Proceedings of Machine Learning Research vol 145:987–1012, 2021 2nd Annual Conference on Mathematical and Scientific Machine Learning

Optimal Policies for a Pandemic: A Stochastic Game Approach and a Deep Learning Algorithm

Yao Xuan YXUAN@MATH.UCSB.EDU
Robert Balkin RBALKIN@UCSB.EDU
Department of Mathematics, University of California, Santa Barbara, CA 93106-3080, USA

Jiequn Han JIEQUNH@PRINCETON.EDU
Department of Mathematics, Princeton University, Princeton, NJ 08544-1000, USA

Ruimeng Hu RHU@UCSB.EDU
Department of Mathematics and Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106-3080, USA

Hector D. Ceniceros CENICEROS@UCSB.EDU
Department of Mathematics, University of California, Santa Barbara, CA 93106-3080, USA

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

Abstract

Game theory has been an effective tool in the control of disease spread and in suggesting optimal policies at both individual and area levels. In this paper, we propose a multi-region SEIR model based on stochastic differential game theory, aiming to formulate optimal regional policies for infectious diseases. Specifically, we enhance the standard epidemic SEIR model by taking into account the social and health policies issued by multiple region planners. This enhancement makes the model more realistic and powerful. However, it also introduces a formidable computational challenge due to the high dimensionality of the solution space brought by the presence of multiple regions. This significant numerical difficulty of the model structure motivates us to generalize the deep fictitious algorithm introduced in [Han and Hu, MSML2020, pp.221–245, PMLR, 2020] and develop an improved algorithm to overcome the curse of dimensionality. We apply the proposed model and algorithm to study the COVID-19 pandemic in three states: New York, New Jersey and Pennsylvania. The model parameters are estimated from real data posted by the Centers for Disease Control and Prevention (CDC). We are able to show the effects of the lockdown/travel ban policy on the spread of COVID-19 for each state and how their policies affect each other.

Keywords: Stochastic differential game, pandemic, optimal policy, enhanced deep fictitious play

1. Introduction

The pandemic of coronavirus disease 2019 (COVID-19) has brought a huge impact on our lives. Based on the CDC Data Tracker, as of early December 2020, there have been more than 15 million confirmed cases of infection and more than 290 thousand cases of death in the United States. Needless to say, the economic impact has also been catastrophic, resulting in unprecedented unemployment and the bankruptcy of many restaurants, recreation centers, shopping malls, etc.

In a classic, compartmental epidemiological model each individual is assigned a label, *e.g.*, **S**usceptible, **E**xposed, **I**nfectious, **R**emoved, **V**accinated. The labels' order shows the flow patterns between the compartments (SIR, SEIR, SIRV models). Other approaches include network

© 2021 Y. Xuan, R. Balkin, J. Han, R. Hu & H.D. Ceniceros.

models, which explicitly include the interaction of individuals, in addition to the modeling of each individual's dynamics, and agent-based models that are useful in informing decision making when accurately calibrated. Moreover, the consideration of pharmaceutical and/or non-pharmaceutical intervention policies naturally couples game theory to epidemiological models by controlling when and how the game is played in such models. For example, in some early studies Bauch et al. (2003); Bauch and Earn (2004), one can use non-repeated games to incorporate game theory into modeling at the individual level, where individuals (known as "players" in the game theory) maximize their gain by weighing the costs and benefits of different strategies. We refer to the review paper Chang et al. (2020) and the references therein for more details.

Differential games, initiated by Isaacs (1965), as an offspring of game theory and optimal control, provide modeling and analysis of conflict in the context of a dynamical system. They have been intensively employed across many disciplines, including management science, economics, social science, biology, military, etc. One of the core objectives in differential games is to compute Nash equilibria that refer to strategies by which no player has an incentive to deviate. However, a major bottleneck comes from the notorious intractability of N-player games, and the direct computation of Nash equilibria is extremely time-consuming and memory demanding. In a series of recent works by Hu (2020); Han and Hu (2020); Han et al. (2020a,b), the deep fictitious play (DFP) theory and algorithms were developed for stochastic differential games (SDG) with a large number of heterogeneous players. The DFP framework embeds the fictitious play idea, introduced by Brown (1949, 1951), into designed architectures of deep neural networks to produce accurate and parallelizable algorithms with convergence analysis, and resolve the intractability issue (curse of dimensionality) caused by the complex modeling and underlying high-dimensional space in SDG.

Building from the DFP theory and algorithms for computing Nash equilibria in SDG, we propose here to strengthen the classical SEIR model by taking into account the social and health policies issued by multiple region planners. We call this new model a stochastic multi-region SEIR model because it couples the stochastic differential game theory with the SEIR model, making it more realistic and powerful. The computational challenge introduced by the high-dimensionality of the multi-region solution space is addressed by generalizing the deep fictitious algorithm proposed by Han and Hu (2020). This new approach leads to an enhanced deep fictitious play algorithm to overcome the curse of dimensionality and further reduce the computational complexity. To showcase the performance of the proposed model and algorithm, we apply them to a case study of the COVID-19 pandemic in three states: New York (NY), New Jersey (NJ), and Pennsylvania (PA). We present the optimal lockdown policy corresponding to the Nash equilibrium of the multi-region SEIR model. We remark that our work is not to predict a pandemic, but to provide a game-themed framework, a deep learning algorithm and possible outcomes for competitive region planners. We hope that information can provide some qualitative guidance for policymakers on the impact of certain policies. The parameters used in the numerical experiments are based on the current knowledge of the Coronavirus, which is still under development. In practice, at the beginning of the Coronavirus, a governor might not be able to trace the infections fully, and infected cases may not be fully identified. All these may lead to the inaccuracy of parameter estimations despite our using of the best available data. Therefore, it may be hard to match the predicted results with practical observation.

The rest of the paper is organized as follows. In Section 2, we propose a novel multi-region epidemiological model and explain how social and health policies issued by region planners are integrated into dynamics, leading to a game feature of the problem. We explain the numerical challenge in Section 3 and propose an enhanced deep fictitious play algorithm for memory and

computational efficiency. Section 4 focuses on a NY-NJ-PA case study with detailed discussions on parameter choices and the resulted optimal policies. We present some concluding remarks in Section 5 and provide technical details in the appendices.

2. Mathematical modeling: A multi-region SEIR model

We consider a pandemic spreading in N geographical regions, and each planner controls the loss of her region by implementing some policies. We aim to study how the region planners' policies affect each other, and the equilibrium policies.

Let us start with a modified version of the very-known epidemic SEIR model (cf. Liu et al. (1987)), where each region's population is assigned to compartments with four labels: **S**usceptible, **E**xposed, **I**nfectious, and **R**emoved. Individuals with different labels denote S: those who are not yet infected; E: who have been infected but are not yet infectious themselves; I: who have been infected and are capable of spreading the disease to those in the susceptible category, and R: who have been infected and then removed from the disease due to recovery or death. The region planners can issue certain policies to mitigate the pandemic, for instance, policies that can help reduce the transmission rates and death rates. Mathematically, denote by $S_t^n, E_t^n, I_t^n, R_t^n$ the *proportion* of population in the four compartments of region n at time t. We consider the following stochastic multi-region SEIR model:

$$\mathrm{d}S_t^n = -\sum_{k=1}^{N} \beta^{nk} S_t^n I_t^k (1-\theta\ell_t^n)(1-\theta\ell_t^k)\,\mathrm{d}t - v(h_t^n) S_t^n\,\mathrm{d}t - \sigma_{s_n} S_t^n\,\mathrm{d}W_t^{s_n}, \tag{1}$$

$$\mathrm{d}E_t^n = \sum_{k=1}^{N} \beta^{nk} S_t^n I_t^k (1-\theta\ell_t^n)(1-\theta\ell_t^k)\,\mathrm{d}t - \gamma E_t^n\,\mathrm{d}t + \sigma_{s_n} S_t^n\,\mathrm{d}W_t^{s_n} - \sigma_{e_n} E_t^n\,\mathrm{d}W_t^{e_n}, \tag{2}$$

$$\mathrm{d}I_t^n = (\gamma E_t^n - \lambda(h_t^n) I_t^n)\,\mathrm{d}t + \sigma_{e_n} E_t^n\,\mathrm{d}W_t^{e_n}, \tag{3}$$

$$\mathrm{d}R_t^n = \lambda(h_t^n) I_t^n\,\mathrm{d}t + v(h_t^n) S_t^n\,\mathrm{d}t, \quad n \in \mathcal{N} := \{1,2,\dots,N\}, \tag{4}$$

where $\boldsymbol{\ell}_t \equiv (\ell_t^1,\dots,\ell_t^N)$ and $\boldsymbol{h}_t \equiv (h_t^1,\dots,h_t^N)$ are policies chosen by the region planners at time t. Each planner n seeks to minimize its region's cost within a period $[0,T]$:

$$J^n(\boldsymbol{\ell},\boldsymbol{h}) := \mathbb{E}\bigg[\int_0^T e^{-rt} P^n\big[(S_t^n + E_t^n + I_t^n)\ell_t^n w + a(\kappa I_t^n \chi + p I_t^n c)\big] + e^{-rt}\eta(h_t^n)^2\,\mathrm{d}t\bigg]. \tag{5}$$

We now give detailed description of this model (1)–(5):

S: β^{nk} denotes the average number of contacts per person per time. The transition rate between S^n and E^n due to contacting infectious people in the region k is proportional to the fraction of those contacts between an infectious and a susceptible individual, which result in the susceptible one becoming infected, *i.e.*, $\beta S_t^n I_t^k$. Although some regions may not be geographically connected, the transmission between the two is still possible due to air travels but is less intensive than the transmission within the region, *i.e.*, $\beta^{nk} > 0$ and $\beta^{nn} \gg \beta^{nk}$ for all $k \neq n$.

$\ell_t^n \in [0,1]$ denotes the decision of the planner n on the fraction of population being locked down at time t. We assume that those in lockdown cannot be infected. However, the policy may only be partially effective as essential activities (food production and distribution, health,

and basic services) have to continue. Here we use $\theta \in [0,1]$ to measure this effectiveness, and the transition rate under the policy $\boldsymbol{\ell}$ thus become $\beta^{nk} S_t^n I_t^k (1-\theta\ell_t^n)(1-\theta\ell_t^k)$. The case $\theta = 1$ means the policy is fully effective.

$h_t^n \in [0,1]$ denotes the effort the planner n decide to put into the health system, which we refer as *health policy*. It will influence the vaccination availability $v(\cdot)$ and the recovery rate $\lambda(\cdot)$ of this model.

$v(h_t^n)$ denotes the vaccination availability of region n at time t. Once vaccinated, the susceptible individuals $v(h_t^n)S_t^n$ become immune to the disease, and join the removed category R_t^n. We model it as an increasing function of h_t^n, and if the vaccine has not been developed yet, we can define $v(x) = 0$ for $x \leq \overline{h}$.

E: γ describes the latent period when the person has been infected but not infectious yet. It is the inverse of the average latent time, and we assume γ to be identical across all regions. The transition between E^n and I^n is proportional to the fraction of exposed, *i.e.*, γE_t^n.

I: $\lambda(\cdot)$ represents the recovery rate. For the infected individuals, a fraction $\lambda(h^n)I^n$ (including both death and recovery from the infection) joins the removed category R^n per time unit. The rate is determined by the average duration of infection D. We model the duration (so does the recovery rate) related to the health policy h_t^n decided by its planner. The more effort put into the region (*i.e.*, expanding hospital capacity, creating more drive-thru testing sites), the more clinical resources the region will have and the more resources will be accessible by patients, which could accelerate the recovery and slow down death. The death rate, denoted by $\kappa(\cdot)$, is crucial for computing the cost of the region n; see the next item.

Cost: Each region planner faces four types of cost. One is the economic activity loss due to the lockdown policy, where w is the productivity rate per individual, and P^n is the population of the region n. The second one is due to the death of infected individuals. Here κ is the death rate which we assume for simplicity to be constant, and χ denotes the cost of each death. The hyperparameter a describes how planners weigh deaths and infections comparing to other costs. The third one is the inpatient cost, where p is the hospitalization rate, and c is the cost per inpatient day. The last term $\eta(h_t^n)^2$ is the grants putting into the health system. We choose a quadratic form to account for diminishing marginal utility (view it from $\eta(h_t^n)^2$ to h_t^n). All costs are discounted by an exponential function e^{-rt}, where r is the risk-free interest rate, to take into account the time preference. Note that region n's cost depends on all regions' policies $(\boldsymbol{\ell}, \boldsymbol{h})$, as $\{I^k, k \neq n\}$ appearing in the dynamics of S^n. Thus we write is $J^n(\boldsymbol{\ell}, \boldsymbol{h})$.

The choices of epidemiological parameters will be discussed in Section 4.1. Next, we summarize the key assumptions in the above model:

1. The dynamics of an epidemic are much faster than the vital (birth and death) dynamics. So vital dynamics are omitted in the above model.

2. The planning is of a short horizon and will be adjusted frequently as the epidemic develop. For simplicity, we assume this is no migration between regions over the time $[0, T]$.

3. Individuals who once recovered from the disease, are immune and free of lockdown policy.

4. The dynamics obeys the conservation law: $S_t^n + E_t^n + I_t^n + R_t^n = P^n$. This means that the process R^n is redundant.

5. The dynamics of S, E and I are subjected to random noise, to account for the noise introduced during data recording, false-positive/negative test results, exceptional cases when recovered individuals become susceptible again, minor individual differences in the latent period, etc.

6. Individuals who are not under lockdown have the same productivity, no matter their categories. We assume this for simplicity remark that this can be improved by assigning different productivity to individuals with or without symptoms.

The above modeling and objectives can be viewed as a stochastic differential game between N players[1]. Here, we view the whole problem as a non-cooperative game, as many regions make decisions individually and indeed even compete for scarce resources (frontline workers, personal protective equipment, etc.) during the outbreak. Each player n controls her states (S^n, E^n, I^n, R^n) through her strategy (ℓ^n, h^n) in order to minimize the associated cost J^n. The optimizers then are interpreted as the optimal lockdown policy and optimal effort putting into the health system.

For a non-cooperative game, one usually refers to Nash equilibrium as a notion of optimality. For completeness, we review the definition here.

Definition 2.1 *A Nash equilibrium is a tuple* $(\boldsymbol{\ell}^*, \boldsymbol{h}^*) = (\ell^{1,*}, h^{1,*}, \ldots, \ell^{N,*}, h^{N,*}) \in \mathbb{A}^N$ *such that*

$$\forall n \in \mathcal{N}, \text{ and } (\ell^n, h^n) \in \mathbb{A}, \quad J^n(\boldsymbol{\ell}^*, \boldsymbol{h}^*) \leq J^n((\boldsymbol{\ell}^{-n,*}, \ell^n), (\boldsymbol{h}^{-n,*}, h^n)),$$

where $\boldsymbol{\ell}^{-n,*}$ *represents strategies of players other than the* n*-th one:*

$$\boldsymbol{\ell}^{-n,*} := [\ell^{1,*}, \ldots, \ell^{n-1,*}, \ell^{n+1,*}, \ldots, \ell^{N,*}],$$

$$\boldsymbol{h}^{-n,*} := [h^{1,*}, \ldots, h^{n-1,*}, h^{n+1,*}, \ldots, h^{N,*}],$$

$\mathbb{A}$ *denotes the set of admissible strategies for each player and* $\mathbb{A}^N$ *is the produce of* N *copies of* $\mathbb{A}$*. For simplicity, we have assumed all players taking actions in the same space.*

In the sequel, to fix the notations, we shall use

- a regular character with a superscript n for an object from player n;
- a boldface character for a collection of objects from all players, *i.e.*, $\boldsymbol{S}_t \equiv [S_t^1, \ldots, S_t^N]^{\mathrm{T}}$;
- a boldface character with a superscript $-n$ for a collection of objects from all players except n, *i.e.*, $\boldsymbol{S}_t^{-n} \equiv [S_t^1, \ldots, S_t^{n-1}, S_t^{n+1}, \ldots, S_t^N]^{\mathrm{T}}$.

A Markovian Nash equilibrium is a Nash equilibrium defined above with $\mathbb{A}$ being the set of Borel measurable functions: $(\ell, h) : [0, T] \times \mathbb{R}^{3N} \to [0, 1]^2$. In other words, the policies (ℓ_t^n, h_t^n) at time t are functions of the time t and the current values of all players' state processes $(\boldsymbol{S}_t, \boldsymbol{E}_t, \boldsymbol{I}_t)$. We omit the dependence on $\boldsymbol{R}_t$ as it is redundant as a consequence of the conservation law.

We derive below the Hamilton-Jacobi-Bellman (HJB) equations characterizing the Markovian Nash equilibrium. To simplify the notation, we first rewrite the dynamics of $(\boldsymbol{S}_t, \boldsymbol{E}_t, \boldsymbol{I}_t)$ defined

1. Henceforth, we shall use *planner* and *player* interchangeably.

in (1)-(2)-(3) into a vector form $\boldsymbol{X}_t \equiv [\boldsymbol{S}_t, \boldsymbol{E}_t, \boldsymbol{I}_t]^{\mathrm{T}} \equiv [S_t^1, \cdots, S_t^N, E_t^1, \cdots, E_t^N, I_t^1, \cdots, I_t^N]^{\mathrm{T}} \in \mathbb{R}^{3N}$. Again, we shall drop the redundant process $\boldsymbol{R}_t$ defined in (4). The dynamics of $\boldsymbol{X}_t$ reads:

$$\mathrm{d}\boldsymbol{X}_t = b(t, \boldsymbol{X}_t, \boldsymbol{\ell}(t, \boldsymbol{X}_t), \boldsymbol{h}(t, \boldsymbol{X}_t))\,\mathrm{d}t + \Sigma(\boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{W}_t,$$

where b, Σ are deterministic functions in $\mathbb{R}^{3N}$ and $\mathbb{R}^{3N\times 2N}$, and $\{\boldsymbol{W}_t\}_{0\le t\le T}$ is a $2N$-dimensional standard Brownian motion. Each player n aims to minimize the expected running cost

$$\mathbb{E}\left[\int_0^T f^n(t, \boldsymbol{X}_t, \ell^n(t, \boldsymbol{X}_t), h^n(t, \boldsymbol{X}_t))\,\mathrm{d}t\right].$$

We defer the specific definitions of $b, \Sigma, \boldsymbol{W}$, and f^n to Appendix A.1 to facilitate the exposition. We now define the value function of player n by

$$V^n(t, \boldsymbol{x}) = \inf_{(\ell^n, h^n)\in\mathbb{A}} \mathbb{E}\left[\int_t^T f^n(s, \boldsymbol{X}_s, \ell^n(s, \boldsymbol{X}_s), h^n(s, \boldsymbol{X}_s))\,\mathrm{d}s | \boldsymbol{X}_t = \boldsymbol{x}\right].$$

By dynamic programming, it solves the following HJB system

$$\begin{cases} \partial_t V^n + \inf\limits_{(\ell^n, h^n)\in[0,1]^2} H^n(t, \boldsymbol{x}, (\boldsymbol{\ell}, \boldsymbol{h})(t, \boldsymbol{x}), \nabla_{\boldsymbol{x}} V^n) + \dfrac{1}{2}\mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}}\mathrm{Hess}_{\boldsymbol{x}} V^n \Sigma(\boldsymbol{x})) = 0, \\ V^n(T, \boldsymbol{x}) = 0, \quad n \in \mathcal{N}, \end{cases} \tag{6}$$

where H^n is the usual Hamiltonian defined by

$$H^n(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h}, \boldsymbol{p}) = b(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h}) \cdot \boldsymbol{p} + f^n(t, \boldsymbol{x}, \ell^n, h^n), \tag{7}$$

∂_t denotes the time derivative, $\nabla_{\boldsymbol{x}} V$ and $\mathrm{Hess}_{\boldsymbol{x}} V$ denote the gradient and the Hessian of the function V with respect to $\boldsymbol{x}$, respectively, and Tr stands for the trace of a matrix.

Finding the optimal policies for N regions is equivalent to solving N-coupled $3N+1$ dimensional nonlinear equations (6). For example, when $N = 3$, each PDE is 10-dimensional and conventional methods start to lose their efficiency. The recently proposed deep learning algorithm in Han and Hu (2020), known as deep fictitious play (DFP), has shown excellent numerical performance in solving high-dimensional, coupled HJB equations with convergence analysis (Han et al., 2020a). In the next section, we will first review and then propose an enhanced version of DFP to tackle some new issues.

3. Numerical methodology: Enhanced deep fictitious algorithm

We first briefly review the deep fictitious play (DFP) algorithm for solving equation (6) and we refer readers to Han and Hu (2020) for full details. With the idea of fictitious play, DFP recasts the N-player game into N decoupled optimization problems, which are solved repeatedly stage by stage. Each individual problem is solved by the deep BSDE method (E et al., 2017; Han et al., 2018). The algorithm starts with some initial guess $(\boldsymbol{\ell}^0, \boldsymbol{h}^0)$, where the superscript 0 stands for stage 0. At the $(m+1)^{th}$ stage, given the optimal policies $(\boldsymbol{\ell}^m, \boldsymbol{h}^m)$ at the previous stage, the algorithm solves the following PDEs

$$\begin{cases} \partial_t V^{n,m+1} + \inf\limits_{(\ell^n, h^n)\in[0,1]^2} H^n(t, \boldsymbol{x}, (\ell^n, \boldsymbol{\ell}^{-n,m}, h^n, \boldsymbol{h}^{-n,m})(t, \boldsymbol{x}), \nabla_{\boldsymbol{x}} V^{n,m+1}) \\ \qquad\qquad + \dfrac{1}{2}\mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}}\mathrm{Hess}_{\boldsymbol{x}} V^{n,m+1}\Sigma(\boldsymbol{x})) = 0, \\ V^{n,m+1}(T, \boldsymbol{x}) = 0, \quad n \in \mathcal{N}, \end{cases} \tag{8}$$

and obtains the $(m+1)^{th}$ stage's optimal strategy by:

$$(\ell^{n,m+1}, h^{n,m+1})(t,\boldsymbol{x}) = \underset{(\ell^n,h^n)\in[0,1]^2}{\arg\min}\ H^n(t,\boldsymbol{x},(\ell^n,\boldsymbol{\ell}^{-n,m},h^n,\boldsymbol{h}^{-n,m})(t,\boldsymbol{x}),\nabla_{\boldsymbol{x}}V^{n,m+1}(t,\boldsymbol{x})). \tag{9}$$

Here, $(\boldsymbol{\ell}^{-n,m},\boldsymbol{h}^{-n,m})$ stands for others' optimal policies from the m^{th} stage and are considered to be fixed functions when solving the PDE at the current stage. In the sequel, to simplify notations we omit the stage label m in the superscript when there is no risk of confusion. To solve (8) at each stage, it is first rewritten in the DFP as

$$\begin{aligned}\partial_t V^n + \frac{1}{2}\mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}}\mathrm{Hess}_{\boldsymbol{x}}V^n\Sigma(\boldsymbol{x})) + \mu^n(t,\boldsymbol{x};\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n})\cdot\nabla_{\boldsymbol{x}}V^n \\ + g^n(t,\boldsymbol{x},\Sigma(\boldsymbol{x})^{\mathrm{T}}\nabla_{\boldsymbol{x}}V^n;\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n}) = 0,\end{aligned} \tag{10}$$

with some functions μ^n and g^n. The solution is then approximated by solving the equivalent BSDE $(\boldsymbol{X}_t^n, Y_t^n, Z_t^n)\in\mathbb{R}^{3N}\times\mathbb{R}\times\mathbb{R}^{2N}$:

$$\begin{cases}\boldsymbol{X}_t^n = \boldsymbol{x}_0 + \displaystyle\int_0^t \mu^n(s,\boldsymbol{X}_s^n;(\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n})(s,\boldsymbol{X}_s^n))\,\mathrm{d}s + \int_0^t \Sigma(\boldsymbol{X}_s^n)\,\mathrm{d}\boldsymbol{W}_s, & (11)\\ Y_t^n = \displaystyle\int_t^T g^n(s,\boldsymbol{X}_s^n,Z_s^n;(\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n})(s,\boldsymbol{X}_s^n))\,\mathrm{d}s - \int_t^T (Z_s^n)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{W}_s, & (12)\end{cases}$$

in the sense of (*cf.* Pardoux and Peng (1992); El Karoui et al. (1997); Pardoux and Tang (1999))

$$Y_t^n = V^n(t,\boldsymbol{X}_t^n)\quad\text{and}\quad Z_t^n = \Sigma(\boldsymbol{X}_t^n)^{\mathrm{T}}\nabla_{\boldsymbol{x}}V^n(t,\boldsymbol{X}_t^n).$$

The high-dimensional BSDE (11)–(12) is tackled by the deep BSDE method proposed in E et al. (2017); Han et al. (2018).

In Han and Hu (2020), the algorithm solves the BSDE by parametrizing $V^n(t,\boldsymbol{x})$ using neural networks (NN) and then obtains the approximate optimal policy by plugging the NN outputs into (9). For memory efficiency, the algorithm only stores the NNs' parameters at the current and the one-step previous stages. This strategy works well for games like the linear-quadratic game, but it would be ineffective if $\boldsymbol{\ell}^{-n}$ or $\boldsymbol{h}^{-n}$ explicitly appears in the minimizer in (9). In this case, when evaluating others' strategy $\boldsymbol{\ell}^{-n}$ or $\boldsymbol{h}^{-n}$ at stage m, it does not only need NNs at stage m but also at stages $m-1$, $m-2$, ..., 0. This means one needs to store NNs' parameters for all the previous stages from $1,\dots,m$, and evaluate the associated output. Therefore, the time complexity of evaluating $(\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n})(s,\boldsymbol{X}_s^n)$ up to stage m is $\mathcal{O}(m^2)$ and the memory complexity is $\mathcal{O}(m)$. This is infeasible in practice, as for real problems it hundreds of stages are needed. To overcome this significant problem, we propose an enhanced version of the original algorithm which reduces the time complexity to $\mathcal{O}(m)$ and the memory complexity to $\mathcal{O}(1)$. We present this new, enhanced algorithm (with pseudocode).

3.1. Algorithm

In order to reduce the computational complexity of evaluating $(\boldsymbol{\ell}^{-n},\boldsymbol{h}^{-n})(s,\boldsymbol{X}_s^n)$ in the situation when $\boldsymbol{\ell}^{-n}$ or $\boldsymbol{h}^{-n}$ explicitly appears in the minimizer in (9), we propose the *Enhanced Deep Fictitious Play* which parametrizes both $V^n(t,\boldsymbol{x})$ and policy $(\ell^n,h^n)(t,\boldsymbol{x})$ by NNs. For simplicity, we state the algorithm based on a generic stochastic differential game, where (possibly high-dimensional) controls are denoted by $\alpha^n(t,\boldsymbol{x})$ for player n.

In each stage of the *Enhanced Deep Fictitious Play*, for each planner n, the loss that our algorithm aims to minimize consists of two parts: (1) the loss related to solving (11)–(12) and (2) the error of approximating the optimal strategy α^n within some hypothesis spaces. The resulted approximation $\tilde{\alpha}^n$ will be used in the next stage of fictitious play:

$$\begin{aligned}
&\inf_{Y_0^n,\tilde{\alpha}^n,\{Z_t^n\}_{0\le t\le T}} \mathbb{E}(|Y_T^n|^2 + \tau\int_0^T \|\alpha^n(s,\boldsymbol{X}_s^n) - \tilde{\alpha}^n(s,\boldsymbol{X}_s^n)\|_2^2\,\mathrm{d}s)\\
\text{s.t. } &\boldsymbol{X}_t^n = \boldsymbol{x}_0 + \int_0^t \mu^n(s,\boldsymbol{X}_s^n;\tilde{\boldsymbol{\alpha}}^{-n}(s,\boldsymbol{X}_s^n))\,\mathrm{d}s + \int_0^t \Sigma(\boldsymbol{X}_s^n)\,\mathrm{d}\boldsymbol{W}_s,\\
&Y_t^n = Y_0^n - \int_0^t g^n(s,\boldsymbol{X}_s^n,Z_s^n;\tilde{\boldsymbol{\alpha}}^{-n}(s,\boldsymbol{X}_s^n))\,\mathrm{d}s + \int_0^t (Z_s^n)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{W}_s,\\
&\alpha^n(s,\boldsymbol{X}_s^n) = \arg\min_{\beta^n} H^n(s,\boldsymbol{X}_s^n,(\beta^n,\tilde{\boldsymbol{\alpha}}^{-n})(s,\boldsymbol{X}_s^n),Z_s^n),\,^{2}
\end{aligned} \tag{13}$$

where $\|\cdot\|_2$ denotes the 2-norm, $\tilde{\boldsymbol{\alpha}}^{-n}$ denotes the collection of approximated optimal controls from the previous stage except player n, and τ is a hyperparameter denoting the weight between two terms in the loss function. As detailed in Section 3.2, the hypothesis space for which we search $\tilde{\alpha}^n$ is characterized by another NN, in addition to the one to approximate Y_0 and $\{Z_t^n\}_{0\le t\le T}$. Although representing $\tilde{\alpha}^n$ with a neural network introduces approximation errors, it allows us to efficiently access the proxy of the optimal strategy $\boldsymbol{\alpha}^{-n}$ in the last stage by calling corresponding networks, instead of storing and calling all the previous strategies $\boldsymbol{\alpha}^{-n,m-1},\ldots,\boldsymbol{\alpha}^{-n,1}$ due to the recursive dependence.

Numerically we solve a discretized version of (13). Given a partition π of size N_T on the time interval $[0,T]$, $0 <= t_0 < t_1 < .... < t_{N_T} = T$, the algorithm reads (to ease the notation, we replace the subscript t_k by k):

$$\inf_{\psi_0\in\mathcal{N}_0^{n'},\{\phi_k\in\mathcal{N}_k^n,\xi_k\in\mathcal{N}_k^{n''}\}_{k=0}^{N_T-1}} \mathbb{E}\{|Y_T^{n,\pi}|^2 + \tau\sum_k \left\|\alpha_k^{n,\pi} - \tilde{\alpha}_k^{n,\pi}(\boldsymbol{X}_k^{n,\pi})\right\|_2^2 \Delta t_k\} \tag{14}$$

$$\begin{aligned}
\text{s.t. } &\boldsymbol{X}_0^{n,\pi} = \boldsymbol{X}_0,\quad Y_0^{n,\pi} = \psi_0\left(\boldsymbol{X}_0^{n,\pi}\right),\quad Z_k^{n,\pi} = \phi_k\left(\boldsymbol{X}_k^{n,\pi}\right),\quad \tilde{\alpha}_k^{n,\pi}(\boldsymbol{X}_k^{n,\pi}) = \xi_k(\boldsymbol{X}_k^{n,\pi}),\\
&\alpha_k^{n,\pi} = \arg\min_{\beta^n} H^n(t_k,\boldsymbol{X}_k^{n,\pi},(\beta^n,\tilde{\boldsymbol{\alpha}}_k^{-n,\pi})(\boldsymbol{X}_k^{n,\pi}),Z_k^{n,\pi}),\quad k=0,\ldots,N_T-1\\
&\boldsymbol{X}_{k+1}^{n,\pi} = \boldsymbol{X}_k^{n,\pi} + \mu^n\left(t_k,\boldsymbol{X}_k^{n,\pi};\tilde{\boldsymbol{\alpha}}_k^{-n,\pi}(\boldsymbol{X}_k^{n,\pi})\right)\Delta t_k + \Sigma\left(t_k,\boldsymbol{X}_k^{n,\pi}\right)\Delta\boldsymbol{W}_k, \qquad (15)\\
&Y_{k+1}^{n,\pi} = Y_k^{n,\pi} - g^n\left(t_k,\boldsymbol{X}_k^{n,\pi},Z_k^{n,\pi};\tilde{\boldsymbol{\alpha}}_k^{-n,\pi}(\boldsymbol{X}_k^{n,\pi})\right)\Delta t_k + \left(Z_k^{n,\pi}\right)^{\mathrm{T}}\Delta\boldsymbol{W}_k, \qquad (16)
\end{aligned}$$

where $\Delta t_k = t_{k+1} - t_k$, $\Delta\boldsymbol{W}_k = \boldsymbol{W}_{t_{k+1}} - \boldsymbol{W}_{t_k}$, and $\mathcal{N}_0^{n'}$, $\{\mathcal{N}_k^n\}_{k=0}^{N_T-1}$, $\{\mathcal{N}_k^{n''}\}_{k=0}^{N_T-1}$ are hypothesis spaces for player n, which will be specified later through neural network structures.

The expectation in (14) is further approximated by Monte Carlo samples of (15)-(16). The parameters in the hypothesis spaces are determined by stochastic gradient descent (SGD) algorithms such that the approximated expectation is minimized, which in turn gives the optimal deterministic functions $(\psi_0^*,\phi_k^*,\xi_k^*)$. We expect that $(\psi_0^*,\phi_k^*,\xi_k^*)$ will approximate $(V^n,\nabla_{\boldsymbol{x}}V^n,\alpha^n)$ well when this proxy of (14) is small. Particularly, $\{\xi_k^*\}_{k=0}^{N_T-1}$ serves as an efficient tool to evaluate the optimal

2. Here we have assumed that the Hamiltonian H^n depends on $\nabla_{\boldsymbol{x}}V$ through $\Sigma^{\mathrm{T}}\nabla_{\boldsymbol{x}}V$.

policy at the current stage for finding Nash equilibrium. Implementation details and the full algorithm are presented in Section 3.2. Note that when $\tau = 0$ and $\tilde{\alpha}$ are replaced by α in the above algorithm, the *Enhanced Deep Fictitious Play* degenerates to the *Deep Fictitious Play* proposed in Han and Hu (2020).

3.2. Implementation

Here we provide some detail to implement the methodology in Section 3.1. First, we specify the hypothesis spaces for neural networks $\mathcal{N}_0^{n'}$,$\{\mathcal{N}_k^n\}_{k=0}^{N_T-1}$, $\{\mathcal{N}_k^{n''}\}_{k=0}^{N_T-1}$, corresponding to V^n, $\nabla_{\boldsymbol{x}} V^n$, α^n (the superscript m is dropped again for simplicity). $V^n(t,x)$ is parametrized directly by a neural network $\mathrm{NN}(t,\boldsymbol{x})$. Corresponding map $\Sigma(\boldsymbol{X})^{\mathrm{T}}\nabla_{\boldsymbol{x}} V^n(t,\boldsymbol{X})$ that defines Z_t^n in the optimization problem (13) could be parametrized by $\Sigma(\boldsymbol{x})\nabla_{\boldsymbol{x}}\mathrm{NN}(t,\boldsymbol{x})$. Naturally, $\Sigma(\boldsymbol{x})\nabla_{\boldsymbol{x}}\mathrm{NN}(t_k,\boldsymbol{x})$ is a hypothesis function in $\mathcal{N}_k^n$. Under this parametrization rule, the hypothesis functions in $\mathcal{N}_0^{n'}$ and $\{\mathcal{N}_k^n\}_{k=0}^{N_T-1}$ share the same set of parameters. The policy function $\alpha^n(t,\boldsymbol{x})$ is parametrized by another neural network $\widetilde{\mathrm{NN}}(t,\boldsymbol{x})$ and then $\widetilde{\mathrm{NN}}(t_k,\boldsymbol{x})$ plays the role of a hypothesis function in $\mathcal{N}_k^{n''}$. In other words, $\{\mathcal{N}_k^{n''}\}_{k=0}^{N_T-1}$ share the same set of neural networks. In a stochastic game of N players, there are $2N$ neural networks in total, with N neural networks corresponding to $V^n(t,\boldsymbol{x})$ and N neural networks corresponding to $\alpha^n(t,\boldsymbol{x})$. At stage m, the N V-networks are trained to approximate the solution of PDE (10) and the N α-networks are trained to approximate the current optimal policy computed by (9) using the optimal strategies in the last stage. The updated neural networks at stage m would be used at stage $m+1$ to simulate paths $\{X_k^{n,\pi}\}_{k=0}^{N_T-1}$ and optimal strategies by (9). In this work, fully connected neural networks with three hidden layers are used.

Second, at each stage, the $2N$ neural networks could be decoupled to N pairs of V-network and α-network based on players. Then, the N pairs of neural networks could be trained in parallel, which dramatically reduces computational time. As Han and Hu (2020) and Seale and Burnett (2006) pointed out, it is not necessary to solve the individual control problem accurately in each stage; the parameters at each stage are updated starting from the optimal parameters in the last stage without re-initialization. This requires only a moderate number of epochs for the stochastic gradient descent at each stage.

The full implementation of *Enhanced Deep Fictitious Play* is shown in Algorithm 1. For simplicity, we state the algorithm based on a generic stochastic differential game.

Due to page limits, the exact choice of NN architectures will be detailed in Appendix B.1. To determine the total stages of fictitious play M, we monitor the relative changes of α^n and V^n, and stop the process when the relative change from stage to stage is below a threshold. Regarding the total number of SGD per stage, as shown in (Han and Hu, 2020, Figure 1), the original DFP is insensitive to the choice of $N_{\text{SGD\_per\_stage}}$. We find the enhanced version sharing the same behavior when apply to the COVID-19 case study. We give more details in Section 4.2, and further experiments regarding different choices of M and $N_{\text{SGD\_per\_stage}}$ in Appendix B.2.

For problems without analytical solutions, one natural concern is the reliability of numerical solutions. Theoretically, the quantity (14) serves as the indicator of the numerical accuracy. In the original DFP where the second term in (14) does not exist, Theorem 3 in Han et al. (2020b) ensures the convergence to the true Nash equilibrium under technical assumptions when (14) is small enough for each fictitious play stage and with sufficiently large M and small Δt_k. In practice, the quantity in (14) is approximated by its Monte Carlo counterpart, which we define as the loss function of our algorithms. Therefore, having small training losses during all stages will ensure

Algorithm 1 Enhanced Deep Fictitious Play for Finding Markovian Nash Equilibrium

Require: N = # of players, N_T = # of subintervals on $[0,T]$, M = # of total stages in fictitious play, N_{sample} = # of sample paths generated for each player at each stage of fictitious play, $N_{\text{SGD\_per\_stage}}$ = # of SGD steps for each player at each stage, N_{batch} = batch size per SGD update, $\boldsymbol{\alpha}^0$: the initial policies that are smooth enough

Initialize N deep neural networks to represent $V^{n,0}$ and N deep neural networks to represent $\alpha^{n,0}, n \in \mathcal{N}$
for $m \leftarrow 1$ to M **do**
  for all $n \in \mathcal{N}$ **do in parallel**
    Generate N_{sample} sample paths $\{\boldsymbol{X}_k^{n,\pi}\}_{k=0}^{N_T}$ according to (15) and the realized approximate optimal policies $\tilde{\boldsymbol{\alpha}}^{-n,m-1}(t_k, \boldsymbol{X}_k^{n,\pi})$ (Remark: $\tilde{\boldsymbol{\alpha}}$ represents social and health policies in the multi-region SEIR model)
    for $e \leftarrow 1$ to $N_{\text{SGD\_per\_stage}}$ **do**
      Update the parameters of the n^{th} V-neural network and α-neural network one step with N_{batch} paths using the SGD algorithm (or its variant), based on the loss function (14)
    end for
    Obtain the approximate optimal policy $\tilde{\alpha}^{n,m}$ represented by the latest policy neural network
  end for
  Collect the approximate optimal policies at stage m: $\tilde{\boldsymbol{\alpha}}^m \leftarrow (\tilde{\alpha}^{1,m}, \ldots, \tilde{\alpha}^{N,m})$
end for
return The approximate optimal policy $\tilde{\boldsymbol{\alpha}}^M$

convergence. Extending Theorem 3 in Han et al. (2020b) to the current setting is beyond the scope of this paper and is left for further work.

4. Application on COVID-19

Our case study is based on COVID-19. We focus mainly on the lockdown/travel ban policy between different regions. Therefore, to simplify the presentation, we omit the health policy h in the following discussion and make $v(\cdot) = v$, $\lambda(\cdot) = \lambda$, and $\eta = 0$. Moreover, as vaccines are not available to the population yet, we let $v(\cdot) = v = 0$.

4.1. Parameter choices

In single-region SEIR models, the transmission rate, β, is the basic reproductive number divided by the length of time an individual is infectious. In our model, we assume that there is a region-independent constant β that underlies the rate of infections for each population. The transmission rates β^{nk} between regions are related to the underlying transmission rate β, and the amount of travel between regions n and k.

To quantify the size of travel between regions, we assume there is a constant fraction of people from region n that travel to region k, f^{nk}, at any given moment in time. We note that realistically one may expect f^{nk} to depend on time and also on the epidemic status of regions n and k. However, for simplicity, we will not consider these scenarios in our numerical experiments. We assume that $f^{nn} \gg f^{nk}$, $f^{nn} \gg f^{kn}$, $\forall k \neq n$, meaning that most of the population n resides in region n at any

given time, and also that most of the people in region n at a given time are from n and not travelers from another region. We will see later that this implies $\beta^{nn} \gg \beta^{nk}\ \forall k \neq n$.

To further clarify the transmission of infection from one region due to another, we need to describe the set of parameters $\{\beta^{nk} : n, k \in \mathcal{N}\}$. Here, β^{nk} represents the rate of transmission from region k to region n. Specifically, β^{nk} is the number of infected people in population n per a contactable, infectious individual in population k per day, assuming that 100% of population n is susceptible. This definition of β^{nk} comes from the derivation of the SDE system itself (1)–(3). Accounting for infection in region n by individuals in region k from travel to both region n and k, we have that

$$\beta^{nk} = \begin{cases} \beta(f^{nk}f^{kk} + f^{kn}f^{nn})\dfrac{P^k}{P^n}, & \text{if } k \neq n \\ \beta(f^{nn})^2, & \text{if } k = n. \end{cases} \tag{17}$$

We defer the detailed derivation of (17) to Appendix A.3.

Therefore, to specify β^{nk}, we need to provide β and f^{nk}. We will specify f^{nk} for New York (NY), New Jersey (NJ), and Pennsylvania (PA) in the next section. To estimate β, we choose a basic reproductive number $R_0 = 2.2$, which is consistent with Fauci et al. (2020), and assume that the length of each individual being infectious is 13 days. More precisely, we assume infectious individuals either recover or die in 13 days. Under these assumptions, we obtain $\beta = \frac{2.2}{13} \approx 0.17$, consistent with Linton et al. (2020) as a 13 day median time until death from illness onset is used.

The infection fatality rate, or IFR, is the fraction of those infected who died from the infection. We choose the IFR to be 0.65% according to the CDC estimate. This is also consistent with Meyerowitz-Katz and Merone (2020), which suggests a point estimate of 0.68%. The assumptions of an IFR of 0.65% and an infectious period of 13 days determines that the recovery rate (including both recovery and death due to infection) is $\lambda = \frac{1}{13} \approx 0.0769$, and the death rate is $\kappa = \frac{(0.65\%)}{13} = 0.0005$. We choose the latent period to be 5 days according to Lauer et al. (2020). This means that the we will have $\gamma = \frac{1}{5}$. Note that this choice has also been used in other models such as Alvarez et al. (2020) and Peng et al. (2020). We assume that the parameters for noise-level $\sigma_{s_n}, \sigma_{e_n}, n \in \mathcal{N}$ are all 0.0002, and the extent to which one adheres to the social distancing policy, θ, is either $\theta = 0.9$ or $\theta = 0.99$.

With most of the parameters for the SDE model (1)-(3) discussed, we now address those specific to defining the cost. Regarding the risk-free-rate r, note that U.S. Treasury yields are historically low and the uncertainty in the current level of inflation. We choose for simplicity that $r = 0$. Also, considering that we are interested in simulations with time periods of less than a year, the discounting is negligible. The parameter w represents the dollar output per individual per day. To estimate w, we use GDP per capita per day, yielding the estimate $w = 172.6$ dollars per person per day. Following Alvarez et al. (2020) and Hall et al. (2020), we use the value of a statistical life, χ, to be 20 times GDP per capita. This results in $\chi = 1.95 \cdot 10^6$ dollars per person. According to the CDC summary of U.S. COVID-19 activity, the hospitalization rate was 228.7 per 100,000 population by 11/14/2020. Thus, we set $p = 228.7 \times 10^{-5}$. The cost per inpatient day is $c = 73300/13$ dollars, estimated according to Health (2020). The attention hyperparameter a takes various values in the case study, and will be specified in Section 4.2.

4.2. NY-NJ-PA COVID-19 case study

In this section, we apply our model (1)–(5) to analyze COVID-19 related policy in three adjacent states: New York, New Jersey, and Pennsylvania. This case study is done over 180 days starting from 03/15/2020, using the Enhanced Deep Fictitious Play algorithm introduced in Section 3 (cf. Algorithm 1) and the parameters discussed in 4.1. The exact formulas of μ^n and h^n in equation (10) are derived in Appendix A.2.

We refer to New York State as region 1, New Jersey State as region 2, and Pennsylvania State as region 3. Their respective populations are $P^1 = 19.54$ million, $P^2 = 8.91$ million, and $P^3 = 12.81$ million. Regarding β^{nk}, $\forall n, k = 1, 2, 3$, we assume that: (a) 90% of any state's population is residing in their state at a given time; (b) the remaining population (travelers) visit the other regions in an equal proportion; and (c) there is no travel outside of the considered regions, *i.e.*, the NY-NJ-PA is a closed system. The reasoning for (c) is that, under our model assuming that infection only occurs in the regions considered, (c) is equivalent to allowing people traveling outside the considered regions, but the travelers cannot be affected. For simplicity, we assume this is the case. Under these assumptions, we will have $f^{nn} = 90\%$ for $n = 1, 2, 3$ and $f^{nk} = 5\%$ for $n \neq k$, and obtain the values of β^{nk} through (17).

Figure 1 presents the equilibrium policy issued by the governors of NY, NJ, and PA when the policy effectiveness is $\theta = 0.99$, *i.e.*, 99% of the population follow the lockdown order. The hyper-parameter is $a = 100$, *i.e.*, each governor values people's death 100 times the lockdown cost. In this scenario, the governors take action at an early stage and soon reach the strictest policy. Once the disease is under control, they may relax the policy later. The percentage of Susceptible, Exposed, Infectious, and Removed stays almost constant in the end. As a comparison, Figure 2 illustrates how the pandemic gets out of control if governors show inaction or issue mild lockdown policies.

Experiment 1: dependence on a. We further analyze how the planners' view on the death of human beings changes their policies. In reality, economic loss is not the only factor the planners concern about. It is also important to mitigate the infections and deaths within the budget and available resources. Different views and values from the planners will lead to different policies. In this experiment, we consider different attitudes towards the infection, especially death caused by COVID-19. This is reflected by the attention hyperparameter a. Large a implies that planners care more about human beings and are willing to spend more effort or endure more economic loss on lockdown to avoid further infection and death. In comparison, smaller a implies that planners care less about infection and death and instead pay more attention to minimizing the total cost.

The numerical results in Figures 3 and 4 are consistent with intuition. With a large a (top-left panels), meaning the planners give more consideration to infection and death, they tend to issue a restrict lockdown policy, which helps slow down the disease spreads and reduce the percentage of infected people. As a becomes smaller (top-right panels), planners weigh more the economic loss and spend fewer efforts on lockdown. When the attention a is small enough, some states even give up controlling the disease spread due to economic concern (bottom panels). As a result, the pandemic would get out of control by the end of the simulation period. This mild lockdown policy leads to a natural spread of disease (also shown in Figure 2).

Experiment 2: dependence on θ. We next analyze how the residents' willingness to comply with the lockdown policy changes the optimal policies and the development of a pandemic. The larger the θ is, the more likely the residents will follow the lockdown policy, and the larger the difference the control makes on the pandemic situation. Conversely, small θ weakens the effect of the lockdown

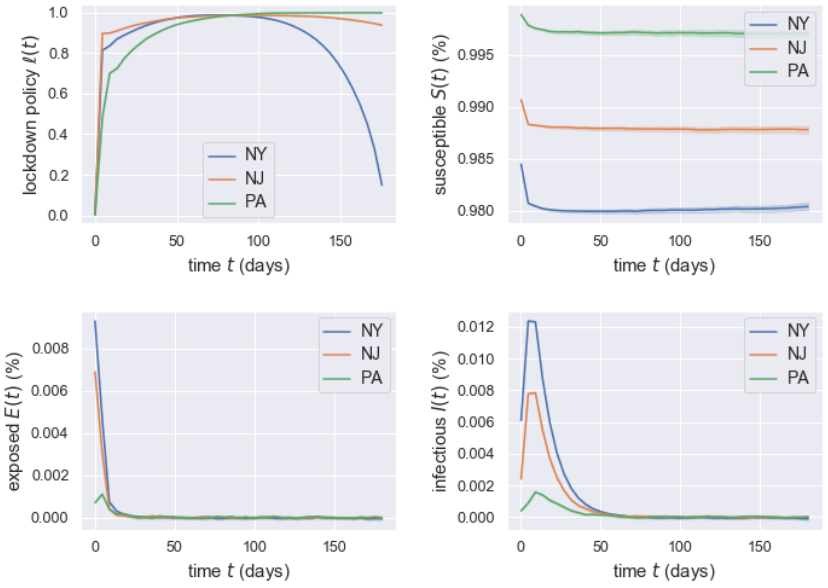

Figure 1: Plots of optimal policies (top-left), Susceptibles (top-right), Exposed (bottom-left) and Infectious (bottom-right) for three states: New York (blue), New Jersey (orange) and Pennsylvania (green). The shaded areas depict the mean and 95% confidence interval over 256 sample paths. Choices of parameters are in Section 4.1, $a = 100$ and $\theta = 0.99$.

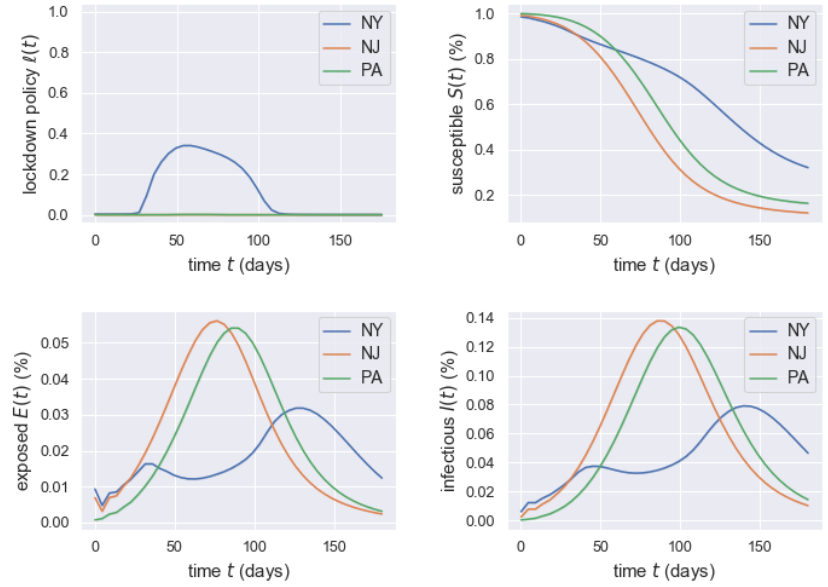

Figure 2: An illustration that governors' inaction or mild control leads to disease spreading.

policy. In the extreme case of $\theta = 0$, no matter how strict the lockdown policy is, the pandemic

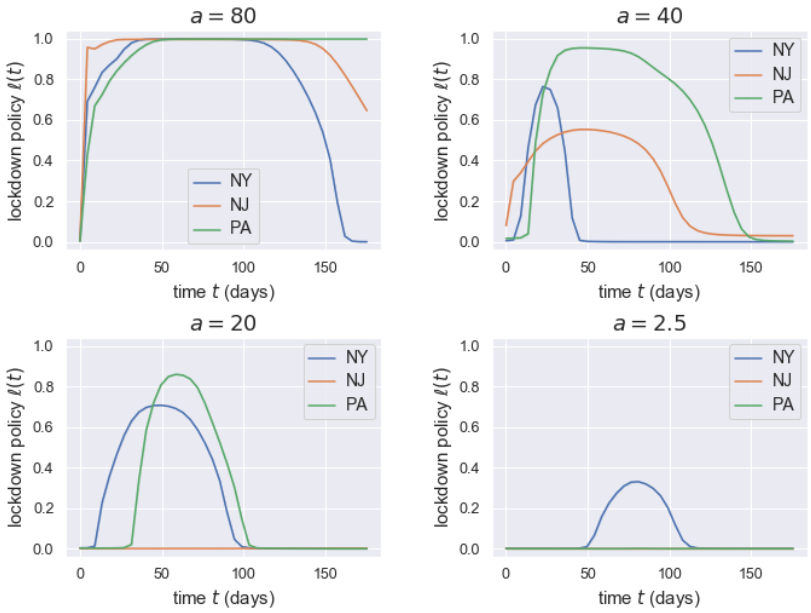

Figure 3: Plots of optimal policies with different choice of a for three states: New York (blue), New Jersey (orange) and Pennsylvania (green), when the lockdown efficiency is $\theta = 0.9$.

will become a natural spread because the control term in (1) disappears. In short, this willingness to policy compliance should be an essential factor in decision-making.

To this end, we compare the optimal policy when $\theta = 0.9$ and $\theta = 0.99$ in Figure 5. Panels (a-d) show the difference of optimal policies $\ell(t)$ and the Susceptible $S(t)$ in the tri-state game under different θ when $a = 50$. In both situations, the pandemic is well-controlled, with the percentage of susceptible people staying stable in the end. Moreover, in the case of $\theta = 0.99$, people are more willing to comply with the policies. Consequently, the planners are allowed to use a less strict lockdown policy as shown in Figure 5(b) compared to 5(a), which saves the lockdown cost. Figure 5 (e-h) shows an interesting case in the comparison of $\theta = 0.9$ and $\theta = 0.99$. In this scenario, with the same attention parameter ($a = 25$), $\theta = 0.9$ leads to a mild lockdown policy, see Figure 5(e), while $\theta = 0.99$ provides a possibility to stop the spread of virus, see Figure 5(f). We believe that the decision when $\theta = 0.9$ is a compromise as the lockdown is not efficient enough to reduce largely the infection and death loss by paying lockdown cost, and also due to the limited simulation period, *i.e.*, the policies could have been different if we had the simulation until the disease dies out. We also believe that the early give-up by NJ drives NY and PA to lift lockdown policies at a later stage, because even NY and PA issue strict policies, they are still facing severe infections from NJ due to its high infected percentages and the existence of travel between states whatever the policy is. So their interventions are not worth the candle. Figure 5(f)(h) further elucidate the importance of residents' support in slowing down the pandemic. Further experiments based on different sets of (a, θ) reveal the possibility of having multiple Nash equilibrium, with more elaboration in Appendix B.3.

To summarize, the numerical experiments illustrate that both the balance of economy and infection/death from the view of plan-makers and the willingness of residents to follow the lockdown

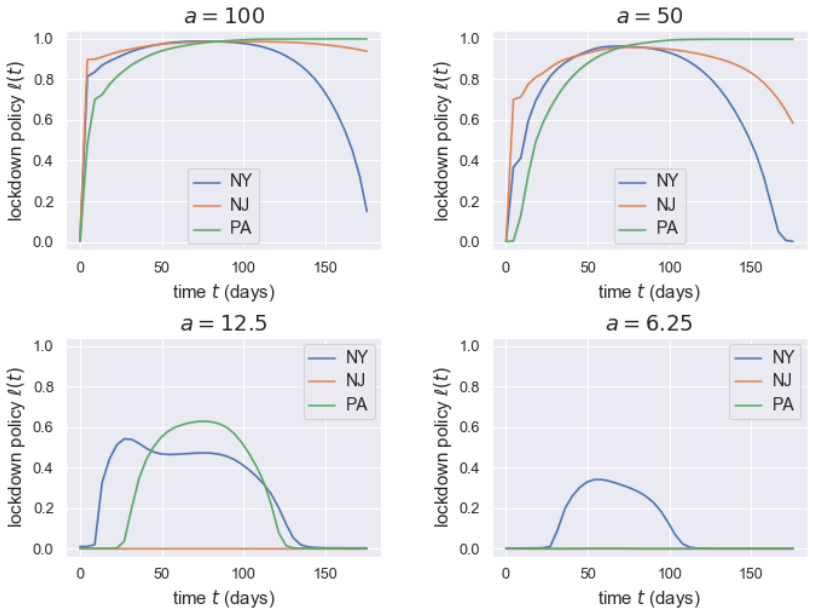

Figure 4: Plots of optimal policies with different choice of a for three states: New York (blue), New Jersey (orange) and Pennsylvania (green), when the lockdown efficiency is $\theta = 0.99$.

policy play an important role in decision-making. In reality, all three states issued stay-at-home orders in March, and attempted to reopen in June. By comparing real world policies and our simulations of $\boldsymbol{\ell}(x)$, we may infer α and θ for NY, NJ, and PA in our model, *i.e.*, $\theta = 0.99$ and $a = 25$.

5. Conclusion

In this paper, we propose a novel multi-region SEIR model to study optimal policies under a pandemic. Our new model, built on game theory, takes into account how the social and health policies issued by multiple region planners affect the progress of infectious diseases. This feature makes the model more realistic and powerful but also introduces a formidable computational challenge due to the high-dimensionality of the solution space and the strong coupling of planners' policies. We propose the enhanced deep fictitious play algorithm to overcome the curse of dimensionality and use the model and algorithm in a case study of the COVID-19 pandemic in three states, New York, New Jersey, and Pennsylvania. The model parameters are estimated from real data posted by the CDC. We are able to show the effect of lockdown/travel ban policy on the spread of COVID-19 for each state and how people's willingness to comply and planners' attitude towards deaths influence the equilibrium strategies as a consequence of the competition between regions. We hope our model can draw more attention to studying optimal interventions in infectious diseases using game theory. Our numerical simulations can shed light on public policies.

In reality, during a pandemic, the planning is usually for short periods and adjusted frequently. This can be modeled using repeated games, and planners may infer other regions' cost functional from past game outcomes. The assumptions that some parameters are identical across different

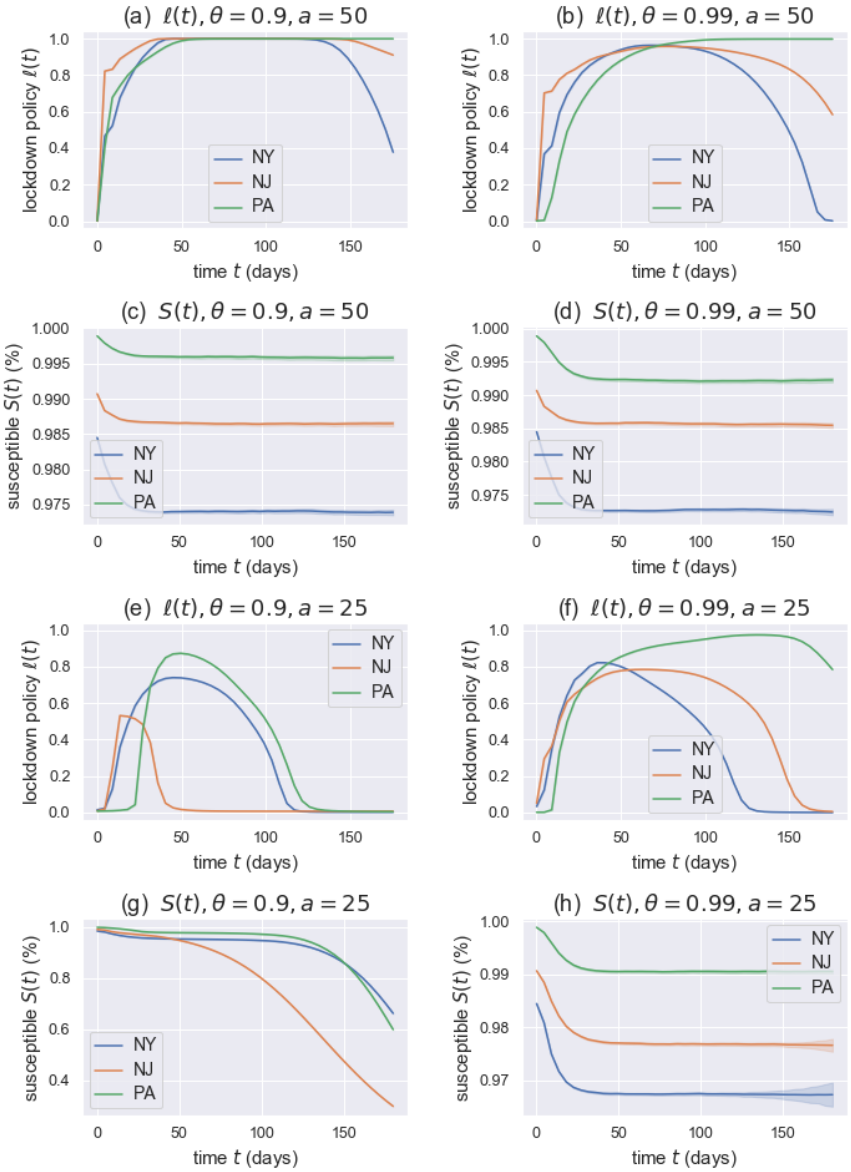

Figure 5: Comparison of optimal policies for three states (NY = blue, NJ = orange, PA = green) and their susceptibles between different policy effectiveness θ and hyperparameter a.

regions can be relaxed and the health policy can also be added for more accurate simulations. These will be left for future work.

Acknowledgments

R.H. was partially supported by NSF grant DMS-1953035, and the Faculty Career Development Award and the Research Assistance Program Award, University of California, Santa Barbara. H.D.C. acknowledges partial support from NSF grant DMS-1818821. This project was jointly supervised

by R.H. and H.D.C.. Y.X. and R.B. have equal contribution as the first authors. J.H. contributes on the algorithms and early discussions of the model setup.

References

F. E. Alvarez, D. Argente, and F. Lippi. A simple planning problem for COVID-19 lockdown. Working Paper 26981, National Bureau of Economic Research, 2020.

C. T. Bauch and D. J. D. Earn. Vaccination and the theory of games. *Proceedings of the National Academy of Sciences*, 101(36):13391–13394, 2004.

C. T. Bauch, A. P. Galvani, and D. J. D. Earn. Group interest versus self-interest in smallpox vaccination policy. *Proceedings of the National Academy of Sciences*, 100(18):10564–10567, 2003.

G. W. Brown. Some notes on computation of games solutions. Technical report, Rand Corp Santa Monica CA, 1949.

G. W. Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951.

S. L. Chang, M. Piraveenan, P. Pattison, and M. Prokopenko. Game theoretic modelling of infectious disease dynamics and intervention methods: a review. *Journal of Biological Dynamics*, 14(1): 1–33, 2020.

W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

N. El Karoui, S. Peng, and M. C. Quenez. Backward stochastic differential equations in finance. *Mathematical Finance*, 7(1):1–71, 1997.

A. S. Fauci, H. C. Lane, and R. R. Redfield. COVID-19 —navigating the uncharted. *New England Journal of Medicine*, 382(13):1268–1269, 2020.

R. E. Hall, C. I Jones, and P. J. Klenow. Trading off consumption and COVID-19 deaths. Working Paper 27340, National Bureau of Economic Research, 2020.

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)*, volume 107, pages 221–245, 2020.

J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

J. Han, R. Hu, and J. Long. Convergence of deep fictitious play for stochastic differential games. *arXiv preprint arXiv:2008.05519*, 2020a.

J. Han, R. Hu, and J. Long. Barron metric for the convergence of empirical distribution. *in preparation*, 2020b.

Fair Health. Costs for a Hospital Stay for COVID-19, 2020. URL https://www.fairhealth.org/article/costs-for-a-hospital-stay-for-covid-19.

R. Hu. Deep fictitious play for stochastic differential games. *Communications in Mathematical Sciences*, 2020.

R. Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. London: John Wiley and Sons, 1965.

S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020.

N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.-M. Jung, B. Yuan, R. Kinoshita, and H. Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 538(9), 2020.

W.-M. Liu, H. W. Hethcote, and S. A. Levin. Dynamical behavior of epidemiological models with nonlinear incidence rates. *Journal of Mathematical Biology*, 25(4):359–380, 1987.

G. Meyerowitz-Katz and L. Merone. A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *medRxiv*, 2020. doi: 10.1101/2020.05.03.20089854.

E. Pardoux and S. Peng. Backward stochastic differential equations and quasilinear parabolic partial differential equations. In *Stochastic Partial Differential Equations and Their Applications*, pages 200–217. Springer, 1992.

E. Pardoux and S. Tang. Forward-backward stochastic differential equations and quasilinear parabolic PDEs. *Probability Theory and Related Fields*, 114(2):123–150, 1999.

L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong. Epidemic analysis of covid-19 in china by dynamical modeling, 2020.

D. Seale and J. Burnett. Solving large games with simulated fictitious play. *International Game Theory Review*, 8(03):437–467, 2006.

Appendix A. Technical Details to the Stochastic Multi-region SEIR Model

A.1. The dynamics of $\boldsymbol{X}_t$ in Section 2

In Section 2, for the ease of notations and clarity of the presentation, we rewrite the dynamics of (S_t^n, E_t^n, I_t^n) defined in (1)–(3) in the vector form

$$\mathrm{d}\boldsymbol{X}_t = b(t, \boldsymbol{X}_t, \boldsymbol{\ell}(t, \boldsymbol{X}_t), \boldsymbol{h}(t, \boldsymbol{X}_t))\,\mathrm{d}t + \Sigma(\boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{W}_t, \tag{18}$$

where $\boldsymbol{X}_t \equiv [\boldsymbol{S}_t, \boldsymbol{E}_t, \boldsymbol{I}_t]^{\mathrm{T}} \equiv [S_t^1, \cdots, S_t^N, E_t^1, \cdots, E_t^N, I_t^1, \cdots, I_t^N]^{\mathrm{T}} \in \mathbb{R}^{3N}$, the Markovian controls $(\boldsymbol{\ell}, \boldsymbol{h})$ are given by $\boldsymbol{\ell}(t, \boldsymbol{x}) = [\ell^1, \ldots, \ell^N]^{\mathrm{T}}(t, \boldsymbol{x})$ and $\boldsymbol{h}(t, \boldsymbol{x}) = [h^1, \ldots, h^N]^{\mathrm{T}}(t, \boldsymbol{x})$. In the sequel, for a vector $\boldsymbol{x} \equiv (\boldsymbol{s}, \boldsymbol{e}, \boldsymbol{i}) \in \mathbb{R}^{3N}$, we shall index them in two ways,

$$(s_1, \cdots, s_N, e_1, \cdots, e_N, i_1, \cdots, i_N) \text{ or } (x_1, \ldots, x_{3N}).$$

and use them interchangeably. The former one emphasizes the dependence on each category, while the later notation is more condensed. Similarly, for partial derivatives, we will have two set of notations

$$(\partial_{s_1}, \cdots, \partial_{s_N}, \partial_{e_1}, \cdots, \partial_{e_N}, \partial_{i_1}, \cdots, \partial_{i_N}) \text{ or } (\partial_{x_1}, \cdots, \partial_{x_{3N}}).$$

We give precise definitions of (18) in this appendix. $b(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h}) = [b_1, \ldots, b_{3N}]^{\mathrm{T}}(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h})$ is a deterministic vector-valued function:

$$b_j(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h}) = \begin{cases} -\sum_{k=1}^{N} \beta^{jk} s_j i_k (1 - \theta \ell^j(t, \boldsymbol{x}))(1 - \theta \ell^k(t, \boldsymbol{x})) - v(h^j(t, \boldsymbol{x})) s_j, & j \in \mathcal{N}, \\ \sum_{k=1}^{N} \beta^{j'k} s_{j'} i_k (1 - \theta \ell^{j'}(t, \boldsymbol{x}))(1 - \theta \ell^k(t, \boldsymbol{x})) - \gamma e_{j'}, & j \in \mathcal{N} + N, \\ \gamma e_{j'} - \lambda(h^{j'}(t, \boldsymbol{x})) i_{j'}, & j \in \mathcal{N} + 2N, \text{ and } j' = j \bmod N. \end{cases}$$

$\Sigma(\boldsymbol{x}) = (\Sigma_{j,k}(\boldsymbol{x}))$ is a matrix-valued deterministic function in $\mathbb{R}^{3N \times 2N}$ with non-zero entries given below:

$$\begin{aligned} &\Sigma_{j,j}(\boldsymbol{x}) = -\sigma_{s_j} s_j, && \Sigma_{j+N,j}(\boldsymbol{x}) = \sigma_{s_j} s_j, \\ &\Sigma_{j+N,j+N}(\boldsymbol{x}) = -\sigma_{e_j} e_j, && \Sigma_{j+2N,j+N}(\boldsymbol{x}) = \sigma_{e_j} e_j, \quad j \in \mathcal{N}. \end{aligned}$$

and $\{\boldsymbol{W}_t\}_{0 \le t \le T}$ is a $2N$-dimensional standard Brownian motion:

$$\boldsymbol{W}_t = [W_t^{s_1}, \cdots, W_t^{s_N}, W_t^{e_1}, \cdots, W_t^{e_N}]^{\mathrm{T}}.$$

Note that all parameters $\beta^{jk}, \gamma, \lambda$, etc., are introduced in Section 2.

According to (5), each region's running cost f^n is

$$f^n(t, \boldsymbol{x}, \boldsymbol{\ell}, \boldsymbol{h}) = e^{-rt} P^n[(s_n + e_n + i_n)\ell^n(t, \boldsymbol{x})w + a(\kappa i_n \chi + p i_n c)] + e^{-rt}\eta(h^n(t, \boldsymbol{x}))^2.$$

A.2. The decoupled HJB equations in the form of (10)

Recall that we aim to solve (8) using the BSDE approach (nonlinear Feynman Kac relation). To this end, in this appendix, we will rewrite it in the form of (10) and identify μ^n and g^n. The first step is to identify the minimizer in the Hamiltonian (7). Keeping in mind that our testing case is COVID-19 where vaccines are not fully developed yet, the term $v(h_t^n) = 0$ is essentially zero in the past. Also, to focus on the lockdown/travel ban policy between different regions (as we did in the case study), we shall exclude the health policy $\boldsymbol{h}(t, \boldsymbol{x})$ from planners' decision problem, *i.e.*, $v(\cdot) = 0$, $\lambda(\cdot) = \lambda$, and $\eta = 0$ in the following derivations, and remove the dependence of $\boldsymbol{h}$ in all relevant functions. We remark that including the health policy $\boldsymbol{h}(t, \boldsymbol{x})$ is a straightforward generalization.

Recall that the Hamiltonian in (8) reads:

$$\begin{aligned} &H^n(t, \boldsymbol{x}, (\ell^n, \boldsymbol{\ell}^{-n,m}), \nabla_{\boldsymbol{x}} V^{n,m+1}) \\ &\quad = b(t, \boldsymbol{x}, (\ell^n, \boldsymbol{\ell}^{-n,m})) \cdot \nabla_{\boldsymbol{x}} V^{n,m+1} + f^n(t, \boldsymbol{x}, \ell^n) \\ &\quad = \sum_{j=1}^{3N} b_j(t, \boldsymbol{x}, (\ell^n, \boldsymbol{\ell}^{-n,m})) \frac{\partial V^{n,m+1}}{\partial x_j} + e^{-rt} P^n[(s_n + e_n + i_n)\ell^n(t, \boldsymbol{x})w + a(\kappa i_n \chi + p i_n c)], \end{aligned}$$

and recall that $\boldsymbol{\ell}^{-n,m} = (\ell^{1,m}, \ldots, \ell^{n-1,m}, \ell^{n+1,m}, \ldots, \ell^{N,m})$ represents the m^{th} stage strategies of all players other than n, which are given functions in this derivation. The first order condition requires for ℓ^n:

$$0 = \sum_{\substack{j=1 \\ j\neq n}}^{N}(1-\theta\ell^{j,m})\left[\beta^{jn}s_j i_n\left(\frac{\partial V^{n,m+1}}{\partial e_j} - \frac{\partial V^{n,m+1}}{\partial s_j}\right) + \beta^{nj}s_n i_j\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right)\right]$$

$$+2(1-\theta\ell^n)\beta^{nn}s_n i_n\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right) - \frac{1}{\theta}e^{-rt}P^n(s_n+e_n+i_n)w.$$

The critical point given by the above equation indeed gives a minimizer of the Hamiltonian, as long as it is in $[0,1]$. Because we can show $\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right) > 0$ by comparing $V^{n,m+1}(t, \boldsymbol{x} + \epsilon_{n+N})$ and $V^{n,m+1}(t, \boldsymbol{x}+\epsilon_n)$ using their definitions, where ϵ_j is a $3N$-vector with only one nonzero entry $\epsilon \ll 1$ at j^{th} position. Intuitively, with all others players' initial condition the same, if player n starts with a higher exposed proportion $e_n + \epsilon$, it will produce more cost, comparing with the same increase proportion still being susceptible $s_n + \epsilon$. To summarize, we deduce the optimal policy for player n at stage $m+1$ is given by:

$$\ell^{n,m+1}(t,\boldsymbol{x}) = \left\{2\beta^{nn}s_n i_n\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right) - \frac{1}{\theta}e^{-rt}P^n(s_n+e_n+i_n)w \right. \tag{19}$$

$$\left. + \sum_{\substack{j=1 \\ j\neq n}}(1-\theta\ell^{j,m})\left[\beta^{jn}s_j i_n\left(\frac{\partial V^{n,m+1}}{\partial e_j} - \frac{\partial V^{n,m+1}}{\partial s_j}\right) + \beta^{nj}s_n i_j\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right)\right]\right\}$$

$$\times\left\{2\theta\beta^{nn}s_n i_n\left(\frac{\partial V^{n,m+1}}{\partial e_n} - \frac{\partial V^{n,m+1}}{\partial s_n}\right)\right\}^{-1} \wedge 1 \vee 0,$$

where we use the conventional notations $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. Plugging (19) into equation (8) and by straightforward computation, one obtains for the $(m+1)^{th}$ stage, $\mu^{n,m+1}(t,\boldsymbol{x};\boldsymbol{\ell}^{-n,m}) = [\mu_1^{n,m+1}, \ldots, \mu_{3N}^{n,m+1}](t,\boldsymbol{x};\boldsymbol{\ell}^{-n,m})^{\mathrm{T}}$ is

$$\mu_j^{n,m+1} = -\beta^{jn}s_j i_n(1-\theta\ell^{j,m}(t,\boldsymbol{x})) - \sum_{\substack{k=1 \\ k\neq n}}^{N}\beta^{jk}s_j i_k(1-\theta\ell^{j,m}(t,\boldsymbol{x}))(1-\theta\ell^{k,m}(t,\boldsymbol{x})),$$

$$j \in \mathcal{N}\setminus n$$

$$\mu_n^{n,m+1} = -\beta^{nn}s_n i_n - \sum_{\substack{k=1 \\ k\neq n}}\beta^{nk}s_n i_k(1-\theta\ell^{k,m}(t,\boldsymbol{x}))$$

$$\mu_{N+j}^{n,m+1} = -\mu_j^{n,m+1} - \gamma e_j, \quad \mu_{2N+j}^{n,m+1} = \gamma e_j - \lambda i_j, \quad j \in \mathcal{N}.$$

To write $g^{n,m+1}$ as a function of $(t, \boldsymbol{x}, z)$, we first compute

$$\Sigma(\boldsymbol{x})^{\mathrm{T}}\nabla_{\boldsymbol{x}}V^n(t,\boldsymbol{x})$$
$$= \left[\sigma_{s_1}s_1\left(\frac{\partial V^n}{\partial e_1} - \frac{\partial V^n}{\partial s_1}\right), \cdots, \sigma_{s_N}s_N\left(\frac{\partial V^n}{\partial e_N} - \frac{\partial V^n}{\partial s_N}\right), \sigma_{e_1}e_1\left(\frac{\partial V^n}{\partial i_1} - \frac{\partial V^n}{\partial e_1}\right), \cdots, \sigma_{e_N}e_N\left(\frac{\partial V^n}{\partial i_N} - \frac{\partial V^n}{\partial e_N}\right)\right]^{\mathrm{T}}.$$

and then $g^{n,m+1}$ is given by:

$$g^{n,m+1}(t,\boldsymbol{x},z;\boldsymbol{\ell}^{-n,m}) = \frac{\theta^2}{\sigma_{s_n}}\beta^{nn}z_n i_n[\ell^{n,m+1}(t,\boldsymbol{x})]^2$$
$$+\left\{e^{-rt}P^n(s_n+e_n+i_n)w - 2\frac{\theta}{\sigma_{s_n}}\beta^{nn}z_n i_n - \sum_{\substack{j=1\\ j\neq n}}^{N}\theta(1-\theta\ell^{j,m}(t,\boldsymbol{x}))\left(\frac{\beta^{nj}}{\sigma_{s_n}}z_n i_j + \frac{\beta^{jn}}{\sigma_{s_j}}z_j i_n\right)\right\}\ell^{n,m+1}(t,\boldsymbol{x})$$
$$+ e^{-rt}P^n a(\kappa i_n\chi + p i_n c).$$

A.3. The derivation of the transmission rate β^{nk} in (17)

We denote by β the underlying transmission rate of the virus, which is assumed to be region independent. This transmission rate is the average number of people infected by an infectious person per day (assuming that the susceptible population is 100%). Thus, if a proportion S of the population is susceptible, then βS represents the average number of people infected per infectious person per day. If there are a total of P people in a population with a fraction I being infectious, then $\beta S(PI)$ is the number of newly infected per day.

Thus, in the context of a single-region SEIR model, the number of newly infected (or the influx to the exposed population) that occurs within $(t, t + \mathrm{d}t)$ is given by $\beta S(t)(I(t)P)\,\mathrm{d}t$. Dividing by P, the influx to the scaled exposed population in $(t, t + \mathrm{d}t)$ is $\beta S(t)I(t)\,\mathrm{d}t$.

Now let us consider the multi-region case and temporarily ignore the effect of lockdown. The term from (1)–(3) that gives the influx to E^n due to infection from I^k is $\beta^{nk}S^n(t)I^k(t)\,\mathrm{d}t$. To determine β^{nk}, we build this exact influx from core assumptions.

First, we quantify the number of people from region n that are infected by those from region k. Specifically, the influx in the interval $(t, t + \mathrm{d}t)$ to the unscaled exposed population n due to transmission from population k is given by

$$\sum_{\ell}(\beta f^{n\ell}S^n(t))(f^{k\ell}I^k(t)P^k)\,\mathrm{d}t, \tag{20}$$

where f^{ij} is the (assumed constant) fraction of people from i currently in region j at any moment in time.

Equation (20) is obtained by summing the number of infections in population n due to population k across each region. The summand represents these infections occurring in region ℓ. This can be seen as the term $f^{n\ell}S^n(t)$ is the proportion of population n that are susceptible and within region ℓ. Of this population, there will be $\beta f^{n\ell}S^n(t)$ infections per infectious individual per day. Since the

number of infectious from k that are in ℓ is $f^{k\ell}I^k(t)P^k$, we have that $(\beta f^{n\ell}S^n(t))(f^{k\ell}I^k(t)P^k)\,\mathrm{d}t$ is the number of new infections in population n due to population k occurring in the region ℓ within the time interval $(t, t+\mathrm{d}t)$.

Let us assume for now that $n \neq k$. Since we assume that $1 > f^{nn} \gg f^{nk}$, the terms in (20) besides the cases where $\ell = n$ or $\ell = k$ are negligible. Removing the negligible terms and dividing by the population P^n, we see that the influx to the scaled exposed population E^n due to transmission from population k over the interval $(t, t+\mathrm{d}t)$ is

$$\frac{P^k}{P^n}\beta(f^{nn}f^{kn} + f^{nk}f^{kk})S^n(t)I^k(t)\,\mathrm{d}t,$$

which is exactly the influx represented by the model of $\beta^{nk}S^n(t)I^k(t)\,\mathrm{d}t$. This verifies the form of β^{nk} for $n \neq k$ in (17). Similarly if we take $n = k$ in (20) and ignore all terms other than $\ell = n(= k)$, then we derive the formula for β^{nn} in (17).

Appendix B. Detailed discussions on the enhanced DFP algorithm

B.1. Implementation details

In the NY-NJ-PA case study, we choose feedforward architectures for both V-networks and α-networks. Both have three hidden layers with a width of 40 neurons. The activation function in each hidden layer is $\tanh(x)$. We do not apply activation function to the output layer of V-networks, and choose sigmoid function $\rho_s(x) = \frac{1}{1+e^{-x}}$ for the α-networks. Other hyperparameters are summarized in the table below.

hyperparameter	lr	M	$N_{\text{SGD\_per\_stage}}$	N_{batch}	N_T	τ
value	5e-4	250	100	256	40	$1e^{-3}$/180

Table 1: Hyperparameters in the case study: lr denotes the learning rate in stochastic gradient descent method, M is the total stages of fictitious play, $N_{\text{SGD\_per\_stage}}$ is the number of stochastic gradient descent done in each minimization problem (14), N_{batch} is batch size in each stochastic gradient descent, N_T is the discretization steps on $[0, T]$, and τ is the weight of the control part in the loss function (14).

B.2. Discussion on the choice of M and $N_{\text{SGD\_per\_stage}}$

We provide further experiments here on various choices of M and $N_{\text{SGD\_per\_stage}}$. In Figure 6, we plot both validation loss and log loss against M for three states, which are produced by evaluating the NNs using unseen data after each fictitious play stage. In each panel, loss curves associated with different number of SGDs per stage are presented in different colors (blue = 50, orange = 75, green = 100, red = 125, purple = 150).

The numerical results show that the validation losses for all states decrease as the number of DFP stages M increases. Moreover, it shows that different $N_{\text{SGD\_per\_stage}}$ generates loss curves with similar patterns. Smaller $N_{\text{SGD\_per\_stage}}$ is more stable on the validation loss of PA. This result is consistent with Han and Hu (2020) and Seale and Burnett (2006), which convey that it is unnecessary to solve the problem extremely accurate in each stage and that a moderate number of $N_{\text{SGD\_per\_stage}}$ is sufficient. As a result, we choose $N_{\text{SGD\_per\_stage}} = 100$ in our case study.

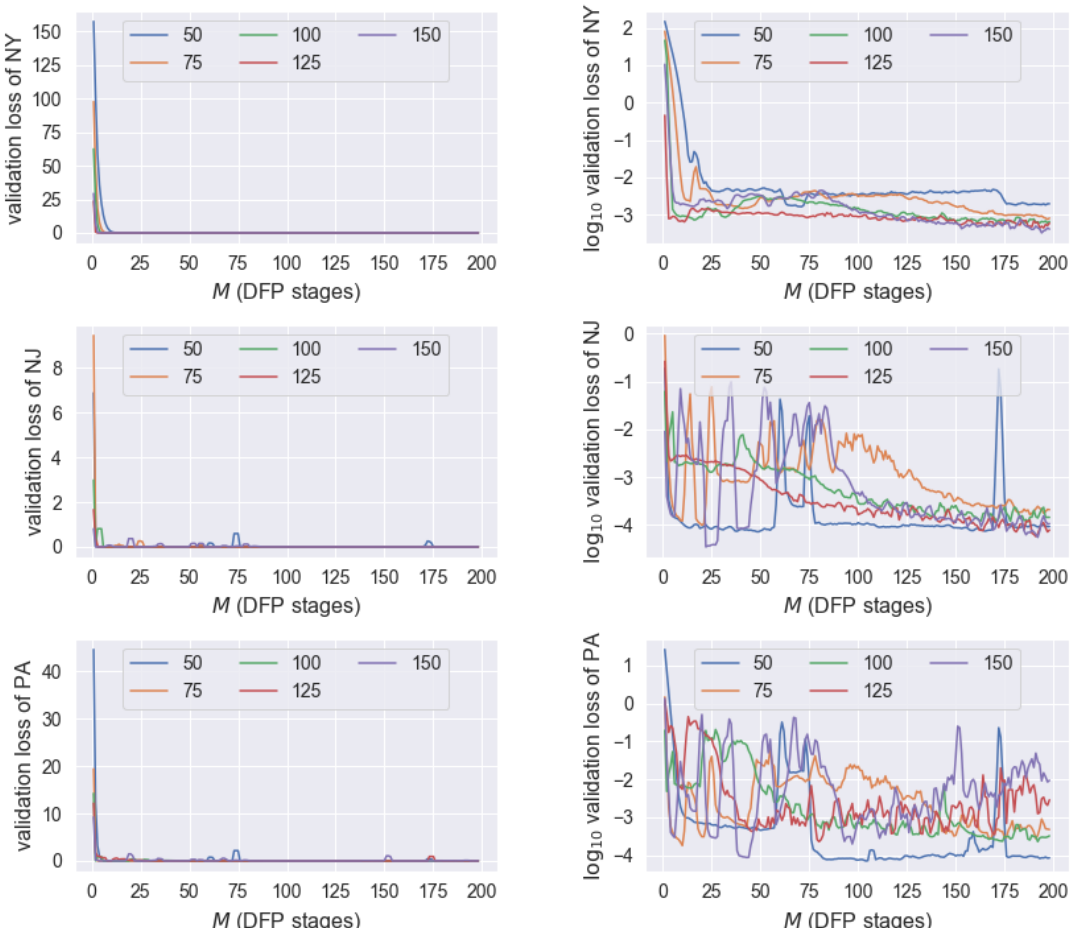

Figure 6: Loss curves of each state. Left: validation losses versus rounds M of the enhanced deep fictitious play; Right: $\log_{10}$ validation loss versus rounds M of the enhanced deep fictitious play. The loss curves with respect to $N_{\text{SGD\_per\_stage}} = 50, 75, 100, 125, 150$ are depicted in blue, orange, green, purple and red. A smoothed moving average with window size 3 is applied in the final plots.

B.3. Stability over different experiments

Here we present experiments to investigate the Nash equilibrium of the model with different combinations of parameters. For each combination of parameters, we use the same hyper-parameters and repeat the experiments several times. We run the algorithm for a certain computational budget, and then filter out the results with a fluctuating loss near the stopping and check the converged equilibrium. In the first combination of parameters, we take $\theta = 0.99$ and $a = 100$, corresponding to the case that a governor weighs the deaths much more than the economic loss and tries to avoid it, and the residents have a strong willingness to follow the governor's policies. Intuitively, the pandemic is possible to get well-controlled. Our numerical experiments confirm this intuition: all converging trails lead to the same Nash equilibrium. A representative plot of $X(t) = (S(t), E(t), I(t))$ is shown in Figure 7.

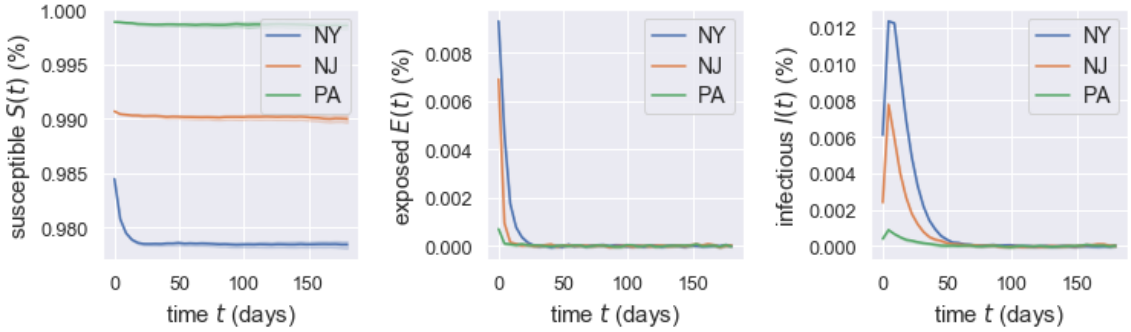

Figure 7: With the parameter combination $\theta = 0.99$, $a = 100$, the algorithm identifies one Nash equilibrium for the NY-NJ-PA case study.

In the second batch of experiments, we take $\theta = 0.9$ and $a = 50$, corresponding to the case that a governor pays less attention to the number of death and the residents are less willing to follow the policies compared to the first batch of experiments. The change leads to the possibility of multiple Nash equilibrium and the pandemic being out of control. In this case, with different NNs' initialization, the algorithm identifies two Nash equilibria: 75% of the experiments converge to the Nash equilibrium that the pandemic gets controlled and 25% of the experiments converge to the other Nash equilibrium where the pandemic gets out of control.

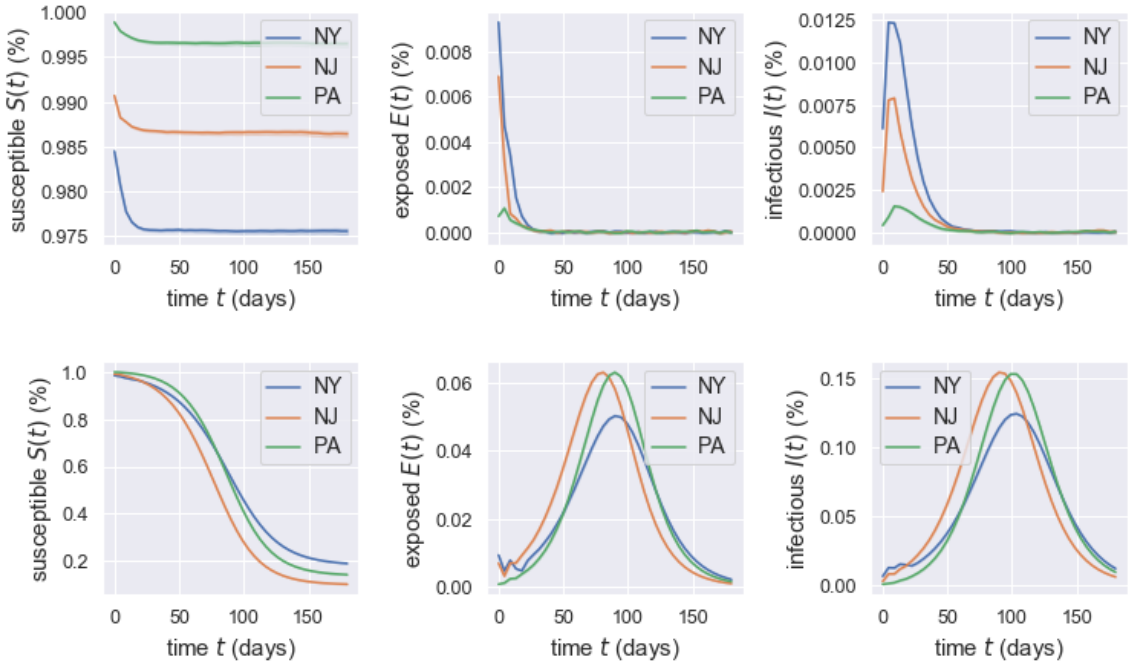

Figure 8: With the parameter combination $\theta = 0.9$, $a = 50$, the algorithm identifies two possible Nash equilibria: an under control one (topic panels, with 75% of the experiments) and on out-of-control one (bottom panels, with 25% of the experiments).

In conclusion, it is possible to have multiple Nash equilibria depending on the parameter chosen in our stochastic multi-region SEIR model. There is usually a single Nash equilibrium for parameters chosen at extreme values, while for the parameters selected in the middle range, there could exist multiple Nash equilibria. When multiple equilibria exist, we conjecture that the possibility to reach a particular one depends on where we start the fictitious play (the initialization of the NNs' parameters).