Frontiers of Mathematical Finance
Vol. 1, No. 2, June 2022, pp. 287–319
doi:10.3934/fmf.2021011

CONVERGENCE OF DEEP FICTITIOUS PLAY FOR STOCHASTIC DIFFERENTIAL GAMES

Jiequn Han

Center for Computational Mathematics
Flatiron Institute, 162 5th Avenue, New York, NY, USA
Department of Mathematics
Princeton University, Princeton, NJ, USA

Ruimeng Hu*

Department of Mathematics, and Department of Statistics and Applied Probability
University of California, Santa Barbara, CA, USA

Jihao Long

The Program in Applied and Computational Mathematics
Princeton University, Princeton, NJ, USA

(Communicated by Xin Guo)

Abstract. Stochastic differential games have been used extensively to model agents' competitions in finance, for instance, in P2P lending platforms from the Fintech industry, the banking system for systemic risk, and insurance markets. The recently proposed machine learning algorithm, deep fictitious play, provides a novel and efficient tool for finding Markovian Nash equilibrium of large N-player asymmetric stochastic differential games [J. Han and R. Hu, Mathematical and Scientific Machine Learning Conference, pages 221-245, PMLR, 2020]. By incorporating the idea of fictitious play, the algorithm decouples the game into N sub-optimization problems, and identifies each player's optimal strategy with the deep backward stochastic differential equation (BSDE) method parallelly and repeatedly. In this paper, we prove the convergence of deep fictitious play (DFP) to the true Nash equilibrium. We can also show that the strategy based on DFP forms an ϵ-Nash equilibrium. We generalize the algorithm by proposing a new approach to decouple the games, and present numerical results of large population games showing the empirical convergence of the algorithm beyond the technical assumptions in the theorems.

1. **Introduction.** Deep neural networks have become popular and powerful tools in scientific computing, for their remarkable performance in approximating high-dimensional functions. Their successes have brought natural applications in stochastic differential games, where high-dimensional optimization problems and/or stochastic differential equations are solved to model and analyze tactical interactions among multiple decision-makers in the context of a random dynamical system. These decision-makers, usually referred to as players or agents, can interact in a

2020 *Mathematics Subject Classification.* Primary: 91A15, 68T07; Secondary: 60H10, 60H30.
Key words and phrases. Deep fictitious play, convergence analysis, stochastic differential games, Markovian Nash equilibrium, backward stochastic differential equations.
R.H. was partially supported by the NSF grant DMS-1953035.
* Corresponding author: Ruimeng Hu.

manner ranging from completely non-cooperative to completely cooperative. The nature of uncertainty makes stochastic differential games appropriate to be used for the study of competitions in finance, *e.g.*, in P2P lending platforms [67, 49] from the Fintech industry and insurance markets [70, 7, 18].

For non-cooperative stochastic differential games, a core problem is to compute the associated Nash equilibrium, which refers to a set of strategies so that when applied, no player will profit from unilaterally changing her own choice. When the games involve heterogeneous agents of moderate size, *e.g.*, $5 \leq N \leq 100$, computing the Nash equilibrium becomes numerically challenging since conventional numerical algorithms lose their efficiency for N beyond 5, and the game with $N \leq 100$ asymmetric players is not yet well approximated by a mean-field framework.

To address the challenge, the authors have recently proposed the deep fictitious play (DFP) algorithms [31], providing a novel and efficient tool for finding Markovian Nash equilibrium of large N-player asymmetric stochastic differential games. However, despite the efficient performance in simulation, the algorithm's theoretical foundation is still lacking, which will be the focus of this paper. In addition, we generalize the previous algorithms, and propose a general two-step scheme: The first step aims to recast the game into N sub-problems that will be repeatedly solved. The desired algorithm requires that, after the recast, the sub-problems are decoupled among different players given the previous stage's solutions, and that their solutions converge to the true Nash equilibrium. Specifically, we propose two options for the first step:

I. Fictitious play. This approach was used in [31], assuming that players are myopic and will choose their best responses against others' previous stage action at every subsequent stage. Therefore each player still faces a nonlinear optimization problem.
II. Policy update. We calculate the game values using all responses from the previous stage, and the current stage responses are determined as if they are the optimizers of the calculated game values.

The second step of the DFP algorithm aims to solve the sub-problems efficiently and accurately. Remark that, due to a large number of players and the high dimensionality of the controlled state process, each sub-problem may still be high-dimensional after the decoupling step. In [31], the Deep BSDE method was employed for each sub-problem, which presents excellent performance. The Deep BSDE method relies on the BSDE representations of semi-linear partial differential equations (PDEs) and deep learning approximations after discretizing the BSDE by an Euler scheme. The method parametrizes the initial position of the backward process and the adjoint process by DNNs, then simulates both processes in a forward manner, aiming to minimize the discrepancy between the terminal value of the backward process and its network approximation. The analysis for the second step shall focus on this method. Meanwhile, we remark that other deep neural networks (DNNs) based algorithms, such as deep learning backward dynamic programming (DBDP) method [42] and deep Galerkin method [66], are also promising choices for solving sub-problems.

Related literature. The theoretical study of differential games was initiated by R. Isaacs in the early 1960s [44]. Later on, to better describe real world's uncertainties, noises are added to the state of the system, and stochastic differential games have now been intensively used across many disciplines. Domains of applications include management science (*e.g.*, operations management, marketing, finance, systemic

risk), economics (*e.g.*, industrial organization, environmental and macroeconomics, production of exhaustible resources), social science (*e.g.*, networks, crowd behavior, congestion), biology (*e.g.*, flocking), and military (*e.g.*, cyber-attacks).

Fictitious play is well documented in the economics literature, as a learning process for finding Nash equilibria. It was firstly proposed by [11, 12] for normal-form games. Since then, there have been extensive studies on the convergence of fictitious play or its variation under different settings; for instance, see [54, 55, 48, 36, 8]. For stochastic differential games, besides [31], the most related work is [41], where fictitious play is used to design numerical algorithms for finding open-loop Nash equilibria. We remark that, the idea of fictitious play is not limited to study the games with a moderate number of heterogeneous players [41, 31], but has also been applied in mean-field games, *e.g.*, see [13, 10, 22].

The proposed policy update for the first step of the DFP algorithm closely follows policy iteration (PI) in spirit, which was initially introduced by Howard [39] for discounted Markovian decision problems (MDP). It consists of two steps: policy evaluation (obtaining the expected reward for a given policy) and policy improvement (updating the policy using the rewards for successor states). PI was later generalized to modified PI in [64], and has remained the method of choice in designing reinforcement learning algorithms, *e.g.*, see [27, 62] and the references therein. Recently this technique has also been utilized to solve mean-field games numerically [29, 65].

The literature of using DNNs for learning high-dimensional function is rich, including methods for solving high-dimensional parabolic PDEs and BSDEs (*e.g.*, the deep BSDE method [20, 32], the DBDP [42, 24], and many others [66, 5, 6, 61, 69, 45]). It also yields algorithms for solving the Schrödinger equation [35, 60, 34], stochastic control problems [30, 56], mean field games [17, 1] and nonlinear optimal stopping problems [40].

Main contribution. The contribution of this paper consists of the following: 1. We propose a general two-step scheme that extends the original deep fictitious play algorithm [31], and provide two options for solving the first step. The proposed algorithm can efficiently solve stochastic differential games with heterogeneous agents of large size (*e.g.*, $5 \leq N \leq 100$), and the presence of common noise. 2. We provide the theoretical foundation for the proposed algorithms. Specifically, we prove that the solutions to the decoupled sub-problems, if solved repeatedly and exactly at each stage, converge to the true Nash equilibrium; that the numerical solutions to each sub-problem tend to be exact as we refine the time step in the Euler scheme; and that the strategy based on numerical solutions forms an ϵ-Nash equilibrium, after running sufficiently many stages and using sufficiently fine time step. 3. We present numerical results showing empirical convergence even beyond the technical assumptions used in the theorems.

The rest of this paper is organized as follows. In Section 2, we give the mathematical formulation of general N-player asymmetric stochastic games in continuous time. The algorithms consisting of the decoupling step and sub-problem-solving step via deep learning are detailed in Section 3. Section 4 provides convergence analysis for the proposed algorithms, followed by numerical examples presented in Section 5. We make conclusive remarks in Section 6.

2. Mathematical formulation.

Throughout the paper, we shall use the following notations:

- A boldface character refers to a collection of objects from all players;
- A regular character with a superscript i refers to an objective from player i (no matter a scalar or a vector) or the i^{th} column of a vector;
- A boldface character with a superscript $-i$ refers to a collection of objects from all players except i;
- The state process $\boldsymbol{X}_t$ introduced below is a common process to all players, and always in boldface.

We consider a general N-player non-zero-sum stochastic differential games. An $\mathbb{R}^n$-valued *common* state process $\boldsymbol{X}_t$ is controlled by a Markovian strategy/policy[1] $\boldsymbol{\alpha}$:

$$\mathrm{d}\boldsymbol{X}_t^{\boldsymbol{\alpha}} = b(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}}))\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}})\,\mathrm{d}\boldsymbol{W}_t, \quad \boldsymbol{X}_0 = \boldsymbol{x}_0, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^N)$ is the collection of all players' $\mathcal{A}^i$-valued strategies. For simplicity, we assume $\mathcal{A}^i = \mathbb{R}^{d_\alpha}$ for $i = 1, 2, \ldots, N$. If not (*e.g.*, some boundedness constraints are put on α^i), we can assume there exist Lipschitz mappings P_α^i from $\mathbb{R}^{d_\alpha}$ to $\mathcal{A}^i$ so that $\mathcal{A}^i = P_\alpha^i(\mathbb{R}^{d_\alpha})$, and all the statements below hold easily with the help of the Lipschitz mappings. The drift and diffusion coefficients b and Σ are deterministic functions of the common state, $b\colon [0,T] \times \mathbb{R}^n \times \mathcal{A} \hookrightarrow \mathbb{R}^n$, $\Sigma\colon [0,T] \times \mathbb{R}^n \hookrightarrow \mathbb{R}^{n\times k}$, where $\mathcal{A} = \otimes_{i=1}^N \mathcal{A}^i = \mathbb{R}^{Nd_\alpha}$ is the space for the joint control $\boldsymbol{\alpha}$, and $\boldsymbol{W}$ is a k-dimensional standard Brownian motion on a filtered probability space $(\Omega, \mathbb{F}, \{\mathcal{F}_t\}_{0\le t\le T}, \mathbb{P})$.

Player i aims at minimizing her expected total cost:

$$\inf_{\alpha^i \in \mathbb{A}^i} \mathbb{E}[\int_0^T f^i(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}))\,\mathrm{d}s + g^i(\boldsymbol{X}_T^{\boldsymbol{\alpha}})] \tag{2}$$

within the set of admissible strategies $\mathbb{A}^i$:

$$\mathbb{A}^i = \left\{\alpha^i(t, \boldsymbol{x}) :\ \text{Borel measurable function } [0,T] \times \mathbb{R}^n \hookrightarrow \mathbb{R}^{d_\alpha}\right\},$$

where the running cost $f^i\colon [0,T]\times\mathbb{R}^n\times\mathcal{A} \hookrightarrow \mathbb{R}$ and the terminal cost $g^i\colon \mathbb{R}^n \hookrightarrow \mathbb{R}$ are deterministic measurable functions. Obviously, the quantity in (2) is also affected by other players' strategies α^j. To emphasize this dependence, we introduce the notation $J_t^i(\alpha^1, \ldots, \alpha^N)$ for the cost of player i starting at t when players choose their strategies $(\alpha^1, \ldots, \alpha^N)$:

$$J_t^i(\alpha^1, \ldots, \alpha^N) \equiv J_t^i(\boldsymbol{\alpha}) := \mathbb{E}[\int_t^T f^i(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}))\,\mathrm{d}s + g^i(\boldsymbol{X}_T^{\boldsymbol{\alpha}})]. \tag{3}$$

In the following sections, we shall present the algorithms for solving the above game and prove its theoretical convergence. In particular, we are interested in finding a Markovian Nash equilibrium (or the Markovian ϵ-Nash equilibrium).

Definition 2.1. A Markovian ϵ-Nash equilibrium is a tuple $\boldsymbol{\alpha}^\epsilon = (\alpha^{1,\epsilon}, \ldots, \alpha^{N,\epsilon}) \in \mathbb{A}$, such that, for non-negative ϵ,

$$\forall i \in \mathcal{I},\ \text{and } \alpha^i \in \mathbb{A}^i, \quad J_0^i(\boldsymbol{\alpha}^\epsilon) - \epsilon \le J_0^i(\alpha^{1,\epsilon}, \ldots, \alpha^{i-1,\epsilon}, \alpha^i, \alpha^{i+1,\epsilon}, \ldots, \alpha^{N,\epsilon}).$$

A Markovian Nash equilibrium, denoted by $\boldsymbol{\alpha}^*$, is equivalent to an ϵ-Nash equilibrium where $\epsilon = 0$. Here $\mathbb{A} = \otimes_{i=1}^N \mathbb{A}^i$ is the product space of $\mathbb{A}^i$, and $\mathcal{I} = \{1, 2, \ldots, N\}$ is the set of all players.

[1]Hereafter, we shall use *strategy* and *policy* interchangeably.

As discussed in [31], the formulation (1)–(2) is less restrictive than the usual case where player i can only control her *private* state. Here, a common state $\boldsymbol{X}_t$ is controlled by all agents, as a common feature in economics literature (see *e.g.*, [19, 63, 50]). Therefore, it is important to include it in our framework, although this will increase the coupling and make the problem harder to solve, both theoretically and numerically. Remark that the difficulty still persists in the limiting problem as $N \to \infty$ with indistinguishable players, when allowing α^i entering into others' states. This is called the extended mean-field game and it has attracted certain attention recently (*e.g.*, [25, 26, 14]). On the other hand, by choosing b and Σ in (1) properly, one can reduce the formulation (1) to the simpler case where each player controls her private state through α^i. For instance, if each player's private state is d-dimensional, we can let $n = dN$, $b = (b^1, \dots, b^\ell, \dots, b^n)$ with $b^\ell \equiv b^\ell(t, \boldsymbol{x}, \alpha^i)$ for $\ell = (i-1)d+1, \dots, id$, then the problem (1)–(2) is the standard modeling in literature in many disciplines including social science, management science and engineering, with the i^{th} player's d-dimensional private state $(X_t^{(i-1)d+1}, \dots, X_t^{id})$ controlled by α^i only.

In the Markovian setting, the value function of player i reads as:

$$V^i(t, \boldsymbol{x}) = \inf_{\alpha^i \in \mathbb{A}^i} \mathbb{E}[\int_t^T f^i(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}))\, \mathrm{d}s + g^i(\boldsymbol{X}_T^{\boldsymbol{\alpha}}) \Big| \boldsymbol{X}_t^{\boldsymbol{\alpha}} = \boldsymbol{x}].$$

Then, to compute the Markovian Nash equilibrium, we apply the dynamic programming principle and obtain a system of Hamilton-Jacobi-Bellman (HJB) equations:

$$\begin{cases} V_t^i + \inf_{\alpha^i \in \mathcal{A}^i} \left\{ b(t, \boldsymbol{x}, \boldsymbol{\alpha}) \cdot \nabla_{\boldsymbol{x}} V^i + f^i(t, \boldsymbol{x}, \boldsymbol{\alpha}) \right\} + \frac{1}{2} \mathrm{Tr}(\Sigma^\intercal \mathrm{Hess}_{\boldsymbol{x}} V^i \Sigma) = 0, \\ V^i(T, \boldsymbol{x}) = g^i(\boldsymbol{x}), \quad i \in \mathcal{I}, \end{cases} \tag{4}$$

where V_t^i, $\nabla_{\boldsymbol{x}} V^i$, $\mathrm{Hess}_{\boldsymbol{x}} V^i$ denote the derivative of V^i with respect to t, the gradient and the Hessian of function V^i with respect to $\boldsymbol{x}$, and Tr denotes the trace of a matrix. Note that the system (4) is coupled, as each minimizer $\alpha^{i,*}$ depends on V^i while b and f^i in (4) depend on all minimizers $\boldsymbol{\alpha}^* = (\alpha^{1,*}, \dots, \alpha^{N,*})$.

Under appropriate conditions, the solution to (4) is related to BSDEs, using nonlinear Feynman-Kac formula (*cf.* [57, 21, 58]). To ease our notations, we prescribe the following the relation on b and Σ.

Assumption 1. *There exists a measurable function ϕ: $[0,T] \times \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}^k$, so that $\Sigma(t, \boldsymbol{x})\phi(t, \boldsymbol{x}, \boldsymbol{\alpha}) = b(t, \boldsymbol{x}, \boldsymbol{\alpha})$ for any $(t, \boldsymbol{x}, \boldsymbol{\alpha}) \in [0,T] \times \mathbb{R}^n \times \mathcal{A}$.*

Consequently, we can define the Hamiltonian function $\boldsymbol{H}(t, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{p}) : [0,T] \times \mathbb{R}^n \times \mathcal{A} \times \mathbb{R}^{k \times N} \to \mathbb{R}^N$ by:

$$\boldsymbol{H} = [H^1, \dots, H^N]^\intercal, \quad H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}, p^i) = \phi(t, \boldsymbol{x}, \boldsymbol{\alpha}) \cdot p^i + f^i(t, \boldsymbol{x}, \boldsymbol{\alpha}), \tag{5}$$

where $p^i \in \mathbb{R}^k$ denotes the i^{th} column of $\boldsymbol{p}$. Using this notation, the HJB system can be rewritten as:

$$V_t^i + \inf_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}, \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i) + \frac{1}{2} \mathrm{Tr}(\Sigma^\intercal \mathrm{Hess}_{\boldsymbol{x}} V^i \Sigma) = 0, \quad \forall i \in \mathcal{I}. \tag{6}$$

To better describe the optimal game policies, we define $\boldsymbol{a}(t, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{p}) : [0,T] \times \mathbb{R}^n \times \mathcal{A} \times \mathbb{R}^{k \times N} \to \mathcal{A}$ by:

$$\boldsymbol{a} = (a^1, \dots, a^N), \quad a^i(t, \boldsymbol{x}, \boldsymbol{\alpha}^{-i}, p^i) = \operatorname*{arg\,min}_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, (\alpha^i, \boldsymbol{\alpha}^{-i}), p^i), \ \forall i \in \mathcal{I}. \tag{7}$$

In other words, a^i is the minimizer of the i^{th} Hamiltonian, emphasizing the dependence on the i^{th} player's game value $\Sigma^\intercal \nabla_{\boldsymbol{x}} V^i$ and others' strategies $\boldsymbol{\alpha}^{-i}$. Then, we define a function $\boldsymbol{\alpha}(t, \boldsymbol{x}, \boldsymbol{p})$ as the fixed point of

$$\boldsymbol{\alpha} = \boldsymbol{a}(t, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{p}). \tag{8}$$

Note that, with the above notations $\boldsymbol{a}$ and $\boldsymbol{\alpha}$, we have assumed the minimizer in (7) exists and is unique, and (8) has a unique fixed point. Later in Assumption 2, we will detail explicit conditions on the model parameters, such that these assumptions are satisfied.

We now state the corresponding BSDE formulation of (4), which is the key component of the algorithm design in Section 3 and the convergence analysis in Section 4. Let $(\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{Z}_t) \in \mathbb{R}^n \times \mathbb{R}^N \times \mathbb{R}^{k\times N}$ be the solution to the following BSDE:

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{x}_0 + \displaystyle\int_0^t \Sigma(s, \boldsymbol{X}_s)\, \mathrm{d}\boldsymbol{W}_s, \\ \boldsymbol{Y}_t = \boldsymbol{g}(\boldsymbol{X}_T) + \displaystyle\int_t^T \check{\boldsymbol{H}}(s, \boldsymbol{X}_s, \boldsymbol{Z}_s)\, \mathrm{d}s - \int_t^T \boldsymbol{Z}_s^\intercal\, \mathrm{d}\boldsymbol{W}_s, \end{cases} \tag{9}$$

where $\check{\boldsymbol{H}}(t, \boldsymbol{x}, \boldsymbol{p}) := \boldsymbol{H}(t, \boldsymbol{x}, \boldsymbol{\alpha}(t, \boldsymbol{x}, \boldsymbol{p}), \boldsymbol{p})$ is the minimized Hamiltonian vector, and $\boldsymbol{g}(\boldsymbol{x}) \equiv [g^1, \dots, g^N]^\intercal(\boldsymbol{x})$ is the vector form of all terminal costs. Then we have the relation:

$$\begin{aligned} &\boldsymbol{Y}_t = [Y_t^1, \dots, Y_t^N]^\intercal, \quad Y_t^i = V^i(t, \boldsymbol{X}_t), \\ &\boldsymbol{Z}_t = [Z_t^1, \dots, Z_t^N], \quad Z_t^i = \Sigma^\intercal(t, \boldsymbol{X}_t) \nabla_{\boldsymbol{x}} V^i(t, \boldsymbol{X}_t), \end{aligned} \tag{10}$$

and the optimal game policy is expressed by $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t)$. Using the relation (10), we notice $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \Sigma^\intercal(t, \boldsymbol{X}_t) \nabla_{\boldsymbol{x}} \boldsymbol{V}(t, \boldsymbol{X}_t))$ where $\boldsymbol{V} := [V^1, \dots, V^N]$[2], and sometimes write $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}^*(t, \boldsymbol{X}_t)$.

Remark 1. Note that the process $\boldsymbol{X}_t$ in (9) does not allude to $b = 0$ in the controlled dynamics $\boldsymbol{X}_t^{\boldsymbol{\alpha}}$ defined in (1). Indeed, it is an auxiliary forward stochastic process derived from the HJB system (6) using the nonlinear Feynman-Kac formula, which is an object different from the controlled process $\boldsymbol{X}_t^{\boldsymbol{\alpha}}$ in equation (1). One, of course, has the flexibility to choose a different forward process with nonzero drift:

$$\begin{cases} \tilde{\boldsymbol{X}}_t = \boldsymbol{x}_0 + \displaystyle\int_0^t \Sigma(s, \tilde{\boldsymbol{X}}_s) \mu(s, \tilde{\boldsymbol{X}}_s)\, \mathrm{d}s + \int_0^t \Sigma(s, \tilde{\boldsymbol{X}}_s)\, \mathrm{d}\boldsymbol{W}_s, \\ \tilde{\boldsymbol{Y}}_t = \boldsymbol{g}(\tilde{\boldsymbol{X}}_T) + \displaystyle\int_t^T \check{\boldsymbol{H}}(s, \tilde{\boldsymbol{X}}_s, \tilde{\boldsymbol{Z}}_s) - \tilde{\boldsymbol{Z}}_s^\intercal \mu(t, \tilde{\boldsymbol{X}}_s)\, \mathrm{d}s - \int_t^T \tilde{\boldsymbol{Z}}_s^\intercal\, \mathrm{d}\boldsymbol{W}_s, \end{cases}$$

and to express the solution to (6) via (10) with all (X, Y, Z) replaced by $(\tilde{X}, \tilde{Y}, \tilde{Z})$. This is essentially rewriting equation (6) to

$$\begin{aligned} V_t^i + \inf_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}, \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i) - \mu(t, \boldsymbol{x}) \cdot \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i + \mu(t, \boldsymbol{x}) \cdot \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i \\ + \frac{1}{2} \mathrm{Tr}(\Sigma^\intercal \mathrm{Hess}_{\boldsymbol{x}} V^i \Sigma) = 0, \quad \forall i \in \mathcal{I}, \end{aligned}$$

and take $\inf_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}, \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i) - \mu(t, \boldsymbol{x}) \cdot \Sigma^\intercal \nabla_{\boldsymbol{x}} V^i$ as the driver. Note that due to the coupling in (6), $\mu(t, \boldsymbol{x})$ needs to be identical across all $i \in \mathcal{I}$. Nevertheless, we think the choice in (9) is the most natural one, without additional knowledge of $\check{\boldsymbol{H}}$.

[2]We use $\nabla_{\boldsymbol{x}} \boldsymbol{V}$ as an $n \times N$ matrix.

A second remark is that, with Assumption 1 one can apply a change of measure and make the controlled dynamics driftless as in (9) under a different measure, which is indeed used in the proof of Theorem 4.

If solving directly, no matter which system ((4) or (9)), one will encounter computational difficulties due to the high dimensionality of $\boldsymbol{X}_t$ or the large number of agents. To overcome this, we propose a two-step scheme in Section 3, where we generalize the idea in [31] and offer two options for the first step. The convergence analysis with appropriate assumptions will be presented in Section 4.

3. Algorithm.

The two-step scheme for solving Markovian Nash equilibrium works as follows. We first decouple the problem (1)–(2) into N independent sub-problems, for which we need to solve repeatedly and can solve in a parallel manner. Since each sub-problem may still be high-dimensional, we then solve each using deep neural networks with a reformulation in backward stochastic differential equations (BSDEs). Next, we describe the algorithms of each step in detail.

3.1. Step I: Decoupling.

This step aims to decentralize the game, converting it into single-agent problems to be solved repeatedly. The algorithms start with an initial guess of the Nash equilibrium $\boldsymbol{\alpha}^0 = [\alpha^{1,0}, \dots, \alpha^{N,0}]$ and produce a sequence of strategies afterward, which we denote by $\boldsymbol{\alpha}^1, \dots \boldsymbol{\alpha}^m, \dots$. The following two options at this step differ in how the sequence is determined. Notationwise, $\boldsymbol{\alpha}^m$ refers to the collection of all players' policies at stage m, and its i^{th} component $\alpha^{i,m}$ refers to player i's choice.

1. **Fictitious Play.** In this option of Step I, at each stage, each player faces an optimization problem (2) while assuming that others are using their strategies from the previous stage as fixed strategies. In other words, at stage $m+1$, $\boldsymbol{\alpha}^m$ is known to all players, and player i's decision problem is

$$\inf_{\alpha^i \in \mathbb{A}^i} J_0^i(\alpha^i; \boldsymbol{\alpha}^{-i,m}), \tag{11}$$

where J_0^i is defined in (3), and the state process $\boldsymbol{X}_t$ follows (1) with $\boldsymbol{\alpha}$ being replaced by $(\alpha^i; \boldsymbol{\alpha}^{-i,m})$. Here $\boldsymbol{\alpha}^{-i,m}$ represents the strategies of all players but player i at stage m, and $(\alpha^i; \boldsymbol{\alpha}^{-i,m})$ is a short notation of $(\alpha^{1,m}, \dots, \alpha^{i-1,m}, \alpha^i, \alpha^{i+1,m}, \alpha^{N,m})$, which emphasis the parameter role of $\boldsymbol{\alpha}^{-i,m}$.

Under the Markovian framework, we denote by $V^{i,m+1}$ the problem value of player i at stage m. Following the idea of fictitious play, it is the solution of the following HJB system

$$\begin{cases} V_t^{i,m+1} + \inf\limits_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, (\alpha^i, \boldsymbol{\alpha}^{-i,m})(t, \boldsymbol{x}), \Sigma^\intercal \nabla_{\boldsymbol{x}} V^{i,m+1}) \\ \qquad\qquad + \dfrac{1}{2} \mathrm{Tr}(\Sigma^\intercal \mathrm{Hess}_{\boldsymbol{x}} V^{i,m+1} \Sigma) = 0, \\ V^{i,m+1}(T, \boldsymbol{x}) = g^i(\boldsymbol{x}). \end{cases} \tag{12}$$

This option, combined with Step II introduced below, is exactly the deep fictitious play algorithm proposed in [31]: at stage $m+1$, players myopically respond to their opponents' policy at stage m without considering all decisions before stage m. As explained in [31, Remarks 3.1 and 3.2], this is a bit discrepant from Brown's original definition [11, 12], where players response take into account all past policies $\boldsymbol{\alpha}^{-i,0}, \dots \boldsymbol{\alpha}^{-i,m}$. And the very reason to use the last stage policy is that, it otherwise requires tracking all past functions

$\boldsymbol{\alpha}^{-i,0}, \dots \boldsymbol{\alpha}^{-i,m}$ due to direct feedback nature, which is computationally infeasible. Serving as the theoretical foundation of [31], we follow the terminology therein and term this decoupling option as fictitious play.

2. **Policy Update.** This is slightly different from the fictitious play, where *every* player follows her strategy from the previous stage to update the problem value. In this case, it is no longer an optimization, but a linear problem for the value function induced by the fix strategy $\boldsymbol{\alpha}^m$:

$$\begin{cases} V_t^{i,m+1} + H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}^m(t,\boldsymbol{x}), \Sigma^\intercal \nabla_{\boldsymbol{x}} V^{i,m+1}) + \dfrac{1}{2}\mathrm{Tr}(\Sigma^\intercal \mathrm{Hess}_{\boldsymbol{x}} V^{i,m+1} \Sigma) = 0, \\ V^{i,m+1}(T, \boldsymbol{x}) = g^i(\boldsymbol{x}). \end{cases} \tag{13}$$

After solving out the decoupled PDE (12) or (13), at the end of stage $m+1$, a policy $\alpha^{i,m+1}$ is determined by

$$\alpha^{i,m+1}(t,\boldsymbol{x}) = \arg\min_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, (\alpha^i, \boldsymbol{\alpha}^{-i,m})(t,\boldsymbol{x}), \Sigma^\intercal \nabla_{\boldsymbol{x}} V^{i,m+1}(t,\boldsymbol{x})), \tag{14}$$

and policies from all players together form $\boldsymbol{\alpha}^{m+1}$.

Note that for fictitious play algorithms, $\alpha^{i,m+1}$ is indeed the optimal strategy of problem (11); while for policy update algorithms, the problem is linear, but we pretend that $V^{i,m+1}$ is the value of an optimization problem, and $\alpha^{i,m+1}$ is determined as if it is an optimizer. In short, the two algorithms differ at how $\boldsymbol{\alpha}^{m+1}$ is updated from $\boldsymbol{\alpha}^m$. When interpreting via BSDEs, the different update rules result in slightly different drivers of the backward components, see equations (15) and (17) below. Nevertheless, the analysis based on the two algorithms presented in Theorems 2 and 3 follows similarly.

3.2. **Step II: Solving each sub-problem via BSDE.** We write down the BSDE counterpart of the sub-problem (12) in the fictitious play:

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{x}_0 + \displaystyle\int_0^t \Sigma(s, \boldsymbol{X}_s)\, \mathrm{d}\boldsymbol{W}_s, \\ Y_t^{i,m+1} = g^i(\boldsymbol{X}_T) + \displaystyle\int_t^T \hat{H}^i(s, \boldsymbol{X}_s, \boldsymbol{\alpha}^{-i,m}(s,\boldsymbol{X}_s), Z_s^{i,m+1})\, \mathrm{d}s - \int_t^T (Z_s^{i,m+1})^\intercal\, \mathrm{d}\boldsymbol{W}_s, \end{cases} \tag{15}$$

where $\hat{H}^i$ is defined by

$$\hat{H}^i(t, \boldsymbol{x}, \boldsymbol{\alpha}^{-i}, p^i) = H^i(t, \boldsymbol{x}, (a^i(t, \boldsymbol{x}, \boldsymbol{\alpha}^{-i}, p^i), \boldsymbol{\alpha}^{-i}), p^i), \tag{16}$$

or the BSDE counterpart of the sub-problem (13) in the policy update:

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{x}_0 + \displaystyle\int_0^t \Sigma(s, \boldsymbol{X}_s)\, \mathrm{d}\boldsymbol{W}_s, \\ Y_t^{i,m+1} = g^i(\boldsymbol{X}_T) + \displaystyle\int_t^T H^i(s, \boldsymbol{X}_s, \boldsymbol{\alpha}^m(s,\boldsymbol{X}_s), Z_s^{i,m+1})\, \mathrm{d}s - \int_t^T (Z_s^{i,m+1})^\intercal\, \mathrm{d}\boldsymbol{W}_s. \end{cases} \tag{17}$$

Here $\boldsymbol{x}_0$ is a random variable whose range covers the states of interest.

Remark 2. Note that, in both BSDEs above, we choose the forward process without a drift term for three reasons: (a) it avoids the involvement of $\boldsymbol{\alpha}^m$, and thus keeps the forward process the same from stage to stage; (b) the BSDEs can be vectorized (cf. (30)) with a single forward process which coincides with the forward component of (9) (corresponding to the true solution), both will facilitate our analysis (c) numerically, this means only one forward process needs to be simulated for

all N sub-problems, which makes one iteration of step I–II more efficient. Once the driftless BSDEs (15) and (17) are solved numerically accurate, it is proved, in theorems in Section 4.3, that the performance on the optimal control process $\boldsymbol{X}^{\boldsymbol{\alpha}^*}$ (cf. (1) with $\boldsymbol{\alpha}^*$ replaced by the Nash equilibrium strategy $\boldsymbol{\alpha}^*$) is also well.

We also remark that both BSDEs are wellposed under Assumptions 1–2, as the drivers H^i and $\hat{H}^i$ are uniformly Lipschitz in p^i and $g^i(\boldsymbol{X}_T)$ is square integrable (cf. [71, Theorem 4.3.1]).

For both sub-problems, the connection between the associated BSDEs and PDEs are the same:

$$Y_t^{i,m+1} = V^{i,m+1}(t, \boldsymbol{X}_t), \quad Z_t^{i,m+1} = \Sigma(t, \boldsymbol{X}_t)^\intercal \nabla_{\boldsymbol{x}} V^{i,m+1}(t, \boldsymbol{X}_t),$$

and according to (14), both optimal policy processes at stage $m+1$ are expressed by

$$\alpha_t^{i,m+1} = a^i(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^{-i,m}, Z_t^{i,m+1}).$$

Therefore, it suffices to solve these two possibly high-dimensional BSDE systems by an efficient algorithm, which we shall describe and call deep BSDE in the sequel. To avoid repetition and cumbersome notation, the algorithms will be presented on a generic BSDE with possibly non-zero drift term:

$$\begin{cases} X_t = x_0 + \displaystyle\int_0^t \mu(s, X_s)\,\mathrm{d}s + \int_0^t \Sigma(s, X_s)\,\mathrm{d}W_s, \\ Y_t = g(X_T) + \displaystyle\int_t^T F(s, X_s, Z_s)\,\mathrm{d}s - \int_t^T Z_s^\intercal\,\mathrm{d}W_s. \end{cases} \tag{18}$$

The algorithm applied to the exact system (15) and (17) will be presented in Section 4.2.

The deep BSDE is firstly introduced in [20], for solving high-dimensional parabolic PDEs. The idea is to solve a variational form of (18) after temporal discretizations using deep neural networks. For a partition π of size N_T on the time interval $[0,T], 0 = t_0 < t_1 < \ldots < t_{N_T} = T$, Δt_k and ΔW_k are short notations for the time and Brownian motion increments respectively, and we denote by $\|\pi\|$ the mesh of this partition:

$$\Delta t_k = t_{k+1} - t_k, \quad \Delta W_k = W_{t_{k+1}} - W_{t_k}, \quad \|\pi\| = \max_{0 \le k \le N_T - 1} \Delta t_k. \tag{19}$$

We also define a step function $\pi(t)$, and a set $\mathcal{T}$ for later use:

$$\pi(t) = t_k \text{ for } t \in [t_k, t_{k+1}), \quad \mathcal{T} := \{t_0, t_1, \ldots, t_{N_T - 1}\}. \tag{20}$$

The deep BSDE method solves the minimization problem:

$$\inf_{\psi_0 \in \mathcal{N}_0',\ \{\phi_k \in \mathcal{N}_k\}_{k=0}^{N_T-1}} \mathbb{E}|g(X_T^\pi) - Y_T^\pi|^2, \tag{21}$$

$$s.t. \quad X_{t_{k+1}}^\pi = X_{t_k}^\pi + \mu(t_k, X_{t_k}^\pi)\Delta t_k + \Sigma(t_k, X_{t_k}^\pi)\Delta W_k, \quad X_0^\pi = x_0, \tag{22}$$

$$Y_{t_{k+1}}^\pi = Y_{t_k}^\pi - F(t_k, X_{t_k}^\pi, Z_{t_k}^\pi)\Delta t_k + (Z_{t_k}^\pi)^\intercal \Delta W_k, \quad Y_0^\pi = \psi_0(X_0^\pi), \tag{23}$$

$$Z_{t_k}^\pi = \phi_k(X_{t_k}^\pi)$$

where $\mathcal{N}_0'$ and $\{\mathcal{N}_k\}_{k=0}^{N_T-1}$ are hypothesis spaces related to deep neural networks, and for brevity, we use the notation X_0^π for $X_{t_0}^\pi$, X_T^π for $X_{t_{N_T}}^\pi$, and the same applies to the process Y^π, Z^π. The goal is to find optimal deterministic maps $\psi_0^*, \{\phi_k^*\}_{k=0}^{N_T-1}$ such that the *loss function* in (21) is minimized. Intuitively, the smaller (21), the better the approximation to the original problem (18). In practice, the expected

value is replaced by the loss of a very deep neural network, which is formed by stacking all the subnetworks $\psi_0, \{\phi_k\}_{k=0}^{N_T-1}$ in sequence according to (23). The loss is computed by generating sample paths of $\{W_{t_k}\}_{k=0}^{N_T}$ and producing (22)–(23). At each stage, there are N losses corresponding to N sub-problems solved by the deep BSDE method.

Now we recall the existing convergence results for the deep BSDE method [33, Theorems 1 and 2] and state them together in the following theorem.

Theorem 1. *For the generic BSDE (18), we assume:*

1. *The functions μ, Σ, g and F satisfy the following Lipschitz condition, for some constant $L > 0$:*
$$|\mu(t,x_1)-\mu(t,x_2)|^2+\|\Sigma(t,x_1)-\Sigma(t,x_2)\|_F^2+|F(t,x_1,p_1)-F(t,x_2,p_2)|^2 + |g(x_1)-g(x_2)|^2 \le L\left[|x_1-x_2|^2+|p_1-p_2|^2\right];$$
2. *The functions μ, Σ and h are all 1/2-Hölder continuous with respect to t. For simplicity, we use K for this Hölder constant;*
3. *We also use K to denote the upper bound of $|\mu(0,0)|^2$, $\|\Sigma(0,0)\|_F^2$, $|F(0,0,0)|^2$ and $|g(0)|^2$.*

Then, we have the following two estimates:

$$\sup_{t\in[0,T]}\mathbb{E}|Y_t-Y^\pi_{\pi(t)}|^2+\int_0^T\mathbb{E}\|Z_t-Z^\pi_{\pi(t)}\|_F^2\,\mathrm{d}t\le C\left[\|\pi\|+\mathbb{E}|g(X_T^\pi)-Y_T^\pi|^2\right], \quad (24)$$

and

$$\inf_{\psi_0\in\mathcal{N}_0',\ \{\phi_k\in\mathcal{N}_k\}_{k=0}^{N_T-1}}\mathbb{E}|g(X_T^\pi)-Y_T^\pi|^2 \le C\Big[\|\pi\|+\inf_{\psi_0\in\mathcal{N}_0',\ \{\phi_k\in\mathcal{N}_k\}_{k=0}^{N_T-1}}\{\mathbb{E}|Y_0-\psi_0(x_0)|^2+\sum_{k=0}^{N_T-1}\mathbb{E}\|\hat{Z}_{t_k}-\phi_k(X^\pi_{t_k})\|_F^2\Delta t_k\}\Big], \quad (25)$$

where $\mathcal{N}_0'$ and $\{\mathcal{N}_k\}_{k=0}^{N_T-1}$ are the hypothesis spaces for neural network architectures to approximate Y_0^π and $Z^\pi_{t_k}$, $\|\pi\|$ and $\pi(t)$ are given in (19)–(20), $\hat{Z}_{t_k}=(\Delta t_k)^{-1}\mathbb{E}[\int_{t_k}^{t_{k+1}}Z_t\,\mathrm{d}t|X^\pi_{t_k}]$, and $C>0$ is a constant only depending on L, T, K and $\mathbb{E}|\boldsymbol{x}_0|^2$.

Remark 3. The first inequality (24) shows that the distance between the true solution of BSDE (18) and the output of the deep BSDE method can be controlled by its loss function. In other words, in practice, the accuracy of the numerical solution is effectively indicated by the value of the loss function. The second inequality (25) states that a small loss function of the deep BSDE method is attainable if the hypothesis spaces ($\mathcal{N}_0'$ and $\{\mathcal{N}_k\}_{k=0}^{N_T-1}$) can approximate specific functions well such that the right-hand side of (25) is small. Neural networks are such hypothesis spaces, ensured by the universal approximation results. For instance, Theorem 2.1 in [3] states that every continuous and piecewise linear function with m-dimensional input can be represented by a deep neural network with ReLu activation function and at most $\lceil 1+\log_2(m+1)\rceil$ layers, which justifies the choice of ψ_0 and ϕ_k being neural networks. For further discussion, we refer to the discussion on page 22 of [33].

Beyond Theorems 1 and 2 in [33], there are still some theoretical issues remaining unresolved regarding the deep BSDE method, which are common in almost all

the algorithms involving deep neural networks: First, it is unclear yet that what types of hypothesis spaces can approximate the specific functions in the deep BSDE method without the curse of dimensionality (*i.e.*, the number of parameters of neural networks grows at most polynomially both in dimension and the reciprocal of the approximation error). Second, even with suitable function spaces, it is hard to guarantee the optimization algorithm can find approximately the minimizer of the highly nonconvex loss function. We refer the interested readers to [20, 32, 33] for more detailed descriptions and theoretical justifications of the deep BSDE method. Details on the implementation in this paper are presented in Section 5.

4. Convergence analysis.

This section will provide the theoretical foundation for the deep fictitious play algorithm. Section 4.1 focuses on the decoupling step. Theorem 2 proves the convergence to the true Nash equilibrium, if the decoupled sub-problems are solved exactly and repeatedly. Section 4.2 focuses on the numerical error on the deep BSDE algorithm for solving each sub-problem. Theorem 3 presents a game version of Theorem 1. Section 4.3 combines the previous results, identifies the ϵ-Nash equilibrium produced by deep fictitious play, and analyzes its numerical performance on the original game.

4.1. Convergence analysis of the decoupling step.

In this section, we will focus on the convergence of the decoupling step, *i.e.*, how the systems defined by PDEs (12) (fictitious play) or (13) (policy update) converge to the system defined by PDEs (4), or equivalently, how the corresponding BSDE systems (15) (fictitious play) or (17) (policy update) converge to the BSDE system (9). Throughout this paper, we shall use the following assumptions.

Assumption 2. *We shall use $|\cdot|$, $\|\cdot\|_F$ and $\|\cdot\|_S$ to denote the Euclidean norm, Frobenius norm and spectral norm, respectively.*

(1) The functions $\phi(t,\boldsymbol{x},\boldsymbol{\alpha}):[0,T]\times\mathbb{R}^n\times\mathcal{A}\to\mathbb{R}^k$, $\Sigma(t,\boldsymbol{x}):[0,T]\times\mathbb{R}^n\to\mathbb{R}^{n\times k}$, $\boldsymbol{f}(t,\boldsymbol{x},\boldsymbol{\alpha})=(f^1,f^2,\dots,f^N)^\intercal(t,\boldsymbol{x},\boldsymbol{\alpha}):[0,T]\times\mathbb{R}^n\times\mathcal{A}\to\mathbb{R}^N$ and $\boldsymbol{g}(\boldsymbol{x})=[g^1,g^2,\dots,g^N]^\intercal(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}^N$ are Lipschitz with respect to $\boldsymbol{x}$ and $\boldsymbol{\alpha}$, with a positive constant L:

$$\begin{aligned}|\phi(t,\boldsymbol{x}_1,\boldsymbol{\alpha}_1)-\phi(t,\boldsymbol{x}_2,\boldsymbol{\alpha}_2)|^2&\le L[|\boldsymbol{x}_1-\boldsymbol{x}_2|^2+|\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2|^2],\\ \|\Sigma(t,\boldsymbol{x}_1)-\Sigma(t,\boldsymbol{x}_2)\|_F^2&\le L|\boldsymbol{x}_1-\boldsymbol{x}_2|^2,\\ |\boldsymbol{f}(t,\boldsymbol{x}_1,\boldsymbol{\alpha}_1)-\boldsymbol{f}(t,\boldsymbol{x}_2,\boldsymbol{\alpha}_2)|^2&\le L[|\boldsymbol{x}_1-\boldsymbol{x}_2|^2+|\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2|^2],\\ |\boldsymbol{g}(\boldsymbol{x}_1)-\boldsymbol{g}(\boldsymbol{x}_2)|^2&\le L|\boldsymbol{x}_1-\boldsymbol{x}_2|^2.\end{aligned}$$

(2) The function $\boldsymbol{a}(t,\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{p})$ given in (7) is well-defined, and is Lipschitz with respect to $\boldsymbol{x}$, $\boldsymbol{\alpha}$ and $\boldsymbol{p}$:

$$\begin{aligned}&|\boldsymbol{a}(t,\boldsymbol{x}_1,\boldsymbol{\alpha}_1,\boldsymbol{p}_1)-\boldsymbol{a}(t,\boldsymbol{x}_2,\boldsymbol{\alpha}_2,\boldsymbol{p}_2)|^2\\ &\qquad\le L(1-a_\alpha)[|\boldsymbol{x}_1-\boldsymbol{x}_2|^2+\|\boldsymbol{p}_1-\boldsymbol{p}_2\|_F^2]+a_\alpha|\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2|^2,\end{aligned}\tag{26}$$

with $0<a_\alpha<1$. Notice that this also implies that $\boldsymbol{\alpha}(t,\boldsymbol{x},\boldsymbol{p})$ defined by (8) exists and is unique, which is Lipschitz with respect to $\boldsymbol{x}$ and $\boldsymbol{p}$:

$$|\boldsymbol{\alpha}(t,\boldsymbol{x}_1,\boldsymbol{p}_1)-\boldsymbol{\alpha}(t,\boldsymbol{x}_2,\boldsymbol{p}_2)|^2\le L[|\boldsymbol{x}_1-\boldsymbol{x}_2|^2+\|\boldsymbol{p}_1-\boldsymbol{p}_2\|_F^2].$$

(3) The functions ϕ and Σ are uniformly bounded:

$$\|\Sigma(t,\boldsymbol{x})\|_S^2\le M,\quad \max_{1\le i\le k}|\phi^i(t,\boldsymbol{x},\boldsymbol{\alpha})|^2\le M.$$

Here ϕ^i denotes the i-th component of ϕ, and M is a positive constant.

(4) *The functions ϕ, Σ, $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{a}$ are all 1/2-Hölder continuous with respect to t. We shall use K as the upper bound of all the Hölder constants.*

(5) *The constant K is also the upper bound of constants $|\boldsymbol{a}(0,0,0,0)|^2$, $|\boldsymbol{f}(0,0,0)|^2$, $|\boldsymbol{g}(0)|^2$, $|\phi(0,0,0)|^2$ and $\|\Sigma(0,0)\|_F^2$.*

Assumption 3. *There exists an adapted solution of the BSDE system (9) such that*

$$\mathbb{E}\Big[\sup_{0\le t\le T}(|\boldsymbol{X}_t|^2+|\boldsymbol{Y}_t|^2)+\int_0^T\|\boldsymbol{Z}_t\|_F^2\,\mathrm{d}t\Big]<+\infty. \tag{27}$$

Moreover, we assume that $\|\boldsymbol{Z}_t\|_S^2\le M'$, $\quad\mathbb{L}\times\mathbb{P}$-a.s..

We remark that Assumption 2 is quite standard in the analysis of stochastic differential games and Assumption 3 can be satisfied in several cases. For instance, Assumption 3 holds true under Assumptions 1 and 2 with small time duration. We provide a detailed proof of this point (Proposition 6) in the appendix. The small time duration assumption is commonly seen in games, for instance in solving mean-field games [15] and the convergence of numerical schemes for mean-field games [4]. Also, through the nonlinear Feynman-Kac formula and the boundedness of Σ in Assumption 2, Assumption 3 is satisfied if the solution to the HJB system (4) is uniformly Lipschitz with respect to $\boldsymbol{x}$. Specifically, with additional assumptions:

$$\boldsymbol{f},\boldsymbol{g}\text{ are bounded, and }\Sigma\Sigma^\intercal\text{ is uniformly nondegenerate,} \tag{28}$$

V^i is continuous and differentiable with continuous bounded gradients on $[0,T]\times\mathbb{R}^n$ (cf. [15, Prop. 2.13]). Therefore, using small time duration result (Proposition 6) on $[T-\delta,T]$ for small δ, one has the uniformly Lipschitz on $[0,T]$ and Assumption 3 is fulfilled under (28). We also point out that Assumption 3 implies that the BSDE system (9) has a unique adapted L^2-integrable solution, see Proposition 7 in Appendix A.

Recalling that m is the stage index in the decoupling step, now we present the main result in this section regarding its convergence.

Theorem 2. *Under Assumptions 1, 2 and 3, for any $\epsilon\in(0,1-a_\alpha)$, there exists a constant $C(\epsilon)>0$ which only depends on T, L, M, M' and ϵ such that*

$$\sup_{0\le t\le T}\mathbb{E}|\boldsymbol{Y}_t^m-\boldsymbol{Y}_t|^2+\int_0^T\mathbb{E}\|\boldsymbol{Z}_t^m-\boldsymbol{Z}_t\|_F^2\,\mathrm{d}t+\int_0^T\mathbb{E}|\boldsymbol{\alpha}_t^m-\boldsymbol{\alpha}_t^*|^2\,\mathrm{d}t \le C(\epsilon)(a_\alpha+\epsilon)^m\int_0^T\mathbb{E}|\boldsymbol{\alpha}_t^0-\boldsymbol{\alpha}_t^*|^2\,\mathrm{d}t,$$

where $(\boldsymbol{Y}_t^m,\boldsymbol{Z}_t^m)$ is defined by $\boldsymbol{Y}_t^m=[Y_t^{1,m},\dots,Y_t^{N,m}]^\intercal$, $\quad\boldsymbol{Z}_t^m=[Z_t^{1,m},\dots,Z_t^{N,m}]$, with $(Y_t^{i,m},Z_t^{i,m})$ from the BSDE systems (15) or (17), $(\boldsymbol{Y}_t,\boldsymbol{Z}_t)$ is defined in (9), $\boldsymbol{\alpha}_t^m=\boldsymbol{\alpha}^m(t,\boldsymbol{X}_t)$ and $\boldsymbol{\alpha}_t^*=\boldsymbol{\alpha}^*(t,\boldsymbol{X}_t)$.

Proof. Theorem 2 states the convergence of both fictitious play (according to (15)) and policy update (according to (17)). The proofs of these two are very similar, and we shall focus on the fictitious play method for brevity.

To perform convergence analysis, we first rewrite the BSDE systems to show the explicit dependence on the players' strategies. For (9), it reads as

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \Sigma(s, \boldsymbol{X}_s)\,\mathrm{d}\boldsymbol{W}_s, \\ \boldsymbol{Y}_t = \boldsymbol{g}(\boldsymbol{X}_T) + \int_t^T \boldsymbol{H}(s, \boldsymbol{X}_s, \boldsymbol{\alpha}_s^*, \boldsymbol{Z}_s)\,\mathrm{d}s - \int_t^T (\boldsymbol{Z}_s)^\intercal\,\mathrm{d}\boldsymbol{W}_s, \\ \boldsymbol{\alpha}_t^* = \boldsymbol{a}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*, \boldsymbol{Z}_t), \end{cases} \tag{29}$$

where $\boldsymbol{H}, \boldsymbol{a}$ are defined in (5) and (7). The rewritten system of (15) is

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \Sigma(s, \boldsymbol{X}_s)\,\mathrm{d}\boldsymbol{W}_s, \\ \boldsymbol{Y}_t^{m+1} = \boldsymbol{g}(\boldsymbol{X}_T) + \int_t^T \hat{\boldsymbol{H}}(s, \boldsymbol{X}_s, \boldsymbol{\alpha}_s^m, \boldsymbol{\alpha}_s^{m+1}, \boldsymbol{Z}_s^{m+1})\,\mathrm{d}s - \int_t^T (\boldsymbol{Z}_s^{m+1})^\intercal\,\mathrm{d}\boldsymbol{W}_s, \\ \boldsymbol{\alpha}_t^{m+1} = \boldsymbol{a}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m, \boldsymbol{Z}_t^{m+1}), \end{cases} \tag{30}$$

where $\hat{\boldsymbol{H}} = [\hat{H}^1, \dots, \hat{H}^N]^\intercal$, $\hat{H}^i(t, \boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{p}) \equiv H^i(t, \boldsymbol{x}, (\gamma^i, \boldsymbol{\xi}^{-i}), p^i)$ and p^i stands for the i^{th} column of $\boldsymbol{p}$. Note that this is slightly an abuse of notation with (16), to show the driver's explicit dependence on $\boldsymbol{\alpha}^{m+1}$. Also note that the rewritten system (30) is simply a condensed form of (15), concatenating all $Y_t^{i,m}$ into $\boldsymbol{Y}_t^m$, without changing its decoupled nature. This will also ease the notation in the following proof.

We now define $\delta\boldsymbol{H}_t^m = \hat{\boldsymbol{H}}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m, \boldsymbol{\alpha}_t^{m+1}, \boldsymbol{Z}_t^{m+1}) - \boldsymbol{H}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*, \boldsymbol{Z}_t)$. Noticing that

$$\begin{aligned} \delta H_t^{i,m} =& \phi(t, \boldsymbol{X}_t, (\alpha_t^{i,m+1}, \boldsymbol{\alpha}_t^{-i,m})) \cdot Z_t^{i,m+1} + f^i(t, \boldsymbol{X}_t, (\alpha_t^{i,m+1}, \boldsymbol{\alpha}_t^{-i,m})) \\ & - \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*) \cdot Z_t^i - f^i(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*) \\ =& \phi(t, \boldsymbol{X}_t, (\alpha_t^{i,m+1}, \boldsymbol{\alpha}_t^{-i,m})) \cdot (Z_t^{i,m+1} - Z_t^i) \\ & + [\phi(t, \boldsymbol{X}_t, (\alpha_t^{i,m+1}, \boldsymbol{\alpha}_t^{-i,m})) - \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m)] \cdot Z_t^i \\ & + [\phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m) - \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*)] \cdot Z_t^i \\ & + [f^i(t, \boldsymbol{X}_t, (\alpha_t^{i,m+1}, \boldsymbol{\alpha}_t^{-i,m})) - f^i(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m)] \\ & + [f^i(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^m) - f^i(t, \boldsymbol{X}_t, \boldsymbol{\alpha}_t^*)]. \end{aligned}$$

Therefore, with Assumptions 2 and 3,

$$\begin{aligned} |\delta\boldsymbol{H}_t^m|^2 \le& C_1 \{ \|\boldsymbol{Z}_t^{m+1} - \boldsymbol{Z}_t\|_F^2 \\ & + \sum_{i=1}^N (|Z_t^i|^2 + 1) |\alpha_t^{i,m+1} - \alpha_t^{i,m}|^2 + \|Z_t\|_S^2 |\boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_t^*|^2 + |\boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_t^*|^2 \} \\ \le& C_2 \left\{ \|\boldsymbol{Z}_t^{m+1} - \boldsymbol{Z}_t\|_F^2 + |\boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_t^*|^2 + |\boldsymbol{\alpha}_t^{m+1} - \boldsymbol{\alpha}_t^*|^2 \right\}, \end{aligned}$$

where C_1, C_2 are two positive constants only depending on L, M and M'.

Next, we define $\delta\boldsymbol{Y}_t^m = \boldsymbol{Y}_t^m - \boldsymbol{Y}_t$, $\delta\boldsymbol{Z}_t^m = \boldsymbol{Z}_t^m - \boldsymbol{Z}_t$, $\delta\boldsymbol{\alpha}_t^m = \boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_t^*$. With (29) and (30), we have

$$\mathrm{d}\delta\boldsymbol{Y}_t^{m+1} = -\delta\boldsymbol{H}_t^m\,\mathrm{d}t + (\delta\boldsymbol{Z}_t^{m+1})^\intercal\,\mathrm{d}\boldsymbol{W}_t.$$

For any $\beta > 0$, by Ito's formula, taking expectation on both sides and integrating from t to T gives

$$\begin{aligned}
&e^{\beta t}\mathbb{E}|\delta \boldsymbol{Y}_t^{m+1}|^2 + \int_t^T e^{\beta s}\mathbb{E}\|\delta \boldsymbol{Z}_s^{m+1}\|_F^2 \,\mathrm{d}s \\
=&\int_t^T e^{\beta s}\mathbb{E}[2\delta \boldsymbol{H}_s^m \cdot \delta \boldsymbol{Y}_s^{m+1} - \beta|\delta \boldsymbol{Y}_s^{m+1}|^2] \,\mathrm{d}s \\
\leq&\frac{1}{\beta}\int_t^T e^{\beta s}\mathbb{E}|\delta \boldsymbol{H}_s^m|^2 \,\mathrm{d}s,
\end{aligned}$$

where the inequality holds because $2\delta \boldsymbol{H}_s^m \cdot \delta \boldsymbol{Y}_s^{m+1} \leq \beta^{-1}|\delta \boldsymbol{H}_s^m|^2 + \beta|\delta \boldsymbol{Y}_s^{m+1}|^2$. Then, taking the supremum with respect to t, we deduce

$$\begin{aligned}
&\sup_{0\leq t\leq T}\mathbb{E}e^{\beta t}|\delta \boldsymbol{Y}_t^{m+1}|^2 + \int_0^T e^{\beta t}\mathbb{E}\|\delta \boldsymbol{Z}_t^{m+1}\|_F^2 \,\mathrm{d}t \\
\leq&\frac{1}{\beta}\int_0^T e^{\beta t}\mathbb{E}|\delta \boldsymbol{H}_t^m|^2 \,\mathrm{d}t \\
\leq&\frac{C_2}{\beta}\int_0^T e^{\beta t}\mathbb{E}[\|\delta \boldsymbol{Z}_t^{m+1}\|_F^2 + |\delta \boldsymbol{\alpha}_t^m|^2 + |\delta \boldsymbol{\alpha}_t^{m+1}|^2] \,\mathrm{d}t.
\end{aligned} \tag{31}$$

Choosing $\beta = C_2$, we can obtain

$$\sup_{0\leq t\leq T}\mathbb{E}|\delta \boldsymbol{Y}_t^{m+1}|^2 \leq e^{C_2 T}\int_0^T [\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 + \mathbb{E}|\delta \boldsymbol{\alpha}_t^{m+1}|^2] \,\mathrm{d}t. \tag{32}$$

For $\beta > C_2$, using inequality (31) again, we have

$$\int_0^T e^{\beta t}\mathbb{E}\|\delta \boldsymbol{Z}_t^{m+1}\|_F^2 \,\mathrm{d}t \leq \frac{C_2}{\beta - C_2}\int_0^T e^{\beta t}[\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 + \mathbb{E}|\delta \boldsymbol{\alpha}_t^{m+1}|^2] \,\mathrm{d}t. \tag{33}$$

The Lipschitz condition of the function $\boldsymbol{a}$ (26) (with constants L and a_α) and estimate (33) give that

$$\begin{aligned}
&\int_0^T e^{\beta t}\mathbb{E}|\delta \boldsymbol{\alpha}_t^{m+1}|^2 \,\mathrm{d}t \\
\leq&\int_0^T e^{\beta t}[L(1-a_\alpha)\mathbb{E}\|\delta \boldsymbol{Z}_t^{m+1}\|_F^2 + a_\alpha\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2] \,\mathrm{d}t \\
\leq&a_\alpha\int_0^T e^{\beta t}\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 \,\mathrm{d}t + \frac{LC_2}{\beta - C_2}\int_0^T e^{\beta t}[\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 + \mathbb{E}|\delta \boldsymbol{\alpha}_t^{m+1}|^2] \,\mathrm{d}t,
\end{aligned}$$

which is equivalent to (further assuming $\beta > (L+1)C_2$)

$$\int_0^T e^{\beta t}\mathbb{E}|\delta \boldsymbol{\alpha}_t^{m+1}|^2 \,\mathrm{d}t \leq \frac{\beta - C_2}{\beta - (L+1)C_2}\left(a_\alpha + \frac{LC_2}{\beta - C_2}\right)\int_0^T e^{\beta t}\mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 \,\mathrm{d}t.$$

For a given $\epsilon \in (0, 1 - a_\alpha)$, we can choose β large enough such that

$$\frac{\beta - C_2}{\beta - (L+1)C_2}\left(a_\alpha + \frac{LC_2}{\beta - C_2}\right) \leq a_\alpha + \epsilon < 1.$$

Then, there exists a constant $C(\epsilon) > 0$ that only depends on T, L, M, M' and ϵ such that

$$\int_0^T \mathbb{E}|\delta \boldsymbol{\alpha}_t^m|^2 \,\mathrm{d}t \leq C(\epsilon)(a_\alpha + \epsilon)^m \int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^0 - \boldsymbol{\alpha}_t^*|^2 \,\mathrm{d}t.$$

Combining the last inequality with inequalities (32) and (33), we obtain our result. □

Remark 4. The convergence in Theorem 2 holds for games with any size of N instead of the focus in the numerical algorithm that is between 5 and 100, and is independent of any numerical scheme.

4.2. **Numerical error analysis.** This section is dedicated to analyzing the numerical error introduced by the deep BSDE method when solving each sub-problem. Specifically, we aim to control the distance between $(\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{Z}_t)$ defined in (9) and the discrete system $(\boldsymbol{X}^{\pi}_{t_k}, \boldsymbol{Y}^{\pi,m}_{t_k}, \boldsymbol{Z}^{\pi,m}_{t_k})$ satisfying:

$$\begin{cases} \boldsymbol{X}^{\pi}_{t_{k+1}} = \boldsymbol{X}^{\pi}_{t_k} + \Sigma(t_k, \boldsymbol{X}^{\pi}_{t_k})\Delta W_k, \quad \boldsymbol{X}^{\pi}_0 = \boldsymbol{x}_0, \\ \boldsymbol{Y}^{\pi,m+1}_{t_{k+1}} = \boldsymbol{Y}^{\pi,m+1}_{t_k} - \boldsymbol{h}^m(t_k, \boldsymbol{X}^{\pi}_{t_k}, \boldsymbol{Z}^{\pi,m+1}_{t_k})\Delta t_k + (\boldsymbol{Z}^{\pi,m+1}_{t_k})^{\intercal}\Delta \boldsymbol{W}_k, \end{cases} \tag{34}$$

where $\boldsymbol{h}^m$ is either $[h^{1,m}, \cdots, h^{N,m}]^{\intercal}$ with $h^{i,m}(t, \boldsymbol{x}, \boldsymbol{p}) = \inf_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, (\alpha^i, \boldsymbol{\alpha}^{-i,\pi,m}(t, \boldsymbol{x})), \boldsymbol{p})$ when decoupled through fictitious play, or $\boldsymbol{h}^m(t, \boldsymbol{x}, \boldsymbol{p}) = \boldsymbol{H}(t, \boldsymbol{x}, \boldsymbol{\alpha}^{\pi,m}(t, \boldsymbol{x}), \boldsymbol{p})$ when decoupled through policy update. As stated in Section 3.2, $\boldsymbol{Y}^{\pi,m+1}_0$ and $\boldsymbol{Z}^{\pi,m+1}_{t_k}$ are paramterized by neural networks,

$$\boldsymbol{Y}^{\pi,m+1}_0 = \boldsymbol{\psi}^{m+1}_0(\boldsymbol{X}^{\pi}_0), \quad \boldsymbol{Z}^{\pi,m+1}_{t_k} = \boldsymbol{\phi}^{m+1}_k(\boldsymbol{X}^{\pi}_{t_k}),$$

where $\boldsymbol{\psi}^m_0$ and $\{\boldsymbol{\phi}^m_k\}_{k=0}^{N_T-1}$ are the optimal deterministic maps determined at stage m that belongs to the hypothesis spaces (*cf.* Section 3.2). Then, the $(m+1)^{th}$-stage policies defined on $\mathcal{T} \times \mathbb{R}^n$ are updated by:

$$\boldsymbol{\alpha}^{\pi,m+1}(t, \boldsymbol{x}) = \boldsymbol{a}(t, \boldsymbol{x}, \boldsymbol{\alpha}^{\pi,m}(t, \boldsymbol{x}), \boldsymbol{\phi}^{m+1}(t, \boldsymbol{x})), \quad \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathbb{R}^n \tag{35}$$

where $\boldsymbol{\phi}^m(t_k, \boldsymbol{x}) = \boldsymbol{\phi}^m_k(\boldsymbol{x})$. Note that the above notation is simply a vector form of the deep BSDE method applied to system (15) or (17). It does not change the decoupling nature of the deep fictitious play algorithm, i.e., each entry $(Y^{i,\pi,m}, Z^{i,\pi,m})$ in $(\boldsymbol{Y}^{\pi,m}, \boldsymbol{Z}^{\pi,m})$ still solves its own problem.

Initially, we hope to apply Theorem 1 to the BSDE system (15) and (17). By the game feature and the decoupling scheme, stage $m+1$'s estimates rely on the regularity of stage m's policy $\boldsymbol{\alpha}^m(t, \boldsymbol{x})$ (see definition in (14)). Specifically, it requires the following condition, in addition to Assumption 2:

$$|\boldsymbol{\alpha}^m(t_1, \boldsymbol{x}_1) - \boldsymbol{\alpha}^m(t_2, \boldsymbol{x}_2)|^2 \leq L[|t_1 - t_2| + |\boldsymbol{x}_1 - \boldsymbol{x}_2|^2].$$

However, in general, this property is not inherited from stage to stage. To circumvent this issue, we introduce a projection operator, which needs to be applied at the end of each stage. Along this line, we need the following assumption.

Assumption 4. *The optimal policy $\boldsymbol{\alpha}^*$ as a function on $[0, T] \times \mathbb{R}^n$ is Lipschitz with respect to $\boldsymbol{x}$ and 1/2-Hölder continuous with respect to t: $|\boldsymbol{\alpha}^*(t_1, \boldsymbol{x}_1) - \boldsymbol{\alpha}^*(t_2, \boldsymbol{x}_2)|^2 \leq L(|t_1 - t_2| + |\boldsymbol{x}_1 - \boldsymbol{x}_2|^2)$. We also assume that $|\boldsymbol{\alpha}^*(t, \boldsymbol{x})|^2 \leq L(1 + |\boldsymbol{x}|^2)$ for any $(t, \boldsymbol{x}) \in [0, T] \times \mathbb{R}^n$.*

The regularity of the Nash equilibrium $\boldsymbol{\alpha}^*$ with respect to $(t, \boldsymbol{x})$ itself is an interesting question that is worth a separate study; for instance, see [37, 38] in different settings. We leave it for future work on checking under what conditions Assumption 4 holds. For partial justification of Assumption 4, we remark that [51, Proposition 3.3] implies the Lipschitz continuity with respect to $\boldsymbol{x}$ but not t.

Recall the set $\mathcal{T}$ containing all endpoints of the size N_T partition π on $[0, T]$ from (20), for any $\eta \geq 0$ we define a Hilbert space on $\mathcal{T} \times \mathbb{R}^n$:

$$\mathcal{H}_\eta^\pi = \Big\{\boldsymbol{\alpha}:\ \text{measurable functions from } \mathcal{T}\times\mathbb{R}^n \text{ to } \mathbb{R}^{Nd_{\boldsymbol{\alpha}}},\ \sum_{k=0}^{N_T-1} e^{\eta t_k}\mathbb{E}|\boldsymbol{\alpha}(t_k,\boldsymbol{X}_{t_k}^\pi)|^2\Delta t_k < +\infty\Big\}, \quad (36)$$

with norm $\|\boldsymbol{\alpha}\|_{\mathcal{H}_\eta^\pi}^2 := \sum_{k=0}^{N_T-1} e^{\eta t_k}\mathbb{E}|\boldsymbol{\alpha}(t_k,\boldsymbol{X}_{t_k}^\pi)|^2\Delta t_k$, and a subset

$$\mathcal{N}^\pi = \Big\{\boldsymbol{\alpha}:\mathcal{T}\times\mathbb{R}^n \hookrightarrow \mathbb{R}^{Nd_{\boldsymbol{\alpha}}}, |\boldsymbol{\alpha}(t_1,\boldsymbol{x}_1)-\boldsymbol{\alpha}(t_2,\boldsymbol{x}_2)|^2 \le L'[|t_1-t_2|+|\boldsymbol{x}_1-\boldsymbol{x}_2|^2],\ |\boldsymbol{\alpha}(t,\boldsymbol{x})|^2 \le L'(1+|\boldsymbol{x}|^2)\Big\}^{3}$$

with a constant $L' \ge L$. Because $\mathcal{N}^\pi$ is a closed convex subset of $\mathcal{H}_\eta^\pi$, the projection $\mathrm{P}_{\mathcal{N}^\pi,\eta}$ from $\mathcal{H}_\eta^\pi$ to $\mathcal{N}^\pi$ exists and does not increase distance (*cf.* [9, Chapter 5, Proposition 5.3]):

$$\|\mathrm{P}_{\mathcal{N}^\pi,\eta}(f_1)-\mathrm{P}_{\mathcal{N}^\pi,\eta}(f_2)\|_{\mathcal{H}_\eta^\pi} \le \|f_1-f_2\|_{\mathcal{H}_\eta^\pi} \quad \forall f_1,f_2\in\mathcal{H}_\eta^\pi. \quad (37)$$

Therefore, we are able to apply the projection operator $\mathrm{P}_{\mathcal{N}^\pi,\eta}$ at the end of each stage, *i.e.*, we change equation (35) to

$$\tilde{\boldsymbol{\alpha}}^{\pi,m+1}(t,\boldsymbol{x}) = \boldsymbol{a}(t,\boldsymbol{x},\boldsymbol{\alpha}^{\pi,m}(t,\boldsymbol{x}),\boldsymbol{\phi}^{m+1}(t,\boldsymbol{x})), \quad (38)$$

$$\boldsymbol{\alpha}^{\pi,m+1} = \mathrm{P}_{\mathcal{N}^\pi,\eta}(\tilde{\boldsymbol{\alpha}}^{\pi,m+1}). \quad (39)$$

By this definition, the numerical solution $\boldsymbol{\alpha}^{\pi,m+1}(t,\boldsymbol{x})$ in fact implicitly depends on the value of η. We suppress this dependence for brevity of notation. The main theorem in this section is as follows.

Theorem 3. *Under Assumptions 1–4, let $\boldsymbol{\alpha}^{\pi,0}:\mathcal{T}\times\mathbb{R}^n\mapsto\mathcal{A}$ be a measurable function satisfying:*

$$|\boldsymbol{\alpha}^{\pi,0}(t_1,\boldsymbol{x}_1)-\boldsymbol{\alpha}^{\pi,0}(t_2,\boldsymbol{x}_2)|^2 \le L'[|t_1-t_2|+|\boldsymbol{x}_1-\boldsymbol{x}_2|^2], \quad |\boldsymbol{\alpha}^{\pi,0}(t,\boldsymbol{x})|^2 \le L'(1+|\boldsymbol{x}|^2). \quad (40)$$

Then, for any $\epsilon\in(0,1-a_\alpha)$, assuming that $\eta>\eta_\epsilon$ in (36), where η_ϵ is a constant depending on T, L, M, M' and ϵ, we have

$$\begin{aligned}\sup_{t\in[0,T]}\mathbb{E}|\boldsymbol{Y}_t-\boldsymbol{Y}_{\pi(t)}^{\pi,m}|^2 + \int_0^T \mathbb{E}\|\boldsymbol{Z}_t-\boldsymbol{Z}_{\pi(t)}^{\pi,m}\|_F^2\,\mathrm{d}t + \int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^*-\boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2\,\mathrm{d}t \\ \le C(\eta,\epsilon)\Big[\|\pi\|+(a_\alpha+\epsilon)^m\int_0^T\mathbb{E}\left|\boldsymbol{\alpha}_t^*-\boldsymbol{\alpha}_{\pi(t)}^{\pi,0}\right|^2\mathrm{d}t \\ +\sum_{j=1}^m(a_\alpha+\epsilon)^{m-j}\mathbb{E}\left|\boldsymbol{g}(\boldsymbol{X}_T^\pi)-\boldsymbol{Y}_T^{\pi,j}\right|^2\Big],\end{aligned} \quad (41)$$

where $(\boldsymbol{X}_{t_k}^\pi,\boldsymbol{Y}_{t_k}^{\pi,m},\boldsymbol{Z}_{t_k}^{\pi,m})$ is defined in (34), $\boldsymbol{\alpha}_{\pi(t)}^{\pi,m}\equiv\boldsymbol{\alpha}_{t_k}^{\pi,m}=\boldsymbol{\alpha}^{\pi,m}(t_k,\boldsymbol{X}_{t_k}^\pi)$ for $t\in[t_k,t_{k+1})$, and $C(\eta,\epsilon)>0$ is a constant depending only on T, L, M, M', K, L', $\mathbb{E}|\boldsymbol{x}_0|^2$, η and ϵ. Here $(\boldsymbol{X}_{t_k}^\pi,\boldsymbol{Y}_{t_k}^{\pi,m},\boldsymbol{Z}_{t_k}^{\pi,m})$ represents either the discrete BSDE system using fictitious play or policy update in the decoupling step, depending on the definition of $\boldsymbol{h}$ in (34).

Next, with a slight abuse of notation (see Remark 5 (2) for details), we define $(\boldsymbol{Y}_t^m,\boldsymbol{Z}_t^m)$ as

$$\boldsymbol{Y}_t^m = [Y_t^{1,m},\dots,Y_t^{N,m}]^\intercal, \quad \boldsymbol{Z}_t^m = [Z_t^{1.m},\dots,Z_t^{N,m}]$$

with $(Y_t^{i,m}, Z_t^{i,m})$ *from the BSDE systems* (15) *in the setting of fictitious play or* (17) *in the setting of policy update, in which the previous stage policy is given by the extension of the numerical approximation in time*

$$\boldsymbol{\alpha}^m(t,\boldsymbol{x}) = \inf_{t'\in\mathcal{T}}[\boldsymbol{\alpha}^{\pi,m}(t',\boldsymbol{x}) + L'|t'-t|^{\frac{1}{2}}]. \tag{42}$$

Then we have the following inequality

$$\begin{aligned}
&\inf_{\boldsymbol{\psi}_0^m\in\mathcal{N}_0',\{\boldsymbol{\phi}_k^m\in\mathcal{N}_k\}_{k=0}^{N_T-1}} \mathbb{E}|\boldsymbol{g}(\boldsymbol{X}_T^\pi) - \boldsymbol{Y}_T^{\pi,m}|^2 \\
&\le C\Big[\|\pi\| + \inf_{\boldsymbol{\psi}_0^m\in\mathcal{N}_0',\{\boldsymbol{\phi}_k^m\in\mathcal{N}_k\}_{k=0}^{N_T-1}} \{\mathbb{E}|\boldsymbol{Y}_0^m - \boldsymbol{\psi}_0^m(\boldsymbol{x}_0)|^2 \\
&\qquad + \sum_{k=0}^{N_T-1} \mathbb{E}\|\hat{\boldsymbol{Z}}_{t_k}^m - \boldsymbol{\phi}_k^m(\boldsymbol{X}_{t_k}^\pi)\|_F^2 \Delta t_k\}\Big],
\end{aligned} \tag{43}$$

where $\mathcal{N}_0'$ *and* $\{\mathcal{N}_k\}_{k=0}^{N_T-1}$ *are hypothesis spaces for neural network architectures to approximate* $\boldsymbol{Y}_0$ *and* $\boldsymbol{Z}_{t_k}$, $\hat{\boldsymbol{Z}}_{t_k}^m = (\Delta t_k)^{-1}\mathbb{E}[\int_{t_k}^{t_{k+1}} \boldsymbol{Z}_t^m \,\mathrm{d}t|\boldsymbol{X}_{t_k}^\pi]$ *and* C *is a constant only depending on* T, L, M, M', K, L' *and* $\mathbb{E}|\boldsymbol{x}_0|^2$. *We still refer* $\mathcal{N}_0'$ *and* $\mathcal{N}_k$ *as the hypothesis spaces for* $\boldsymbol{\psi}_0^m : \mathbb{R}^n \to \mathbb{R}^N$, $\boldsymbol{\phi}_k^m : \mathbb{R}^n \to \mathbb{R}^{k\times N}$, *without introducing superscript* m *to indicate the stage.*

Remark 5. We have the following remarks regarding Theorem 3:

(1) The interpretation of Theorem 3 is similar to that of Theorem 1. The first inequality (41) shows that the distance between the true solution of BSDE (9) and the output of the deep BSDE method at stage m can be controlled together by the mesh size, the error of the initial policy and the loss functions achieved at all the previous stages. The second inequality (43) states that the loss function of deep BSDE method at each stage is small if the approximation capability of the parametric function spaces ($\mathcal{N}_0'$ and $\{\mathcal{N}_k\}_{k=0}^{N_T-1}$) is high. The overall message conveyed in Theorem 3 is that, if the deep BSDE method can solve each sub-problem accurately enough, the deep fictitious play method will produce a strategy close to the Nash equilibrium.

(2) Note that there is a slight abuse of notation in the statement of Theorem 3, since $(\boldsymbol{Y}_t^m, \boldsymbol{Z}_t^m)$ and $\boldsymbol{\alpha}^m$ have already been introduced in Sections 4.1 and 3.1 (cf. equations (30) and (14)), as the theoretical solution from the decoupling step at stage m. In this section, to avoid introducing further complicated notations, we still refer $(\boldsymbol{Y}_t^m, \boldsymbol{Z}_t^m)$ as the theoretical solution depending on $\boldsymbol{\alpha}^{m-1}$, but $\boldsymbol{\alpha}^{m-1}$ is the interpolation (42) of the deep BSDE solution $\boldsymbol{\alpha}^{\pi,m-1}$ at stage $m-1$. Nevertheless, the relation between $(\boldsymbol{Y}^m, \boldsymbol{Z}^m)$ and the interpolated strategy $\boldsymbol{\alpha}^{m-1}$ in Theorem 3 remains the same as the relation between $(\boldsymbol{Y}^m, \boldsymbol{Z}^m)$ and the exact strategy $\boldsymbol{\alpha}^{m-1}$ in Theorem 2, thus some estimates follow using the same derivations as in the proof of Theorem 2. In particular, we can obtain that there exists positive constants β_0 and C only depending on T, L, M and M' such that for any $\beta > \beta_0$,

$$\sup_{0\le t\le T} \mathbb{E}|\boldsymbol{Y}_t - \boldsymbol{Y}_t^m|^2 \le C\int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}^{m-1}(t,\boldsymbol{X}_t)|^2 \,\mathrm{d}t, \tag{44}$$

$$\int_0^T e^{\beta t}\mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_t^m\|_F^2 \,\mathrm{d}t \le \frac{C}{\beta-\beta_0}\int_0^T e^{\beta t}\mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}^{m-1}(t,\boldsymbol{X}_t)|^2 \,\mathrm{d}t. \tag{45}$$

where $\boldsymbol{\alpha}^{m-1}$ follows (42) and is the interpolation of strategies computed numerically at stage $m-1$.

(3) The interpolation (42) is indeed needed to apply Theorem 1 on $(\boldsymbol{Y}^{\pi,m}_{\pi(t)}, \boldsymbol{Z}^{\pi,m}_{\pi(t)})$ and $(\boldsymbol{Y}^m_t, \boldsymbol{Z}^m_t)$. The particular form (42) ensures that the Hölder continuity, as a prerequisite of Theorem 1, is preserved after the extension (cf. [53]).

(4) The inequality (43) shares the same implication as (25) in Theorem 1. That is, a small loss is attainable if the hypothesis spaces $\mathcal{N}_0'$ and $\mathcal{N}_k$ can approximate particular functions well, and this is feasible for deep neural networks given universal approximation results. Therefore, the choice of using feed-forward neural networks as hypothesis spaces $\mathcal{N}_0'$ and $\mathcal{N}_k$ is justified.

Proof. Throughout this proof, we will use C to denote a positive constant depending only on T, L, M, M', K, L' and $\mathbb{E}|\boldsymbol{x}_0|^2$ and use $C(\cdot)$ to denote a positive constant depending on all the above constants and the arguments represented by $\cdot$. Both C and $C(\cdot)$ may vary from line to line.

Since $\boldsymbol{\alpha}^{\pi,m} \in \mathcal{N}^\pi$, with conditions (40) and (42), we obtain (cf. [53])

$$\begin{aligned} &|\boldsymbol{\alpha}^m(t_1,\boldsymbol{x}_1) - \boldsymbol{\alpha}^m(t_2,\boldsymbol{x}_2)|^2 \le L'[|t_1-t_2| + |\boldsymbol{x}_1-\boldsymbol{x}_2|^2], \\ &|\boldsymbol{\alpha}^m(t,\boldsymbol{x})| \le L'(1+\sqrt{\|\pi\|}+|\boldsymbol{x}|^2). \end{aligned} \tag{46}$$

Thus, the inequality (43) follows from Theorem 1.

We next prove the inequality (41). As before, we will focus on proving the case of fictitious play in the sequel, and we claim that the statements also hold for policy update using a similar argument.

Recalling the $\{\boldsymbol{X}^\pi_{t_k}\}_{0\le k\le N_T-1}$ in (34), we then define the Euler-type scheme for BSDE system (9) as follows:

$$\begin{cases} \boldsymbol{Y}^\pi_{t_k} = \mathbb{E}[\boldsymbol{Y}^\pi_{t_{k+1}}|X_{t_k}] + \check{\boldsymbol{H}}(t_k, \boldsymbol{X}^\pi_{t_k}, \boldsymbol{Z}^\pi_{t_k})\Delta t_k, \quad \boldsymbol{Y}^\pi_T = g(\boldsymbol{X}^\pi_T), \\ \boldsymbol{Z}^\pi_{t_k} = \dfrac{1}{\Delta t_k}\mathbb{E}[(\boldsymbol{Y}^\pi_{t_{k+1}})^\intercal \Delta \boldsymbol{W}_k | \boldsymbol{X}^\pi_{t_k}], \quad \forall k = 0,1,\dots,N_T-1. \end{cases}$$

With Assumptions 1 and 2, classical estimations of the discretization error gives

$$\sup_{0\le t\le T}[\mathbb{E}|\boldsymbol{X}_t - \boldsymbol{X}^\pi_{\pi(t)}|^2 + \mathbb{E}|\boldsymbol{Y}_t - \boldsymbol{Y}^\pi_{\pi(t)}|^2] + \int_0^T \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}^\pi_{\pi(t)}\|_F^2 \,\mathrm{d}t \le C\|\pi\|. \tag{47}$$

For the $\boldsymbol{Z}$-part error, we decompose it into two terms by the Cauchy-Schwartz inequality:

$$\int_0^T \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}^{\pi,m+1}_{\pi(t)}\|_F^2 \,\mathrm{d}t \le 2\int_0^T \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}^{m+1}_t\|_F^2 \,\mathrm{d}t + 2\int_0^T \mathbb{E}\|\boldsymbol{Z}^{m+1}_t - \boldsymbol{Z}^{\pi,m+1}_{\pi(t)}\|_F^2 \,\mathrm{d}t. \tag{48}$$

A similar inequality can be written on the $\boldsymbol{Y}$-part error. For both of them, the second term is taken care by applying Theorem 1 to $(\boldsymbol{Y}^m_t, \boldsymbol{Z}^m_t)$. More precisely, applying Theorem 1 with Assumptions 2–4, one has:

$$\begin{aligned} \sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{Y}^{m+1}_t - \boldsymbol{Y}^{\pi,m+1}_{\pi(t)}|^2 + \int_0^T \mathbb{E}\|\boldsymbol{Z}^{m+1}_t - \boldsymbol{Z}^{\pi,m+1}_{\pi(t)}\|_F^2 \,\mathrm{d}t \\ \le C\left[\|\pi\| + \mathbb{E}|\boldsymbol{Y}^{\pi,m+1}_T - \boldsymbol{g}(\boldsymbol{X}^\pi_T)|^2\right], \end{aligned} \tag{49}$$

where $(\boldsymbol{Y}^{\pi,m}_{\pi(t)}, \boldsymbol{Z}^{\pi,m}_{\pi(t)})$ is defined in (34). For the first term in (48), we recall the inequality (45) (choosing $\beta = \beta_0 + 1$) and deduce

$$\int_0^T \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_t^{m+1}\|_F^2 \,\mathrm{d}t \le C\int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}^m(t,\boldsymbol{X}_t)|^2 \,\mathrm{d}t \le C[\int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t + \|\pi\|], \quad (50)$$

where we have used

$$\int_0^T \mathbb{E}|\boldsymbol{\alpha}_{\pi(t)}^{\pi,m} - \boldsymbol{\alpha}^m(t,\boldsymbol{X}_t)|^2 \,\mathrm{d}t \le C\|\pi\| \quad (51)$$

as a consequence of (46) and (47). Combining (48)–(50), we claim that

$$\int_0^T \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_{\pi(t)}^{\pi,m+1}\|_F^2 \,\mathrm{d}t \le C[\int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t + \|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^\pi)|^2]. \quad (52)$$

Using equations (44), (49) and (51), we can similarly obtain that

$$\sup_{0\le t\le T} \mathbb{E}|\boldsymbol{Y}_t - \boldsymbol{Y}_{\pi(t)}^{\pi,m+1}| \le C[\int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t + \|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^\pi)|^2]. \quad (53)$$

We next derive an estimate that is useful in controlling the $\boldsymbol{\alpha}$-part error. We first require $\eta_\epsilon > \beta_0$, then we have

$$\begin{aligned}
&\sum_{k=0}^{N_T-1} e^{\eta t_k}\mathbb{E}\|\boldsymbol{Z}_{t_k}^\pi - \boldsymbol{Z}_{t_k}^{\pi,m+1}\|_F^2 \Delta t_k = \int_0^T e^{\eta\pi(t)}\mathbb{E}\|\boldsymbol{Z}_{\pi(t)}^\pi - \boldsymbol{Z}_{\pi(t)}^{\pi,m+1}\|_F^2 \,\mathrm{d}t \\
&\le 3\int_0^T e^{\eta t}[\mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_t^{m+1}\|_F^2 + \mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_{\pi(t)}^\pi\|_F^2 + \mathbb{E}\|\boldsymbol{Z}_t^{m+1} - \boldsymbol{Z}_{\pi(t)}^{\pi,m+1}\|_F^2] \,\mathrm{d}t \\
&\le 3\int_0^T e^{\eta t}\mathbb{E}\|\boldsymbol{Z}_t - \boldsymbol{Z}_t^{m+1}\|_F^2 \,\mathrm{d}t + C(\eta)\left[\|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^\pi)|^2\right] \\
&\le \frac{C}{\eta - \beta_0}\int_0^T e^{\eta t}\mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t + C(\eta)\left[\|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^\pi)|^2\right], \quad (54)
\end{aligned}$$

where we have used the Cauchy-Schwart inequality, inequalities (45), (47), (49) and (51).

For the $\boldsymbol{\alpha}$-part error, it suffices to control $\int_0^T e^{\eta t}\mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t$ and we plan to

(1) express $\mathrm{I} := \sum_{k=0}^{N_T-1} e^{\eta t_k}\mathbb{E}|\boldsymbol{\alpha}^* - \tilde{\boldsymbol{\alpha}}^{\pi,m+1}|^2(t_k, \boldsymbol{X}_{t_k}^\pi)\Delta t_k$ in terms of $\int_0^T e^{\eta t}\mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t$;
(2) obtain the estimate of $\mathrm{II} := \sum_{k=0}^{N_T-1} e^{\eta t_k}\mathbb{E}|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{\pi,m+1}|^2(t_k, \boldsymbol{X}_{t_k}^\pi)\Delta t_k \le \mathrm{I}$ by the property (37) of $P_{\mathcal{N}^\pi}$;
(3) take care the difference between the $\boldsymbol{\alpha}$-part error and II by III which is defined by:

$$\mathrm{III} := \int_0^T e^{\eta t}\mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}^*(\pi(t), \boldsymbol{X}_{\pi(t)}^\pi)|^2 \,\mathrm{d}t \le C(\eta)\|\pi\|. \quad (55)$$

Step (2) follows from the fact that $\boldsymbol{\alpha}^{\pi,m+1}$ is defined as the projection of $\tilde{\boldsymbol{\alpha}}^{\pi,m+1}$ into $\mathcal{N}^\pi$, and that $\boldsymbol{\alpha}^* \in \mathcal{N}^\pi$ if viewed as a function on $\mathcal{T}\times\mathbb{R}^n$. Step (3) is a consequence of Assumption 4 and (47). So it remains to address step (1).

To this end, we define $\boldsymbol{\alpha}_{t_k}^{\pi,*} = \boldsymbol{\alpha}(t_k, \boldsymbol{X}_{t_k}^{\pi}, \boldsymbol{Z}_{t_k}^{\pi})$, then $\boldsymbol{\alpha}_{t_k}^{\pi,*} = \boldsymbol{a}(t_k, \boldsymbol{X}_{t_k}^{\pi}, \boldsymbol{\alpha}_{t_k}^{\pi,*}, \boldsymbol{Z}_{t_k}^{\pi})$, and $\tilde{\boldsymbol{\alpha}}_{t_k}^{\pi,m} = \tilde{\boldsymbol{\alpha}}^{\pi,m}(t_k, \boldsymbol{X}_{t_k}^{\pi})$, then $\tilde{\boldsymbol{\alpha}}_{t_k}^{\pi,m+1} = \boldsymbol{a}(t_k, \boldsymbol{X}_{t_k}^{\pi}, \boldsymbol{\alpha}_{t_k}^{\pi,m}, \boldsymbol{Z}_{t_k}^{\pi,m+1})$. Thus, for any $\lambda > 0$, using the AM-GM inequality

$$\begin{aligned}
\mathrm{I} \leq & (1+\lambda^{-1}) \sum_{k=0}^{N_T-1} e^{\eta t_k} \mathbb{E}|\boldsymbol{\alpha}_{t_k}^{\pi,*} - \tilde{\boldsymbol{\alpha}}_{t_k}^{\pi,m+1}|^2 \Delta t_k \\
& + C(\lambda) \sum_{k=0}^{N_T-1} e^{\eta t_k} \mathbb{E}|\boldsymbol{\alpha}_{t_k}^{\pi,*} - \boldsymbol{\alpha}^*(t_k, \boldsymbol{X}_{t_k}^{\pi})|^2 \Delta t_k \\
:= & (1+\lambda^{-1})\mathrm{I}^{(1)} + C(\lambda)\mathrm{I}^{(2)}. \qquad (56)
\end{aligned}$$

For term $\mathrm{I}^{(1)}$, using (54) and the Lipschitz condition of $\boldsymbol{a}$ in Assumption 2(2), we obtain

$$\begin{aligned}
\mathrm{I}^{(1)} \leq & \sum_{k=0}^{N_T-1} e^{\eta t_k} \left[L\mathbb{E}\|\boldsymbol{Z}_{t_k}^{\pi} - \boldsymbol{Z}_{t_k}^{\pi,m+1}\|_F^2 + a_\alpha \mathbb{E}|\boldsymbol{\alpha}_{t_k}^{\pi,*} - \boldsymbol{\alpha}_{t_k}^{\pi,m}|^2 \right] \Delta t_k \\
\leq & [a_\alpha(1+\lambda^{-1}) + \frac{C}{\eta - \beta_0}] \int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t \\
& + C(\lambda,\eta) \left[\|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^{\pi})|^2 \right],
\end{aligned}$$

where we remove $C(\cdot)$'s dependence on a_α using $a_\alpha < 1$ and we have also used

$$\int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,*}|^2 = \int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t) - \boldsymbol{\alpha}(\pi(t), \boldsymbol{X}_{\pi(t)}^{\pi}, \boldsymbol{Z}_{\pi(t)}^{\pi}|^2 \,\mathrm{d}t \leq C(\eta)\|\pi\|.$$

Combining the last inequality with (55) yields the estimate $\mathrm{I}^{(2)} \leq C(\eta)\|\pi\|$. Now plugging the estimates of $\mathrm{I}^{(1)}$ and $\mathrm{I}^{(2)}$ into (56) and following step (1)–(3), we obtain:

$$\begin{aligned}
& \int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m+1}|^2 \,\mathrm{d}t \\
\leq & (1+\lambda^{-1}) \int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{\pi,m+1}|^2(\pi(t), \boldsymbol{X}_{\pi(t)}^{\pi}) \,\mathrm{d}t + C(\lambda)\mathrm{III} \\
\leq & (1+\lambda^{-1}) e^{\eta\|\pi\|}\mathrm{II} + C(\lambda)\mathrm{III} \leq (1+\lambda^{-1})\mathrm{II} + C(\lambda)\mathrm{III} + C(\lambda,\eta)\|\pi\| \\
\leq & (1+\lambda^{-1})^2 \mathrm{I}^{(1)} + C(\lambda)(\mathrm{I}^{(2)} + \mathrm{III}) + C(\lambda,\eta)\|\pi\| \\
\leq & (1+\lambda^{-1})^2 [a_\alpha(1+\lambda^{-1}) + \frac{C}{\eta - \beta_0}] \int_0^T \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t \\
& + C(\lambda,\eta) \left[\|\pi\| + \mathbb{E}|\boldsymbol{Y}_T^{\pi,m+1} - \boldsymbol{g}(\boldsymbol{X}_T^{\pi})|^2 \right],
\end{aligned}$$

where we have used $e^{\eta\|\pi\|}\mathrm{II} \leq \mathrm{II} + C(\eta)\|\pi\|\mathrm{II} \leq \mathrm{II} + C(\eta)\|\pi\| \sum_{k=0}^{N_T-1} (\mathbb{E}|\boldsymbol{X}_{t_k}^{\pi}|^2 + 1)\Delta t_k \leq \mathrm{II} + C(\eta)\|\pi\|$. Let λ and η_ϵ be large enough such that $(1+\lambda^{-1})^2[a_\alpha(1+\lambda^{-1}) + \frac{C}{\eta_\epsilon - \beta_0}] \leq a_\alpha + \epsilon$, then for $\eta > \eta_\epsilon$,

$$\begin{aligned}
\int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,m}|^2 \,\mathrm{d}t \leq & C(\eta,\epsilon)[\|\pi\| + (a_\alpha + \epsilon)^m \int_0^T e^{\eta t} \mathbb{E}|\boldsymbol{\alpha}_t^* - \boldsymbol{\alpha}_{\pi(t)}^{\pi,0}|^2 \,\mathrm{d}t \\
& + \sum_{j=1}^{m} (a_\alpha + \epsilon)^{m-j} \mathbb{E}|\boldsymbol{g}(\boldsymbol{X}_T^{\pi}) - \boldsymbol{Y}_T^{\pi,j}|^2].
\end{aligned}$$

Combining the above inequality with inequality (52) and (53), we obtain our result. □

Here are some remarks regarding Theorem 3 on its implication for numerical algorithms. The primary concern is how we can implement the projection mapping in practice if wished. Note that we choose 1/2-Hölder continuity in time in Assumption 4 for the generality of the result, although numerically it is challenging to guarantee the Hölder continuity. If we replace that with the Lipschitz continuity in time, as a more restrictive condition, and instead consider the projection onto the space with the Lipschitz continuity, the estimates still hold. Accordingly, there are some practical approaches in the literature on ensuring the Lipschitz continuity of deep neural networks that can be introduced in our algorithms. For instance, [23] gives an efficient and accurate estimation of Lipschitz constants for deep Neural networks, and [59] further extends it for robust training with regularization to keep the Lipschitz constant of neural networks small. In practice, Wasserstein GAN [2, 28] has shown remarkable performance when using weight clipping as a loose but efficient way to impose the Lipschitz constraint. Therefore we can leverage similar techniques to keep the Lipschitz regularity during the training of the deep fictitious play. Also, notice that in the above, we define a single projection $\mathcal{N}^\pi$ from the space of all players' strategies $\boldsymbol{\alpha}^{\pi,m}$, in consideration of the simplicity of the statement. One can also use the projection of $\alpha^{i,\pi,m}$ for each player with possibly easier numerical implementation and the same theoretical guarantee.

4.3. **On the ϵ-Nash equilibrium.** This section combines the previous analysis, identifies the ϵ-Nash equilibrium produced by the deep fictitious play, and evaluates its performance on the original game.

Theorem 4. *Under Assumptions 1–4, if $\hat{\boldsymbol{\alpha}}$ is a policy function on $[0,T]\times\mathbb{R}^n$ and Lipschitz in $\boldsymbol{x}$, and*

$$\int_0^T \mathbb{E}|\boldsymbol{\alpha}^*(t,\boldsymbol{X}_t)-\hat{\boldsymbol{\alpha}}(t,\boldsymbol{X}_t)|^2\,\mathrm{d}t \le \epsilon, \tag{57}$$

where $\boldsymbol{X}_t$ is the forward component of (9), then

(1) Given $\hat{\boldsymbol{\alpha}}$, the game values produced by $\hat{\boldsymbol{\alpha}}$ are near the Nash equilibrium, i.e.,

$$|\tilde{\boldsymbol{J}}_0(\hat{\boldsymbol{\alpha}})-\boldsymbol{J}_0(\boldsymbol{\alpha}^*)|^2 \le C\epsilon, \text{ and } |\boldsymbol{J}_0(\hat{\boldsymbol{\alpha}})-\boldsymbol{J}_0(\boldsymbol{\alpha}^*)|^2 \le C\epsilon, \tag{58}$$

where $\tilde{\boldsymbol{J}}_0(\hat{\boldsymbol{\alpha}})=[\tilde{J}_0^1(\hat{\boldsymbol{\alpha}}),\dots,\tilde{J}_0^N(\hat{\boldsymbol{\alpha}})]$ with $\tilde{J}_0^i(\hat{\boldsymbol{\alpha}}):=\inf_{\beta^i\in\mathbb{A}^i} J_0^i(\beta^i,\hat{\boldsymbol{\alpha}}^{-i})$, $\boldsymbol{J}_0(\hat{\boldsymbol{\alpha}})=[J_0^1(\hat{\boldsymbol{\alpha}}),\dots,J_0^N(\hat{\boldsymbol{\alpha}})]$ with J_0^i defined in (3). Thus, there exists $0<\epsilon_i\ll 1$ such that $\sum_{i=1}^N\epsilon_i^2\le C\epsilon$ and

$$J_0^i(\beta^i,\hat{\boldsymbol{\alpha}}^{-i}) \ge J_0^i(\hat{\boldsymbol{\alpha}})-\epsilon_i, \quad \forall\beta^i\in\mathbb{A}^i \text{ and } i\in\mathcal{I}. \tag{59}$$

Here C is a constant depending on T, L, M and M' which may vary from line to line in the proof.

(2) The generated game paths $\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}$ are close to the paths $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}$ associated with the Nash equilibrium:

$$\mathbb{E}\sup_{0\le t\le T}|\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}-\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}|^2 \le C(\lambda)\epsilon^\lambda,$$

where $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}$ and $\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}$ follow (1) with the true Nash equilibrium strategy $\boldsymbol{\alpha}^*$ and $\hat{\boldsymbol{\alpha}}$. Here λ is an arbitrary constant in $(0,1)$, and $C(\lambda)$ is a constant depending on T, L, M, M' and λ.

Immediately, we have the following corollary.

Corollary 1. *Under Assumptions 1–4, assuming the sub-problems (34) are solved accurately enough at all stages, i.e.,*

$$\mathbb{E}|\boldsymbol{g}(X_T^\pi) - \boldsymbol{Y}_T^{\pi,j}|^2 \le C\epsilon^2, \quad \forall j \le m, \tag{60}$$

here C is a constant depending only on T, L, M, M', K, L' and $\mathbb{E}|\boldsymbol{x}_0|^2$. Then, for sufficiently large m and small mesh size $\|\pi\|$, the strategy $\boldsymbol{\alpha}^m$ defined in (42), as an interpolated policy based on the deep fictitious play, forms an ϵ-Nash equilibrium.

Proof. This follows from (59) in Theorem 4, with the assumptions satisfied according to equations (41), (46) and (51). □

Remark 6. As mentioned in Remark 3, there are still some theoretical issues unsolved regarding the approximation error and optimization of the deep BSDE method. The analysis of the deep fictitious play method has similar issues that remain open. To circumvent these issues and have a rigorous statement for ϵ-Nash equilibrium, we introduce assumption (60). In practice, an observable proxy of (60) is the training loss of the deep BSDE method evaluated by its Monte Carlo counterpart.

Proof of Theorem 4. The proof of item (1) relies on the estimates of BSDEs presented previously. Let $(\boldsymbol{X}_t, Y_t^{i,\mathrm{FP}}, Z_t^{i,\mathrm{FP}})$ solve (15) with $\boldsymbol{\alpha}^{-i,m}$ replaced by $\hat{\boldsymbol{\alpha}}^{-i}$, where the superscript FP is used to emphasize that the fictitious play strategy in adopted at the decoupling step. By the nonlinear Feynman-Kac formula (cf. [57, 21, 58]) and the associated HJB equation, we have $\mathbb{E}[Y_0^{i,\mathrm{FP}}] = \tilde{J}_0^i(\hat{\boldsymbol{\alpha}})$. Therefore, we have

$$|\tilde{\boldsymbol{J}}_0(\hat{\boldsymbol{\alpha}}) - \boldsymbol{J}_0(\boldsymbol{\alpha}^*)|^2 = |\mathbb{E}[\boldsymbol{Y}_0^{\mathrm{FP}}] - \mathbb{E}[\boldsymbol{Y}_0]|^2 \le \mathbb{E}|\boldsymbol{Y}_0^{\mathrm{FP}} - \boldsymbol{Y}_0|^2. \tag{61}$$

To bound the above term, we claim a stronger result:

$$\sup_{0\le t\le T} \mathbb{E}|\boldsymbol{Y}_t^{\mathrm{FP}} - \boldsymbol{Y}_t|^2 + \int_0^T \mathbb{E}\|\boldsymbol{Z}_t^{\mathrm{FP}} - \boldsymbol{Z}_t\|_F^2 \,\mathrm{d}t \le C\epsilon, \tag{62}$$

where $(\boldsymbol{Y}_t, \boldsymbol{Z}_t)$ solves (9), $\boldsymbol{Y}_t^{\mathrm{FP}} = [Y_t^{1,\mathrm{FP}}, \dots, Y_t^{N,\mathrm{FP}}]^\intercal$, and $\boldsymbol{Z}_t^{\mathrm{FP}} = [Z_t^{i,\mathrm{FP}}, \dots, Z_t^{N,\mathrm{FP}}]^\intercal$, as a consequence of (44), (45) and (57).

If we let $(\boldsymbol{X}_t, Y_t^{i,\mathrm{PU}}, Z_t^{i,\mathrm{PU}})$ solve (17) with $\boldsymbol{\alpha}^m$ replaced by $\hat{\boldsymbol{\alpha}}$, where the superscript PU emphasizes that policy update is used at the decoupling step, an argument similar to (61) and (62) can give the second inequality in (58). Then (59) is obtained by observing $J_0^i(\beta^i, \hat{\boldsymbol{\alpha}}^{-i}) \ge \tilde{J}_0^i(\hat{\boldsymbol{\alpha}}), \quad \forall \beta^i \in \mathbb{A}^i$, and $|\tilde{\boldsymbol{J}}_0(\hat{\boldsymbol{\alpha}}) - \boldsymbol{J}_0(\hat{\boldsymbol{\alpha}})|^2 \le C\epsilon$.

We now prove item (2). Under the standing assumptions, we first observe that $b^1(t,\boldsymbol{x}) := b(t,\boldsymbol{x},\boldsymbol{\alpha}^*(t,\boldsymbol{x})) = \Sigma(t,\boldsymbol{x})\phi(t,\boldsymbol{x},\boldsymbol{\alpha}^*(t,\boldsymbol{x}))$ and $b^2(t,\boldsymbol{x}) := b(t,\boldsymbol{x},\hat{\boldsymbol{\alpha}}(t,\boldsymbol{x})) = \Sigma(t,\boldsymbol{x})\phi(t,\boldsymbol{x},\hat{\boldsymbol{\alpha}}(t,\boldsymbol{x}))$ are Lipschitz in $\boldsymbol{x}$. Thus $\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}$ is well-defined, and the standard estimates in SDE gives (cf. [71, Theorem 3.2.4])

$$\mathbb{E}\sup_{0\le t\le T}|\boldsymbol{X}_t^{\boldsymbol{\alpha}^*} - \boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}|^2 \le C\mathbb{E}\left[\left(\int_0^T |b^1(t,\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - b^2(t,\boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|\,\mathrm{d}t\right)^2\right]. \tag{63}$$

To bound the right-hand side above with the condition (57), let us define a new probability measure $\mathbb{Q}$, and denote by $\mathcal{Z}$ the Radon-Nikodym derivative: $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \equiv \mathcal{Z} := \exp\left\{-\int_0^T \phi_t^{\boldsymbol{\alpha}^*}\cdot \mathrm{d}\boldsymbol{W}_t - \frac{1}{2}\int_0^T |\phi_t^{\boldsymbol{\alpha}^*}|^2\,\mathrm{d}t\right\}$, where $\phi_t^{\boldsymbol{\alpha}^*} := \phi(t, X_t^{\boldsymbol{\alpha}^*}, \boldsymbol{\alpha}^*(t, X_t^{\boldsymbol{\alpha}^*}))$. By Assumptions 2, the Novikov condition is fulfilled. Thus $\mathbb{Q} \sim \mathbb{P}$, and $\boldsymbol{W}_\cdot^{\mathbb{Q}} :=$

$\boldsymbol{W}_\cdot + \int_0^\cdot \phi_t^{\boldsymbol{\alpha}^*}\,\mathrm{d}s$ is a standard Brownian motion under $\mathbb{Q}$. In particular, the process $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}$ can be rewritten as $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*} = \boldsymbol{x}_0 + \int_0^t \Sigma(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}^*})\,\mathrm{d}\boldsymbol{W}_s^{\mathbb{Q}}$, and immediately from (57) we have

$$\int_0^T \mathbb{E}_{\mathbb{Q}}|\boldsymbol{\alpha}^*(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - \hat{\boldsymbol{\alpha}}(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|^2\,\mathrm{d}t \le \epsilon, \tag{64}$$

where we denote by $\mathbb{E}_{\mathbb{Q}}$ the expectation under $\mathbb{Q}$. We next compute a bound for $\mathcal{Z}^{-\gamma}$ under $\mathbb{Q}$, for $\gamma > 2$:

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[\mathcal{Z}^{-\gamma}] =& \mathbb{E}_{\mathbb{Q}}\left[\exp\left\{\gamma\int_0^T \phi_t^{\boldsymbol{\alpha}^*}\cdot\,\mathrm{d}\boldsymbol{W}_t^{\mathbb{Q}} - \frac{\gamma}{2}\int_0^T |\phi_t^{\boldsymbol{\alpha}^*}|^2\,\mathrm{d}t\right\}\right]\\
\le& \mathbb{E}_{\mathbb{Q}}^{1/2}\left[\exp\left\{2\gamma\int_0^T \phi_t^{\boldsymbol{\alpha}^*}\cdot\,\mathrm{d}\boldsymbol{W}_t^{\mathbb{Q}} - 2\gamma^2\int_0^T |\phi_t^{\boldsymbol{\alpha}^*}|^2\,\mathrm{d}t\right\}\right]\\
&\times \mathbb{E}_{\mathbb{Q}}^{1/2}\left[\exp\left\{(2\gamma^2-\gamma)\int_0^T |\phi_t^{\boldsymbol{\alpha}^*}|^2\,\mathrm{d}t\right\}\right]\\
\le& e^{CT(\gamma^2-\frac{1}{2}\gamma)},
\end{aligned}$$

where $\mathbb{E}_{\mathbb{Q}}^p$ denotes $(\mathbb{E}_{\mathbb{Q}}[\cdot])^p$, and we use the Cauchy-Schwartz inequality, the martingale property, and the boundedness of ϕ (Assumption 2).

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_0^T |b^1(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - b^2(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|\,\mathrm{d}t\right)^2\right]\\
\le& \mathbb{E}_{\mathbb{Q}}^{1-\frac{1}{\gamma}}\left[\left(\int_0^T |b^1(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - b^2(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|\,\mathrm{d}t\right)^{\frac{2\gamma}{\gamma-1}}\right]\mathbb{E}_{\mathbb{Q}}^{\frac{1}{\gamma}}[\mathcal{Z}^{-\gamma}]\\
\le& C(\gamma)\mathbb{E}_{\mathbb{Q}}^{1-\frac{2}{\gamma}}\left[\left(\int_0^T |b^1(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - b^2(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|\,\mathrm{d}t\right)^2\right]\\
&\times \mathbb{E}_{\mathbb{Q}}^{\frac{1}{\gamma}}\left[\left(\int_0^T |b^1(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - b^2(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|\,\mathrm{d}t\right)^4\right]\\
\le& C(\gamma)\mathbb{E}_{\mathbb{Q}}^{1-\frac{2}{\gamma}}\left[\int_0^T |\boldsymbol{\alpha}^*(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*}) - \hat{\boldsymbol{\alpha}}(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}^*})|^2\,\mathrm{d}t\right] \le C(\gamma)\epsilon^{1-\frac{2}{\gamma}},
\end{aligned}$$

where we have consecutively used Hölder's inequality, the estimate of $\mathbb{E}_{\mathbb{Q}}[\mathcal{Z}^{-\gamma}]$, Hölder's inequality again, the Lipschitz property of $\phi(t, \boldsymbol{x}, \boldsymbol{\alpha})$, the boundedness of Σ and b (Assumption 2), and the estimate (64). Here $C(\gamma)$ is a constant depending on the T, L, M, M' and γ, which may vary from line to line. With (63) and noticing $0 < 1 - \frac{2}{\gamma} < 1$ we conclude. □

In practice, the game is played on $\mathcal{T}$ rather than $[0, T]$. Therefore, we define a discrete version of the stochastic differential game (1)–(2) and evaluate the performance of $\boldsymbol{\alpha}^{\pi,m}$ in section 4.2 on the discrete game. To be precise, given a policy function $\boldsymbol{\alpha}^\pi$ on $\mathcal{T}\times\mathbb{R}^n$, we define the discrete state process $\boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi}$ and discrete individual cost functional $J_0^{\pi,i}(\boldsymbol{\alpha}^\pi)$ as follows

$$\begin{aligned}
\boldsymbol{X}_{t_{k+1}}^{\pi,\boldsymbol{\alpha}^\pi} &= \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi} + b(t_k, \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi}, \boldsymbol{\alpha}^\pi(t_k, \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi}))\Delta t_k + \Sigma(t_k, \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi})\Delta\boldsymbol{W}_k, \quad (65)\\
J_0^{\pi,i}(\boldsymbol{\alpha}^\pi) &= \mathbb{E}[\sum_{j=0}^{N_T-1} f^i(t_k, \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi}, \boldsymbol{\alpha}^\pi(t_k, \boldsymbol{X}_{t_k}^{\pi,\boldsymbol{\alpha}^\pi}))\Delta t_k + g^i(\boldsymbol{X}_T^{\pi,\boldsymbol{\alpha}^\pi})],
\end{aligned}$$

with $\boldsymbol{X}_0^{\pi,\boldsymbol{\alpha}^\pi} = \boldsymbol{x}_0$. Note that when there are both π and $\boldsymbol{\alpha}$ in the superscript of $\boldsymbol{X}$, it refers to the discrete version of the original state (1), and when there is only π in the superscript, it refers to the discrete version of $\boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \Sigma(s, \boldsymbol{X}_s)\,\mathrm{d}\boldsymbol{W}_s$. We then state a discrete version of Theorem 4.

Theorem 5. *Under Assumptions 1–4, if $\hat{\boldsymbol{\alpha}}^\pi$ is a policy function on $\mathcal{T}\times\mathbb{R}^n$, Lipschitz in $\boldsymbol{x}$ and Hölder continuous with t: $|\hat{\boldsymbol{\alpha}}^\pi(t_1,\boldsymbol{x}_1) - \hat{\boldsymbol{\alpha}}^\pi(t_2,\boldsymbol{x}_2)|^2 \le L'[|t_1-t_2| + |\boldsymbol{x}_1-\boldsymbol{x}_2|^2]$, and*

$$\int_0^T \mathbb{E}|\boldsymbol{\alpha}^*(t,\boldsymbol{X}_t) - \hat{\boldsymbol{\alpha}}^\pi(\pi(t),\boldsymbol{X}^\pi_{\pi(t)})|^2\,\mathrm{d}t \le \epsilon, \tag{66}$$

then

(1) The value of the discrete game produced by $\hat{\boldsymbol{\alpha}}^\pi$ is close to the one associated with the Nash equilibrium of the continuous game, i.e.,

$$|\boldsymbol{J}_0^\pi(\hat{\boldsymbol{\alpha}}^\pi) - \boldsymbol{J}_0(\boldsymbol{\alpha}^*)|^2 \le C[\epsilon + \|\pi\|],$$

where $\boldsymbol{J}_0^\pi(\hat{\boldsymbol{\alpha}}^\pi) = [J_0^{\pi,1}(\hat{\boldsymbol{\alpha}}^\pi),\dots,J_0^{\pi,N}(\hat{\boldsymbol{\alpha}}^\pi)]$. Moreover, there exists $0<\epsilon_i\ll 1$ such that $\sum_{i=1}^N \epsilon_i^2 \le C[\epsilon+\|\pi\|]$ and $J_0^i(\beta^i,\hat{\boldsymbol{\alpha}}^{\pi,-i}) \ge J_0^{\pi,i}(\hat{\boldsymbol{\alpha}}^\pi) - \epsilon_i$, $\forall\beta^i\in\mathbb{A}^i$ and $i\in\mathcal{I}$. Here C is a constant depending on T, L, M, M', K, L' and $\mathbb{E}|\boldsymbol{x}_0|^2$, which may vary from line to line in the proof.

(2) The generated game paths $\boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{t_k}$ are close to the paths $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}$ associated with the Nash equilibrium:

$$\mathbb{E}\sup_{0\le t\le T}|\boldsymbol{X}_t^{\boldsymbol{\alpha}^*} - \boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{\pi(t)}|^2 \le C(\lambda)[\epsilon+\|\pi\|]^\lambda,$$

where $\boldsymbol{X}_t^{\boldsymbol{\alpha}^*}$ follows (1) with the true Nash equilibrium strategy $\boldsymbol{\alpha}^*$ and $\boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{t_k}$ follows (65), and $C(\lambda)$ is a constant depending on T, L, M, M', K, L', $\mathbb{E}|\boldsymbol{x}_0|^2$, and λ (an arbitrary constant in $(0,1)$).

Proof. Let $\hat{\boldsymbol{\alpha}}(t,\boldsymbol{x}) = \inf_{t'\in\mathcal{T}}[\hat{\boldsymbol{\alpha}}^\pi(t',\boldsymbol{x}) + L'|t'-t|^{\frac{1}{2}}]$, then with an argument similar to that in Theorem 3, $\hat{\boldsymbol{\alpha}}$ satisfies:

$$|\hat{\boldsymbol{\alpha}}(t_1,\boldsymbol{x}_1) - \hat{\boldsymbol{\alpha}}(t_2,\boldsymbol{x}_2)|^2 \le L'[|t_1-t_2| + |\boldsymbol{x}_1-\boldsymbol{x}_2|^2]. \tag{67}$$

By (47), (66) and (67), we have

$$\begin{aligned}&\int_0^T \mathbb{E}|\boldsymbol{\alpha}^* - \hat{\boldsymbol{\alpha}}|^2(t,\boldsymbol{X}_t)\,\mathrm{d}t \\ \le& 2\int_0^T \mathbb{E}|\boldsymbol{\alpha}^*(t,\boldsymbol{X}_t) - \hat{\boldsymbol{\alpha}}^\pi(\pi(t),\boldsymbol{X}^\pi_{\pi(t)})|^2\,\mathrm{d}t + 2\int_0^T \mathbb{E}|\hat{\boldsymbol{\alpha}}(t,\boldsymbol{X}_t) - \hat{\boldsymbol{\alpha}}^\pi(\pi(t),\boldsymbol{X}^\pi_{\pi(t)})|^2\,\mathrm{d}t \\ \le& C[\|\pi\|+\epsilon].\end{aligned} \tag{68}$$

By the regularity of $\hat{\boldsymbol{\alpha}}$ (*c.f.* (67)) and the standard estimates of the Euler Scheme of SDE (*c.f.* [47, Theorem 10.2.2]), we can obtain

$$\mathbb{E}\sup_{0\le t\le T}|\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}} - \boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{\pi(t)}|^2 \le C\|\pi\|. \tag{69}$$

Observing that

$$\begin{aligned}&\boldsymbol{J}_0^\pi(\hat{\boldsymbol{\alpha}}^\pi) - \boldsymbol{J}_0(\hat{\boldsymbol{\alpha}}) \\ =&\mathbb{E}\int_0^T [\boldsymbol{f}(\pi(t),\boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{\pi(t)},\hat{\boldsymbol{\alpha}}^\pi(\pi(t),\boldsymbol{X}^{\pi,\hat{\boldsymbol{\alpha}}^\pi}_{\pi(t)})) - \boldsymbol{f}(t,\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}},\hat{\boldsymbol{\alpha}}(t,\boldsymbol{X}_t^{\hat{\boldsymbol{\alpha}}}))]\,\mathrm{d}t\end{aligned}$$

$$+\,\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_T^{\pi,\hat{\boldsymbol{\alpha}}^\pi}) - \boldsymbol{g}(\boldsymbol{X}_T^{\hat{\boldsymbol{\alpha}}})],$$

with (67), (69) and Assumption 2, one has

$$|\boldsymbol{J}_0^\pi(\hat{\boldsymbol{\alpha}}^\pi) - \boldsymbol{J}_0(\hat{\boldsymbol{\alpha}})|^2 \le C\|\pi\|. \tag{70}$$

Finally, with (68), (69), (70) and Theorem 4, we reach all the conclusions of this theorem. □

5. **Numerical results.** We supplement our theoretical analysis with some numerical results in a symmetric game. We shall mainly focus on how deep BSDE performs when combined with policy update strategy in the decoupling step, *i.e.*, when solving (17). The same example using fictitious play strategy has been studied in [31] to where we refer readers for further details. As for numerical performances, We did not observe a prominent difference between the two decoupling methods. For games with asymmetric players implemented by the deep fictitious play, we refer to a recent work on modeling pandemic policies [68].

The example we present here is an inter-bank game concerning the systemic risk [16]. Assume an inter-bank market with N banks, and denote by $X_t^i \in \mathbb{R}$ the log-monetary reserves of bank i at time t. Its dynamics are modeled as the following diffusion processes,

$$\mathrm{d}X_t^i = [a(\overline{X}_t - X_t^i) + \alpha_t^i]\,\mathrm{d}t + \sigma(\rho\,\mathrm{d}W_t^0 + \sqrt{1-\rho^2}\,\mathrm{d}W_t^i), \quad \overline{X}_t = \frac{1}{N}\sum_{i=1}^N X_t^i, \quad i \in \mathcal{I}.$$

Here $a(\overline{X}_t - X_t^i)$ represents the rate at which bank i borrows from or lends to other banks in the lending market, while α_t^i denotes its control rate of cash flows to a central bank. The standard Brownian motions $\{W_t^i\}_{i=0}^N$ are independent, in which $\{W_t^i, i \ge 1\}$ stands for the idiosyncratic noises and W_t^0 denotes the systemic shock, or so-called common noise in the general context. To describe the model in the form of (1), we concatenate the log-monetary reserves X_t^i of N banks to form $\boldsymbol{X}_t^{\boldsymbol{\alpha}} = [X_t^1, \dots, X_t^N]^\intercal$. The associated drift term and diffusion term are defined as

$$b(t, \boldsymbol{x}, \boldsymbol{\alpha}) = [a(\bar{x} - x^1) + \alpha^1, \dots, a(\bar{x} - x^N) + \alpha^N]^\intercal \in \mathbb{R}^{N\times 1},$$
$$\Sigma(t, x) = [\sigma\rho\mathbf{1}_N, \sigma\sqrt{1-\rho^2}\mathbf{I}_N] \in \mathbb{R}^{N\times(N+1)},$$

and $\boldsymbol{W}_t = (W_t^0, \dots, W_t^N)$ is $(N+1)$-dimensional, where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x^i$, $\mathbf{1}_N$ is the N-vector of ones and $\mathbf{I}_N$ is the $N \times N$ identity matrix. The cost functional (3) that player i wishes to minimize has the form

$$f^i(t, \boldsymbol{x}, \boldsymbol{\alpha}) = \frac{1}{2}(\alpha^i)^2 - q\alpha^i(\bar{x} - x^i) + \frac{\epsilon}{2}(\bar{x} - x^i)^2, \quad g^i(\boldsymbol{x}) = \frac{c}{2}(\bar{x} - x^i)^2.$$

Under such specifications, the solution of this game admits a quadratic form whose coefficient functions can be solved from a Riccati equation. We direct the interested readers to [16, 31] for the detailed interpretation of this model and the explicit characterization of the solution. Note that such a setting does not satisfy Assumption 2. However we can still observe convergence in the numerical experiment, showing the robustness of the proposed algorithms and potential improvement of our theoretical analysis. We also remark that Assumption 2 can be satisfied if one truncates the f^i and g^i functions at large constants.

In our numerical computation, we choose $N = 10$, $T = 1$, $a = 0.1$, $q = 0.1$, $c = 0.5$, $\epsilon = 0.5$, $\rho = 0.2$, $\sigma = 1$. We discretize the time $[0, T]$ into $N_T = 40$ intervals and specify the hypothesis spaces $\mathcal{N}_0^{i'}$ and $\{\mathcal{N}_k^i\}_{k=0}^{N_T-1}$ for each player i as follows.

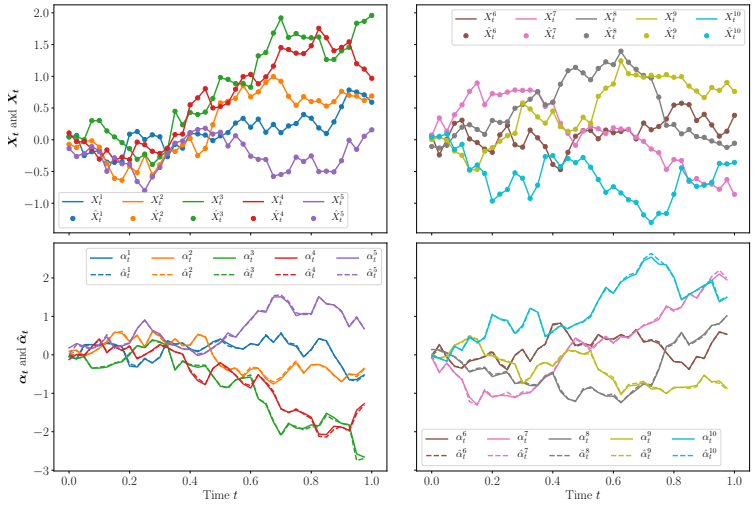

FIGURE 1. A sample path for all $N = 10$ players in the inter-bank game, obtained from decoupling the problem by policy update and solving the sub-problems with the Deep BSDE method. Top: the optimal state process X_t^i (solid lines) and its neural networks approximation $\hat{X}_t^i$ (circles), under the same realized path of Brownian motion. Bottom: comparisons of the strategies α_t^i and $\hat{\alpha}_t^i$ (dashed lines).

We parametrize $V^i(0, \boldsymbol{x})$ (the superscript m is dropped again for simplicity) with a neural network, denoted by $\mathrm{Net}_V(\boldsymbol{x})$, as the space $\mathcal{N}_0^{i'}$ of Y_0^i. We also parametrize $\nabla_{\boldsymbol{x}} V^i(t, \boldsymbol{x})$ with another network, denoted by $\mathrm{Net}_{\nabla V}(t, \boldsymbol{x})$, as the space of $\{\mathcal{N}_k^i\}_{k=0}^{N_T-1}$, in which the timestamp t_k is provided as another dimension of the input. This choice is in consistence with our theoretical analysis in Theorem 3 involving the linear interpolation of the strategy in time.

In this numerical example, we use fully-connected feedforward networks to instantiate both $\mathrm{Net}_V(\boldsymbol{x})$ and $\mathrm{Net}_{\nabla V}(t, \boldsymbol{x})$. Since the problem is homogeneous among all players, we let two networks share the same parameters among all players and only solve player 1's problem for updating the parameters. Both networks consist of three hidden layers with a width of 40. The activation function is hyperbolic tangent, and the technique of batch normalization [43] is adopted right after each linear transformation and before activation. For simplicity, we do not impose the projection procedure discussed in Section 4.2.

Regarding the optimization, the loss function in Deep BSDE is differentiable with respect to the network parameters. We can use backpropagation to derive the gradient of the loss function with respect to all the parameters in the neural networks and employ stochastic gradient descent (SGD) to optimize all the parameters. In this work, we use Adam optimizer [46] to optimize network parameters with constant learning rate 5e-4 and batch size 256. The parameters are updated by 30000 steps in total.

To implement the algorithm, we also need to specify the distribution of the initial state $\boldsymbol{x}_0$ in (15) or (17). We follow the same way as in [31]. Each component of $\boldsymbol{x}_0$, as the initial state of each player, is sampled independently from the uniform

distribution on $[-\delta_0, \delta_0]$. δ_0 is chosen such that in the process driven by the optimal policy $\boldsymbol{\alpha}^*$, the standard deviation of each component is approximately δ_0. In other words, δ_0 is determined as a fixed-point. The rationale for such a procedure is to make sure the data generated for the learning is representative enough in the whole state space.

Note that our technical assumptions are not strictly satisfied in this example, since $\boldsymbol{f}, \boldsymbol{g}$ are not Lipschitz continuous, ϕ is not uniform bounded, and T is not sufficiently small. Nevertheless, our numerical results show that the deep BSDE method can solve this game when combined with policy update. In particular, we compute the relative error of controls (proportional to the gradient of value function):

$$\text{RSE} = \frac{\sum_{\substack{0\le k\le N_T-1\\ 1\le j\le J}} \left(\nabla_{\boldsymbol{x}} V^1(t_k, \boldsymbol{x}_{t_k}^{(j)}) - \nabla_{\boldsymbol{x}} \widehat{V}^1(t_k, \boldsymbol{x}_{t_k}^{(j)})\right)^2}{\sum_{\substack{0\le k\le N_T-1\\ 1\le j\le J}} \left(\nabla_{\boldsymbol{x}} V^1(t_k, \boldsymbol{x}_{t_k}^{(j)}) - \overline{\nabla_{\boldsymbol{x}} V}^1\right)^2},$$

where V^1 is the true solution (of player 1), $\hat{V}^1$ is the prediction from the neural networks, and $\bar{V}^1$ (*resp.* $\overline{\nabla_{\boldsymbol{x}} V}^1$) is the average of V^1 (*resp.* $\nabla_{\boldsymbol{x}} V^1$) evaluated at all the indices j, k. To compute the relative error, we generate $J = 256$ ground truth sample paths $\{\boldsymbol{x}_{t_k}^{(j)}\}_{k=0}^{N_T-1}$ using Euler scheme based on (17) and the true optimal strategy. Note that the superscript (j) here does not mean the player index, but the j^{th} path for all players. The final RSE for the Deep BSDE method is 0.27%. Figures 1 presents one sample path for each player of the optimal state process X_t^i and the optimal control α_t^i *vs.* their approximations $\hat{X}_t^i, \hat{\alpha}_t^i$, with good agreement.

6. **Conclusion.** In this paper, we established the theoretical foundation for the deep fictitious play algorithm for finding Markovian Nash equilibrium proposed in [31]. Specifically, we proved the following three things: 1. The solutions of the decoupled sub-problems, if solved exactly and repeatedly, converge to the true Nash equilibrium; 2. The numerical error of each sub-problem, if solved by deep BSDE individually and repeatedly, converges to zero subject to the universal approximation capacity of neural networks; 3. The interpolated strategy based on the deep fictitious play algorithm forms a ϵ-Nash equilibrium, after sufficiently many stages m and with sufficiently small mesh $\|\pi\|$. We also generalize the algorithm by proposing a new approach to decouple the games, and present a numerical example in the end to show the empirical convergence beyond the technical assumptions used in the theorems. In the future, with this solidly established theory of deep fictitious play, we aim to study the competitions in finance, including P2P lending platforms from the Fintech industry and insurance markets. We also plan to generalize the theory and algorithm to stochastic differential games with delays.

Acknowledgments. R.H. was partially supported by the NSF grant DMS-1953035, and the Faculty Career Development Award, the Research Assistance Program Award, the Early Career Faculty Acceleration funding and the Regents' Junior Faculty Fellowship at the University of California, Santa Barbara.

REFERENCES

[1] A. Angiuli, J.-P. Fouque and M. Laurière, Unified reinforcement Q-learning for mean field game and control problems, arXiv:2006.13912, 2020.

[2] M. Arjovsky, S. Chintala and L. Bottou, Wasserstein generative adversarial networks, In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PLMR*, 2017, 214–223.
[3] R. Arora, A. Basu, P. Mianjy and A. Mukherjee, Understanding deep neural networks with rectified linear units, arXiv preprint, arXiv:1611.01491, 2016.
[4] E. Bayraktar, A. Budhiraja and A. Cohen, A numerical scheme for a mean field game in some queueing systems based on Markov chain approximation method, *SIAM J. Control Optim.*, **56** (2018), 4017–4044.
[5] C. Beck, S. Becker, P. Cheridito, A. Jentzen and A. Neufeld, Deep splitting method for parabolic PDEs, *SIAM J. Sci. Comput.*, **43** (2021), A3135–A3154.
[6] C. Beck, W. E and A. Jentzen, Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations, *J. Nonlinear Sci.*, **29** (2019), 1563–1619.
[7] A. Bensoussan, C. C. Siu, S. C. P. Yam and H. Yang, A class of non-zero-sum stochastic differential investment and reinsurance games, *Automatica J. IFAC*, **50** (2014), 2025–2037.
[8] U. Berger, Fictitious play in $2 \times n$ games, *J. Econom. Theory*, **120** (2005), 139–154.
[9] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext. Springer, New York, 2011.
[10] A. Briani and P. Cardaliaguet, Stable solutions in potential mean field game systems, *NoDEA Nonlinear Differential Equations Appl.*, **25** (2018), Paper No. 1, 26 pp.
[11] G. W. Brown, *Some Notes on Computation of Games Solutions*, Technical report, Rand Corp Santa Monica CA, 1949.
[12] G. W. Brown, Iterative solution of games by fictitious play, *Activity Analysis of Production and Allocation*, **13** (1951), 374–376.
[13] P. Cardaliaguet and S. Hadikhanloo, Learning in mean field games: The fictitious play, *ESAIM Control Optim. Calc. Var.*, **23** (2017), 569–591.
[14] P. Cardaliaguet and C.-A. Lehalle, Mean field game of controls and an application to trade crowding, *Math. Financ. Econ.*, **12** (2018), 335–363.
[15] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.
[16] R. Carmona, J.-P. Fouque and L.-H. Sun, Mean field games and systemic risk, *Commun. Math. Sci.*, **13** (2015), 911–933.
[17] P. Casgrain, B. Ning and S. Jaimungal, Deep Q-learning for Nash equilibria: Nash-DQN, arXiv:1904.10554, 2019.
[18] S. Chen, H. Yang and Y. Zeng, Stochastic differential games between two insurers with generalized mean-variance premium principle, *Astin Bull.*, **48** (2018), 413–434.
[19] E. J. Dockner, S. Jørgensen, N. V. Long and G. Sorger, *Differential Games in Economics and Management Science*, Cambridge University Press, 2000.
[20] W. E, J. Han and A. Jentzen, Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations, *Commun. Math. Stat.*, **5** (2017), 349–380.
[21] N. El Karoui, S. Peng and M. C. Quenez, Backward stochastic differential equations in finance, *Math. Finance*, **7** (1997), 1–71.
[22] R. Elie, J. Pérolat, M. Laurière, M. Geist and O. Pietquin, On the convergence of model free learning in mean field games, *AAAI-20 Technical Tracks 5*, Vol. 34, 2020. arXiv:1907.02633.
[23] M. Fazlyab, A. Robey, H. Hassani, M. Morari and G. Pappas, Efficient and accurate estimation of Lipschitz constants for deep neural networks, In *Advances in Neural Information Processing Systems*, (2019), 11427–11438.
[24] M. Germain, H. Pham and X. Warin, Deep backward multistep schemes for nonlinear PDEs and approximation error analysis, arXiv preprint, arXiv:2006.01496, 2020.
[25] D. A. Gomes, S. Patrizi and V. Voskanyan, On the existence of classical solutions for stationary extended mean field games, *Nonlinear Anal.*, **99** (2014), 49–79.
[26] D. A. Gomes and V. K. Voskanyan, Extended deterministic mean-field games, *SIAM J. Control Optim.*, **54** (2016), 1030–1055.
[27] A. Gosavi, A reinforcement learning algorithm based on policy iteration for average reward: Empirical results with yield management and convergence analysis, *Machine Learning*, **55** (2004), 5–29.
[28] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, Improved training of wasserstein gans, In *Advances in Neural Information Processing Systems*, (2017), 5767–5777.

[29] X. Guo, A. Hu, R. Xu and J. Zhang, Learning mean-field games, *Advances in Neural Information Processing Systems*, **32** (2019), 4966–4976.
[30] J. Han and W. E, Deep learning approximation for stochastic control problems, arXiv:1611.07422, 2016.
[31] J. Han and R. Hu, Deep fictitious play for finding Markovian Nash equilibrium in multi-agent games, in *Proceedings of The First Mathematical and Scientific Machine Learning Conference (MSML)*, **107** (2020), 221–245.
[32] J. Han, A. Jentzen and W. E, Solving high-dimensional partial differential equations using deep learning, *Proc. Natl. Acad. Sci. USA*, **115** (2018), 8505–8510.
[33] J. Han and J. Long, Convergence of the deep BSDE method for coupled FBSDEs, *Probab. Uncertain. Quant. Risk*, **5** (2020), Paper No. 5, 33 pp.
[34] J. Han, J. Lu and M. Zhou, Solving high-dimensional eigenvalue problems using deep neural networks: A diffusion Monte Carlo like approach, *J. Comput. Phys.*, **423** (2020), 109792, 13 pp.
[35] J. Han, L. Zhang and W. E, Solving many-electron Schrödinger equation using deep neural networks, *J. Comput. Phys.*, **399** (2019), 108929, 8 pp.
[36] J. Hofbauer and W. H. Sandholm, On the global convergence of stochastic fictitious play, *Econometrica*, **70** (2002), 2265–2294.
[37] U. Horst, Stability of linear stochastic difference equations in strategically controlled random environments, *Adv. in Appl. Probab.*, **35** (2003), 961–981.
[38] U. Horst, Stationary equilibria in discounted stochastic games with weakly interacting players, *Games Econom. Behav.*, **51** (2005), 83–108.
[39] R. A. Howard, *Dynamic Programming and Markov Processes*, John Wiley, 1960.
[40] R. Hu, Deep learning for ranking response surfaces with applications to optimal stopping problems, *Quant. Finance*, **20** (2020), 1567–1581.
[41] R. Hu, Deep fictitious play for stochastic differential games, *Commun. Math. Sci.*, **19** (2021), 325–353.
[42] C. Huré, H. Pham and X. Warin, Deep backward schemes for high-dimensional nonlinear PDEs, *Math. Comp.*, **89** (2020), 1547–1579.
[43] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, In *International Conference on Machine Learning*, (2015), 448–456.
[44] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*, John Wiley & Sons, Inc., New York-London-Sydney 1965
[45] S. Ji, S. Peng, Y. Peng and X. Zhang, Three algorithms for solving high-dimensional fully-coupled FBSDEs through deep learning, *IEEE Intelligent Systems*, **35** (2020), 71–84.
[46] D. Kingma and J. Ba, Adam: A method for stochastic optimization, In *Proceedings of the International Conference on Learning Representations*, 2015.
[47] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer-Verlag, Berlin, 1992.
[48] V. Krishna and T. Sjöström, On the convergence of fictitious play, *Math. Oper. Res.*, **23** (1998), 479–511.
[49] H. Liu, H. Qiao, S. Wang and Y. Li, Platform competition in peer-to-peer lending considering risk control ability, *European J. Oper. Res.*, **274** (2019), 280–290.
[50] N. V. Long, Dynamic games in the economics of natural resources: A survey, *Dyn. Games Appl.*, **1** (2011), 115–148.
[51] J. Ma, P. Protter and J. Yong, Solving forward-backward stochastic differential equations explicitly-a four step scheme, *Probab. Theory Related Fields*, **98** (1994), 339–359.
[52] J. Ma and J. Zhang, Representation theorems for backward stochastic differential equations, *Ann. Appl. Probab.*, **12** (2002), 1390–1418.
[53] E. J. McShane, Extension of range of functions, *Bull. Amer. Math. Soc.*, **40** (1934), 837–842.
[54] P. Milgrom and J. Roberts, Adaptive and sophisticated learning in normal form games, *Games Econom. Behav.*, **3** (1991), 82–100.
[55] D. Monderer and L. S. Shapley, Fictitious play property for games with identical interests, *J. Econom. Theory*, **68** (1996), 258–265.
[56] T. Nakamura-Zimmerer, Q. Gong and W. Kang, Adaptive deep learning for high dimensional Hamilton-Jacobi-Bellman equations, *SIAM J. Sci. Comput.*, **43** (2021), A1221–A1247.
[57] É. Pardoux and S. Peng, Backward stochastic differential equations and quasilinear parabolic partial differential equations, in *Stochastic Partial Differential Equations and their Applications*, 200–217. Springer, 1992.

[58] E. Pardoux and S. Tang, Forward-backward stochastic differential equations and quasilinear parabolic PDEs, *Probab. Theory Related Fields*, **114** (1999), 123–150.
[59] P. Pauli, A. Koch, J. Berberich, P. Kohler and F. Allgöwer, Training robust neural networks using Lipschitz bounds, *2021 American Control Conference (ACC)*, (2021), 2595–2600.
[60] D. Pfau, J. S. Spencer, A. G. D. G. Matthews and W. M. C. Foulkes, Ab-initio solution of the many-electron Schrödinger equation with deep neural networks, *Phys. Rev. Research*, **2** (2020), 033429.
[61] H. Pham, X. Warin and M. Germain, Neural networks-based backward scheme for fully nonlinear PDEs, *Partial Differ. Equ. Appl.*, **2** (2021), Paper No. 16, 24 pp.
[62] W. B. Powell and J. Ma, A review of stochastic algorithms with continuous value function approximation and some new approximate policy iteration algorithms for multidimensional continuous applications, *J. Control Theory Appl.*, **9** (2011), 336–352.
[63] A. Prasad and S. P. Sethi, Competitive advertising under uncertainty: A stochastic differential game approach, *J. Optim. Theory Appl.*, **123** (2004), 163–185.
[64] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 1994.
[65] C. Simone, C. Fabio and G. Alessandro, A policy iteration method for mean field games, *ESAIM: Control, Optimisation and Calculus of Variations*, **27** (2021).
[66] J. Sirignano and K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, *J. Comput. Phys.*, **375** (2018), 1339–1364.
[67] Z. Wei and M. Lin, Market mechanisms in online peer-to-peer lending, *Management Science*, **63** (2017), 4236–4257.
[68] Y. Xuan, R. Balkin, J. Han, R. Hu and H. D. Ceniceros, Optimal policies for a pandemic: A stochastic game approach and a deep learning algorithm, *Proceedings of The Second Mathematical and Scientific Machine Learning Conference (MSML)*, **145** (2022), 987-1012.
[69] B. Yu, X. Xing and A. Sudjianto, Deep-learning based numerical BSDE method for barrier options, Available at *SSRN*. arXiv:1904.05921, 2019.
[70] X. Zeng, A stochastic differential reinsurance game, *J. Appl. Probab.*, **47** (2010), 335–349.
[71] J. Zhang, *Backward Stochastic Differential Equations: From Linear to Fully Nonlinear Theory*, Springer, 2017.

Received September 2021; revised March 2022; early access May 2022.

E-mail address: jiequnhan@gmail.com
E-mail address: rhu@ucsb.edu
E-mail address: jihaol@princeton.edu

Appendix A. List of some important notations. General Setting

- $\mathcal{A}^i = \mathbb{R}^{d_\alpha}$: the space of player i's strategy α^i
- $\boldsymbol{\alpha} = (\alpha^1, \dots, \alpha^N)$: a collection of all players' strategies
- $\mathcal{A} = \otimes_{i=1}^N \mathcal{A}^i = \mathbb{R}^{Nd_\alpha}$: the space for the joint control $\boldsymbol{\alpha}$
- $\boldsymbol{\alpha}^{-i} = (\alpha^1, \dots, \alpha^{i-1}, \alpha^{i+1}, \dots, \alpha^N)$: a collection of all players' strategies except player i
- $\boldsymbol{X}_t^{\boldsymbol{\alpha}} \in \mathbb{R}^n$: the common state process controlled by a collection of Markovian strategies $\boldsymbol{\alpha}$, defined in (1)
- $V^i(t, \boldsymbol{x})$: the value function of player i
- $\boldsymbol{V} = [V^1, \dots, V^N]$: the value functions of all players
- $\boldsymbol{p} = [p^1, \dots, p^N] \in \mathbb{R}^{k \times N}$: adjoint variables
- $H^i(t, \boldsymbol{x}, \boldsymbol{\alpha}, p^i) = \phi(t, \boldsymbol{x}, \boldsymbol{\alpha}) \cdot p^i + f^i(t, \boldsymbol{x}, \boldsymbol{\alpha})$: the Hamiltonian function of player i, defined in (5)
- $\boldsymbol{H} = [H^1, \dots, H^N]^\intercal$: the collection of all players' Hamiltonian functions, defined in (5)
- $a^i(t, \boldsymbol{x}, \boldsymbol{\alpha}^{-i}, p^i) = \arg\min_{\alpha^i \in \mathcal{A}^i} H^i(t, \boldsymbol{x}, (\alpha^i, \boldsymbol{\alpha}^{-i}), p^i)$: the minimizer of the i^{th} Hamiltonian, defined in (7)

- $\hat{H}^i(t,\boldsymbol{x},\boldsymbol{\alpha}^{-i},p^i) = H^i(t,\boldsymbol{x},(a^i(t,\boldsymbol{x},\boldsymbol{\alpha}^{-i},p^i),\boldsymbol{\alpha}^{-i}),p^i)$: the optimized Hamiltonian for player i given other players' strategies $\boldsymbol{\alpha}^{-i}$, defined in (16)
- $\hat{H}^i(t,\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{p}) \equiv H^i(t,\boldsymbol{x},(\gamma^i,\boldsymbol{\xi}^{-i}),p^i)$: used in (30), a slightly abuse of notation with (16)
- $\boldsymbol{a} = (a^1,\dots,a^N)$: the collection of all players' Hamiltonian minimizers
- $\boldsymbol{\alpha}(t,\boldsymbol{x},\boldsymbol{p})$: the fixed point of $\boldsymbol{\alpha} = \boldsymbol{a}(t,\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{p})$, defined in (8)
- $\check{\boldsymbol{H}}(t,\boldsymbol{x},\boldsymbol{p}) := \boldsymbol{H}(t,\boldsymbol{x},\boldsymbol{\alpha}(t,\boldsymbol{x},\boldsymbol{p}),\boldsymbol{p})$: the minimized Hamiltonian vector
- $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}(t,\boldsymbol{X}_t,\Sigma^\intercal(t,\boldsymbol{X}_t)\nabla_{\boldsymbol{x}}\boldsymbol{V}(t,\boldsymbol{X}_t))$ or $\boldsymbol{\alpha}_t^* = \boldsymbol{\alpha}^*(t,\boldsymbol{X}_t)$: the Nash equilibrium strategy processes of the N-player game
- $|\cdot|$, $\|\cdot\|_F$ and $\|\cdot\|_S$: the Euclidean norm, Frobenius norm and spectral norm

Algorithm

- $\boldsymbol{\alpha}^m = (\alpha^{1,m},\dots,\alpha^{N,m})$: all players' strategies at stage m
- $(\alpha^i;\boldsymbol{\alpha}^{-i,m})$: a short notation of $(\alpha^{1,m},\dots,\alpha^{i-1,m},\alpha^i,\alpha^{i+1,m},\alpha^{N,m})$
- $V^{i,m+1}$: the problem value of player i at stage $m+1$ given others' strategies $\boldsymbol{\alpha}^m$, defined in (12) for fictitious play (depending only on $\boldsymbol{\alpha}^{-i,m}$) and (13) for policy update
- $\alpha^{i,m+1}(t,\boldsymbol{x}) = \arg\min_{\alpha^i\in\mathcal{A}^i} H^i(t,\boldsymbol{x},(\alpha^i,\boldsymbol{\alpha}^{-i,m})(t,\boldsymbol{x}),\Sigma^\intercal\nabla_{\boldsymbol{x}}V^{i,m+1}(t,\boldsymbol{x}))$: the update rule based on the value function solved from fictitious play or policy update
- π: a partition of size N_T on the time interval $[0,T]$, $0 = t_0 < t_1 < \ldots < t_{N_T} = T$
- $\|\pi\| := \max_{0\le k\le N_T-1}\Delta t_k$: the step size of the time partition where $\Delta t_k = t_{k+1} - t_k$
- $\mathcal{T} = \{t_0,t_1,\dots,t_{N_T-1}\}$: the set of time grids from the partition π
- $\pi(t) = t_k$ for $t\in[t_k,t_{k+1})$: a step function associated with the partition π
- $\boldsymbol{Y}_t^m = [Y_t^{1,m},\dots,Y_t^{N,m}]^\intercal$: the collection of backward processes $Y_t^{i,m}$ at stage m, defined in (15) for fictitious play or (17) for policy update, rewritten in the vector form in (30)
- $\boldsymbol{X}_{t_k}^\pi$: the discretized path of $\boldsymbol{X}$ according to the time partition π, defined in (34)
- $\boldsymbol{Y}_{t_k}^{\pi,m}$: the discretized path of $\boldsymbol{Y}$ at stage m, defined in (34)
- $\boldsymbol{X}_0^\pi = \boldsymbol{X}_{t_0}^\pi, \boldsymbol{X}_T^\pi = \boldsymbol{X}_{t_{N_t}}^\pi$: short notations
- $\mathcal{H}_\eta^\pi$: a Hilbert space on $\mathcal{T}\times\mathbb{R}^n$ with parameter $\eta\ge 0$, defined in (36)
- $\mathcal{N}^\pi$: a closed convex subset of $\mathcal{H}_\eta^\pi$, defined below (36)
- $\mathrm{P}_{\mathcal{N}^\pi,\eta}$: the projection from $\mathcal{H}_\eta^\pi$ to $\mathcal{N}^\pi$
- $\boldsymbol{\alpha}^{\pi,m+1}(t,\boldsymbol{x})$: $(m+1)^{th}$-stage policies defined on $\mathcal{T}\times\mathbb{R}^n$, whose update rules are defined in (38)–(39)
- $\boldsymbol{\alpha}^m(t,\boldsymbol{x}) = \inf_{t'\in\mathcal{T}}[\boldsymbol{\alpha}^{\pi,m}(t',\boldsymbol{x}) + L'|t'-t|^{\frac{1}{2}}]$: the extension of the numerical approximation $\boldsymbol{\alpha}^{\pi,m}(t,x)$ in time, defined in (42)

Appendix B. **Supporting Propositions for Assumption 3.** We prove the following propositions in this section.

Proposition 6. *Under Assumptions 1 and 2, there exists a constant $T_0 > 0$ only depending on L and M, such that Assumption 3 is satisfied when $T\le T_0$.*

Proposition 7. *Under Assumptions 3, the BSDE system (9) has a unique adapted solution satisfying inequality (27).*

Proof of Proposition 6. Fix $M_z > 0$, we use $\mathrm{P}_{M_z}(\boldsymbol{Z})$ to denote the projection from $\mathbb{R}^{k\times N}$ to $\{\boldsymbol{Z}\in\mathbb{R}^{k\times N} : \|\boldsymbol{Z}\|_S^2 \le M_z\}$, and use $\mathrm{P}_{M_z}(\boldsymbol{Z})^i$ for its i^{th} column. Let $\tilde{\boldsymbol{H}} = [\tilde{H}^1,\dots,\tilde{H}^N]^\intercal$ with $\tilde{H}^i(t,\boldsymbol{x},\boldsymbol{p}) = \phi(t,\boldsymbol{x},\boldsymbol{\alpha}(t,\boldsymbol{x},\boldsymbol{p}))\cdot \mathrm{P}_{M_z}(\boldsymbol{p})^i + f^i(t,\boldsymbol{x},\boldsymbol{\alpha}(t,\boldsymbol{x},\boldsymbol{p}))$, then its Lipschitz constants with respect to $\boldsymbol{x}$ and $\boldsymbol{p}$ are computed by

$$
\begin{aligned}
&|\tilde{\boldsymbol{H}}(t,\boldsymbol{x}_1,\boldsymbol{p}_1) - \tilde{\boldsymbol{H}}(t,\boldsymbol{x}_2,\boldsymbol{p}_2)|^2\\
\le& 3M_z|\phi(t,\boldsymbol{x}_1,\boldsymbol{\alpha}(t,\boldsymbol{x}_1,\boldsymbol{p}_1)) - \phi(t,\boldsymbol{x}_2,\boldsymbol{\alpha}(t,\boldsymbol{x}_2,\boldsymbol{p}_2))|^2\\
&+ 3M\|\mathrm{P}_{M_z}(\boldsymbol{p}_1) - \mathrm{P}_{M_z}(\boldsymbol{p}_2)\|_F^2\\
&+ 3|\boldsymbol{f}(t,\boldsymbol{x}_1,\boldsymbol{\alpha}(t,\boldsymbol{x}_1,\boldsymbol{p}_1)) - \boldsymbol{f}(t,\boldsymbol{x}_2,\boldsymbol{\alpha}(t,\boldsymbol{x}_2,\boldsymbol{p}_2))|^2\\
\le& 3(M_zL+L)|\boldsymbol{x}_1-\boldsymbol{x}_2|^2 + 3M\|\boldsymbol{p}_1-\boldsymbol{p}_2\|_F^2\\
&+ 3(M_zL+L)|\boldsymbol{\alpha}(t,\boldsymbol{x}_1,\boldsymbol{p}_1) - \boldsymbol{\alpha}(t,\boldsymbol{x}_2,\boldsymbol{p}_2)|^2\\
\le& 3(M_z+1)(L^2+L)|\boldsymbol{x}_1-\boldsymbol{x}_2|^2 + 3[M+(M_zL+1)L]\|\boldsymbol{p}_1-\boldsymbol{p}_2\|_F^2.
\end{aligned}
$$

Now define $M_z = 2M(L+1)$ and $\overline{M} = 3(2ML+2M+1)(L^2+L)$. Consider the solution $(\boldsymbol{X}_s^{t,\boldsymbol{x}}, \tilde{\boldsymbol{Y}}_s^{t,\boldsymbol{x}}, \tilde{\boldsymbol{Z}}_s^{t,\boldsymbol{x}})$ to the following BSDE system

$$
\begin{cases}
\boldsymbol{X}_s^{t,\boldsymbol{x}} = \boldsymbol{x} + \displaystyle\int_t^s \Sigma(u,\boldsymbol{X}_u^{t,\boldsymbol{x}})\,\mathrm{d}\boldsymbol{W}_u,\\
\tilde{\boldsymbol{Y}}_s^{t,\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{X}_T^{t,\boldsymbol{x}}) + \displaystyle\int_s^T \tilde{\boldsymbol{H}}(u,\boldsymbol{X}_u^{t,\boldsymbol{x}},\tilde{\boldsymbol{Z}}_u^{t,x})\,\mathrm{d}u - \int_s^T (\tilde{\boldsymbol{Z}}_u^{t,\boldsymbol{x}})^\intercal\,\mathrm{d}\boldsymbol{W}_u,
\end{cases}
$$

for any $(t,\boldsymbol{x})\in[0,T]\times\mathbb{R}^n$. Then, for any $t_0\in[0,T]$ and $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbb{R}^n$, we define $\delta\boldsymbol{X}_t$, $\delta\boldsymbol{Y}_t$, $\delta\boldsymbol{Z}_t$, $\delta\boldsymbol{H}_t$ and $\delta\Sigma_t$ as follows:

$$
\begin{aligned}
&\delta\boldsymbol{X}_t = \boldsymbol{X}_t^{t_0,\boldsymbol{x}_1} - \boldsymbol{X}_t^{t_0,\boldsymbol{x}_2}, \quad \delta\boldsymbol{Y}_t = \tilde{\boldsymbol{Y}}_t^{t_0,\boldsymbol{x}_1} - \tilde{\boldsymbol{Y}}_t^{t_0,\boldsymbol{x}_2},\\
&\delta\boldsymbol{Z}_t = \tilde{\boldsymbol{Z}}_t^{t_0,\boldsymbol{x}_1} - \tilde{\boldsymbol{Z}}_t^{t_0,\boldsymbol{x}_2},\\
&\delta\boldsymbol{H}_t = \tilde{\boldsymbol{H}}(t,\boldsymbol{X}_t^{t_0,\boldsymbol{x}_1},\tilde{\boldsymbol{Z}}_t^{t_0,\boldsymbol{x}_1}) - \tilde{\boldsymbol{H}}(t,\boldsymbol{X}_t^{t_0,\boldsymbol{x}_2},\tilde{\boldsymbol{Z}}_t^{t_0,\boldsymbol{x}_2}),\\
&\delta\Sigma_t = \Sigma(t,\boldsymbol{X}_t^{t_0,\boldsymbol{x}_1}) - \Sigma(t,\boldsymbol{X}_t^{t_0,\boldsymbol{x}_2}).
\end{aligned}
$$

Then, we have $\mathrm{d}\delta\boldsymbol{X}_t = \delta\Sigma_t\,\mathrm{d}\boldsymbol{W}_t$, and $\mathrm{d}\delta\boldsymbol{Y}_t = -\delta\boldsymbol{H}_t\,\mathrm{d}t + (\delta\boldsymbol{Z}_t)^\intercal\,\mathrm{d}\boldsymbol{W}_t$. Using Itô's lemma and taking the expectation on both sides yields

$$
\mathbb{E}|\delta\boldsymbol{X}_t|^2 = |\boldsymbol{x}_1-\boldsymbol{x}_2|^2 + \int_{t_0}^t \mathbb{E}\|\delta\Sigma_s\|_F^2\,\mathrm{d}s \le |\boldsymbol{x}_1-\boldsymbol{x}_2|^2 + L\int_{t_0}^t \mathbb{E}|\delta\boldsymbol{X}_s|^2\,\mathrm{d}s,
$$

and by Grönwall's inequality, we have $\mathbb{E}|\delta\boldsymbol{X}_t|^2 \le e^{L(t-t_0)}|\boldsymbol{x}_1-\boldsymbol{x}_2|^2$. Similarly, we deduce that

$$
\begin{aligned}
&\mathbb{E}|\delta\boldsymbol{Y}_t|^2\\
=&\mathbb{E}|\delta\boldsymbol{Y}_T|^2 + \int_t^T \mathbb{E}\left[2\delta\boldsymbol{H}_s\cdot\delta\boldsymbol{Y}_s - \|\delta\boldsymbol{Z}_s\|_F^2\right]\,\mathrm{d}s\\
\le& L\mathbb{E}|\delta\boldsymbol{X}_T|^2 + \int_t^T \left\{\overline{M}\mathbb{E}|\delta\boldsymbol{Y}_s|^2 + (\overline{M})^{-1}\left[\overline{M}\mathbb{E}|\delta\boldsymbol{X}_s|^2 + \overline{M}\mathbb{E}\|\delta\boldsymbol{Z}_s\|_F^2\right] - \mathbb{E}\|\delta\boldsymbol{Z}_s\|_F^2\right\}\,\mathrm{d}s\\
\le& (L+T-t_0)e^{L(T-t_0)}|\boldsymbol{x}_1-\boldsymbol{x}_2|^2 + \overline{M}\int_t^T \mathbb{E}|\delta\boldsymbol{Y}_s|^2\,\mathrm{d}s,
\end{aligned}
$$

and by Grönwall's inequality, we have $|\delta\boldsymbol{Y}_{t_0}|^2 \le (L+T-t_0)e^{(\overline{M}+L)(T-t_0)}|\boldsymbol{x}_1-\boldsymbol{x}_2|^2$. Following the argument in [52, Theorem 3.1], we define $\boldsymbol{u}(t,\boldsymbol{x}) = \tilde{\boldsymbol{Y}}_t^{t,\boldsymbol{x}}$ and deduce $|\boldsymbol{u}(t_0,\boldsymbol{x}_1) - \boldsymbol{u}(t_0,\boldsymbol{x}_2)|^2 \le (L+T)e^{(\overline{M}+L)T}|\boldsymbol{x}_1-\boldsymbol{x}_2|^2$. Therefore, we claim

$$
\|\nabla_{\boldsymbol{x}}\boldsymbol{u}(t,\boldsymbol{x})\|_S^2 \le (L+T)e^{(\overline{M}+L)T} \text{ a.s. with the Lebesgue measure on } \mathbb{R}^n,
$$

for all $t \in [0,T]$. Also noticing that $\tilde{\boldsymbol{Z}}_t^{0,\boldsymbol{x}_0} = (\Sigma^\intercal \nabla_{\boldsymbol{x}} \boldsymbol{u})(t, \boldsymbol{X}_t)$ $\mathbb{P}$-a.s. (*cf.* [52, Theorem 3.1]),

$$\|(\Sigma^\intercal \nabla_{\boldsymbol{x}} \boldsymbol{u})(t,\boldsymbol{x})\|_S^2 \le \|\Sigma(t,\boldsymbol{x})\|_S^2 \|\nabla_{\boldsymbol{x}} \boldsymbol{u}(t,\boldsymbol{x})\|_S^2 \le M(L+T)e^{(\overline{M}+L)T},$$

and the law of $\boldsymbol{X}_t$ is absolute continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, we can get

$$\|\tilde{\boldsymbol{Z}}_t^{0,\boldsymbol{x}_0}\|_S^2 \le M(L+T)e^{(\overline{M}+L)T} = M(L+T)e^{[3(2ML+2M+1)(L^2+L)+L]T} := M(T),$$

$\mathbb{P}$-a.s., $\forall t \in [0,T]$. Therefore, if $M(T) \le 2M(L+1)$, $(\boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_t^{0,\boldsymbol{x}_0}, \tilde{\boldsymbol{Z}}_t^{0,\boldsymbol{x}_0})$ is the desired solution to the BSDE system (9) with $\|\tilde{\boldsymbol{Z}}_t^{0,\boldsymbol{x}_0}\|_S^2 \le 2M(L+1)$, which can be fulfilled if T is small enough. □

Proof of Proposition 7. Let $(\boldsymbol{X}_t, \boldsymbol{Y}_t', \boldsymbol{Z}_t')$ be another adapted solution of the BSDE system (9) satisfying inequality (27). Define $\delta\boldsymbol{Y}_t = \boldsymbol{Y}_t' - \boldsymbol{Y}_t$, $\delta\boldsymbol{Z}_t = \boldsymbol{Z}_t' - \boldsymbol{Z}_t$ and $\delta\boldsymbol{H}_t = \check{\boldsymbol{H}}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t') - \check{\boldsymbol{H}}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t)$. Using Itô's lemma, taking expectation on both side and using $\delta\boldsymbol{Y}_T = 0$, we deduce that

$$\begin{aligned}\mathbb{E}|\delta\boldsymbol{Y}_t|^2 + \int_t^T \mathbb{E}\|\delta\boldsymbol{Z}_s\|_F^2 \,\mathrm{d}s &= 2\int_t^T \mathbb{E}[\delta\boldsymbol{H}_s \cdot \delta\boldsymbol{Y}_s]\,\mathrm{d}s \\ &\le \lambda \int_t^T \mathbb{E}|\delta\boldsymbol{Y}_s|^2\,\mathrm{d}s + \lambda^{-1}\int_t^T \mathbb{E}|\delta\boldsymbol{H}_s|^2\,\mathrm{d}s, \quad (71)\end{aligned}$$

for any $\lambda > 0$. By

$$\begin{aligned}\delta\boldsymbol{H}_t =& \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t')) \cdot \boldsymbol{Z}_t' - \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t) \cdot \boldsymbol{Z}_t \\ &+ \boldsymbol{f}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t')) - \boldsymbol{f}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t)) \\ =& \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t')) \cdot (\boldsymbol{Z}_t' - \boldsymbol{Z}_t) \\ &+ [\phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t')) - \phi(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t)] \cdot \boldsymbol{Z}_t \\ &+ \boldsymbol{f}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t')) - \boldsymbol{f}(t, \boldsymbol{X}_t, \boldsymbol{\alpha}(t, \boldsymbol{X}_t, \boldsymbol{Z}_t)),\end{aligned}$$

we have $|\delta\boldsymbol{H}_t|^2 \le L_z\|\delta\boldsymbol{Z}_t\|_F^2$ with $L_z = 3[M + M'L^2 + L^2]$. Taking $\lambda = L_z$ in (71), we deduce that $\mathbb{E}|\delta\boldsymbol{Y}_t|^2 \le L_z\int_t^T \mathbb{E}|\delta\boldsymbol{Y}_s|^2\,\mathrm{d}s$ and therefore $\boldsymbol{Y}_t' \equiv \boldsymbol{Y}_t$ by Grönwall's inequality. We then have $\int_0^T \mathbb{E}\|\delta\boldsymbol{Z}_t\|_F^2\,\mathrm{d}t = 0$ from the first equality in (71), which implies $\boldsymbol{Z}_t' \equiv \boldsymbol{Z}_t$. □