = \inf\Big\{ \Big(\int \|x - y\|^p \mathrm{d}\gamma(x, y) \Big)^{1/p} \Big| \, \gamma \in \mathscr{P}_p(\mathbb{R}^D \times \mathbb{R}^D), x_\sharp \gamma = \mu, y_\sharp \gamma = \nu \Big\}, \quad (15)$$

where the infimum is taken over all the coupling of μ and ν. Here we denote by $x_\sharp \gamma$ and $y_\sharp \gamma$ the marginal distributions of γ with respect to x and y, respectively. We call $\mathcal{M}_p = (\mathscr{P}_p(\mathbb{R}^D), \mathcal{W}_p)$ the Wasserstein-p space. For any $1 \le p \le q$, due to the relation that $\mathbb{E}[|X|^p]^{1/p} \le \mathbb{E}[|X|^q]^{1/q}$, we have that $W_p(\mu, \nu) \le W_q(\mu, \nu)$ for two measures μ, ν. In this paper, we focus on the cases when $p = 1, 2$. Without further clarification, we refer to the distance with $p = 2$ as the Wasserstein distance in the remainder of the text.

The Wasserstein-2 space $\mathcal{M}_2 = (\mathscr{P}_2(\mathbb{R}^D), \mathcal{W}_2)$ can be viewed as an infinite-dimensional Riemannian manifold (Villani, 2008). Formally, the tangent space at point $\rho \in \mathscr{P}_2(\mathbb{R}^D)$ is defined as

$$\mathrm{Tan}_\rho\big(\mathscr{P}_2(\mathbb{R}^D)\big) = \Big\{ v \in L^2(\rho) \Big| \int \langle v, u \rangle d\rho = 0, \forall u \in L^2(\rho) \text{ s.t. } \mathrm{div}(u\rho) = 0 \Big\}.$$

Then, for any absolutely continuous curve $\rho : [0, 1] \to \mathscr{P}_2(\mathbb{R}^D)$ on the Wasserstein-2 space, there exists a family of vector fields $v_t \in \mathrm{Tan}_{\rho_t}(\mathscr{P}_2(\mathbb{R}^D))$ such that the continuity equation

$$\partial_t \rho_t + \mathrm{div}(v_t \rho_t) = 0 \quad (16)$$

holds in the sense of distributions. For any two absolutely continuous curves $\rho, \tilde{\rho} : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$, we define the inner product between $\partial_t \rho_t, \partial_t \tilde{\rho}_t$ for any $t \in [0,1]$ as follows,

$$\langle \partial_t \rho_t, \partial_t \tilde{\rho}_t \rangle_{\rho_t} = \int \langle v_t, \tilde{v}_t \rangle \mathrm{d}\rho_t, \tag{17}$$

where $\langle v_t, \tilde{v}_t \rangle$ is the inner product over $\mathbb{R}^D$, (ρ_t, v_t) and $(\tilde{\rho}_t, \tilde{v}_t)$ satisfy the continuity equation in (16). Note that (17) yields a Riemannian metric over $\mathcal{M}_2$. Furthermore, the Riemannian metric induces a norm $\|\partial_t \rho_t\|_{\rho_t} = \langle \partial \rho_t, \partial_t \rho_t \rangle_{\rho_t}^{1/2}$.

3. Algorithms

In this section, we introduce the stochastic gradient descent-ascent algorithm (SGDA) and its mean-field limit, which is characterized by the continuity equation.

Stochastic Gradient Descent-Ascent Algorithm. We solve the minimax optimization problem in (10) via SGDA. Recall that in the minimax objective, we have two functions simultaneously involved, where the primal function f represents the true model of interest and the dual function g represents an adversarial player. Specifically, we parameterize both f and g with neural networks with width N and parameters $\boldsymbol{\theta} = (\theta^1, \theta^2, \ldots, \theta^N) \in \mathbb{R}^{D \times N}$ and $\boldsymbol{\omega} = (\omega^1, \omega^2, \ldots, \omega^N) \in \mathbb{R}^{D \times N}$

$$f(\cdot; \boldsymbol{\theta}) = \frac{\alpha}{N} \sum_{i=1}^{N} \phi(\cdot; \theta^i), \quad g(\cdot; \boldsymbol{\omega}) = \frac{\alpha}{N} \sum_{i=1}^{N} \psi(\cdot; \omega^i). \tag{18}$$

where we use bold symbols $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ to denote the whole parameter used by each neural net and unbold symbols θ and ω to denote the parameter used by each neuron. Here, $\phi(\cdot; \theta) : \mathcal{W} \times \mathbb{R}^D \to \mathbb{R}$, $\psi(\cdot; \omega) : \mathcal{Z} \times \mathbb{R}^D \to \mathbb{R}$ are the functions for neurons. In particular, we can recover the general setting of two-layer neural networks parameterization for f and g when we choose ϕ, ψ to be the following specific form,

$$\phi(w; \beta, W) = \beta \cdot \sigma_f(w; W), \quad \psi(z; \beta, W) = \beta \cdot \sigma_g(z; W),$$

where $\sigma_f : \mathcal{W} \times \mathbb{R}^D \to \mathbb{R}$, $\sigma_g : \mathcal{Z} \times \mathbb{R}^D \to \mathbb{R}$ are activation functions with input w and z respectively and parameters W. We note that it's not necessary to choose the same width N for f and g, and activation functions σ_f, σ_g need not have the same parameter dimension D. Here we use the same width N and parameter dimension D to keep notations simple as these won't affect the validity of the results presented in this paper.

Besides, we have also introduced a scaling factor $\alpha > 0$ in (18). Setting the scaling parameter $\alpha = \sqrt{N}$ in (18) recovers the neural tangent kernel regime (Jacot et al., 2018). Setting the parameter $\alpha = 1$ recovers the mean-field regime (Mei et al., 2018, 2019). In a discrete-time finite-width (DF) scenario, at the kth iteration, the primal function f and adversarial player g are updated as follows,

$$\begin{aligned} &\text{DF-GD:} \ \ \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \cdot g(z_k; \omega_k) \cdot \nabla_{\boldsymbol{\theta}} \Phi(x_k, z_k; f(\cdot; \boldsymbol{\theta}_k)) - \eta\lambda \cdot \nabla_{\boldsymbol{\theta}} \Psi(x_k, z_k; f(\cdot; \boldsymbol{\theta}_k)), \\ &\text{DF-GA:} \ \ \boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k + \eta \cdot \Phi(x_k, z_k; f(\cdot; \boldsymbol{\theta}_k)) \cdot \nabla_{\boldsymbol{\omega}} g(z_k; \boldsymbol{\omega}_k) - \eta \cdot g(z_k; \boldsymbol{\omega}_k) \cdot \nabla_{\boldsymbol{\omega}} g(z_k; \boldsymbol{\omega}_k), \end{aligned} \tag{19}$$

where $\boldsymbol{\theta}_k, \boldsymbol{\omega}_k$ denotes the state of the parameters at iteration k, $\eta > 0$ is the step-size, and the data samples $\{(x_k, z_k)\}_{k=0}^{\infty}$ are collected by independently sampling from the data distribution $\mathcal{D}$. When f, g are two-layered neural networks with width N, we can plug in the form for f, g as is described in (18). The update for the parameter of i-th neuron at k-th iteration can be further specified as follows,

$$\theta_{k+1}^i = \theta_k^i - \eta\alpha\epsilon \cdot g(z_k; \boldsymbol{\omega}_k)\nabla_\theta \Phi(x_k, z_k; \phi(\cdot, \theta_k^i)) - \eta\lambda\epsilon \cdot \frac{\delta\Psi(x_k, z_k; f(\cdot, \boldsymbol{\theta}_k))}{\delta f}\nabla_\theta \phi(x_k; \theta_k^i),$$
$$\omega_{k+1}^i = \omega_k^i + \eta\alpha\epsilon \cdot \Phi(x_k, z_k; f(\cdot, \boldsymbol{\theta}_k))\nabla_\omega \psi(z_k; \omega_k^i) - \eta\alpha\epsilon \cdot g(z_k; \boldsymbol{\omega}_k)\nabla_\omega \psi(z_k; \omega_k^i), \tag{20}$$

where $\boldsymbol{\theta}_k = (\theta_k^1, \theta_k^2, \ldots, \theta_k^N)$ and $\boldsymbol{\omega}_k = (\omega_k^1, \omega_k^2, \ldots, \omega_k^N)$, $\delta\Psi/\delta f$ denotes the variation of Ψ with respect to f. Here, α is the neural network scaling parameter and $\epsilon = 1/N$ is the stepsize scale. Both α and ϵ show up in (20) due to the finite width parameterization of two-layered neural networks described in (18).

For a given space $\mathcal{S}$, let $\mathcal{H}$ denote a class of functions from $\mathcal{S}$ to $\mathbb{R}$. For a functional defined over $\mathcal{H}$, $F : \mathcal{H} \to \mathbb{R}$, its variation at $f \in \mathcal{H}$ is a function $\frac{\delta F}{\delta f} : \mathcal{S} \to \mathbb{R}$, such that for any test function $h \in \mathcal{H}$,

$$\Big[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(f + \varepsilon h)\Big]_{\varepsilon=0} = \int_{\mathcal{S}} \frac{\delta F}{\delta f}(s) \cdot h(s)\ \mathrm{d}s. \tag{21}$$

We initialize the parameters with $\theta_0^i \sim \mu_0$ and $w_0^i \sim \nu_0$, with $\mu_0, \nu_0 = \mathcal{N}(0, I_D)$ be standard Gaussian distribution in $\mathbb{R}^D$. In addition, to keep track of the evolution of the parameter distribution, we denote the empirical distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ at the kth iteration by,

$$\hat{\mu}_k(\theta) = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_k^i}(\theta), \quad \hat{\nu}_k(\omega) = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{\omega}_k^i}(\omega),$$

where δ is the Dirac mass function.

Mean-Field (MF) Limit. To analyze the convergence of the Stochastic Gradient Descent-Ascent Algorithm for solving functional conditional moment equations with neural networks, we employ an analysis that studies the mean-field limit regime (Mei et al., 2018, 2019) of the discrete-time dynamics described in (19). Here, by the mean-field limit, we are referring to an infinite-width limit, i.e., when $N \to \infty$ for the neural network width and a continuous time, i.e., $t = k\epsilon$ where the step scale $\epsilon \to 0$ in (20). In what follows, we introduce the mean-field limit of the SGDA dynamics, which refers to the infinite-width and continuous limit of (20). For $\boldsymbol{\theta} = \{\theta^i\}_{i=1}^N$ and $\boldsymbol{\omega} = \{\omega^i\}_{i=1}^N$ independently sampled respectively from $\mu, \nu \in \mathscr{P}(\mathbb{R}^D)$, we can write the infinite width limit of neural networks used in (18) as

$$f(\cdot; \mu) = \alpha \int \phi(\cdot; \theta)\mu(\mathrm{d}\theta), \quad g(\cdot; \nu) = \alpha \int \psi(\cdot; \omega)\nu(\mathrm{d}\omega). \tag{22}$$

From now on, we denote by μ_t the distribution of θ_t^i and ν_t the distribution of ω_t^i for the infinite-width and continuous limit of the neural networks at time t. For notational simplicity, we overload the notation of the objective function in (10) via $\mathcal{L}(\mu, \nu) = \mathcal{L}(f(\cdot; \mu), g(\cdot; \nu))$. This is to further emphasize the dependence of objective $\mathcal{L}$ on (μ, ν) when we parameterize

the function pair (f, g) using distributions on the parameter space. By Otto's calculus (Villani, 2008), the mean-field limit of the update direction takes the following form,

$$
\begin{aligned}
v^f(\theta;\mu,\nu) &= -\nabla_\theta \frac{\delta \mathcal{L}(\mu,\nu)}{\delta \mu}(\theta) \\
&= \alpha \mathbb{E}_{\mathcal{D}}\Big[-g(Z;\nu)\big\langle \frac{\delta \Phi(X,Z;f(\cdot;\mu))}{\delta f}, \nabla_\theta \phi(\cdot;\theta)\big\rangle_{L^2} - \lambda \Big\langle \frac{\delta \Psi(X,Z;f(\cdot;\mu))}{\delta f}, \nabla_\theta \phi(\cdot;\theta)\Big\rangle_{L^2}\Big], \\
v^g(\omega;\mu,\nu) &= \nabla_\omega \frac{\delta \mathcal{L}(\mu,\nu)}{\delta \nu}(\omega) \\
&= \alpha \mathbb{E}_{\mathcal{D}}\Big[\Phi(X,Z;f(\cdot,\mu))\nabla_\omega \psi(Z;\omega) - g(Z;\nu)\nabla_\omega \psi(Z;\omega)\Big]. \qquad (23)
\end{aligned}
$$

Here $\langle \cdot, \cdot \rangle_{L_2}$ is the inner product on $L^2(\mathcal{X} \times \mathcal{Z})$ with respect to the Lebesgue measure. Recall that $\mathcal{D}$ is the data distribution of random variables $(X, Z) \in \mathcal{X} \times \mathcal{Z}$, we denote by $\rho_{\mathcal{X},\mathcal{Z}}$ the density of $\mathcal{D}$ with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Z}$ and we use $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ to represent the inner product on $L^2(\mathcal{X} \times \mathcal{Z})$ with respect to the probability distribution $\mathcal{D}$. That is to say, for any two function $h_1, h_2 \in L^2(\mathcal{X} \times \mathcal{Z})$, $\langle h_1, h_2 \rangle_{\mathcal{D}} = \int_{\mathcal{X}\times\mathcal{Z}} h_1 h_2 \, \mathrm{d}\rho_{\mathcal{X},\mathcal{Z}}$.

We will also slightly abuse this notation and use $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ to denote the inner product on sub-spaces of $L^2(\mathcal{X} \times \mathcal{Z})$, with the measure being the marginals of $\mathcal{D}$ on these sub-spaces. In (23), $\delta\Phi/\delta f$ and $\delta\Psi/\delta f$ is the variation of Φ and Ψ over f under $\langle \cdot, \cdot \rangle_{L^2}$, where the test functions are chosen over the function class $\mathcal{F}$. In the same way, $\delta\mathcal{L}/\delta\mu$ and $\delta\mathcal{L}/\delta\nu$ respectively denote the variation of the objective $\mathcal{L}$ with respect to distributions μ and ν under $\langle \cdot, \cdot \rangle_{L^2}$, following definition in (21) with the test function chosen over $\mathscr{P}(\mathcal{X} \times \mathcal{Z})$. We also remark that we can also define the variation under $\langle \cdot, \cdot \rangle_{\mathcal{D}}$, which will only differ from the variation under $\langle \cdot, \cdot \rangle_{L^2}$ by a constant function factor that corresponds to the density of the marginals of $\mathcal{D}$. Then, the mean-field limit of the SGDA update in (19) is characterized by the continuity equation, which is a system of PDEs given by,

$$
\partial_t \mu_t(\theta) = -\eta \cdot \mathrm{div}_\theta\big(\mu_t(\theta) v^f(\theta;\mu_t,\nu_t)\big), \ \partial_t \nu_t(\omega) = -\eta \cdot \mathrm{div}_\omega\big(\nu_t(\omega) v^g(\omega;\mu_t,\nu_t)\big), \qquad (24)
$$

where div_θ, div_ω denotes the divergence with respect to θ, ω respectively. Note that the initialization μ_0 and ν_0 are the same as the initialization of the discrete-time dynamics in (20), i.e. $\mu_0 = \mathcal{N}(0, I_D)$, $\nu_0 = \mathcal{N}(0, I_D)$ are taken to be the distribution of standard Gaussian random variables in $\mathbb{R}^D$.

4. Main Results

In this section, we introduce the main theoretical results of the stochastic gradient descent-ascent dynamics. We first present the assumptions in §4.1. Then, in §4.2, we show that the SGDA dynamics converge to a mean-field limit as the network size N goes to infinity and the step size scale ϵ goes to zero. Finally, in §4.3 we prove that the mean-field limiting dynamics converge to a globally optimal solution of the primal objective J under proper assumptions. Moreover, we will show that the mean-field dynamics learns a data-dependent representation that is $\mathcal{O}(\alpha^{-1})$ away from the initial representation.

4.1 Assumptions

We introduce three main assumptions in this work. Assumptions 1 and 2 discuss the richness and regularity of the two-layered neural network function class, which will be the space in which we search for solutions to the minimax optimization problem. Assumption 3 presents regularity conditions on the data space and the smoothness of the functionals.

We start by discussing the two-layered neural network function class. Consider the neuron function ϕ and ψ with the following form,

$$\phi(w;\theta) = b(\beta)\cdot\sigma(\tilde{\theta}^\top(w,1)), \quad \psi(z;\omega) = b(\beta)\cdot\sigma(\tilde{\omega}^\top(z,1)), \tag{25}$$

where $\theta = (\beta,\tilde{\theta}) \in \mathbb{R}\times\mathbb{R}^{1+\dim(\mathcal{W})}$, $\omega = (\beta,\tilde{\omega}) \in \mathbb{R}\times\mathbb{R}^{1+\dim(\mathcal{Z})}$ contains the parameters in the output layer and the hidden layer, $b:\mathbb{R}\to\mathbb{R}$ is an odd re-scaling function and $\sigma:\mathbb{R}\to\mathbb{R}$ is the activation function. Note that such a form of activation function satisfies the condition of the universal function approximation theorem (Theorem 3.1 in Pinkus (1999)) if σ is not a polynomial. For notational simplicity, we write $\sigma(w;\tilde{\theta}) = \sigma(\tilde{\theta}^\top(w,1))$. The re-scaling function $b:\mathbb{R}\to\mathbb{R}$ is introduced to ensure that the value of the neural network is upper-bounded. When $b(\mathbb{R}) = (-B_0, B_0)$, the function class induced by the neural network in (22) is equivalent to the following class,

$$\mathcal{F} = \Big\{ f:\mathcal{W}\to\mathbb{R} \,\Big|\, f(w) = \int \beta'\cdot\sigma(w;\tilde{\theta})\,\mathrm{d}\mu(\beta',\tilde{\theta}), \mu\in\mathscr{P}_2\big((-B_0,B_0)\times\mathbb{R}^{d+1}\big)\Big\}, \tag{26}$$

where $d = \dim(\mathcal{W})$. This captures a rich function class due to the universal function approximation theorem (Barron, 1993; Pinkus, 1999). We remark that we introduce the re-scaling function $b(\beta)$ in (25) to avoid the study of the space of probability measures over $(-B_0,B_0)\times\mathbb{R}^{d+1}$, which has a boundary and thus lacks regularity in the study of optimal transport. Moreover, note that a scaling hyperparameter $\alpha > 0$ is introduced in the definition of the mean-field neural nets in (22). When $\alpha > 1$, this results in an over-parameterization effect. In brief, α controls the error between the $(f(\cdot;\mu_t), g(\cdot;\mu_t))$ and optimizer (f^*, g^*) according to Theorem 7. Furthermore, the over-parameterization scale α has an influence through Lemma 6, which shows that the Wasserstein distance between the Gaussian initialization (μ_0,ν_0) and the optimal distribution (μ^*,ν^*) is upper-bounded by $\mathcal{O}(\alpha^{-1})$. Next, we impose the following regularity assumptions on the neural network functions ϕ and ψ.

Assumption 1 (Regularity of Neural Networks) *There exist fixed finite constants $B_0 > 0$, $B_1 > 0$ and $B_2 > 0$ such that*

$$\begin{aligned} &|\phi(w;\theta)| \le B_0, \quad \big\|\nabla_\theta\phi(w;\theta)\big\| \le B_1, \quad \big\|\nabla^2_{\theta\theta}\phi(w;\theta)\big\|_F \le B_2, \qquad \forall w\in\mathcal{W},\ \theta\in\mathbb{R}^D,\\ &|\psi(z;\omega)| \le B_0, \quad \big\|\nabla_\omega\psi(z;\omega)\big\| \le B_1, \quad \big\|\nabla^2_{\omega\omega}\phi(z;\omega)\big\|_F \le B_2, \qquad \forall z\in\mathcal{Z},\ \omega\in\mathbb{R}^D, \end{aligned}$$

where $\nabla^2_{\theta\theta}, \nabla^2_{\omega\omega}$ denotes the hessian with respect to θ and ω respectively, $\|\cdot\|$ denotes the vector $2-$norm, and $\|\cdot\|_F$ denotes the matrix Frobenius norm. Moreover, the rescaling function $b:\mathbb{R}\to\mathbb{R}$ is odd and its range satisfies that $b(\mathbb{R}) = (-B_0, B_0)$.

Assumption 1 is satisfied by a broad class of neuron functions. For example, it is satisfied when we set the activation function $\sigma(x) = \mathrm{sigmoid}(x)$ and rescaling function $b(\beta) = \tanh(\beta)$.

We also impose the following assumption regarding the realizability of the saddle point solution (f^*, g^*) to (10).

Assumption 2 (Realizability) *The saddle point solution (f^*, g^*) of (10) belongs to the function class defined in (26), i.e., $f^*, g^* \in \mathcal{F}$.*

In general, problem (10) may not admit a saddle point within the given neural network function class. Therefore, Assumption 2 is introduced to guarantee that the discussion in this paper is meaningful. By the universal function approximation theorem (Barron, 1993; Pinkus, 1999), the function class defined in (26) captures a rich class of functions. Therefore, this assumption is quite general and does not restrict the applicability of our results.

We impose the following set of conditions on the data space $\mathcal{X}$ and $\mathcal{Z}$, and on the integrability of the functionals Φ and Ψ and their variations.

Assumption 3 (Data regularity and Functional Integrability)

(i) *the data space $\mathcal{X} \times \mathcal{Z}$ is compact, in the sense that there exists a positive constant $C_1 > 0$ such that for any data tuple $(x, z) \in \mathcal{X} \times \mathcal{Z}$, it satisfies that $\|(x, z)\| \leq C_1$. Moreover, the data distribution $\mathcal{D}$ admits a positive, smooth density $\rho_{\mathcal{D}}$ with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Z}$.*

(ii) *For the functionals $\Phi(x, z; f)$, there exists a positive constant $C_2 > 0$ such that*

$$\int_{\mathcal{W}} \Big| \frac{\delta \Phi(x, z; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq C_2, \quad \forall (x, z) \in \mathcal{X} \times \mathcal{Z}.$$

(iii) *We assume that $\int_{\mathcal{W}} \frac{\delta \Psi(x,z;f)}{\delta f}(w')\mathrm{d}w'$, as a linear functional of f, is upper-bounded by a constant times the values of f. That is, there exists $w \in \mathcal{W}$ as a part of the data tuple (x, z) and a positive constant $C_\Psi > 0$ such that*

$$\int_{\mathcal{W}} \Big| \frac{\delta \Psi(x, z; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq C_\Psi \cdot \big| f(w) \big|.$$

(iv) *The variations of the minimax objective $\mathcal{L}(f, g)$ with respect to f and g are continuous functions on $\mathcal{W}$ and $\mathcal{Z}$. That is,*

$$\frac{\delta \mathcal{L}(f, g)}{\delta f} \in \mathscr{C}(\mathcal{W}), \quad \frac{\delta \mathcal{L}(f, g)}{\delta g} \in \mathscr{C}(\mathcal{Z}).$$

Item (i) of Assumption 3 restricts our scenarios to data spaces with bounded values and smooth densities for technical reasons. Items (ii) and (iii) of Assumption 3 are integrability conditions that we additionally require to avoid discussion of improper functionals that potentially have singularities with exploding values. Item (iv) is a smoothness condition that involves the variation of the minimax objective averaged over data to be continuous on the respective space. A sufficient condition for item (iv) to hold is the continuity of the variation of Φ and Ψ with respect to f averaged under the marginal of $\mathcal{D}$ on $\mathcal{W}$. We will use this sufficient condition to verify item (iv) in practice. These are mild conditions that are satisfied by many applications in machine learning, causal inference, and statistics.

4.2 Convergence of SGDA dynamics to the Mean-Field Limit

In the following proposition, we show that the empirical distribution of the parameters $\hat{\mu}_k$ and $\hat{\nu}_k$ weakly converges to the mean-field limit in (24) as the width N goes to infinity and the stepsize scale ϵ goes to zero. Let $\rho_t(\theta, \omega) := \mu_t(\theta) \otimes \nu_t(\omega)$, where (μ_t, ν_t) is the PDE solution to the continuous deterministic dynamics in (24) and $\hat{\rho}_k := N^{-1} \cdot \sum_{i=1}^{N} \delta_{\theta_k^i} \cdot \delta_{\omega_k^i}$ corresponds to the empirical distribution of $(\boldsymbol{\theta}_k, \boldsymbol{\omega}_k)$, which is k-th iterate of the discrete time stochastic dynamics in (20) with stepsize scale ϵ. The following proposition proves that the PDE solution ρ_t in (24) well approximates the discrete time stochastic gradient descent-ascent dynamics in (20).

Proposition 4 (Convergence of SGDA to Mean-Field Limit) *Let $\{\rho_t\}_{t\geq 0}$ be solution to (24) with $\rho_0 = \mathcal{N}(0, I_D) \otimes \mathcal{N}(0, I_D)$, $\{\hat{\rho}_k\}_{k\geq 0}$ be solution to (20) with $\hat{\rho}_0 = \mathcal{N}(0, I_D) \otimes \mathcal{N}(0, I_D)$. Under Assumptions 1 and 3, $\widehat{\rho}_{\lfloor t/\epsilon \rfloor}$ converges weakly to ρ_t as $\epsilon \to 0^+$ and $N \to \infty$. It holds for any Lipschitz continuous, bounded function $F : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ that*

$$\lim_{\epsilon \to 0^+, N \to \infty} \int F(\theta, \omega) \mathrm{d}\widehat{\rho}_{\lfloor t/\epsilon \rfloor}(\theta, \omega) = \int F(\theta, \omega) \mathrm{d}\rho_t(\theta, \omega).$$

Proof See §C for a detailed proof. ■

The proof of Proposition 4 is based on the propagation of chaos (Mei et al., 2018, 2019; Araújo et al., 2019; Zhang et al., 2020; Sznitman, 1991). In this proposition, we establish a weak convergence for any fixed t, which allows us to convert the discrete-time SGDA dynamics over a finite-dimensional parameter space to its continuous-time, infinite-dimensional counterpart in Wasserstein space, in which the training is amenable to analysis since our infinitely wide neural network $f(\cdot; \mu)$ and $g(\cdot; \nu)$) in (22) are linear in μ and ν respectively.

4.3 Global Optimality and Convergence of the Mean-Field Limit

In this section, we introduce our main results, which characterize the global optimality and convergence of mean-field neural networks parameterized by the parameter distribution $\rho_t = (\mu_t, \nu_t)$. The proof contains two steps. We first demonstrate that it is sufficient to find a stationary point of the Wasserstein gradient flow defined in (24) to solve the minimax optimization problem in (10). Then, we characterize the convergence of ρ_t to the stationary point. Before presenting the two stages of the proof, we need to further clarify the notion of stationarity for the Wasserstein gradient flow. We introduce the following definition,

Definition 5 (Stationary point of Wasserstein Gradient Flow) *A distribution pair (μ, ν) is called a stationary point of the Wasserstein gradient flow (24) if it satisfies*

$$v^f(\theta; \mu, \nu) = v^g(\omega; \mu, \nu) = 0, \quad \forall \theta, \omega \in \mathbb{R}^D.$$

From Definition 5, the stationary point of Wasserstein gradient flow (24) is a distribution pair (μ, ν), at which the associated vector field $(v^f(\cdot; \mu, \nu), v^g(\cdot; \mu, \nu))$ is a zero function on the parameter space $\mathbb{R}^D \times \mathbb{R}^D$. Moreover, for the Wasserstein gradient flow following vector field (v^f, v^g) and initial condition (μ, ν), the solution to its associated continuity equation (μ_t, ν_t) is a constant flow such that for all $t \geq 0$, $\mu_t = \mu, \nu_t = \nu$. Now, we have the following important supporting lemma that characterizes the relation between stationary points of Wasserstein gradient flow (24) and saddle points of (10).

Lemma 6 *Under Assumptions 1 and 2, the following two properties hold.*

(i) Suppose that (μ^*, ν^*) is a stationary point of the Wasserstein gradient flow as is defined in Definition 5. Then, the corresponding function $(f(\cdot;\mu^*), g(\cdot;\nu^*))$ is the saddle point of the objective function $\mathcal{L}(f,g)$ defined in (10).

(ii) There exists a stationary distribution pair (μ^*, ν^*) and constant $\bar{D} > 0$ such that

$$W_2(\mu_0, \mu^*) \le \alpha^{-1}\bar{D}, \quad W_2(\nu_0, \nu^*) \le \alpha^{-1}\bar{D}.$$

Proof See §B.1 for a detailed proof. ■

Lemma 6 demonstrates that the stationary point of the Wasserstein gradient flow in (24) achieves global optimality as a solution to the minimax objective (10). Lemma 6 allows us to bypass the hardness of solving the nonconvex-nonconcave optimization problem (10) of finding saddle points in the space of neural network parameters $(\boldsymbol{\theta}, \boldsymbol{\omega})$ by searching for a stationary point of the Wasserstein gradient flow instead. Moreover, there exist good pairs of stationary points that are close to the Gaussian initialization (μ_0, ν_0), with Wasserstein distance upper bounded by order $\mathcal{O}(\alpha^{-1})$.

We are now ready to show our main results. The following theorem characterizes the global optimality and convergence of the Wasserstein gradient flow ρ_t.

Theorem 7 (Global Convergence to Saddle Point) *Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (24) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 = \mathcal{N}(0, I_D)$, (f^*, g^*) the saddle point of the minimax objective $\mathcal{L}(f,g)$ in (10). Under Assumptions 1, 2, and 3, it holds that*

$$\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi\big(X, Z; f(\cdot;\mu_t) - f^*(\cdot)\big) + \big(g(Z;\nu_t) - g^*(Z)\big)^2\Big] \le \mathcal{O}(T^{-1} + \alpha^{-1}). \tag{27}$$

Proof See §B.2 for a detailed proof. ■

Theorem 7 says that the optimality gap between $(f(\cdot;\mu_t), g(\cdot;\nu_t))$ and (f^*, g^*), quantified by the Ψ-induced distance and L^2 distance respectively, decays to zero at a sublinear rate in terms of time T up to the error of $\mathcal{O}(\alpha^{-1})$, where $\alpha > 0$ is the scaling parameter in (18) and (22). To prove the convergence, we construct a potential $V(\mu,\nu) = \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi\big(X, Z; f(\cdot;\mu) - f^*(\cdot)\big) + \big(g(Z;\nu) - g^*(Z)\big)^2\Big]$, with $V(\mu,\nu) = 0$ if and only if $(\mu,\nu) = (\mu^*,\nu^*)$. Such a potential characterizes the saddle point of the minimax objective. We demonstrate that the Wasserstein gradient flow decreases the potential at a sublinear rate, thereby suggesting convergence of the gradient flow to the saddle point.

Moreover, a varying α allows a trade-off between the error of order $\mathcal{O}(\alpha^{-1})$ in the optimality gap and the maximum deviation between ρ_t and the Gaussian initialization ρ_0 for all t. In the proof of item (ii) of Lemma 6, we proved that the deviation of ρ_t from ρ_0 quantified by the Wasserstein distance is of order $\mathcal{O}(\alpha^{-1})$. Regarding representation learning, this suggests that SGDA induces a data-dependent representation that is significantly different from the initialization. Choosing a small α of order $\mathcal{O}(1)$ will correspond to the mean-field regime (Mei et al., 2018, 2019) that allows ρ_t to move further away from the initialization, with the

potential drawback of yielding a large error of order $\mathcal{O}(\alpha^{-1})$. On the other hand, choosing a large α of order $\mathcal{O}(\sqrt{N})$ will correspond to the NTK regime (Jacot et al., 2018), and this causes the Wasserstein flow ρ_t to stay close to the initial distribution ρ_0 along the trajectory, inducing a data-independent representation. Our analysis produces richer results as we can flexibly choose a scale for α that is different from $\mathcal{O}(1)$ or $\mathcal{O}(\sqrt{N})$. Theorem 7 also implies a trade-off between the strength of representation learning and the speed of optimization. When α is large, the optimization proceeds faster; meanwhile, the representation remains nearly unchanged relative to its initialization, mimicking the phenomenon observed in NTK analysis. We can also select α to grow at a much slower rate. One choice could be $\alpha = \mathcal{O}(\log T)$. With this option, although the gradient flow takes longer to reach optimality, the feature movement along the trajectory may also be more significant.

As we have commented before, an important class of regularizer $\Psi(X, Z; f)$ is the L^2 regularizer. In this scenario, the left-hand side of (27) should be understood as a weighted L^2 distance between the gradient flow iterate at time t to the optimal solution (f^*, g^*). As T and α go to infinity, such a distance will shrink to 0, thus the gradient flow converges globally in the minimal distance sense to the optimal solution. Due to this observation, in the sequel, we will discuss several additional results in the case where the regularizer $\Psi(X, Z; f)$ is strongly convex, in the sense that it's bounded below by a quadratic function. We formalize the additional constraint in this case with the following assumption,

Assumption 8 (Strong Convexity) *The regularizer $\Psi(X, Z; f)$ is c_Ψ-strongly convex, in the sense that there exists a constant $c_\Psi > 0$ such that for any $f \in \mathcal{F}$,*

$$\Psi(x, z; f) \geq c_\Psi \cdot |f(w)|^2, \quad \forall (x, z) \in \mathcal{X} \times \mathcal{Z},$$

where $w \in \mathcal{W}$ is part of the data tuple (x, z).

Assumption 8 implies that regularizer $\Psi(X, Z; f)$ is equivalent to a quadratic regularizer because Ψ is simultaneously bounded above and below by quadratic functionals. We have the following strengthened version of Theorem 7 in such case,

$$\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[\lambda c_\Psi \cdot \big(f(\cdot;\mu_t) - f^*(\cdot)\big)^2 + \big(g(Z;\nu_t) - g^*(Z)\big)^2\Big] \leq \mathcal{O}(T^{-1} + \alpha^{-1}). \tag{28}$$

Equation (28) shows that the iterates $(f(\cdot;\mu_t), g(\cdot;\nu_t))$ converges to the saddle point solution (f^*, g^*) as a weighted L^2 distance decays to zero at a sublinear rate up to an error of $\mathcal{O}(\alpha^{-1})$. With Assumption 2, the saddle point f^* is the global optimizer of the primal functional $J(f)$ defined in (4). Therefore, as a direct consequence of Theorem 7, when the regularizer Ψ is strongly convex, $f(\cdot;\mu_t)$ converges globally to f^* at a sublinear rate in terms of T up to an error of $\mathcal{O}(\alpha^{-1})$.

Under Assumption 8, we can also quantify the optimality gap between $J(f_t)$ and $J(f^*)$, in terms of the minimal distance $\inf_{t\in[0,T]} J(f_t) - J(f^*)$. The following theorem characterize the global convergence of $J(f_t)$ to $J(f^*)$,

Theorem 9 (Global Convergence to Primal Solution) *Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (24) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 =$*

$\mathcal{N}(0, I_D)$. *Under Assumptions 1, 2, 3 and 8, let* $f_t = f(\cdot; \mu_t)$, *it holds that*

$$\inf_{t\in[0,T]} J(f_t) - J(f^*) \leq \mathcal{O}(T^{-1/2} + \alpha^{-1/2}),$$

where f^* *is the global minimizer of the objective function defined in* (4).

Proof See §B.3 for a detailed proof. ■

Theorem 9 proves that under the additional strong convexity assumption on the regularizer $\Psi(X, Z; f)$, the primal objective $J(f_t)$ as is defined in (4) decays to zero at rate of $T^{-1/2}$ in terms of time horizon T, up to an error of $\mathcal{O}(\alpha^{-1/2})$. Here we use f^* to denote the global minimizer instead of the saddle point. However, this will not create any confusion since for each f^* global minimizer of the primal objective (4), we can find $g^* \in \mathcal{F}$ such that (f^*, g^*) is a saddle point of (10).

5. Applications

In this section, we present the applications of Theorem 7 and Theorem 9 to several special cases of the functional conditional moment equation, such as the problem of policy evaluation, instrumental variables regression, and asset pricing. In Section 2.2, we already discussed why these problems are special cases of functional conditional moment equations, Theorem 7 and Theorem 9 are potentially feasible to apply. We will recall the problem settings and examine the technical assumptions for these cases.

5.1 Application 1: Policy Evaluation

Let $\mathcal{D}$ denote the joint distribution of the state-action tuple (S, A, S') under policy π. In this scenario, the endogenous variable $X = S'$ is the next state while the exogenous variable $Z = S$ is the current state. Therefore, $\mathcal{X} = \mathcal{S}$, $\mathcal{Z} = \mathcal{S}$ and $\mathcal{W} = \mathcal{S}$. We attempt to estimate the value function V, which is defined on $\mathcal{W} = \mathcal{S} \to \mathbb{R}$. The functional Φ and regularizer Ψ adopted in this case are,

$$\Phi(s', s; f) = r + \gamma \cdot f(s') - f(s), \quad \Psi(s', s; f) = f(s')^2.$$

Here, the regularizer we adopt is a L^2 regularizer that penalizes the squared value of the estimator evaluated at the next state s'. With these specific choices of functional Φ and regularizer Ψ, the SGDA algorithm recovers the Gradient Temporal Difference Learning (GTD) algorithm (Wai et al., 2020). Therefore, applying our general framework to the problem of policy evaluation contributes to the reinforcement learning literature by providing an analysis of the neural GTD algorithm in the mean-field regime. Before presenting the theoretical results, we first verify that Assumption 3 and Assumption 8 hold.

Verify item (i) of Assumption 3. For item (i) of Assumption 3, it's reasonable to assume that $\|(x, z)\| \leq 1$ since we can always re-scale the state space without changing the nature of the problem, the compactness assumption is inherently satisfied.

Verify item (ii) of Assumption 3. For item (ii) of Assumption 3, we first compute the variation of the functional Φ and Ψ,

$$\frac{\delta\Phi(s', s; f)}{\delta f}(w') = \gamma\delta_{s'}(w') - \delta_s(w'), \quad \frac{\delta\Psi(s', s; f)}{\delta f}(w') = 2f(s')\delta_{s'}(w').$$

Therefore, the desired integrability conditions hold since

$$\int_{\mathcal{W}} \Big| \frac{\delta \Phi(s', s; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq \gamma + 1, \quad \int_{\mathcal{W}} \Big| \frac{\delta \Psi(s', s; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq 2 \cdot |f(s')|. \tag{29}$$

Verify item (iii) of Assumption 3. For item (iii) of Assumption 3, we choose $w = s'$, $C_\Psi = 2$. The desired condition holds due to (29).

Verify item (iv) of Assumption 3. For item (iv) of Assumption 3, we first compute the variations of $\mathcal{L}(f, g)$ in explicit forms,

$$\frac{\delta \mathcal{L}(f,g)}{\delta f}(w') = \mathbb{E}_{S|S'} \Big[\gamma \cdot g(S) \,\Big|\, S' = w' \Big] - g(w')\rho_s(w') + 2\lambda \cdot f(w')\rho_{S'}(w'), \quad \forall w' \in \mathcal{S},$$
$$\frac{\delta \mathcal{L}(f,g)}{\delta g}(z') = \mathbb{E}_{S'|S} \Big[r + \gamma \cdot f(s') \,\Big|\, S = z' \Big] - f(z') - g(z')\rho_S(z'), \quad \forall z' \in \mathcal{S},$$

where ρ_S, $\rho_{S'}$ denotes the density of the marginal distribution of $\mathcal{D}$ with respect to the current state S and next state S' respectively. Due to the item (i) of Assumption 3, the variations of $\mathcal{L}$ with respect to f and g are both continuous since the density of the conditional transition $S' \,\big|\, S$ and $S \,\big|\, S'$ are both smooth, and the functions f, g are also continuous by construction. Therefore, item (iv) is satisfied.

Verify Assumption 8. For Assumption 8, we choose $c_\Psi = 1$ and $w = s'$. The desired condition holds by definition of our choice of regularizer $\Psi(s', s; f) = f(s')^2$.

We have checked that the technical Assumption 3 and Assumption 8 hold for the case of policy evaluation. Assumption 3 allows us to apply Theorem 7. This implies the global convergence of the estimated value function to the minimizer of the primal objective (4) applied in this case. The convergence is quantified in a weighted L^2 distance. Additionally, Assumption 8 enables us to apply Theorem 9 and further characterize this convergence in terms of the optimality gap between the primal objective values. We summarize the conclusions in the following corollary.

Corollary 10 (Global Convergence of Mean-field Neural Nets in Policy Evaluation) *Let f^* be the minimizer of primal objective $J(f)$ defined in (2.1) with $\Phi(S', S; f) = r + \gamma \cdot f(S') - f(S)$, $\mathcal{R}(f) = \mathbb{E}_{\mathcal{D}}[f(S')^2]$. Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (24) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 = \mathcal{N}(0, I_D)$. Under Assumption 1, 2, 3, and 8, it holds that*

$$\inf_{t \in [0,T]} \mathbb{E}_{\mathcal{D}} \Big[\big(f(S'; \mu_t) - f^*(S') \big)^2 \Big] \leq \mathcal{O}(T^{-1} + \alpha^{-1}),$$
$$\inf_{t \in [0,T]} J(f(\cdot; \mu_t)) - J(f^*(\cdot)) \leq \mathcal{O}(T^{-1/2} + \alpha^{-1/2}).$$

Proof We apply Theorem 7 and Theorem 9 to the setting of policy evaluation. As we have examined above, the Assumption 1, 2, 3, and 8 are all satisfied. Thus, by Theorem 7 and Theorem 9, the desired results hold directly. ■

Corollary 10 proves that in the setting of policy evaluation, the L^2 distance between the mean-field neural network $f(\cdot; \mu_t)$ at time t and the global minimizer f^* decays to

zero at a sub-linear rate, up to an error of order $\mathcal{O}(\alpha^{-1})$. Moreover, the optimality gap $\inf_{t\in[0,T]} J(f(\cdot;\mu_t)) - J(f^*(\cdot))$ in terms of primal objective values decays to zero at the rate of $\mathcal{O}(T^{-1/2})$, up to an error $\mathcal{O}(\alpha^{-1/2})$ caused by overparameterization. Corollary 10 allows us to efficiently and globally solve the policy evaluation problem using overparameterized two-layer neural networks. We also remark that in this scenario, the primal objective $J(f)$ is known in the reinforcement learning literature as the regularized mean-squared Bellman error (MSBE). As we have noted before, in the context of policy evaluation, applying the SGDA algorithm to neural network function classes is equivalent to applying the neural GTD algorithm. Therefore, Corollary 10 states that, in the mean-field regime, the neural GTD algorithm converges globally to the minimizer at a sublinear rate up to an additional overparameterization error $\mathcal{O}(\alpha^{-1})$. The neural GTD algorithm also reduces regularized MSBE at the rate of $\mathcal{O}(T^{-1/2})$ up to an additional overparameterization error $\mathcal{O}(\alpha^{-1/2})$. Moreover, the global convergence of mean-field neural networks also implies the global convergence of the discrete dynamics in (20) due to the proximity between the discrete dynamics and continuous dynamics, which is proved in Proposition 4.

5.2 Application 2: Nonparametric Instrumental Variables Regression

Let $\mathcal{D}$ denote the joint distribution of the endogenous variable X, the exogenous variable Z, and the observed outcome Y. In this scenario, the endogenous variable is defined in space $\mathcal{X}$, the exogenous variable is defined in space $\mathcal{Z}$, and $\mathcal{W} = \mathcal{X}$. We attempt to estimate the model function f_0, which is defined on $\mathcal{W} = \mathcal{X} \to \mathbb{R}$. The functional Φ and regularizer Ψ adopted in this case are,

$$\Phi(x, z; f) = y - f(x), \quad \Psi(x, z; f) = f(x)^2.$$

Here, the regularizer we adopt is an L^2 regularizer that penalizes the squared value of the model function's estimator evaluated at the endogenous variable x. We examine Assumption 3 and Assumption 8 in order to apply results from Section 4.3.

Verify item (i) of Assumption 3. For item (i) of Assumption 3, the NPIV problem with compact data space captures a large class of important applications, the scenarios considered are still general while imposing this assumption.

Verify item (ii) of Assumption 3. For item (ii) of Assumption 3, we first compute the variation of the functional Φ and Ψ,

$$\frac{\delta\Phi(x, z; f)}{\delta f}(w') = -\delta_x(w'), \quad \frac{\delta\Psi(x, z; f)}{\delta f}(w') = 2f(x)\delta_x(w').$$

Therefore, the desired integrability conditions hold since

$$\int_{\mathcal{W}} \Big|\frac{\delta\Phi(x, z; f)}{\delta f}(w')\Big| \mathrm{d}w' \le 1, \quad \int_{\mathcal{W}} \Big|\frac{\delta\Psi(x, z; f)}{\delta f}(w')\Big| \mathrm{d}w' \le 2 \cdot |f(x)|. \tag{30}$$

Verify item (iii) of Assumption 3. For item (iii) of Assumption 3, we choose $w = x$, $C_\Psi = 2$. The desired condition holds due to (30).

Verify item (iv) of Assumption 3. For item (iv) of Assumption 3, we first compute the variations of $\mathcal{L}(f,g)$ in explicit forms,

$$\frac{\delta \mathcal{L}(f,g)}{\delta f}(w') = \mathbb{E}_{Z|X}\Big[-g(Z) \,\Big|\, X = w' \Big] + 2\lambda \cdot f(w')\rho_X(w'), \quad \forall w' \in \mathcal{X},$$
$$\frac{\delta \mathcal{L}(f,g)}{\delta g}(z') = \mathbb{E}_{X|Z}\Big[Y - f(X) \,\Big|\, Z = z' \Big] - g(z')\rho_Z(z'), \quad \forall z' \in \mathcal{Z},$$

where ρ_X, ρ_Z denotes the density of the marginal distribution of $\mathcal{D}$ with respect to the endogenous variable X and the exogenous variable Z respectively. Due to the item (i) of Assumption 3, the variations of $\mathcal{L}$ with respect to f and g are both continuous since the density of the conditional transition $Z \,\big|\, X$ and $X \,\big|\, Z$ are both smooth, and the functions f, g are also continuous by construction. Therefore, item (iv) is satisfied.

Verify Assumption 8. For Assumption 8, we choose $c_\Psi = 1$ and $w = s'$. The desired condition holds by definition of our choice of regularizer $\Psi(x, z; f) = f(x)^2$.

We have verified that the technical Assumptions 3 and 8 hold in the case of nonparametric instrumental variables regression. Theorem 7 can be applied in this case due to the establishment of Assumption 3. This implies that the estimated model function globally converges to the minimizer of the primal objective. The convergence is quantified in a weighted L^2 distance. The choice of quadratic regularizer implies the establishment of Assumption 8, which further enables us to apply Theorem 9 and characterize the convergence in terms of primal objective value. We summarize the conclusions in the following corollary.

Corollary 11 (Global Convergence of Mean-field Neural Nets in NPIV) *Let f^* be the minimizer of primal objective $J(f)$ defined in (2.1) with $\Phi(X, Z; f) = Y - f(X)$, $\mathcal{R}(f) = \mathbb{E}_{\mathcal{D}}[f(X)^2]$. Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (24) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 = \mathcal{N}(0, I_D)$. Under Assumption 1, 2, 3, and 8, it holds that*

$$\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[(f(X;\mu_t) - f^*(X))^2\Big] \le \mathcal{O}(T^{-1} + \alpha^{-1}),$$
$$\inf_{t\in[0,T]} J(f(\cdot;\mu_t)) - J(f^*(\cdot)) \le \mathcal{O}(T^{-1/2} + \alpha^{-1/2}).$$

Proof We apply Theorem 7 and Theorem 9 to the setting of NPIV. As we have examined above, the Assumption 1, 2, 3, and 8 are all satisfied. Thus, by Theorem 7 and Theorem 9, the desired results hold directly. ■

Corollary 11 proves that in the setting of NPIV, the L^2 distance between the mean-field neural network $f(\cdot;\mu_t)$ at time t and the global minimizer f^* decays to zero at a sub-linear rate, up to an error of order $\mathcal{O}(\alpha^{-1})$. Moreover, the optimality gap $\inf_{t\in[0,T]} J(f(\cdot;\mu_t)) - J(f^*(\cdot))$ decays to zero at the rate of $\mathcal{O}(T^{-1/2})$, up to an error $\mathcal{O}(\alpha^{-1/2})$. Corollary 11 allows us to solve the NPIV problem globally using overparameterized two-layer neural networks. We also want to remark that when the true model function is linear in the input, we recover the setting of instrumental variables regression as an important special instance of NPIV. Therefore, Corollary 11 also implies that IV regression can be solved efficiently globally using overparameterized two-layer neural networks.

5.3 Application 3: Asset Pricing

Let $\mathcal{D}$ denote the joint distribution of the growth-return tuple $(c_t, \tilde{r}_{t+1}, c_{t+1})$. In this scenario, the exogenous variable $Z = c_t$ is the consumption growth at the current time t, and the endogenous variable $X = c_{t+1}$ is the consumption growth at the next time $t+1$. Therefore, $\mathcal{X} = \mathcal{Z} = \mathcal{C}$, $\mathcal{W} = \mathcal{C}$, where $\mathcal{C}$ is the space of consumption growth and is also a compact subset of $\mathbb{R}$. Here, we consider the scenario where the modified return $\tilde{r}_{t+1}$ is also bounded for all $t \geq 0$, i.e., $\|\tilde{r}_{t+1}\| \leq R$ for some $R > 0$. We attempt to estimate the function f_0, which is defined on $\mathcal{W} = \mathcal{S} \to \mathbb{R}$. The functional Φ and regularizer Ψ adopted in this case are,

$$\Phi(c_{t+1}, c_t; f) = \tilde{r}_{t+1} \cdot f(c_{t+1}) - f(c_t), \quad \Psi(c_{t+1}, c_t; f) = f(c_{t+1})^2.$$

Here, the regularizer we adopt is an L^2 regularizer that penalizes the squared value of the estimator evaluated at the next period's consumption growth, c_{t+1}. Before presenting the theoretical results, we first verify that Assumption 3 and Assumption 8 hold.

Verify item (i) of Assumption 3. For item (i) of Assumption 3, since we assume that the space of consumption growth $\mathcal{C}$ is a compact subset of $\mathbb{R}$, therefore there exists $C_1 > 0$ such that for all $t \geq 0$, $\|(c_{t+1}, c_t)\| \leq C_1$. Moreover, it is reasonable to assume that consumption growth is bounded, since the data often fluctuate within certain regimes in practice.

Verify item (ii) of Assumption 3. For item (ii) of Assumption 3, we first compute the variation of the functional Φ and Ψ,

$$\frac{\delta\Phi(c_{t+1}, c_t; f)}{\delta f}(w') = \tilde{r}_{t+1} \cdot \delta_{c_{t+1}}(w') - \delta_{c_t}(w'), \quad \frac{\delta\Psi(c_{t+1}, c_t; f)}{\delta f}(w') = 2f(c_{t+1}) \cdot \delta_{c_{t+1}}(w').$$

Therefore, the desired integrability condition holds since,

$$\int_{\mathcal{W}} \Big| \frac{\delta\Phi(c_{t+1}, c_t; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq R + 1, \quad \int_{\mathcal{W}} \Big| \frac{\delta\Psi(c_{t+1}, c_t; f)}{\delta f}(w') \Big| \mathrm{d}w' \leq 2 \cdot |f(c_{t+1})|. \quad (31)$$

Verify item (iii) of Assumption 3. For item (iii) of Assumption 3, we choose $w = \tilde{c}$, $C_\Psi = 2$. The desired property holds due to (31).

Verify item (iv) of Assumption 3. For item (iv) of Assumption (3), we first compute the variations of $\mathcal{L}(f, g)$ in explicit forms,

$$\frac{\delta\mathcal{L}(f,g)}{\delta f}(w') = \mathbb{E}_{c_t|c_{t+1}}\Big[\tilde{r}_{t+1} \cdot g(c_t) \,\Big|\, \tilde{c}_t = w'\Big] - g(w')\rho_{c_t}(w') + 2\lambda \cdot f(w')\rho_{c_{t+1}}(w'), \quad \forall w' \in \mathcal{S},$$
$$\frac{\delta\mathcal{L}(f,g)}{\delta g}(z') = \mathbb{E}_{c_{t+1}|c_t}\Big[\tilde{r}_{t+1} \cdot f(c_{t+1}) \,\Big|\, c_t = z'\Big] - f(z') - g(z')\rho_{c_t}(z'), \quad \forall z' \in \mathcal{S},$$

where $\rho_{c_t}, \rho_{c_{t+1}}$ denotes the density of the marginal distribution of $\mathcal{D}$ with respect to the current time consumption growth c_t and the next time consumption growth c_{t+1} respectively. The variations of $\mathcal{L}$ with respect to f and g are both continuous since the density of the conditional transition $c_{t+1} \,\big|\, c_t$ and $c_t \,\big|\, c_{t+1}$ are both smooth, and the function f, g are also continuous by construction. Therefore, item (iv) is satisfied.

Verify Assumption 8. For Assumption 8, we choose $c_\Psi = 1$ and $w = c_{t+1}$. The desired condition holds by definition of our choice of regularizer $\Psi(c_{t+1}, c_t; f) = f(c_{t+1})^2$.

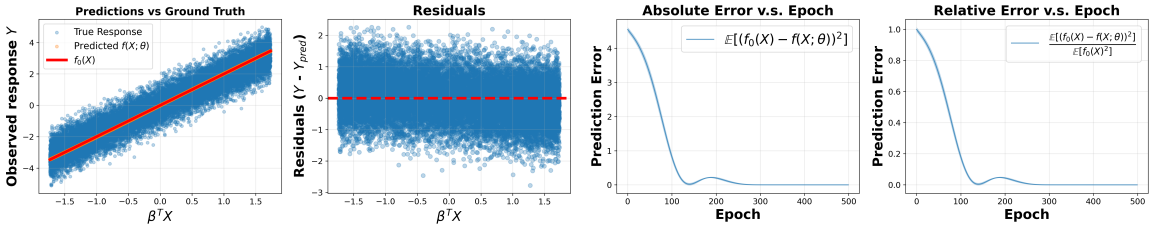

Figure 1: Results on NPIV with $f_0(X) = \alpha \cdot \beta^\top X$ with $\alpha = 2$, $\beta = \mathbf{1}$.

We have verified that the technical Assumptions 3 and 8 hold in the context of asset pricing with the CCAPM model. Theorem 7 can be applied in this case due to the establishment of Assumption 3. This implies that the estimated function globally converges to the minimizer of the primal objective. The convergence is quantified in a weighted L^2 distance. Since Assumption 8 holds, we can apply Theorem 9 and characterize the convergence in terms of primal objective value. We summarize the conclusions in the following corollary.

Corollary 12 (Global Convergence of Mean-field Neural Nets in Asset Pricing) *Let f^* be the minimizer of primal objective $J(f)$ defined in (2.1) with $\Phi(c_{t+1}, c_t; f) = \tilde{r}_{t+1} \cdot f(c_{t+1}) - f(c_t)$, $\mathcal{R}(f) = \mathbb{E}_{\mathcal{D}}[f(c_{t+1})^2]$. Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (24) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 = \mathcal{N}(0, I_D)$. Under Assumption 1, 2, 3, and 8, it holds that*

$$\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[(f(c_{t+1};\mu_t) - f^*(c_{t+1}))^2\Big] \le \mathcal{O}(T^{-1} + \alpha^{-1}),$$
$$\inf_{t\in[0,T]} J(f(\cdot;\mu_t)) - J(f^*(\cdot)) \le \mathcal{O}(T^{-1/2} + \alpha^{-1/2}).$$

Proof We apply Theorem 7 and Theorem 9 to the setting of asset pricing. As we have examined above, the Assumption 1, 2, 3, and 8 are all satisfied. Thus, by Theorem 7 and Theorem 9, the desired results hold directly. ■

Corollary 12 proves that in the setting of asset pricing, the L^2 distance between the mean-field neural network $f(\cdot;\mu_t)$ at time t and the global minimizer f^* decays to zero at a sub-linear rate, up to an error of order $\mathcal{O}(\alpha^{-1})$. Moreover, the optimality gap $\inf_{t\in[0,T]} J(f(\cdot;\mu_t)) - J(f^*(\cdot))$ decays to zero at the rate of $\mathcal{O}(T^{-1/2})$, up to an error $\mathcal{O}(\alpha^{-1/2})$. Corollary 12 allows us to solve the CCAPM model globally by estimating the nonparametric structural demand function with overparameterized two-layer neural networks. Since the return on investment is linked to the marginal utility of consumption through the CCAPM equation, we can price assets fairly by considering consumption risk and using marginal utility information.

6. Numerical Experiments

In this section, we numerically validate the effectiveness of our proposed algorithm using the nonparametric instrumental variable regression (NPIV) example in Subsection 5.2, with a variety of linear or nonlinear structural functions f_0 to be learned.

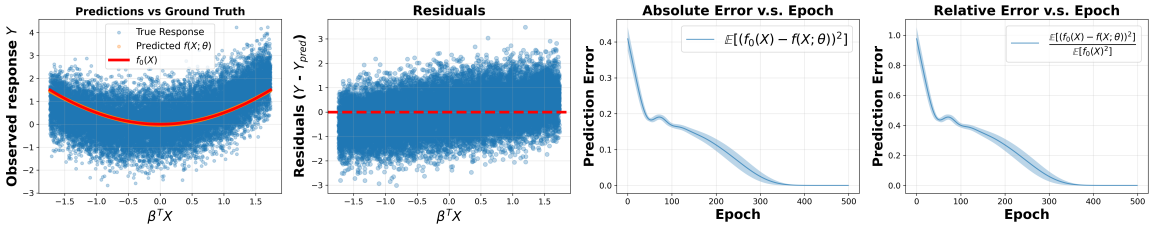

Figure 2: Results on NPIV with $f_0(X) = \alpha \cdot (\beta^\top X)^2$ with $\alpha = 0.5$, $\beta = \mathbf{1}$.

6.1 NPIV experiment setup

The statistical model we considered in the numerical experiment is described as,

$$\begin{aligned} Y &= f_0(X) + \Gamma_1 \varepsilon_{\text{conf}} + \varepsilon \\ X &= \frac{X'}{\|X'\|} \quad X' = \Gamma_2 Z + \varepsilon_{\text{conf}} \quad Z \sim \text{Uniform}(\mathbb{S}^{p-1}) \\ \varepsilon_{\text{conf}} &\sim \mathcal{N}(0, \sigma_{\text{conf}}^2 I_p) \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{32}$$

where $X \in \mathbb{S}^{p-1} \subset \mathbb{R}^p$ is the endogenous variable, $Z \in \mathbb{S}^{q-1} \subset \mathbb{R}^q$ is the instrumental (exogenous) variable, $Y \in \mathbb{R}$ is the observed response, $\Gamma_1 \in \mathbb{R}^{1\times p}$, $\Gamma_2 \in \mathbb{R}^{p\times q}$, $f : \mathbb{R}^p \to \mathbb{R}$ is the structural function to predict, and ε, $\varepsilon_{\text{conf}}$, Z are independent.
We let $p = 3$ and $q = 10$, resulting in the endogenous variable X lies on $\mathbb{R}^3$ while the exogenous variable Z lies on $\mathbb{R}^{10}$. For Γ_1 and Γ_2, we sample their value element-wise from Gaussian distributions and fix their values when generating the (X, Z, Y) data pair. Specifically, $\Gamma_1 \in \mathbb{R}^3$, with $(\Gamma_1)_i \sim \mathcal{N}(0, \frac{1}{9})$, and $\Gamma_2 \in \mathbb{R}^{10\times 3}$, with $(\Gamma_2)_{i,j} \sim \mathcal{N}(0, 1)$. We set $\sigma_{\text{conf}}^2 = 1$ and $\sigma^2 = 1/2$. To demonstrate the robustness of our approach, we consider the following three types of true structural functions f_0:

(1) $f_0(X) = \alpha \cdot \beta^\top X$,

(2) $f_0(X) = \alpha \cdot (\beta^\top X)^2$,

(3) $f_0(X) = \alpha_1 \beta_1^\top X + \alpha_2 \beta_2^\top X / \big(1 + \exp(-\beta_2^\top X)\big)$.

Note that in model (32), we have $\|X\|_2 = \|Z\| = 1$, satisfying the data compactness assumption in Assumption 3. It is also straightforward to verify that problem (32) satisfies the conditional moment equation for NPIV, which is $\mathbb{E}[Y - f(X)|Z] = 0$, due to the independence assumption between Z and $\varepsilon_{\text{conf}}$, ε. However, such a problem cannot be solved directly by regressing Y over X, due to the presence of $\varepsilon_{\text{conf}}$. This promotes the use of minimax optimization, with the formulation introduced in Subsection 5.2, specifically, with the functionals $\Phi(x, z; f) = y - f(x)$ and $\Psi(x, z; f) = f(x)^2$.

6.2 Experiment Results

For each of the above true structural functions f_0, we generate 20000 synthetic data points following the NPIV model (32), where we split it into a training set of 18000 data points and a test set of 2000 data points using a split ratio of 9 : 1.

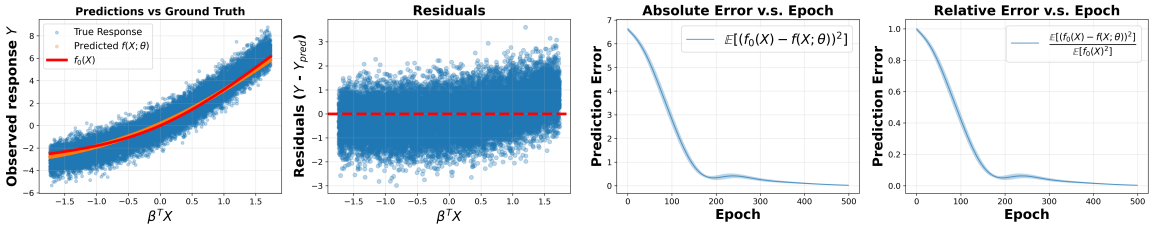

Figure 3: Results on NPIV with $f_0(X) = \alpha_1 \beta_1^\top X + \dfrac{\alpha_2 \beta_2^\top X}{1 + \exp(-\beta_2^\top X)}$, with $\alpha_1 = 0.1$, $\alpha_2 = 4$, $\beta_1 = \beta_2 = \mathbf{1}$.

For the training and across all experiments, we use $\lambda_f = 10^{-3}$ as the regularization strength, and parameterize both the primal function f and the dual function g with finite-width two-layer neural networks. Each network has a width of 500, which we have found to be sufficient for parameterizing a broad class of functions. We use a batch size of 1000, a learning rate of 10^{-4} to update both f and g, and run the optimization for 500 epochs. During training, we track the absolute and relative prediction errors of the learned function at epoch t with respect to the ground truth by computing their mean squared error over the test split of 2000 data points. For each experiment, we repeat 10 times to demonstrate the consistency of the results. All computations are performed on standard consumer-grade CPUs.

The results are shown in Figures 1–3 respectively for the three structural functions f_0. From the figures, it is evident that SGDA well recovers the unknown true structural function f_0. This demonstrates the effectiveness of SGDA for solving functional conditional moment equations via minimax optimization. Moreover, despite the theoretical requirement of two-layer infinite-width neural nets, in practice we use only two-layer neural networks with widths up to 500, which can be optimized quickly and efficiently without incurring expensive computational costs. These numerical findings further support the practicality and effectiveness of SGDA in solving difficult functional-space ill-posed inverse problems.

To investigate how the computational cost of Neural SGDA scales with the width of the neural networks, we conduct a separate ablation study in which we train neural networks with widths ranging from 10 to 1000 on the same NPIV dataset with the true structural function being $f_0(X) = 0.5 \cdot (\mathbf{1}^\top X)^2$, for a fixed number of 500 epochs. We record the time elapsed for each training run and the corresponding error against the neural network width used in the experiment. The reported error is the $\mathbb{E}[(f_0(X) - f(X;\theta)^2]$, where the expectation is approximated using the empirical measure of 2000 data points in the test split of the dataset, unseen during the training. The results are demonstrated in Figure 4. As the plot shows, although training time approximately scales linearly with network width, the overall computational cost remains affordable, even for wide neural networks. It takes approximately 140 seconds to fully optimize a network with a width of 1000 on standard consumer-grade GPUs, highlighting the algorithm's computational efficiency. From the plot, we can also see that the error plateaus once the network width exceeds 500. This is also consistent with our choice of neural network width in the previous experiments.

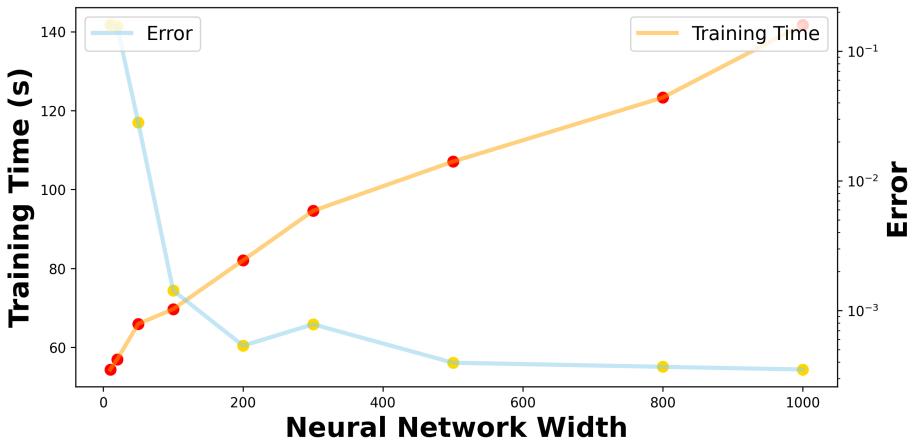

Figure 4: Growth trend of training time and decay rate of error $\mathbb{E}[(f_0(X) - f(X;\theta)^2]$ against the increase of neural network width on NPIV with $f_0(X) = 0.5 \cdot (\mathbf{1}^\top X)^2$.

7. Conclusion

In this paper, we focus on the minimax optimization problem derived from solving functional conditional moment equations using overparameterized two-layer neural networks. For such a problem, we first prove that the stochastic gradient descent-ascent algorithm converges to a mean-field limit as the stepsize goes to zero and the network width goes to infinity. In this mean-field limit, the optimization dynamics is characterized by a Wasserstein gradient flow in the space of probability distributions. We further establish the global convergence of the Wasserstein gradient flow and prove that the feature representation induced by the neural networks may move a considerable distance from its initial value. We further apply our general results to policy evaluation with high-dimensional state space, nonparametric instrumental variables regression with high-dimensional endogenous and exogenous variables, and asset pricing with a nonparametric structural demand function. Our analysis opens avenues for studying functional minimax optimization problems with more complicated objectives, such as nonlinear functional conditional moment equations. We leave the study of the algorithm's convergence properties in this general setting to future research. This setting includes nonparametric quantile instrumental variables regression as a leading and important application.

Appendix A. Notations

Category	Notation	Location	Description
Basic	$\mathcal{L}, J$	Section 2.1	Minimax and primal obj.
	X, x		Endogenous variables
	Z, z		Exogenous variables
	W, w		Input variables of f
	f, g		primal & dual functions
	f^*, g^*		optimal primal & dual functions
	Φ		Moment equation functional
	Ψ		Regularization functional
	$\frac{\delta\Phi}{\delta f}, \frac{\delta\Psi}{\delta f}$		First variation of functionals
Algorithms	$\boldsymbol{\theta}$, $\boldsymbol{\omega}$	Section 3	network parameters
	$f(\cdot;\boldsymbol{\theta}), g(\cdot;\boldsymbol{\omega})$		finite-width neural network
	$\boldsymbol{\theta}_k$, $\boldsymbol{\omega}_k$		SGDA iterates at step k
	ϵ, η		discrete/continuous time step size scale
	μ, ν, ρ		mean-field dist.
	μ^*, ν^*, ρ^*		optimal mean-field dist.
	$f(\cdot;\mu), g(\cdot;\nu)$		infinite-width neural network
	v^f, v^g		vector field of WGF
	μ_t, ν_t, ρ_t		WGF iterates at time t
Assumptions	σ, ψ	Equation (25)	neuron functions
	B_0, B_1, B_2	Assumption 1	Neuron func. regularity constants
	C_1, C_2, C_Ψ, c_Ψ	Assumption 3 & 8	Data & functional regularity constants
Convergence	α	Equation (22)	Over-paramterization scaling factor
	$\{\beta_s\}_{s\in[0,1]}$	Equation (49)	Wasserstein geodesics between ρ_t,ρ^*
	u_s	Equation (49)	Velocity field associated with β_s
	t^*, t_*	Equation (66), (67)	First hitting times
	$\mathcal{V}(\theta, \omega; \cdot)$	Equation (50)	WGF potential
Propagation of Chaos	$\theta_i(k), \omega_i(k), \hat{V}_k^f, \hat{V}_k^g$	Equation. (82)	Stochastic GDA
	$\breve{\theta}_i(k), \breve{\omega}_i(k), \hat{v}^f, \hat{v}^g$	Equation (83)	Population GDA
	$\tilde{\theta}_i(t), \tilde{\omega}_i(t), \hat{v}^f, \hat{v}^g$	Equation (84)	Continuous-time Population GDA
	$\bar{\theta}_i(t), \bar{\omega}_i(t), v^f, v^g$	Equation (85)	Ideal Particles

Table 1: A summary of used notations with their location and description

Appendix B. Proof of Main Results

In this section, we provide proofs for the main theorems and technical lemmas in our work.

B.1 Proof of Lemma 6

Proof of (i). The proof for Claim (i) will be two-stage. First, we will show that if a function pair (f^*, g^*) is a stationary point for $\mathcal{L}(f, g)$ with respect to (f, g), then it's a saddle point for the same objective as well. Then we will show that the distribution pair (μ^*, ν^*) being a stationary point of $\mathcal{L}(\mu, \nu)$ implies that the corresponding (f^*, g^*) is a stationary point for $\mathcal{L}(f, g)$, which concludes the claim. We will start with the first part. We define the following functional $\mathcal{L}_1$ and $\mathcal{L}_2$,

$$\mathcal{L}_1(f, g) = \mathbb{E}_{\mathcal{D}}\Big[g(Z) \cdot \Phi(X, Z; f)\Big], \quad \mathcal{L}_2(f, g) = \mathbb{E}_{\mathcal{D}}\Big[-1/2 \cdot g(Z)^2 + \lambda\Psi(X, Z; f)\Big].$$

We see that the minimax objective in (10) is indeed the sum of such two functionals,

$$\mathcal{L}(f,g) = \mathcal{L}_1(f,g) + \mathcal{L}_2(f,g).$$

For any function pair (f,g), we can verify that the following chain of equalities holds,

$$\max_{g'}\ \mathcal{L}(f,g') - \min_{f'}\ \mathcal{L}(f',g) = \max_{g'}\ \big(\mathcal{L}(f,g') - \mathcal{L}(f,g)\big) + \max_{f'}\ \big(\mathcal{L}(f,g) - \mathcal{L}(f',g)\big). \quad (33)$$

We considered the function space $L^2(\mathcal{W})$ and $L^2(\mathcal{Z})$ equipped with inner product $\langle\cdot,\cdot\rangle_{L^2}$, which are also Hilbert spaces. Since $\mathcal{X}\times\mathcal{Z}$ are compact, continuous function f and g parameterized in the form of (22) are square-integrable, thus naturally belong to $L^2(\mathcal{W})$ and $L^2(\mathcal{Z})$.

For a fixed f, $\mathcal{L}_1(f,g)$ is a continuous linear functional in g defined on $L^2(\mathcal{Z})$. Thus, there exists function h_f in $L_2(\mathcal{Z})$ such that $\mathcal{L}_1(f,g) = \big\langle h_f, g\big\rangle_{L^2}$. Similarly, for a fixed g, $\mathcal{L}_1(f,g)$ is a continuous linear functional in f, thus there exists function h_g in $L_2(\mathcal{W})$ such that $\mathcal{L}_1(f,g) = \big\langle h_g, f\big\rangle_{L^2}$. In fact, h_f and h_g matches the variation of $\mathcal{L}_1$ with respect to g and f.

$$h_f = \frac{\delta\mathcal{L}_1(f,g)}{\delta g}, \quad h_f = \frac{\delta\mathcal{L}_2(f,g)}{\delta f}.$$

Since $\mathcal{L}_2$ is a concave functional with respect to g, we apply Jensen's inequality and it holds that,

$$\begin{aligned}\mathcal{L}(f,g') - \mathcal{L}(f,g) &= \mathcal{L}_1\big(f,g'\big) - \mathcal{L}_1\big(f,g\big) + \mathcal{L}_2\big(f,g'\big) - \mathcal{L}_2\big(f,g\big)\\ &\le \Big\langle \frac{\delta\mathcal{L}_1(f,g)}{\delta g}, g'-g\Big\rangle_{L^2} + \Big\langle \frac{\delta\mathcal{L}_2(f,g)}{g}, g'-g\Big\rangle_{L^2}\\ &= \Big\langle \frac{\delta\mathcal{L}(f,g)}{\delta g}, g'-g\Big\rangle_{L^2}. \end{aligned} \quad (34)$$

Follow a similar reasoning, using the fact that $\mathcal{L}_1$ is a linear functional with respect to f and $\mathcal{L}_2$ is a convex functional with respect to f, it holds that

$$\mathcal{L}(f,g) - \mathcal{L}(f',g) \le \Big\langle \frac{\delta\mathcal{L}(f,g)}{\delta f}, f-f'\Big\rangle_{L^2}. \quad (35)$$

Plugging (34) and (35) into (33), we re-write the minimax expression in (33) using the variation of $\mathcal{L}(f,g)$, the following inequality holds,

$$\max_{g'}\ \mathcal{L}(f,g') - \min_{f'}\ \mathcal{L}(f',g) \le \max_{f',g'} \Big\langle \frac{\delta\mathcal{L}(f,g)}{\delta g}, g'-g\Big\rangle_{L^2} + \Big\langle \frac{\delta\mathcal{L}(f,g)}{\delta f}, f-f'\Big\rangle_{L^2}. \quad (36)$$

Thus, if (f^*,g^*) is the stationary point, i.e.,

$$\frac{\delta\mathcal{L}(f^*,g^*)}{\delta f} = \frac{\delta\mathcal{L}(f^*,g^*)}{\delta g} = 0, \quad \text{a.s.}, \quad (37)$$

then (36) suggests that for such stationary point (f^*,g^*), for any function pair (f',g'), the following inequality holds,

$$\max_{g'}\ \mathcal{L}(f^*,g') - \min_{f'}\ \mathcal{L}(f',g^*) \le 0. \quad (38)$$

Equation (38) proves that (f^*, g^*) is a saddle point for the minimx objective $\mathcal{L}(f, g)$. Therefore, the stationarity of (f^*, g^*) implies that it's a saddle point for objective $\mathcal{L}(f, g)$.

Now, we proceed to show the second stage of the proof. We now show that if (μ^*, ν^*) is the stationary point of $\mathcal{L}$, i.e., $v^f(\cdot; \mu^*, \nu^*) = v^g(\cdot; \mu^*, \nu^*) = 0$, the corresponding function pair $(f(\cdot; \mu^*), g(\cdot; \nu^*))$ is the stationary point of $\mathcal{L}(f, g)$ with respect to (f, g). We recall that the correspondence between (μ^*, ν^*) and $(f(\cdot; \mu^*), g(\cdot; \nu^*))$ is through (22). Let (μ^*, ν^*) be a stationary point of (10), that is

$$\nabla_\theta \frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \mu}(\theta) = \nabla_\omega \frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \nu}(\omega) = 0, \quad \forall \theta, \omega \in \mathbb{R}^D \tag{39}$$

We can also compute the variation of $\mathcal{L}(\mu, \nu)$ explicitly.

$$\frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \mu}(\theta) = \mathbb{E}_{\mathcal{D}}\Big[\alpha\Big\langle g(Z; \nu^*) \cdot \frac{\delta \Phi(X, Z; f(\cdot; \mu^*))}{\delta f} + \lambda \cdot \frac{\delta \Psi(X, Z; f(\cdot; \mu^*))}{\delta f}, \phi(\cdot; \theta)\Big\rangle_{L^2}\Big],$$
$$\frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \nu}(\omega) = \mathbb{E}_{\mathcal{D}}\Big[\alpha\big(\Phi(X, Z; f(\cdot, \mu^*)) - g(Z; \nu^*)\big) \cdot \psi(Z; \omega)\Big].$$

By the oddness of b in Assumption 1, we have that $\phi(\cdot; \mathbf{0}) = 0$, This implies that the variation of $\mathcal{L}(\mu^*, \nu^*)$ with respect to μ and ν are 0 when $\theta = \omega = \mathbf{0}$, i.e.,

$$\frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \mu}(\mathbf{0}) = \frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \nu}(\mathbf{0}) = 0.$$

Combined with (39), we deduced that

$$\frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \mu}(\theta) = \frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \nu}(w) = 0 \quad \forall \theta, \omega \in \mathbb{R}^D.$$

Note that we can expand the variation of $\mathcal{L}$ with respect to μ,

$$\alpha\Big\langle \frac{\delta \mathcal{L}(f(\cdot; \mu^*), g(\cdot; \nu^*))}{\delta f}, \phi(\cdot; \theta)\Big\rangle_{L^2} = \frac{\delta \mathcal{L}(\mu^*, \nu^*)}{\delta \mu}(\theta) = 0. \tag{40}$$

By the universal function approximation theorem (Lemma 23), since $\frac{\mathcal{L}(f(\cdot;\mu^*), g(\cdot;\nu^*))}{\delta f}$ is in $\mathscr{C}(\mathcal{W})$ as is assumed in item (iv) of Assumption 3, there exists $\{\phi_n\}_{n=1}^\infty \in \mathcal{G}(\phi)$ such that $\phi_n \to \frac{\mathcal{L}(f(\cdot;\mu^*), g(\cdot;\nu^*))}{\delta f}$ uniformly. Here, $\mathcal{G}(\phi)$ denotes the space of functions that are linearly spanned by $\phi(\cdot, \theta)$ By (40), it holds that

$$\Big\langle \frac{\mathcal{L}(f(\cdot; \mu^*), g(\cdot; \nu^*))}{\delta f}(\cdot), \phi_n(\cdot)\Big\rangle_{L^2} = 0. \tag{41}$$

Following a similar strategy, we can show that there exists $\{\psi_n\}_{n=1}^\infty \in \mathcal{G}(\psi)$ such that $\psi_n \to \frac{\mathcal{L}(f(\cdot;\mu^*), g(\cdot;\nu^*))}{\delta g}$, where for each ψ_n, it holds that

$$\Big\langle \frac{\mathcal{L}(f(\cdot; \mu^*), g(\cdot; \nu^*))}{\delta g}(\cdot), \psi_n(\cdot)\Big\rangle_{L^2} = 0. \tag{42}$$

We take the limit of (41) and (42) by passing $n \to \infty$ and conclude,

$$\frac{\delta \mathcal{L}(f(\cdot;\mu^*), g(\cdot;\nu^*))}{\delta f} = 0, \quad \frac{\mathcal{L}(f(\cdot;\mu^*), g(\cdot;\nu^*))}{\delta g}(\cdot) = 0. \quad \text{a.s.} \tag{43}$$

Equation (43) proves that if (μ^*, ν^*) is a stationary point of the Wasserstein gradient flow, then the associated function pair $(f(\cdot;\mu^*, g(\cdot;\nu^*)))$ is a stationary point of the minimax objective $\mathcal{L}(f, g)$, which matches the conditions we conclude in (37). Therefore, we prove that $(f(\cdot;\mu^*), g(\cdot;\nu^*))$ is a saddle point of the minimax objective $\mathcal{L}(f, g)$. We complete the proof of item (i).

Proof of (ii). We now show that there exists good solution pair (μ^*, ν^*) that is both optimal as well as close to initialization (μ_0, ν_0) in Wasserstein distance. By Assumption 2, there exists distribution $\mu^\dagger, \nu^\dagger \in \mathscr{P}_2(\mathbb{R}^D)$ such that the optimal solution to the optimization problem (10)(f^*, g^*) satisfies the following,

$$f^*(w) = \int \phi(w;\theta) \mathrm{d}\mu^\dagger(\theta), g^*(z) = \int \psi(z;\omega) \mathrm{d}\nu^\dagger(\omega), \quad \forall w \in \mathcal{W}, z \in \mathcal{Z}$$

Recall that $\alpha > 0$ is the scaling parameter in neural network parameterization. We can construct (μ^*, ν^*) using a convex combination of $(\mu^\dagger, \nu^\dagger)$ and the initialization (μ_0, ν_0),

$$\mu^*(\theta) = \alpha^{-1}\mu^\dagger(\theta) + (1 - \alpha^{-1})\mu_0(\theta), \quad \nu^*(w) = \alpha^{-1}\nu^\dagger(\omega) + (1 - \alpha^{-1})\nu_0(\omega). \tag{44}$$

We claim that (μ^*, ν^*) constructed in (44) satisfies all the desired requirements. Since μ_0, ν_0 are standard Gaussian distribution, the integration of $\phi(\cdot;\theta)$ with respect to μ_0 and $\psi(\cdot;\omega)$ with respect to ν_0 are identically 0 due to oddness of neuron functions,

$$\int_{\mathcal{W}} \phi(w;\theta) \mathrm{d}\mu_0(\theta) = 0, \quad \int_{\mathcal{Z}} \psi(z;\omega) \mathrm{d}\nu_0(\omega) = 0. \quad \forall w \in \mathcal{W}, z \in \mathcal{Z}$$

Thus, the expressions for (f^*, g^*) are simplified to

$$f^*(w) = \alpha \int \phi(w;\theta) \mathrm{d}\mu^*(\theta), \quad g^*(z) = \alpha \int \psi(z;\omega) \mathrm{d}\nu^*(\omega).$$

By Talagrand's inequality (Lemma 27), the following chain of inequalities holds,

$$\begin{aligned}
\mathcal{W}_2(\mu_0, \mu^*)^2 &\le 2D_{\mathrm{KL}}(\mu^* \,\|\, \mu_0) \le D_{\chi^2}(\mu^* \,\|\, \mu_0) \\
&= \int \left(\frac{\mu^*(\theta)}{\mu_0(\theta)} - 1\right)^2 \mathrm{d}\mu_0(\theta) = \int \left(\frac{(1-\alpha^{-1}) \cdot \mu_0(\theta) + \alpha^{-1} \cdot \mu^\dagger(\theta)}{\mu_0(\theta)} - 1\right)^2 \mathrm{d}\mu_0(\theta) \\
&= \alpha^{-2} D_{\chi^2}(\mu^\dagger \,\|\, \mu_0).
\end{aligned} \tag{45}$$

A similar bound on $\mathcal{W}_2(\nu_0, \nu^*)$ also applies,

$$\mathcal{W}_2(\nu_0, \nu^*)^2 \le \alpha^{-2} D_{\chi^2}(\nu^\dagger \,\|\, \nu_0). \tag{46}$$

Let $\bar{D} = \max\{D_{\chi^2}(\mu^\dagger \,\|\, \mu_0)^{1/2}, D_{\chi^2}(\nu^\dagger \,\|\, \nu_0)^{1/2}\}$, we conclude the proof of item (ii).

B.2 Proof of Theorem 7

By Lemma 6, there exists distribution (μ^*, ν^*) that is a stationary point of Wasserstein gradient flow (24) and simultaneously satisfying the distance bound in item (ii) of Lemma 6. For such (μ^*, ν^*), we denote $\rho^*(\theta, \omega) = \mu^*(\theta)\nu^*(\omega)$ as their product measure. Moreover, for any distribution pair (μ, ν), we use $\rho(\theta, \omega) = \mu(\theta)\nu(\omega)$ as their product measure for simplicity. To rewrite the Wasserstein gradient flow for (μ, ν) into the flow for ρ, we define vector the stacked vector field v as,

$$v(\theta, \omega; \mu, \nu) = \big(v^f(\theta; \mu, \nu), v^g(\omega; \mu, \nu)\big). \tag{47}$$

Following from Lemma 24, (45), and (46), it holds that $\mathcal{W}_2(\rho^*, \rho_0) \le \alpha^{-1}\bar{D}$, where $\bar{D}$ is defined in Lemma 6. Note that

$$f(w; \mu) = \alpha \int \phi(w; \theta)\mu(\theta)\mathrm{d}\theta = \alpha \int \phi(w; \theta)\rho(\theta, \omega)\mathrm{d}(\theta, \omega), \quad \forall w \in \mathcal{W},$$
$$g(z; \nu) = \alpha \int \psi(z; \omega)\nu(\omega)\mathrm{d}\omega = \alpha \int \psi(z; \omega)\rho(\theta, \omega)\mathrm{d}(\theta, \omega), \quad \forall z \in \mathcal{Z}.$$

Thus, we overload the notation to write $f(\cdot; \rho) = f(\cdot; \mu)$ and $g(\cdot; \rho) = g(\cdot; \nu)$ for $\rho \in \mathscr{P}_2(\mathbb{R}^D \times \mathbb{R}^D)$. By writing $\rho_t = (\mu_t, \nu_t)$, the update in (24) takes the following form

$$\partial_t \rho_t(\theta, \omega) = -\mathrm{div}\big(\rho_t(\theta, \omega)v(\theta, \omega; \rho_t)\big), \quad \rho_0 = (\mu_0, \nu_0).$$

Before we prove Theorem 7, we first show the following important technical lemma.

Lemma 13 *We assume $\mathcal{W}_2(\rho_t, \rho^*) \le 2\mathcal{W}_2(\rho_0, \rho^*)$. Under Assumptions 1, 2, 3, it holds that*

$$\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t, \rho^*)^2}{\mathrm{d}t} \le -\eta \cdot \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi\big(X, Z; f(\cdot; \mu_t) - f^*(\cdot)\big) + \big(g(Z; \nu_t) - g^*(Z)\big)^2\Big] + C_* \cdot \eta\alpha^{-1}. \tag{48}$$

where $C_* > 0$ is a constant depending on B_0, B_1, B_2, λ, and $\bar{D}$

Proof Let $\{\beta_s\}_{s\in[0,1]}$ be the geodesic connecting ρ_t and ρ^* with $\beta_0 = \rho_t$ and $\beta_1 = \rho^*$. Let u be the corresponding veclocity field such that $\partial_s\beta_s = -\mathrm{div}(\beta_s u_s)$. By the first variation formula of Wasserstein distance in Lemma 25, it holds that

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t, \rho^*)^2}{\mathrm{d}t} &= -\eta\big\langle v(\cdot; \rho_t), u_0\big\rangle_{\rho_t} = -\eta\big\langle v(\cdot; \rho^*), u_1\big\rangle_{\rho^*} + \eta\int_0^1 \partial_s\big\langle v(\cdot; \beta_s), u_s\big\rangle_{\beta_s}\mathrm{d}s \qquad (49)\\ &= \eta\underbrace{\int_0^1 \big\langle \partial_s v(\cdot; \beta_s), u_s\big\rangle_{\beta_s}\mathrm{d}s}_{\text{(i)}} + \eta\underbrace{\int_0^1\int \big\langle v(\theta, \omega; \beta_s), \partial_s(u_s(\theta, w)\beta_s(\theta, \omega))\big\rangle\mathrm{d}(\theta, \omega)\mathrm{d}s}_{\text{(ii)}}.\end{aligned}$$

where the notation $\big\langle h_1, h_2\big\rangle_\rho = \int h_1 \cdot h_2\mathrm{d}\rho$ for any distribution ρ and functions h_1, h_2. We will provide bounds for term (i) and (ii) separately in the sequel.

Upper bounding term (i). For term (i) of (49), by the definitions of v, v^f, and v^g in (47) and (23), we have that

$$\begin{aligned}
&\partial_s v^f(\theta,\omega;\beta_s)\\
&=\alpha\partial_s\mathbb{E}_{\mathcal{D}}\Big[-g(Z;\beta_s)\Big\langle\frac{\delta\Phi(X,Z;f(\cdot;\beta_s))}{\delta f},\nabla_\theta\phi(\cdot;\theta)\Big\rangle_{L^2}-\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\beta_s))}{\delta f},\nabla_\theta\phi(\cdot;\theta)\Big\rangle_{L^2}\Big]\\
&=\alpha\nabla_\theta\mathbb{E}_{\mathcal{D}}\Big[-g(Z;\partial_s\beta_s)\Big\langle\frac{\delta\Phi(X,Z;f(\cdot;\beta_s))}{\delta f},\phi(\cdot;\theta)\Big\rangle_{L^2}-\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\partial_s\beta_s))}{\delta f},\phi(\cdot;\theta)\Big\rangle_{L^2}\Big].
\end{aligned}$$

where the second inequality holds since $\frac{\delta\Phi(X,Z;f)}{\delta f}$ a constant, $s-$independent function, $\frac{\delta\Psi(X,Z;f)}{\delta f}$ is linear in f, and $\partial_s f(\cdot;\beta_s),\partial_s g(\cdot;\beta_s)$ satisfies

$$\begin{aligned}
\partial_s f(w;\beta_s)&=\int\partial_s\big(\phi(w;\theta)\beta_s(\theta,\omega)\big)\mathrm{d}(\theta,\omega)=\int\phi(w;\theta)\partial_s\beta_s\mathrm{d}(\theta,\omega)=f(w;\partial_s\beta_s),\quad\forall w\in\mathcal{W}\\
\partial_s g(z;\beta_s)&=\int\partial_s\big(\psi(z;\omega)\beta_s(\theta,\omega)\big)\mathrm{d}(\theta,\omega)=\int\psi(z;\omega)\partial_s\beta_s\mathrm{d}(\theta,\omega)=g(z;\partial_s\beta_s),\quad\forall z\in\mathcal{Z}
\end{aligned}$$

A similar computation for $\partial_s v^g(\theta,\omega;\beta_s)$ gives

$$\partial_s v^g(\theta,\omega;\beta_s)=\alpha\nabla_\omega\mathbb{E}_{\mathcal{D}}\Big[\tilde{\Phi}(X,Z;f(\cdot,\partial_s\beta_s))\phi(Z;\omega)-g(Z;\partial_s\beta_s)\phi(Z;\omega)\Big]$$

We recall that $\tilde{\Phi}(x,z;f)=\Phi(x,z;f)-\Phi(x,z;\mathbf{0})$ is the linear component in Φ. We note that the variation of $\tilde{\Phi}$ is the same as the variation of Φ with respect to f, $\frac{\delta\Phi(X,Z;f)}{\delta f}=\frac{\delta\tilde{\Phi}(X,Z;f)}{\delta f}$. We define the potential $\mathcal{V}(\theta,\omega;\partial_s\beta_s)$ as

$$\begin{aligned}
&\mathcal{V}(\theta,\omega;\partial_s\beta_s)\\
&=\mathbb{E}_{\mathcal{D}}\Big[g(Z;\partial_s\beta_s)\Big\langle\frac{\delta\Phi(X,Z;f(\cdot;\beta_s))}{\delta f},\phi(\cdot;\theta)\Big\rangle_{L^2}+\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\partial_s\beta_s))}{\delta f},\phi(\cdot;\theta)\Big\rangle_{L^2}\Big]\\
&\qquad-\mathbb{E}_{\mathcal{D}}\Big[\tilde{\Phi}(X,Z;f(\cdot,\partial_s\beta_s))\psi(Z;\omega)-g(Z;\partial_s\beta_s)\psi(Z;\omega)\Big]
\end{aligned}\tag{50}$$

Then, the vector field $\partial_s v(\theta,\omega;\beta_s)$ is the gradient of such potential $\mathcal{V}(\theta,\omega;\partial_s\beta_s)$

$$\partial_s v(\theta,\omega;\beta_s)=\begin{pmatrix}\partial_s v^f(\theta;\beta_s)\\ \partial_s v^g(\omega;\beta_s)\end{pmatrix}=-\alpha\nabla\mathcal{V}(\theta,w;\partial_s\beta_s),$$

where the gradient operator $\nabla=(\nabla_\theta,\nabla_\omega)$. Then, by Stoke's formula and integration by parts, we have

$$\begin{aligned}
\big\langle\partial_s v(\cdot;\beta_s),u_s\big\rangle_{\beta_s}&=-\int\alpha\nabla\mathcal{V}(\theta,\omega;\partial_s\beta_s)u_s(\theta,w)\beta_s(\theta,w)\mathrm{d}(\theta,w)\\
&=\int\alpha\mathcal{V}(\theta,w;\partial_s\beta_s)\mathrm{div}(u_s\beta_s)\mathrm{d}(\theta,w)=-\int\alpha\mathcal{V}(\theta,w;\partial_s\beta_s)\partial_s\beta_s\mathrm{d}(\theta,w)
\end{aligned}$$

Integrating potential $\mathcal{V}$ with respect to $\partial_s\beta_s$ simplied the expression to

$$\begin{aligned}
&\int \alpha\mathcal{V}(\theta,\omega;\partial_s\beta_s)\partial_s\beta_s \mathrm{d}(\theta,\omega)\\
&\quad= \mathbb{E}_{\mathcal{D}}\Big[g(Z;\partial_s\beta_s)\Big\langle\frac{\delta\Phi(X,Z;f(\cdot;\beta_s))}{\delta f}, \int \alpha\phi(\cdot;\theta)\partial_s\beta_s(\mathrm{d}\theta)\Big\rangle_{L^2}\Big]\\
&\qquad+ \mathbb{E}_{\mathcal{D}}\Big[\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\partial_s\beta_s))}{\delta f}, \int \alpha\phi(\cdot;\theta)\partial_s\beta_s(\mathrm{d}\theta)\Big\rangle_{L^2}\Big]\\
&\qquad- \mathbb{E}_{\mathcal{D}}\Big[\tilde{\Phi}(X,Z;f(\cdot,\partial_s\beta_s))\int \alpha\psi(Z;\omega)\partial_s\beta_s(\mathrm{d}\omega) - g(Z;\partial_s\beta_s)\int \alpha\psi(Z;\omega)\partial_s\beta_s(\mathrm{d}\omega)\Big]\\
&\quad= \mathbb{E}_{\mathcal{D}}\Big[\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\partial_s\beta_s))}{\delta f}, f(\cdot;\partial_s\beta_s)\Big\rangle_{L^2} + g(Z;\partial_s\beta_s)^2\Big]. \qquad (51)
\end{aligned}$$

By convexity of $\Psi(x,z;f)$ and $\Psi(x,z;\mathbf{0})=0$ for all $(x,z)\in\mathcal{X}\times\mathcal{Z}$, it holds that

$$\Psi(x,z;f(\cdot;\partial_s\beta_s)) \le \Big\langle\frac{\delta\Psi(x,z;f(\cdot;\partial_s\beta_s))}{\delta f}, f(\cdot;\partial_s\beta_s)\Big\rangle_{L^2}, \quad \forall (x,z)\in\mathcal{X}\times\mathcal{Z}. \qquad (52)$$

Integrating (51) with respect to $s\in[0,1]$, we have that

$$\begin{aligned}
\int_0^1 \big\langle\partial_s v(\cdot;\beta_s), u_s\big\rangle_{\beta_s}\mathrm{d}s &= -\int_0^1 \mathbb{E}_{\mathcal{D}}\Big[\lambda\Big\langle\frac{\delta\Psi(X,Z;f(\cdot;\partial_s\beta_s))}{\delta f}, f(\cdot;\partial_s\beta_s)\big\rangle_{L^2} + g(Z;\partial_s\beta_s)^2\Big]\mathrm{d}s\\
&\le -\int_0^1 \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot;\partial_s\beta_s)) + g(Z;\partial_s\beta_s)^2\Big]\mathrm{d}s\\
&\le -\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot;\rho_t) - f(\cdot;\rho^*)) + \Big(g(Z;\rho_t) - g(Z;\rho^*)\Big)^2\Big]\\
&= -\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot;\rho_t) - f^*(\cdot)) + \big(g(Z;\rho_t) - g^*(Z)\big)^2\Big]. \qquad (53)
\end{aligned}$$

where the first inequality holds due to (52), and the second holds by Jensen's inequality.

Upper bounding term (ii). By Lemma 28, for term (ii) in (49), it holds that

$$\begin{aligned}
&\int\big\langle v(\theta,\omega;\beta_s), \partial_s(u_s(\theta,\omega)\beta_s(\theta,\omega))\big\rangle\mathrm{d}(\theta,\omega)\\
&\quad= \int\big\langle\nabla v(\theta,\omega;\beta_s), u_s(\theta,\omega)\otimes u_s(\theta,\omega)\beta_s(\theta,\omega)\big\rangle\mathrm{d}(\theta,\omega)\\
&\quad\le \sup_{\theta,\omega}\|\nabla v(\theta,\omega;\beta_s)\|_F\cdot\|u_s\|_{\beta_s}^2. \qquad (54)
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Since u_s is the velocity field corresponding to the geodesic connecting ρ^*, by assumptions, it holds that

$$\|u_s\|_{\beta_s}^2 = \mathcal{W}_2(\rho_t,\rho^*)^2 \le 4\mathcal{W}_2(\rho_0,\rho^*)^2 = 4\alpha^{-2}\bar{D}^2 = \mathcal{O}(\alpha^{-2}) \qquad (55)$$

On the other hand, by the definition of v in (47), we have that

$$\|\nabla v(\theta,\omega;\beta_s)\|_F^2 = \|\nabla_\theta v^f(\theta;\beta_s)\|_F^2 + \|\nabla_\omega v^g(\omega;\beta_s)\|_F^2 \qquad (56)$$

By the definition of v^f in (23), we have that

$$\begin{aligned}\big\|\nabla_\theta v^f(\theta;\beta_s)\big\|_F &\le \alpha\cdot\mathbb{E}_{\mathcal{D}}\Big[\Big|g(Z;\beta_s)\cdot\int_{\mathcal{W}}\frac{\delta\Phi(X,Z;f(\cdot;\beta_s))}{\delta f}(w')\mathrm{d}w'\Big|\Big]\cdot\sup_{w\in\mathcal{W}}\big\|\nabla^2_{\theta,\theta}\phi(w;\theta)\big\|_F^2\\ &\quad+\alpha\cdot\mathbb{E}_{\mathcal{D}}\Big[\lambda\cdot\Big|\int_{\mathcal{W}}\frac{\delta\Psi(X,Z;f(\cdot;\beta_s))}{\delta f}(w')\mathrm{d}w'\Big|\Big]\cdot\sup_{w\in\mathcal{W}}\big\|\nabla^2_{\theta,\theta}\phi(w;\theta)\big\|_F^2\\ &\le \alpha B_2\cdot\mathbb{E}_{\mathcal{D}}\Big[\lambda C_\Psi\big|f(W;\beta_s)\big|+C_2\big|g(Z;\beta_s)\big|\cdot\big|f(W;\beta_s)\big|\Big]. \end{aligned}\tag{57}$$

where the first inequality follows from Assumption 1, and second inequality comes from the integrability conditions in Assumption 3. Thus, it suffices to upper bound $f(w;\beta_s)$ and $g(z;\beta_s)$ for all $(w,z)\in\mathcal{W}\times\mathcal{Z}$. For $f(w;\beta_s)$, we have that

$$\begin{aligned}\big|f(w;\beta_s)\big| &= \alpha\cdot\Big|\int\phi(w\theta)\mathrm{d}\beta_s(\theta,\omega)\Big| = \alpha\cdot\Big|\int\phi(w;\theta)\ \mathrm{d}(\beta_s-\rho_0)(\theta,\omega)\Big|\\ &\le \alpha B_1\cdot\mathcal{W}_1(\beta_s,\rho_0)\le\alpha B_1\cdot\mathcal{W}_2(\beta_s,\rho_0).\end{aligned}\tag{58}$$

Moreover, it holds that

$$\mathcal{W}_2(\beta_s,\rho_0)\le\mathcal{W}_2(\beta_s,\rho^*)+\mathcal{W}_2(\rho^*,\rho_0)\le\mathcal{W}_2(\rho_t,\rho^*)+\mathcal{W}_2(\rho_0,\rho^*)\le 3\alpha^{-1}\bar{D},\tag{59}$$

where the second inequality follows from the fact that $\beta_s, s\in[0,1]$ is the geodesic connecting ρ_t and ρ^* and the last inequality follows from (ii) in Lemma 6. Plugging (59) into (58), we have that

$$\big|f(w;\beta_s)\big|\le\mathcal{O}(1),\quad\forall w\in\mathcal{W}.\tag{60}$$

Through a similar argument, such an upper bound can also be established for $g(z;\beta_s)$ for all $z\in\mathcal{Z}$,

$$\big|g(z;\beta_s)\big|\le\mathcal{O}(1),\quad z\in\mathcal{Z}.\tag{61}$$

Plugging (60) and (61) into (57), we establish an upper bound for $\big\|\nabla_\theta v^f(\theta;\beta_s)\big\|_F$,

$$\big\|\nabla_\theta v^f(\theta;\beta_s)\big\|_F\le\mathcal{O}(\alpha).\tag{62}$$

Similarly, by the definition of v^g in (23) we have that

$$\begin{aligned}\big\|\nabla_\omega v^g(\omega;\beta_s)\big\|_F &\le \alpha\cdot\mathbb{E}_{\mathcal{D}}\Big[\big|\Phi(X,Z;f(\cdot;\beta_s))\big|+\big|g(Z;\beta_s)\big|\Big]\cdot\sup_{z\in\mathcal{Z}}\big\|\nabla^2_{\omega,\omega}\psi(z;\omega)\big\|_F^2\\ &\le \alpha B_2\cdot\Big(\mathbb{E}_{\mathcal{D}}\Big[\big|\Phi(X,Z;\mathbf{0})\big|+C_2\big|f(W;\beta_s)\big|+\big|g(Z;\beta_s)\big|\Big)=\mathcal{O}(\alpha).\end{aligned}\tag{63}$$

Combining the bound from (62) and (63) and plugging into(56), it holds that

$$\Big\|\nabla v(\theta,\omega;\beta_s)\Big\|_F^2=\Big\|\nabla_\theta v^f(\theta;\beta_s)\Big\|_F^2+\Big\|\nabla_\omega v^g(\omega;\beta_s)\Big\|_F^2\le\mathcal{O}(\alpha^2).\tag{64}$$

Equation (55) and (64) provide upper bounds on the two terms involved in (54). Plugging the upper bounds that we have achieved, it holds that

$$\int\Big\langle v(\theta,w;\beta_s),\partial_s(u_s(\theta,\omega)\beta_s(\theta,\omega))\Big\rangle\mathrm{d}(\theta,\omega)\le\mathcal{O}(\alpha^{-1}).\tag{65}$$

Now combining (53) and (65), we have that

$$\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{\mathrm{d}t} \leq -\eta\cdot\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\rho_t)-f(\cdot;\rho^*))+\big(g(Z;\rho_t)-g(Z;\rho^*)\big)^2\Big]+C_*\cdot\eta\cdot\alpha^{-1}.$$

where $C_*=C_*\big(B_0,B_1,B_2,C,\lambda,\bar{D}\big)>0$ is a constant. This completes the proof of Lemma 13. ■

We are now ready to present the proof of Theorem 7 with the help of Lemma 13.

Proof We define

$$t^*=\inf\Big\{\tau\in\mathbb{R}_+\,\Big|\,\mathbb{E}_{\mathcal{D}}\big[\lambda\Psi(X,Z;f(\cdot,\rho_\tau)-f(\cdot;\rho^*))+\big(g(Z;\rho_\tau)-g(Z;\rho^*)\big)^2\big]<C_*\cdot\alpha^{-1}\Big\} \tag{66}$$

Also, we define

$$t_*=\inf\Big\{\tau\in\mathbb{R}_+\,\Big|\,\mathcal{W}_2(\rho_\tau,\rho^*)>2\mathcal{W}_2(\rho_0,\rho^*)\Big\} \tag{67}$$

In other words, (48) of Lemma 13 holds for $t\leq t_*$, and for $0\leq t\leq\min\{t_*,t^*\}$, we have

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{\mathrm{d}t} &\leq -\eta\cdot\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\rho_t)-f(\cdot;\rho^*))+\big(g(Z;\rho_t)-g(Z;\rho^*)\big)^2\Big]+C_*\cdot\eta\alpha^{-1}\\ &\leq 0\end{aligned}$$

We now show that $t_*>t^*$ by contradiction. By the continuity of $\mathcal{W}_2(\rho_t,\rho^*)^2$ with respect to t Ambrosio et al. (2008), since $\mathcal{W}_2(\rho_0,\rho^*)<2\mathcal{W}_2(\rho_0,\rho^*)$, it holds that $t_*>0$. Let's assume $t_*\leq t^*$, then $t_*=\min\{t_*,t^*\}$. Thus, by (48), (66), (67), it holds that for $0\leq t\leq t_*$ that

$$\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{\mathrm{d}t}\leq 0$$

which further implies that $\mathcal{W}_2(\rho_{t_*},\rho^*)\leq\mathcal{W}_2(\rho_0,\rho^*)$. This contradicts the definition of t_* in (67). Thus, it holds that $t_*\geq t^*$, which implies that (48) of Lemma 13 holds for any $0\leq t\leq t^*$. We now discuss two different situations.

Scenario (i) If $t_*\leq T$, then it holds that

$$\begin{aligned}&\inf_{t\in[0,T]}\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\mu_t)-f^*)+\big(g(Z;\nu_t)-g^*\big)^2\Big]\\ &\quad\leq\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\mu_{t_*})-f^*)+\big(g(Z;\nu_{t_*})-g^*\big)^2\Big]\\ &\quad<C_*\alpha^{-1}=\mathcal{O}(T^{-1}+\alpha^{-1}).\end{aligned} \tag{68}$$

Therefore, (68) implies Theorem 7 in this scenario.

Scenario (ii) If $t_*>T$, then (48) in Lemma 13 holds for $0\leq t\leq T$. Re-arranging the terms, we have the following inequality for all $0\leq t\leq T$,

$$\mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\mu_t)-f^*)+\big(g(Z;\nu_t)-g^*\big)^2\Big]\leq-\eta^{-1}\cdot\frac{1}{2}\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{\mathrm{d}t}+C_*\cdot\alpha^{-1} \tag{69}$$

This further suggests the following upper bound,

$$\begin{aligned}
&\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\mu_t)-f^*)+\big(g(Z;\nu_t)-g^*\big)^2\Big] \\
&\quad\le T^{-1}\cdot\int_0^T \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi(X,Z;f(\cdot,\mu_t)-f^*)+\big(g(Z;\nu_t)-g^*\big)^2\Big]\mathrm{d}t \\
&\quad\le 1/2\cdot\eta^{-1}\cdot T^{-1}\cdot\mathcal{W}_2(\rho_0,\rho^*)^2+C_*\cdot\alpha^{-1} \\
&\quad\le 1/2\cdot\alpha^{-2}\cdot\bar{D}^2\cdot\eta^{-1}\cdot T^{-1}+C_*\cdot\alpha^{-1}=\mathcal{O}(T^{-1}+\alpha^{-1}),
\end{aligned} \tag{70}$$

where the second inequality comes from integrating (69) in for $t\in[0,T]$, the third inequality comes from (ii) in Lemma 6 and last equality comes from setting η to α^{-2}. Therefore, (70) implies Theorem 7 in this scenario.

Based on the discussion of scenarios (i) and (ii) above, we finish the proof of Theorem 7. ■

B.3 Proof of Theorem 9

Proof We now prove Theorem 9. For notation simplicity, we denote $f_t=f(\cdot;\mu_t)$ as the estimator at time t. Recall the definition of $J(f)$ from (4) and $\bar{\delta}(z;f)$ from (3).

$$J(f)=\mathbb{E}_{\mathcal{D}}\big[1/2\cdot\bar{\delta}(Z;f)^2+\lambda\cdot\Psi(X,Z;f)\big],\quad \bar{\delta}(z;f)=\mathbb{E}_{X|Z}\big[\Phi(X,Z;f)\,\big|\,Z=z\big].$$

Plugging the definition of $J(f)$, it holds that

$$\begin{aligned}
&\inf_{t\in[0,T]} J(f_t)-J(f^*) \\
&\quad=\inf_{t\in[0,T]}\mathbb{E}_{\mathcal{D}}\Big[1/2\cdot\Big(\bar{\delta}(Z,f_t)^2-\bar{\delta}(Z,f^*)^2\Big)+\lambda\Big(\Psi(X,Z;f_t)-\Psi(X,Z;f^*)\Big)\Big].
\end{aligned} \tag{71}$$

Similar to the proof of Theorem 7, we define t_* as,

$$t_*=\inf\Big\{\tau\in\mathbb{R}_+\,\Big|\,\mathcal{W}_2(\rho_\tau,\rho^*)>2\mathcal{W}_2(\rho_0,\rho^*)\Big\}.$$

We will upper-bound the term in (71) separately in two different scenarios, depending on the value of t_* compared with T.

Scenario (i) If $t_*\le T$, then we have that

$$\inf_{t\in[0,T]} J(f_t)-J(f^*)\le J(f_{t_*})-J(f^*). \tag{72}$$

In order to upper-bound right-hand side of (72), we need to uniformly upper-bound $f_{t_*}(w)$ and $f^*(w)$ for all $w\in\mathcal{W}$. For $f_{t_*}(w)=f(w;\mu_{t_*})$, we have that

$$\begin{aligned}
\sup_{w\in\mathcal{W}}|f(w;\mu_{t_*})| &= \alpha\cdot\sup_{w\in\mathcal{W}}\Big|\int\phi(w;\theta)\mathrm{d}\mu_{t_*}(\theta)\Big|=\alpha\cdot\sup_{w\in\mathcal{W}}\Big|\int\phi(w;\theta)\mathrm{d}(\mu_{t_*}-\mu_0)(\theta)\Big| \\
&\le\alpha B_1\cdot\mathcal{W}_1(\mu_{t_*},\mu_0)\le\alpha B_1\cdot\mathcal{W}_2(\mu_{t_*},\mu_0)\ \le\alpha B_1\cdot\Big(\mathcal{W}_2(\rho_{t_*},\rho^*)+\mathcal{W}_2(\rho_0,\rho^*)\Big) \\
&\le 3B_1\cdot\bar{D}=\mathcal{O}(1).
\end{aligned} \tag{73}$$

where the first inequality follows from Lemma 29, the second inequality follows from Lemma 24. The last inequality follows from (ii) in Lemma (6) and definition of t_*. For f^*, a similar chain of inequalities would apply,

$$\begin{aligned}\sup_{w\in\mathcal{W}}|f(w;\mu^*)| &= \alpha\cdot\sup_{w\in\mathcal{W}}\Big|\int\phi(w;\theta)\mathrm{d}\mu^*(\theta)\Big| = \alpha\cdot\sup_{w\in\mathcal{W}}\Big|\int\phi(w;\theta)\mathrm{d}(\mu^*-\mu_0)(\theta)\Big|\\ &\le \alpha B_1\cdot\mathcal{W}_1(\mu^*,\mu_0)\le\alpha B_1\cdot\mathcal{W}_2(\mu^*,\mu_0)\ \le\alpha B_1\cdot\mathcal{W}_2(\rho^*,\rho_0)\\ &\le B_1\cdot\bar{D}=\mathcal{O}(1).\end{aligned}\tag{74}$$

With uniform bounds on f_{t_*} and f^*, we are now ready to upper-bound $\inf_{t\in[0,T]}J(f_t)-J(f^*)$ through upper-bounding $J(f_{t_*})-J(f^*)$,

$$\begin{aligned}&J(f_{t_*})-J(f_*)\qquad\qquad(75)\\ &\le\mathbb{E}_{\mathcal{D}}\Big[\bar{\delta}(Z;f_{t_*})\cdot\mathbb{E}_{X|Z}\big[\tilde{\Phi}(X,Z;f_{t_*}-f^*)|Z\big]+\lambda\cdot\Big\langle\frac{\delta\Psi(X,Z;f_{t_*})}{\delta f},f_{t_*}-f^*\Big\rangle_{L^2}\Big]\\ &\le\Big(\sup_{x,z}|\Phi(x,z;0)|+C_\Phi\cdot\sup_{w\in\mathcal{W}}|f(w;\mu_{t_*})|\Big)\cdot\mathbb{E}_{\mathcal{D}}\Big[C_\Phi\cdot|f(W;\mu_{t_*})-f(W;\mu^*)|\Big]\\ &\quad+\lambda C_\Psi\cdot\mathbb{E}_{\mathcal{D}}\Big[|f(W;\mu_{t_*})-f(W;\mu^*)|\Big]\\ &\le B_*\cdot\mathbb{E}_{\mathcal{D}}\Big[|f(W;\mu_{t_*})-f(W;\mu^*)|\Big]\le B_*\cdot\Big(\mathbb{E}_{\mathcal{D}}\Big[\lambda|f(W;\mu_{t_*})-f(W;\mu^*)|^2\Big]\Big)^{1/2}\\ &\le B_*\cdot\Big(\mathbb{E}_{\mathcal{D}}\Big[\Psi(X,Z;f_{t_*}-f^*)\Big]\Big)^{1/2}\le B_*\cdot\alpha^{-1/2},\qquad(76)\end{aligned}$$

where $B_*=B_*(\Phi,c_\phi,C_\Phi,C_\Psi,\lambda,C,B_1,\bar{D},C_*)>0$ is a constant and its values changes from line to line. The second inequality follows from (73) and (74). The last inequality follows from (68) in the proof of Theorem (7). Therefore, in this scenario, we have that

$$\inf_{t\in[0,T]}J(f_t)-J(f_*)\le J(f_{t_*})-J(f_*)\le\mathcal{O}(T^{-1/2}+\alpha^{-1/2}).\tag{77}$$

Equation (77) concludes the proof of Theorem 9 in the scenario of $t_*\le T$.

Scenario (ii) If $t_*>T$, by definition of t_*, we have that

$$\mathcal{W}_2(\mu_t,\mu^*)\le\mathcal{W}_2(\rho_t,\rho^*)\le 2\mathcal{W}_2(\rho_0,\rho^*)=2\alpha\cdot\bar{D},\quad\forall\,t\in[0,T].$$

Following the same arguments in (73) and (74), we have a uniform upper-bound for f_t for all $t\in[0,T]$ and f^* that writes,

$$\sup_{w\in\mathcal{W}}\big|f(w;\mu_t)\big|+\big|f(w;\mu^*)\big|\le 4B_1\cdot\bar{D}=\mathcal{O}(1),\quad\forall\,t\in[0,T].$$

Following the same derivation of (75), we have that

$$\begin{aligned}\inf_{t\in[0,T]}J(f_t)-J(f_*)&\le B_*\cdot\inf_{t\in[0,T]}B_*\cdot\Big(\mathbb{E}_{\mathcal{D}}\Big[\Psi(X,Z;f_t-f^*)\Big]\Big)^{1/2}\\ &\le B_*\cdot\Big(\inf_{t\in[0,T]}\mathbb{E}_{\mathcal{D}}\Big[\Psi(X,Z;f_t-f^*)\Big]\Big)^{1/2}\\ &\le B_*\cdot\Big(T^{-1}\cdot\int_0^T\mathbb{E}_{\mathcal{D}}\Big[\Psi(X,Z;f_t-f^*)\Big]\mathrm{d}t\Big)^{1/2}\\ &\le B_*\cdot\sqrt{1/2\cdot\bar{D}^2\cdot T^{-1}+C_*\cdot\alpha^{-1}}=\mathcal{O}(T^{-1/2}+\alpha^{-1/2}),\end{aligned}\tag{78}$$

where the last inequality follows from (69) and (70) in the proof of Theorem 7. Equation (78) concludes the proof of Theorem (9) in the scenario of $t_* > T$.

Based on the discussion of scenarios (i) and (ii) above, we finish the proof of Theorem 9. ■

Appendix C. Mean Field Limit of Neural Networks

In this section, we prove Proposition 4. The formal version is presented as follows. Let $\rho_t(\theta,\omega) = \mu_t(\theta) \otimes \nu_t(\omega)$, where (μ_t, ν_t) is the PDE solution in (24) and $\hat{\rho}_k(\theta,\omega) = N^{-1} \cdot \sum_{i=1}^N \delta_{\theta_k^i}(\theta) \cdot \delta_{\omega_k^i}(\omega)$ is the empirical distribution of $(\boldsymbol{\theta}_k, \boldsymbol{\omega}_k)$. Here we omit the dependence of the empirical distribution $\hat{\rho}_k$ on N and stepsize scale ϵ for notational simplicity.

Proposition 14 (Formal Version of Proposition 4) *Let $h : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ by any continuous function such that $\|h\|_\infty \le 1$ and $\mathrm{Lip}(h) \le 1$. Under Assumption 1, 3, with probability at least $1 - 5\delta$, it holds that*

$$\sup_{\substack{k \le T/\epsilon \\ (k\in\mathbb{N})}} \left| \int h(\theta,w) \mathrm{d}\rho_{k\epsilon}(\theta,w) - \int h(\theta,w) \mathrm{d}\widehat{\rho}_k(\theta,w) \right| \le B e^{BT} \Big(\sqrt{\log(N/\delta)/N} + \sqrt{\epsilon \cdot (D + \log(N/\delta))} \Big).$$

Here B is a constant that depends on $\alpha, \eta, \lambda, B_0, B_1$ and B_2.

The proof of Proposition 14 based heavily on Mei et al. (2018, 2019); Araújo et al. (2019); Zhang et al. (2020), which make use of the propagation of chaos arguments in Sznitman (1991). Recall that $(v^f(\cdot;\rho), v^g(\cdot;\rho))$ is the a vector field defined as,

$$\begin{aligned} v^f(\theta;\rho) &= \alpha \mathbb{E}_{\mathcal{D}} \Big[-g(Z;\rho) \Big\langle \frac{\delta \Phi(X,Z;f(\cdot;\rho))}{\delta f}, \nabla_\theta \phi(\cdot;\theta) \Big\rangle_{L^2} - \lambda \Big\langle \frac{\delta \Psi(X,Z;f(\cdot;\rho))}{\delta f}, \nabla_\theta \phi(\cdot;\theta) \Big\rangle_{L^2} \Big], \\ v^g(w;\rho) &= \alpha \mathbb{E}_{\mathcal{D}} \Big[\Phi(X,Z;f(\cdot,\rho)) \nabla_\omega \psi(Z;\omega) - g(Z;\rho) \nabla_\omega \psi(Z;\omega) \Big]. \end{aligned} \tag{79}$$

From now on, we equivalently write $\theta_k^i = \theta_i(k)$, $\omega_k^i = \omega_i(k)$ to emphasize the dependence on iterations. For abbreviation, we denote $\theta^{(N)}(k) = \{\theta_i(k)\}_{i=1}^N$ and $\omega^{(N)}(k) = \{\omega_i(k)\}_{i=1}^N$. We recall the finite-width representation of $f(\cdot;\theta^{(N)})$ and $g(\cdot;\omega^{(N)})$ are,

$$f(\cdot,\theta^{(N)}) = \frac{\alpha}{N} \cdot \sum_{i=1}^N \phi(\cdot;\theta_i), \qquad g(\cdot,\omega^{(N)}) = \frac{\alpha}{N} \cdot \sum_{i=1}^N \psi(\cdot;\omega_i).$$

Correspondingly, we defined the finite-width counter-part of v^f and v^g as following,

$$\begin{aligned} \hat{v}^f(\theta;\theta^{(N)},\omega^{(N)}) = \alpha \mathbb{E}_{\mathcal{D}} \Big[&- g(Z;\omega^{(N)}) \Big\langle \frac{\delta \Phi(X,Z;f(\cdot;\theta^{(N)}))}{\delta f}, \nabla_\theta \phi(\cdot;\theta) \Big\rangle_{L^2} \\ &- \lambda \Big\langle \frac{\delta \Psi(X,Z;f(\cdot;\theta^{(N)}))}{\delta f}, \nabla_\theta \phi(\cdot;\theta) \Big\rangle_{L^2} \Big], \\ \hat{v}^g(w;\theta^{(N)},\omega^{(N)}) = \alpha \mathbb{E}_{\mathcal{D}} \Big[&\Phi(X,Z;f(\cdot,\theta^{(N)})) \nabla_\omega \psi(Z;\omega) - g(Z;\omega^{(N)}) \nabla_\omega \psi(Z;\omega) \Big]. \end{aligned} \tag{80}$$

And we also defined the stochastic counterpart,

$$\hat{V}_k^f(\theta;\theta^{(N)},w^{(N)}) = \alpha\Big[-g(z_k;\omega^{(N)})\Big\langle\frac{\delta\Phi(x_k,z_k;f(\cdot;\theta^{(N)}))}{\delta f},\nabla_\theta\phi(\cdot;\theta)\Big\rangle_{L^2} - \lambda\Big\langle\frac{\delta\Psi(x_k,z_k;f(\cdot;\theta^{(N)}))}{\delta f},\nabla_\theta\phi(\cdot;\theta)\Big\rangle_{L^2}\Big],$$
$$\hat{V}_k^g(\omega;\theta^{(N)},w^{(N)}) = \alpha\Big(\Phi(x_k,z_k;f(\cdot;\theta^{(N)}))\nabla_\omega\psi(z_k;\omega) - g(z_k;\omega^{(N)})\nabla_\omega\psi(z_k;\omega)\Big). \quad (81)$$

where $(x_k,z_k)\sim\mathcal{D}$. Following from Mei et al. (2019); Araújo et al. (2019), we consider the following four dynamics.

- **Stochastic Gradient Descent Ascent (SGDA).** We consider the following SGDA dynamics for $\theta^{(N)}(k)$ and $\omega^{(N)}(k)$, where $k\in\mathbb{N}$, with $\theta_i(0)\overset{\text{i.i.d.}}{\sim}\mu_0, w_i(0)\overset{\text{i.i.d.}}{\sim}\nu_0$ ($i\in[N]$) as its initialization,

$$\theta_i(k+1) = \theta_i(k) + \eta\epsilon\cdot\hat{V}_k^f(\theta_i(k);\theta^{(N)}(k),\omega^{(N)}(k)),$$
$$\omega_i(k+1) = \omega_i(k) + \eta\epsilon\cdot\hat{V}_k^g(\omega_i(k);\theta^{(N)}(k),\omega^{(N)}(k)). \quad (82)$$

  Note that this dynamics is equivalent to (20).

- **Population Gradient Descent Ascent (PGDA).** We consider the following population gradient descent ascent dynamics for $\breve{\theta}^{(N)}(k)$ and $\breve{\omega}^{(N)}(k)$, where $k\in\mathbb{N}$, with $\breve{\theta}_i(0)=\theta_i(0)$, $\breve{\omega}_i(0)=\omega_i(0)$ ($i\in[N]$) as its initialization,

$$\breve{\theta}_i(k+1) = \breve{\theta}_i(k) + \eta\epsilon\cdot\hat{v}^f(\breve{\theta}_i(k);\breve{\theta}^{(N)}(k),\breve{\omega}^{(N)}(k)),$$
$$\breve{\omega}_i(k+1) = w_i(k) + \eta\epsilon\cdot\hat{v}^g(\breve{\omega}_i(k);\breve{\theta}^{(N)}(k),\breve{\omega}^{(N)}(k)). \quad (83)$$

- **Continuous-time Population Gradient Descent Ascent (CTPGDA).** We consider the following continuous time population gradient descent ascent dynamics for $\tilde{\theta}^{(N)}(t)$ and $\tilde{\omega}^{(N)}(t)$, where $t\in\mathbb{R}_+$, with $\tilde{\theta}_i(0)=\theta_i(0)$, $\tilde{\omega}_i(0)=\omega_i(0)$ ($i\in[N]$) as initialization,

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\theta}_i(t) = \eta\cdot\hat{v}^f(\tilde{\theta}_i(t);\tilde{\theta}^{(N)}(t),\tilde{\omega}^{(N)}(t)), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\tilde{\omega}_i(t) = \eta\cdot\hat{v}^g(\tilde{\omega}_i(t);\tilde{\theta}^{(N)}(t),\tilde{\omega}^{(N)}(t)). \quad (84)$$

- **Ideal particle (IP).** We consider the following ideal particle dynamics for $\bar{\theta}^{(N)}(t)$ and $\bar{w}^{(N)}(t)$, where $t\in\mathbb{R}_+$, with $\bar{\theta}_i(0)=\theta_i(0)$, $\bar{w}_i(0)=w_i(0)$ ($i\in[N]$) as initialization,

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\theta}_i(t) = \eta\cdot v^f(\bar{\theta}_i(t);\rho_t), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{\omega}_i(t) = \eta\cdot v^g(\bar{\omega}_i(t);\rho_t). \quad (85)$$

We aim to prove that $\hat{\rho}_k = N^{-1}\cdot\sum_{i=1}^N\delta_{\theta_i(k)}\cdot\delta_{w_i(k)}$ weakly converges to $\rho_{k\epsilon}$. For any continuous function h that satisfies the assumptions of Proposition 14, using the IP, CTPGDA, and

PGDA dynamics as interpolating dynamics, we have,

$$
\begin{aligned}
&\overbrace{\left|\int h(\theta,\omega)\mathrm{d}\rho_{k\epsilon}(\theta,\omega)-\int h(\theta,\omega)\mathrm{d}\hat{\rho}_k(\theta,\omega)\right|}^{\text{PDE}-\text{SGDA}} \\
&\quad\leq \underbrace{\left|\int h(\theta,\omega)\mathrm{d}\rho_{k\epsilon}(\theta)-N^{-1}\cdot\sum_{i=1}^{N}h\left(\bar{\theta}_i(k\epsilon),\bar{\omega}_i(k\epsilon)\right)\right|}_{\text{PDE}-\text{IP}}+\underbrace{\left\|(\bar{\theta},\bar{\omega})^{(N)}(k\epsilon)-(\tilde{\theta},\tilde{\omega})^{(N)}(k\epsilon)\right\|_{(N)}}_{\text{IP}-\text{CTPGDA}} \\
&\quad+\underbrace{\left\|(\tilde{\theta},\tilde{\omega})^{(N)}(k\epsilon)-(\breve{\theta},\breve{\omega})^{(N)}(k)\right\|_{(N)}}_{\text{CTPGDA}-\text{PGDA}}+\underbrace{\left\|(\breve{\theta},\breve{\omega})^{(N)}(k)-(\theta,\omega)^{(N)}(k)\right\|_{(N)}}_{\text{PGDA}-\text{SGDA}}. \qquad (86)
\end{aligned}
$$

The last inequality follows from the fact that $\mathrm{Lip}(h)\leq 1$. Here the norm $\|\cdot\|_{(N)}$ denotes the supremum norm over the sequence of vectors $(\theta,w)^{(N)}=\{(\theta_i,w_i)\}_{i=1}^N$,

$$
\left\|(\theta,\omega)^{(N)}\right\|_{(N)}=\sup_{i\in[N]}\left\|(\theta_i,\omega_i)\right\|. \qquad (87)
$$

In what follows, we define $B>0$ as a constant with its value varying from line to line. We establish the following lemmas as upper bounds of the four terms on the right-hand side of (86).

Lemma 15 (Upper Bound of $\mathrm{PDE}-\mathrm{IP}$) *Under Assumption 1 and 3, with probability at least $1-\delta$, it holds that*

$$
\sup_{t\in[0,T]}\left|\int h(\theta,\omega)\mathrm{d}\rho_t(\theta,\omega)-N^{-1}\sum_{i=1}^{N}h\big(\bar{\theta}_i(t),\bar{\omega}_i(t)\big)\right|\leq B\cdot\sqrt{\log(NT/\delta)/N}. \qquad (88)
$$

Lemma 16 (Upper Bound of $\mathrm{IP}-\mathrm{CTPGDA}$) *Under Assumption 1 and 3, with probability at least $1-2\delta$, it holds that*

$$
\sup_{t\in[0,T]}\left\|(\bar{\theta},\bar{\omega})^{(N)}(t)-(\tilde{\theta},\tilde{\omega})^{(N)}(t)\right\|_{(N)}\leq B\cdot e^{BT}\cdot\sqrt{\log(N/\delta)/N}. \qquad (89)
$$

Lemma 17 (Upper Bound of $\mathrm{CTPGDA}-\mathrm{PGDA}$) *Under Assumption 1 and 3, it holds that*

$$
\sup_{k\leq T/\epsilon}\left\|(\tilde{\theta},\tilde{\omega})^{(N)}(k\epsilon)-(\breve{\theta},\breve{\omega})^{(N)}(k))\right\|_{(N)}\leq B\cdot e^{BT}\cdot\epsilon. \qquad (90)
$$

Lemma 18 (Upper Bound of $\mathrm{PGDA}-\mathrm{SGDA}$) *Under Assumption 1 and 3, with probability at least $1-2\delta$, it holds that*

$$
\sup_{k\leq T/\epsilon}\left\|(\breve{\theta},\breve{\omega})^{(N)}(k))-(\theta,w)^{(N)}(k)\right\|_{(N)}\leq B\cdot e^{BT}\cdot\sqrt{\epsilon\cdot(D+\log(N/\delta)}. \qquad (91)
$$

With these lemmas, we are now ready to present the proof of Proposition 14.

Proof See §C.1.1, C.1.2, C.1.3, C.1.4 for detailed proofs for Lemma 15 to Lemma 18.

Plug in (88), (90), (90) and (91) to (86) and condition on the intersection of events in Lemma 15, 16, 17 and 18, we have that

$$\left|\int h(\theta,\omega)\mathrm{d}\rho_{k\epsilon}(\theta,\omega) - \int h(\theta,\omega)\mathrm{d}\hat{\rho}_k(\theta,\omega)\right| \leq B\cdot e^{BT}\cdot\left(\sqrt{\log(N/\delta)/N} + \sqrt{\epsilon\cdot(D+\log(N/\delta))}\right),$$

with probability at least $1-5\delta$. Thus, we complete the proof of Proposition 14. ■

C.1 Proofs of Lemmas 15-18

In this section, we present the proofs of Lemmas 15-18, which based heavily on Mei et al. (2018, 2019); Araújo et al. (2019); Zhang et al. (2020). The required supporting technical lemmas are in §D. The constant B presented in the proof is a positive constant whose values varies from line to line for notational simplicity.

C.1.1 Proof of Lemma 15

Proof We first consider the ideal particle dynamics in (85). It holds that $\bar{\theta}_i(t)\sim\mu_t, \bar{\omega}_i(t)\sim\nu_t,\ (i\in[N])$ (Proposition 8.1.8 in Ambrosio et al. (2008)). Since the randomness of $\bar{\theta}_i(t)$ and $\bar{\omega}_i(t)$ comes from $\theta_i(0)$ and $\omega_i(0)$ respectively while $\theta_i(0)$ and $\omega_i(0)$ $(i\in[N])$ are independent, $\bar{\theta}_i(t)\overset{\text{i.i.d.}}{\sim}\mu_t$, $\bar{\omega}_i(t)\overset{\text{i.i.d.}}{\sim}\nu_t$ $(i\in[N])$. Due to independence of $\bar{\theta}_i(t)$ and $\bar{\omega}_i(t)$, we also have $(\bar{\theta}_i(t),\bar{\omega}_i(t))\overset{\text{i.i.d.}}{\sim}\rho_t$ $(i\in[N])$. This implies the following,

$$\mathbb{E}_{\rho_t}\Big[N^{-1}\cdot\sum_{i=1}^N h(\bar{\theta}_i(t),\bar{w}_i(t))\Big] = \int h(\theta,\omega)\mathrm{d}\rho_t(\theta,\omega).$$

For notational simplicity, we denote $\gamma_i=(\theta_i,\omega_i)$, similar notations also generalize to $\bar{\gamma}_i,\tilde{\gamma}_i,\check{\gamma}_i$. Let $\gamma^{1,(N)}=\{\gamma_1,\ldots,\gamma_i^1\ldots,\gamma_N\}$ and $\gamma^{2,(N)}=\{\gamma_1,\ldots,\gamma_i^2,\ldots,\gamma_N\}$ be two sets of variables that only differ in the i-th element. Then, by the assumption that $\|f\|_\infty\leq 1$, we have the following bounded difference property,

$$\left|N^{-1}\sum_{j=1}^N h(\gamma_j^1) - N^{-1}\sum_{j=1}^N h(\gamma_j^2)\right| = N^{-1}\cdot|h(\gamma_i^1)-h(\gamma_i^2)|\leq 2/N.$$

Applying McDiarmid's inequality (Wainwright, 2019), we have for a fixed $t\in[0,T]$ that

$$\mathbb{P}\left(\left|N^{-1}\sum_{i=1}^N h\left(\bar{\gamma}_i(t)\right) - \int h(\gamma)\mathrm{d}\rho_t(\gamma)\right|\geq p\right)\leq\exp\left(-Np^2/4\right). \tag{92}$$

Moreover, we have for any $s,t\in[0,T]$ that,

$$\begin{aligned}
&\left|\left|N^{-1}\sum_{i=1}^{N}h\left(\bar{\gamma}_i(t)\right)-\int h(\gamma)\mathrm{d}\rho_t(\gamma)\right|-\left|N^{-1}\sum_{i=1}^{N}h\left(\bar{\gamma}_i(s)\right)-\int h(\gamma)\mathrm{d}\rho_s(\gamma)\right|\right|\\
&\quad\le\left|N^{-1}\sum_{i=1}^{N}h\left(\bar{\gamma}_i(t)\right)-N^{-1}\sum_{i=1}^{N}h\left(\bar{\gamma}_i(s)\right)\right|+\left|\int h(\gamma)\mathrm{d}\rho_t(\gamma)-\int h(\gamma)\mathrm{d}\rho_s(\gamma)\right|\\
&\quad\le\left\|\bar{\gamma}^{(N)}(t)-\bar{\gamma}^{(N)}(s)\right\|_{(N)}+\mathcal{W}_1\left(\rho_t,\rho_s\right)\le\left\|\bar{\gamma}^{(N)}(t)-\bar{\gamma}^{(N)}(s)\right\|_{(N)}+\mathcal{W}_2\left(\rho_t,\rho_s\right)\\
&\quad\le\left\|\bar{\theta}^{(N)}(t)-\bar{\theta}^{(N)}(s)\right\|_{(N)}+\left\|\bar{w}^{(N)}(t)-\bar{w}^{(N)}(s)\right\|_{(N)}+\mathcal{W}_2\left(\mu_t,\mu_s\right)+\mathcal{W}_2\left(\nu_t,\nu_s\right).
\end{aligned}$$

where the second inequality follows from the fact that $\mathrm{Lip}(h)\le 1$ and Lemma 29. The last inequality follows from the definition of $\gamma^{(N)}$, (87) and Lemma 24. Applying (122), (124) of Lemma 20, we have for any $s,t\in[0,T]$ that

$$\left|\left|N^{-1}\sum_{i=1}^{N}h(\bar{\gamma}_i(t))-\int h(\gamma)\mathrm{d}\rho_t\right|-\left|N^{-1}\cdot\sum_{i=1}^{N}h(\bar{\gamma}_i(s))-\int h(\gamma)\mathrm{d}\rho_s\right|\right|\le B\cdot\left|t-s\right|.$$

Apply the union bound to (92) for $t\in\iota\cdot\{0,1,\ldots,\lfloor T/\iota\rfloor\}$, we have that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|N^{-1}\sum_{i=1}^{N}h\left(\bar{\gamma}_i(t)\right)-\int h(\gamma)\mathrm{d}\rho_t(\gamma)\right|\ge p+B\cdot\iota\right)\le(T/\iota+1)\cdot\exp\left(-Np^2/4\right).$$

Setting $\iota=N^{-1/2}$ and $p=B\cdot\sqrt{\log(NT/\delta)/N}$, we have that

$$\sup_{t\in[0,T]}\left|N^{-1}\sum_{i=1}^{N}h\left(\bar{\theta}_i(t),\bar{\omega}_i(t)\right)-\int h(\theta,\omega)\mathrm{d}\rho_t\right|\le B\cdot\sqrt{\log(NT/\delta)/N}.$$

with probability at least $1-\delta$. Thus, we complete the proof of Lemma 15. ■

C.1.2 Proof of Lemma 16

Following from the definition of $\tilde{\theta}_i(t)$, $\tilde{w}_i(t)$ and $\bar{\theta}_i(t)$, $\bar{w}_i(t)$ in (84) and (85). We have for any $i\in[N]$ and $t\in[0,T]$ that

$$\begin{aligned}
\left\|\bar{\theta}_i(t)-\tilde{\theta}_i(t)\right\|&\le\int_0^t\left\|\frac{\mathrm{d}\tilde{\theta}_i(s)}{\mathrm{d}s}-\frac{\mathrm{d}\bar{\theta}_i(s)}{\mathrm{d}s}\right\|\mathrm{d}s\\
&\le\eta\cdot\int_0^t\left\|\hat{v}^f(\tilde{\theta}_i(s);\tilde{\theta}^{(N)}(s),\tilde{\omega}^{(N)}(s))-\hat{v}^f(\bar{\theta}_i(s);\bar{\theta}^{(N)}(s),\bar{\omega}^{(N)}(s))\right\|\mathrm{d}s\\
&\quad+\eta\cdot\int_0^t\left\|\hat{v}^f(\bar{\theta}_i(s);\bar{\theta}^{(N)}(s),\bar{\omega}^{(N)}(s))-v^f(\bar{\theta}_i(s);\rho_s)\right\|\mathrm{d}s\\
&\le B\cdot\int_0^t\left\|\bar{\theta}^{(N)}(s)-\tilde{\theta}^{(N)}(s)\right\|_{(N)}+\left\|\bar{\omega}^{(N)}(s)-\tilde{\omega}^{(N)}(s)\right\|_{(N)}\mathrm{d}s\\
&\quad+\eta\cdot\int_0^t\left\|\hat{v}^f(\bar{\theta}_i(s);\bar{\theta}^{(N)}(s),\bar{\omega}^{(N)}(s))-v^f(\bar{\theta}_i(s);\rho_s),\right\|\mathrm{d}s
\end{aligned}\tag{93}$$

where the last inequality follows from (118) of Lemma 19. Similarly, we have that

$$\begin{aligned}\big\|\bar{\omega}_i(t)-\tilde{\omega}_i(t)\big\| \;\le B\cdot\int_0^t\Big\|\bar{\theta}^{(N)}(s)-\tilde{\theta}^{(N)}(s)\Big\|_{(N)}+\Big\|\bar{\omega}^{(N)}(s)-\tilde{\omega}^{(N)}(s)\Big\|_{(N)}\mathrm{d}s\\ +\eta\cdot\int_0^t\Big\|\hat{v}^g(\bar{\omega}_i(s);\bar{\theta}^{(N)}(s),\bar{\omega}^{(N)}(s))-v^g(\bar{\omega}_i(s);\rho_s)\Big\|\mathrm{d}s,\end{aligned}\tag{94}$$

where the inequality follows from (119). We now upper-bound the second term of (93) and (94). We start with (93). Following from the definition of v^f and $\hat{v}^f$ in (79) and (80), we have for any $s\in[0,T]$ and $i\in[N]$ that

$$\Big\|\hat{v}^f(\bar{\theta}_i(s);\bar{\theta}^{(N)}(s),\bar{w}^{(N)}(s))-v^f(\bar{\theta}_i(s);\rho_s)\Big\|=\alpha^2\cdot\Big\|N^{-1}\cdot\sum_{j=1}^N Z_i^j(s)\Big\|,\tag{95}$$

where $Z_i^j(s)$ is given by,

$$\begin{aligned}Z_i^j(s)=\mathbb{E}_{\mathcal{D}}\bigg[\Big\langle\Big(\int\psi(Z;\omega)\mathrm{d}\nu_s(\omega)-\psi(Z;\bar{\omega}_j(s))\Big)\cdot\frac{\delta\Phi(X,Z;f)}{\delta f},\nabla_\theta\phi(\cdot;\bar{\theta}_i(s))\Big\rangle_{L^2}\\ +\lambda\cdot\Big\langle\frac{\delta\Psi(X,Z;\int\phi(\cdot;\theta)\mathrm{d}\mu_s(\theta))}{\delta f}-\frac{\delta\Psi(X,Z;\phi(\cdot;\bar{\theta}_j(s)))}{\delta f},\nabla_\theta\phi(\cdot;\bar{\theta}_i(s))\Big\rangle_{L^2}\bigg].\end{aligned}$$

Following from Assumption 1 and 3, we have that $\|Z_i^j(s)\|\le B$. When $j\neq i$, since $\bar{\theta}_j(s)\overset{\text{i.i.d.}}{\sim}\mu_s,\bar{\omega}_j(s)\overset{\text{i.i.d.}}{\sim}\nu_s$ ($j\in[N]$), it holds that $\mathbb{E}[Z_i^j(s)\,|\,\bar{\theta}_i(s)]=0$. Following from Lemma 21, we have for fixed $s\in[0,T]$ and $i\in[N]$ that

$$\begin{aligned}\mathbb{P}\bigg(\Big\|N^{-1}\cdot\sum_{j\neq i}Z_i^j(s)\Big\|\ge B\cdot\Big(N^{-1/2}+p\Big)\bigg)&=\mathbb{E}\Big[\mathbb{P}\Big(\Big\|N^{-1}\cdot\sum_{j\neq i}Z_i^j(s)\Big\|\ge B\cdot\Big(N^{-1/2}+p\Big)\;\Big|\;\bar{\theta}_i(s)\Big)\Big]\\ &\le\exp\big(-Np^2\big).\end{aligned}\tag{96}$$

From Lemma 29 and (124) of Lemma 20, we have that

$$\sup_{w\in\mathcal{W}}\Big|\int\phi(w;\theta)\mathrm{d}\mu_s(\theta)-\int\phi(w;\theta)\mathrm{d}\mu_t(\theta)\Big|\le B\cdot\mathcal{W}_1(\mu_s,\mu_t)\le B\cdot\mathcal{W}_2(\mu_s,\mu_t)\le B\cdot\big|s-t\big|.$$

Following from Assumption 1 and 3, Lemma 20, we have for any $s,t\in[0,T]$ that,

$$\bigg|\Big\|N^{-1}\cdot\sum_{j\neq i}Z_i^j(s)\Big\|-\Big\|N^{-1}\cdot\sum_{j\neq i}Z_i^j(t)\Big\|\bigg|\le B\cdot\big|t-s\big|.$$

Applying the union bound to (96) for $i\in[N]$ and $t\in\iota\cdot\{0,1,\ldots,\lfloor T/\iota\rfloor\}$, we have that

$$\mathbb{P}\Big(\sup_{\substack{i\in[N]\\ s\in[0,T]}}\big\|N^{-1}\cdot\sum_{j\neq i}Z_i^j(s)\big\|\ge B\cdot\Big(N^{-1/2}+p\Big)+B\iota\Big)\le N\cdot(T/\iota+1)\cdot\exp\big(-Np^2\big).$$

Setting $\iota = N^{-1/2}$ and $p = B \cdot \sqrt{\log(NT/\delta)/N}$, we have that

$$\sup_{\substack{i\in[N]\\ s\in[0,T]}} \left\| N^{-1} \cdot \sum_{j\neq i} Z_i^j(s) \right\| \leq B \cdot \sqrt{\log(NT/\delta)/N}. \tag{97}$$

with probability at least $1-\delta$. Following from Assumption 1, when $i = j$, $\|N^{-1} Z_i^i(s)\| \leq B/N$ in (95). Plugging (97) into (95), with probability at least $1-\delta$, we have that

$$\begin{aligned}\sup_{\substack{i\in[N]\\ s\in[0,T]}} \left\| \hat{v}^f(\bar{\theta}_i(s); \bar{\theta}^{(N)}(s), \bar{\omega}^{(N)}(s)) - v^f(\bar{\theta}_i(s); \rho_s) \right\| &\leq \sup_{i\in[N], s\in[0,T]} \alpha^2 \cdot \left\| N^{-1} \sum_{j=1}^{N} Z_i^j(s) \right\| \\ &\leq B \cdot \sqrt{\log(NT/\delta)/N}. \end{aligned} \tag{98}$$

Through similar arguments, with probability at least $1-\delta$, the second term of (94) holds

$$\sup_{\substack{i\in[N]\\ s\in[0,T]}} \left\| \hat{v}^g(\bar{w}_i(s); \bar{\theta}^{(N)}(s), \bar{\omega}^{(N)}(s)) - v^g(\bar{\omega}_i(s); \rho_s) \right\| \leq B \cdot \sqrt{\log(NT/\delta)/N}. \tag{99}$$

Now, conditioning on the intersection of event in (98) and event in (99), the following holds simultaneously for any $t \in [0, T]$

$$\left\| \tilde{\theta}^{(N)}(t) - \bar{\theta}^{(N)}(t) \right\|_{(N)} \leq B \cdot \int_0^t \left\| \tilde{\theta}^{(N)}(s) - \bar{\theta}^{(m)}(s) \right\|_{(N)} \mathrm{d}s + BT \cdot \sqrt{\log(NT/\delta)/N} \tag{100}$$

$$\left\| \tilde{\omega}^{(N)}(t) - \bar{\omega}^{(N)}(t) \right\|_{(N)} \leq B \cdot \int_0^t \left\| \tilde{\omega}^{(N)}(s) - \bar{\omega}^{(N)}(s) \right\|_{(N)} \mathrm{d}s + BT \cdot \sqrt{\log(NT/\delta)/N} \tag{101}$$

Summing (100) and (101) and applying Gronwall's Lemma (Holte, 2009), with probability at least $1-2\delta$, for any $t \in [0, T]$, it holds that

$$\begin{aligned}\left\| \tilde{\theta}^{(N)}(t) - \bar{\theta}^{(N)}(t) \right\|_{(N)} + \left\| \tilde{\omega}^{(N)}(t) - \bar{\omega}^{(N)}(t) \right\|_{(N)} &\leq B \cdot e^{Bt} \cdot 2BT \cdot \sqrt{\log(NT/\delta)/N} \\ &\leq B \cdot e^{BT} \cdot \sqrt{\log(N/\delta)/N}. \end{aligned} \tag{102}$$

The last inequality holds since B as a constant represents values changing from line to line. Therefore, equation (102) implies (89). Thus, we complete the proof of Lemma 16.

C.1.3 Proof of Lemma 17

By the definition of $\hat{v}^f, \hat{v}^g$ in (80), $\breve{\theta}_i(t), \breve{\omega}_i(t)$ in (83), $\tilde{\theta}_i(t), \tilde{\omega}_i(t)$ in (84), it holds that the distances $\left\| \widetilde{\theta}_i(k\epsilon) - \breve{\theta}_i(k) \right\|$ and $\|\widetilde{\omega}_i(k\epsilon) - \breve{\omega}_i(k)\|$ satisfy

$$\begin{aligned}&\left\| \widetilde{\theta}_i(k\epsilon) - \breve{\theta}_i(k) \right\| \\ &\quad\leq \eta \cdot \int_0^{k\epsilon} \left\| \widehat{v}^f\left(\widetilde{\theta}_i(s); \widetilde{\theta}^{(N)}(s), \widetilde{\omega}^{(N)}(s)\right) - \widehat{v}^f\left(\widetilde{\theta}_i(\lfloor s/\epsilon \rfloor \epsilon); \widetilde{\theta}^{(N)}(\lfloor s/\epsilon \rfloor \epsilon), \widetilde{\omega}^{(N)}(\lfloor s/\epsilon \rfloor \epsilon)\right) \right\| \mathrm{d}s \\ &\qquad + \eta \cdot \sum_{\ell=0}^{k-1} \left\| \widehat{v}^f\left(\widetilde{\theta}_i(\ell\epsilon); \widetilde{\theta}^{(N)}(\ell\epsilon), \widetilde{\omega}^{(N)}(\ell\epsilon)\right) - \widehat{v}^f\left(\breve{\theta}_i(\ell); \breve{\theta}^{(N)}(\ell), \breve{\omega}^{(N)}(\ell)\right) \right\| \\ &\quad\leq B \cdot k \cdot \epsilon^2 + B \cdot \sum_{\ell=0}^{k-1} \Big(\left\| \widetilde{\theta}^{(N)}(\ell\epsilon) - \breve{\theta}^{(N)}(\ell) \right\|_{(N)} + \left\| \widetilde{\omega}^{(N)}(\ell\epsilon) - \breve{\omega}^{(N)}(\ell) \right\|_{(N)} \Big). \end{aligned} \tag{103}$$

$$
\begin{aligned}
&\|\widetilde{\omega}_i(k\epsilon) - \breve{\omega}_i(k)\| \\
&\quad \le \eta \cdot \int_0^{k\epsilon} \left\| \widehat{v}^g \left(\widetilde{\omega}_i(s); \widetilde{\theta}^{(N)}(s), \widetilde{\omega}^{(N)}(s) \right) - \widehat{v}^g \left(\widetilde{\omega}_i(\lfloor s/\epsilon \rfloor \epsilon); \widetilde{\theta}^{(N)}(\lfloor s/\epsilon \rfloor \epsilon), \widetilde{\omega}^{(N)}(\lfloor s/\epsilon \rfloor \epsilon) \right) \right\| \mathrm{d}s \\
&\qquad + \eta \cdot \sum_{\ell=0}^{k-1} \left\| \widehat{v}^g \left(\widetilde{w}_i(\ell\epsilon); \widetilde{\theta}^{(N)}(\ell\epsilon), \widetilde{\omega}^{(N)}(\ell\epsilon) \right) - \widehat{v}^g \left(\breve{\omega}_i(\ell); \breve{\theta}^{(N)}(\ell), \breve{\omega}^{(N)}(\ell) \right) \right\| \\
&\quad \le B \cdot k \cdot \epsilon^2 + B \cdot \sum_{\ell=0}^{k-1} \Big(\left\| \widetilde{\theta}^{(N)}(\ell\epsilon) - \breve{\theta}^{(N)}(\ell) \right\|_{(N)} + \left\| \widetilde{\omega}^{(N)}(\ell\epsilon) - \breve{\omega}^{(N)}(\ell) \right\|_{(N)} \Big). \qquad (104)
\end{aligned}
$$

where (103) follows from (118) of Lemma 19 and (123) of Lemma 20, (104) follows from (119) of Lemma 19 and (123) of Lemma 20. Combining the inequalities in (103) and (104), it holds for any $k \le T/\epsilon$ ($k \in \mathbb{N}$) that

$$
\begin{aligned}
&\left\| \widetilde{\theta}^{(N)}(k\epsilon) - \breve{\theta}^{(N)}(k) \right\|_{(N)} + \left\| \widetilde{\omega}^{(N)}(k\epsilon) - \breve{\omega}^{(N)}(k) \right\|_{(N)} \\
&\qquad \le 2BT\epsilon + B \cdot \sum_{\ell=0}^{k-1} \left\| \widetilde{\theta}^{(m)}(\ell\epsilon) - \breve{\theta}^{(N)}(\ell) \right\|_{(N)} + B \cdot \sum_{\ell=0}^{k-1} \left\| \widetilde{\omega}^{(N)}(\ell\epsilon) - \breve{\omega}^{(N)}(\ell) \right\|_{(N)}. \qquad (105)
\end{aligned}
$$

Applying the discrete Gronwall's lemma (Holte, 2009) to (105) , we have that

$$
\sup_{\substack{k \le T/\epsilon \\ (k \in \mathbb{N})}} \left\| \widetilde{\theta}^{(N)}(k\epsilon) - \breve{\theta}^{(N)}(k) \right\|_{(N)} + \left\| \widetilde{\omega}^{(N)}(k\epsilon) - \breve{\omega}^{(N)}(k) \right\|_{(N)} \le 2B^2 \cdot T \cdot \epsilon \cdot e^{BT} \le B \cdot e^{BT} \cdot \epsilon,
$$

where the inequalities hold since we allow the value of B to vary from line to line. Thus, we complete the proof of Lemma 17.

C.1.4 Proof of Lemma 18

Proof Let $\mathcal{G}_k = \sigma(\theta^{(N)}(0), w^{(N)}(0), u_0, \ldots, u_k)$ be the $\sigma-$algebra generated by $\theta^{(N)}(0), w^{(N)}(0)$ and $u_\ell = (x_\ell, z_\ell)$ ($\ell \le k$). Following from the definition of $\hat{V}_k^f, \hat{V}_k^g$ and $\hat{v}^f, \hat{v}^g$ in (80) and (81), we have for any $i \in [N]$ and $k \in \mathbb{N}_+$ that

$$
\begin{aligned}
\mathbb{E}\Big[\hat{V}_k^f(\theta_i(k); \theta^{(N)}(k), \omega^{(N)}(k)) \,\big|\, \mathcal{G}_{k-1} \Big] &= \hat{v}^f(\theta_i(k); \theta^{(N)}(k), \omega^{(N)}(k)), \\
\mathbb{E}\Big[\hat{V}_k^g(\omega_i(k); \theta^{(N)}(k), \omega^{(N)}(k)) \,\big|\, \mathcal{G}_{k-1} \Big] &= \hat{v}^g(\omega_i(k); \theta^{(N)}(k), \omega^{(N)}(k)).
\end{aligned}
$$

Recall the definition of $\theta^{(N)}, \omega^{(N)}$ and $\breve{\theta}^{(N)}, \breve{\omega}^{(N)}$ as the SGDA and PGDA dynamics defined in (82) and (83). We have for any $i \in [N]$, $k \in \mathbb{N}_+$ that

$$
\begin{aligned}
&\left\| \breve{\theta}_i(k) - \theta_i(k) \right\| \\
&\quad \le \eta\epsilon \cdot \left\| \sum_{\ell=0}^{k-1} X_i(\ell) \right\| + \eta\epsilon \cdot \sum_{\ell=0}^{k-1} \left\| \widehat{v}^f \left(\breve{\theta}_i(\ell); \breve{\theta}^{(N)}(\ell), \breve{\omega}^{(N)}(\ell) \right) - \widehat{v}^f \left(\theta_i(\ell); \theta^{(N)}(\ell), \omega^{(N)}(\ell) \right) \right\| \\
&\quad \le \eta\epsilon \cdot \|A_i(k)\| + B\epsilon \cdot \sum_{\ell=0}^{k-1} \left\| \breve{\theta}^{(m)}(\ell) - \theta^{(m)}(\ell) \right\|_{(N)} + \left\| \breve{\omega}^{(N)}(\ell) - \omega^{(N)}(\ell) \right\|_{(N)}, \qquad (106)
\end{aligned}
$$

where the last inequality follows from (118) of Lemma 19. $X_i(\ell)$ and $A_i(k)$ are defined as,

$$X_i(\ell) = \widehat{V}_\ell^f\left(\theta_i(\ell); \theta^{(N)}(\ell), \omega^{(N)}(\ell)\right) - \mathbb{E}\left[\widehat{V}_\ell^f\left(\theta_i(\ell); \theta^{(N)}(\ell), \omega^{(N)}(\ell)\right) \mid \mathcal{G}_{\ell-1}\right] \quad \forall \ell \geq 1,$$
$$X_i(0) = 0, \quad A_i(k) = \sum_{\ell=0}^{k-1} X_i(\ell).$$

Following from (117) of Lemma 19, it holds that $\|X_i(\ell)\| \leq B$, thus the stochastic process $\{A_i(k)\}_{k\in\mathbb{N}_+}$ is a martingale with $\|A_i(k) - A_i(k-1)\| \leq B$. Applying the Azuma-Hoeffding bound in Lemma 22, we have that

$$\mathbb{P}\Big(\max_{\substack{k\leq T/\epsilon \\ (k\in\mathbb{N}_+)}} \|A_i(k)\| \geq B\cdot\sqrt{T/\epsilon}\cdot(\sqrt{D}+p)\Big) \leq \exp\left(-p^2\right). \tag{107}$$

Apply the union bound to (107) for $i \in [N]$, we have that

$$\mathbb{P}\Big(\max_{\substack{i\in[N] \\ k\leq T/\epsilon,(k\in\mathbb{N}_+)}} \|A_i(k)\| \geq B\cdot\sqrt{T/\epsilon}\cdot(\sqrt{D}+p)\Big) \leq N\cdot\exp\left(-p^2\right).$$

Setting $p = \sqrt{\log(N/\delta)}$, with probability at least $1-\delta$, it holds that

$$\|A_i(k)\| \leq B\cdot\sqrt{T/\epsilon}\cdot(\sqrt{D}+\sqrt{\log(N/\delta)}), \quad \forall i\in[N], k\leq T/\epsilon\,(k\in\mathbb{N}_+). \tag{108}$$

Plug (108) into (106) and taking supremum norm over $i \in [N]$, we have that

$$\left\|\breve{\theta}^{(N)}(k) - \theta^{(N)}(k)\right\|_{(N)} \leq B\epsilon\cdot\sum_{\ell=0}^{k-1}\left(\left\|\breve{\theta}^{(m)}(\ell) - \theta^{(m)}(\ell)\right\|_{(N)} + \left\|\breve{\omega}^{(N)}(\ell) - \omega^{(N)}(\ell)\right\|_{(N)}\right) + B\cdot\sqrt{T\epsilon}\cdot(\sqrt{D}+\sqrt{\log(N/\delta)}). \tag{109}$$

Through similar arguments, for $\breve{w}_i(k)$ and $w_i(k)$, with probability at least $1-\delta$,

$$\left\|\breve{\omega}^{(N)}(k) - \omega^{(N)}(k)\right\|_{(N)} \leq B\epsilon\cdot\sum_{\ell=0}^{k-1}\left(\left\|\breve{\theta}^{(m)}(\ell) - \theta^{(m)}(\ell)\right\|_{(N)} + \left\|\breve{\omega}^{(N)}(\ell) - \omega^{(N)}(\ell)\right\|_{(N)}\right) + B\cdot\sqrt{T\epsilon}\cdot(\sqrt{D}+\sqrt{\log(N/\delta)}). \tag{110}$$

Conditioning on the intersection of event in (109) and event in (110), summing (109), (110), and applying the discrete Gronwall's lemma (Holte, 2009), for any $k \leq T/\epsilon, k \in \mathbb{N}_+$, the following inequality holds with probability at least $1-2\delta$,

$$\begin{aligned}\left\|\breve{\theta}^{(N)}(k) - \theta^{(N)}(k)\right\|_{(N)} + \left\|\breve{\omega}^{(N)}(k) - \omega^{(N)}(k)\right\|_{(N)} &\leq B\cdot e^{Bk\epsilon}\cdot B\cdot\sqrt{T\epsilon}\cdot(\sqrt{D}+\sqrt{\log(N/\delta)}) \\ &\leq B\cdot e^{BT}\cdot\sqrt{\epsilon\cdot(D+\log(N/\delta))}.\end{aligned}$$

Here the last inequality holds since we allow the value of B to vary from line to line. Thus, we complete the proof of Lemma 18. ■

Appendix D. Supporting Lemmas

D.1 Supporting Lemmas for §C

In what follows, we presented the technical lemmas heavily used in § C. We recall the definition of v^f, v^g, $\hat{v}^f, \hat{v}^g$ and $\hat{V}_k^f, \hat{V}_k^g$ as in (79), (80), and (81) respectively. Let $B > 0$ be a constant depending on $\alpha, \eta, B_0, B_1, B_2, C$, whose value varies from line to line. Recall that $f(\cdot; \theta^{(N)})$ and $g(\cdot; \omega^{(N)})$ are the finite width representation with parameters $\theta^{(N)}, \omega^{(N)}$, whose definitions are given by

$$f(\cdot; \theta^{(N)}) = \frac{\alpha}{N} \cdot \sum_{i=1}^{N} \phi(\cdot; \theta_i), \quad g(\cdot; \omega^{(N)}) = \frac{\alpha}{N} \cdot \sum_{i=1}^{N} \psi(\cdot; \omega_i).$$

Lemma 19 *Under Assumption 1 and 3, it holds that for any $\theta^{(N)} = \{\theta_i\}_{i=1}^N$, $\underline{\theta}^{(N)} = \{\underline{\theta}_i\}_{i=1}^N$, $w^{(N)} = \{w_i\}_{i=1}^N$, $\underline{w}^{(N)} = \{\underline{w}_i\}_{i=1}^N$, that, $f(\cdot; \theta^{(N)})$ and $g(\cdot; \omega^{(N)})$ are uniformly bounded and Lipschitz in θ, ω respectively, which is given by the following,*

$$\sup_{w \in \mathcal{W}} \big| f(w; \theta^{(N)}) \big| + \sup_{z \in \mathcal{Z}} \big| g(z; \omega^{(N)}) \big| \leq B, \tag{111}$$

$$\sup_{w \in \mathcal{W}} \big| f(w; \theta^{(N)}) - f(w; \underline{\theta}^{(N)}) \big| \leq B \cdot \big\| \theta^{(N)} - \underline{\theta}^{(N)} \big\|_{(N)}, \tag{112}$$

$$\sup_{z \in \mathcal{Z}} \big| g(z; \omega^{(N)}) - g(z; \underline{\omega}^{(N)}) \big| \leq B \cdot \big\| \omega^{(N)} - \underline{\omega}^{(N)} \big\|_{(N)}. \tag{113}$$

Recall the definition of $\hat{v}^f, \hat{v}^g$ and $\hat{V}_k^f, \hat{V}_k^g$ in (80), (81), the finite width representation of the velocity field and its stochastic counter-part, when evaluated at arbitrary θ_i, ω_i, are also uniformly bounded and lipschitz in θ, ω respectively. This means for $\hat{V}_k^f, \hat{V}_k^g$, the following inequalities hold,

$$\big\| \hat{V}_k^f(\theta_i; \theta^{(N)}, \omega^{(N)}) \big\| + \big\| \hat{V}_k^g(\omega_i; \theta^{(N)}, w^{(N)}) \big\| \leq B, \tag{114}$$

$$\big\| \hat{V}_k^f(\theta_i; \theta^{(N)}, \omega^{(N)}) - \hat{V}_k^f(\underline{\theta}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)}) \big\| \leq B \cdot \Big(\big\| \theta^{(N)} - \underline{\theta}^{(N)} \big\|_{(N)} + \big\| \omega^{(N)} - \underline{\omega}^{(N)} \big\|_{(N)} \Big), \tag{115}$$

$$\big\| \hat{V}_k^g(\omega_i; \theta^{(N)}, w^{(N)}) - \hat{V}_k^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)}) \big\| \leq B \cdot \Big(\big\| \theta^{(N)} - \underline{\theta}^{(N)} \big\|_{(N)} + \big\| \omega^{(N)} - \underline{\omega}^{(N)} \big\|_{(N)} \Big). \tag{116}$$

A similar series of inequalities also hold for $\hat{v}^f, \hat{v}^g$,

$$\big\| \hat{v}^f(\theta_i; \theta^{(N)}, \omega^{(N)}) \big\| + \big\| \hat{v}^g(\omega_i; \theta^{(N)}, \omega^{(N)}) \big\| \leq B, \tag{117}$$

$$\big\| \hat{v}^f(\theta_i; \theta^{(N)}, \omega^{(N)}) - \hat{v}_k^f(\underline{\theta}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)}) \big\| \leq B \cdot \Big(\big\| \theta^{(N)} - \underline{\theta}^{(N)} \big\|_{(N)} + \big\| \omega^{(N)} - \underline{\omega}^{(N)} \big\|_{(N)} \Big), \tag{118}$$

$$\big\| \hat{v}^g(\omega_i; \theta^{(N)}, \omega^{(N)}) - \hat{v}^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)}) \big\| \leq B \cdot \Big(\big\| \theta^{(N)} - \underline{\theta}^{(N)} \big\|_{(N)} + \big\| \omega^{(N)} - \underline{\omega}^{(N)} \big\|_{(N)} \Big). \tag{119}$$

As a corollary of the inequalities stated above, the uniform bounds in fact hold for any $f, g \in \mathcal{F}$, which says,

$$\sup_{w \in \mathcal{W}} \big| f(w) \big| + \sup_{z \in \mathcal{Z}} \big| g(z) \big| \leq B. \tag{120}$$

Similarly, the uniform bounds also hold for the velocity field v^f, v^g, such that for any $\rho \in \mathscr{P}_2(\mathbb{R}^D \times \mathbb{R}^D)$, it holds that

$$\left\|v^f(\theta;\rho)\right\| + \left\|v^g(\omega;\rho)\right\| \leq B. \tag{121}$$

Proof We will prove these results separately.

(i) Proof of (111), (112), and (113)

For (111) of Lemma 19, since ϕ, ψ are bounded as is assumed in Assumption 1, we have for any $w \in \mathcal{W}, z \in \mathcal{Z}$, any $\theta^{(N)}$ and $\omega^{(N)}$ that

$$\left|f(w;\theta^{(N)})\right| + \left|g(z;\omega^{(N)})\right| \leq \alpha \cdot N^{-1} \sum_{i=1}^{N} \left|\phi(w;\theta_i)\right| + \left|\psi(z;\omega_i)\right| \leq B.$$

For (112), and (113) of Lemma 19, since for any $w \in \mathcal{W}$, $z \in \mathcal{Z}$, $\phi(w;\theta)$ has a bounded gradient in θ, $\psi(z;\omega)$ has a bounded gradient in ω. The uniform upper bound of the gradient controls the Lipschitz constant of the function, thus it holds for any $w \in \mathcal{W}, z \in \mathcal{Z}$, any $\theta^{(N)}, \underline{\theta}^{(N)}$ and $\omega^{(N)}, \underline{\omega}^{(N)}$ that

$$\begin{aligned}
\left|f(w;\theta^{(N)}) - f(w;\underline{\theta}^{(N)})\right| &\leq \alpha N^{-1} \cdot B_1 \sum_{i=1}^{N} \left|\theta_i - \underline{\theta}_i\right| \leq B \cdot \left\|\theta^{(N)} - \underline{\theta}^{(N)}\right\|_{(N)}, \\
\left|g(z;\omega^{(N)}) - g(z;\underline{\omega}^{(N)})\right| &\leq \alpha N^{-1} \cdot B_1 \sum_{i=1}^{N} \left|\omega_i - \underline{\omega}_i\right| \leq B \cdot \left\|\omega^{(N)} - \underline{\omega}^{(N)}\right\|_{(N)}.
\end{aligned}$$

(ii) Proof of (114), (115) and (116)

For (114) of Lemma 19, recall the definition of $\hat{V}_k^f$, $\hat{V}_k^g$ in (81), for any $\theta^{(N)}$ and $\omega^{(N)}$,

$$\begin{aligned}
\left\|\hat{V}_k^f(\theta_i;\theta^{(N)},\omega^{(N)})\right\| &\leq \alpha \cdot \sup_{w\in\mathcal{W}} \left\|\nabla_\theta \phi(w;\theta_i)\right\| \cdot \sup_{z\in\mathcal{Z}} \left|g(z;\omega^{(N)})\right| \cdot \int_{\mathcal{W}} \left|\frac{\delta\Phi(x_k,z_k,f(\cdot;\theta^{(N)}))}{\delta f}(w')\right| \mathrm{d}w' \\
&\quad + \alpha \cdot \sup_{w\in\mathcal{W}} \left\|\nabla_\theta \phi(w;\theta_i)\right\| \cdot \lambda \cdot \int_{\mathcal{W}} \left|\frac{\delta\Psi(x_k,z_k,f(\cdot;\theta^{(N)}))}{\delta f}(w')\right| \mathrm{d}w' \leq B, \\
\left\|\hat{V}_k^g(\omega_i;\theta^{(N)},\omega^{(N)})\right\| &\leq \alpha \cdot \left(\left|\Phi(x_k,z_k;f(\cdot;\theta^{(N)}))\right| + \sup_{z\in\mathcal{Z}} \left|g(z;\omega^{(N)})\right|\right) \cdot \sup_{z\in\mathcal{Z}} \left\|\nabla_\omega \psi(z;\omega_i)\right\| \leq B.
\end{aligned}$$

For notational simplicity, we further define

$$\begin{aligned}
u^f(\theta^{(N)},w^{(N)}) &= -\alpha g(z_k;\omega^{(N)}) \cdot \frac{\delta\Phi(x_k,z_k;f(\cdot;\theta^{(N)}))}{\delta f} - \alpha\lambda \cdot \frac{\delta\Psi(x_k,z_k;f(\cdot;\theta^{(N)}))}{\delta f}, \\
u^g(\theta^{(N)},w^{(N)}) &= \alpha\Phi(x_k,z_k;f(\cdot;\theta^{(N)})) - \alpha g(z_k;\omega^{(N)}).
\end{aligned}$$

For (115) of Lemma 19, following from Assumption 3 and the definition of $\hat{V}_k^f$ in (81), we have for any $\theta^{(N)}, \underline{\theta}^{(N)}$ and $\omega^{(N)}, \underline{\omega}^{(N)}$ that

$$\begin{aligned}
&\left\|\hat{V}_k^f(\theta_i;\theta^{(N)},\omega^{(N)}) - \hat{V}_k^f(\underline{\theta}_i;\underline{\theta}^{(N)},\underline{\omega}^{(N)})\right\| \\
&\quad\leq \left\|\hat{V}_k^f(\theta_i;\theta^{(N)},\omega^{(N)}) - \hat{V}_k^f(\theta_i;\underline{\theta}^{(N)},\underline{\omega}^{(N)})\right\| + \left\|\hat{V}_k^f(\theta_i;\underline{\theta}^{(N)},\underline{\omega}^{(N)}) - \hat{V}_k^f(\underline{\theta}_i;\underline{\theta}^{(N)},\underline{\omega}^{(N)})\right\| \\
&\quad\leq \left|u^f(\theta^{(N)},\omega^{(N)}) - u^f(\underline{\theta}^{(N)},\underline{\omega}^{(N)})\right| \cdot \sup_{w\in\mathcal{W}} \left\|\nabla_\theta \phi(w;\theta_i)\right\| + \left\|\left\langle u^f(\underline{\theta}^{(N)},\underline{\omega}^{(N)}), \nabla_\theta\phi(\cdot;\theta_i) - \nabla_\theta\phi(\cdot;\underline{\theta}_i)\right\rangle_{L^2}\right\|.
\end{aligned}$$

Moreover, $u^f(\theta^{(N)}, \omega^{(N)})$ is also Lipschitz in $(\theta^{(N)}, \omega^{(N)})$ since

$$\begin{aligned}|u^f(\theta^{(N)}, \omega^{(N)}) - u^f(\underline{\theta}^{(N)}, \underline{\omega}^{(N)})| &\leq B \cdot |f(w_k; \theta^{(N)}) - f(w_k; \underline{\theta}^{(N)})| + B \cdot |g(z_k; \omega^{(N)}) - g(z_k; \underline{\omega}^{(N)})| \\ &\leq B \cdot \Big(\big\|\theta^{(N)} - \underline{\theta}^{(N)}\big\|_{(N)} + \big\|\omega^{(N)} - \underline{\omega}^{(N)}\big\|_{(N)} \Big),\end{aligned}$$

where the second inequality is achieved by applying (112), (113). Therefore, the fact that $\hat{V}_k^f(\theta_i; \theta^{(N)}, \omega^{(N)})$ is Lipschitz in $(\theta^{(N)}, \omega^{(N)})$ is due to $\|\nabla_\theta \phi(w; \theta_i)\|$ and $\big|\int u^f(\theta^{(N)}, \omega^{(N)})(w')\mathrm{d}w'\big|$ is uniformly bounded.

For (116) of Lemma 19, following from Assumption 3 and the definition of $\hat{V}_k^g$ in (81), through a similar argument as is in the proof of (115), we have for any $\theta^{(N)}, \underline{\theta}^{(N)}$ and $\omega^{(N)}, \underline{\omega}^{(N)}$ that

$$\begin{aligned}&\Big\|\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)}) - \hat{V}_k^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\Big\| \\ &\quad\leq \Big\|\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)}) - \hat{V}_k^g(\omega_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\Big\| + \Big\|\hat{V}_k^g(\omega_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)}) - \hat{V}_k^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\Big\| \\ &\quad\leq |u^g(\theta^{(N)}, \omega^{(N)}) - u^f(\underline{\theta}^{(N)}, \underline{\omega}^{(N)})| \cdot \sup_{z \in \mathcal{Z}} \|\nabla_\omega \psi(z; \omega_i)\| + \Big\|\Big\langle u^g(\underline{\theta}^{(N)}, \underline{\omega}^{(N)}), \nabla_\omega \psi(\cdot; \omega_i) - \nabla_\omega \psi(\cdot; \underline{\omega}_i)\Big\rangle_{L^2}\Big\|.\end{aligned}$$

Again, $u^g(\theta^{(N)}, \omega^{(N)})$ is Lipschitz in $(\theta^{(N)}, \omega^{(N)})$ since

$$\begin{aligned}|u^g(\theta^{(N)}, \omega^{(N)}) - u^g(\underline{\theta}^{(N)}, \underline{\omega}^{(N)})| &\leq B \cdot |f(w_k; \theta^{(N)}) - f(w_k; \underline{\theta}^{(N)})| + B \cdot |g(z_k; \omega^{(N)}) - g(z_k; \underline{\omega}^{(N)})| \\ &\leq B \cdot \Big(\big\|\theta^{(N)} - \underline{\theta}^{(N)}\big\|_{(N)} + \big\|\omega^{(N)} - \underline{\omega}^{(N)}\big\|_{(N)} \Big).\end{aligned}$$

Therefore, the Lipschtizness of $\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)})$ in $(\theta^{(N)}, \omega^{(N)})$ comes from $\|\nabla_\omega \psi(z; \omega_i)\|$ and $\big|\int u^g(\theta^{(N)}, \omega^{(N)})(z')\mathrm{d}z'\big|$ is uniformly bounded.

(iii) Proof of (117)**,** (118)**, and** (119)

Equations (117), (118), (119) of Lemma 19 for $\hat{v}^f$ and $\hat{v}^g$ follow from the fact that

$$\hat{v}^f(\theta_i; \theta^{(N)}, \omega^{(N)}) = \mathbb{E}_{\mathcal{D}}\Big[\hat{V}_k^f(\theta_i; \theta^{(N)}, \omega^{(N)})\Big], \quad \hat{v}^g(\omega_i; \theta^{(N)}, w^{(N)}) = \mathbb{E}_{\mathcal{D}}\Big[\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)})\Big].$$

Therefore, (117) follows from (114) and triangle inequality,

$$\big\|\hat{v}^f(\theta_i; \theta^{(N)}, \omega^{(N)})\big\| + \big\|\hat{v}^g(\omega_i; \theta^{(N)}, \omega^{(N)})\big\| \leq \mathbb{E}_{\mathcal{D}}\Big[\big\|\hat{V}_k^f(\theta_i; \theta^{(N)}), \omega^{(N)})\big\|\Big] + \mathbb{E}_{\mathcal{D}}\Big[\big\|\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)})\big\|\Big] \leq B.$$

Equations (118) and (119) follows from (115), (116) and triangle inequality,

$$\begin{aligned}\big\|\hat{v}^f(\theta_i; \theta^{(N)}, \omega^{(N)}) - \hat{v}_k^f(\underline{\theta}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\big\| &\leq \mathbb{E}_{\mathcal{D}}\Big[\big\|\hat{V}_k^f(\theta_i; \theta^{(N)}, \omega^{(N)}) - \hat{V}_k^f(\underline{\theta}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\big\|\Big] \\ &\leq B \cdot \Big(\big\|\theta^{(N)} - \underline{\theta}^{(N)}\big\|_{(N)} + \big\|\omega^{(N)} - \underline{\omega}^{(N)}\big\|_{(N)} \Big), \\ \big\|\hat{v}^g(\omega_i; \theta^{(N)}, \omega^{(N)}) - \hat{v}_k^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\big\| &\leq \mathbb{E}_{\mathcal{D}}\Big[\big\|\hat{V}_k^g(\omega_i; \theta^{(N)}, \omega^{(N)}) - \hat{V}_k^g(\underline{\omega}_i; \underline{\theta}^{(N)}, \underline{\omega}^{(N)})\big\|\Big] \\ &\leq B \cdot \Big(\big\|\theta^{(N)} - \underline{\theta}^{(N)}\big\|_{(N)} + \big\|\omega^{(N)} - \underline{\omega}^{(N)}\big\|_{(N)} \Big).\end{aligned}$$

(iv) Proof of (120)**, and** (121)

Equation (120) follows from the definition of $\mathcal{F}$ in (26) and the uniform bounds of neuron functions ϕ and ψ. For any $f, g \in \mathcal{F}$, there exists probability measures $\hat{\mu}, \hat{\nu}$ over the parameter space such that

$$f(w) = \int \phi(w;\theta)\hat{\mu}(\mathrm{d}\theta), \quad g(z) = \int \psi(z;\omega)\hat{\nu}(\mathrm{d}\omega), \quad \forall w \in \mathcal{W}, z \in \mathcal{Z}.$$

We apply the triangle inequality and achieve,

$$\sup_{w\in\mathcal{W}} |f(w)| + \sup_{z\in\mathcal{Z}} |g(z)| \leq \int \sup_{w\in\mathcal{W}} |\phi(w;\theta)|\hat{\mu}(\mathrm{d}\theta) + \int \sup_{z\in\mathcal{Z}} |g(z)||\psi(z;\omega)|\hat{\nu}(\mathrm{d}\omega) \leq B.$$

Equation (121) follows from the definition of v^f, v^g in (79) and the proof of (114) and (117). Proof of (121) is the same as the proof for (114) and (117), except for the fact that a uniform bound is needed for the infinite width representation of f and g, which is proved in (120).

Based on proofs for items (i), (ii), (iii), and (iv) above, we finish the proof of Lemma (19). ■

Now, recall ρ_t is the PDE solution to (24), $\bar{\theta}^{(N)}(t), \bar{w}^{(N)}(t)$ is the IP dynamics defined in (85), $\tilde{\theta}^{(N)}(t), \tilde{w}^{(N)}(t)$ is the CTPGDA dynamics defined in (84). We have the following lemma that also bound the difference of iterates for IP, CTPGDA dynamics between time s and t.

Lemma 20 *Under Assumption 1 and 3, it holds for any $s, t \in [0, T]$ that,*

$$\big\|\bar{\theta}^{(N)}(t) - \bar{\theta}^{(N)}(s)\big\|_{(N)} + \big\|\bar{\omega}^{(N)}(t) - \bar{\omega}^{(N)}(s)\big\|_{(N)} \leq B \cdot \big|t - s\big|, \tag{122}$$

$$\big\|\tilde{\theta}^{(N)}(t) - \tilde{\theta}^{(N)}(s)\big\|_{(N)} + \big\|\tilde{\omega}^{(N)}(t) - \tilde{\omega}^{(N)}(s)\big\|_{(N)} \leq B \cdot \big|t - s\big|, \tag{123}$$

$$\mathcal{W}_2(\mu_t, \mu_s)) + \mathcal{W}_2(\nu_t, \nu_s)) \leq B \cdot \big|t - s\big|. \tag{124}$$

Proof For (122) of Lemma 20, by the definition of $\bar{\theta}_i(t)$ and $\bar{\omega}_i(t)$ in (85) and (121) of Lemma 19, we have for any $s, t \in [0, T]$ and $i \in [N]$ that

$$\big\|\bar{\theta}_i(t) - \bar{\theta}_i(s)\big\| \leq \eta \cdot \int_s^t \big\|v^f(\bar{\theta}_i(\tau); \rho_\tau)\big\|\mathrm{d}\tau \leq B \cdot \big|t - s\big|$$

$$\big\|\bar{\omega}_i(t) - \bar{\omega}_i(s)\big\| \leq \eta \cdot \int_s^t \big\|v^g(\bar{\omega}_i(\tau); \rho_\tau)\big\|\mathrm{d}\tau \leq B \cdot \big|t - s\big|$$

Similarly, for (123) of Lemma 20, by the definition of $\tilde{\theta}_i(t)$ and $\tilde{\omega}_i(t)$ in (84), and (117) of Lemma 19, we have for any $s, t \in [0, T]$ and $i \in [N]$,

$$\big\|\tilde{\theta}_i(t) - \tilde{\theta}_i(s)\big\| \leq B \cdot \big|t - s\big|, \quad \big\|\tilde{\omega}_i(t) - \tilde{\omega}_i(s)\big\| \leq B \cdot \big|t - s\big|.$$

For (124) of Lemma 20, following from the fact that $\bar{\theta}_i(t) \overset{\text{i.i.d.}}{\sim} \mu_t$, $\bar{\omega}_i(t) \overset{\text{i.i.d.}}{\sim} \nu_t$ and the definition of $\mathcal{W}_2$ in (15), it holds that for any $s, t \in [0, T]$ that

$$\mathcal{W}_2(\mu_t, \mu_s) \leq \mathbb{E}\Big[\big\|\bar{\theta}_i(t) - \bar{\theta}_i(s)\big\|^2\Big]^{1/2} \leq B \cdot |t - s|$$

$$\mathcal{W}_2(\nu_t, \nu_s) \leq \mathbb{E}\Big[\big\|\bar{\omega}_i(t) - \bar{\omega}_i(s)\big\|^2\Big]^{1/2} \leq B \cdot |t - s|$$

Therefore, we complete the proof of Lemma 20. ■

Lemma 21 *Let $\{X_i\}_{i=1}^N$ be i.i.d. random variables with $\|X_i\| \leq \xi$ and $\mathbb{E}[X_i] = 0$. Then it holds for any $p > 0$, there exists $C > 0$ being an absolute constant that*

$$\mathbb{P}\left(\left\|N^{-1} \cdot \sum_{i=1}^{N} X_i\right\| \geq C\xi \cdot \left(N^{-1/2} + p\right)\right) \leq \exp\left(-Np^2\right),$$

Proof See Lemma 30 in Mei et al. (2019) ■

Lemma 22 (Azuma-Hoeffding bound) *Let $X_k \in \mathbb{R}^D$ be a martingale with respect to the filtration $\mathcal{G}_k$ $(k \geq 0)$ with $X_0 = 0$. We assume for $\xi > 0$ and any $\lambda \in \mathbb{R}^D$ that,*

$$\mathbb{E}\left[\exp\left(\langle \lambda, X_k - X_{k-1}\rangle\right) \mid \mathcal{G}_{k-1}\right] \leq \exp\left(\xi^2 \cdot \|\lambda\|^2/2\right)$$

Then it holds that, with $C > 0$ being an absolute constant.

$$\mathbb{P}\left(\max_{\substack{k \leq n \\ (k \in \mathbb{N})}} \|X_k\| \geq C\xi \cdot \sqrt{n} \cdot (\sqrt{D} + p)\right) \leq \exp\left(-p^2\right)$$

Proof See Lemma 31 in Mei et al. (2019) and Lemma A.3 in Araújo et al. (2019). ■

Appendix E. Technical Results

E.1 Universal Function Approximation Theorem

In what follows, we introduce the universal function approximation theorem (Pinkus, 1999). For any given activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we consider the following function class,

$$\mathcal{G}(\sigma) = \Big\{\sum_{i=1}^{r} c_i \sigma(x^\top w^i + \theta_i) \,\Big|\, c_i, \theta_i \in \mathbb{R}, w^i \in \mathbb{R}^d\Big\}.$$

We denote by $\mathscr{C}(\mathbb{R}^d)$ the class of continuous functions over $\mathbb{R}^d$. Then, the following theorem holds.

Lemma 23 (Universal Function Approximation Theorem, Theorem 3.1 in Pinkus (1999)) *If the activation function $\sigma \in \mathscr{C}(\mathbb{R})$ is not a polynomial, the function class $\mathcal{G}(\sigma)$ is dense in $\mathscr{C}(\mathbb{R}^d)$ in the topology of uniform convergence on a compact set.*

E.2 Wasserstein Space

We use the definition of absolutely continuous curves in $\mathscr{P}_2(\mathbb{R}^D)$ in Ambrosio et al. (2008) and introduce the following lemmas.

Lemma 24 *For any probability measures $\mu, \nu, \mu', \nu' \in \mathscr{P}_2(\mathbb{R}^D)$, it holds that*

$$\mathcal{W}_2(\mu \otimes \nu, \mu' \otimes \nu')^2 \leq \mathcal{W}_2(\mu, \mu')^2 + \mathcal{W}_2(\nu, \nu')^2.$$

Lemma 25 (First Variation Formula, Theorem 8.4.7 in Ambrosio et al. (2008)) *Given $\nu \in \mathscr{P}_2(\mathbb{R}^D)$ and an absolutely continuous curve $\mu : [0,T] \to \mathscr{P}_2(\mathbb{R}^D)$, let $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ be the geodesic connecting μ_t and ν. It holds that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\mu_t,\nu)^2}{2} = -\langle \dot{\mu}_t, \dot{\beta}_0 \rangle_{\mu_t}.$$

where $\dot{\mu}_t = \partial_t \mu_t$, $\dot{\beta}_0 = \partial_s \beta_s|_{s=0}$.

Lemma 26 (Benamou-Brenier formula, Proposition 2.30 in Ambrosio and Gigli (2013)) *Let $\mu^0, \mu^1 \in \mathscr{P}_2(\mathbb{R}^D)$. Then, it holds that*

$$\mathcal{W}_2(\mu^0,\mu^1) = \inf\left\{ \int_0^1 \|\dot{\mu}_t\|_{\mu_t} \,\mathrm{d}t \,\middle|\, \mu : [0,1] \to \mathscr{P}_2(\mathbb{R}^D), \mu_0 = \mu^0, \mu_1 = \mu^1 \right\}.$$

Lemma 27 (Talagrand's Inequality, Corollary 2.1 in Otto and Villani (2000)) *Let ν be $N(0, \kappa \cdot I_D)$. It holds for any $\mu \in \mathscr{P}_2(\mathbb{R}^D)$ that*

$$\mathcal{W}_2(\mu,\nu)^2 \le 2D_{\mathrm{KL}}(\mu \,\|\, \nu)/\kappa.$$

Lemma 28 (Eulerian Representation of Geodesics, Proposition 5.38 in Villani (2003)) *Let $\beta : [0,1] \to \mathscr{P}_2(\mathbb{R}^D)$ be a geodesic and u be the corresponding vector field such that $\partial_t \beta_t = -\mathrm{div}(\beta_t \cdot u_t)$. It holds that*

$$\partial_t(\beta_t \cdot u_t) = -\mathrm{div}(\beta_t \cdot u_t \otimes u_t).$$

where $\otimes$ is the outer product of two vectors.

Lemma 29 (Dual Representation of the first order Wasserstein Distance, Villani (2008)) *The first order Wasserstein distance has the following dual representation form*

$$\mathcal{W}_1(\mu,\nu) = \sup\left\{ \int f(x)\mathrm{d}(\mu-\nu)(x) \,\middle|\, f : \mathbb{R}^D \to \mathbb{R} \text{ that is 1-Lipschitz continuous} \right\}$$

for any two probability measures $\mu, \nu \in \mathscr{P}_1(\mathbb{R}^D)$.

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, Alexander Gasnikov, and Dmitry Kovalev. Accelerated methods for composite non-bilinear saddle point problem. *arXiv preprint arXiv:1906.03620*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

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

Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, pages 1–155. Springer, 2013.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: In metric spaces and in the space of probability measures*. Springer, 2008.

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

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019a.

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

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International conference on artificial intelligence and statistics*, pages 172–235. PMLR, 2023.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, pages 11312–11322, 2019.

Yang Cai, Siddharth Mitra, Xiuyuan Wang, and Andre Wibisono. Convergence of the min-max langevin dynamics and algorithm for zero-sum games. *arXiv preprint arXiv:2412.20471*, 2024.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Xiaohong Chen and Sydney C Ludvigson. Land of addicts? an empirical investigation of habit-based asset pricing models. *Journal of Applied Econometrics*, 24(7):1057–1093, 2009.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Xiaohong Chen and Zhengling Qi. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pages 3558–3582. PMLR, 2022.

Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27:497–521, 2011.

Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33:13363–13373, 2020a.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv preprint arXiv:2002.04026*, 2020b.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

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

S. Darolles, Y. Fan, J. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79:1541–1566, 2011.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33: 12248–12262, 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

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

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

Cong Fang, Hanze Dong, and Tong Zhang. Over parameterized two-level neural networks can learn near optimal feature representations. *arXiv preprint arXiv:1910.11508*, 2019.

Cong Fang, Hanze Dong, and Tong Zhang. Mathematical models of overparameterized neural networks. *Proceedings of the IEEE*, 109(5):683–703, 2021a.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021b.

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. In *International Conference on Algorithmic Learning Theory*, pages 465–487. PMLR, 2022.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, 201(1):373–407, 2023.

Saeed Hajizadeh, Haihao Lu, and Benjamin Grimmer. On the linear convergence of extra-gradient methods for nonconvex–nonconcave minimax problems. *INFORMS Journal on Optimization*, 6(1):19–31, 2024.

Peter Hall and Joel Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33:2904–2929, 2005.

Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *Journal of Machine Learning Research*, 25(2):1–86, 2024.

John M Holte. Discrete Gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré, 2021.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International conference on machine learning*, pages 4542–4551. PMLR, 2020.

Minhui Huang, Xuxing Chen, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684*, 2022.

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Lower bounds and conditioning of differentiable games. *arXiv preprint arXiv:1906.07300*, page 31, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, 2018.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *Conference on Learning Theory*, pages 4362–4415. PMLR, 2022.

Juno Kim, Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Symmetric mean-field langevin dynamics for distributional minimax problems. *arXiv preprint arXiv:2312.01127*, 2023.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Chris Junchi Li, Huizhuo Yuan, Gauthier Gidel, Quanquan Gu, and Michael Jordan. Nesterov meets optimism: rate-optimal separable minimax optimization. In *International Conference on Machine Learning*, pages 20351–20383. PMLR, 2023.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth nonconvex-nonconcave minimax optimization: Primal-dual balancing and iteration complexity analysis. *arXiv preprint arXiv:2209.10825*, 2022.

Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation models: An adversarial approach. *Advances in Neural Information Processing Systems*, 33:8947–8958, 2020.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020b.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020a.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep resnet and beyond: Towards provable optimization via overparameterization from depth, 2020b.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

Luo Luo, Guangzeng Xie, Tong Zhang, and Zhihua Zhang. Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

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

Whitney Newey and James Powell. Instrumental variables estimation for nonparametric models. *Econometrica*, 71:1565–1578, 2003.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Dmitrii M Ostrovskii, Babak Barazandeh, and Meisam Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, 2021a.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021b.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, pages 1–14, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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

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

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

Cédric Villani. *Optimal transport: Old and new*. Springer, 2008.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Provably efficient neural GTD for off-policy learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Guangzeng Xie, Luo Luo, Yijiang Lian, and Zhihua Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *International Conference on Machine Learning*, pages 10504–10513. PMLR, 2020a.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020b.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020.

Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Mean field langevin actor-critic: Faster convergence and global optimality beyond lazy learning. In *Forty-first International Conference on Machine Learning*, 2024.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021a.

Yufeng Zhang, Qi Cai, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Can temporal-difference and q-learning learn representation? A mean-field theory. *arXiv preprint arXiv:2006.04761*, 2020.

Yufeng Zhang, Siyu Chen, Zhuoran Yang, Michael Jordan, and Zhaoran Wang. Wasserstein flow meets replicator dynamics: A mean-field analysis of representation learning in actor-critic. *Advances in Neural Information Processing Systems*, 34:15993–16006, 2021b.

Renbo Zhao. A primal-dual smoothing framework for max-structured non-convex optimization. *Mathematics of operations research*, 2023.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761. PMLR, 2022.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.