Pandemic Control, Game Theory and Machine Learning

Yao Xuan * Robert Balkin † Jiequn Han ‡ Ruimeng Hu §
Hector D. Ceniceros ¶

August 15, 2022

COVID-19 and Control Policies

The coronavirus disease 2019 (COVID-19) pandemic has brought an enormous impact on our lives. Based on data from World Health Organization, as of May 2022, there have been more than 520 million confirmed cases of infection and more than 6 million deaths globally; In the United States, there have been more than 83 million confirmed cases of infection and more than one million cases of death. 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.

Control policies play a crucial role in the alleviation of the COVID-19 pandemic. For example, lockdown and work-from-home policies and mask requirements on public transport and public areas have been proved to be effective in stopping the spreading of COVID-19. On the other hand, governors also have to be aware of the economic activity loss due to these pandemic control policies. Therefore, a thorough understanding of the evolution of COVID-19 and the corresponding decision-making provoked by such a virus will be beneficial for future events and in other interconnected systems around the world.

Epidemiology

Epidemiology is the science of analyzing the distribution and determinants of health-related states and events in specified populations. It is also the application of this study to the control of health problems. Infectious diseases are one of this kind, including the ongoing novel coronavirus (COVID-19).

Since March 2020, when the World Health Organization declared the COVID-19 outbreak a global pandemic, epidemiologists have made tremendous efforts to understand how COVID-19 infections emerge and spread and how they may be prevented and controlled. Many epidemiological methods involve mathematical tools, e.g., using causal inference to identify causative agents and factors for its propagation, and molecular methods to simulate disease transmission dynamics.

The first epidemic model concerning epidemic spreading dates back to 1760 by Daniel Bernoulli [Ber60]. Since then, many papers have been dedicated to this field and, later on, to epidemic control. Among control strategies, the quarantine, firstly introduced in 1377 in Dubrovnik on Croatias Dalmatian Coast [GB97], has shown as a powerful component of the public health response to emerging and reemerging infectious diseases. However, quarantine and other measures for controlling epidemic diseases have always been controversial due to the potentially raised political, ethical, and socioeconomic issues.

*Yao Xuan contributed to the project when he was a Ph.D. student at the University of California, Santa Barbara. His email address is yxscience@gmail.com.

†Robert Balkin is a current graduate student supervised by Ruimeng Hu and Hector D. Ceniceros at the University of California, Santa Barbara. His email address is rbalkin@ucsb.edu.

‡Jiequn Han is a research fellow at the Flatiron Institute. His email address is jiequnhan@gmail.com.

§Ruimeng Hu is an assistant professor at the University of California, Santa Barbara. Her email address is rhu@ucsb.edu.

¶Hector D. Ceniceros is a full professor at the University of California, Santa Barbara. His email address is ceniceros@ucsb.edu.

Such complication naturally calls for the inclusion of decision-making in epidemic control, as it helps to answer how to take *optimal* actions to balance public interest and individual rights. But not until recent years have there been some research studies in this direction. Moreover, when multiple authorities are involved in the decision-making process, it is challenging to analyze how to collectively or competitively make decisions due to the difficulty of solving this high-dimensional problem.

In this article, we focus on the decision-making development for the intervention of COVID-19, aiming to provide mathematical models and efficient numerical methods, and justifications for related policies that have been implemented in the past and explain how the authorities' decisions affect their neighboring regions from a game theory viewpoint.

Mathematical models

In a classic, compartmental epidemiological model, each individual in a geographical region is assigned a label, e.g., **S**usceptible, **E**xposed, **I**nfectious, **R**emoved, **V**accinated. Different labels represent different status – **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, **R**: who have been infected and then removed from the disease due to recovery or death, and **V**: who have been vaccinated and are immune to the infection. As COVID-19 progressed, it was learned that spread from asymptomatic cases was an important driving force. More refined models may further split **I** into mild-symptomatic/asymptomatic individuals who are in-home for recovery and serious-symptomatic ones that need hospitalization. We point to [AZM+20] which considers a similar problem in the optimal control setting, which includes asymptomatic individuals and the effect of impulses.

Individuals transit between these compartments, and the labels' order in a model indicates the flow patterns between the compartments. For instance, in a simple SEIR model [LHL87] (see also Figure 1a), a susceptible becomes exposed after close contact with infected ones; exposed individuals become infectious after a latency period; and infected ones become removed afterward due to recovery or death. Let $S(t)$, $E(t)$, $I(t)$ and $R(t)$ be the proportion of population of each compartment at time t, the following differential equations provide the mathematical model:

$$\begin{aligned} \dot{S}(t) &= -\beta S(t)I(t), \\ \dot{E}(t) &= \beta S(t)I(t) - \gamma E(t), \\ \dot{I}(t) &= \gamma E(t) - \lambda I(t),\ \dot{R}(t) = \lambda I(t), \end{aligned} \tag{1}$$

where β is the average number of contacts per person per time, γ describes the latent period when the person has been infected but not yet infectious, and λ represents the recovery rate measuring the proportion of people recovered or dead from infected population.

Many infections, such as measles and chickenpox, confer long-term, if not lifelong, immunity, while others, such as influenza, do not. As evidenced by numerous epidemiological and clinical studies analyzing possible factors for COVID reinfections, COVID-19 falls precisely into the second category [NBN22]. Mathematically, this can be taken into account by adding a transition $I \to S$.

Though deterministic models such as (1) have received more attention in the literature, mainly due to their tractability, stochastic models have some advantages. The epidemic-spreading progress is by nature stochastic. Moreover, introducing stochasticity to the system could account for numerical and empirical uncertainties, and also provide probabilistic predictions, i.e., a range of possible scenarios associated with their likelihoods. This is crucial for understanding the uncertainties in the estimates.

One class of stochastic epidemic models uses continuous-time Markov chains, where the state process takes discrete values but evolves in continuous time and is Markovian. In a simple Stochastic SIS (susceptible-infectious-susceptible) model [KL89] with a population of N individuals, let X_t be the number of infected individuals at time t, β the rate of infected individuals infecting those susceptible, and λ the rate that an infected individual recovers and becomes susceptible again. The transition probabilities

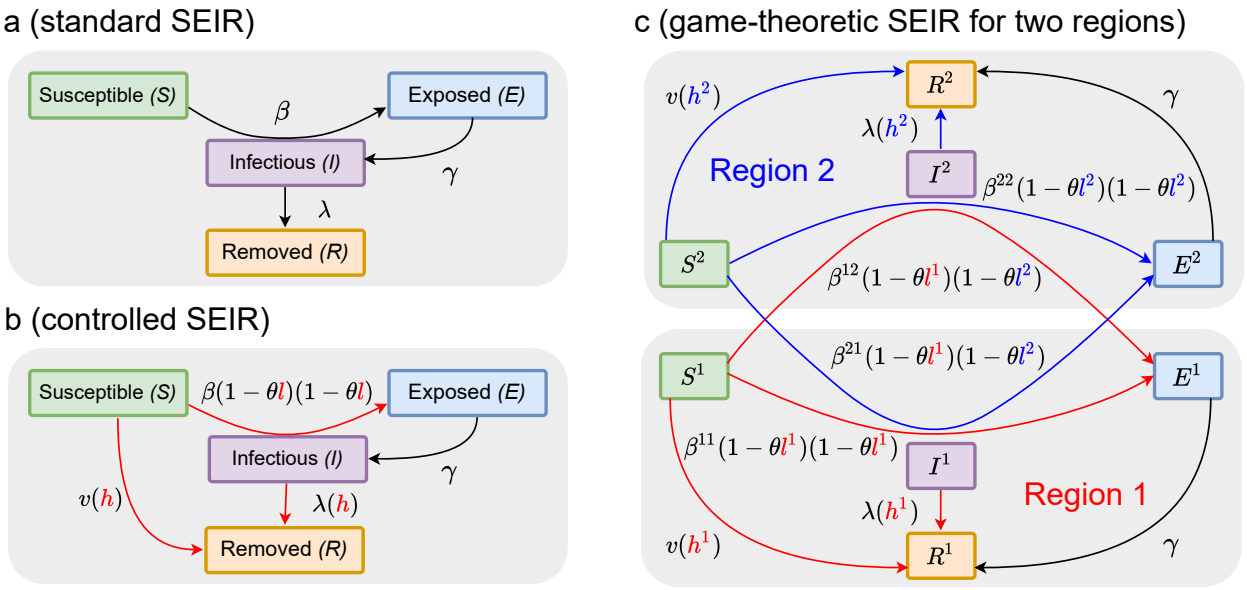

Figure 1: (a) A simple SEIR model: susceptible individuals become exposed after close contact with infected ones; those exposed become infectious after a latency period; and those infected become removed afterward due to recovery or death; (b) Controlled SEIR model: the planner chooses the level of nonpharmaceutical policies (lockdown or work from home) ℓ and pharmaceutical policies (effort of vaccination development or distribution) h affecting the transitions such that only $(1-\theta\ell(t))$ of the original susceptible and infectious individuals can contact each other, and affecting the recovery rate $\lambda(h)$ from infectious individuals to removed ones, here θ is used describe the effectiveness of policy ℓ; (c) An illustration of the game-theoretic SEIR model for two regions.

among states n, $n+1$, $n-1$ are

$$\begin{aligned}
\mathbb{P}(X_{t+\Delta t} = n+1 | X_t = n) &\approx \frac{\beta}{N} n(N-n)\Delta t, \\
\mathbb{P}(X_{t+\Delta t} = n-1 | X_t = n) &\approx \lambda n \Delta t, \\
\mathbb{P}(X_{t+\Delta t} = n | X_t = n) &\approx 1 - \left(\frac{\beta}{N} n(N-n) + \lambda n \right) \Delta t.
\end{aligned}$$

Another way to construct a stochastic model is by introducing white noise W_t in (1) [TBV05, All08], which we shall mainly consider in this paper and describe in details in the later section.

Control of disease spread

After modeling how diseases are transmitted through a population, epidemiologists then design corresponding control measures and recommend health-related policies to the region planner.

In general, there are two types of interventions: pharmaceutical interventions (PIs), such as getting vaccinated and taking medicines, and nonpharmaceutical interventions (NPIs), such as requiring mandatory social distancing, quarantining infected individuals, and deploying protective resources. For the ongoing COVID-19, intervention policies that have been implemented include, but are not limited to, issuing lockdown or work-from-home policies, developing vaccines, and later expanding equitable vaccine dis-

tribution, providing telehealth programs, deploying protective resources and distributing free testing kits, educating the public on how the virus transmits, and focusing on surface disinfection.

Mathematically, this can be formulated as a control problem: the planner chooses the level of each policy affecting the transitions in (1) such that the region's overall cost is minimized. Generally, NPIs help mitigate the spread by lowering the infection rate β, e.g., a lockdown or work-from-home policy $\ell(t)$ implemented at time t modifies the transition to

$$\dot{S}(t) = -\beta(1-\theta\ell(t))S(t)(1-\theta\ell(t))I(t),$$

meaning that only $(1-\theta\ell(t))$ of the original susceptible and infectious individuals can contact each other where θ describes the effectiveness of ℓ [AAL20] (see Figure 1b). PIs such as taking preventive medicines, if available, will also lower the infection rate β, while using antidotes will increase the recovery rate λ. The modeling of vaccinations is more complex. Depending on the target disease, it may reduce β (less chance to be infected) or increase λ (faster recovery). It may even create a new compartment "Vaccinated" in which individuals can not be infected and which is an absorbing state if life-long immunity is gained.

A region planner, taking into account the interventions' effects on the dynamics (1), decides on policy by weighing different costs. These costs may include the economic loss due to decrease in productivity during a lockdown, the economic value of life due to death of infected individuals and other social-welfare costs due to the aforementioned measurements.

Game-theoretic SEIR Model

Game theory studies the strategic interactions among rational players and has applications in all fields of social science, computer science, financial mathematics, and epidemiology. A game is non-cooperative if players cannot form alliances or if all agreements need to be self-enforcing. Nash equilibrium is the most common kind of self-enforcing agreement [Nas51], in which a collective strategy emerges from all players in the game to which no one has an incentive to deviate unilaterally.

Nowadays, as the world is more interconnected than ever before, one region's epidemic policy will inevitably influence the neighboring regions. For instance, in the US, decisions made by the governor of New York will affect the situation in New Jersey, as so many people travel daily between the two states. Imagine that both state governors make decisions representing their own benefits, take into account others' rational decisions, and may even compete for the scarce resources (e.g., frontline workers and personal protective equipment). These are precisely the features of a non-cooperative game. Computing the Nash equilibrium from such a game will provide valuable, qualitative guidance and insights for policymakers on the impact of specific policies.

We now introduce a multi-region stochastic SEIR model [XBH+22] to capture the game features in epidemic control. We give an illustration for two regions in Figure 1c. Each region's population is divided into four compartments: **S**usceptible, **E**xposed, **I**nfectious, and **R**emoved. Denote by $S_t^n, E_t^n, I_t^n, R_t^n$ the *proportion* of the population in the four compartments of the region n at time t. They satisfy the following stochastic differential equations (SDEs), which have included interventions (PIs and NPIs), stochastic factors, and game features,

$$\begin{aligned}
\mathrm{d}S_t^n &= -\sum_{k=1}^{N} \beta^{nk} S_t^n I_t^k (1-\theta\ell_t^n)(1-\theta\ell_t^k)\,\mathrm{d}t \\
&\quad - v(h_t^n) S_t^n\,\mathrm{d}t - \sigma_{s_n} S_t^n\,\mathrm{d}W_t^{s_n}, && (2)\\
\mathrm{d}E_t^n &= \sum_{k=1}^{N} \beta^{nk} S_t^n I_t^k (1-\theta\ell_t^n)(1-\theta\ell_t^k)\,\mathrm{d}t && (3)\\
&\quad - \gamma E_t^n\,\mathrm{d}t + \sigma_{s_n} S_t^n\,\mathrm{d}W_t^{s_n} - \sigma_{e_n} E_t^n\,\mathrm{d}W_t^{e_n}, \\
\mathrm{d}I_t^n &= (\gamma E_t^n - \lambda(h_t^n) I_t^n)\,\mathrm{d}t + \sigma_{e_n} E_t^n\,\mathrm{d}W_t^{e_n}, && (4)\\
\mathrm{d}R_t^n &= \lambda(h_t^n) I_t^n\,\mathrm{d}t + v(h_t^n) S_t^n\,\mathrm{d}t, && (5)
\end{aligned}$$

where $n \in \mathcal{N} := \{1, 2, \ldots, N\}$ is the collection of N regions, W_t with different superscripts indicate white noise for a compartment in a specific region, $\boldsymbol{\ell}_t \equiv (\ell_t^1, \ldots, \ell_t^N)$ and $\boldsymbol{h}_t \equiv (h_t^1, \ldots, h_t^N)$ are NPIs and PIs chosen by the region planners at time t. The planner of region n minimizes its region's cost within a period

$[0,T]$:

$$J^n(\boldsymbol{\ell},\boldsymbol{h}) := \mathbb{E}\Big[\int_0^T e^{-rt}P^n\big[(S_t^n+E_t^n+I_t^n)\ell_t^n w + a(\kappa I_t^n\chi + pI_t^n c)\big] + e^{-rt}\eta(h_t^n)^2\,\mathrm{d}t\Big]. \tag{6}$$

We explain the model (2)–(6) in details:

S. In (2), β^{nk} denotes the average number of contacts of infected people in region k with susceptible ones in region n per time unit. Although some regions may not be geographically connected, the transmission between the two is still possible due to air travel, but is still less intensive than the transmission within the region, i.e., $\beta^{nk}>0$ and $\beta^{nn}\gg\beta^{nk}$ for all $k\neq n$. The decision for NPIs of region n's planner is given by $\ell_t^n\in[0,1]$. In particular, it represents the fraction of the population being under NPIs (such as social distancing) at time t. We assume that those under interventions 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. We use $\theta\in[0,1]$ to measure this effectiveness. The transition rate under policy $\boldsymbol{\ell}$ thus become $\beta^{nk}S_t^nI_t^k(1-\theta\ell_t^n)(1-\theta\ell_t^k)$. The case $\theta=1$ means the policy is fully effective. One can also view θ as the level of public compliance.

The planner of region n also makes the decision $h_t^n\in[0,1]$. This represents the effort, at time t, that the planner puts into PIs. We refer to this term, h_t^n, as the *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. In this model, we assume that 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. This assumption is not very consistent with COVID-19 but reasonable for a short-term decision-making problem. We model it as an increasing function of h_t^n, and if the vaccine has not yet been developed, we can define $v(x)=0$ for $x\leq\overline{h}$.

E. In (3), γ describes the latent period when the person is infected but is not yet infectious. 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 individuals, i.e., γE_t^n.

I and R. In (4) and (5), $\lambda(\cdot)$ 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 and the recovery rate related to the health policy h_t^n decided by its planner. The more effort put into the region (e.g., expanding hospital capacity and 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.

Cost. In (6), 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 economic cost of each death. The hyperparameter a describes how planners weigh deaths and infections as compared to other costs. The third one is the in-patient cost, where p is the hospitalization rate, and c is the cost per in-patient per day. The last term $\eta(h_t^n)^2$ quantifies the grants for health policies. We choose a quadratic form so that the function is concave in h_t^n. This is to account for the law of diminishing marginal utility: the marginal utility from each additional unit declines as investment increases. 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 it as $J^n(\boldsymbol{\ell},\boldsymbol{h})$.

The above model (2)–(5) is by no doubt a prototype, and one can generalize it by considering reinfections (adding transmission from R^n to S^n), asymptomatic population (adding asymptomatic compartment A^n),

different control policy for S^n and I^n (using ℓ_S and ℓ_I in (2)–(3)), different fatality rates for young and elder population (introducing κ_Y and κ_E in (6)).

Nash equilibria and the HJB system

As explained above, the interaction between region planners can be viewed as a non-cooperative game, when Nash equilibrium is the notion of optimality.

Definition 1. *A Nash equilibrium (NE) is a tuple* $(\boldsymbol{\ell}^*, \boldsymbol{h}^*) = (\ell^{1,*}, h^{1,*}, \dots, \ell^{N,*}, h^{N,*}) \in \mathbb{A}^N$ *such that* $\forall n \in \mathcal{N}$ *and* $(\ell^n, h^n) \in \mathbb{A}$,

$$J^n(\boldsymbol{\ell}^*, \boldsymbol{h}^*) \le J^n((\boldsymbol{\ell}^{-n,*}, \ell^n), (\boldsymbol{h}^{-n,*}, h^n)),$$

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

$$\boldsymbol{\ell}^{-n,*} := [\ell^{1,*}, \dots, \ell^{n-1,*}, \ell^{n+1,*}, \dots, \ell^{N,*}] \in \mathbb{A}^{N-1}.$$

Here $\mathbb{A}$ *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 that all players take actions in the same space.*

Under proper conditions, the NE is obtained by solving N-coupled Hamilton-Jacobi-Bellman (HJB) equations via dynamic programming [CD18, Section 2.1.4]. To simplify the notation, we concatenate the states into a vector form $\boldsymbol{X}_t \equiv [\boldsymbol{S}_t, \boldsymbol{E}_t, \boldsymbol{I}_t]^{\mathrm{T}} \equiv [S_t^1, \cdots, S_t^N, E_t^1, \cdots, E_t^N, I_t^1, \cdots, I_t^N]^{\mathrm{T}} \in \mathbb{R}^{3N}$, and denote its dynamics by

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

For the sake of simplicity, we omit the actual definition of b, f^n and Σ and refer [XBH+22] for further details. Let $V^n(t, \boldsymbol{x})$ be the minimized cost defined in (6) if the system starts at $\boldsymbol{X}_t = x$. Then, V^n, $n = 1, \dots, N$ solves

$$\partial_t V^n + \inf_{(\ell^n, h^n) \in [0,1]^2} H^n(t, \boldsymbol{x}, (\boldsymbol{\ell}, \boldsymbol{h})(t, \boldsymbol{x}), \nabla_{\boldsymbol{x}} V^n) + \frac{1}{2}\mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}}\mathrm{Hess}_{\boldsymbol{x}} V^n \Sigma(\boldsymbol{x})) = 0, \quad (7)$$

with $V^n(T, \boldsymbol{x}) = 0$, where H^n is the usual Hamiltonian defined by

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

Enhanced Deep Fictitious Play

Solving for the NE of the game is equivalent to solving the N-coupled HJB equations of dimension $(3N + 1)$ defined in Equation (7). Due to the high dimensionality, this is a formidable numerical challenge. We overcome this through a deep learning methodology we call *Enhanced Deep Fictitious Play*, being broadly motivated by the method of fictitious play introduced by Brown [Bro51].

Deep Learning. Deep learning leverages a class of computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction [LBH15]. Deep neural networks are effective tools for approximating unknown functions in high-dimensional space. In recent years, we have witnessed noticeable success in a marriage of deep learning and computational mathematics to solve high-dimensional differential equations. Specifically, deep neural networks show strong capability in solving stochastic control and games [HJE18, HL22]. Below, we take a simple example to illustrate how a deep neural network is determined for function approximation.

Suppose we would like to approximate a map $y = f(x)$ by a neural network $\mathcal{NN}(x, \boldsymbol{w})$ in which one seeks to obtain appropriate parameters of the network, $\boldsymbol{w}$, through a process called training. This consists of minimizing a loss function that measures the discrepancies between the approximation and true values over the so-called training set $\{x_i\}_{i=1}^N$. Such a loss function has the general form

$$L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i), \mathcal{NN}(x_i, \boldsymbol{w})) + \lambda \mathcal{R}(\boldsymbol{w}),$$

where $\mathcal{R}(\boldsymbol{w})$ is a regularization term on the parameters. The first term $L_i(f(x_i), \mathcal{NN}(x_i, \boldsymbol{w}))$ ensures that the predictions of $\mathcal{NN}(x_i, \boldsymbol{w})$ match approximately the true value $f(x_i)$ on the training set $\{x_i\}_{i=1}^N$. Here, L_i could be a direct distance like the L^p norm or error terms derived from some complex simulations associated with $f(x_i)$ and $\mathcal{NN}(x_i, \boldsymbol{w})$. The hyperparameter λ characterizes the relative importance between the two terms in $L(\boldsymbol{w})$. To find an

optimal set of parameters $\boldsymbol{w}^*$, one solves the problem of minimizing $L(\boldsymbol{w})$ by the stochastic gradient descent (SGD) method [BCN18]. Regarding the architecture of $\mathcal{NN}(x, \boldsymbol{w})$, there is a wide variety of choices depending on the problem, for example fully connected neural networks, convolutional neural networks, recurrent neural networks, and transformers. In this work, we choose fully connected neural networks to approximate the solution and constructed the loss function by simulating the backward differential equations corresponding to the HJB equations.

Enhanced Deep Fictitious Play. Note that the HJB system (7) is difficult to solve due to the high dimensionality of the N-coupled equations. What if we could decouple the system to N separate equations, each of which is easier to solve? This is the central idea of *fictitious play*, where we update our approximations to the optimal policies of each player iteratively stage by stage. In each stage, instead of updating the approximations of all the players together by solving the giant system, we do it separately and parallelly. Each player solves for her own optimal policy assuming that the other players are taking their approximated optimal strategies from the last stage. Let us denote the optimal policy and corresponding value function of the single player n in stage m as $\alpha^{n,m}$ and $V^{n,m}$, respectively, and the collection of these two quantities for all the players as $\boldsymbol{\alpha}^m = (\alpha^{1,m}, ..., , \alpha^{N,m})$ and $\boldsymbol{V}^m = (V^{1,m}, ..., V^{N,m})$. Finally, let us denote the optimal policies and corresponding value functions for all the players except for player n as $\boldsymbol{\alpha}^{-n,m} = (\alpha^{1,m}, ..., \alpha^{n-1,m}, \alpha^{n+1,m}, ..., , \alpha^{N,m})$ and $\boldsymbol{V}^{-n,m} = (V^{1,m}, ..., V^{n-1,m}, V^{n+1,m}, ..., , V^{N,m})$, where $\alpha^{n,m}$ is a concatenation of lockdown policies and vaccination policies, *i.e.*, $(\ell^{n,m}, h^{n,m})$. At stage $m+1$, we can solve for the optimal policy and value function of player n given other players are taken the known policies $\boldsymbol{\alpha}^{-n,m}$ and the corresponding value $\boldsymbol{V}^{-n,m}$. The logic of fictitious play is shown in Figure 2, where players iteratively decide optimal policies in stage $m+1$, based on other players' optimal policies in stage m. This is slightly different than the usual simultaneous fictitious play, where the belief is described by the time average of past play and the distinction is further discussed in [HH20].

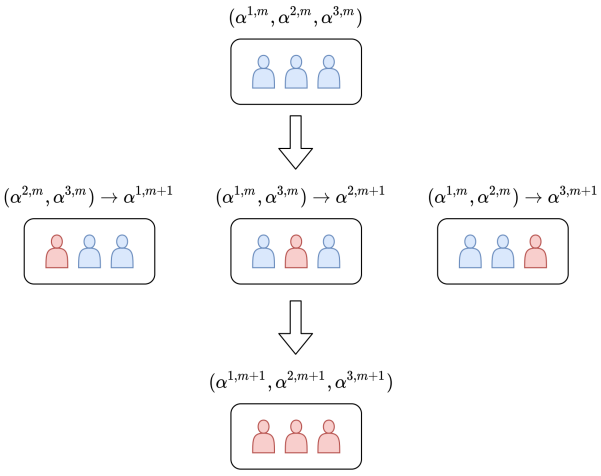

Figure 2: Schematic plot of fictitious play: each player derives optimal policies at stage $m+1$ assuming other players take optimal strategies at stage m.

The *Enhanced Deep Fictitious Play (DFP)* algorithm we have designed, built from the Deep Fictitious Play (DFP) algorithm [HH20], reduces time cost from $\mathcal{O}(M^2)$ to $\mathcal{O}(M)$ and memory cost from $\mathcal{O}(M)$ to $\mathcal{O}(1)$, with M as the total number of fictitious play iterations.

We illustrate one stage of enhanced deep fictitious play in Figure 3. At the $(m+1)^{th}$ stage, given the optimal policies $\boldsymbol{\alpha}^m$ at the previous stage, for $n = 1, \ldots, N$, the algorithm solves the following partial differential equations (PDEs),

$$\begin{aligned} &\partial_t V^{n,m+1} \\ &+ \inf_{\alpha^n \in [0,1]^2} H^n(t, \boldsymbol{x}, (\alpha^n, \boldsymbol{\alpha}^{-n,m})(t, \boldsymbol{x}), \nabla_{\boldsymbol{x}} V^{n,m+1}) \\ &\qquad + \frac{1}{2} \mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}} \mathrm{Hess}_{\boldsymbol{x}} V^{n,m+1} \Sigma(\boldsymbol{x})) = 0, \end{aligned} \tag{8}$$

with $V^{n,m+1}(T, \boldsymbol{x}) = 0$, and obtains the optimal

strategy of the $(m+1)^{th}$ stage:

$$\alpha^{n,m+1} = \arg\min_{\alpha^n \in [0,1]^2} H^n(t, \boldsymbol{x}, (\alpha^n, \boldsymbol{\alpha}^{-n,m})(t, \boldsymbol{x}), \nabla_{\boldsymbol{x}} V^{n,m+1}(t, \boldsymbol{x})).$$

For simplicity of notations, we omit the stage number m in the superscript in the following discussions. The solution to Equation (8) is approximated by solving the equivalent backward stochastic differential equations (BSDEs) using neural networks [HJE18]:

$$\begin{cases} \boldsymbol{X}_t^n = \boldsymbol{x}_0 + \int_0^t \mu^n(s, \boldsymbol{X}_s^n; \boldsymbol{\alpha}^{-n}(s, \boldsymbol{X}_s^n))\,\mathrm{d}s \\ \qquad + \int_0^t \Sigma(\boldsymbol{X}_s^n)\,\mathrm{d}\boldsymbol{W}_s, \qquad (9) \\ Y_t^n = \int_t^T g^n(s, \boldsymbol{X}_s^n, Z_s^n; \boldsymbol{\alpha}^{-n}(s, \boldsymbol{X}_s^n))\,\mathrm{d}s \\ \qquad - \int_t^T (Z_s^n)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{W}_s. \qquad (10) \end{cases}$$

The nonlinear Feynman-Kac formula [PP92] yields:

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

Here μ^n and g^n are derived by rewriting (8) to $\partial_t V^n + \frac{1}{2}\mathrm{Tr}(\Sigma(\boldsymbol{x})^{\mathrm{T}} \mathrm{Hess}_{\boldsymbol{x}} V^n \Sigma(\boldsymbol{x})) + \mu^n(t, \boldsymbol{x}; \boldsymbol{\alpha}^- n) \cdot \nabla_{\boldsymbol{x}} V^n + g^n(t, \boldsymbol{x}, \Sigma(\boldsymbol{x})^{\mathrm{T}} \nabla_{\boldsymbol{x}} V^n; \boldsymbol{\alpha}^{-n}) = 0$; see [XBH+22, Appendix A.2]. Notice that, we parametrized V^n by neural networks (denote as V-networks) so Y_t^n and Z_t^n could all be computed by a function of V-networks. The loss function to update the V-network is constructed by simulating the BSDE along the time axis and penalizing the difference between the true terminal value and the simulated terminal value based on neural networks of Y.

In Enhanced DFP, we further parameterize α^n (denote as α-networks). In each stage, the loss function with respect to the V-network and the α-network of player n is defined by the weighted sum of two terms: the loss related to BSDE (9)–(10) to approximate its solution and the error of approximating the optimal strategy α^n by α-networks. We denote $\|\cdot\|_2$ as the 2-norm, α^n and $\tilde{\alpha}^n$ as the derived and approximated optimal control of player n in the current stage, $\tilde{\boldsymbol{\alpha}}^{-n} = (\tilde{\alpha}^{1,m}, ..., \tilde{\alpha}^{n-1,m}, \tilde{\alpha}^{n+1,m}, ..., \tilde{\alpha}^{N,m})$ as the collection of approximated optimal controls from the last stage except player n, and τ as a hyperparameter balancing the two types of errors in the loss function. Then the Enhanced DFP solves

$$\begin{aligned} \inf_{Y_0^n, \tilde{\alpha}^n, \{Z_t^n\}_{0 \le t \le T}} & \mathbb{E}(|Y_T^n|^2 \\ & + \tau \int_0^T \|\alpha^n(s, \boldsymbol{X}_s^n) - \tilde{\alpha}^n(s, \boldsymbol{X}_s^n)\|_2^2\,\mathrm{d}s) \\ \text{s.t. } \boldsymbol{X}_t^n &= \boldsymbol{x}_0 + \int_0^t \mu^n(s, \boldsymbol{X}_s^n; \tilde{\boldsymbol{\alpha}}^{-n}(s, \boldsymbol{X}_s^n))\,\mathrm{d}s \\ & + \int_0^t \Sigma(\boldsymbol{X}_s^n)\,\mathrm{d}\boldsymbol{W}_s, \\ Y_t^n &= Y_0^n - \int_0^t g^n(s, \boldsymbol{X}_s^n, Z_s^n; \tilde{\boldsymbol{\alpha}}^{-n}(s, \boldsymbol{X}_s^n))\,\mathrm{d}s \\ & + \int_0^t (Z_s^n)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{W}_s, \\ \alpha^n(s, \boldsymbol{X}_s^n) &= \arg\min_{\beta^n} H^n(s, \boldsymbol{X}_s^n, (\beta^n, \tilde{\boldsymbol{\alpha}}^{-n})(s, \boldsymbol{X}_s^n), Z_s^n). \end{aligned} \qquad (11)$$

In each stage, there are two types of optimal strategies for player n: 1. the *derived* optimal strategy α^n by solving $\arg\min_{\beta^n} H^n$ in the last equation of (11); 2. the *approximated* optimal strategy $\tilde{\alpha}^n$ also known as α-networks for reducing the non-trivial cost of evaluating α^n. Take stage $m+1$ as an example, $\alpha^{n,m+1}$ depends on players' last stage optimal policies $\boldsymbol{\alpha}^{-n,m}$ which in turn depends on $\boldsymbol{\alpha}^{-n,m-1}$. The evaluation of the current stage strategy $\alpha^{n,m+1}$ actually requires the recursive iteration of optimal strategies from all previous stages. Enhanced DFP unblocks the computation bottleneck by introducing approximated optimal strategy $\tilde{\alpha}^n$, which approximates α^n. Although representing α^n with a neural network $\tilde{\alpha}^n$ introduces approximation errors, it allows us to efficiently access the proxy of the optimal strategy $\boldsymbol{\alpha}^{-n}$ in the current 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. This is the key factor that Enhanced Deep Fictitious Play addresses leading to reduction in both time and memory complexity compared to Deep Fictitious Play.

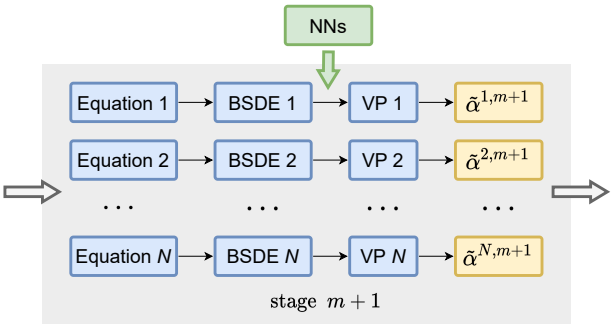

Figure 3: Illustration of one stage of *enhanced deep fictitious play*. At the $(m+1)^{th}$ stage, one needs to solve the PDEs (8), which is approximated by solving the BSDEs (9)-(10). Then with the help of neural networks, one solves the variational problem (VP) given by Equation (11) to get the optimal strategy.

To implement the loss function defined in (11), we discretize and simulate the BSDE by Euler's method with a partition on a time interval $[0, T]$. The expectation in the loss function is approximated by Monte Carlo samples of $\boldsymbol{W}_s$ in the stochastic process. Then, we use the stochastic gradient descent method to update V^n and $\tilde{\alpha}^n$ in the current stage for player n. In parallel, we update the V-network and α-network for each player. The updated networks of each player will be observable for other players in future stages.

A Regional COVID-19 Study

In this section, we apply the multi-region stochastic SEIR model (2)–(6) to analyze optimal COVID-19 policies in three adjacent states: New York, New Jersey and Pennsylvania. This case study focuses on 180 days starting from 03/15/2020, and solves for the optimal policies of the three states corresponding to Nash Equilibrium by the Enhanced Deep Fictitious Play Algorithm. We denote New York (NY) as region 1, New Jersey (NJ) as region 2 and Pennsylvania (PA) as region 3, with population $P_1 = 19.54$ million, $P_2 = 8.91$ million, and $P_3 = 12.81$ million during the case study time range, respectively. We assume that (a) 90% of any state's population resides in their own state at a given time; (b) the remaining population(travellers) visit the other states at an equal chance; (c) there is no travel outside of the three states, that is, NY-NJ-PA is a closed system. The parameters in (2)-(6) are estimated based on the above assumptions and public information about COVID-19: $\beta = 0.17, \kappa = \frac{0.65\%}{13}, \lambda = \frac{1}{13}, \gamma = \frac{1}{5}, p = 228.7 \times 10^{-5}, c = 73300/13$. Other parameters in the model are chosen at: $r = 0, w = 172.6, \chi = 1.96 \times 10^6$. The hyperparameters, θ and a, which represent policy effectiveness and planners' views on the death of human beings will change the optimal policies. For results including vaccination controls, we point to [OS21], which considers an optimal control problem for vaccines and testing of COVID-19. However, in the time period we study, vaccination was not available, so we ignore the health policy h and mainly solve for the lockdown policy of each state.

Figure 4 shows the Nash equilibrium policies in NY, NJ, and PA in a setting where the policy effectiveness is $\theta = 0.99$, i.e., 99% of the residents will follow the lockdown orders. The weight parameter quantifying each planner's view is $a = 100$, i.e., each governor values human life 100 times more than the economic value of a human life. The resulting Nash equilibrium of this scenario corresponds to the planners taking action at an early stage by implementing strict lockdown policies and later relaxing the policy as the infections improve. In the end, the percentage of Susceptible, Exposed, Infectious, and Removed stays almost constant. The pandemic will be significantly mitigated in this scenario of proactive lockdown for both planners and residents. As a comparison, [XBH+22, Figure 2] illustrates a scenario of how the pandemic gets out of control if governors show inaction or issue mild lockdown policies.

Acknowledgements

The contents of this article are based on the authors' previous publication [XBH+22]. R.H. was partially supported by the NSF grant DMS-1953035, the Faculty Career Development Award, the Research Assistance Program Award, the Early Career Faculty Acceleration funding, and the Regents' Junior Faculty

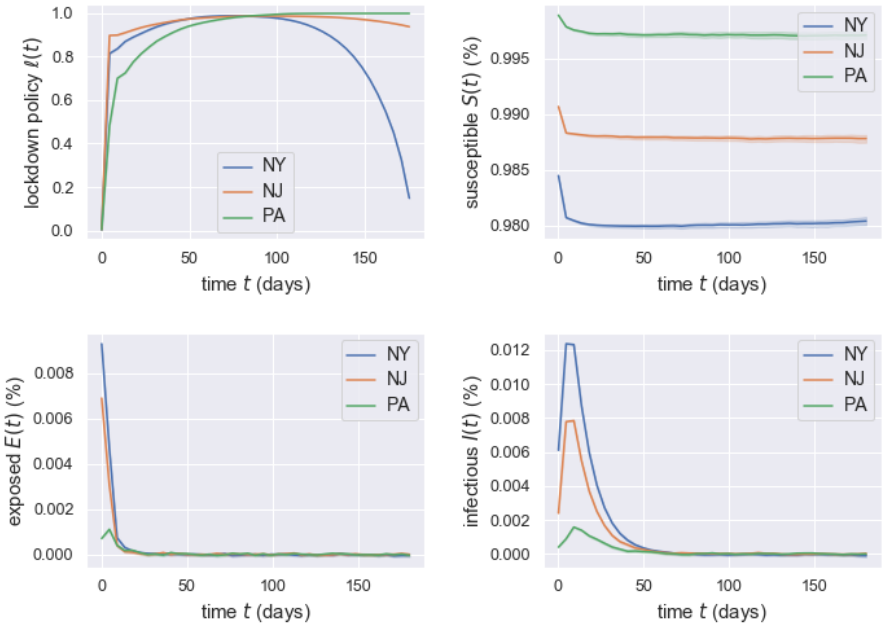

Figure 4: Plots of optimal policies (top-left), Susceptibles (top-right), Exposed (bottom-left) and Infectious (bottom-right) for three states: New York (blue), New Jersey (orange) and Pennsylvania (green). The shaded areas depict the mean and 95% confidence interval over 256 sample paths. Choices of parameters are $a = 100$ and $\theta = 0.99$.

Fellowship at University of California, Santa Barbara. H.D.C. gratefully acknowledges partial support from the NSF grant DMS-1818821.

References

[AAL20] Fernando E Alvarez, David Argente, and Francesco Lippi, *A simple planning problem for covid-19 lockdown*, National Bureau of Economic Research, 2020.

[All08] Linda J. S. Allen, *An introduction to stochastic epidemic models*, Mathematical epidemiology, 2008, pp. 81–130. MR2428373

[AZM+20] Zohreh Abbasi, Iman Zamani, Amir Hossein Amiri Mehra, Mohsen Shafieirad, and Asier Ibeas, *Optimal control design of impulsive SQEIAR epidemic models with application to COVID-19*, Chaos Solitons Fractals **139** (2020), 110054, 20. MR4121530

[BCN18] Léon Bottou, Frank E. Curtis, and Jorge Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev. **60** (2018), no. 2, 223–311. MR3797719

[Ber60] Daniel Bernoulli, *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir*, Histoire de l'Acad., Roy. Sci.(Paris) avec Mem (1760), 1–45.

[Bro51] George W. Brown, *Iterative solution of games by fictitious play*, Activity Analysis of Production and Allocation, 1951, pp. 374–376. MR0056265

[CD18] René Carmona and François Delarue, *Probabilistic theory of mean field games with applications. I*, Probability Theory and Stochastic Modelling, vol. 83, Springer, Cham, 2018. Mean field FBSDEs, control, and games. MR3752669

[GB97] MD Grmek and C Buchet, *The beginnings of maritime quarantine*, Man, health and the sea. Paris: Honoré Champion (1997), 39–60.

[HH20] J. Han and R. Hu, *Deep fictitious play for finding Markovian Nash equilibrium in multi-agent games*, Proceedings of the first mathematical and scientific machine learning conference (MSML), 2020, pp. 107:221–245.

[HJE18] Jiequn Han, Arnulf Jentzen, and Weinan E, *Solving high-dimensional partial differential equations using deep learning*, Proc. Natl. Acad. Sci. USA **115** (2018), no. 34, 8505–8510. MR3847747

[HL22] Ruimeng Hu and Mathieu Lauriere, *Recent developments in machine learning methods for stochastic control and games*, Available at SSRN: https://ssrn.com/abstract=4096569 (2022).

[KL89] Richard J. Kryscio and Claude Lefèvre, *On the extinction of the S-I-S stochastic logistic epidemic*, J. Appl. Probab. **26** (1989), no. 4, 685–694. MR1025386

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, *Deep learning*, nature **521** (2015), no. 7553, 436–444.

[LHL87] Wei Min Liu, Herbert W. Hethcote, and Simon A. Levin, *Dynamical behavior of epidemiological models with nonlinear incidence rates*, J. Math. Biol. **25** (1987), no. 4, 359–380. MR908379

[Nas51] John Nash, *Non-cooperative games*, Ann. of Math. (2) **54** (1951), 286–295. MR43432

[NBN22] Peter Nordström, Marcel Ballin, and Anna Nordström, *Risk of sars-cov-2 reinfection and covid-19 hospitalisation in individuals with natural and hybrid immunity: a retrospective, total population cohort study in sweden*, The Lancet Infectious Diseases **22** (2022), no. 6, 781–790.

[OS21] Alberto Olivares and Ernesto Staffetti, *Optimal control-based vaccination and testing strategies for covid-19*, Computer Methods and Programs in Biomedicine **211** (2021), 106411.

[PP92] É. Pardoux and S. Peng, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, Stochastic partial differential equations and their applications (Charlotte, NC, 1991), 1992, pp. 200–217. MR1176785

[TBV05] Elisabetta Tornatore, Stefania Maria Buccellato, and Pasquale Vetro, *Stability of a stochastic SIR system*, Physica A: Statistical Mechanics and its Applications **354** (2005), 111–126.

[XBH+22] Yao Xuan, Robert Balkin, Jiequn Han, Ruimeng Hu, and Hector 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), 2022, pp. 145:987–1012.