

# MEAN-FIELD NEURAL DIFFERENTIAL EQUATIONS: A GAME-THEORETIC SEQUENCE PREDICTION

Sungwoo Park\* Byungseung Kong

Korea University

{sungwoo\_park, xncb135}@korea.ac.kr

## ABSTRACT

We propose a novel class of neural differential equation models called *mean-field continuous sequence predictors* (MFPs) for efficiently generating continuous sequences with potentially infinite-order complexity. To address complex inductive biases in time-series data, we employ mean-field dynamics structured through carefully designed graphons. By reframing continuous sequence prediction as mean-field games, we utilize a fictitious play strategy integrated with gradient-descent techniques. This approach exploits the stochastic maximum principle to determine the Nash equilibrium of the system. Both empirical evidence and theoretical analysis highlight the unique advantages of our framework, where a collective of continuous predictors achieves highly accurate predictions and consistently outperforms benchmark prior works.

## 1 INTRODUCTION

Modeling spatiotemporal processes is central to understanding and predicting the behavior of complex systems that evolve across time and space. Recent work on neural differential equation models (Chen et al., 2019; Tzen & Raginsky, 2019) has shown that such architectures can effectively capture spatiotemporal dynamics in a wide range of applications, including generative modeling (Song et al., 2021), quantitative finance (Cohen et al., 2023), and physics-informed learning (Iakovlev et al., 2024). However, most of these approaches are formulated and evaluated under fixed, finitely sampled time grids, and therefore provide only limited theoretical insight into the following question about inherently continuous sequences: *How can we systematically model continuous-time sequences as the temporal discretization is refined and the effective number of events becomes very large?* In this work, we address this question by directly formulating the data dynamics in continuous time and then studying their behavior in the regime of increasingly fine temporal granularity. To obtain a tractable and theoretically grounded framework, we cast the prediction problem into the setting of *mean-field games* (Lasry & Lions, 2007), which gives rise to an infinite-dimensional predictive decision-making model that extends existing neural differential equation approaches (Tzen & Raginsky, 2019) to the analysis of continuous-time sequences.

The mean-field principle, a core philosophy in various scientific domains including neuroscience (Faugeras et al., 2009), statistical physics (Negele, 1982), and economics (Carmona, 2020; Cardaliaguet & Lehalle, 2018), serves as a powerful tool to model and analyze a large number of interacting agents, who behave in a manner that can be described as tragically rational within the decentralized coalition to satisfy *Nash equilibrium*. In this state of the mean-field regime, a continuum of infinitely many agents individually governs the dynamics of partially observed historical sequential data and collectively interacts with the others to make optimal group decisions for the prediction of future events. The foundational principle of this game-theoretic interpretation of the predictive system can be stated as follows: We reconstruct the continuous-time sequence prediction problem under the formal lens of mean-field games to gain powerful generalization capabilities in continuous sequence modeling. Stemming from the principle, we offer two main contributions:

- We extend conventional neural differential equation models by introducing mean-field principles, providing a new approach for modeling continuous sequences. This framework represents the stochastic spatiotemporal dynamics of an infinite continuum of agents and is rooted in hypotheses

---

\*Corresponding author.

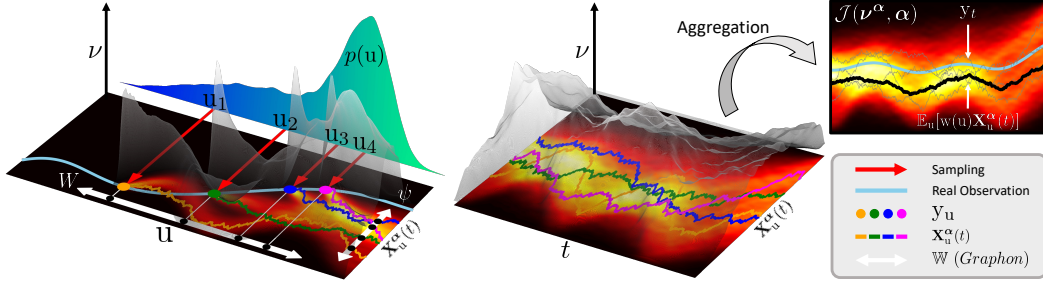


Figure 1: **(Left)**. The mean-field predictors are conditioned on a set of labeled past observations  $\{u_n\}_{n \leq N=4} \sim p(u)$ . Each spatiotemporal dynamic is interconnected via the neural graphon  $\mathbb{W}_\alpha$ , which leverages inductive biases tailored for continuous sequential data. **(Right)**. The collective decisions of a coalition of mean-field predictors are calibrated to approximate **(black trajectory)** the target future event interval.

from time series analysis (e.g., seasonality). We demonstrate that our proposed method surpasses state-of-the-art benchmarks in continuous sequence prediction tasks including variants of state-space models.

- We propose a gradient-based mean-field FBSDE approach that provides feasible computational complexity for approximating Nash equilibria in mean-field games. Building on the concentration of empirical measures and the propagation of chaos property in the mean-field regime, our theoretical findings clarify the number of past observations on the generalization performance of the mean-field system. We demonstrate that the coalition produces increasingly accurate and reliable predictions.

**Problem Setup.** Given past observations  $\{y_u\}$  in the interval  $u \in \mathbb{O} \subset \mathbb{T} = [0, T]$ , the primary objective of the continuous sequence prediction task is to accurately forecast future events  $\{y_t\}$  within the interval  $t \in \mathbb{T} \setminus \mathbb{O}$ . The continuous sequences  $\{y_u, y_t\} : [0, T] \rightarrow \mathbb{R}^d$  are continuously defined and share three notable properties: (1) *Irregularity*. The temporal granularity between spatio-temporal states in sequential data varies. (2) *Non-uniformity*. The cardinality of sequences exhibits stochastic and non-uniform behavior, fluctuating each time it is sampled from the dataset. (3) *Temporal scalability*. The sequential data spans multiple time scales, encompassing both short-term fluctuations and extended temporal ranges.

## 2 MEAN-FIELD CONTINUOUS SEQUENCE PREDICTORS

This section starts by introducing a stochastic differential equation model designed to depict infinite-order continuous signals, incorporating graphon structures for feature interactions.

**Definition 2.1.** (Mean-field Graphon SDEs) For the Markovian feedback controls  $\alpha : \mathcal{T} \times \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^d$  (i.e.,  $\alpha := \alpha(t, x; \theta)$ ) and continuous labels  $v \sim p(u)$ , we propose the  $\mathbb{R}^d$ -valued controlled stochastic differential equations called a mean-field graphon dynamics defined as follows:

$$d\mathbf{X}_u^\alpha(t) = \langle \mathbb{W}_\alpha[\nu_v(t)](u), \psi \rangle(\mathbf{X}_u^\alpha(t), \alpha)dt + \mathbf{b}(t, \mathbf{X}_u^\alpha(t), \alpha)dt + \sigma_t dW_t^u, \quad \mathbf{X}_u^\alpha(0) := y_u, \quad (1)$$

where a probability measure  $\nu := \{\nu_v(t)\}_{(v,t) \in \mathbb{O} \times \mathbb{T}}$  serves as a concise representation of the law of dynamics, and  $y_u \sim p(u, y)$  denotes a continuous representation of past observations.

The mean-field dynamics presented in Definition 2.1 involves three terms on the right-hand side, with an emphasis on important notions (A) *mean-field predictors* and (B) *neural graphons*.

(A) *Mean-field Predictor*. The proposed dynamical system incorporates two types of continuity encoding: *locality* (i.e.,  $t$ ) and *labeling* (i.e.,  $u$ ). The state variable  $\mathbf{X}_u^\alpha(t)$ , termed a *continuum of predictors* or *mean-field predictors (MFPs)*, represent a continuous set of information flows, each labeled by  $u \sim p(u)$  and initialized from the past observation,  $\mathbf{X}_u^\alpha(0) = y_u \sim p(u, y)$ . For instance, a continuum of predictors for the sequence of infinite i.i.d labels  $\mathbf{u}_\infty := \{u_n \sim p(u); n \leq N \rightarrow \infty\}$  in the mean-field regime  $\mathbf{X}_{\mathbf{u}_\infty}^\alpha(0)$  can be interpreted as being conditioned on the **past observational interval**, i.e., the support of the label distribution  $p(u)$ , with their future causal outcomes, producing  $\mathbf{X}_{\mathbf{u}_\infty}^\alpha(t)$  at **future event interval** being obtained from the dynamics in Eq (1).

The suggested model effectively handles continuous signals by ensuring both input and output are processed continuously. Within this setting, the closed Markovian control process  $\alpha(\cdot; \theta) \in \mathbb{A}$ ,

parameterized by neural networks  $\theta \in \Theta$ , *neural agents*, controls the state dynamics  $\mathbf{X}_{u_\infty}^\alpha(t)$ . Fig 1 depicts illustrative examples of how the proposed mean-field predictors are conditioned (**left**), propagated (**mid**), and utilized to produce future prediction (**right**). The overarching goal is then to calibrate the trajectory of predictors by determining the optimal neural agent  $\alpha^*$  that closely approximates the target future interval, e.g.,  $\mathbb{E}_t[\|\mathbb{E}_{u_\infty} \mathbf{X}_{u_\infty}^{\alpha^*}(t) - y_t\|_E^2] \approx 0$ , where decision aggregation  $w : \mathbb{O} \rightarrow [0, 1]$  captures the collective behavior of mean-field predictors. Section 3 will present a systematic algorithm to fulfill this objective.

(B) *Neural Graphon*. It is widely recognized in the literature that fundamental assumptions of inductive biases, such as *temporal decay*, *cycles*, and *seasonality* are vital for effective time series modeling. To incorporate these our mean-field system, we introduce a *neural graphon*, a graphon structure parameterized with neural networks, capturing the inherent heterogeneity among predictors.

**Definition 2.2.** (*Neural Graphon*) A graphon is a set of symmetric integrable function  $W : \mathbb{O}^2 \rightarrow \mathbb{R}$  equipped with  $\mathbb{L}^2$ -norm. For a probability measure  $\mu$  defined on  $\mathbb{O} \times \mathbb{R}^d$  with bounded second moment, we define a measure-valued function  $\mathbb{W}_\alpha[\mu](\cdot) : \mathbb{O} \rightarrow \mathcal{M}^a$  and a continuous symmetric function  $\psi_\alpha := \psi(y, x, \alpha) := H_\psi(\alpha) \text{Proj}(y - x)$  such that the first term in right-hand side of Eq (1) is defined as  $\langle \mathbb{W}_\alpha[\mu](u), \psi_\alpha \rangle(y, \alpha) := \mathbb{E}_{v \sim p(v), x \sim \mu} [W_\alpha(u, v) \psi_\alpha(y, x)] \in \mathbb{R}^d$ .

<sup>a</sup>Please refer to Section A.2 for the details.

For two tuples  $(x, u) \sim \nu_u \otimes p(u)$  and  $(y, v) \sim \nu_v \otimes p(v)$ , a symmetric function  $\psi$  estimates scaled relative dissimilarity between *spatial features*  $x$  and  $y$ . The neural agent, i.e.,  $H_\psi(\alpha)$ , then adjusts the importance of dissimilarity by rescaling projected vectors, i.e., **Proj**. Meanwhile, the neural graphon  $W$  encodes a degree of interaction between temporal variables  $u$  and  $v$ . Among the various graphon designs available, we propose two structures informed by inductive biases specific to continuous time series. Note that the key distinction from conventional methods is that our approach *directly models inductive biases in the data space*  $\mathbb{R}^d$ , rather than in latent feature spaces, facilitated by the graphon structure.

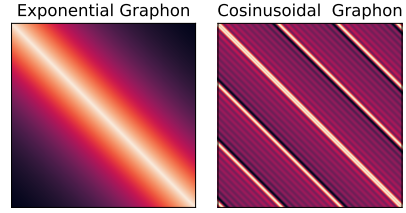


Figure 2: Visualization of Graphons.

*Exponential Graphon*. In the first graphon structure, we incorporate *temporal decay* (Che et al., 2018) assumption on spatiotemporal variables, which suggests that the influence of the past event decreases exponentially as time deviations increase. An example depicted in Fig 2 illustrates an exponential graphon, highlighting that events occurring close in time often show strong interactions. Here, the function i.e.,  $W_1 : \mathbb{A} \rightarrow \mathbb{R}^+$  determines the interaction magnitude. Subsequently, we introduce an exponential graph that diminishes the influence of events that are temporally distant;  $W_\alpha(u, v) := W_1(\alpha) \exp(-T^{-1} \Delta_u)$  with  $\Delta_u := |u - v|$ .

*Cosinusoidal Graphon*. The second graphon is intended to highlight the continuous *cyclic* assumption (Oreshkin et al., 2020), which characterizes the periodic aspects of time-series data. To embody this assumption, we conduct an eigen-decomposition of the proposed graphon operator on  $\mathbb{L}^2(\mathbb{O})$ , employing sinusoidal eigen-functions (i.e.,  $\{\psi_l\}$ ) and different frequency modes for the eigenvalues (i.e.,  $\{\lambda_l\}$ ), as suggested by Gao & Caines (2019);  $\mathbb{W} = \text{Id} + \sum_{k, l \in \mathbb{Z}_+} \lambda_l \varphi_l$ , where  $\{\varphi_l\} \subset \{\text{Id}, \sqrt{2} \cos 2\pi k, \sqrt{2} \sin 2\pi k\}$  and  $\{\lambda_l\} \subset \{a_0, b_k/2\}$ . We utilize neural networks to parameterize the graphon operator, substituting the Fourier coefficients  $\{\text{Id}, \lambda_l\}$  with equivalent neural networks, specifically  $W_0, W_{1,l}, W_{2,l} : \mathbb{A} \rightarrow \mathbb{R}^+$ . To illustrate different periodicities, we introduce  $\mathfrak{f}(l) \in \{1/2, 1/4, 1/8\}_{l \leq L}$ , which denotes a pre-determined series of frequencies. Consequently, we define a *cosinusoidal graphon* as follows:

$$W_\alpha(u, v) = W_0(\alpha) + \sum_{l \in \{1, \dots, L\}} W_{1,l}(\alpha) \cos(\cdot) + W_{2,l}(\alpha) \sin(\cdot). \quad (2)$$

where  $(\cdot) := 2\pi \mathfrak{f}(l) \Delta_u / |\mathbb{O}|$ . Note that we limit the summation to finite modes (i.e.,  $L$ ) for computational tractability. Fig 2 illustrates periodic interaction magnitudes for a predefined frequency setup. Further details on the implementation and their analysis can be found in the Appendix.

### 3 TRAINING MEAN-FIELD NEURAL NETWORKS

In the prior section, we introduced mean-field continuous sequence predictors based on SDEs that incorporate spatio-temporal interactions. Given that the mean-field system in Eq (2.1) is characterized as *controlled SDEs* with neural agents, we can define the objective function as a *stochastic control problem*. More precisely, our main aim is to reduce the cost functional  $\mathcal{J}$  formulated for training neural agents for sequence prediction and to derive the *value function*  $\mathcal{V}$ :

**Definition 3.1.** (Cost functional)<sup>a</sup> For the given neural graphon  $\mathbb{W}_\alpha$ , and fixed set of admissible controls  $\mathbb{A}$ , the cost functional is defined as follows:

$$\mathcal{V} := \inf_{\alpha \in \mathbb{A}} \mathcal{J}(\nu^\alpha, \alpha) = \inf_{\alpha \in \mathbb{A}} \mathbb{E}_{\alpha, \nu, t} [\|\mathbb{E}_{u \sim p(u)} \mathbf{X}_u^\alpha(t) - y_t\|_E^2 + \mathbf{G}^\alpha]. \quad (3)$$

where  $\mathbf{G}^\alpha := \mathbf{G}(\mathbf{X}_u^\alpha(T), \nu^\alpha)$  represents the terminal cost at time  $t = T$ , and  $w : \mathbb{O} \rightarrow [0, 1]$  is a decision aggregation function, satisfying  $\int w(u) du = 1$ .

<sup>a</sup>Please refer to Section A.4 for the details on the definition.

To predict future values, mean-field predictors operate by generating a unified measure, specifically referred to as a temporal marginal of predictors  $\mathbb{E}_{u \sim p(u)} \mathbf{X}_u^\alpha(t)$ . Here, the expectation accounts for the label  $u$  by amalgamating weighted outputs (*i.e.*,  $w$ ) from a range of predictors  $u \sim p(u) := w_\#[\text{Unif}(\mathbb{O})](u)$ <sup>1</sup>, aiming to approximate the target continuous interval  $\{y_t\}_{t \in \mathbb{T}}$ . Figure 1 (**right**) presents a demonstration of the decision-making mechanism. In pursuit of producing target intervals, neural agents are conditioned to extract the *value function*  $\mathcal{V}$ , which describes the state where a continuum of players unites to collaboratively forecast the most favorable future occurrences.

The difficulty in addressing this problem arises because the neural agent affects the population of predictors  $\nu^\alpha$ , which, in turn, persistently alters the individual state variables through interactions facilitated by the neural graphon. In the literature, these types of problems are typically described as (*graphon*) *mean-field games* (Lasry & Lions, 2007; Caines & Huang, 2021). In this study, we propose a novel methodology to cast **the continuous sequence prediction problem through the lens of mean-field games**. Our main aim is subsequently to identify the most suitable *optimal control*  $\alpha^*$  that fosters the optimal response in the recursive interaction between  $\mathcal{V}$  and  $\nu^\alpha$ . In our analysis, we explore the derivation of exact solutions  $(\mathcal{V}, \nu^{\alpha^*})$  from optimal control profiles over time, by studying the subsequent system of PDEs within the mean-field regime:

**Definition 3.2.** (Forward-Backward PDE System). For the obtained optimal neural agent  $\alpha^*$ , exact solutions of the value function in Eq (3) can be obtained by solving the following system of PDEs:

$$\begin{aligned} \partial_t \mathcal{V}(t, x) + \sigma_t^2 / 2 \Delta \mathcal{V}(t, x) + H(t, x, \partial_x \mathcal{V}, \nu_u(t), \alpha^*) &= 0, \\ \partial_t \nu_u^{\alpha^*}(t) - \sigma_t^2 / 2 \Delta \nu_u^{\alpha^*}(t) + \nabla \cdot \left[ \left( \mathbf{b}_W(x, \nu_u^{\alpha^*}(t), \alpha^*) + \mathbf{b}(t, x, \alpha^*) \right) \nu_u^{\alpha^*}(t) \right] &= 0, \end{aligned}$$

where  $\Delta$  and  $\nabla \cdot$  denotes Laplacian and divergence operators, respectively. The stochastic Hamiltonian system  $H$  is given by

$$H(t, x_u, a, \nu, \alpha) := (\mathbf{b}_W(x_u, \nu, \alpha) + \mathbf{b}(t, x_u, \alpha)) \cdot a + \|\mathbb{E}_{u \sim p(u)} x_u - y_t\|^2,$$

where  $\mathbf{b}_W(x, \nu, \alpha) := \langle \mathbb{W}_\alpha[\nu](u), \psi \rangle(x, \alpha)$  is the graphon interaction term in Definition 2.2.

A system of decoupled PDEs consists of the *Hamilton-Jacobi-Bellman (HJB)* equation and the *Fokker-Planck-Kolmogorov (FPK)* equation, which individually describes the propagation rules of the state variable and the value function over time. In mean-field equilibrium states, a set of PDEs are coupled as the law of the state variables  $\text{Law}(\mathbf{X}_u^\alpha(t))$  matches  $\nu_u(t)$  with marginal errors. This specific mathematical constraint can be formally expressed in the following definition:

<sup>1</sup>Here,  $f_\# \mu$  denotes a push-forward probability measure of  $\mu$  through function  $f$ .

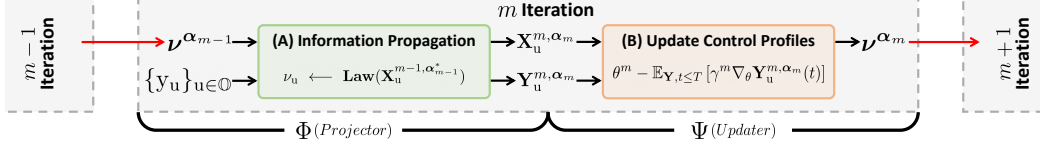


Figure 3: Illustrative Algorithm of the Gradient System of FBSDEs.

**Definition 3.3.** (*Mean-field  $\epsilon$ -Nash Equilibrium*). We say that a continuous flow of measure  $\nu_u(\cdot)$  is an  $\epsilon$ -equilibrium of (graphon) mean-field games if there exists a numerical constant  $\epsilon > 0$  such that the following inequality holds:  $\sup_{u,t} [\mathcal{W}_2^2(\nu_u(t), \text{Law}(\mathbf{X}_u^{\alpha^*}(t)))] \lesssim \mathcal{O}(\epsilon)$ , where  $\alpha^* \in \mathbb{A}$  is an optimal control of the problem in Eq. (3).

The mean-field equilibrium described in Definition 3.3 characterizes a scenario where a continuum of predictors is not incentivized to modify their policies  $\alpha^*$  to non-optimal counterpart  $\beta$ , which induces marginal errors, i.e.,  $\mathcal{J}(\nu^\beta, \beta) \geq \mathcal{J}(\nu^{\alpha^*}, \alpha^*)$ . Here, the law of optimal mean-field predictors closely approximates the population  $\nu_u$  with marginal errors  $\epsilon$ . This coupling integrates the Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck-Kolmogorov (FPK) equations, forming a *master equation*. Several numerical methods exist to approximate solutions to mean-field games including fixed-point iterations (Lauriere, 2021), and fictitious play (Min & Hu, 2021). However, these methodologies are typically constrained to linear quadratic dynamics, leading to computational intractability when confronting non-linearity (e.g., neural networks). Additionally, numerical simulations for obtaining analytic solutions of this system of PDEs present significant challenges due to the curse of dimensionality in high-dimensional data spaces. The following section is dedicated to addressing these issues by leveraging the deep neural architecture.

### 3.1 GRADIENT SYSTEM OF NEURAL FBSDEs

Inspired by computational algorithms designed for fictitious play (Cardaliaguet & Hadikhannloo, 2017), we explore a *gradient descent*-based algorithm, which enables us to tackle solving MFGs by fusing deep neural architectures. To be more specific, we propose a gradient system of *forward-backward stochastic differential equations* (Bensoussan et al., 2013), which is adapted for reflecting the update of neural agents with respect to the gradient descent algorithm.

**Definition 3.4.** (*Gradient System of FBSDEs*)<sup>a</sup>. For the fixed flow of measures  $\nu_u(\cdot) : \mathbb{T} \rightarrow \mathcal{P}_2$  and the fixed label  $u$  at each stage  $m$ , we consider a family of processes  $(\mathbf{X}_u(t), \mathbf{Y}_u(t), \mathbf{Z}_u(t))$  that solves forward-backward stochastic differential equations with respect to the proposed graphon system in Eq (1) given as follows:

$$\begin{aligned} d\mathbf{X}_u^{m, \alpha_m}(t) &= \mathbf{b}_W^m dt + \mathbf{b}^m dt + \sigma_t dW_t^u, \quad d\mathbf{Y}_u^{m, \alpha_m}(t) = -H^m dt - \mathbf{Z}_t^m \cdot dW_t^u, \\ \alpha_{m+1} &:= \alpha(t, \mathbf{X}_u^{m, \alpha_m}; \theta^m - \mathbb{E}_{\mathbf{Y}, t \leq T} [\gamma^m \nabla_\theta \mathbf{Y}_u^{m, \alpha_m}(t)]), \quad \nu_u = \text{Law}(\mathbf{X}_u^{m-1, \alpha_{m-1}^*}), \end{aligned}$$

where  $\gamma^m > 0$  is a learning rate of the gradient descent at  $m$ -th stage, and  $\{\alpha_m\}_m \subset \mathbb{A}$  is a set of admissible neural agents. Then, the triplet can be identified with  $(\mathbf{Y}_u(t), \mathbf{Y}_u(T), \mathbf{Z}_u(t)) = (\mathcal{J}, \mathbf{G}, (\partial_x \mathcal{J}) \sigma_t^{-1})$ .

<sup>a</sup>For the detailed description of the FBSDE system, please refer to the Definition A.2

The proposed gradient system can be decomposed by iterating a two-step procedure, i.e., (A) and (B), over a total of  $M$  stages. Fig 3 illustrates the evolution of the mean-field predictors related to the updated parameters of neural agents  $\alpha_m$  across different stages  $m$ . The details of the two-step procedure are specified below.

**(A) Information Propagation.** Initially, the system disseminates the information to a continuum of players by utilizing the population information of the previous stage, where the forward and backward system of SDEs propagates information relating to the updated population,  $\nu_u$ .

$$\nu_u \leftarrow \text{Law}(\mathbf{X}_u^{m-1, \alpha_{m-1}^*}), \quad (\mathbf{X}_u^m, \mathbf{Y}_u^m) \sim \text{Law}(\mathbf{X}_u^m | \nu_u) \otimes \text{Law}(\mathbf{Y}_u^m | \nu_u). \quad (4)$$

Note that the backward dynamics is propagated in **reverse** direction starting from its terminal state  $\mathbf{Y}_u(T) = \mathbf{G}$  while the forward dynamics evolve in the **forward** direction from the initial state. This shows that the proposed FBSDEs parallel the PDE system described in Definition 3.2.

**(B) Update Control Profiles.** In the subsequent step, the neural agent  $\alpha^m$  is updated with respect to its parameter  $\theta^m$  following the steepest direction of minimizing the values of backward dynamics  $\mathbf{Y}_u^m$ . The backward dynamics, associated with the cost functional  $\mathcal{J}$  as described in Proposition A.2, guide the updates of the parameters, allowing the mean-field predictors to gradually approximate the target interval. Since we have proposed an iterative algorithm to solve MFGs, the remaining part aims to provide convergence guarantees and highlight optimality conditions.

Proposition A.4 guarantees that the gradient system in Definition A.2 induces optimal neural agents  $\alpha^*$ , which yield a feasible value function (i.e.,  $\mathbf{Y}_u^m(0) \xrightarrow{m \rightarrow \infty} \mathcal{V}$ ) where the optimality of the control is represented in the sense of the *Pontryagin stochastic maximum principle* (Yong & Zhou, 2012). Specifically, we have the following two results:

$$\lim_{m \rightarrow \infty} H(\cdot, \alpha_m) \approx \inf_{\alpha \in \mathbb{A}} H(\cdot, \alpha), \quad dt \otimes d\nu, \quad \mathcal{V} \approx \mathbf{Y}_u^\infty(0) = \mathcal{J}(\nu^{\alpha^\infty}, \alpha_\infty). \quad (5)$$

The result illuminates that a pair  $(\lim_{m \rightarrow \infty} \alpha^m = \alpha^*, \lim_{m \rightarrow \infty} \nu^{\alpha_m} = \nu^{\alpha^*})$  solves both HJB and FPK equations in Def 3.2, assuring stochastic optimality. Having obtained the value function, the next step is to provide an explicit estimation of margin  $\varepsilon$  in the convergence of mean-field equilibrium.

**Convergence to Mean-field Equilibrium.** To rigorously analyze the convergence to equilibrium in a distributional sense, we define two distinct operators,  $\Phi$  and  $\Psi : \mathcal{M} \rightarrow \mathcal{M}$ , referred to as the *projector* and *updater*, respectively. Each operator corresponds to one of the two steps mentioned earlier, as illustrated in Fig. 3:

$$\begin{aligned} \Phi(\nu^{\alpha_m}) &:= \{\mathbf{Law}(\mathbf{X}_u^{\alpha_m}(t))\}_{\nu=\nu^{\alpha_{m-1}^*}; t \in \mathbb{T}, u \in \mathbb{O}}. \\ \Psi(\nu^{\alpha_{m-1}^*}) &:= \{\nu^{\alpha_m} = \alpha_{m-1}^*; \mathcal{V} = \mathcal{J}(\nu^{\alpha_{m-1}^*}, \alpha_{m-1}^*)\}. \end{aligned}$$

It can be easily verified that the composition of these operators at stage  $m$  maps the previous state's population to the next stage i.e.,  $\Phi \circ \Psi(\nu^{m-1}) = \nu^m$ . Proposition 3.5 asserts that the population  $\{\nu^{\alpha_m}\}_{m \leq M}$  generated by the proposed algorithm begins to converge in the Wasserstein metric as the stages  $m$  increase.

**Proposition 3.5. (informal)** *For arbitrary  $u \sim p(u)$  and  $t \in \mathbb{T}$ , the  $m$ -fold of composition  $\Phi \circ \Psi$  induces convergent behavior of squared 2-Wasserstein distance:*

$$\mathcal{W}_2^2([\Phi \circ \Psi]^{om}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{om}(\nu^{\alpha_0})) \lesssim \sup_{t \in \mathbb{T}} \|\nabla_\theta \mathbf{Y}^m\|_E \cdot \mathbb{O}(\gamma^m, C) := \varepsilon_m \xrightarrow{m \rightarrow \infty} 0. \quad (6)$$

where a numerical constant  $C$  is dependent on  $M, b_0, C_1, H_\psi, \mathbf{Lip}_b$  and  $\mathfrak{m}_2, |\mathbb{O}|, e^{-|\mathbb{O}|}, \mathbf{Lip}_W, \mathfrak{h}(\alpha) = \|W_\alpha\|_{\mathfrak{g}}$  is a cut-norm of the proposed graphons (i.e., exponential, cosinusoidal)

Proposition 3.5 reveals two theoretical implications regarding the convergence property. First, the proposed gradient system converges in a distributional sense, as the Wasserstein distance between the populations  $([\Phi \circ \Psi]^{om}(\nu^{\alpha_1}) = \nu^{\alpha_{m+1}}$  and  $([\Phi \circ \Psi]^{om}(\nu^{\alpha_0}) = \nu^{\alpha_m})$ , governed by the gradient norm of the backward dynamics, is expected to decrease as  $m$  increases. In other words,  $\{\Phi \circ \Psi\}^{om}$  is a Cauchy sequence in  $\mathcal{M}$ , ensuring the convergent behavior of the proposed training scheme. Second, the proposed gradient system ensures the convergence of the dynamics for the upper bounds  $\varepsilon_m$ . It is important to note that the inequality in Eq (6) is an equivalent expression of the **mean-field Nash  $\varepsilon_m$ -equilibrium** described in Definition 3.3. In this context, the neural agent with greater capacity (i.e., a smaller radius  $\mathbf{r}_m$  of the metric balls in Eq (43)) further tightens the upper bound. In conclusion, the findings from Proposition 3.5 validate that the proposed gradient system efficiently utilizes neural networks to solve mean-field games in continuous sequence prediction.

## 4 SAMPLING MEAN-FIELD PREDICTORS

In this section, we propose the numerical algorithm for sampling the proposed mean-field predictors and provide a theoretical analysis of the sample complexity error and the asymptotic convergence of empirical estimation for mean-field predictors. Inspired by the Euler-Maruyama approach for McKean-Vlasov type Reisinger & Stockinger (2022), we introduce an Euler-Maruyama method tailored for graphon-interacting particle systems to sample a series of mean-field predictors.

Algorithm 1 in Section A.5 details the computational procedure for sampling these mean-field predictors. We assume that  $\alpha^* := \alpha(\cdot; \theta^*)$  is nearly optimal, as defined by the  $\varepsilon$ -Nash equilibrium

derived from the mean-field gradient system associated with FBSDEs. The infinite-dimensional characteristics of the proposed system introduce inherent complexity challenges when mean-field predictors are sampled and applied to finite-dimensional, real-world datasets. When these sampled mean-field predictions aim to approximate their mean-field limits, an important question regarding sample complexity surfaces. To thoroughly investigate, we begin by defining the probabilistic description for both the sampled and model dynamics, as elaborated below:

$$\text{MFPs in Alg. 1 : } \nu_t^N := \frac{1}{N} \sum_i^N \delta_{\mathbf{X}_i^n(t)}, \quad \text{MFPs with } \infty\text{-order : } \hat{\mu}_t := \mathbb{E}_{\mathbf{u} \sim p(\mathbf{u})}[\nu_{\mathbf{u}}(t)]. \quad (7)$$

where  $\mathbf{X}_i^n(t) \sim \nu_t^N$  is sampled predictors, which can be obtained from implementing the Algorithm 1 and the weighted sum  $\Lambda_t$  approximates true collective prediction made by mean-field predictors  $\mathbb{E}_{\mathbf{u}} \mathbf{X}_{\mathbf{u}}^\alpha(t) \sim \hat{\mu}_t$  in Eq (3).

**Proposition 4.1.** (*Sampling Complexity*) For arbitrary  $\mathbf{u} \in \mathbb{O}$ , let  $\nu_t^N, \hat{\mu}_t$  be probability measures defined in Eq (7). Then, there exist numerical constants  $\mathfrak{c}_4, \mathfrak{c}_7, \mathfrak{c}_8, \mathfrak{c}_9 > 0, w > 0$  and  $\kappa > 0$  such that the probability of squared 2-Wasserstein distance can be controlled as follows:

$$\sup_{t \in \mathbb{T}} \mathbb{P} [\mathcal{W}_2^2(\nu_t^N, \hat{\mu}_t) \geq \epsilon] \leq \mathfrak{A}(\mathfrak{B} + \mathfrak{C} + \mathfrak{D}), \quad \mathfrak{A} := \frac{2\mathfrak{c}_7^{3/2}}{\kappa} \exp(\mathfrak{c}_4 e^{\frac{1}{2}\mathfrak{c}_1 T}) (e^{\kappa T} - 1) \vee \mathfrak{c}_9 \exp(-4\mathfrak{c}_8),$$

$$\mathfrak{B} := \frac{e^{-N\epsilon^2/4\mathfrak{c}}}{\epsilon^2}, \quad \mathfrak{C} := \frac{1}{72^4 \epsilon \sqrt{N}}, \quad \mathfrak{D} := \frac{e^{-N\epsilon}}{N} \left( 1 - \frac{128\omega\mathfrak{h}(\alpha)}{N} \right)^{-d/8}.$$

The established inequality presents the relation between squared 2-Wasserstein distance and the number of samples  $N$ , the dimensionality of the data distribution  $d$ . The proof primarily draws on the findings presented in Bolley et al. (2007). It is important to note that the result also guarantees the proposed system benefits from the *propagation of chaos* (Chaintron & Diez, 2022), validating the asymptotic behavior of the sampled predictions generated by the mean-field predictors:

$$\sup_{t \in \mathbb{T}} \lim_{k \rightarrow \infty} \mathcal{W}_2^2(\text{Law}(\mathbf{X}_{i_1}^n, \dots, \mathbf{X}_{i_k}^n), \otimes_{\{j=1, \dots, k\}} \nu_{j/n}(t)) \leq \Omega(N, k) \rightarrow 0$$

Proposition 4.1, A.5 and the inequality above align with the intuition that *as the number of predictors  $N$  increases (and dimensionality  $d$ ), the sampled dynamics converges more closely to the mean-field limit  $\hat{\mu}_t$  and  $\nu_{\mathbf{u}}(t)$* . In particular, the right-hand side of the inequality is influenced by two exponentially decaying terms, while the other term decreases at a polynomial rate, both showing short-tailed concentration with respect to the number of past observations. Overall, our theoretical findings highlight advantages of capitalizing on mean-field games: Rational individuals (*i.e.*,  $\delta_{\mathbf{X}_i^n(t)}$ ) satisfying Nash equilibrium and conditioned on partial information (*i.e.*,  $\mathbf{X}_i^n(0) = y_{i/n}$ ) forms a coalition (*i.e.*,  $\nu_t^N$ ), and the group decision is progressively refined to collaboratively solve the continuous sequence prediction problem. **As the coalition size increases, the resulting predictions become progressively more precise and reliable.** In Section 6, we conduct an ablation study to numerically verify these theoretical findings.

## 5 RELATED WORK

**Neural Differential Equation Models.** In recent years, neural differential equation models have gained attention for their ability to capture the dynamics of complex continuous sequences. Latent ODEs (Rubanova et al., 2019) extend standard RNNs to handle continuous signals by integrating neural ODEs with them. Kidger et al. (2020) introduced differential equation models based on controlled differential equations (Neural CDE) to address a key limitation of neural ODEs, where solutions depend solely on initial conditions and not on subsequent observations. Recently, Contiformer (Chen et al., 2024) was developed, combining neural ODEs and Transformers into a single framework. Another line of research integrates stochasticity by utilizing SDEs, particularly for time-series applications. For instance, Latent SDE (Li et al., 2020) encodes sequential data in the latent space using neural SDEs, while MaSDE (Park et al., 2023) employs a concept of stochastic differential games to analyze time series. Koshizuka & Sato (2023) proposed a regularized neural SDE based on the Lagrangian Schrödinger bridge, and Oh et al. (2024) introduced three stable types (classes) of neural SDEs: Langevin-type SDE, Linear Noise SDE, and Geometric SDE.



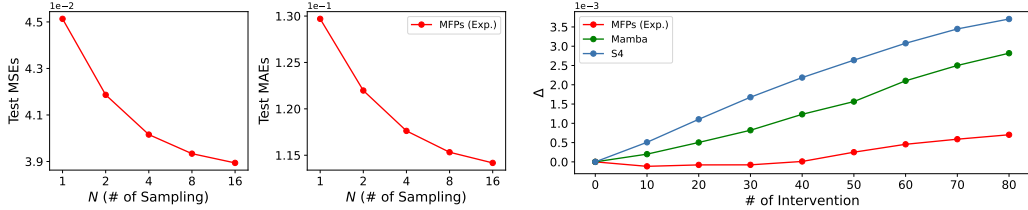


Figure 4: Ablation studies on EigenWorm dataset. **(Left)** Sensitivity analysis on the sample complexity. **(Right)** Robust analysis on post-intervention of non-informative signals.

Methods	MIT Humanoid Robot		MIMIC-II		Beijing Air Quality	
	MSE	MAE	MSE	MAE	MSE	MAE
Neural Laplace	8.11±0.25	17.03±0.33	7.76±0.04	18.70±0.08	3.21±0.12	11.45±0.23
MaSDEs	16.51±0.21	27.89±0.30	8.41±0.06	20.67±0.08	3.47±0.03	13.13±0.07
CRU	32.08±5.07	42.50±3.90	13.09±0.31	24.68±0.47	3.48±0.06	12.76±0.19
Latent SDE	6.01±0.14	15.94±0.14	8.04±0.02	19.63±0.06	3.29±0.03	11.99±0.07
Neural LSDE	6.80±0.14	16.51±0.08	7.93±0.05	19.09±0.07	3.74±0.04	11.98±0.15
CONTIME	6.88±0.29	16.60±0.25	12.29±0.14	25.26±0.12	5.15±0.17	15.86±0.27
Contiformer	5.94±0.23	15.29±0.26	7.90±0.12	19.05±0.18	3.25±0.10	11.48±0.16
S4	5.59±0.16	13.98±0.19	13.24±0.01	24.79±0.30	3.95±0.15	12.35±0.17
Mamba	5.21±0.09	13.71±0.15	13.23±0.02	24.76±0.19	3.68±0.14	11.56±0.24
Jamba	5.13±0.13	13.32±0.20	9.71±0.09	21.37±0.06	4.03±0.10	13.04±0.20
MFPs (Exp.)	<b>3.31±0.30</b>	<b>10.12±0.22</b>	<u>7.51±0.08</u>	<b>18.59±0.11</b>	<u>2.98±0.15</u>	<u>10.06±0.31</u>
MFPs (Cosin.)	<u>3.91±0.07</u>	<u>11.43±0.07</u>	<b>7.51±0.06</b>	<u>18.60±0.10</u>	<b>3.13±0.07</b>	<b>11.38±0.08</b>

Table 1: Mean Squared Errors (MSEs) and Mean Absolute Errors (MAEs) in various continuous sequence prediction tasks. The top and second-top scores in each dataset are highlighted in bold and underlined, respectively. Each metric is scaled by  $10^{-2}$ .

**Mean-field Principles in Generative Models.** Recent works utilized the mean-field principle to model the infinitely many random particles in high-dimensional data space, where they interact with each other. In Liu et al. (2022), the Schrödinger bridge was incorporated to address mean-field games in order to approximate data distributions for large populations. Park et al. (2024) introduced the concept of propagation of chaos to generate data structures with exchangeable high cardinality such as 3D point clouds.

## 6 EXPERIMENTAL RESULTS

**Datasets.** In the experiments, we evaluate our results against benchmarks using the following datasets: (i) MIT Humanoid Robot (Li et al., 2024), (ii) MIMIC-II (Silva et al., 2012), (iii) Beijing Air Quality (Zhang et al., 2017), and (iv) EigenWorm (Bagnall et al., 2018). The MIT Humanoid Robot dataset contains the robot’s state trajectories during various activities, such as running, jogging, and stepping in place, with 27 features describing these states. The MIMIC-II dataset, from the PhysioNet Challenge 2012, consists of time series data with 41 features representing the first 48 hours of a patient’s ICU admission (e.g.,  $\text{SaO}_2$  and cholesterol levels). The Beijing Air Quality dataset contains time series data for six air pollution indicators, collected from 12 different locations in Beijing. The EigenWorm dataset comprises six features that characterize worm motion by projecting their shapes onto the six principal eigenworms, providing continuous sequences of 1500 length. To ensure consistent training, we apply either min-max and z-score normalization on each data instance.

**Benchmarks.** Given our focus on continuous sequence modeling, the benchmark baselines consist of various continuous models, including Neural Laplace (Holt et al., 2022), MaSDEs (Park et al., 2023), CRU (Schirmer et al., 2022), Latent SDE (Li et al., 2020), Neural LSDE (Oh et al., 2024), CONTIME (Jhin et al., 2024), and Contiformer (Chen et al., 2024). To further enhance the baselines, we also incorporate continuous state-space models, such as S4 (Gu et al., 2022), Mamba (Gu & Dao, 2024) and Jamba Lieber et al. (2024). Performance evaluation is carried out using mean squared error (MSE) and mean absolute error (MAE) metrics. Each model is executed five times, with the average scores and standard deviations reported.

**Quantitative Results.** The main table presents a performance comparison with benchmark methodologies across three datasets. The results show that the proposed MFPs consistently outperform other benchmarks by significant margins on all datasets. Notably, conventional neural differential equation models perform reasonably well on the MIMIC-II and BAQD datasets, where sequences are



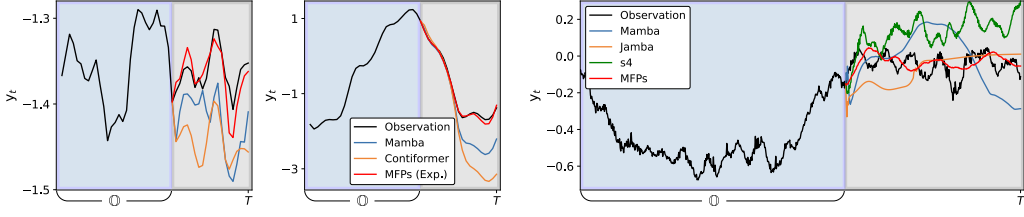


Figure 5: Qualitative results on Prediction (Left) MIT Humanoid Robot dataset (Right) EigenWorm datasets.

irregularly sampled with missing values. However, they exhibit a performance drop on the MIT Humanoid Robot dataset, likely due to their limitations in handling complex spatio-temporal dynamics. In contrast, state-space models excel on the MIT Humanoid Robot dataset but experience a decline in performance on the other two datasets, indicating their limitations in dealing with irregularly sampled sequences. Figure 4 (right) illustrates the qualitative prediction results on the MIT Humanoid Robot dataset. As shown, our MFPs deliver superior performance compared to the other models.

#### Ablation Study I: Long-term Prediction.

The first study aims to demonstrate the efficacy of utilizing mean-field principles for accurate long-term predictions. We compare the MFPs against variants of state-space models, recognized for their capability in managing long-range dependencies Zhan et al. (2024) for long sequences. Table ?? along with Fig. 4 provide both quantitative and qualitative comparisons with benchmarks for long-term prediction tasks using the EigenWorm dataset. Evidently, our MFPs achieve significant advances over other benchmarks.

Methods	EigenWorm	
	MSE	MAE
S4	14.16±0.18	28.69±0.21
Mamba	15.79±1.03	29.80±1.06
Jamba	17.63±1.09	31.96±1.04
MFPs	<b>12.52±0.16</b>	<b>26.61±0.29</b>

Table 2: Long-term prediction on EigenWorm.

**Ablation Study II: Sample Complexity.** To empirically validate the theoretical results derived in Section 4, we conduct an ablation study examining how predictive accuracy scales with the number of sampled predictors. As depicted in Fig. 4 (Left), our results align well with the concentration bounds formalized in Proposition 4.1, confirming that the empirical error diminishes as the coalition size of predictors increases. This behavior is consistent with the propagation-of-chaos property of mean-field systems, whereby larger predictor ensembles more faithfully approximate the infinite-agent limit distribution. In practice, this translates to monotonic improvements in both MSE and MAE as  $N$  increases from 1 to 16, suggesting that additional predictors systematically enhance stability and reliability of forecasts. Nevertheless, these gains must be balanced against the computational overhead introduced during inference.

**Ablation Study III: Noise Robustness.** We perform a robustness study to assess the impact of non-informative noisy signal (*i.e.*, white noise) interventions in past observations. Specifically, we inject the Gaussian random noises with variance  $\sigma_{\text{noise}} = 0.3$  to derive the distributional shift of test continuous-time sequences and corrupt the test data,  $\hat{p}(u, y) = p(u, y) \otimes \mathcal{N}(\mathbf{0}_d, \sigma_{\text{noise}} \mathbf{I}_d)$ , where  $\otimes$  is a convolution operation. Fig 4 (right) shows a uniform performance degradation (*i.e.*,  $\Delta$ ) with an increasing number of past observations corrupted by non-informative noisy signals. As can be seen, our MFPs exhibit robust performance against noise interventions, as Mamba experiences sharp declines in accuracy under high levels of noise. Our MFPs under the coalition, trained on the original clean sequence  $p(u, y)$ , neutralizes the influence of individuals conditioned on noisy signals  $\hat{p}(u, y)$ , thereby preserving the Nash equilibrium, resulting the robust generalization performance.

## 7 CONCLUSION

This paper introduces *mean-field continuous sequence predictors*, a novel class of neural SDE model for the efficient generation of continuous sequences, which can possess infinite-order complexity. We recast the time-series prediction problem as a mean-field game and adopt a fictitious play approach, integrated with a gradient-descent-based method, to leverage the stochastic maximum principle and identify the Nash equilibrium of the system. Both empirical and theoretical results reveal the distinctive features of our MFPs, where the coalition of a continuum of predictors generates accurate predictions and consistently surpasses benchmark performance.

## ACKNOWLEDGEMENT

This work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2026-RS-2020-II201819)

## REFERENCES

- Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018, 2018. URL <https://arxiv.org/abs/1811.00075>.
- Erhan Bayraktar and Ruoyu Wu. Stationarity and uniform in time convergence for the graphon particle system. *Stochastic Processes and their Applications*, 150:532–568, 2022.
- Erhan Bayraktar and Ruoyu Wu. Graphon particle system: Uniform-in-time concentration bounds. *Stochastic Processes and their Applications*, 156(C), 2023.
- Erhan Bayraktar, Suman Chakraborty, and Ruoyu Wu. Graphon mean-field systems. *The Annals of Applied Probability*, 33(5):3587–3619, 2023.
- Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.
- François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.
- François Bolley. Quantitative concentration inequalities on sample path space for mean field interaction. *ESAIM: Probability and Statistics*, pp. 192–209, 2010.
- Amarjit Budhiraja and Wai-Tong Louis Fan. Uniform in time interacting particle approximations for nonlinear equations of patlak-keller-segel type. 2017.
- Peter E Caines and Minyi Huang. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.
- Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.
- Pierre Cardaliaguet and Charles-Albert Lehalle. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12:335–363, 2018.
- Rene Carmona. Applications of mean field games in financial engineering and economic theory. *arXiv preprint arXiv:2012.05237*, 2020.
- René Carmona and François Delarue. Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, 51(4):2705–2734, 2013.
- René Carmona, François Delarue, et al. *Probabilistic theory of mean field games with applications I-II*. Springer, 2018.
- Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. i. models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022.
- Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.
- Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2019.
- Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling, 2024.
- Samuel N Cohen, Christoph Reisinger, and Sheng Wang. Arbitrage-free neural-sde market models. *Applied Mathematical Finance*, 30(1):1–46, 2023.
- Christian Fabian, Kai Cui, and Heinz Koepl. Learning sparse graphon mean field games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4486–4514. PMLR, 2023.

- Olivier D Faugeras, Jonathan D Touboul, and Bruno Cessac. A constructive mean-field analysis of multi population neural networks with random synaptic weights and stochastic inputs. *Frontiers in computational neuroscience*, 3:323, 2009.
- Shuang Gao and Peter E Caines. Spectral representations of graphons in very large network systems control. In *2019 IEEE 58th conference on decision and Control (CDC)*, pp. 5068–5075. IEEE, 2019.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.
- Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.
- Samuel Holt, Zhaozhi Qian, and Mihaela van der Schaar. Neural laplace: Learning diverse classes of differential equations in the laplace domain, 2022.
- Valerii Iakovlev, Markus Heinonen, and Harri Lähdesmäki. Learning space-time continuous latent neural pdes from partially observed states. *Advances in Neural Information Processing Systems*, 36, 2024.
- Sheo Yon Jhin, Seojin Kim, and Noseong Park. Addressing prediction delays in time series forecasting: A continuous gru approach with derivative regularization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 33 of *KDD '24*, pp. 1234–1245. ACM, August 2024. doi: 10.1145/3637528.3671969.
- Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series, 2020.
- Takeshi Koshizuka and Issei Sato. Neural lagrangian schrödinger bridge: Diffusion modeling for population dynamics, 2023.
- Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2 (1):229–260, 2007.
- Mathieu Lauriere. Numerical methods for mean field games and mean field type control. *Mean field games*, 78(221-282), 2021.
- Chenhao Li, Elijah Stanger-Jones, Steve Heim, and Sangbae Kim. Fld: Fourier latent dynamics for structured motion representation and learning, 2024.
- Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations, 2020.
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model, 2024. URL <https://arxiv.org/abs/2403.19887>.
- Guan-Hong Liu, Tianrong Chen, Oswin So, and Evangelos A. Theodorou. Deep generalized schrödinger bridge, 2022.
- Ming Min and Ruimeng Hu. Signed deep fictitious play for mean field games with common noise. In *International Conference on Machine Learning*, pp. 7736–7747. PMLR, 2021.
- Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.
- John W Negele. The mean-field theory of nuclear structure and dynamics. *Reviews of Modern Physics*, 54(4):913, 1982.
- YongKyung Oh, Dongyoung Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data, 2024.

- Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.
- Sungwoo Park, Byoungwoo Park, Moontae Lee, and Changhee Lee. Neural stochastic differential games for time-series analysis. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27269–27293. PMLR, 23–29 Jul 2023.
- Sungwoo Park, Dongjun Kim, and Ahmed Alaa. Mean-field chaos diffusion models, 2024.
- William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.
- Christoph Reisinger and Wolfgang Stockinger. An adaptive euler–maruyama scheme for mckean–vlasov sdes with super-linear growth and application to the mean-field fitzhugh–nagumo model. *Journal of Computational and Applied Mathematics*, 400:113725, 2022.
- Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent odes for irregularly-sampled time series, 2019.
- Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units, 2022.
- Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting In-Hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012. *Comput Cardiol (2010)*, 39:245–248, 2012.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit, 2019.
- Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5(Jun):669–695, 2004.
- Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.
- Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 2012.
- Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, and Yanzhi Wang. Rethinking token reduction for state space models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proc Math Phys Eng Sci*, 473(2205):20170457, September 2017.

## A APPENDIX

### A.1 NOTATION TABLE

For convenience and to improve readability, Table 3 summarizes the main symbols used throughout the paper, together with brief descriptions of their roles in the mean-field formulation and learning algorithm.

Table 3: Summary of frequently used notation.

Symbol	Meaning
$\mathbb{T} = [0, T]$	Global time horizon of a sequence instance
$\mathcal{O} \subset \mathbb{T}$	Past observation interval used as input
$u$	Time label for past observations, $u \in \mathcal{O}$
$t$	Prediction time index, $t \in \mathbb{T}$
$p(u)$	Distribution of observation times on $\mathcal{O}$
$p(u, y)$	Joint distribution of times and observed values
$y_u, y_t$	Observed value at time $u$ or $t$
$d$	Data dimension of each observation
$X_u^\alpha(t)$	Mean field predictor at $(u, t)$ under control $\alpha$
$\nu_u(t)$	Law of $X_u^\alpha(t)$ at time $t$ for label $u$
$\nu^\alpha$	Flow of laws $\{\nu_u(t)\}_{u,t}$ induced by control $\alpha$
$\hat{\mu}_t$	Mean field limit obtained by averaging over $u$ at time $t$
$N$	Number of sampled predictors used in the finite system
$\nu_t^N$	Empirical law of $N$ predictors at time $t$
$\mathcal{P}_2(\mathbb{R}^d)$	Probability measures with finite second moment on $\mathbb{R}^d$
$W_\alpha(u, v)$	Neural graphon weight between labels $u$ and $v$
$\psi_\alpha(y, x)$	Feature interaction function between states $y$ and $x$
$b(t, x, \alpha)$	Drift of the controlled SDE at $(t, x)$
$\sigma_t$	Diffusion scale of the SDE at time $t$
$W_t$	Brownian motion driving the stochastic dynamics
$\alpha(t, x; \theta)$	Feedback control policy parameterized by $\theta$
$\theta$	Trainable neural network parameters
$\mathcal{A}$	Set of admissible control policies
$\Lambda_t$	Aggregated prediction obtained by averaging over labels $u$
$J(\nu^\alpha, \alpha)$	Cost functional of the mean field control problem
$V(t, x)$	Value function associated with the control problem
$H(t, x, a, \nu, \alpha)$	Stochastic Hamiltonian in the Pontryagin principle
$W_2(\cdot, \cdot)$	2-Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^d)$
$W_{t,\mathcal{M}}(\cdot, \cdot)$	Generalized Wasserstein distance between flows of laws
$\Phi$	Projector operator that updates the state law forward in time
$\Psi$	Updater operator that refines the control via FBSDE
$m$	Fictitious play or gradient descent iteration index

## A.2 MATHEMATICAL BACKGROUNDS AND DEFINITIONS

This section includes brief summary of the mathematical backgrounds, omitted notations and definitions in the manuscript. Throughout, **bold-face** notation will be employed without loss of generality to omit subscript and superscript indices of mathematical entities where suitable for the sake of simplicity.

**Generalized Wasserstein Distance.** Recall the definition of space for probability measures that consist of generic path measures with finite second moments,

$$\begin{aligned}\hat{\mathcal{M}} &:= \{\boldsymbol{\nu} = (\nu_u : u \in \mathbb{O}) \in [C([0, T], \mathbb{R}^d)]^{\mathbb{O}}; u \mapsto \nu_u \in \mathcal{P}(C([0, T], \mathbb{R}^d)) \text{ is measurable}\}, \\ \tilde{\mathcal{M}} &:= \{\boldsymbol{\nu}; \sup_{u \in \mathbb{O}} \int \|\mathbf{X}_u(\cdot)\|^2 d\nu_u(\mathbf{X}_u(\cdot)) < \infty\}.\end{aligned}$$

For the arbitrary elements  $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{M} := \hat{\mathcal{M}} \cap \tilde{\mathcal{M}}$ , let us consider  $\mathcal{M}$  equipped with the generalized 2-Wasserstein metric as

$$\mathcal{W}_{t, \mathcal{M}}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \sup_{u \in \mathbb{O}} \left[ \inf_{\Pi} \mathbb{E} \left( \sup_{s \leq t} \|\mathbf{X}_u(s) - \mathbf{Y}_u(s)\|^2 \right) \right]^{1/2}, \quad \begin{cases} \mathbf{Law}(\mathbf{X}_u) = P_u^{-1} \circ \boldsymbol{\mu}, \\ \mathbf{Law}(\mathbf{Y}_u) = P_u^{-1} \circ \boldsymbol{\nu}, \end{cases} \quad (8)$$

where  $\Pi$  is a coupling between two probability measures and  $P_u$  denotes a canonical projection onto the interval  $\mathbb{O}$ . Followed by the Kantorovich-Rubinstein duality, definition in Eq (8) can be further modified as

$$L\mathcal{W}_{t, \mathcal{M}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \geq \sup_{u \in \mathbb{O}} \sup_{f \in \mathbf{Lip}(L)} \left| \int_{\mathbb{R}^d} f d(\mu_{u, t} - \nu_{u, t}) \right|, \quad \boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{M}. \quad (9)$$

Note that the inner supremum is taken over a family of  $L$ -Lipschitz real-valued continuous functions.

**Cut Norm of Graphon.** The cut-norm measures the *discrepancy* between two graphons over all possible cuts of the square of  $\mathbb{O}$ . Formally, for a graphon  $W : \mathbb{O} \times \mathbb{O} \rightarrow \mathbb{R}^+$ , the cut-norm is defined as:

$$\|W\|_{\mathbf{g}} := \sup_{A, B \subset \mathbb{O}} \left| \int_{A \times B} W(u, v) du dv \right|, \quad (10)$$

where the supremum is taken over all measurable subsets  $A$  and  $B$ . The definition illustrates that the cut-norm quantifies the maximum deviation of  $W$  from zero over any rectangle  $\mathbb{O}^2$ . Given the definition, one defines the metric called *cut distance*:

$$d_{\mathbf{g}}^q(W_1, W_2) = \|W_1 - W_2\|_{\mathbf{g}}^q \quad (11)$$

The cut distance measures how close two graphons are after optimally aligning their domains. If the cut distance between two graphons  $W_1$  and  $W_2$  is small, the graphs they represent are structurally similar.

**(Exponential AM-GM Inequality).** For the arbitrary random variables  $X, Y$  and positive constants  $a, b > 0$ , the expectation can be decomposed as follows:

$$\mathbb{E}[\exp(aX^2 + bY^2)] \leq (2\mathbb{E}[\exp(2X^2)])^{1/2} (2\mathbb{E}[\exp(2Y^2)])^{1/2}. \quad (12)$$

**(Arithmetic AM-GM Inequality).** For arbitrary positive constants  $x, y, w > 0$ , we have

$$xy \leq \omega x + \frac{1}{4\omega} y. \quad (13)$$



### A.3 ASSUMPTION

Without additional information, we make the following assumptions in this paper.

1. **(H1)**. There exists a finite collection of intervals  $\{O_k; k \in \{1, \dots, N\}\}$  for arbitrary  $N \in \mathbb{N}^+$  such that  $\cup_k^N O_k = \mathbb{O}$ . Then we assume the following:
  - For each  $k$ , the initial datum of the graphon system is set with the data distribution  $\nu_u$ :  $O_k \ni u \mapsto \mu_u(0) := \nu_u \in \mathcal{P}_2$ , where the mapping assigns to independent measures.
2. **(H2)**. For each  $k$  and  $O_k \ni u$ , there exists a constant  $C_1$  such that we have probability  $\nu_{u,s}[\sup_{x \in \mathbb{R}^d} \|\mathbf{Y}(\omega)\| \|x - \mathbf{Y}\|^{-p} \leq C_1]$  almost surely for all  $p \in \mathbb{N}^+$ , and the second moment (*i.e.*,  $m_2$ ) of  $\nu_{u,s}$  is bounded.
3. **(H3)**. The Lipschitz constants of the functions in modeling of graphons  $W_{(\cdot)} : \mathbb{L}_2(\nu_u(t)) \supset \mathbb{A} \rightarrow \mathbb{R}^+$  are bounded above. The parameterized Markovian feedback controls are Lipschitz in parameters:

$$|W_{(\cdot)}(\alpha) - W_{(\cdot)}(\beta)| \leq \mathbf{Lip}_W \|\alpha - \beta\|_{\nu_u(t)}, \quad (14)$$

$$\{\|\alpha - \beta\|_{\nu_u(t)}, \|\alpha(t, x, \theta_\alpha) - \alpha(t, x, \theta_\beta)\|\} \leq \mathbf{Lip}_\theta \|\theta_\alpha - \theta_\beta\|_E \quad (15)$$

The drift function is Lipschitz continuous and dissipative, ensuring that the constant  $c_1$  is well-defined.

$$\|(\mathbf{b}, \mathbf{b}_W)(t, x, \alpha) - (\mathbf{b}, \mathbf{b}_W)(t, y, \beta)\| \leq \mathbf{Lip}_b (\|x - y\|_E + \|\alpha - \beta\|_{\nu_u(t)}). \quad (16)$$

$$c_1 := \inf_{x, y} -(x - y) \cdot [(\mathbf{b}, \mathbf{b}_W)(x) - (\mathbf{b}, \mathbf{b}_W)(y)] / \|x - y\|_E^2 \quad (17)$$

4. **(H4)**. The maximal rank of embedding of neural agents in  $\mathbb{A}$  is  $d'$ .

$$\mathbb{T} \times \mathbb{R}^d \times \Theta \mapsto \alpha \in \mathbb{A} \hookrightarrow \mathbb{L}_2(\nu). \quad (18)$$

### A.4 PROOFS

#### A.4.1 STOCHASTIC OPTIMAL CONTROL, MEAN-FIELD FBSDEs

Before presenting the main proofs, this section offers a detailed analysis of how the proposed mean-field games can be formulated.

**Weak Formulation of Mean-field Games.** We start by explicating on the rigorous definition of forward mean-field dynamic in Eq. (1) cost functional in Eq. (3) and gradient system of FBSDEs in Propsoition A.2, followed by a brief summary of how forward-backward SDEs are formulated in the context of stochastic optimal control problems. To this end, let us first define the primitive process  $\bar{\mathbf{X}}_t$ , which solves the following SDE for a fixed label  $u$ :

$$d\bar{\mathbf{X}}_u(t) = \sigma_t dB_t^u, \quad \bar{\mathbf{X}}_0(t) = y_t. \quad (19)$$

where  $B_t^u$  is a Brownian motion under probability measure  $\mathbb{P}$ . Given the square of volatility term  $\sigma_t^2$  is bounded below some constant, we introduce the probability measure  $\mathbb{P}^{\mu, \alpha}$ , which can be derived by the following Radon-Nikodym derivative:

$$\frac{d\mathbb{P}^{\mu, \alpha}}{d\mathbb{P}} = \mathcal{E} \left( \int_0^{(\cdot)} \sigma_t^{-1} (\mathbf{b}_W(\bar{\mathbf{X}}_u(t), \nu, \alpha) + \mathbf{b}(t, \bar{\mathbf{X}}_u(t), \alpha)) \cdot dB_t^u \right) \Big|_{t=T}. \quad (20)$$

where  $\mathcal{E}$  denotes a Doléans-Dade exponential of a martingale. Applying Girsanov's theorem, we have the Brownian motion  $W^{\mu, \alpha}$  under the probability measure  $\mathbb{P}^{\mu, \alpha}$ :

$$W_t^{\mu, \alpha} = B_t^u - \int_{\mathbb{T}} \sigma_s^{-1} (\mathbf{b}_W(\bar{\mathbf{X}}_u(s), \nu, \alpha) + \mathbf{b}(s, \bar{\mathbf{X}}_u(s), \alpha)) ds. \quad (21)$$

Then, the primitive process can be rewritten as follows almost surely  $\mathbb{P}^{\mu, \alpha}$ ,

$$d\bar{\mathbf{X}}_u(t) = (\mathbf{b}_W(\bar{\mathbf{X}}_u(t), \nu, \alpha) + \mathbf{b}(t, \bar{\mathbf{X}}_u(t), \alpha)) dt + \sigma_t dW_t^{\mu, \alpha}. \quad (22)$$

By suppressing the objects in upper-scripts for simplicity, with the notation  $W_t^u = W_t^{u,\mu,\alpha}$  and  $\mathbf{X}_u(t) = \bar{\mathbf{X}}_u(t)$ , one can recover the original mean-field forward SDE defined in Eq (1). Note that this formulation reveals that the process  $\bar{\mathbf{X}}_u(t)$  is a weak solution under  $\mathbb{P}^{\mu,\alpha}$ , and the cost functional can be posed as follows:

$$\mathcal{J}(\nu^\alpha, \alpha) = \int_{\mathbb{T}} \mathbb{E}_{\alpha, \nu} [\|\mathbb{E}_{u \sim p(u)} \mathbf{X}_u^\alpha(t) - y_t\|_E^2 + \mathbf{G}(\mathbf{X}_u^\alpha(T), \nu^\alpha)] dt, \quad (23)$$

where the expectation  $\mathbb{E}_{\alpha, \nu}$  is taken with respect to  $\mathbb{P}^{\mu,\alpha}$ . Note that the cost functional in Eq. (3) is an alternative form of Eq. (23). Next, we reformulate the approximation of mean-field games with graphon in the probabilistic sense. Let  $\alpha = \alpha(t, x; \theta) := \hat{\alpha}(t, x, \bar{\mu}, \bar{\epsilon}) := \hat{\alpha}$  be an extended control with fixed arguments  $\bar{\mu}, \bar{\epsilon}$ . For the fixed  $\nu_u^{\alpha^*}$  a.e.,  $u \sim \nu_{\text{Unif}}$  associated with the optimal control  $\hat{\alpha}^*$ , let us consider a Hamiltonian-Jacobi-Bellman equation (HJBE), having a classical value function  $\mathcal{V}$ :

$$\partial_t \mathcal{V}(t, x) + \frac{1}{2} \text{Tr}[\sigma_t^2 \partial_{xx}^2 \mathcal{V}(t, x)] + H\left(t, x, \nu_u^{\alpha^*}, \partial_x \mathcal{V}(t, x), \hat{\alpha}^*(t, x, \nu_u^{\alpha^*}, \partial_x \mathcal{V}(t, x))\right) = 0, \quad (24)$$

Then, forward-backward SDEs associated with the Hamiltonian system in (24) can be described in the Proposition A.1:

**Proposition A.1.** (*Weak Formulation: Forward-Backward SDEs I*) (Carmona & Delarue, 2013) For the fixed flow of measures  $\nu_u(\cdot) : \mathbb{T} \rightarrow \mathcal{P}_2$  and the fixed label  $u$ , let  $(\mathbf{X}_u(t), \mathbf{Y}_u(t), \mathbf{Z}_u(t))$  be a family of processes that solves forward-backward stochastic differential equations with respect to the proposed graphon system in Eq (1) given as follows:

$$d\mathbf{X}_u(t) = (\mathbf{b}_W(\mathbf{X}_u(t), \nu_u, \hat{\alpha}^*) + \mathbf{b}(t, \mathbf{X}_u(t), \hat{\alpha}^*)) dt + \sigma_t dW_t^u, \quad (25)$$

$$d\mathbf{Y}_u(t) = -H(t, \mathbf{X}_u(t), \mathbf{Y}_u(t), \nu_u, \hat{\alpha}^*) dt + \mathbf{Z}_u(t) \cdot dW_t^u. \quad (26)$$

where  $\mathbf{b}_W(\mathbf{x}, \nu, \alpha) := \langle \mathbb{W}_\alpha[\nu](u), \psi \rangle(\mathbf{x}, \alpha)$  is the graphon interaction term, and terminal constraint is given as  $\mathbf{Y}_u(T) = \mathbf{G}(\mathbf{X}_T, \nu_T)$ . Then, under the mild assumption (e.g., smooth boundness of  $\partial_x \mathcal{V}$  and  $\partial_{xx} \mathcal{V}$ ), there exist solutions of stochastic optimal control of the following minimization problem:

$$\inf_{\alpha \in \mathbb{A}} \mathcal{J}(\nu^\alpha, \alpha) = \mathbf{Y}_u(0). \quad (27)$$

For the closed Markovian control such as neural control introduced in Section 2, the solution to adjoint process  $\mathbf{Z}_u(t)$  can be defined as stated in Definition A.2. By rewriting forward-backward SDEs in Eq (25) and Eq (26) for non-optimal neural controls  $\alpha$  (i.e., neural networks) which are updated via gradient descent, we can recover the proposed gradient system of FBSDEs in Definition 3.2.

**Definition A.2.** (*Gradient System of FBSDEs*). For the fixed flow of measures  $\nu_u(\cdot) : \mathbb{T} \rightarrow \mathcal{P}_2$  and the fixed label  $u$  at each stage  $m$ , we consider a family of processes  $(\mathbf{X}_u(t), \mathbf{Y}_u(t), \mathbf{Z}_u(t))$  that solves forward-backward stochastic differential equations with respect to the proposed graphon system in Eq (1) given as follows:

$$d\mathbf{X}_u^{m, \alpha_m}(t) = \mathbf{b}_W(\mathbf{X}_u^{m, \alpha_m}(t), \nu_u, \alpha_m) dt + \mathbf{b}(t, \mathbf{X}_u^{m, \alpha_m}(t), \alpha_m) dt + \sigma_t dW_t^u, \quad (28)$$

$$d\mathbf{Y}_u^{m, \alpha_m}(t) = -H(t, \mathbf{X}_u^{m, \alpha_m}(t), \mathbf{Y}_u^{m, \alpha_m}(t), \nu_u, \alpha_m) dt - \mathbf{Z}_t^m \cdot dW_t^u, \quad (29)$$

$$\alpha_{m+1} := \alpha(t, \mathbf{X}_u^{m, \alpha_m}; \theta^m - \mathbb{E}_{\mathbf{Y}, t \leq T} [\gamma^m \nabla_\theta \mathbf{Y}_u^{m, \alpha_m}(t)]) \in \mathbb{A}, \quad (30)$$

$$\nu_u = \text{Law}(\mathbf{X}_u^{m-1, \alpha_{m-1}^*}), \quad (31)$$

where  $\gamma^m > 0$  is a learning rate of gradient descent, and  $\mathbb{A}$  is a set of admissible neural agents. Then, we have  $(\mathbf{Y}_u(t), \mathbf{Y}_u(T), \mathbf{Z}_u(t)) = (\mathcal{J}, \mathbf{G}, (\partial_x \mathcal{J}) \sigma_t^{-1})$ .

## A.4.2 ANALYSIS ON STOCHASTIC OPTIMALITY AND CONVERGENCE

**Stochastic Optimality.** In the following, we introduce the second type of forward-backward SDEs, which is based on the principles of stochastic maximum principle:

**Proposition A.3.** (*Stochastic Maximum Principle: Forward-Backward SDEs II*) (Bensoussan et al., 2013) For the fixed flow of measures  $\nu_u(\cdot) : \mathbb{T} \rightarrow \mathcal{P}_2$  and the fixed label  $u$ , let  $(\mathbf{X}_u(t), \mathbf{Y}_u^{\text{MP}}(t), \mathbf{Z}_u^{\text{MP}}(t))$  be a family of processes that solves forward-backward stochastic differential equations with respect to the proposed graphon system in Eq (1) given as follows:

$$\begin{aligned} d\mathbf{X}_u(t) &= (\mathbf{b}_W(\mathbf{X}_u(t), \nu_u, \hat{\alpha}^*) + \mathbf{b}(t, \mathbf{X}_u(t), \hat{\alpha}^*)) dt + \sigma_t dW_t^u, \\ d\mathbf{Y}_u^{\text{MP}}(t) &= -\partial_x H(t, \mathbf{X}_u(t), \mathbf{Y}_u^{\text{MP}}(t), \nu_u, \hat{\alpha}^*) dt + \mathbf{Z}_t^{\text{MP}} \cdot dW_t^u. \end{aligned}$$

For the progressively measurable admissible Markovian neural control  $\beta$  under the mild assumption (e.g., smooth boundness of  $\partial_x \mathcal{V}$  and  $\partial_{xx} \mathcal{V}$ ), there exists a constant  $\tau_{\text{SMP}} > 0$  such that the following inequality holds:

$$\mathcal{J}(\nu^{\hat{\alpha}^*}, \hat{\alpha}^*) + \tau_{\text{SMP}} \int_{\mathbb{T}} \|\hat{\alpha}^* - \beta\|_{\nu} dt \leq \mathcal{J}(\nu^{\beta}, \beta). \quad (32)$$

*Remark.* Note that the backward dynamics  $\mathbf{Y}_u^{\text{MP}}$  differs from the original backward dynamics  $\mathbf{Y}_u$  in Definition (A.2) as the dynamics is designed to be associated with *Pontryagin stochastic maximum principle*. This principle plays a central role in the proof of Proposition A.4, demonstrating the stochastic optimality of neural agents in the following section.

In what it follows, we demonstrate that the stochastic optimality of the proposed gradient system can be guaranteed under the specific conditions required for constructing the control set in Prop A.4.

**Proposition A.4.** (*Maximum Principle of Graphon Mean-field System*) Assume that there exists a constant  $K_H$  such that  $\|\partial_{\alpha} \mathbf{H}\|_{E\|_{\infty, \nu}} \leq K_H$ . Then, there exists a convex set of admissible neural agents  $\alpha_m \in \mathbb{A}$  such that the following relation holds:

$$D_{\alpha} \mathcal{J}(\nu^{\alpha_m}, \alpha_m) := \lim_{\varepsilon \rightarrow 0} \frac{d}{d\varepsilon} \mathcal{J}[\alpha_m + \varepsilon(\alpha_m - \alpha_{m-1})] \xrightarrow{m \rightarrow \infty} 0. \quad (33)$$

Furthermore, the sequence of control profile  $\{\alpha_m\}$  leads to the minimization of the stochastic Hamiltonian system in terms of *Pontryagin maximum principle*:

$$\lim_{m \rightarrow \infty} H(t, \mathbf{X}_u^m(t), \mathbf{Y}_u^{m, \text{MP}}(t), \nu_u, \alpha_m) = \inf_{\alpha \in \mathbb{A}} H(t, \mathbf{X}_u(t), \mathbf{Y}_u^{\text{MP}}(t), \nu_u, \alpha), \quad dt \otimes d\mathbb{P} - a.e. \quad (34)$$

where the population is set to  $\nu_u = \Psi(\nu^{\alpha_{m-1}}) := \Psi(\nu^{\alpha_{m-1}})$ . In other words, the value function can be derived by the proposed gradient system of FBSDEs:

$$\mathcal{V} := \inf_{\alpha \in \mathbb{A}} \mathcal{J}(\nu^{\alpha}, \alpha) = \lim_{m \rightarrow \infty} \mathcal{J}(\nu^{\alpha_m}, \alpha_m). \quad (35)$$

*Proof.* We divide the proof into two separate steps.

**1. Computation of Gâteaux derivative  $D_{\alpha} \mathcal{J}$ .** The aim of the first step is to provide an explicit computation of the Gâteaux derivative of cost functional (value function) with respect to the neural agent. To achieve this, we introduce the variation equation  $i_u$  and its associated gradient system of SDEs with fixed  $\beta$ :

$$d\mathbf{Y}_u^{m, \text{MP}}(t) = -\partial_x H(t, \mathbf{X}_u^m(t), \mathbf{Y}_u^{m, \text{MP}}(t), \nu_u, \hat{\alpha}_m) dt + \mathbf{Z}_t^{m, \text{MP}} \cdot dW_t^u, \quad (36)$$

$$di_u(t) = [(\partial_x \mathbf{b}_W + \partial_x \mathbf{b})i_u(t)] dt + [(\partial_{\alpha} \mathbf{b}_W + \partial_{\alpha} \mathbf{b})\beta_m] dt, \quad (37)$$

$$dj_u(t) := d[i_u(t) \cdot \mathbf{Y}_u^{m, \text{MP}}(t)] dt \in \mathbb{R}^d. \quad (38)$$

Let  $\Upsilon_{\alpha}(m, \epsilon) := \alpha_m + \epsilon \beta_m$  represent the infinitesimal changes of the admissible neural agent  $\alpha^m$  in the direction of  $\beta_m := \alpha_{m-1} - \alpha_m$ . To feasibly select the convex combination  $\Upsilon_{\alpha}(m, \epsilon)$  for any

$m$  and  $\epsilon \in [0, 1]$ , both neural agents need to lie within some convex set  $\mathbb{A}_m$ . For now, we assume that there exists a convex set  $\mathbb{A}_m$  that includes  $\alpha$  and  $\beta$ . The explicit form of this set will be clarified in the subsequent step. Given the definition, we compute the derivative as follows:

$$\begin{aligned} D_{\alpha} \mathcal{J}(\nu^{\alpha_m}, \alpha_m) &= \frac{d}{d\epsilon} \mathcal{J}(\nu^{\Upsilon_{\alpha}(m, \epsilon)}, \Upsilon_{\alpha}(m, \epsilon))|_{\epsilon=0} \\ &= \mathbb{E} \left[ \int_{\mathbb{T}} [\mathbf{i}_u(t) \partial_x f + \beta_m \partial_{\alpha} f] dt + \mathbf{i}_u(T) \partial_x \mathbf{G} \right], \end{aligned} \quad (39)$$

where we denote  $f(t, x, \alpha) = \|\mathbb{E}_u[x^{\alpha}(t)] - y_t\|^2$ . While  $\mathbf{i}_u(T) \partial_x \mathbf{G}$  can be identified with  $\mathbf{j}_u(T)$ , we apply the product rule to the third dynamics  $d\mathbf{j}_u$  in Eq (38) to have variational form to induce  $\mathbf{j}_u(T)$ :

$$\begin{aligned} d\mathbf{j}_u(t) &= [\mathbf{Y}_u(t) \cdot d\mathbf{i}_u(t)]dt + [\mathbf{i}_u(t) \cdot d\mathbf{Y}_u^{\text{MP}}(t)]dt + \text{Tr}[d\mathbf{Y}_u^{\text{MP}}(t) \otimes d\mathbf{i}_u(t)] \\ &= \int_0^T \mathbf{Y}_u^{\text{MP}}(t) \cdot (\partial_x \mathbf{b}_W + \partial_x \mathbf{b}) \beta_m + \mathbf{Y}_u^{\text{MP}}(t) \cdot (\partial_{\alpha} \mathbf{b}_W + \partial_{\alpha} \mathbf{b}) \mathbf{i}_u(t) dt \\ &= \int_0^T \partial_x \mathbf{G} \cdot (\partial_x \mathbf{b}_W + \partial_x \mathbf{b}) \beta_m + \partial_x \mathbf{G} \cdot (\partial_{\alpha} \mathbf{b}_W + \partial_{\alpha} \mathbf{b}) \mathbf{i}_u(t) dt. \end{aligned} \quad (40)$$

Combining Eq (39) with Eq (40), and Cauchy–Schwarz inequality gives explicit form for the Gâteaux derivative of objective functional.

$$\begin{aligned} D_{\alpha} \mathcal{J}(\Upsilon_{\alpha}(m, \epsilon)) &= \mathbb{E} \left[ \int_{\mathbb{T}} \partial_{\alpha} H(t, \mathbf{X}_u^m, \mathbf{Y}_u^{m, \text{MP}}(t), \Psi(\nu^{\alpha_{m-1}}), \alpha_m) dt \cdot \beta_m \right] \\ &\leq \mathbb{E} \left[ \int_{\mathbb{T}} \|\partial_{\alpha} H(t, \mathbf{X}_u^m, \mathbf{Y}_u^{m, \text{MP}}(t), \Psi(\nu^{\alpha_{m-1}}), \alpha_m)\|_E \cdot \|\beta_m\|_E dt \right] \\ &\leq \|\|\partial_{\alpha} \mathbf{H}^m\|_E\|_{\infty} \cdot \|\|\beta_m\|_E\|_1, \end{aligned} \quad (41)$$

where  $\|\cdot\|_p$  denotes  $L_p$ -norm, and the last inequality is obtained by applying Hölder’s inequality with the conjugate pair  $(p = \infty, q = 1)$ . Then, we have

$$\begin{aligned} \|\|\beta_m\|_E\|_1 &= \|\|\alpha_m - \alpha_{m-1}\|_E\|_1 := \|\|\alpha(t, \mathbf{X}_u^m, \theta^m) - \alpha(t, \mathbf{X}_u^m, \theta^{m-1})\|_E\|_1 \\ &\leq \gamma^{m-1} \text{Lip}_{\alpha} \mathbb{E} \delta_{\theta} \mathbf{Y}^{m-1}, \quad \delta_{\theta} \mathbf{Y}^{m-1} := \|\nabla_{\theta} \mathbf{Y}_u^{m-1, \alpha_{m-1}}(t)\|_E. \end{aligned} \quad (42)$$

**2. Construction of  $\mathbb{A}$ .** Next, we define the explicit form of the control set  $\mathbb{A}_m$ . The constructed control set must meet two conditions: (1) it must be convex, and (2) the right-hand side of the inequality in Eq (41) must converge. For properly dealing with the first condition, let us consider a metric ball  $\mathbf{B}_m$  in  $L_1$  space as follows:

$$\mathbf{B}_m := B(\alpha_{m-1}, \mathbf{r}_m) \in \mathbb{L}_1, \quad (43)$$

$$\mathbf{r}_m := r_{u, t, m} = \varepsilon \gamma^{m-1} \text{Lip}_{\alpha} \delta_{\theta} \mathbf{Y}_u^{m-1}(t), \quad \varepsilon \in [0, 1]. \quad (44)$$

Since any arbitrary metric ball is convex and the calculated reverse direction of gradient guarantees local minimum at each stage, the setup of the proposed metric ball ensures the well-definedness of Gâteaux derivative in Eq (41) and local optimality at each stage  $m$ .

Let  $\lambda_{\max}^m(\alpha)$  be an eigenvalue with respect to the principal direction of Hessian for cost functional, *i.e.*,  $\text{Hess}_{\theta} \mathcal{J}(\nu^{\alpha}, \alpha(\cdot; \theta))$ . Consider another control set  $\mathbb{C}_m := \{\alpha_{m-1}; \lambda_{\max}^{m-1}(\alpha_{m-1}) \leq (\gamma^{m-1})^{-1}\}$ . The conventional analysis of gradient descent gives the following inequality on  $\mathbb{C}_m$ :

$$\mathbb{E} \mathbf{Y}^{m, \alpha_m} \leq \mathbb{E} \mathbf{Y}^{m-1, \alpha_{m-1}} - \frac{1}{2} (2\gamma^{m-1} - (\gamma^{m-1})^2 \lambda_{\max}^{m-1}(\alpha_{m-1})) (\mathbb{E} \delta_{\theta} \mathbf{Y}^{m-1})^2. \quad (45)$$

While the second term in right-hand side of Eq (45) is non-negative, the sequence of expectations for the backward dynamics is non-increasing, demonstrating that  $\lim_{m \rightarrow \infty} D_{\alpha} \mathcal{J} \leq \lim_{m \rightarrow \infty} \mathbb{E} \delta_{\theta} \mathbf{Y}^{m-1} = 0$  when the infinite sequence  $\{\alpha_m\}$  lies within  $\lim_{m \rightarrow \infty} \mathbb{C}_m$ . To inherit aforementioned properties lying in both control profiles for all  $m$ , we define  $\mathbb{A}_m := \bigsqcup_{m \geq m} (\mathbf{B}_m \cap \mathbb{C}_m)$ , where  $\mathbb{A} = \lim_{m \rightarrow \infty} \mathbb{A}_m$ . The result directly follows from findings in the stochastic maximum

principle (SMP) (Carmona et al., 2018; Bensoussan et al., 2013), ensuring the equivalence of the following relation:

$$\mathbb{E} \partial_{\alpha} \mathbf{H}(\cdot, \alpha^*) \cdot \beta_m = 0 \quad \longleftrightarrow \quad \alpha^* = \arg \inf_{\alpha \in \mathbb{A}} \mathbf{H}(\cdot, \alpha). \quad (46)$$

Note that this equivalence relation is applicable only when  $\mathbb{A}$  is constructed in the manner previously specified.  $\square$

#### A.4.3 CONVERGENCE OF GRADIENT SYSTEM OF FBSDEs, MEAN-FIELD EQUILIBRIUM

As we have formally defined the stochastic optimal control problem and established the corresponding optimality conditions, this section delves into the detailed rationale of how the proposed gradient descent-based FBSDEs achieve the Nash equilibrium. We will prove Proposition 3.5 through the following steps:

1. For the arbitrary probability measures (*i.e.*,  $\mu^\beta, \nu^\alpha$ ) associated with fixed Markovian controls  $\alpha$  and  $\beta$ , we first establish that the upper bounds of the generalized Wasserstein distance remain stable when two measure-valued operators  $\Phi$  and  $\Psi$  are composed repeatedly:

$$\mathcal{W}_{t, \mathcal{M}}([\Phi \circ \Psi]^{om}(\mu^\beta), [\Phi \circ \Psi]^{om}(\nu^\alpha)) \xrightarrow{m \geq M} 0. \quad (47)$$

2. Consequently, we reparameterize reference measures ( $\mu^\beta, \nu^\alpha$ ) with the laws of inferred mean-field forward dynamics in Eq 1 at subsequent stages (*i.e.*,  $\nu^{\alpha^m}, \nu^{\alpha^{m+1}}$ ), proving the convergence towards mean-field Nash equilibrium.

**Proposition 3.5.** *With the assumptions explored in the previous proof, for the fixed label  $u \sim p(u)$ , the  $m$ -fold of composition  $\Phi \circ \Psi$  induces convergent behavior of generalized Wasserstein distance:*

$$\begin{aligned} \mathcal{W}_2([\Phi \circ \Psi]^{om}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{om}(\nu^{\alpha_0}))^2 &\leq \sup_{t \in \mathbb{T}} \mathcal{W}_{t, \mathcal{M}}([\Phi \circ \Psi]^{om}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{om}(\nu^{\alpha_0}))^2 \\ &\leq \lim_{M \rightarrow \infty} \frac{C(T)^M (\sup_t \sup_m \mathbf{r}_m)^M - 1}{C(T) (\sup_t \sup_m \mathbf{r}_m) - 1} + \frac{(C'T)^M}{M!} \sup_{t \in \mathbb{T}} \mathcal{W}_{t, \mathcal{M}}(\nu^{\alpha_1}, \nu^{\alpha_0})^2 \xrightarrow{M \rightarrow \infty} 0. \end{aligned} \quad (48)$$

where a numerical constant  $C$  is dependent on  $b_0, C_1, H_\psi, \mathbf{Lip}_b, \mathbf{m}_2, |\mathbb{O}|, e^{-|\mathbb{O}|}, \mathbf{h}, \mathbf{Lip}_W$ . In other words,  $[\Phi \circ \Psi]^{om}$  is a Cauchy sequence on  $\mathcal{M}$ , and the proposed gradient system converges.

*Proof.* Recall the definition of controlled graphon system that the particle dynamics at time  $t$  with distinctive controls  $\alpha$  and  $\beta$  can be presented as follows:

$$\begin{aligned} \mathbf{X}_u^{\nu, \alpha}(t) &= \mathbf{X}_u^{\nu, \alpha}(0) + \int_0^t \langle \mathbb{W}_\alpha[\nu_{\nu, s}], \psi \rangle (\mathbf{X}_u^\alpha(s)) ds + \int_0^t \mathbf{b}(s, \mathbf{X}_u^\alpha(s), \alpha) ds + \int_0^t \sigma_s dW_s^u, \\ \mathbf{X}_u^{\mu, \beta}(t) &= \mathbf{X}_u^{\mu, \beta}(0) + \int_0^t \langle \mathbb{W}_\beta[\mu_{\nu, s}], \psi \rangle (\mathbf{X}_u^\beta(s)) ds + \int_0^t \mathbf{b}(s, \mathbf{X}_u^\beta(s), \beta) ds + \int_0^t \sigma_s dW_s^u. \end{aligned}$$

Given the dynamics above, the property of measure projection  $\Psi$  induces the upper bound of generalized Wasserstein distance as follows:

$$\begin{aligned} \mathcal{W}_{t, \mathcal{M}}(\Phi(\mu^\beta), \Phi(\nu^\alpha))^2 &\leq \mathbb{E} \left[ \sup_{s \leq t} \|\mathbf{X}_u^{\mu, \beta}(s) - \mathbf{X}_u^{\nu, \alpha}(s)\|^2 \right] \\ &\leq b_0 \mathbb{E} \left[ \int_0^t \int_{\mathbb{O}} \int_{\mathbb{R}^d} \psi(\mathbf{X}_u^{\mu, \beta}(s), \mathbf{Y}) W_\beta(u, v) d\mu_{\nu, s}(\mathbf{Y}) \right. \\ &\quad \left. - \int_{\mathbb{R}^d} \psi(\mathbf{X}_u^{\nu, \alpha}(s), \hat{\mathbf{Y}}) W_\alpha(u, v) d\nu_{\nu, s}(\hat{\mathbf{Y}}) \|^2 d\nu_{\text{Unif}} ds \right] \\ &\quad + b_0 \mathbb{E} \left[ \int_0^t \|\mathbf{b}(s, \mathbf{X}_u^{\mu, \alpha}(s), \alpha) - \mathbf{b}(s, \mathbf{X}_u^{\nu, \beta}(t), \beta)\|^2 ds \right] \\ &\leq 3b_0 (\text{I} + \text{II} + \text{III}) + b_0 \text{IV}, \end{aligned} \quad (49)$$

where the first and second inequalities are induced from Holder's inequality and the Burkholder-Davis-Gundy (Chaintron & Diez, 2022) with some constant  $b_0 > 0$ . Following the assumptions in Section A.3 and the modeling of graphons in Section 2, the first term (*i.e.*, I) can be upper-bounded in the following estimation.

$$\begin{aligned} \text{I} &:= \mathbb{E} \left[ \int_0^t \int_{\mathbb{O}} \left\| \int_{\mathbb{R}^d} [\psi(\mathbf{X}_u^{\nu, \alpha}(s), \hat{\mathbf{Y}}) - \psi(\mathbf{X}_u^{\mu, \beta}(s), \hat{\mathbf{Y}})] W_{\alpha}(u, v) d\nu_{v, s}(\hat{\mathbf{Y}}) \right\|^2 d\nu_{\text{Unif}} ds \right] \\ &\leq \text{Lip}(\psi)^2 \mathbb{E} \left[ \int_0^t \int_{\mathbb{O}} W_{\alpha}^2(u, v) \int_{\mathbb{R}^d} \|\mathbf{X}_u^{\nu, \alpha}(s) - \mathbf{X}_u^{\mu, \beta}(s)\|^2 d\nu_{v, s}(\hat{\mathbf{Y}}) d\nu_{\text{Unif}} ds \right]. \end{aligned} \quad (50)$$

Given the fixed control  $\alpha = \bar{\alpha}$ , optimizing the last inequality requires estimating the (local) Lipschitz continuity of positional encoding  $\psi$ :

$$\begin{aligned} \text{Lip}(\psi(\cdot, \hat{\mathbf{Y}})) &\leq \sup_{\mathbf{x} \in \mathbb{R}^d \setminus \{\hat{\mathbf{Y}}\}} \|\nabla \psi(\mathbf{x}, \hat{\mathbf{Y}})\| \\ &\leq H_{\psi}(\bar{\alpha}) \sup_{\mathbf{x} \in \mathbb{R}^d \setminus \{\hat{\mathbf{Y}}\}} \mathfrak{a}^{-2} \left\| \left( \mathbf{I}_d - \frac{2(\mathbf{x} - \hat{\mathbf{Y}}) \otimes_E (\mathbf{x} - \hat{\mathbf{Y}})}{\mathfrak{a}^2} \right) \right\|. \end{aligned} \quad (51)$$

where  $\mathfrak{a} = \|\mathbf{x} - \hat{\mathbf{Y}}\|$  and  $\otimes_E$  denotes the Euclidean outer product. Following by the assumption (H2), Grönwall's inequality with the fact that  $\text{spec}(\nabla \psi) := \lambda_1 \leq \max(1, -1)\mathfrak{a}^{-2}$ , we have

$$\text{I} \leq C_1^2 H_{\psi}^2(\bar{\alpha}) \mathfrak{h}(\beta) \mathbb{E} \left[ \int_0^t \sup_{r \leq s} \|\mathbf{X}_u^{\nu, \alpha}(r) - \mathbf{X}_u^{\mu, \beta}(r)\|^2 ds \right]. \quad (52)$$

Since each component  $\psi_i$  possesses the same spectral norm as  $\psi$ , the second term can be upper-bounded with the improved definition of generalized Wasserstein distance in Eq (9):

$$\begin{aligned} \text{II} &:= \mathbb{E} \left[ \int_0^t \int_{\mathbb{O}} \left\| \int_{\mathbb{R}^d} \psi(\mathbf{X}_u^{\mu, \beta}(s), \hat{\mathbf{Y}}) W_{\beta}(u, v) d[\nu_{v, s} - \mu_{v, s}](\hat{\mathbf{Y}}) \right\|^2 d\nu_{\text{Unif}} ds \right] \\ &\leq d|\mathbb{O}| C_1^2 \mathbb{E} \left[ \sup_{u \in \mathbb{O}} \max_{i \in \{1, \dots, d\}} \int_0^t \left| \int_{\mathbb{R}^d} \frac{\psi_i}{C_1}(\mathbf{X}_u^{\mu, \beta}(s), \cdot) W_{\beta}(u, v) d[\nu_{v, s} - \mu_{v, s}] \right|^2 ds \right] \\ &\leq d|\mathbb{O}| C_1^2 \mathfrak{h}(\beta) \int_0^t \mathcal{W}_{s, \mathcal{M}}(\mu^{\beta}, \nu^{\alpha})^2 ds. \end{aligned} \quad (53)$$

Regarding the third term (*i.e.*, III), we have

$$\begin{aligned} \text{III} &:= \mathbb{E} \left[ \int_0^t \int_{\mathbb{O}} \left\| \int_{\mathbb{R}^d} \psi(\mathbf{X}_u^{\mu, \beta}(s), \hat{\mathbf{Y}}) |W_{\beta} - W_{\alpha}| d\nu_{v, s}(\hat{\mathbf{Y}}) \right\|^2 d\nu_{\text{Unif}} ds \right] \\ &\leq (2C_1^2 \mathfrak{m}_2 H_{\psi} + 1) \int_0^t \int_{\mathbb{O}^2} |W_{\beta} - W_{\alpha}|^2 d\nu_{\text{Unif}}^{\otimes 2}(u, v) ds \\ &\leq (2C_1^2 \mathfrak{m}_2 H_{\psi} + 1) |\mathbb{T}| d_{\mathfrak{g}}^2(W_{\beta}, W_{\alpha}). \end{aligned} \quad (54)$$

The upper-bound of last term can be directly obtained by the Lipschitz condition.

$$\begin{aligned} \text{IV} &:= \mathbb{E} \left[ \int_0^t \|b(s, \mathbf{X}_u^{\mu, \alpha}(s), \alpha) - b(s, \mathbf{X}_u^{\nu, \beta}(t), \beta)\|^2 ds \right] \\ &\leq \text{Lip}_b \mathbb{E} \left[ \int_0^t \sup_{r \leq s} \|\mathbf{X}_u^{\mu, \alpha}(r) - \mathbf{X}_u^{\nu, \beta}(r)\|^2 ds \right] + \text{Lip}_b \int_0^t \sup_{r \leq s} \|\alpha - \beta\|_{\nu_u(s)}^2 ds. \end{aligned} \quad (55)$$

By replacing each term with numerical constants  $C_3, C_4, C_5$  in the aggregation of all four terms, we finally have the following upper-bounds related to  $d_g, \mathcal{W}_{\mathcal{M}}$  and  $L_2$ -norm:

$$\begin{aligned} \mathbb{E} \left[ \sup_{s \leq t} \|\mathbf{X}_u^{\mu, \beta}(s) - \mathbf{X}_u^{\nu, \alpha}(s)\|^2 \right] &\leq 3b_0 (\text{I} + \text{II} + \text{III}) + b_0(\text{IV}) \\ &\leq \underbrace{b_0(3C_1 H_\psi(\bar{\alpha}) \mathfrak{h}(\beta) + \mathbf{Lip}_b)}_{:= \log(C_3/t)} \mathbb{E} \left[ \int_0^s \sup_{r \leq s} \|\mathbf{X}_u^{\nu, \alpha}(r) - \mathbf{X}_u^{\nu, \beta}(r)\|^2 dr \right] \\ &\quad + \underbrace{(6b_0 C_1^2 \mathfrak{m}_2 H_\psi + 3b_0) |\mathbb{T}|}_{:= C_4} d_g^2(W_\beta, W_\alpha) \\ &\quad + \underbrace{\max(3b_0 d |\mathbb{O}| C_1 \mathfrak{h}(\beta), \mathbf{Lip}_b)}_{:= C_5} \left( \int_0^t \sup_{r \leq s} \|\alpha - \beta\|_{\nu_u(r)}^2 + \mathcal{W}_{s, \mathcal{M}}(\mu^\beta, \nu^\alpha)^2 ds \right). \end{aligned} \quad (56)$$

Applying Grönwall's inequality to the above result in Eq (56) and the first inequality in Eq (49) shows that there exists a constant  $C' = 3 \max(C_3, C_4, C_5)$  such that

$$\mathcal{W}_{t, \mathcal{M}}(\Phi(\mu^\beta), \Phi(\nu^\alpha))^2 \leq C' \left( d_g^2(W_\beta, W_\alpha) + \int_0^t \sup_{r \leq s} \|\alpha - \beta\|_{\nu_u(r)}^2 + \mathcal{W}_{s, \mathcal{M}}(\mu^\beta, \nu^\alpha)^2 ds \right). \quad (57)$$

Next, the aim is to show the upper-bound of  $d_g^2, \|\alpha - \beta\|_\nu^2$  and  $\mathcal{W}_{\cdot, \mathcal{M}}$ . To proceed, let us first examine the upper bounds of the cut norms for both exponential and cosinusoidal graphons as follows:

$$\sup_{A, B} \left| \int_{A \times B} W_\alpha(u, v) du dv \right|^2 \leq \mathfrak{h}(\alpha) = \begin{cases} W_0^2 + 2W_0(W_{1,l} + W_{2,l}) + (2/L)(\sum_l W_{1,l} + W_{2,l})^2 \\ (T/2)W_1^2 (e^{-2T^{-1}|\mathbb{O}|} - 1). \end{cases} \quad (58)$$

Modifying the upper-bound in Eq (58) by replacing  $W_\alpha$  with  $\delta W := W_\alpha - W_\beta$ , one can derive the following

$$d_g^2(W_\beta, W_\alpha) = \|\delta W\|_g^2 \leq \max \left( 11\mathbf{Lip}_W, (T/2)(e^{-2T^{-1}|\mathbb{O}|} - 1) \right) \|\alpha - \beta\|_\nu^2. \quad (59)$$

At each stage  $\{m\}_{1 \leq m \leq M}$  with the given sequence of probability measures  $\{\nu^{\alpha_m}\}_{1 \leq m \leq M}$ , we substitute  $\Phi(\mu^\beta)$  and  $\Phi(\nu^\alpha)$  in Eq (57) with  $\Phi \circ \Psi(\nu^{\alpha_{m+1}})$  and  $\Phi \circ \Psi(\nu^{\alpha_m})$ , respectively. Then, one can derive the following relation:

$$\begin{aligned} \mathcal{W}_{t, \mathcal{M}}(\Phi \circ \Psi(\nu^{\alpha_m}), \Phi \circ \Psi(\nu^{\alpha_{m-1}}))^2 &= \mathcal{W}_{t, \mathcal{M}}(\mathbf{Law}(\mathbf{X}^{\nu^{\alpha_{m+1}^*}, \alpha_{m+1}^*}), \mathbf{Law}(\mathbf{X}^{\nu^{\alpha_m^*}, \alpha_m^*}))^2 \\ &\leq C' \left( d_g^2(W_{\alpha_{m+1}^*}, W_{\alpha_m^*}) + \int_0^t \sup_{r \leq s} \|\alpha_{m+1}^* - \alpha_m^*\|_{\nu_u(r)}^2 + \mathcal{W}_{s, \mathcal{M}}(\nu^{\alpha_{m+1}^*}, \nu^{\alpha_m^*})^2 ds \right) \\ &\leq C' \left( \max \left( t + 11\mathbf{Lip}_W, t + (T/2)(e^{-2T^{-1}|\mathbb{O}|} - 1) \right) \right) \sup_t \|\alpha(t, \cdot, \theta^{m+1}) - \alpha(t, \cdot, \theta^m)\|_{\nu_u(t)}^2 \\ &\quad + C' \int_0^t \mathcal{W}_{s, \mathcal{M}}(\nu^{\alpha_{m+1}^*}, \nu^{\alpha_m^*})^2 ds \\ &\leq \underbrace{C' \left( \max \left( t + 11\mathbf{Lip}_W, t + (T/2)(e^{-2T^{-1}|\mathbb{O}|} - 1) \right) \right)}_{:= C(t) \leq C(T)} \left( \sup_t r_m \right) \\ &\quad + C' \int_0^t \mathcal{W}_{s, \mathcal{M}}(\nu^{\alpha_{m+1}^*}, \nu^{\alpha_m^*})^2 ds, \end{aligned}$$

where the radius of metric ball (*i.e.*,  $r_m := r_{u, t, m}$ ) was defined in the proof of Proposition A.4. In the first equality, the controls  $\alpha$  are replaced with their optimal profiles  $\alpha^*$  following the definition of the operator  $\Psi$  in. To set up the subsequent stage, we substitute a pair of controls  $(\alpha_{m+1}^*, \alpha_m^*)$  with  $(\alpha_{m+1}, \alpha_m)$  again. Next, we show the stability of the result obtained above for  $M$ -th stage by



observing the upper bound of  $M$ -fold of the operator composition.

$$\begin{aligned}
& \mathcal{W}_{t,\mathcal{M}}([\Phi \circ \Psi]^{\circ M}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{\circ M}(\nu^{\alpha_0}))^2 \\
& \leq C(t) \sup_t \mathbf{r}_m + C' \int_0^t \mathcal{W}_{s_0,\mathcal{M}}([\Phi \circ \Psi]^{\circ M-1}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{\circ M-1}(\nu^{\alpha_0}))^2 ds^0 \\
& \quad \vdots \\
& \leq \sum_{m=1}^M (C(t) \sup_t \mathbf{r}_m)^m + (C')^M \int_0^{s_0} \dots \int_0^{s_M} \mathcal{W}_{s_m,\mathcal{M}}(\nu^{\alpha_{m+1}}, \nu^{\alpha_m})^2 d[\Pi^M](s^0, \dots, s^M).
\end{aligned} \tag{60}$$

where  $d\Pi^m := ds^0 \otimes \dots \otimes ds^m$  denotes  $m$ -product of Lebesgue measures  $\{ds^m\}_{1 \leq m \leq M}$ . Finally, we deduce that the supremum of the left-hand side can be controlled by

$$\begin{aligned}
& \lim_{M \rightarrow \infty} \sup_{t \in \mathbb{T}} \mathcal{W}_{t,\mathcal{M}}([\Phi \circ \Psi]^{\circ m}(\nu^{\alpha_1}), [\Phi \circ \Psi]^{\circ m}(\nu^{\alpha_0}))^2 \leq \\
& \quad + \frac{C(T)^M (\sup_t \sup_m \mathbf{r}_m)^M - 1}{C(T) (\sup_t \sup_m \mathbf{r}_m) - 1} + \frac{(C'T)^M}{M!} \sup_{t \in \mathbb{T}} \mathcal{W}_{t,\mathcal{M}}(\nu^{\alpha_1}, \nu^{\alpha_0})^2 \rightarrow 0.
\end{aligned} \tag{61}$$

where the learning rate  $\gamma^m$  is chosen such that  $\sup_t \sup_m \mathbf{r}_m \leq 1$  remains sufficiently small, and the last term in the inequality can be derived by modifying the following

$$\sup_{t \in \mathbb{T}} \mathcal{W}_{t,\mathcal{M}}(\Phi^{\circ m}(\nu^{\alpha_1}), \Phi^{\circ m}(\nu^{\alpha_0}))^2 \leq (C')^M \int_0^T \frac{(T-s)}{(m-1)!} \mathcal{W}_{s,\mathcal{M}}(\nu^{\alpha_1}, \nu^{\alpha_0})^2 ds. \tag{62}$$

The inequality in Eq (61) demonstrates that the sequence of operator compositions  $\{[\Phi \circ \Psi]^{\circ m}\}_{m \leq M} : \mathcal{M} \rightarrow \mathcal{M}$  forms a Cauchy sequence, confirming the convergence of the proposed gradient system in the distributional sense.  $\square$

#### A.4.4 SAMPLING ERRORS OF MEAN-FIELD PREDICTORS

Though not presented in the manuscript, the following result implies key theoretical conclusions: It demonstrates that the estimation errors for the neural agent, introduced by the sampled mean-field predictors (empirical measure) at the  $m$ -th gradient descent step, are kept within acceptable margins.

**Proposition A.5.** (Worst-case Estimation Error of Neural Agents) Let  $\mathbb{Q}_n := \mathbb{Q}_n(\mathbf{u}, t) = (1/n) \sum_i \delta_{\mathbf{x}_{\mathbf{u}_i}^\alpha(t)}$  and  $\mathbb{Q} := \nu_{\mathbf{u}}(t)$  be empirical laws of mean-field predictors and their corresponding mean-field limit. Then, the worst-case estimation error can be upper bounded with probability at least  $1 - \delta$ :

$$\begin{aligned}
& \sup_{\alpha_m \in \mathbb{A}} \left\| \int \alpha^m d(\mathbb{Q}_n - \mathbb{Q}) \right\|_E^2 \leq \sqrt{\frac{32T^3(1 + \mathfrak{m}_2)^2}{n} \ln \left( \frac{1}{\delta} \right)} \\
& \quad + 4 \left( \sqrt{\frac{32}{n}} 2^{(3d-2)/2} \left( \varepsilon \gamma^{m-1} \mathbf{Lip}_\alpha \|\nabla_\theta \mathbf{Y}_{\mathbf{u}}^{m-1, \alpha_{m-1}}(t)\|_E \right)^{d/2} \frac{d+2}{4(d-2)} \right)^{(d/2+2)^{-1}}.
\end{aligned} \tag{63}$$

**Remark.** While the admissible control set  $\mathbb{A}$  guarantees the diminishing behavior of  $\|\nabla_\theta \mathbf{Y}_{\mathbf{u}}^{m-1, \alpha_{m-1}}(t)\|_E$ , the second term in Eq (63) approaches zero as  $m$  becomes large, even when  $n$  is small.

*Proof.* The proof follows the standard convergence analysis of empirical processes. Let us fix the temporal variable  $t$  and the labels of mean-field predictors  $\mathbf{u}$ . Then, one can show that the supremum of Euclidean norm can be decomposed as follows:

$$\sum_j^d \sup_{\pi_j \circ \alpha \in \mathbb{A}_m^j} |\mathbb{E}_{\mathbb{Q}_n} \pi_j \circ \alpha_m - \mathbb{E}_{\mathbb{Q}} \pi_j \circ \alpha_m| \leq \sum_j^d \sup_{g \in \mathbb{A}_m^j} |\mathbb{E}_{\mathbb{Q}_n} g - \mathbb{E}_{\mathbb{Q}} g| := \Gamma_{\mathbb{A}_m^j}(\mathbb{Q}_n, \mathbb{Q}), \tag{64}$$

where  $\Gamma_{\mathbb{A}_m^j}$  denotes the *integral probability metric* (Müller, 1997) with respect to the set  $\mathbb{A}_m^j$  which consists of  $j$ -th component of neural agents at  $m$ -th stage. Note that the the supremum in the second term is taken for all function  $g$  lying in the set of parameterized function, *i.e.*, neural agent. Let us define  $\mathbf{p}, \mathbf{q} : (\mathbb{R}^d)^n \rightarrow \mathbb{R}$  such that  $(\mathbf{X}_{u_1}^\alpha(t), \dots, \mathbf{X}_{u_n}^\alpha(t)) \xrightarrow{\mathbf{p}} \sup_g |(1/n) \sum_i g(\mathbf{X}_{u_i}^\alpha(t)) - \mathbb{E}_{\mathbb{Q}} g|$ , and  $(\mathbf{X}_{u_1}^\alpha(t), \dots, \mathbf{X}_{u_n}^\alpha(t)) \xrightarrow{\mathbf{q}} \mathbb{E}_{\sigma} \sup_g |(1/n) \sum_i \sigma_i g(\mathbf{X}_{u_i}^\alpha(t))|$  where  $\{\sigma_i\}_{i \leq n}$  is a set of i.i.d Rademacher random variables. Then both  $\mathbf{p}$  and  $\mathbf{q}$  satisfies the following inequality:

$$\sup_t \max_{i \in \{1, \dots, n\}} |(\mathbf{p}, \mathbf{q})(\mathbf{X}_{u_1}^\alpha(t), \dots, \mathbf{X}_{u_{i-1}}^\alpha(t), \mathbf{x}', \mathbf{X}_{u_{i+1}}^\alpha(t), \dots, \mathbf{X}_{u_n}^\alpha(t)) - (\mathbf{p}, \mathbf{q})(\mathbf{X}_{u_1}^\alpha(t), \dots, \mathbf{X}_{u_n}^\alpha(t))| \leq \frac{4T \sup_{\mathbf{x}, t} \alpha(t, \mathbf{x}; \theta)}{n}. \quad (65)$$

Following by the McDiarmid's inequality with respect to  $\mathbf{p}$ , we have two concentration inequalities:

$$\exp\left(\frac{-n\varepsilon^2}{8T^2 \sup_{\mathbf{x}, t} \alpha(t, \mathbf{x}; \theta)^2}\right) \geq \begin{cases} \mathbb{P}(\mathbf{p} - \mathbb{E}\mathbf{p} \geq \varepsilon) \\ \mathbb{P}(\mathbf{q} - \mathbb{E}\mathbf{q} \geq \varepsilon). \end{cases} \quad (66)$$

By applying the symmetrization inequality (Wellner et al., 2013), we have the following inequality with probability at least  $1 - \delta$

$$\begin{aligned} \Gamma_{\mathbb{A}_m^j}(\mathbb{Q}_n, \mathbb{Q}) &\leq \mathbb{E}\Gamma_{\mathbb{A}_m^j}(\mathbb{Q}_n, \mathbb{Q}) + \sqrt{\frac{8T^2 \sup_{\mathbf{x}, t} \alpha(t, \mathbf{x}; \theta)^2}{n} \ln\left(\frac{1}{\delta}\right)} \\ &\leq 2\mathbb{E}\mathbb{E}_{\sigma} \left[ \sup_{g \in \mathbb{A}_m^j} \left| \frac{1}{n} \sum_i \sigma_i g(\mathbf{X}_{u_i}^\alpha(t)) \right| + \sqrt{\frac{8T^2 \sup_{\mathbf{x}, t} \alpha(t, \mathbf{x}; \theta)^2}{n} \ln\left(\frac{1}{\delta}\right)} \right] \\ &\leq 2\mathbb{E}_{\sigma} \left[ \underbrace{\sup_{g \in \mathbb{A}_m^j} \left| \frac{1}{n} \sum_i \sigma_i g(\mathbf{X}_{u_i}^\alpha(t)) \right|}_{\mathcal{R}_m(\mathbb{A}_m^j, \{\mathbf{X}_{u_n}^\alpha(t)\})} \right] + \sqrt{\frac{32T^3(1 + \mathbf{m}_2)^2}{n} \ln\left(\frac{1}{\delta}\right)} \end{aligned} \quad (67)$$

where the outer expectation is taken with respect to the randomness of mean-field predictors in the second line, and we apply McDiarmid's inequality in Eq (66) again to derive the last line. With the covering number of the Hilbert space for the  $\mathbb{L}_2$ -norm, we get

$$\begin{aligned} \mathcal{R}_m(\mathbb{A}_m^j, \{\mathbf{X}_{u_n}^\alpha(t)\}) &\leq \mathbb{E}_{\sigma} \left[ \sup_{g \in \mathbb{A}_m^j} \left| \frac{1}{n} \sum_i \sigma_i g(\mathbf{X}_{u_i}^\alpha(t)) \right| \right] \\ &\leq \inf_{\epsilon > 0} \left\{ 2\epsilon + \sqrt{\frac{32}{n}} \int_{\epsilon/4}^{\infty} \sqrt{\mathcal{H}(\tau, \mathbb{A}_m^j, \mathbb{L}_2(\mathbb{Q}_n))} \right\} \\ &\leq \inf_{\epsilon > 0} \left\{ 2\epsilon + \sqrt{\frac{32}{n}} \int_{\epsilon/4}^{\infty} \left( \frac{2\mathbf{r}_m}{\tau} \right)^{d/2} d\tau \right\} \\ &\leq \inf_{\epsilon > 0} \left\{ 2\epsilon + \sqrt{\frac{32}{n}} (2\mathbf{r}_m)^{d/2} (\epsilon/4)^{-d/2+1} (d/2 - 1)^{-1} \right\} \\ &\leq \inf_{\epsilon > 0} \left\{ 2\epsilon + \sqrt{\frac{32}{n}} 2^{(3d-2)/2} \mathbf{r}_m^{d/2} \epsilon^{-d/2-1} (d-2)^{-1} \right\} \\ &= 4 \left( \sqrt{\frac{32}{n}} 2^{(3d-2)/2} (\epsilon \gamma^{n-1} \mathbf{Lip}_{\alpha} \delta_{\theta} Y_u^{m-1}(t))^{d/2} \frac{d+2}{4(d-2)} \right)^{(d/2+2)^{-1}}, \end{aligned} \quad (68)$$

where we assume the data dimensionality is  $d > 2$ . The second line is a direct consequence of Theorem 16 (von Luxburg & Bousquet, 2004), the second inequality can be derived from the fact that  $\mathbb{Q}_n$  is an empirical measure, and  $\mathbb{A}_m^j$  is a metric ball of radius  $\mathbf{r}_m$  embedded on finite-dimensional Hilbert space following by (H4). By setting  $d = d'$ , the last result comes from the definition of radius  $\mathbf{r}_m$ .  $\square$

**Proposition 4.2.** (*Sampling Complexity*) Let  $\nu_t^N, \hat{\mu}_t$  probability measures defined in Eq (7). Then, there exist numerical constants  $\mathfrak{c}, \mathfrak{c}_7, \mathfrak{c}_8, \mathfrak{c}_9 > 0, w > 0$  and  $\kappa > 0$  such that the probability of squared 2-Wasserstein distance can be controlled as follows:

$$\mathbb{P} [W_2^2(\nu_t^N, \hat{\mu}_t) \geq \epsilon] \leq \mathfrak{A} \left( \frac{1}{\epsilon^2} e^{-N\epsilon^2/4\mathfrak{c}} + \frac{1}{N} e^{-N\epsilon} \left( 1 - \frac{128\omega\mathfrak{h}(\alpha)}{N} \right)^{-d/8} + \frac{1}{72^4\epsilon\sqrt{N}} \right), \quad (69)$$

$$\mathfrak{A} = \max \left( \mathfrak{c}_9, \frac{2\mathfrak{c}_7^{3/2}}{\kappa} \exp(\mathfrak{c}_4 e^{\frac{1}{2}\mathfrak{c}_1 T}) (e^{\kappa T} - 1), \mathfrak{c}_9 \exp(-4\mathfrak{c}_8) \right), \quad (70)$$

where  $\mathfrak{u} \in \mathbb{O}$ ,  $t \in \mathbb{T}$  is arbitrary and  $\mathfrak{h}(\alpha) = \|W_\alpha\|_{\mathfrak{g}}$  is a cut-norm<sup>a</sup> of the proposed graphons (i.e., exponential, cosinusoidal).

<sup>a</sup>Eq. 58 clarifies the explicit upper-bound of the cut-norm for the proposed graphons.

**Remark.** The approach used in the proof to establish the concentration bound is largely inspired by the series of works on the measure concentration (Bolley et al., 2007; Budhiraja & Fan, 2017; Bayraktar & Wu, 2022; Bayraktar et al., 2023; Bayraktar & Wu, 2023), with slight modifications tailored to the structure of the proposed mean-field system. We intentionally omit some parts of the proofs in this work that have already been covered in the reference.

*Proof.* We divide the proof into separate steps.

**1. Estimation of Concentration Inequality.** For the controlled mean-field system via neural agents  $\alpha$ , fix the the population  $\nu^\alpha$  and its related control  $\mathbf{X}_u^{\nu, \alpha} = \mathbf{X}_u$  and let  $\mathfrak{u} = i/n$  for the moment. First, let us define the following probability measures:

$$\nu_t^n := \frac{1}{n} \sum_i \delta_{\mathbf{X}_i^n(t)}, \bar{\nu}_t^n := \frac{1}{n} \sum_i \delta_{\mathbf{X}_{(i/n)}(t)}, \hat{\mu}_t = \int \nu_u(t) p(d\mathfrak{u}), \bar{\mu}_t^n = \frac{1}{n} \sum_i \nu_{\mathfrak{u}=i/n}(t). \quad (71)$$

Then, we analyze the law of difference between the following two mean-field dynamics:

$$\begin{aligned} \mathbf{X}_u(t) &= \mathbf{X}_u(0) + \int_0^t \langle \mathbb{W}_\alpha[\nu_{v,s}], \psi \rangle (\mathbf{X}_u(s)) ds + \int_0^t \mathbf{b}(s, \mathbf{X}_u(s), \alpha) ds + \int_0^t \sigma_s dW_s^u, \\ \mathbf{X}_i^n(t) &= \mathbf{X}_{(i/n)}(0) + \int_0^t \langle \mathbb{W}_\alpha[\delta_{v,s}], \psi \rangle (\mathbf{X}_i^n(s)) ds + \int_0^t \mathbf{b}(s, \mathbf{X}_i^n(s), \alpha) ds + \int_0^t \sigma_s dW_s^{(i/n)}. \end{aligned}$$

Given that fact that the expectation of Ito's differential for mean-square error can be expressed as  $d_{\mathbf{I}} \|A(t)\|^2 = 2\langle A(t), \mathbf{m}_A \rangle dt + 2\sigma A(t) dW_t + \sigma^2 dt$  where  $\mathbb{R}^+ \ni \sigma$  and  $\mathbf{m}_A$  are compensate and martingale part of  $A(t)$ , we get

$$\begin{aligned} d_{\mathbf{I}} \|\mathbf{X}_{(i/n)}(t) - \mathbf{X}_i^n(t)\|_E^2 &= 2\delta\mathbf{X}(t) \cdot (\mathbf{b}(s, \mathbf{X}_i^n(s), \alpha) - \mathbf{b}(s, \mathbf{X}_{i/n}(s), \alpha)) dt \\ &\leq \left( \frac{1}{n} \sum_{j=1}^n W_\alpha \left( \frac{i}{n}, \frac{j}{n} \right) \psi_\alpha(\mathbf{X}_i^n(t), \mathbf{X}_j^n(t)) - \hat{\mathbb{E}} \left[ W_\alpha \left( \frac{i}{n}, v \right) \psi_\alpha(\mathbf{X}_{(i/n)}(t), x) \right] \right) \\ &\quad \cdot 2\delta\mathbf{X}(t) dt \end{aligned} \quad (72)$$

where we denote  $\hat{\mathbb{E}} := \mathbb{E}_{v \sim p(v), x \sim \nu_{v=j/n}(t)}$  and  $p(v) := w_\#[\mathbf{Unif}(\mathbb{O})]$ ,  $\delta\mathbf{X}(t) := \mathbf{X}_{(i/n)}(t) - \mathbf{X}_i^n(t)$ . Then, the dissipativity assumption gives

$$d_{\mathbf{I}} \|\delta\mathbf{X}(t)\|_E^2 \leq \text{I} + \text{II} + \text{III} + \text{IV} \quad (73)$$

For simplicity let us denote  $W^{i,j} := W_\alpha(i/n, i/j)$ , and  $W^{i,v} := W_\alpha(i/n, v)$ . Using the dissipativity of the proposed drift function. For the second first, one can get

$$\text{I} := 2\delta\mathbf{X}(t) \cdot (\mathbf{b}(s, \mathbf{X}_i^n(s), \alpha) - \mathbf{b}(s, \mathbf{X}_{i/n}(s), \alpha)) \leq -\mathfrak{c}_1 \|\delta\mathbf{X}(t)\|_E^2 \quad (74)$$

By adding and subtracting new terms, we have

$$\begin{aligned} \text{II} &:= \left( \frac{1}{n} \sum_j^n W^{i,j} [\psi_\alpha(\mathbf{X}_i^n, \mathbf{X}_j^n) - \psi_\alpha(\mathbf{X}_{(i/n)} - \mathbf{X}_{(j/n)})] \right) \cdot \delta \mathbf{X}(t) \\ &\leq \frac{\text{Lip}_b}{n} \sum_j^n |\delta \mathbf{X}(t)| (|\delta \mathbf{X}(t)| + |\mathbf{X}_j^n(t) - \mathbf{X}_{(j/n)(t)}|) \end{aligned} \quad (75)$$

Similarly, the second term can be upper-bounded as follows:

$$\begin{aligned} \text{III} &:= \left( \frac{1}{n} \sum_j^n W^{i,j} [\psi_\alpha(\mathbf{X}_{(i/n)}, \mathbf{X}_{(j/n)}) - \mathbb{E}_{\nu_{j/n}(t)} \psi_\alpha(\mathbf{X}_i^n, \cdot)] \right) \cdot \delta \mathbf{X}(t) \\ &\leq |\delta \mathbf{X}(t)| \cdot \|W^{i,j}\|_\infty \|\mathcal{F}_{\text{III}}^i\|_E^2. \end{aligned} \quad (76)$$

By adding and subtracting the term  $W_{i,j} \mathbb{E} \psi_\alpha(\mathbf{X}_{i/n}(t), \cdot)$ , the fourth term can be improved as

$$\begin{aligned} \text{IV} &:= \left( \frac{1}{n} \sum_j^n \left[ W^{i,j} \int \psi_\alpha(\mathbf{X}_{i/n}(t), \cdot) d\nu_{i/n}(t) - \int W^{i,v} \psi_\alpha(\mathbf{X}_{i/n}(t), \cdot) d\nu_v(t) \right] \right) \cdot \delta \mathbf{X}(t) \\ &\leq \frac{1}{n} \sum_j^n \|W^{i,j}\|_\infty (C_1 \mathcal{W}_2(\nu_{i/n}(t), \nu_v(t)) + n_2 d_g(W^{i,j}, W^{i,v})) \\ &\leq |\delta \mathbf{X}(t)| \cdot \|W^{i,j}\|_\infty \|\mathcal{F}_{\text{IV}}^i\|_E^2 \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad (77)$$

Note that the last inequality tends to zero for large enough  $n$ . Aggregating all the terms and using the fact that  $g'(t) \leq ag(t) + b$  implies  $g(t) \leq \int e^{-a(t-s)} b ds$  and  $d/dt \|g(t)\|_E^2 \leq (1/2)g(t)^{-1/2} \dot{g}(t)$ , where  $g(t) := (1/n) \sum_i^n \|\delta \mathbf{X}(t)\|_E^2$  and  $a = (2\text{Lip}_b - c_1)$ ,  $b := b(\mathcal{F}_{\text{III}}^i, \mathcal{F}_{\text{IV}}^i)$ , we have

$$\begin{aligned} \mathcal{W}_2^2(\nu_t^n, \bar{\nu}_t^n) &\leq \frac{1}{n} \sum_i^n \|\delta \mathbf{X}(t)\|_E^2 \\ &\leq \int_0^t e^{-(4\text{Lip}_b - 2c_1)(t-s)} \left( \sup_{i',j'} \|W^{i',j'}\|_\infty^2 \frac{1}{n} \sum_i^n \left\| \mathcal{F}_{\text{III}}^i\|_E^2 + \|\mathcal{F}_{\text{IV}}^i\|_E^2 \right\|^2 \right) ds. \\ &\leq \underbrace{\int_0^t e^{-(4\text{Lip}_b - 2c_1)(t-s)} \left( \sup_{i',j'} \|W^{i',j'}\|_\infty^2 \frac{1}{n} \sum_i^n \|\mathcal{F}_{\text{III}}^i\|_E^2 + \|\mathcal{F}_{\text{IV}}^i\|_E^2 \right) ds}_{:= \text{V} + \text{VI}} \end{aligned} \quad (78)$$

where the first inequality follows from the estimation of Wasserstein distance for empirical measures, and the last inequality can be derived by applying AM-GM inequality.

$$\mathbb{P} [W_2^2(\nu_t^n, \hat{\mu}_t) \geq \epsilon] \leq \mathbb{P} \left[ \underbrace{W_2^2(\bar{\nu}_t^n, \hat{\mu}_t) \geq \epsilon/2}_{:= \text{VII}} \right] + \mathbb{P}[\text{V} \geq \epsilon/4] + \underbrace{\mathbb{P}[\text{VI} \geq \epsilon/4]}_{=0, n \gg N}, \quad (79)$$

where the last term vanishes for small enough  $\epsilon$ , with large  $N$ .

**2. Estimation of Exponential**  $e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}$ . In this step, we derive the upper bound of the exponential for the square norm of mean-field predictors. We first apply the Ito's lemma to  $e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}$  for arbitrary scalar  $\lambda_{\text{exp}} > 0$  and observe that

$$d_{\mathbf{I}} e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} = \lambda_{\text{exp}} e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} (2\mathbf{X}_u \cdot (\mathbf{b} + \mathbf{b}_W) dt + \sigma_t (d + 2\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2) dt + \sigma_t dB_u). \quad (80)$$

where gradient and Laplace of exponential can be calculated as  $\nabla e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} = 2\lambda_{\text{exp}} e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}$  and  $\Delta e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} = 2\lambda_{\text{exp}} e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} (d + 2\lambda_{\text{exp}} e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2})$ . Taking expectation on both sides with the dissipative condition, we can show that there exist constants  $\mathbf{c}_2 = 2\lambda_{\text{exp}}(-\mathbf{c}_1 + \sigma_t \lambda_{\text{exp}})$ ,  $\mathbf{c}_3 = \lambda_{\text{exp}} \sigma_t d$  that directly gives following two inequalities

$$d_{\mathbf{I}} \mathbb{E}[e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}] \leq \mathbb{E} \left[ e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2} (\mathbf{c}_2 \|\mathbf{X}_u(t)\|_E^2 + \mathbf{c}_3) \right] dt + \mathbb{E} \left[ \int M_s dt \right], \quad (81)$$

$$\sup_{t \leq T} \|\mathbf{X}_u(t)\|_E^2 \leq \sup_{t \leq T} \|y_u\|^2 + N_t + \mathbf{c}_1 \int_0^t \|\mathbf{X}_u(s)\|_E^2 ds \leq \mathbf{c}_4 e^{\mathbf{c}_1 T}. \quad (82)$$

where the second inequality is a direct consequence of Grownall's inequality, and  $M_t$  and  $N_t$  denote some martingale. Applying Grownall's inequality again, we have the desired result.

$$d_{\mathbf{I}} \mathbb{E}[e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}] \leq (\mathbf{c}_5 + \mathbf{c}_6 \mathbb{E}[e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}]) dt, \quad (83)$$

$$\mathbb{E}[e^{\lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2}] \leq (\exp(\lambda_{\text{exp}} \|y_u\|_E^2) + \mathbf{c}_5) \exp(\mathbf{c}_6 T) \leq (\mathbf{c}_7)^2. \quad (84)$$

where we used inequality  $e^a + e^b \leq \exp(\max(a, b) + \ln(1 + \exp(-|a - b|))) = (\mathbf{c}_7)^2$  such that  $a = \lambda_{\text{exp}} \|\mathbf{X}_u(t)\|_E^2 + \mathbf{c}_6 T$ ,  $b = \ln \mathbf{c}_5 + \mathbf{c}_6 T$ . Note that the upper-bound of the term  $\exp(\lambda_{\text{exp}} \|y_u\|_E^2)$  at initial time  $t = 0$  determines the exponential integrability of the right-hand side above.

**3. Estimation of Probability**  $\mathbb{P}[\mathbf{V} \geq \epsilon/4]$ . By the exponential Markov inequality with some constant  $\lambda > 0$ , Jensen's inequality, we obtain

$$\begin{aligned} \mathbb{P}[\mathbf{V} \geq \epsilon/4] &:= \mathbb{P} \left[ \int_0^t e^{-(4\mathbf{Lip}_b - 2\mathbf{c}_1)(t-s)} \left( \sup_{i', j'} \|W^{i', j'}\|_\infty^2 \frac{1}{n} \sum_i \|\mathcal{F}_{\text{III}}^i\|_E^2 \right) ds > \epsilon/4 \right] \\ &\leq \frac{1}{n} \sum_i e^{-\lambda \epsilon/4} \mathbb{E} \left[ \int_0^t e^{-(4\mathbf{Lip}_b - 2\mathbf{c}_1)(t-s)} \right. \\ &\quad \cdot \exp \left( \lambda \mathbf{h}(\alpha) \left\| \frac{1}{n} \sum_j \psi_\alpha(\mathbf{X}_{(i/n)}, \mathbf{X}_{(j/n)}) - \mathbb{E}_{\nu_{j/n}(t)} \psi_\alpha(\mathbf{X}_{(i/n)}, \cdot) \right\|_E^2 \right) ds \Big]. \end{aligned} \quad (85)$$

Note that  $\|\psi_\alpha(x, y)\|_E \leq \mathbf{Lip}_\psi (\|x\|_E + \|y\|_E)$  have linear growth for all  $x, y \in \mathbb{R}^d$  by the assumptions.

$$\begin{aligned} &\mathbb{E} \left[ \exp \left( \lambda \mathbf{h}(\alpha) \left\| \frac{1}{n} \sum_j \psi_\alpha(\mathbf{X}_{(i/n)}, \mathbf{X}_{(j/n)}) - \mathbb{E}_{\nu_{j/n}(t)} \psi_\alpha(\mathbf{X}_{(i/n)}, \cdot) \right\|_E^2 \right) \right] \\ &\leq \mathbb{E} \left[ \exp \left( \frac{2\lambda \mathbf{h}(\alpha) \mathbf{Lip}_\psi}{n} \|\mathbf{X}_{(i/n)}\|_E^2 + 2\lambda \mathbf{h}(\alpha) \left\| \frac{1}{n} F_\psi \right\|_E^2 \right) \right] \\ &\leq \left( 2\mathbb{E} \left[ \exp \left( \frac{4\lambda \mathbf{h}(\alpha) \mathbf{Lip}_\psi}{n} \|\mathbf{X}_{(i/n)}\|_E^2 \right) \right] \right)^{1/2} \left( 2\mathbb{E} \left[ \exp \left( 2\zeta \left\| \frac{1}{n} F_\psi \right\|_E^2 \right) \right] \right)^{1/2} \end{aligned} \quad (86)$$

where the last inequality can be derived by applying exponential AM-GM inequality

$$\begin{aligned}
& \mathbb{E} \left[ \exp \left( 2\zeta \left\| \frac{1}{n} F_\psi \right\|_E^2 \right) \right] = \mathbb{E} \left[ \exp \left( \left\| \frac{2\sqrt{\zeta}}{n} \mathbf{Z} \right\|_E \cdot \|F_\psi\|_E \right) \right] \\
& \leq \mathbb{E} \left[ \exp \left( \omega \left\| \frac{2\sqrt{\zeta}}{n} \mathbf{Z} \right\|_E^2 + \frac{1}{4\omega} \|F_\psi\|_E^2 \right) \right] \\
& \leq \left( 2\mathbb{E} \left[ \exp \left( \frac{8\omega\zeta}{n^2} \|\mathbf{Z}\|_E^2 \right) \right] \right)^{1/2} (2\mathbb{E} [\exp ((10n) \cdot \mathbf{Lip}_\psi \|F_\psi\|_E^2)])^{1/2} \exp(\mathbf{c}_4 e^{\mathbf{c}_1 T}) \\
& \leq 2\mathbf{c}_7 \left( 1 - \frac{16\omega\zeta}{n^2} \right)^{-\frac{d}{4}} \cdot \exp(\mathbf{c}_4 e^{\mathbf{c}_1 T}),
\end{aligned} \tag{87}$$

where  $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$  is a standard Gaussian random vector. The last inequality is a direct consequence of the property of the moment generation function. The second line can be deduced from the fact that the discretized predictors  $\mathbf{X}_{(i/n)}$  and  $\mathbf{X}_{(j/n)}$  are i.i.d with the selection of  $\omega > 0$ ,  $\lambda_{\text{exp}}$  and  $\zeta$  satisfying the following:

$$\frac{1}{4\omega} \|F_\psi\|_E^2 \leq n \cdot \mathbf{Lip}_\psi (5\|\mathbf{X}_{(i/n)}\|_E^2 + \exp(\mathbf{c}_4 e^{\mathbf{c}_1 T})) \tag{88}$$

$$\lambda_{\text{exp}} := \max \left( \frac{4\lambda\mathfrak{h}(\alpha)\mathbf{Lip}_\psi}{n}, (10n)\mathbf{Lip}_\psi \right). \tag{89}$$

$$\zeta := 2\lambda\mathfrak{h}(\alpha) > 0 \tag{90}$$

By aggregating all the terms, we finally have

$$\begin{aligned}
& \mathbb{E} \left[ \exp \left( \lambda\mathfrak{h}(\alpha) \left\| \frac{1}{n} \sum_j \psi_\alpha(\mathbf{X}_{(i/n)}, \mathbf{X}_{(j/n)}) - \mathbb{E}_{\nu_{j/n}(t)} \psi_\alpha(\mathbf{X}_{(i/n)}, \cdot) \right\|_E^2 \right) \right] \\
& \leq 2\mathbf{c}_7^{3/2} \left( 1 - \frac{16\omega\zeta}{n^2} \right)^{-\frac{d}{8}} \cdot \exp(\mathbf{c}_4 e^{\frac{1}{2}\mathbf{c}_1 T})
\end{aligned} \tag{91}$$

Thus, the probability of  $V$  larger than threshold  $\epsilon/4$  can be written as follows:

$$\mathbb{P}[V \geq \epsilon/4] \leq \frac{2}{\kappa n} e^{-n\epsilon} \mathbf{c}_7^{3/2} \left( 1 - \frac{16\omega\zeta}{n^2} \right)^{-\frac{d}{8}} \cdot \exp(\mathbf{c}_4 e^{\frac{1}{2}\mathbf{c}_1 T}) (e^{\kappa T} - 1), \tag{92}$$

$$\kappa = -(4\mathbf{Lip}_b - 2\mathbf{c}_1), \quad \lambda = 4n. \tag{93}$$

**4. Estimation of Probability**  $\mathbb{P}[\text{VII} \geq \epsilon/2]$ . Now, it remains to establish the upper bound of the probability related to VII. We modify the standard estimation of concentration probabilities of empirical measures as outlined in Bolley (2010). By the triangle inequality, the probability can be decomposed as

$$\mathbb{P} \left[ \text{VII} \geq \frac{\epsilon}{2} \right] \leq \mathbb{P} \left[ \sup_{\substack{h\Delta \leq t \leq (h+1)\Delta \\ 0 \leq h \leq \bar{M}-1}} \mathcal{W}_2^2(\bar{\nu}_t^n, \bar{\nu}_{h\Delta}^n) \geq \frac{\epsilon}{6} \right] + \mathbb{P} \left[ \sup_{0 \leq h \leq \bar{M}-1} \mathcal{W}_2^2(\bar{\nu}_{h\Delta}^n, \bar{\mu}_{h\Delta}^n) \geq \frac{\epsilon}{6} \right] \tag{94}$$

where the temporal interval can be also decomposed as  $\mathbb{T} = [0, \Delta] \cup [\Delta, 2\Delta] \cup \dots \cup [(M-1)\Delta, T] \subseteq \bigcup_{h=0}^{M-1} [h\Delta, (h+1)\Delta]$ . The first term of the right-hand side above can be bounded as

$$\begin{aligned}
& \mathbb{P} \left[ \sup_{h\Delta \leq t \leq (h+1)\Delta} \mathcal{W}_2^2(\bar{\nu}_{t_1}^n, \bar{\nu}_{t_2}^n) \geq \frac{\epsilon}{6} \right] \leq \mathbb{P} \left[ \frac{1}{n} \sup_{0 \leq t_1 \leq t_2 \leq t} \|\mathbf{X}_{i/n}(t_1) - \mathbf{X}_{i/n}(t_2)\|_E^2 \geq \frac{\epsilon}{6} \right] \\
& \leq \exp \left( -n \sup_{\zeta > 0} \left( \epsilon\zeta - \log \mathbb{E} \exp \left( \zeta \sup_{0 \leq t_1 \leq t_2 \leq t} \|\mathbf{X}_{i/n}(t_1) - \mathbf{X}_{i/n}(t_2)\|_E^2 \right) \right) \right)
\end{aligned} \tag{95}$$

The first line is induced as any measures  $\nu_{(\cdot)}^n$  are empirical, and the next line can be induced by using Chebyshev's exponential inequality and the independence of the mean-field predictor. Denoting

$\delta \mathbf{X}_{(i/n)} := \sup_{0 \leq t_1 \leq t_2 \leq t} \|\mathbf{X}_{i/n}(t_1) - \mathbf{X}_{i/n}(t_2)\|_E^2$  for any  $t_1 \leq t_2 \in \mathbb{T}$ , we can further improve the right hand side by showing

$$\mathbb{E} \exp(\zeta \delta \mathbf{X}_{(i/n)}) \leq \exp(\zeta^2 \mathbf{c}_8) \exp(2\zeta \delta \mathbf{X}_{(i/n)}) \leq \exp(\zeta^2 \mathbf{c}_8) (1 + \hat{C} \Delta), \quad (96)$$

where we used the fact that  $ax \leq a^2b + 2ax$  for all  $a, b, x \geq 0$ . In order to show the upper bound of the first term in the last inequality (96), we used the result (4.6) Bolley (2010) tailored to our case under the assumption made in Section A.3 for fixed  $u$  and  $\alpha$ . Combining results, we have

$$\begin{aligned} \mathbb{P} \left[ \sup_{\substack{h\Delta \leq t \leq (h+1)\Delta \\ 0 \leq h \leq \bar{M}-1}} \mathcal{W}_2^2(\bar{\nu}_t^n, \bar{\nu}_{h\Delta}^n) \geq \frac{\epsilon}{6} \right] &\leq \bar{M} \exp \left( -n \sup_{\zeta > 0} \left( \epsilon \zeta - \zeta^2 \mathbf{c}_8 - \log(1 + \hat{C} \Delta) \right) \right) \\ &\leq \bar{M} \exp \left( -\frac{n\epsilon^2}{4\mathbf{c}_8} - \log(1 + \hat{C} \Delta) \right) \leq \frac{\mathbf{c}_9}{\epsilon^2} \exp \left( -\frac{n\epsilon^2 + 1}{4\mathbf{c}_8} \right), \quad \begin{cases} \Delta = \exp(4\mathbf{c}_8^{-1}) \hat{C}^{-1}, \\ \bar{M} \leq \mathbf{c}_9/\epsilon^2. \end{cases} \end{aligned} \quad (97)$$

For the second term of the right-hand side in (94), we first apply Boole's inequality of events to have

$$\begin{aligned} \mathbb{P} \left[ \sup_{0 \leq h \leq \bar{M}-1} \mathcal{W}_2^2(\bar{\nu}_{h\Delta}^n, \hat{\mu}_{h\Delta}) \geq \frac{\epsilon^2}{36} \right] &\leq \overbrace{\mathbb{P} \left[ \sup_{0 \leq h \leq \bar{M}-1} \mathcal{W}_2^2(\bar{\mu}_{h\Delta}^n, \hat{\mu}_{h\Delta}) \geq \frac{\epsilon^2}{72} \right]}^{\rightarrow 0, n \gg N} \\ &\quad + \mathbb{P} \left[ \sup_{0 \leq h \leq \bar{M}-1} \mathcal{W}_2^2(\bar{\nu}_{h\Delta}^n, \bar{\mu}_{h\Delta}^n) \geq \frac{\epsilon^2}{72} \right] \\ &\leq \frac{\bar{M}\epsilon}{(72)^4 \sqrt{n}} \leq \frac{\mathbf{c}_9}{(72)^4 \epsilon \sqrt{n}}. \end{aligned} \quad (98)$$

The second inequality can be deduced by the result of Theorem 1.5 Bolley (2010) with  $d \leq d' = 4$ ,  $(0, 1) \ni \hat{\delta} = 2, p = 2, q = 4$ . Then, there exists a constant  $n_0 > 0$  such that  $n \geq n_0 \max(\epsilon^{-16}, \epsilon)$  for any  $\epsilon > 0$  and

$$\sup_{\substack{t \in \mathbb{T} \\ i \leq N}} \mathbb{P} \left[ W_2^2(\delta \mathbf{X}_{(i/n)}(t)), \nu_{(i/n)}(t) \geq \frac{\epsilon^2}{72} \right] \leq \frac{\epsilon}{(72)^4 \sqrt{n}}. \quad (99)$$

where the quantity in (100) can be derived by proceeding similarly as in Step 2.

$$\sup_{\substack{t \in \mathbb{T} \\ i \leq N}} \mathbb{E} [\|\mathbf{X}_{(i/n)}(t)\|_E^4] \leq \infty \quad (100)$$

The first term in the first inequality is direct consequence of following result:

$$\mathbb{E} [\|\mathbf{X}_{(i/n)}(t) - \mathbf{X}_{(i/n)}(s)\|_E^2] \propto |t - s|^2. \quad (101)$$

Combining all the results for the probability bounds of V, VII for deduce the upper bound in (79),

$$\begin{aligned} \mathbb{P} [W_2^2(\nu_t^n, \hat{\mu}_t) \geq \epsilon] &\leq \frac{\mathbf{c}_9}{(72)^4 \epsilon \sqrt{n}} + \frac{\mathbf{c}_9}{\epsilon^2} \exp(-4\mathbf{c}_8) \exp \left( -\frac{n\epsilon^2}{4\mathbf{c}_8} \right) \\ &\quad + \frac{2}{\kappa n} e^{-n\epsilon} \mathbf{c}_7^{3/2} \left( 1 - \frac{128\omega \mathfrak{h}(\alpha)}{n} \right)^{-\frac{d}{8}} \cdot \exp(\mathbf{c}_4 e^{\frac{1}{2}\mathbf{c}_1 T}) (e^{\kappa T} - 1). \end{aligned} \quad (102)$$

By setting  $\mathfrak{A}$  as follows, the proof is complete.

$$\mathfrak{A} = \max \left( \mathbf{c}_9, \frac{2\mathbf{c}_7^{3/2}}{\kappa} \exp(\mathbf{c}_4 e^{\frac{1}{2}\mathbf{c}_1 T}) (e^{\kappa T} - 1), \mathbf{c}_9 \exp(-4\mathbf{c}_8) \right). \quad (103)$$

□



## A.5 ALGORITHM

**Algorithm 1** Sampling Mean-field Continuous Sequence Predictors

---

```

while  $t \in \mathbb{T}$  do                                ▷ Graphon Mean-field Euler-Maruyama Sampling
  while  $i \leq N$  do
     $\{y_{u_i}\}_{i \leq N} \sim p(u, y), \Delta_t \sim p(\Delta_t), U \sim \text{Unif}(\mathbb{O}), t \sim p(t).$ 
     $\alpha_i = \alpha(t, \mathbf{X}_i^n; \theta^*), W_{ij} = W_{\alpha_i}(\lceil nu_i \rceil / n, \lceil nv_j \rceil / n), \psi_{ij} = \psi_{\alpha_i}(\mathbf{X}_i^n(t), \mathbf{X}_j^n(t)).$     (104)
     $\mathbf{X}_i^n(t + \Delta_t) = \mathbf{X}_i^n(t) + \frac{1}{n} \sum_j W_{ij} \psi_{ij} \Delta_t + \mathbf{b}(t, \mathbf{X}_i^n, \alpha_i) \Delta_t + \mathcal{N}(0_d, \sigma_t \Delta_t \mathbf{I}_d).$     (105)
  end while                                          ▷ Predict Subsequent Future Event
  if  $t \in \mathbb{T} \setminus \mathbb{O}$  then
     $\Lambda_{t+\Delta_t} = \sum_i^K w(U, \lceil nu_i \rceil / n) \mathbf{X}_i^{n, \alpha_i}(t + \Delta_t) \approx \mathbb{E}_{u \sim p(u)} \mathbf{X}_u^\alpha(t + \Delta_t)$ 
  end if
end while

```

---

**Graphon Mean-field Euler-Maruyama Sampling.** Algorithm 1 describes a discretization of the proposed infinite-order mean-field system. For a set of sampled temporal states, the proposed sampling method in Eq. (104) firstly projects the original graphon  $W_\alpha$  and interaction functions  $\psi$  onto their discrete counterparts  $W_{ij}$  and  $\psi_{ij}$ , which are referred to as *step graphons* Fabian et al. (2023) and *step interactions* in the literature. In the second phase, once the projections have been obtained, the Euler-Maruyama method is utilized to sample the trajectories of mean-field SDEs, effectively propagating information. In the prediction interval *i.e.*,  $\mathbb{T} \setminus \mathbb{O}$ , the aggregation function  $w$  is utilized to integrate the sampled particles, facilitating the generation of a forecast.

**Continuity of Temporal States.** It is worth noting that every temporality integrated into the proposed framework is completely non-uniform, as minimal temporal granularity (*i.e.*,  $\Delta_t$ ), local and global temporal states (*i.e.*,  $u, t$ ) are distributed to their corresponding probability densities defined on the continuous interval, resulting continuous representations of sequences. To operate in the described continuous setting, the neural network architecture in our framework is temporally resolution-free, which differs by exiting benchmarks, *e.g.*, Contiformer. The following list summarizes

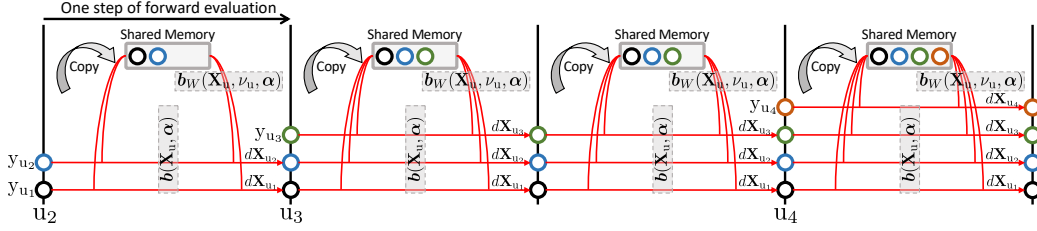


Figure 6: Parallel Computation in Sampling Mean-field Predictors

#### A.6 EXPERIMENTAL DETAILS

**Experimental Setup.** We consider  $\mathbb{T} := [0, T]$  as the entire temporal interval for each sequence instance and use the first  $\alpha\%$  of observations,  $[0, (\alpha T/100)]$ , to predict the remaining  $(1 - \alpha)\%$ ,  $[(\alpha T/100), T]$ . For this study,  $T$  is set to 100 for the MIT Humanoid Robot dataset, 48 for MIMIC-II, 72 for the Beijing Air Quality dataset, and 1200 for Eigenworm, with  $\alpha = 80$  in all cases. For every dataset we construct such input output windows on the continuous trajectories and apply per-feature  $z$ -score normalization using statistics computed on the training split, which are then reused for validation and test, and all models are evaluated primarily by mean squared error (MSE) on the prediction horizon. Our MFP is trained with the Adam optimizer using batch size 128 and learning rate  $10^{-4}$ ; the neural graphon and control networks together contain on the order of one million parameters, while competing methods use the architectures and hyperparameters recommended in their original papers rather than being artificially matched by parameter count. All models, including baselines, are trained for 10 000 epochs without early stopping, and reported results are averaged over multiple runs with different random seeds on a single GPU. In our mean-field game interpretation, the fictitious-play index  $m$  coincides with the number of Adam gradient steps on the control parameters  $\alpha$ , so that these 10 000 epochs realize 10 000 fictitious-play updates and provide a high-resolution approximation of the HJB component in Eq. (5) for this high-dimensional setting.

**Model Architecture.** In each forward step of  $\mathbf{X}_u(t)$  from  $t$  to  $t + \Delta t$ , a neural network takes  $\mathbf{X}_u(t)$ ,  $t$ , and  $u$  as inputs and outputs  $\mathbf{b}(\cdot, \alpha)$ ,  $\mathbf{W}(\alpha)$ , and  $w$ . In the first stage of the neural network,  $\mathbf{X}_u(t)$  and  $t$  are concatenated into a single vector, which is then projected into a hidden vector via a multilayer perceptron (MLP). This hidden vector is subsequently passed through a computation block consisting of several MLP layers with skip connections. Finally, after the computation block, the hidden vector is projected into  $\mathbf{b}(\cdot, \alpha)$ ,  $\mathbf{W}(\alpha)$ , and  $w$  using respective MLPs. In our architecture, each MLP is composed of two linear layers, with a Swish activation function positioned between them.

To process the labeling information  $u$  in the neural network, we apply adaptive normalization (Peebles & Xie, 2023). Specifically, instead of using fixed scale and shift parameters in the normalization layers of  $\alpha(\cdot; \theta)$ , we regress these parameters based on  $u$ . The adaptive normalization layers are placed between MLP layers. We find that this conditioning mechanism effectively incorporates the labeling information, outperforming the approach of simply concatenating  $u$  into input vectors.

After obtaining outputs from the neural networks, we evaluate  $\mathbf{b}_W(\cdot, \alpha)$  for forward evaluation of SDEs. To derive  $\mathbf{b}_W(\cdot, \alpha)$ , we compute an exponential or cosine graphon  $W$  using  $u$  and  $v$  where  $v < t$ . Next we calculate the projection  $\text{Proj}(x - y) := (x - y) / \|x - y\|$  with  $x = \mathbf{X}_u(t)$  and  $y = \mathbf{X}_{v < t}(t)$ . These values are then integrated into with  $\mathbf{W}(\alpha)$  using Eq (2) into  $\mathbf{W}_\alpha$  and  $\psi_\alpha$ , finally leading to  $\mathbf{b}_W(\cdot, \alpha) = \sum_{v < t} \psi_\alpha(\mathbf{X}_u(t), \mathbf{X}_v(t)) \mathbf{W}_\alpha(u, v)$ . After forward evaluation, we utilize  $w$  to aggregate predictors by applying softmax. (i.e.,  $\Lambda_t = \sum_{v < t} \text{Softmax}(w(u, v); \{w(t, u)\}_{u < t}) \mathbf{X}_u(t)$  where  $\text{Softmax}(x \in S; S)$  represents the value of  $x$  after applying the softmax operation to the entire set  $S$  which includes  $x$ .)

**Parallel Computation.** Since the direct application of Alg. 1 is computationally intractable for large particle count  $N$ , we introduce novel parallel computing to efficiently sample proposed mean-field predictors, as described in Fig ?? . At each step of forward evaluation, given all predictors  $\mathbf{X}_u^\alpha$ , each predictor can be processed independently using Eq (1). In other words, no predictor needs to wait for the others to complete their forward evaluation. By taking advantage of this property, at time  $t$ , we store all predictors with  $u \leq t$  in the shared memory and forward predictors one step in parallel. This

parallel implementation significantly decreases empirical computation time by reducing the number of iterations for forward evaluation from  $\mathcal{O}(SN)$  to  $\mathcal{O}(S)$  where  $S$  is the number of steps for forward evaluation and  $N$  is the number of sampled observations.