

Optimal incentives to mitigate epidemics: A Stackelberg mean field game approach

Alexander Aurell* René Carmona* Gökçe Dayanıklı* Mathieu Laurière*

Abstract

Motivated by the models of epidemic control in large populations, we consider a Stackelberg mean field game model between a principal and a mean field of agents whose states evolve in a finite state space. The agents play a non-cooperative game in which they control their rates of transition between states to minimize an individual cost. The principal influences the nature of the resulting Nash equilibrium through incentives so as to optimize its own objective. We analyze this game using a probabilistic approach. We then propose an application to an epidemic model of SIR type in which the agents control the intensities of their interactions, and the principal is a regulator acting with non pharmaceutical interventions. To compute the solutions, we propose an innovative numerical approach based on Monte Carlo simulations and machine learning tools for stochastic optimization. We conclude with numerical experiments illustrating the impact of the agents' and the regulator's optimal decisions in two specific models: a basic SIR model with semi-explicit solutions and a more complex model with a larger state space.

Keywords. SIR epidemics, Mean field game, Stackelberg equilibrium, Machine learning

AMS subject classifications. 92D30, 49N90, 91A13, 91A15, 62M45.

Acknowledgments. This work was done with the support of NSF DMS-1716673, ARO W911NF-17-1-0578, and AFOSR # FA9550-19-1-0291.

1 Introduction

Non pharmaceutical interventions such as the reduction of social interactions are powerful measures to limit the spread of an ongoing epidemic. Containment and suppression of disease spread are crucial factors in order to avoid overwhelming the health care system. However, even in the midst of pandemics, some individuals still refuse to comply with guidelines such as social distancing or mask wearing. From a global perspective, this could push the equilibrium behavior of the population to exceed the limits of the health care system. For this reason, responsible authorities have a keen interest in the design of incentive systems that are acceptable to individuals and sufficiently strong to induce them to successfully combat the epidemic.

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (aaurell@princeton.edu, rcarmona@princeton.edu, gokced@princeton.edu, lauriere@princeton.edu).

In a mathematical model, we would like the decision maker to take both the global state of the society and the individuals' behaviors into account before deciding on an incentive policy. The intractability of large interacting dynamical systems usually prevents that kind of analysis. In this work, we analyze a Stackelberg game between a principal agent representing the regulatory authority, and a field of individuals providing the societal response to the principal's policy. We use the probabilistic approach to mean field games because it provides both a macroscopic description of the state of the population and a microscopic analysis of the behavior of a representative single individual.

Mean Field Game (MFG) models study the equilibrium between a representative player and the distribution of the other players' states and actions. Mean field equilibria are simpler to identify and compute than equilibria of large populations. Moreover, they provide approximate Nash equilibria for certain games with a large but finite number of players. The framework has found numerous applications, from the analysis of growth models in macro-economics, to crowd motion and energy production. Here, we investigate an application to epidemic control.

Compartmental models in epidemic research are in many cases large population limits of interacting Markov chains. Each Markov chain represents the individual's state of health, the transitions between states occurring with rates depending on the global state of the population, the proportions of individuals in different states to be more specific. This form of interaction is clearly in the purview of mean field models. Incorporating in the model the opportunity for individuals to choose their behaviors and control their contributions to the spread of the disease, the interacting system can be analyzed as a MFG. However, a natural choice of controls yields what is often called an extended MFG, where players interact not only through the distribution of their states, but instead through the joint distribution of their states and actions. Next, we give a hands-on example of the extended aspect of our model.

1.1 The SIR extended MFG with contact factor control

In order to provide a motivating example, we consider the simplest compartmental model in epidemics, the classical SIR model. First, we outline how to construct it as a large population limit. Secondly, we comment on why an extended MFG formulation is relevant for the large population limit problem if players use what we will call a contact factor control to reduce the risk of disease spread.

Before moving to the example, we need to introduce some notation. Consider N individuals, each of whom transitions between the states Susceptible (S), Infected (I), and Removed (R). An individual in state R has either gained permanent immunity, or is deceased. Denote the state of individual $j \in \{1, \dots, N\}$ at time t by X_t^j , and let $p_t^N = (p_t^N(S), p_t^N(I), p_t^N(R)) := (\frac{1}{N} \sum_{j=1}^N \mathbb{1}_i(X_t^j))_{i \in \{S, I, R\}}$ be the vector of proportions of individuals in each state, in other words, the empirical distribution of the state at time t .

A susceptible individual might meet infected individuals, possibly resulting in disease transmission. Encounters occur pairwise and randomly throughout the population. Their intensity is denoted by $\beta > 0$. The number of encounters with infected individuals during a small time interval $[t - \Delta t, t)$ is proportional to the the proportion of the population in state I at t . Hence the transition from state S to I happens with intensity $\beta p_t^N(I)$. Upon infection an individual starts the path to recovery. The transition from state I to R happens after an exponentially distribution time with rate γ . The state R is absorbing. To summarize, the transition rate matrix, which is common to all agents of the population, is at time t ,

$$Q(p_t^N) = \begin{bmatrix} -\beta p_t^N(I) & \beta p_t^N(I) & 0 \\ 0 & -\gamma & \gamma \\ 0 & 0 & 0 \end{bmatrix}. \quad (1.1)$$

As $N \rightarrow \infty$, $(p_t^N)_t$ converges in probability to the unique solution to

$$\dot{p}_t = p_t Q(p_t), \quad p_0 = p^0, \quad (1.2)$$

if the initial configuration is sampled from a symmetric probability measure with marginals equal to p^0 , a classical result found for example in [31]. Scaling p_t by a population size N , *i.e.*, letting $Np_t =: (S(t), I(t), R(t))$. we retrieve the standard formulation of the SIR model,

$$\begin{cases} \dot{S}(t) = -\frac{\beta}{N}I(t)S(t), & S(0) = Np^0(S), \\ \dot{I}(t) = \frac{\beta}{N}I(t)S(t) - \gamma I(t), & I(0) = Np^0(I), \\ \dot{R}(t) = \rho\gamma I(t), & R(0) = Np^0(R). \end{cases} \quad (1.3)$$

Now, let us assume that each individual has the option to control the intensity they seek or try to avoid interacts with others. Instances occur when individuals try to lower the risk of disease transmission by, *e.g.*, avoiding to ride public transportation at congested hours, shopping online, or wearing protective equipment.

The probability of the spread of the disease is likely to be a non-linear function of the joint effort of the individuals that interact. Say, for example, that two individuals meet and both have the option to wear a protective face mask. The absolute decrease in risk of transmission is not necessarily equal for each additional mask that is worn. Motivated by this observation, we assume that the individuals' efforts to reduce spread affect the probability of transmission in a multiplicative way: in each encounter the probability of disease spread is scaled by each of the agents effort. We view an individual's effort to meet someone as their control. We often call it their *contact factor* because this effort enters as a factor in the contact rate between individuals of specific states.

Assuming that the meeting frequency is β , that the pairing is random, that the disease spreads from infected agents to susceptible, and that the spread probability is scaled by the effort intensity of the search for meetings, the transition rate for individual j , currently susceptible, to the state of infected is

$$\beta\alpha_t^j \frac{1}{N} \sum_{k=1}^N \alpha_t^k \mathbf{1}_I(X_{t-}^k), \quad (1.4)$$

where α_t^k denotes the (contact factor) action of individual $k \in \{1, \dots, N\}$ at time t , selected from set A of admissible actions. Along the lines of the heuristics of MFG theory, we anticipate that in an appropriate approximation of our interacting system in the limit $N \rightarrow \infty$, the representative agent transitions from susceptible to infected with rate

$$\beta\alpha_t \int_A a\rho_t(da, I), \quad (1.5)$$

where ρ_t is the joint distribution of action and state of the representative agent in a suitable probability space (rigorously defined in the next section). The joint action-state distribution is often referred to as the extended mean field in the MFG literature. To summarize, the representative agent transitions between states S , I , and R according to the rate matrix $Q(t, \alpha_t, \rho_t)$,

$$Q(t, \alpha, \rho) = \begin{bmatrix} \cdots & \beta\alpha_t \int_A a\rho_t(da, I) & 0 \\ 0 & \cdots & \gamma \\ \eta & 0 & \cdots \end{bmatrix}, \quad (1.6)$$

where $\beta, \gamma, \eta \in \mathbb{R}_+$ are non-controlled constants, and as usual, the diagonal terms \cdots should be replaced by the negative of the sum of the entries in the same row. See Fig. 1 for a diagram of the dynamics.

The representative agent is incentivized by a regulator to choose their contact factor close to a level determined by the regulator. The level and incentive can vary between the susceptible, infected, and recovered parts of the population, as the state of an agent naturally influences their contribution to the

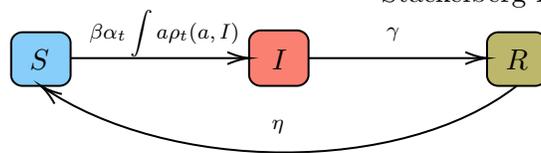


Figure 1: SIR model with extended mean-field interactions corresponding to the Q -matrix (1.6). overall societal risks in an epidemic (which the regulator aims to mitigate). Moreover, the representative agent faces a cost of inconvenience for being sick. In particular, let us consider a model where a representative agent pays per unit of time a running cost given as

$$\frac{c_\lambda}{2} \left(\lambda_t^{(S)} - \alpha_t \right)^2 \mathbb{1}_S(x) + \left(\frac{1}{2} \left(\lambda_t^{(I)} - \alpha_t \right)^2 + c_I \right) \mathbb{1}_I(x) + \frac{1}{2} \left(\lambda_t^{(R)} - \alpha_t \right)^2 \mathbb{1}_R(x), \quad (1.7)$$

where $c_\lambda, c_I \in \mathbb{R}_+$ are constants, $\lambda^{(\cdot)}$ are the socialization levels recommended by the regulator, and α_t is the contact factor of the representative player at time t .¹ In this model, which we will revisit in the section on numerical experiments, a $\lambda^{(S)}$ valued close to 0 can be interpreted as a recommendation for low levels of social interactions (or more generally a high level of cautiousness through non-pharmaceutical interventions, such as hand cleaning, lockdown, mandatory mask wearing etc.) for all susceptible individuals. Conversely, $\lambda^{(S)}$ approximately equal to 1 amounts to recommending the regular level of social interactions (i.e. no restriction). Finally, we assume that the agent also receives a terminal utility $U(\xi)$ depending on the agent's behavior during the time interval $[0, T]$, payed by the regulator as an incentive to follow the recommended socialization protocols.

The problem is then two-fold. First, the regulator announces a policy (λ, ξ) so as to minimize an objective function which involves the state of the population (*e.g.*, the proportion of infected people). The minimization is constrained, not all policies will be accepted by the population and the regulator is optimizing only over acceptable ones. Acceptable here does not mean a complete commitment to following the recommendations, but rather accepting the penalty structure for deviations. The way the population reacts to a given policy (λ, ξ) is through a Nash equilibrium in which each agent tries to optimize their own individual cost induced by (1.7) and their terminal payment utility $U(\xi)$. If the representative agent's expected cost at the Nash equilibrium is above some threshold, the policy is deemed unacceptable and the population rejects the policy altogether (the policy is not feasible). The regulator, to find an optimal feasible policy, needs to understand how the population reacts to each policy. The two nested problems comprise a so-called Stackelberg game: the regulator's problem is a constrained optimization problem driven by the Nash equilibrium of the player population.

The two components λ and ξ of the regulator's policy play different roles. The process λ is used to incentivize the agents to adopt a certain cautiousness level over time. Practically, deviations from λ are penalized. The terminal payment ξ is used to reward participation in the incentive structure. Later, we will see how the regulator decides on a terminal reward or payment such that the representative agent does not reject the proposed incentive scheme altogether. That is, the terminal payment's role is to make the policy feasible in the sense that it is a reasonable compensation to the agent for complying with the cautiousness level recommendations.

¹The running cost (1.7) is far from the only possible model choice within the framework presented in this paper, but one that facilitates evaluation of the performance of the proposed numerical method. In a more general setting, the running cost could be a function depending also on the extended mean field interactions.

1.2 Related literature

1.2.1 Discrete state space MFG

The behavior of the population of agents, amongst whom disease spread takes place, is in this paper modeled by a discrete space MFG. MFGs were first developed for continuous state space [33, 34, 26]. Soon after works on discrete state spaces followed [22, 30, 23]. Amongst the many contributions to the field of discrete state MFGs we note the minor-major player model [6], the probabilistic approach [9], the master equation approach [2], and the extended game [8]. Mean-field optimal control, risk-sensitive control, and zero-sum games are treated in [15, 13, 14] which cover cases of unbounded jump intensities.

1.2.2 Compartmental models and MFG in epidemics

Games and optimal control in compartmental models have been studied intensively for a long time. This literature review focuses on other work within the mean-field approach, which has gained attention in the last decade. Efforts to model the control of disease spread range from strategies for social contacts to vaccination.

Our work falls within a category of models where agents attempt to suppress the risk of disease spread. In [19] a deterministic mean-field game is studied where the agents control the contact rate, which is proportional to the risk of disease spread. The agents are penalized if they get infected prior to some terminal time horizon, which introduces a stopping time component to the game similar to that in evacuation problems. A contact rate common to all agents and all states is found such that the agents are in an MFG equilibrium with the crowd. In [27] a Stackelberg game where the epidemic evolves in the population of agents, modeled as a MFG, according to a compartmental model is considered. The agents collaborate to find the best contact rate to suppress the epidemic. The compartmental models considered in the paper are stochastic and the uncertainty in the model is controlled by the principal through testing policies. Their goal is to mitigate the saturation of intensive care units. This problem was studied from the point of view of optimal control in [11], where numerical results show that it is optimal to isolate infected individuals so as to maintain a basic reproduction rate close to 1. In [12], SIR and SEIR² models where agents control the contact rate are studied and the author compares the MFG equilibrium, the socially optimal strategy, and unconstrained disease spread.

Vaccination is a powerful tool when available. However, it is not considered in this paper. With a MFG formulation of the SIR model, In vaccination strategies in a society of non-cooperative individuals are studied. The authors extend the model to include limited vaccination capacity [32], limited persistence [39], and vital dynamics [28].³ Vaccination has also been studied with MFG-based SIR models in [18, 21], their focus being the loss of efficiency in the mean-field game compared to optimal vaccination policies.

Spatial distance naturally mitigates the risks of the pandemic and [42] uses a mean-field type game to take the spatial features of disease spread into account (and many more features, *e.g.*, physical and social status of the agent). In [37], the authors consider three crowds, each corresponding to a state in the SIR model, which evolve spatially. The pandemic risks are mitigated by a central planner who controls of spatial velocity of the agents. The multi-population mean-field optimal control problem is studied.

1.2.3 Contract theory and Stackelberg MFG

Contract theory studies the interaction of a principal and an agent, where the former proposes a contract to the latter, who decides whether or not they should work for the principal and receive a reward. The

²The SEIR model includes the additional “Exposed” state, modeling the incubation period before the agent transitions to the “Infected” state.

³Persistence here refers to immunity to reinfection and when this is limited the agents will eventually become vulnerable again. The SIR model with vital dynamics includes births and deaths.

principal tries to anticipate the decision of the agent and to design an attractive contract while still trying to maximize their profit. Solutions to this type of problems are typically studied using the concept of Stackelberg equilibrium. In [25], continuous time method is used to study this type of problems, and in [40, 41] dynamic programming and martingale optimality principles are used to characterize the solution in the framework of optimal control theory. These ideas are generalized in [16]. In [17] the solution in a general class of principal-agent problem is characterized by the stochastic maximum principle.

In the context of MFGs, problems with a principal and a mean-field of agents have been studied in [20] in the continuous state space setting and in [7] for finite state spaces. The theory has been extended in several directions, including problems with delayed information [3]. This type Stackelberg mean field models have found applications for instance to advertising [38], where the principal plays the role of the advertiser and the population of agents decides whether they want to buy a product. Stackelberg equilibria with a mean-field of agents have also been applied in the context of epidemic containment: the aforementioned [7] proposes an application with two cities where the agents can move between cities and the principal can influence the quality of healthcare, while in [27] the authors consider a model where the principal can choose a tax policy and a testing policy.

1.3 Contributions and paper structure

The scientific contribution of this work is two-fold. Firstly, we move beyond current theory and consider a Stackelberg game between a principal and an extended MFG. A common assumption in extended MFGs is that any dependency on the joint distribution of action and state only involves dependencies on the marginal distributions. We avoid this assumption in order to capture the epidemiological aspects outlined in Section 1.1. The trade-off is that, at some points in the paper, we need to make assumptions about the existence and uniqueness of mean field Nash equilibria. We work in the weak probabilistic formulation of the problem and we allow the player’s action to depend on their state. Our numerical experiments show that contact factors do differ between the compartments of the population. To the best of our knowledge, compartmental models with applications towards epidemics which incorporate at the same time a non-cooperative population and a regulator have not yet been suggested or studied in the literature.

Secondly, we propose an innovative numerical scheme based on neural networks, and validate its performance on simple examples for which we can derive semi-explicit solutions. To obtain a problem amenable to numerical treatment by optimization procedures, we first rewrite the principal’s problem under the constraint of the mean field Nash equilibrium as an optimal control problem with two forward stochastic equations. Then, the numerical scheme relies on the approximation of the population by an interacting particle system and the approximation of the controls by neural networks, including the principal’s policy. The optimization of the principal’s cost is then performed using a variant of stochastic gradient descent to update the neural networks’ parameters.

We carry out multiple numerical experiments studying model characteristics and policy impact. As a first step towards understanding how a regulator should design containment policies, we test the population’s reaction to various policies. We show that when the agents minimize their own cost and behave as in a Nash equilibrium, they adopt some level of cautiousness, which reduces the severity of the epidemic compared to an unconstrained *free spread* scenario. Moreover, we show that taking early action (*e.g.*, deciding on an early lockdown) has a bigger impact on disease spread than a strategic action taken later. In a second numerical test, we solve the full Stackelberg game problem (where the regulator optimizes over the policies to minimize its own cost) for both the SIR-based example of Section 1.1 and an extended model with two more additional states for the agents (*E*: Exposed and *D*: Deceased). For the latter, we show that if agents are not feeling safe enough, they are able to rationally choose lower contact levels than the recommended levels by the regulator.

The rest of the paper is structured as follows. In Section 2 the Stackelberg game between a principal and a non-cooperative population is introduced and analyzed. In Section 3 the details of the numerical

approach are presented. Finally, Section 4 contains the evaluation of the numerical method and further simulations. All proofs have been postponed to appendices.

2 The model

2.1 Preliminaries

We adopt the following notation throughout the paper: m is a finite integer corresponding to the number of states, $E := \{e_1, \dots, e_m\}$ is a state space where $e_i \in \mathbb{R}^m$ is the basis vector in direction i , $A := [0, 1]$ is an action space, and $\mathcal{R} := \mathcal{P}(A \times E)$ is the set of Borel probability measures on $A \times E$. We endow A with the Euclidean metric $|\cdot|$, E with a bounded discrete metric, and $A \times E$ with the 1-product metric. We will identify the set $\mathcal{P}(E)$ with the m -dimensional simplex and use the Euclidean metric $\|\cdot\|$ to measure distances on $\mathcal{P}(E)$ (the choice of metric on $\mathcal{P}(E)$ is irrelevant since all metrics derived from norms on $\mathcal{P}(E)$ are equivalent). We endow \mathcal{R} with the 1-Wasserstein metric $W_{\mathcal{R}}$ which is well-defined on \mathcal{R} since $A \times E$ is compact.

Let $T > 0$ be a constant corresponding to a finite time horizon. Let Λ be the set of measurable \mathbb{R}_+^m -valued functions with domain $[0, T]$ and let $M(\mathcal{R})$ and $M(\mathcal{P}(E))$ be the set of measurable mappings from $[0, T]$ to \mathcal{R} and to $\mathcal{P}(E)$, respectively. Let $Q : [0, T] \times A \times \mathcal{R} \mapsto \mathbb{R}^{m \times m}$ be a bounded measurable function such that $Q(t, a, \rho)$ is a transition rate matrix, also called Q -matrix,⁴ for all $(t, a, \rho) \in [0, T] \times A \times \mathcal{R}$.

A process $(X_t)_{t \in [0, T]}$ will in short-hand be denoted \mathbf{X} . Let Ω be the space of càdlàg functions $\omega : [0, T] \rightarrow E$ and from now on let \mathbf{X} be the canonical process, $X_t(\omega) = \omega(t)$. Denote by $\mathbb{F} := (\mathcal{F}_t)_{t \in [0, T]}$ the natural filtration generated by \mathbf{X} , with $\mathcal{F}_t := \sigma(\{X_s, s \leq t\})$ and $\mathcal{F} := \mathcal{F}_T$, and by \mathbb{A} the collection of \mathbb{F} -predictable processes α with values in A . For any probability measure \mathbb{Q} on (Ω, \mathcal{F}) we denote by $\mathbb{E}^{\mathbb{Q}}$ expectation under \mathbb{Q} .

On $(\Omega, \mathbb{F}, \mathcal{F})$ we consider the probability measure \mathbb{P} under which the law of X_0 is $p^0 \in \mathcal{P}(E)$ and \mathbf{X} is a continuous time Markov chain with transition rate from e_i to e_j equal to 1 if $(i, j) \in G \subset \{1, \dots, m\}^2$, otherwise zero. Here G represents a graph of states on which a typical agent evolves. Denote the corresponding Q -matrix by Q^0 . We let, for $i = 1, \dots, m$,

$$\psi(e_i) := \text{diag}(Q^0 e_i) - Q^0 \text{diag}(e_i) - \text{diag}(e_i) Q^0, \quad t \in [0, T], \quad (2.1)$$

and let $\psi_t = \psi(X_{t-})$. Denote by ψ_t^+ the Moore-Penrose generalized inverse of the matrix ψ_t . Expectation under \mathbb{P} is abbreviated to \mathbb{E} .

We denote by \mathcal{H}^2 the set of \mathbb{F} -adapted and real-valued càdlàg processes \mathbf{Y} such that $\mathbb{E}[\int_0^T Y_t^2 dt] < +\infty$ and by \mathcal{H}_X^2 the set of \mathbb{F} -adapted and \mathbb{R}^m -valued left-continuous processes \mathbf{Z} such that $\mathbb{E}[\int_0^T \|Z_t\|_{X_{t-}}^2 dt] < +\infty$. The seminorms $\|\cdot\|_{e_i}$, $i = 1, \dots, m$, and the stochastic seminorm $\|\cdot\|_{X_{t-}}$ are defined by

$$\|z\|_{e_i}^2 := z^* \psi(e_i) z, \quad \|z\|_{X_{t-}}^2 := z^* \psi_t z, \quad z \in \mathbb{R}^m. \quad (2.2)$$

Hereinafter we use a superscript $*$ to denote the transpose of a vector or a matrix.

2.2 The Stackelberg extended MFG in a general setting

We consider a society made up of a population of non-cooperative players and one principal agent. We begin by focusing on the game between the members of the population. As is common in the MFG paradigm, a representative player takes the role of any individual in the population. Given knowledge of how the population and the principal agent act over time, the representative player optimizes their cost functional.

⁴That is, $q(t, i, j, a, \rho) \geq 0$ for all $1 \leq i, j, \leq m$ and $\sum_{j \neq i} q(t, i, j, a, \rho) = -q(t, i, i, a, \rho)$, where $q(t, i, j, a, \rho)$ is the element element at row i and column j of $Q(t, a, \rho)$.

To use strategy $\alpha = (\alpha_t)_{t \in [0, T]} \in \mathbb{A}$ the representative player pays the expected total cost

$$J^{\lambda, \xi}(\alpha, \rho) := \mathbb{E}^{\mathbb{Q}^{\alpha, \rho}} \left[\int_0^T f(t, X_t, \alpha_t, \rho_t; \lambda_t) dt - U(\xi) \right], \quad (2.3)$$

where (λ, ξ) is the principal's policy choice, $f : [0, T] \times E \times A \times \mathcal{R} \rightarrow \mathbb{R}$ is a running cost which depends on the policy λ , $\rho = (\rho_t)_{t \in [0, T]} \in M(\mathcal{R})$ is a flow of measures in \mathcal{R} representing the joint state-control distribution in the population, and $\mathbb{Q}^{\alpha, \rho}$ is a probability measure over (Ω, \mathcal{F}) . The notation will be our convention throughout the paper whenever there is no possibility for confusion. The canonical process \mathbf{X} appearing in (2.3) models the representative player's dynamics under the probability measure $\mathbb{Q}^{\alpha, \rho}$. Under $\mathbb{Q}^{\alpha, \rho}$, \mathbf{X} is a pure-jump process with transition rate matrix $Q(t, \alpha_t, \rho_t)$ at time t .⁵

The agent is truly representative if their joint distribution of action and state agrees with the population. The consistency condition in MFG assures just this, see (ii) in definition 2.1 where the equilibrium notion in the population's problem is formalized.

Definition 2.1. *If the pair $(\hat{\alpha}, \hat{\rho}) \in \mathbb{A} \times M(\mathcal{R})$ satisfies*

- (i) $\hat{\alpha} = \arg \inf_{\alpha \in \mathbb{A}} J^{\lambda, \xi}(\alpha, \hat{\rho})$;
- (ii) $\forall t \in [0, T] : \hat{\rho}_t = \mathbb{Q}^{\hat{\alpha}, \hat{\rho}} \circ (\hat{\alpha}_t, X_t)^{-1}$,

we say that $(\hat{\alpha}, \hat{\rho})$ is a mean-field Nash equilibrium given the contract (λ, ξ) . We denote by $\mathcal{N}(\lambda, \xi)$ the set of such mean field Nash equilibria.

We state in Proposition 2.7 below that (under suitable assumptions) $(\hat{\alpha}, \hat{\rho}) \in \mathcal{N}(\lambda, \xi)$ if $(\mathbf{Y}, \mathbf{Z}, \hat{\alpha}, \hat{\rho}, \mathbb{Q})$ is a solution to the following equation⁶ under \mathbb{P}

$$\begin{cases} Y_t = U(\xi) + \int_t^T \hat{H}(s, X_{s-}, Z_s, \hat{\rho}_s) ds - \int_t^T Z_s^* d\mathcal{M}_s, \\ \mathcal{E}_t = 1 + \int_0^t \mathcal{E}_{s-} X_{s-}^* (Q(s, \hat{\alpha}_t, \hat{\rho}_s) - Q^0) \psi_s^+ d\mathcal{M}_s, \\ \hat{\rho}_t = \mathbb{Q} \circ (\hat{\alpha}_t, X_t)^{-1}, \quad \frac{d\mathbb{Q}}{d\mathbb{P}} = \mathcal{E}_T, \quad \hat{\alpha}_t = \hat{a}(t, X_{t-}, Z_t, \hat{\rho}_t), \end{cases} \quad (2.4)$$

where \hat{H} is the minimized Hamiltonian of the representative player and \hat{a} is the minimizer, defined in (2.9) below. The solution has a (λ, ξ) -dependence (entering the problem through $U(\xi)$ and the Hamiltonian) which we suppress to alleviate the notation.

The principal's problem is to find the policies that yield the most favorable configuration of minor players in terms of their cost. By using policies (λ, ξ) as incentives, the principal can modify the set of mean-field Nash equilibria $\mathcal{N}(\lambda, \xi)$ and hence exert influence over the population's behavior. In the sequel, unless otherwise mentioned, we consider the following class of policies for the principal.

Definition 2.2. *A policy (λ, ξ) is admissible if the deterministic mapping $\lambda \in \Lambda$, the real-valued random variable ξ is \mathcal{F} -measurable, and that $\mathcal{N}(\lambda, \xi)$ is a singleton. We denote the set of admissible policies by \mathcal{C} .*

⁵Existence of the measure $\mathbb{Q}^{\alpha, \rho}$ is granted by Girsanov Theorem under some conditions, see for example [8] and the references therein. The hypothesis on Q stated in Section 2.2 is strong enough for existence to hold.

⁶We define a solution to (2.4) in line with [8, Def. 2]: the tuple $(\mathbf{Y}, \mathbf{Z}, \alpha, \rho, \mathbb{Q})$ is a solution to the McKean-Vlasov BSDE (2.4) if $\mathbf{Y} \in \mathcal{H}^2$, $\mathbf{Z} \in \mathcal{H}_X^2$, $\alpha \in \mathbb{A}$, $\rho \in M(\mathcal{R})$, \mathbb{Q} is a probability measure on (Ω, \mathcal{F}) , and (2.4) is satisfied \mathbb{P} -a.s. for all $t \in [0, T]$.

To use an admissible policy $(\lambda, \xi) \in \mathcal{C}$ the principal pays the cost where $\hat{p}_t^{\lambda, \xi}(e_i) = \hat{\rho}_t^{\lambda, \xi}(A, e_i)$, $i = 1, \dots, m$, and $(\hat{\alpha}^{\lambda, \xi}, \hat{\rho}^{\lambda, \xi}) = \mathcal{N}(\lambda, \xi)$.

The last aspect of the problem is a walk-away option of the minor players: all Nash equilibria are disregarded in which the representative agent's expected total cost is higher than the reservation threshold κ . The principal's optimization problem is

$$V(\kappa) := \inf_{(\lambda, \xi) \in \mathcal{C}} \inf_{J^{\lambda, \xi}(\mathcal{N}(\lambda, \xi)) \leq \kappa} J(\lambda, \xi). \quad (2.5)$$

2.3 Analysis of the Stackelberg extended MFG

In this section we will state results under the hypotheses presented below. While setting the hypotheses, we also make the notation used in the previous section precise.

Hypothesis 2.3 (Structure and regularity of the Q -matrix).

(i) *There exists constants $C_1, C_2 > 0$ such that for all $(t, i, j, \alpha, \rho) \in [0, T] \times G \times A \times \mathbb{R}$ we have $0 < C_1 < q(t, i, j, \alpha, \rho) < C_2$. For all $(i, j) \in \{1, \dots, m\}^2 \setminus G$, $q(t, i, j, \alpha, \rho) = 0$ for $t \in [0, T]$, $\alpha \in A$, $\rho \in \mathcal{R}$.*

(ii) *There exists a constant $C > 0$ such that for all $t \in [0, T]$, $(i, j) \in G$, $\alpha, \alpha' \in A$, and $\rho, \rho' \in \mathcal{R}$, we have*

$$|q(t, i, j, \alpha, \rho) - q(t, i, j, \alpha', \rho')| \leq C(|\alpha - \alpha'| + W_{\mathcal{R}}(\rho, \rho')). \quad (2.6)$$

Hypothesis 2.4 (Regularity of the running cost).

There exists a constant $C > 0$ such that for all $(t, i, \ell) \in [0, T] \times \{1, \dots, m\} \times \mathbb{R}_+^m$, $\alpha, \alpha' \in A$, $p, p' \in \mathcal{P}(E)$, $\rho, \rho' \in \mathcal{R}$, we have

$$|f(t, e_i, \alpha, \rho; \ell) - f(t, e_i, \alpha', \rho'; \ell)| \leq C(|\alpha - \alpha'| + W_{\mathcal{R}}(\rho, \rho')). \quad (2.7)$$

Given a policy $(\lambda, \xi) \in \mathcal{C}$, the Hamiltonian for the representative player's optimization problem is the function $H : [0, T] \times E \times \mathbb{R}^m \times A \times \mathcal{R} \rightarrow \mathbb{R}$

$$H : (t, x, z, \alpha, \rho) \mapsto x^* (Q(t, \alpha, \rho) - Q^0) z + f(t, x, \alpha, \rho; \lambda_t). \quad (2.8)$$

The representative player's reduced Hamiltonian in state e_i is $H_i : (t, z, \alpha, \rho) \mapsto H(t, e_i, z, \alpha, \rho)$, $i = 1, \dots, m$.

Hypothesis 2.5 (Minimizer of the Hamiltonian).

(i) *For any $t \in [0, T]$, $i \in \{1, \dots, m\}$, $z \in \mathbb{R}^m$ and $\rho \in \mathcal{R}$, the mapping $\alpha \mapsto H_i(t, z, \alpha, \rho)$ admits a unique minimizer which we denote by $\hat{\alpha}_i(t, z, \rho)$.*

(ii) *$\hat{\alpha}_i$ is measurable on $[0, T] \times \mathbb{R}^m \times \mathcal{R}$ for every $i \in \{1, \dots, m\}$.*

With the minimizers at hand we define the representative player's optimized Hamiltonian \hat{H} and the optimizer $\hat{\alpha}$ as

$$\hat{H}(t, x, z, \rho) := \sum_{i=1}^m 1_{e_i}(x) \hat{H}_i(t, z, \rho), \quad \hat{\alpha}(t, x, z, \rho) := \sum_{i=1}^m 1_{e_i} \hat{\alpha}_i(t, z, \rho), \quad (2.9)$$

where $\hat{H}_i(t, z, \rho) = H_i(t, z, \hat{\alpha}_i(t, z, \rho), \rho)$.

In the notation, we intentionally differentiate between a strategy and the function minimizing the Hamiltonian by denoting the former with the greek letter α and the latter with the hatted latin letter $\hat{\alpha}$. By evaluating the function $\hat{\alpha}$ as in (2.4) we get an admissible strategy of feedback form (feedback on state, aggregate, and joint distribution). Later, in Proposition 2.7, we study how $\hat{\alpha}$ can be used to construct a mean-field Nash equilibrium.

Hypothesis 2.6 (Regularity of the Hamiltonian minimizer).

There exists a constant $C > 0$, independent of the principal's policy, such that for all $(t, i, \rho) \in [0, T] \times \{1, \dots, m\} \times \mathcal{R}$ and $z, z' \in \mathbb{R}^m$:

$$|\hat{a}_i(t, z, \rho) - \hat{a}_i(t, z', \rho)| \leq C \|z - z'\|_{e_i}. \quad (2.10)$$

The following result provides necessary and sufficient conditions for a mean-field Nash equilibrium. The proof follows the lines of [7, Thm. 1].

Proposition 2.7. Assume that Hypothesis 2.3–2.6 hold true. If (2.4) admits a solution $(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \mathbb{Q})$ then $(\boldsymbol{\alpha}, \boldsymbol{\rho})$ is a mean-field Nash equilibrium (according to Definition 2.1). Conversely, if $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\rho}})$ is a mean-field Nash equilibrium then (2.4) admits a solution $(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \mathbb{Q})$ such that $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}, d\mathbb{P} \otimes dt$ -a.s., and $\rho_t = \hat{\rho}_t, dt$ -a.e.

Given $\mathbf{Z} \in \mathcal{H}_X^2$, $\boldsymbol{\lambda} \in \Lambda$, and real-valued \mathcal{F}_0 -measurable Y_0 , consider under \mathbb{P} :

$$\left\{ \begin{array}{l} Y_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = Y_0 - \int_0^t \hat{H}(s, X_{s-}, Z_s, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}) ds + \int_0^t Z_s^* d\mathcal{M}_s, \\ \mathcal{E}_t = 1 + \int_0^t \mathcal{E}_{s-} X_{s-}^* (Q(s, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}) - Q^0) \psi_s^+ d\mathcal{M}_s, \\ \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = \mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} \circ (\hat{\alpha}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, X_t)^{-1}, \quad \frac{d\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}}{d\mathbb{P}} = \mathcal{E}_T, \\ \hat{\alpha}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = \hat{a}(t, X_{t-}, Z_t, \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}), \quad \hat{p}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}(\cdot) = \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}(A, \cdot). \end{array} \right. \quad (2.11)$$

These are the same equations as (2.4), except that the dynamic of \mathbf{Y} is written in the forward direction of time. Here Y_0 is fixed instead of Y_T .

Hypothesis 2.8 (Regularity of the principal's cost).

- (i) The function $U : \mathbb{R} \rightarrow \mathbb{R}$ is invertible.
- (ii) c_0, f_0 are measurable on $[0, T] \times \mathbb{R}^3$.

Consider the following optimal control problem

$$\begin{aligned} \tilde{V}(\kappa) := \inf_{Y_0: \mathbb{E}[Y_0] \leq \kappa} \inf_{\substack{\mathbf{Z} \in \mathcal{H}_X^2 \\ \boldsymbol{\lambda} \in \Lambda}} \mathbb{E}^{\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}} \left[\int_0^T \left(c_0 \left(t, \hat{p}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} \right) + f_0(t, \lambda_t) \right) dt \right. \\ \left. + C_0 \left(\hat{p}_T^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} \right) + U^{-1} \left(-Y_T^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} \right) \right], \end{aligned} \quad (2.12)$$

under the dynamic constraint (2.11) under \mathbb{P} (the dynamic under $\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}$ is given below in (3.1)). The optimization is now performed not only over the principal's control policy, $\boldsymbol{\lambda}$, but also over the initial condition Y_0 and the \mathbf{Z} component. Since we want to find a solution to (2.4), the terminal Y_T must, by definition of the representative agent's problem, equal the utility $U(\xi)$ of the terminal payment ξ . This remark allows us to remove ξ from the principal's problem, replacing it by $U^{-1}(Y_T)$.

Proposition 2.9. If Hypothesis 2.3–2.6, 2.8 hold true, then $\tilde{V}(\kappa) = V(\kappa)$.

The proof follows the lines of [7, Thm. 2].

Our final result says that certain extended MFGs of the type (2.3) are equivalent to regular MFGs (where there is no dependence on the distribution of player actions) in the sense that the two problem's

Nash equilibria are the same. The key property of games with this feature is that their Hamiltonian and transition rate matrix evaluated at the mean-field Nash equilibrium are functions of the equilibrium state distribution, not the full joint distribution of equilibrium state and control. The property is formalized in the following hypothesis:

Hypothesis 2.10 (Properties for Nash equilibria simplification).

- (i) *There exists a unique solution $(\hat{Y}, \hat{Z}, \hat{\alpha}, \hat{\rho}, \hat{Q})$ to (2.4).*
- (ii) *There exists measurable functions $\bar{a}_i : [0, T] \times \mathbb{R}^m \times \mathcal{P}(E) \rightarrow A$, $\bar{f} : [0, T] \times E \times A \times \mathbb{R}^m \times A \times \mathcal{P}(E) \rightarrow \mathbb{R}$ and $\bar{Q} : [0, T] \times A \times \mathcal{P}(E) \rightarrow \mathbb{R}^m \times \mathcal{R} \times \mathbb{R}_+^m$:*

$$\begin{aligned} \hat{a}_i(t, z, \hat{\rho}_t) &= \bar{a}_i(t, z, \hat{p}_t), \\ f(t, e_i, \hat{a}_i(t, z, \hat{\rho}_t), \hat{\rho}_t; \ell) &= \bar{f}(t, e_i, \bar{a}_i(t, z, \hat{p}_t), \hat{p}_t; \ell), \\ Q(t, \hat{a}_i(t, z, \hat{\rho}_t), \hat{\rho}_t) &= \bar{Q}(t, \bar{a}_i(t, z, \hat{p}_t), \hat{p}_t), \end{aligned} \quad (2.13)$$

where $\hat{p}_t(e_i) := \hat{\rho}_t(A, e_i)$, $i = 1, \dots, m$.

- (iii) *There exists constants C_1 and C_2 , independent of the principal's policy, such that for all $(t, i) \in [0, T] \times \{1, \dots, m\}$, $z, z' \in \mathbb{R}^m$ and $p, p' \in \mathcal{P}(E)$:*

$$|\bar{a}_i(t, z, p) - \bar{a}_i(t, z', p')| \leq C_1 \|z - z'\|_{e_i} + (C_1 + C_2 \|z\|_{e_i}) \|p - p'\|. \quad (2.14)$$

Assuming that hypothesis 2.10 is true we define the non-extended mean field Nash equilibrium of the game as follows:

Definition 2.11. *Let $(\alpha, p) \in \mathbb{A} \times M(\mathcal{P}(E))$ and denote by $\mathbb{Q}^{\alpha, p} \in \mathcal{P}(\Omega)$ the measure such that the coordinate process X_t has transition rate matrix $\bar{Q}(t, \alpha_t, p_t)$ under $\mathbb{Q}^{\alpha, p}$. Assume that $(\bar{\alpha}, \bar{p}) \in \mathbb{A} \times M(\mathcal{P}(E))$ satisfies*

- (i) $\bar{\alpha} = \arg \inf_{\alpha \in \mathbb{A}} \mathbb{E}^{\mathbb{Q}^{\alpha, \bar{p}}} \left[\int_0^T \bar{f}(t, X_t, \alpha_t, \bar{p}_t) dt - U(\xi) \right]$,
- (ii) $\forall t \in [0, T], i \in \{1, \dots, m\} : \bar{p}_t(i) = \mathbb{Q}^{\bar{\alpha}, \bar{p}}(X_t = e_i)$.

Then $(\bar{\alpha}, \bar{p})$ is called a non-extended mean field Nash equilibrium.

The following result allows a simplification of the Nash equilibrium through a non-extended problem. The proof is found in Appendix A.

Proposition 2.12. *Assume Hypothesis 2.3–2.5, 2.10 to be true. Denote the tuple of Hypothesis 2.10(i) by $(\hat{Y}, \hat{Z}, \hat{\alpha}, \hat{\rho}, \hat{Q})$. The pair $(\hat{\alpha}, \hat{\rho})$ is a mean-field Nash equilibrium. Let \hat{p}_t be the the E -marginal of $\hat{\rho}_t$ and let $(\bar{\alpha}, \bar{p})$ be a non-extended mean field Nash equilibrium, satisfying Definition 2.11. Then $\hat{p}_t = \bar{p}_t$ for dt -a.e. $t \in [0, T]$ and $\hat{\alpha}_t = \bar{\alpha}_t d\mathbb{P} \otimes dt$ -a.e..*

3 Numerical approach

In this section, we propose a numerical method to solve the Stackelberg equilibrium. This requires finding the optimal policy of the principal and the associated mean-field Nash equilibrium for the population of agents. The principal's policy influences the Nash equilibrium in a rather intricate way. For this reason, we depart from existing numerical methods for finite state mean field games such as [1], and we propose a probabilistic method in which Monte Carlo samples are generated to train neural networks approximating the optimal controls, including the principal's policy.

3.1 Monte Carlo simulation

From Proposition 2.9, we know that solving the original Stackelberg MFG problem amounts to solving an optimal control problem in which the state can be viewed as (\mathbf{X}, \mathbf{Y}) and has a forward dynamics: under \mathbb{P} , \mathbf{X} is a continuous time Markov chain with Q -matrix Q^0 and \mathbf{Y} satisfies (2.11). We recall that the controls are $\mathbf{Z} \in \mathcal{H}_X^2$, $\boldsymbol{\lambda} \in \Lambda$, and a real-valued \mathcal{F}_0 -measurable random variable Y_0 . This problem involves the state and action distribution. We replace this distribution by an empirical distribution obtained with an interacting system of particles and we discretize the time integral. We first note that, given a triple of controls $(\mathbf{Z}, \boldsymbol{\lambda}, Y_0)$, the dynamic (2.11) can be written as:

$$\begin{cases} Y_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = Y_0 - \int_0^t f(s, X_{s-}, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}; \lambda_s) ds + \int_0^t Z_s^* d\mathcal{M}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \\ \mathcal{E}_t = 1 + \int_0^t \mathcal{E}_{s-} X_{s-}^* (Q(s, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}) - Q^0) \psi_s^+ d\mathcal{M}_s, \\ \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = \mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} \circ (\hat{\alpha}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, X_t)^{-1}, \quad \frac{d\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}}{d\mathbb{P}} = \mathcal{E}_T, \\ \hat{\alpha}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = \hat{\alpha}(t, X_{t-}, Z_t, \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}; \lambda_s), \quad \hat{p}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}(\cdot) = \hat{\rho}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}(A, \cdot), \end{cases} \quad (3.1)$$

where the process $\mathcal{M}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}$ is defined by:

$$\mathcal{M}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0} = \mathcal{M}_t - \int_0^t X_{s-}^* (Q(s, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}) - Q^0) ds,$$

is a $\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}$ -martingale. Furthermore, from this definition we see that the canonical process \mathbf{X} satisfies, under $\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}$,

$$X_t = X_0 + \int_0^t X_{s-}^* Q(s, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}) ds + \mathcal{M}_t^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}. \quad (3.2)$$

In other words, under the probability measure $\mathbb{Q}^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}$, the intensity rate of \mathbf{X} is given by $Q(s, \hat{\alpha}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0}, \hat{\rho}_s^{\mathbf{Z}, \boldsymbol{\lambda}, Y_0})$.

To simulate Monte Carlo trajectories, we will use the expressions (3.1)–(3.2). For simplicity of the implementation, we assume that given any admissible policy $(\boldsymbol{\lambda}, \xi)$, we can express the induced equilibrium control $\hat{\alpha}$ as a function of $\hat{\rho}$, the flow of second marginals of the equilibrium distribution flow $\hat{\rho}$. To wit, for every $(\boldsymbol{\lambda}, \xi)$, denoting $(\mathbf{Y}, \mathbf{Z}, \hat{\alpha}, \hat{\rho}, \mathbb{Q})$ the solution to (2.4), we assume that there exists $\tilde{a} : [0, T] \times E \times \mathbb{R}^m \times \mathcal{P}(E) \rightarrow \mathbb{R}$ such that

$$\hat{\alpha}_t = \tilde{a}(t, X_t, Z_t, \hat{\rho}_t),$$

where $\hat{\rho}_t = \hat{\rho}_t(A, \cdot)$ is the state marginal of $\hat{\rho}$. This is automatically true if the minimizer \hat{a} of the Hamiltonian is independent of the first marginal of ρ , i.e.,

$$\hat{a}_i(t, z, \rho) = \tilde{a}_i(t, z, \rho(A, \cdot)),$$

for a function $\tilde{a}_i : [0, T] \times \mathbb{R}^m \times \mathcal{P}(E) \rightarrow \mathbb{R}$, which is often assumed to be true in extended MFGs, see e.g. [8, Assumption 3.4] in the finite space setting and [35, 36] in the continuous space setting. However, it also holds in more general situations, for instance when the equilibrium control can be expressed in terms of the solution to a forward-backward PDE system [24, 10, 29] in the continuous space setting. In such cases, the equilibrium control is expressed as a feedback function of (t, x) related to the backward PDE, and the distribution of actions can be recovered from this feedback control and the state distribution related to the forward PDE.

We now present the scheme with a finite number of particles and discrete time steps. We consider $N > 0$ particles and denote $\llbracket N \rrbracket = \{1, \dots, N\}$ the set of indexes. Given measurable control functions

$z : [0, T] \times E \rightarrow \mathbb{R}^m, \lambda : [0, T] \rightarrow \mathbb{R}_+^m, y_0 : E \rightarrow \mathbb{R}$, we construct trajectories $(X_{t_n}^i, Y_{t_n}^i)_{n,i=1,\dots,N}$. After initialization, we proceed iteratively for every $n \geq 1$ while $t_n \leq T$: $X_{t_n}^i$ is sampled with Q -matrix given by

$$Q_{t_n}^i := Q(t_n, \alpha_{t_n}^i, \bar{\rho}_{t_n}^N),$$

where

$$\bar{\rho}_{t_n}^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X_{t_n}^i, \alpha_{t_n}^i)}$$

is the empirical action-state distribution, and

$$\alpha_{t_n}^i = \check{a}(t, X_{t_n}^i, z(t_n, X_{t_n}^i), \bar{\rho}_{t_n}^N), \quad \text{where } \bar{p}_{t_n}^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_{t_n}^i}.$$

Based on (3.2), we define:

$$\Delta \mathcal{M}_{t_n}^i = -(X_{t_{n+1}}^i - X_{t_n}^i) - (X_{t_n}^i)^* Q_{t_n}^i (t_{n+1} - t_n),$$

and then, based on (3.1), we let: $Y_0^i = y(X_0^i)$ and for $n \geq 0$,

$$Y_{t_{n+1}}^i = Y_{t_n}^i - f(t_n, X_{t_n}^i, \alpha_{t_n}^i, \bar{\rho}_{t_n}^N; \lambda(t_n))(t_{n+1} - t_n) + z(t_n, X_{t_n}^i)^* \Delta \mathcal{M}_{t_n}^i.$$

Here t_n corresponds to the time of the n -th jump in the particle system $(\mathbf{X}^i)_{i=1,\dots,N}$. In the implementation, these trajectories are constructed using a time marching procedure from time 0 until time T , with steps corresponding to jumps. For more details, see Algorithm 1.

3.2 Approximation based on neural networks

In order to have a problem amenable to numerical treatment, we replace the controls $(\mathbf{Z}, \boldsymbol{\lambda}, Y_0)$ by parameterized functions $z_{\theta_1} : [0, T] \times E \rightarrow \mathbb{R}^m, \lambda_{\theta_2} : [0, T] \rightarrow \mathbb{R}_+^m$, and $y_{0,\theta_3} : E \rightarrow \mathbb{R}$ with respective parameters $\theta_1, \theta_2, \theta_3$. Since the number of states m is potentially large, we choose to use neural networks and, to be specific, in the implementation we take feedforward fully connected neural networks. Then, taking into account the time discretization and the approximation using a finite number of particles as described above, instead of (2.12), and considering a finite number M of Monte Carlo samples, the goal is now to minimize over $\theta = (\theta_1, \theta_2, \theta_3)$ the objective function:

$$\begin{aligned} \mathbb{J}^N(\theta) = & \frac{1}{M} \sum_{j=1}^M \left[\sum_{n=0}^{n_{tot}-1} \left(c_0 \left(t_n, \bar{p}_{t_n}^{j,N,\theta} \right) + f_0(t_n, \lambda_{\theta_2}(t_n)) \right) (t_{n+1} - t_n) \right. \\ & \left. + C_0 \left(\bar{p}_T^{j,N,\theta} \right) + \frac{1}{N} \sum_{i=1}^N U^{-1} \left(-Y_T^{j,i,\theta} \right) \right] \end{aligned} \quad (3.3)$$

where, for $j = 1, \dots, M$, $(\mathbf{Y}^{j,i,\theta})_{i \in \llbracket N \rrbracket}$ and $\bar{\mathbf{p}}^{j,N,\theta}$ are constructed by Algorithm 1 using $(z, \lambda, y_0) = (z_{\theta_1}, \lambda_{\theta_2}, y_{0,\theta_3})$. Intuitively, in the limit when M, N and the number of parameters θ go to infinity, we would expect $\inf_{\theta} \mathbb{J}^N(\theta)$ to converge to the principal's optimal cost.

To optimize over $\theta = (\theta_1, \theta_2, \theta_3)$, we rely on a variant of stochastic gradient descent (namely the Adaptive Moment Estimation algorithm). This kind of methods is particularly well suited to the minimization of \mathbb{J}^N in (3.3) since on the one hand the number of parameters in deep neural networks is potentially large and on the other hand this cost is written as an expectation and can thus be computed using Monte Carlo samples. Our method can be viewed as an adaptation of the second algorithm in [5]

to the finite state case and its generalization to the Stackelberg setting with a principal. More precisely, we introduce for a sample $S = (X_{t_n}^i, Y_{t_n}^i, Z_{t_n}^i)_{n=0, \dots, n_{tot}, i \in [N]}$ of a population of N particles:

$$\begin{aligned} \mathbb{J}_S^N(\theta) = & \sum_{n=0}^{n_{tot}-1} \left(c_0 \left(t_n, \bar{p}_{t_n}^{N, \theta} \right) + f_0(t_n, \lambda_{\theta_2}(t_n)) \right) (t_{n+1} - t_n) \\ & + C_0 \left(\bar{p}_T^{N, \theta} \right) + \frac{1}{N} \sum_{i=1}^N U^{-1} \left(-Y_T^{i, \theta} \right) \end{aligned} \quad (3.4)$$

where $\bar{p}_{t_n}^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_{t_n}^i}$. Note that (3.3) is simply the average of \mathbb{J}_S^N over M samples (here one sample corresponds to the trajectories of one population). See Algorithm 2 for more details on the stochastic gradient descent (SGD) method applied in the context of Stackelberg mean field game.

Remark 3.1. *If the regulator is not optimizing its own outcomes, in other words if the policies set by the regulator, (λ, ξ) , are exogenous, the neural network approach can be still used. Since λ is known in Section 3.1, we can write system (3.1) controlled only by Z and Y_0 . Then in Section 3.2, we replace controls (Z, Y_0) by parameterized functions $z_{\theta_1} : [0, T] \times E \rightarrow \mathbb{R}^m$ and $y_{\theta_2} : E \rightarrow \mathbb{R}$ with respective parameters θ_1, θ_2 . Then considering a finite number M of Monte Carlo samples with N particles in each, the goal is to minimize over $\theta = (\theta_1, \theta_2)$ the loss function:*

$$\mathbb{L}^N(\theta) = \frac{1}{M} \sum_{j=1}^M \left[\left(U(\xi) - \frac{1}{N} \sum_{i=1}^N Y_T^{j, i, \theta} \right)^2 \right] \quad (3.5)$$

4 Experimental results

4.1 SIR equilibrium with a fixed policy

Our first numerical experiment is an extended MFG with SIR dynamics. The principal is not active in this experiment; we think of the principal as a static regulator who has beforehand declared a fixed policy. The purpose is two-fold. Firstly, since a semi-explicit solution is attainable the numerical method's accuracy can be evaluated (see Fig. 2 and 3), and secondly, we illuminate the game-aspect of the model and how the agents are prone to “cheat” a little and use a less conservative socialization protocol than the incentivized one.

We begin by recalling the problem introduced in Section 1.1 and we focus here on the limiting problem with a continuum of agents. The representative agent evolves according to the rate matrix $Q(t, \alpha_t, \rho_t)$ given in (1.6) and the running cost in (1.7). As stated in Section 1.1, the running cost does not depend on the extended mean field, however the dynamic does, making the problem an extended MFG. The representative agent's terminal utility is $U(\xi) = \xi$.

Given (λ, ξ) , the mean field Nash equilibrium can be reduced to the solution of a system of forward-backward ODEs as we explain in more details below. We solve the ODEs by iteratively solving the forward and the backward equations, until the distance between two iterates is small enough.

Since the neural network based method we propose is new, we consider a testbed on which we can assess its correctness using a well-studied approach, which serves as a benchmark. For any given (λ, ξ) , the mean field Nash equilibrium can be reduced to the solution of a system of forward-backward ODEs, in the spirit of, for example, [4, Section 7.2.2]. The forward equation describes the evolution of the

Algorithm 1: Monte Carlo simulation an interacting batch

- 1 **Input:** Transition rate matrix Q ; number of particles N ; time horizon T ; initial distribution p_0 ; control functions λ, y_0, z
 - 2 **Output:** Approximate sampled trajectories of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ solving (3.1)–(3.2)
 - 1: Let $n = 0, t_0 = 0$; pick $X_0^i \sim p^0$ i.i.d and set $Y_0^i = y_0(X_0^i), i \in \llbracket N \rrbracket$
 - 2: **while** $t_n \leq T$ **do**
 - 3: Set $Z_{t_n}^i = z(t_n, X_{t_n}^i), \alpha_{t_n}^i = \check{a}(t_n, X_{t_n}^i, Z_{t_n}^i, p_{t_n}), i \in \llbracket N \rrbracket$
 - 4: Let $\bar{\rho}_{t_n}^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X_{t_n}^i, \alpha_{t_n}^i)}$ and $\bar{p}_{t_n}^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_{t_n}^i}$
 - 5: Pick $(T^{i,e})_{e \in E, i \in \llbracket N \rrbracket}$ i.i.d. with exponential distribution of parameter 1
 - 6: Set the holding times: $\tau^{i,e} = T^{i,e} / Q_{X_{t_n}^i, e}(t_n, \alpha_{t_n}^i, \bar{\rho}_{t_n}^N), i \in \llbracket N \rrbracket, e \in E$
 - 7: Let $e_\star^i = \arg \min_{e \in E} \tau^{i,e}$ and $\tau_\star^i = \tau^{i, e_\star^i} = \min_{e \in E} \tau^{i,e}, i \in \llbracket N \rrbracket$
 - 8: Let $i_\star = \arg \min_{i \in \llbracket N \rrbracket} \tau_\star^i$ be the first particle to jump
 - 9: Let $\Delta t = \tau_\star^{i_\star}$ be the time increment
 - 10: Set $X_{t_n+\Delta t}^{i_\star} = e_\star^{i_\star}$, and for every $i \neq i_\star$, set $X_{t_n+\Delta t}^i = X_{t_n}^i$
 - 11: Let $\Delta M_{t_n}^i = X_{t_n+\Delta t}^i - X_{t_n}^i - (X_{t_n}^i)^* Q(t_n, \alpha_{t_n}^i, \bar{\rho}_{t_n}^N) \Delta t, i \in \llbracket N \rrbracket$
 - 12: Let $Y_{t_n+\Delta t}^i = Y_{t_n}^i - f(t, X_{t_n}^i, \alpha_{t_n}^i, \bar{\rho}_{t_n}^N; \lambda(t_n)) \Delta t + (Z_{t_n}^i)^* \Delta M_{t_n}^i, i \in \llbracket N \rrbracket$
 - 13: Set $n = n + 1$ and $t_n = t_{n-1} + \Delta t$
 - 14: **end while**
 - 15: Set $n_{tot} = n, t_{n_{tot}} = T, (X_{t_{n_{tot}}}^i, Y_{t_{n_{tot}}}^i, Z_{t_{n_{tot}}}^i) = (X_{t_{n_{tot}-1}}^i, Y_{t_{n_{tot}-1}}^i, Z_{t_{n_{tot}-1}}^i)$
 - 16: **return** $(X_{t_n}^i, Y_{t_n}^i, Z_{t_n}^i)_{n=0, \dots, n_{tot}, i \in \llbracket N \rrbracket}$ and $(t_n)_{n=0, \dots, n_{tot}}$
-

Algorithm 2: SGD for Stackelberg Mean Field Game

- 1 **Input:** Initial parameter θ_0 ; number of iterations K ; sequence $(\beta_k)_{k=0, \dots, K-1}$ of learning rates; transition rate matrix Q ; number of particles N ; time horizon T ; initial distribution p_0
 - 2 **Output:** Approximation of θ^* minimizing \mathbb{J}^N defined by (3.3)
 - 1: **for** $k = 0, 1, 2, \dots, K - 1$ **do**
 - 2: Sample $S = (X_{t_n}^i, Y_{t_n}^i, Z_{t_n}^i)_{n=0, \dots, n_{tot}, i \in \llbracket N \rrbracket}$ and $(t_n)_{n=0, \dots, n_{tot}}$ using Algorithm 1 with control functions $(z, \lambda, y_0) = (z_{\theta_{k,0}}, \lambda_{\theta_{k,1}}, y_{0, \theta_{k,2}})$ and parameters: transition rate matrix Q ; number of particles N ; time horizon T ; initial distribution p_0
 - 3: Compute the gradient $\nabla \mathbb{J}_S^N(\theta_k)$ of $\mathbb{J}_S^N(\theta_k)$ defined by (3.4)
 - 4: Set $\theta_{k+1} = \theta_k - \beta_k \nabla \mathbb{J}_S^N(\theta_k)$
 - 5: **end for**
 - 6: **return** θ_K
-

Table 1: Parameter values, policies, and contact factor in the four numerical experiments on the SIR extended MFG. The regulator declares a fixed policy (λ, ξ) .

Test case	Contact factor	ξ	$\lambda_t^{(S)}$	$\lambda_t^{(I)}$	$\lambda_t^{(R)}$
Free spread	Constant	0	1	1	1
No lockdown	MF Nash eq.	0	1	1	1
Late lockdown	MF Nash eq.	0	$1 - 0.3\mathbb{1}_{t>40}$	$0.9 - 0.3\mathbb{1}_{t>40}$	1
Early lockdown	MF Nash eq.	0	$1 - 0.3\mathbb{1}_{t\leq 10}$	$0.9 - 0.3\mathbb{1}_{t\leq 10}$	1

Parameter	T	p^0	c_λ	c_I	β	γ	η
Value in tests	50	(0.9, 0.1, 0)	10	1	0.25	0.1	0

population distribution while the backward characterizes the value function of an infinitesimal agent:

$$\begin{aligned}
\dot{p}_t(S) &= -\beta \left(\lambda_t^{(S)} + \frac{\beta}{c_\lambda} \lambda_t^{(I)} p_t(I) (u_t(S) - u_t(I)) \right) \lambda^{(I)} p_t(S) p_t(I) + \eta p_t(R), \\
\dot{p}_t(I) &= \beta \left(\lambda_t^{(S)} + \frac{\beta}{c_\lambda} \lambda_t^{(I)} p_t(I) (u_t(S) - u_t(I)) \right) \lambda^{(I)} p_t(S) p_t(I) - \gamma p_t(I), \\
\dot{p}_t(R) &= \gamma p_t(I) - \eta p_t(R) \\
\dot{u}_t(S) &= \beta \lambda^{(S)} \lambda^{(I)} p_t(I) (u_t(S) - u_t(I)) + \frac{1}{2c_\lambda} (\beta \lambda^{(I)} p_t(S) (u_t(S) - u_t(I)))^2 \\
\dot{u}_t(I) &= \gamma (u_t(I) - u_t(R)) - c_I \\
\dot{u}_t(R) &= \kappa (u_t(R) - u_t(S)) \\
u_T(e) &= 0, \quad p_0(e) = p_0^e, \quad e \in \{S, I, R\},
\end{aligned} \tag{4.1}$$

They are, in the finite state MFG, the counterparts to Kolmogorov-Fokker-Planck and Hamilton-Jacobi-Bellman partial differential equations arising in continuous space MFGs. For more details on the derivation of these ODEs for a (slightly different) class of finite-state MFGs, we refer the reader to [4, Section 7.2.2].

The six equations are coupled, reflecting the fact that an agent cannot compute their value function (and their optimal control) without knowing the population evolution when in Nash equilibrium, and the population distribution cannot be computed without knowing the controls chosen by the agents. For this reason, we cannot solve one equation before the other. We propose to solve this system by solving iteratively the forward and the backward ODEs in turn, plugging the solution of the previous iteration in the equation at the current iteration. To implement this strategy, we discretize time and replace the distribution and the value function by vectors (see Algorithm 3). In our implementation, we used an explicit Euler scheme.

When we analyze Figure 2 and Figure 3 (also Figure 4 and 5), we see that neural network based method accurately computes the results obtained by the ODE method.

Algorithm 3: ODE Approach for the finite-state Mean Field Game

-
- 1 **Input:** Time horizon T ; Time Increments Δt ; Initial discretized flow of state distribution and value function $\mathbf{p}^{(0)} = \{p_0, p_{\Delta t}, p_{2\Delta t}, \dots, p_T\}$ and $\mathbf{u}^{(0)} = \{u_0, u_{\Delta t}, u_{2\Delta t}, \dots, u_T\}$; initial state distribution p_0 ; terminal condition of value function u_T ; Tolerance τ
 - 2 **Output:** Equilibrium discretized flow of state distribution and corresponding value function: $(\hat{p}_0, \hat{p}_{\Delta t}, \dots, \hat{p}_T)$ and $(\hat{u}_0, \hat{u}_{\Delta t}, \dots, \hat{u}_T)$
 - 1: $k \leftarrow 0$
 - 2: **while** $\|\mathbf{p}^{(k)} - \mathbf{p}^{(k-1)}\| > \tau$ or $\|\mathbf{u}^{(k)} - \mathbf{u}^{(k-1)}\| > \tau$ **do**
 - 3: Compute $\mathbf{p}^{(k+1)}$ solving the forward equation in (4.1) with \mathbf{u} replaced by $\mathbf{u}^{(k)}$
 - 4: Compute $\mathbf{u}^{(k+1)}$ solving the backward equation in (4.1) with \mathbf{p} replaced by $\mathbf{p}^{(k+1)}$
 - 5: Update $\mathbf{p}^{(k-1)} \leftarrow \mathbf{p}^{(k)}$, $\mathbf{p}^{(k)} \leftarrow \mathbf{p}^{(k+1)}$, $\mathbf{u}^{(k-1)} \leftarrow \mathbf{u}^{(k)}$, $\mathbf{u}^{(k)} \leftarrow \mathbf{u}^{(k+1)}$
 - 6: $k \leftarrow k + 1$
 - 7: **end while**
 - 8: **return** $\mathbf{p}^{(k)}$ and $\mathbf{u}^{(k)}$
-

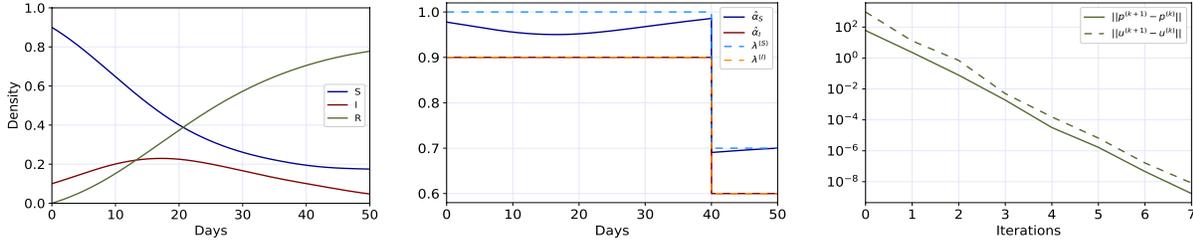


Figure 2: Late lockdown with the ODE solver. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the solver (right).

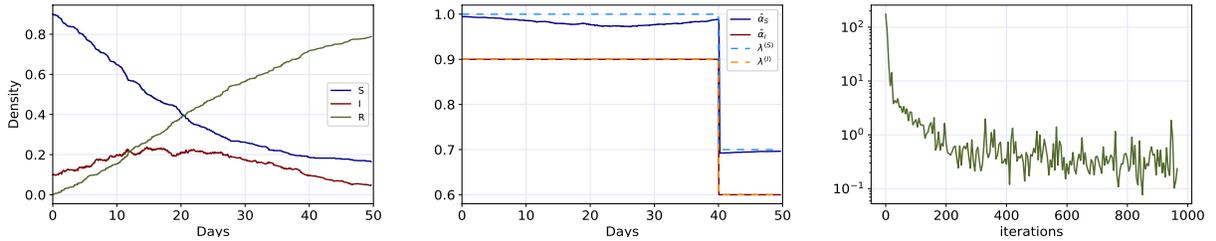


Figure 3: Late lockdown with Algorithm 2. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the loss value (right).

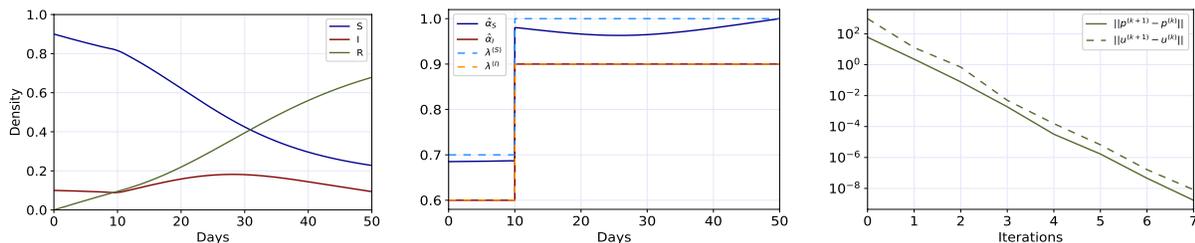


Figure 4: Early lockdown with the ODE solver. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the solver (right).

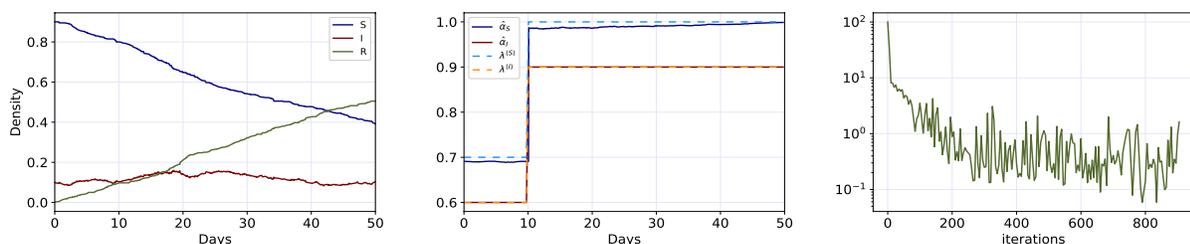


Figure 5: Early lockdown with Algorithm 2. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the loss value (right).

To illustrate the impact of the agents' optimization and the impact of the regulator's choice of policy, we consider four test cases, presented in Table 1. The parameters related to the dynamics are chosen according to the following assumptions on COVID-19 pandemic: Since the average recovery duration is around 10 days, the recovery rate is taken as $\gamma = 0.1$ 1/days; furthermore, to our up to date knowledge, the reinfection possibility is lower in the first 3 months and for later times it is uncertain.⁶ Therefore, in some experiments in this paper, the reinfection rate η is taken 0 or 0.01. Finally, since it is hard to estimate infection rate β directly, we used the estimates on Basic Reproduction number (R_0) of COVID-19. The CDC uses $R_0 = 2.5$ in the "Current Best Estimate" scenario in its simulations.⁷ Therefore, we use $\beta = R_0 \times \gamma = 0.25$ in our simulations. For the parameters related to the cost function of agents, we choose $c_I = 1$ and $c_\lambda = 10$ to balance the powers of the different cost terms. Since c_I stands alone while c_λ is multiplied with a term likely to be much more smaller than 1, we decide to use a higher c_λ value. Further, we also rule out extreme cases. For example, if c_I is taken to be too dominant then susceptible people will decide to minimize their contact factor, which leads to unrealistic results. Also, c_I is not taken to be very small (≈ 0), since that means sickness does not come with an added negative effect. We know from our experience during COVID-19 pandemic that being sick has both social and economic burden.

⁶<https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html>

⁷<https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html>

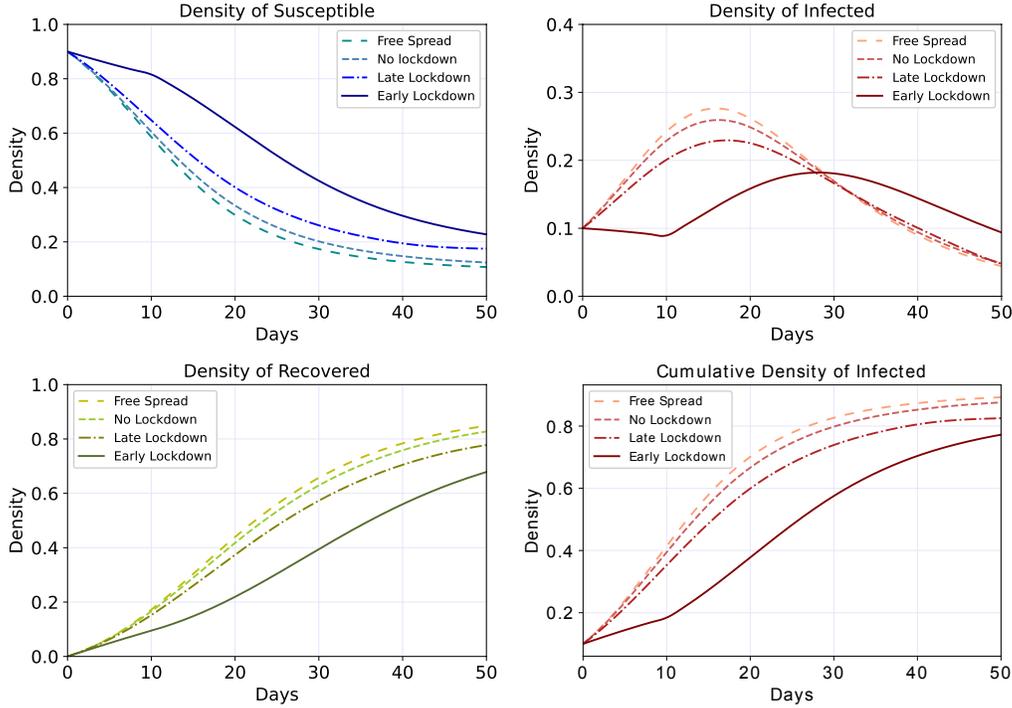


Figure 6: Evolution of the population state distribution in the four test cases, obtained using an ODE solver.

Fig. 6 displays the evolution of the population’s state distribution in each of the four test cases. In it, it is worth to note that the proportion of infected is decreasing from first test case to the last one, which can be interpreted in the following way: In the Nash equilibrium the agents take action to reduce the risk of being infected compared with letting the epidemic spread freely; imposing $\lambda^{(I)} < 1$ encourages the players to be more cautious and hence decreases further infections; last, recommending a $\lambda^{(I)}$ that is low in the beginning and then relaxing it (early lockdown case) helps avoiding the first peak (around 15 days) but leads to another peak later (around 28 days). We can further infer that the cumulative number of infected people can be decreased the most with the early lockdown policy. Therefore, we conclude that early actions are more effective at decreasing the severeness of the disease.

4.2 A semi-explicitly solvable Stackelberg game

We now add the regulator’s optimization to the previous example, making it a Stackelberg game. The regulator pays a cost that is increasing with the number of infections and any deviation of the issued policy $\lambda = (\lambda_t^{(S)}, \lambda_t^{(I)}, \lambda_t^{(R)})_{t \in [0, T]}$ from some endogenously recommended levels, $\bar{\lambda}$. There are multiple ways we can think of the latter: as levels recommended by the health authorities such as C.D.C.; an average of what other regulators are doing (*e.g.*, other countries’ regulations)⁸; or budgetary constraints. We will take on the viewpoint that $\bar{\lambda}$ is a health authority recommendation to the regulating body that

⁸A scenario with multiple competing regulators is a highly interesting and relevant problem. During the past year, we have seen such opposition between US states, and EU member states. We imagine that in such a model we would sometimes see alignment of recommendations and other times specialization. However, the case is beyond the scope of this paper.

is “the government”.

Turning to the specifics, we set $C_0(p) = 0$ and we assume that the government minimizes the proportion of the infected people over time and tries to set socialization levels close to the levels recommended by the health authorities:

$$c_0(t, p) = c_{\text{Inf}} p(I)^2, \quad f_0(t, \lambda) = \sum_{i \in \{S, I, R\}} \frac{\bar{\beta}^{(i)}}{2} \left(\lambda^{(i)} - \bar{\lambda}^{(i)} \right)^2 \quad (4.2)$$

for constant $\bar{\lambda}, \bar{\beta} \in \mathbb{R}_+^m$ and $c_{\text{Inf}} > 0$. The next proposition provides a semi-explicit construction of the optimal contract and the mean-field Nash equilibrium in this case. A proof is presented in appendix B.

Proposition 4.1. *Consider the Stackelberg game of this section. Let $\tilde{H} : [0, T] \times \mathcal{P}(E) \times \mathbb{R}^m \times A \times \mathbb{R}_+^m \rightarrow \mathbb{R}$ be defined by:*

$$\begin{aligned} \tilde{H}(t, \pi, y, \tilde{\alpha}, \lambda) := & (y(I) - y(S)) \beta \lambda^{(I)} \alpha \pi(I) \pi(S) + (y(R) - y(I)) \gamma \\ & + (y(S) - y(R)) \eta + c_0(t, \pi) + f_0(t, \lambda) \\ & + \frac{c_\lambda}{2} \left(\lambda^{(S)} - \tilde{\alpha} \right)^2 \pi(S) + c_I \pi(I). \end{aligned} \quad (4.3)$$

Let $(\hat{\alpha}(\pi_t, y_t), \hat{\lambda}^{(S)}(\pi_t, y_t), \hat{\lambda}^{(I)}(\pi_t, y_t), \hat{\lambda}^{(R)}(\pi_t, y_t))$ be the solution to

$$(\nabla_{\tilde{\alpha}}, \nabla_{\lambda}) \tilde{H}(t, \pi_t, y_t, \hat{\alpha}(\pi_t, y_t), \hat{\lambda}(\pi_t, y_t)) = 0, \quad (4.4)$$

assumed to exist uniquely and be admissible, and (π, y) solves

$$\begin{aligned} \dot{\pi}_t &= \nabla_y \tilde{H}(t, \pi_t, y_t, \hat{\alpha}(\pi_t, y_t), \hat{\lambda}(\pi_t, y_t)), \quad \pi_0 = p_0, \\ \dot{y}_t &= -\nabla_\pi \tilde{H}(t, \pi_t, y_t, \hat{\alpha}(\pi_t, y_t), \hat{\lambda}(\pi_t, y_t)), \quad y_T = 0. \end{aligned} \quad (4.5)$$

Denote by $(\hat{\pi}, \hat{y})$ the solution to (4.5) and let us define the processes $\hat{\alpha} \in \mathbb{A}$ and $\hat{Z} \in \mathcal{H}_X^2$ by:

$$\begin{aligned} \hat{\alpha}_t &= \hat{\alpha}(\hat{\pi}_t, \hat{y}_t) \mathbb{1}_S(X_{t-}) + \hat{\lambda}^{(I)}(\hat{\pi}_t, \hat{y}_t) \mathbb{1}_I(X_{t-}) + \hat{\lambda}^{(R)} \mathbb{1}_R(X_{t-}) \\ \hat{Z}_t &= \left(\frac{c_\lambda \left(\hat{\alpha}(\hat{\pi}_t, \hat{y}_t) - \hat{\lambda}^{(S)}(\hat{\pi}_t, \hat{y}_t) \right)}{\beta \hat{\lambda}^{(I)}(\hat{\pi}_t, \hat{y}_t) \hat{\pi}_t(I)} \mathbb{1}_S(X_{t-}) \mathbb{1}(\lambda^{(I)} > 0), 0, 0 \right). \end{aligned} \quad (4.6)$$

Now let $y^0 \in \mathbb{R}^m$ be such that $p_0^* y^0 \leq \kappa$. We then define the random variable $\hat{\xi}$ almost surely by the Stieltjes integral

$$\hat{\xi} := -X_0^* y_0 + \int_0^T \left(f(t, X_{t-}, \hat{\alpha}_t, \hat{\pi}_t) + X_{t-}^* \bar{Q}(t, \hat{\alpha}_t, \hat{\pi}_t) \hat{Z}_t \right) dt - \int_0^T \hat{Z}_t^* dX_{t-}. \quad (4.7)$$

Then $(\hat{\lambda}, \hat{\xi})$ is an optimal contract. Moreover, under the optimal contract, every agent adopts the strategy where they pick the control $\hat{\alpha}_t$ and the flow of distribution of agents' states is $\hat{\pi}$.

The next numerical experiment is a comparison of Algorithm 2 and the semi-explicit solution of Proposition 4.1. The results are presented in Fig. 7–8 and the parameters used in the simulation are found in Table 2. The additional parameters are chosen as follows: For $\bar{\lambda}$, we assumed that health authorities recommend a stricter policy for the infected people than susceptible and recovered people; therefore $\bar{\lambda}^{(I)}$ is equal to 0.7, while $\bar{\lambda}^{(S)}$ and $\bar{\lambda}^{(R)}$ are equal to 1. Further, for the government it is more important to follow the guidelines for infected people and if there is no reinfection as in this experiment here it is not important to follow the guidelines for the recovered people; therefore, $\bar{\beta}^{(i)}$ are taken 0.2,

1 and 0, respectively for $i \in \{S, I, R\}$. The value of the loss in the numerical scheme of Algorithm 2 converges to the optimal value from the semi-explicit solution. Furthermore, the population dynamics are very similar in both solutions. As for the controls, we see that in both cases, the agents tend to follow closely the regulator’s policy. The policies $\lambda^{(S)}, \lambda^{(I)}$ output by Algorithm 2 seem to capture the average value of the policies given by the semi-explicit solution. Since the loss value is very close to the optimal one, we deduce that these policies are approximately optimal.

Table 2: Parameter values for the SIR Stackelberg MFG experiment.

T	p^0	c_λ	c_I	c_{Inf}	$\bar{\beta}$	$\bar{\lambda}$	β	γ	η	κ
30	(0.9, 0.1, 0)	10	0.5	1	(0.2, 1, 0)	(1, 0.7, 0)	0.25	0.1	0	0

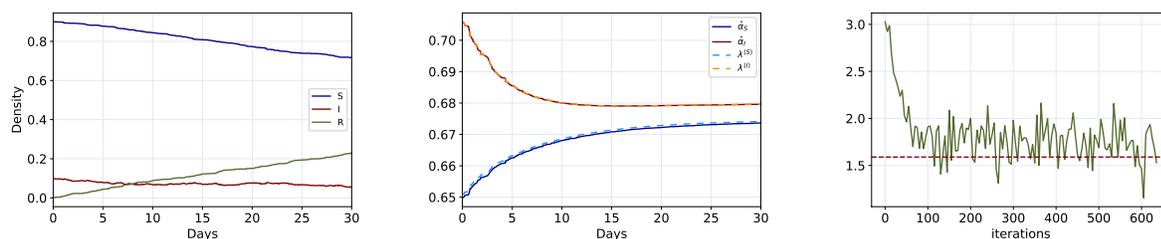


Figure 7: SIR Stackelberg mean field game with Algorithm 2. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the loss value (right). Here, green line refers to the loss function found by the neural network based approach and the red line shows the optimal loss value.

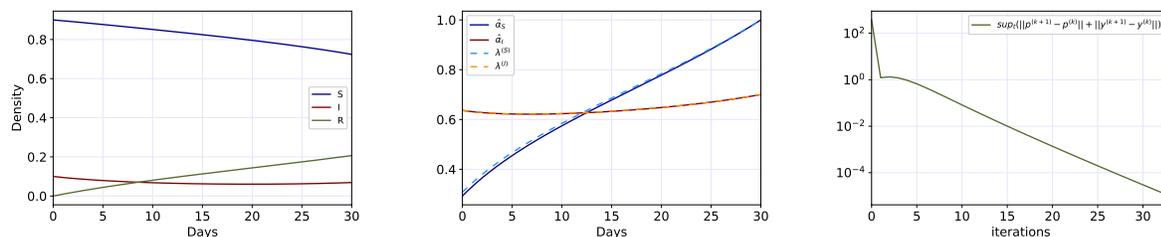


Figure 8: SIR Stackelberg mean field game with ODE solver. Evolution of the population state distribution (left), evolution of the controls (middle), convergence of the solver (right).

4.3 A more complex model: SEIRD

To illustrate the flexibility and scalability of the proposed numerical method, we now consider a more complex model. On top of the S , I and R states considered above, we add two new states: Exposed (E) and Dead (D). An individual is in state E when it has been infected but is not yet infectious. Hence the agents evolve from S to E and then I , and the infection rate from S to E depends on the proportion of the infected people. From the point of view of the dynamics, the state D is absorbing but it is important

for the cost functions discussed below. Now R is interpreted as recovered. We consider the states in the order: S, E, I, R, D . A representative agent evolves according to the rate matrix $Q(t, \alpha_t, \rho_t)$,

$$Q(t, \alpha, \rho) = \begin{bmatrix} \cdots & \beta\alpha_t \int_A a\rho_t(da, I) & 0 & 0 & 0 \\ 0 & \cdots & \epsilon & 0 & 0 \\ 0 & 0 & \cdots & \gamma & \delta \\ \eta & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots \end{bmatrix}, \quad (4.8)$$

where $\beta, \gamma, \eta, \Lambda, \delta \in \mathbb{R}_+$ are constants. See Fig. 9 for a diagram of the dynamics.

The cost of the agents generalizes the previous example (1.7) by incorporating terms related to the new states as follows:

$$f(t, x, \alpha, \rho; \lambda) = \frac{c_\lambda}{2} \left(\lambda^{(S)} - \alpha \right)^2 \mathbb{1}_S(x) + \frac{1}{2} \left(\lambda^{(E)} - \alpha \right)^2 \mathbb{1}_E(x) + \left(\frac{1}{2} \left(\lambda^{(I)} - \alpha \right)^2 + c_I \right) \mathbb{1}_I(x) + \frac{1}{2} \left(\lambda^{(R)} - \alpha \right)^2 \mathbb{1}_R(x) + c_D \mathbb{1}_D(x), \quad (4.9)$$

where $c_\lambda, c_I, c_D \in \mathbb{R}_+$ are constants. We note that the final term in (4.9) represents a cost of passing due to the disease (transitioning to the absorbing state D) and a preference for doing so as late as possible. The terminal payment utility is $U(\xi) = \xi$. Further, we also modify the cost of the regulator:

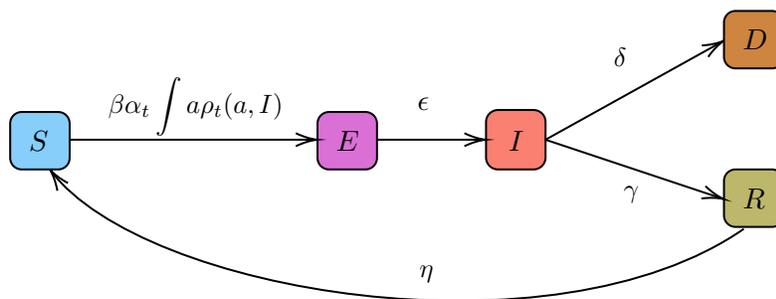
$$c_0(t, p) = c_{inf} p(I)^2, \quad f_0(t, \lambda) = \sum_{i \in \{S, E, I, R\}} \frac{\bar{\beta}^{(i)}}{2} \left(\lambda^{(i)} - \bar{\lambda}^{(i)} \right)^2, \quad C_0(p) = c_d p(D), \quad (4.10)$$

for constant $\bar{\lambda}, \bar{\beta} \in \mathbb{R}_+^m$ and $c_{inf}, c_d > 0$. Compared to previous experiments, the regulator is now paying an additional terminal cost $C_0(p)$ depending on the proportion of deceased people at the end of the time horizon. We set the coefficients c_d and c_D to high values to put more importance to the cost terms related to death. Further we assumed that mortality rate δ is 1%.

As it can be seen in the top plots in Figure 10, playing the mean field Nash Equilibrium under the control of regulator flattens the curve of infections compared to free spread. We also see that when susceptible players do not feel safe enough with the recommended socialization levels they use a lower contact rate. This showcases the ability of the model to capture a population that is more risk-averse than their regulator. On a final note, this experiment shows that agents may not follow exactly what regulator proposes; therefore, the regulator should not assume during policy optimization that the population will strictly obey the announced policy.

Table 3: Parameter values for the SEIRD Stackelberg MFG experiment.

T	p^0	c_λ	c_I	c_{Inf}	$\bar{\beta}$	$\bar{\lambda}$
30	(0.9, 0, 0.1, 0, 0)	10	1	1	(0.2, 0.2, 1, 0)	(1, 1, 0.7, 1)
β	γ	η	ϵ	δ	c_d	c_D
0.25	0.1	0.01	2	0.01	20	20

Figure 9: SEIRD model corresponding to the Q -matrix (4.8).

5 Conclusions

We have studied a Stackelberg mean field game with finite state space using a probabilistic approach. Compared with deterministic approaches using systems of ODEs, this approach has the advantage of describing the evolution of the system from the point of view of a typical (infinitesimal) agent. The theoretical contributions of this paper relate to formulating the Stackelberg game between a principal and a non-cooperative population. To cover the case of populations given by extended MFGs we establish the results in Section 2.2. We have then applied this class of models to a problem of epidemic containment with and SIR-type dynamics. In contrast with the existing literature, our model incorporates at the same time a non-cooperative population and a regulator such as a government. This problem, which can be viewed as a control problem under a constraint given by a Nash equilibrium, is complex because the regulator's decisions influence only indirectly the population's equilibrium. A naive approach would have been to solve a Nash equilibrium for each choice of the regulator's policy but this is computationally prohibitive for our model. Thus, building on the probabilistic approach, we introduced a numerical method based on neural network approximation and Monte Carlo simulations to compute the optimal policy (see Algorithm 1 and 2). We presented several numerical examples and for some of them we managed to derive semi-explicit solutions which can be used as benchmarks (see Proposition 4.1). In particular, we have shown the difference between the uncontrolled scenario, the Nash equilibrium without regulator's intervention, and equilibria arising from early or late lockdowns. We have also shown that the numerical scheme can approximately learn the regulator's optimal policy, including a non-trivial model in Section 4.3. Overall, the numerical approach is able to capture well the evolution of the epidemic in society and provides a satisfactory approximation of the optimal policy of the regulator in presence of a large number of non-cooperative agents.

Several directions are left for future work. For example, on a theoretical side, we may analyze Stackelberg mean field games with interaction through the joint state-action distribution under weaker assumptions. From an applied viewpoint, we may consider more complex finite-state models for instance to add a regulation aspect to SIR-like models appeared recently in the epidemiological literature. We believe that the derivation of the semi-explicit solutions can be generalized to more evolved compartmental models. On the numerical side, the algorithm we proposed is based on approximation by neural networks and could potentially handle even more complex models. This aspect is important for applications to epidemiological models. Some realistic features we could consider in future work are for instance age structure or geography with a network of cities. In both cases, the population is split into more sub-groups, which increases the number of possible states. Another important point is the impact of testing on the ability to take optimal decisions. Indeed, testing is directly related to the uncertainty of the

Stackelberg MFG for Epidemic Control

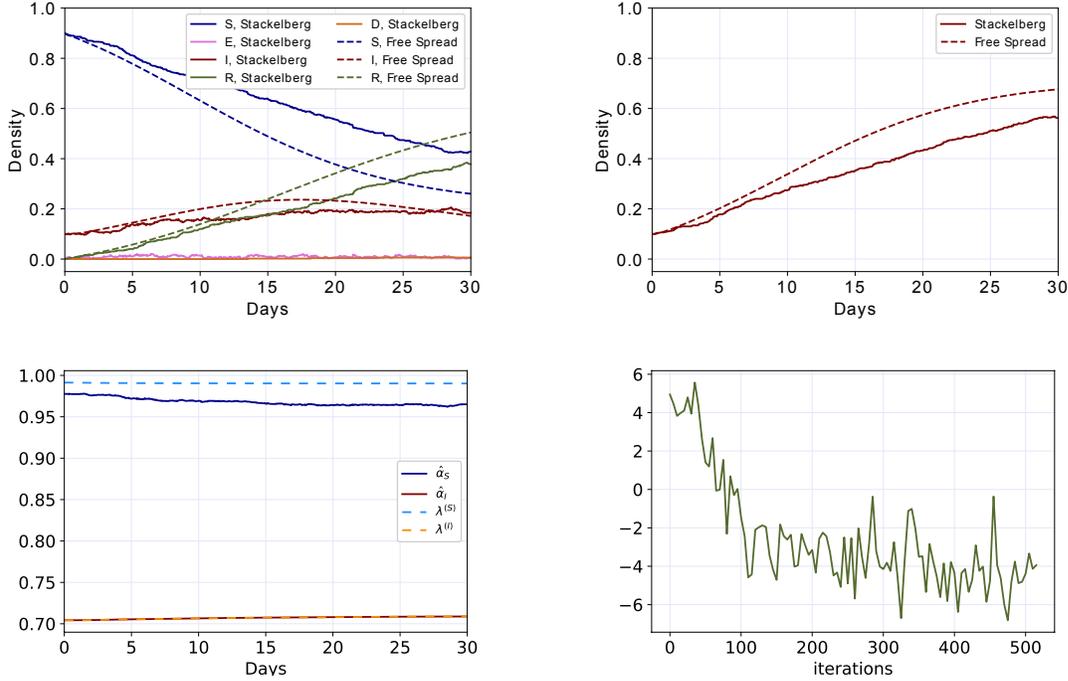


Figure 10: SEIRD Stackelberg MFG with Algorithm 2 in comparison with free spread SEIRD dynamics. Comparison of the Evolution of the state distribution (top left), Comparison of the Cumulative Density of Infected people under Stackelberg MFG and Free Spread (top right); evolution of the controls (bottom left), convergence of the loss (bottom right).

regulator on the population distribution (e.g., to know what is the current proportion of infected people). For this aspect, it seems that a probabilistic approach like the one we adopted is particularly well-suited.

A Proofs for Section 2

Propositions 2.7 and 2.9 are adaptations of [7, Thm. 1] and [7, Thm. 2], respectively, to the extended case, *i.e.*, to the case where the players interact through the joint distribution of the states and the actions and not only through the distribution of the states. For the sake of brevity we omit the proofs.

Proof of Proposition 2.12. We define the representative agent’s Hamiltonian in the regular mean-field game $h(t, x, z, \alpha, p)$ in line with H , but with the “overlined” functions \bar{f}, \bar{Q} replacing f and Q . It follows from the assumptions that $\bar{a}_i(t, z, p)$ is the minimizer of $\alpha \mapsto h(t, e_i, z, \alpha, p)$ for $(t, i, z, p) \in [0, T] \times \{1, \dots, m\} \times \mathbb{R}^m \times \mathcal{P}(E)$. Let $\bar{H}(t, e_i, z, p) := h(t, e_i, z, \bar{a}_i(t, z, p), p)$.

By Proposition 2.12, the pair $(\hat{\alpha}, \hat{p})$ is a mean-field Nash equilibrium. Turning to non-extended mean-field Nash equilibrium, there exists a pair (α', p') satisfying Definition 2.11, see for example [8, Thm 4.1]. Consequently, by [7, Thm. 1], there exists a solution $(Y', Z', \alpha', p', Q')$ to the system (under

\mathbb{P})

$$\begin{cases} Y'_t = U(\xi) + \int_t^T \bar{H}(s, X_{s-}, Z'_s, p'_s) ds - \int_t^T (Z'_s)^* d\mathcal{M}_s \\ \mathcal{E}'_t = 1 + \int_0^t \mathcal{E}'_{s-} X_{s-}^* (\bar{Q}(s, \alpha'_s, p'_s) - Q^0) \psi_s^+ d\mathcal{M}_s, \\ \alpha'_t = \bar{a}(t, X_{t-}, Z'_t, p'_t), \quad p'_t = Q' \circ (X_t)^{-1}, \quad \frac{dQ'}{d\mathbb{P}} = \mathcal{E}'_T, \end{cases} \quad (\text{A.1})$$

such that $\alpha' = \bar{\alpha} d\mathbb{P} \otimes dt$ -a.e. and $p'_t = \bar{p}_t dt$ -a.e..

The claim $\hat{p}_t = \bar{p}_t dt$ -a.e. $t \in [0, T]$ is equivalent to

$$\mathbb{E}^{\hat{\mathbb{Q}}}[X_t] - \mathbb{E}^{Q'}[X_t] = \mathbb{E}^{\mathbb{P}}[(\hat{\mathcal{E}}_t - \mathcal{E}'_t)X_t] = 0, \quad dt\text{-a.e. } t \in [0, T]. \quad (\text{A.2})$$

Let $\hat{\sigma}_t$ and σ'_t be the volatility (viewed as row vectors in \mathbb{R}^m) of $\hat{\mathcal{E}}_t$ and \mathcal{E}'_t , respectively,

$$\hat{\sigma}_t = \hat{\mathcal{E}}_{t-} X_{t-}^* (\bar{Q}(t, \hat{\alpha}_t, \hat{p}_t) - Q^0) \psi_t^+, \quad \sigma'_t = \mathcal{E}'_{t-} X_{t-}^* (\bar{Q}(t, \alpha'_t, p'_t) - Q^0) \psi_t^+, \quad (\text{A.3})$$

and let $\Delta \hat{\mathcal{E}}_t := \hat{\mathcal{E}}_t - \hat{\mathcal{E}}_{t-}$. In the same way we define ΔX_t and $\Delta \mathcal{E}'_t$. Since

$$\Delta \hat{\mathcal{E}}_t = \hat{\sigma}_t \Delta X_t, \quad \Delta \mathcal{E}'_t = \sigma'_t \Delta X_t, \quad (\text{A.4})$$

we have by Ito's formula that under \mathbb{P} ,

$$\begin{aligned} d\|(\hat{\mathcal{E}}_t - \mathcal{E}'_t)X_t\|^2 &= 2 \left(\hat{\mathcal{E}}_{t-} - \mathcal{E}'_{t-} \right)^2 X_{t-}^* Q^0 X_{t-} dt + 2 \left(\hat{\mathcal{E}}_{t-} - \mathcal{E}'_{t-} \right) X_{t-}^* \psi_t (\hat{\sigma}_t - \sigma'_t)^* dt \\ &\quad + \text{Tr} \left[\left(\Sigma_t + (\hat{\mathcal{E}}_t - \mathcal{E}'_t) I_m \right) \psi_t \left(\Sigma_t + (\hat{\mathcal{E}}_t - \mathcal{E}'_t) I_m \right)^* \right] dt + d\widetilde{\mathcal{M}}_t, \end{aligned} \quad (\text{A.5})$$

where $\widetilde{\mathcal{M}}$ is a \mathbb{P} -martingale, I_m is the identity matrix in $\mathbb{R}^{m \times m}$ and

$$\mathbb{R}^{m \times m} \ni \Sigma_t := [X_t(i) (\hat{\sigma}_t(j) - \sigma'_t(j))]_{ij} \quad (\text{A.6})$$

Firstly, we have that $\text{Tr}[\psi_t] = \sum_{i=1}^m [Q^0]_{X_{t-}, i} = 0$, secondly,

$$\text{Tr}[\Sigma_t \psi_t] = \text{Tr}[\psi_t \Sigma_t^*] = (m-1) (\hat{\sigma}_t - \sigma'_t) X_{t-}, \quad (\text{A.7})$$

and finally, $\text{Tr}[\Sigma_t \psi_t \Sigma_t^*] = (m-1) ((\hat{\sigma}_t - \sigma'_t) X_{t-})^2$. We gather that

$$\begin{aligned} &\text{Tr} \left[\left(\Sigma_t + (\hat{\mathcal{E}}_t - \mathcal{E}'_t) I_{m \times m} \right) \psi_t \left(\Sigma_t + (\hat{\mathcal{E}}_t - \mathcal{E}'_t) I_{m \times m} \right)^* \right] \\ &= 2(m-1) (\hat{\sigma}_t - \sigma'_t) (\hat{\mathcal{E}}_{t-} - \mathcal{E}'_{t-}) X_{t-} + (m-1) ((\hat{\sigma}_t - \sigma'_t) X_{t-})^2. \end{aligned} \quad (\text{A.8})$$

Using our calculations above, Young's inequality, Gronwall's lemma, and the Lipschitz continuity of \bar{Q} and \bar{a} (see Hypotheses 2.3 and 2.6), we get that for dt -a.e. $t \in [0, T]$,

$$\begin{aligned} \mathbb{E}^{\mathbb{P}} \left[\|(\hat{\mathcal{E}}_t - \mathcal{E}'_t)X_t\|^2 \right] &\leq C \mathbb{E}^{\mathbb{P}} \left[\int_0^t |\bar{a}(s, X_{s-}, \hat{Z}_s, \hat{p}_s) - \bar{a}(s, X_{s-}, Z'_s, p'_s)|^2 ds \right] \\ &\leq C \mathbb{E}^{\mathbb{P}} \left[\int_0^t \left(\|\hat{Z}_s - Z'_s\|_{X_{s-}}^2 + \|\hat{p}_s - p'_s\|^2 \right) ds \right] \end{aligned} \quad (\text{A.9})$$

for some positive constant C . The estimate holds only dt -a.s. since

$$(\hat{\mathcal{E}}_{t-} - \mathcal{E}'_{t-})X_{t-} = (\hat{\mathcal{E}}_t - \mathcal{E}'_t)X_t, \quad (\text{A.10})$$

an equality we need in order to apply Gronwall's lemma, holds only dt -a.e. $t \in [0, T]$. The constant C depends on $\mathbb{E}^{\mathbb{P}} \left[\int_0^t \mathcal{E}_{t-}^2 dt \right]$, $\mathcal{E} \in \{\hat{\mathcal{E}}, \mathcal{E}'\}$, which is bounded since Q, \bar{Q}, Q^0 , and X_{t-} are bounded. After one more use of Gronwall's lemma, we arrive to

$$\mathbb{E}^{\mathbb{P}} \left[\|(\hat{\mathcal{E}}_t - \mathcal{E}'_t) X_t\|^2 \right] \leq C \mathbb{E}^{\mathbb{P}} \left[\int_0^t \|\hat{Z}_s - Z'_s\|_{X_{s-}}^2 ds \right], \quad dt\text{-a.e. } t \in [0, T]. \quad (\text{A.11})$$

Consider the difference process $(\delta \mathbf{Y}, \delta \mathbf{Z}) := (\hat{\mathbf{Y}} - \mathbf{Y}', \hat{\mathbf{Z}} - \mathbf{Z}')$. The BSDE estimate of [43, Thm. 4.2.3] yields the second inequality below, the first one follows by the equivalence of norms on \mathbb{R}^m :

$$\mathbb{E}^{\mathbb{P}} \left[\int_0^t \|\hat{Z}_s - \bar{Z}_s\|_{X_{s-}}^2 ds \right] \leq C \mathbb{E}^{\mathbb{P}} \left[\int_0^T \|\delta Z_t\|^2 dt \right] \leq C \int_0^T \|\hat{p}_t - p'_t\|^2 dt. \quad (\text{A.12})$$

After one final application of Gronwall's lemma we see that $\hat{p}_t = p'_t$ for dt -a.e. $t \in [0, T]$ and therefore $\hat{p}_t = \bar{p}_t$ for dt -a.e. $t \in [0, T]$.

Finally, we compare the controls and get

$$|\hat{\alpha}_t - \bar{\alpha}_t| \leq C \left(\|\hat{Z}_t - Z'_t\|_{X_{t-}} + \|\hat{p}_t - p'_t\| \right), \quad d\mathbb{P} \otimes dt\text{-a.s.} \quad (\text{A.13})$$

where we used Hypothesis 2.10 to link $\hat{\alpha}$ and $\bar{\alpha}$, and to exploit the Lipschitz continuity of $\bar{\alpha}$. Using the previous calculations, we conclude that $\hat{\alpha}_t = \bar{\alpha}_t$, $d\mathbb{P} \otimes dt$ -a.s.. \square

B Proof for Section 4

Proof of Proposition 4.1. In the setting of the example Hypotheses 2.3–2.5, 2.8, and 2.10 hold true. The minimizers of the reduced Hamiltonians are

$$\begin{aligned} \hat{a}_S(t, z, \rho) &= \lambda_t^{(S)} + \frac{\beta}{c\lambda} \left(\int_A a \rho_t(da, I) \right) (z(S) - z(I)), \\ \hat{a}_S(t, z, \rho) &= \lambda_t^{(I)}, \quad \hat{a}_S(t, z, \rho) = \lambda_t^{(R)}. \end{aligned} \quad (\text{B.1})$$

Imposing the consistency condition on (B.1) we see that the game satisfies hypotheses 2.10(ii) and 2.10(iii). Thus, since we assume hypothesis 2.10(i), Proposition 2.10 says that the mean field Nash equilibrium is almost surely equal to the solution of the regular mean-field game with transition rate matrix

$$\bar{Q}(t, \alpha, p) = \begin{bmatrix} \dots & \beta \alpha \lambda_t^{(I)} p(I) & 0 \\ 0 & \dots & \gamma \\ \eta & 0 & \dots \end{bmatrix}, \quad (\text{B.2})$$

and the minimizers of the reduced Hamiltonians for this regular mean field game are

$$\begin{aligned} \bar{a}_S(t, z, \rho) &= \lambda_t^{(S)} + \frac{\beta}{c\lambda} \lambda_t^{(I)} p(I) (z(S) - z(I)), \\ \bar{a}_S(t, z, \rho) &= \lambda_t^{(I)}, \quad \bar{a}_S(t, z, \rho) = \lambda_t^{(R)}. \end{aligned} \quad (\text{B.3})$$

The principal's problem, after the same rewriting that yielded (2.11), reads

$$\begin{aligned}
 V(\kappa) &= \inf_{\mathbb{E}[Y_0] \leq \kappa} \inf_{\substack{\mathbf{z} \in \mathcal{H}_X^2 \\ \boldsymbol{\lambda} \in \Lambda}} \mathbb{E}^{\mathbb{Q}^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}} \left[\int_0^T \left(c_0(t, p_t^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}) + f_0(t, \lambda_t) \right) dt - Y_T^{\mathbf{z}, \boldsymbol{\lambda}, Y_0} \right] \\
 &= -\kappa + \inf_{\substack{\mathbf{z} \in \mathcal{H}_X^2 \\ \boldsymbol{\lambda} \in \Lambda}} \mathbb{E}^{\mathbb{Q}^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}} \left[\int_0^T \left(c_0(t, p_t^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}) + f_0(t, \lambda_t) \right. \right. \\
 &\quad \left. \left. + \frac{c_\lambda}{2} \left(\lambda_t^{(S)} - \bar{a}_S(t, Z_t, p_t^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}) \right)^2 \mathbb{1}_S(X_{t-}) + c_I \mathbb{1}_I(X_{t-}) \right) dt \right],
 \end{aligned} \tag{B.4}$$

where $-\kappa$ is achieved by $\mathbb{E}[Y_0]$.

The principal's problem $V(\kappa)$ can be recast as an optimization problem over $\mathbb{A} \times \Lambda$. Given $\boldsymbol{\alpha} \in \mathbb{A}$ let $\bar{\mathbf{Z}}$ be given by, for now formally,

$$\bar{Z}_t(\boldsymbol{\alpha}) := \left(\frac{c_\lambda(\alpha_t - \lambda_t^{(S)})}{\beta \lambda_t^{(I)} p_t^{\bar{\mathbf{Z}}(\boldsymbol{\alpha}), \boldsymbol{\lambda}, Y_0}(I)} \mathbb{1}_S(X_{t-}) \mathbb{1}(\lambda_t^{(I)} > 0), 0, 0 \right), \quad t \in [0, T]. \tag{B.5}$$

Lemma B.1. *Given $\boldsymbol{\alpha} \in \mathbb{A}$ and $\boldsymbol{\lambda} \in \Lambda$, let*

$$\Psi(\mathbf{z}) = \left(\frac{c_\lambda(\alpha_t - \lambda_t^{(S)})}{\beta \lambda_t^{(I)} \Phi_t^{\mathbf{z}}(I)} \mathbb{1}_S(X_{t-}) \mathbb{1}(\lambda_t^{(I)} > 0), 0, 0 \right)_{t \in [0, T]} \tag{B.6}$$

for $\mathbf{z} \in \mathcal{H}_X^2$, where $\Phi_t^{\mathbf{z}}$ is given by

$$\Phi_t^{\mathbf{z}} = p_0 + \int_0^t \mathbb{E}^{\mathbb{Q}^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}} [\bar{Q}^*(s, \bar{\alpha}(s, X_{s-}, z_s, \Phi_s^{\mathbf{z}}), \Phi_s^{\mathbf{z}}) X_{s-}] ds. \tag{B.7}$$

Then Ψ is well defined as a mapping from \mathcal{H}_X^2 to itself and there exists a unique fixed point to Ψ . We denote the fixed point of Ψ by $\bar{\mathbf{Z}}(\boldsymbol{\alpha})$ and $\Phi_t^{\bar{\mathbf{Z}}(\boldsymbol{\alpha})}$ by $p_t^{\bar{\mathbf{Z}}(\boldsymbol{\alpha}), \boldsymbol{\lambda}, Y_0}$. Furthermore, for all $t \in [0, T]$,

$$\bar{a}(t, X_{t-}, \bar{\mathbf{Z}}_t(\boldsymbol{\alpha}), p_t^{\bar{\mathbf{Z}}(\boldsymbol{\alpha}), \boldsymbol{\lambda}, Y_0}) = \alpha_t \mathbb{1}_S(X_{t-}) + \lambda_t^{(I)} \mathbb{1}_I(X_{t-}) + \lambda_t^{(R)} \mathbb{1}_R(X_{t-}). \tag{B.8}$$

Proof. The first observation is there exists a uniform lower bound c such that $\Phi_t^{\mathbf{z}}(i) \geq c > 0$ for $i \in \{S, I, R\}$ and $t \in (0, T]$. This grants the first assertion of the lemma, that $\Psi(\mathbf{z}) \in \mathcal{H}_X^2$ for all $\mathbf{z} \in \mathcal{H}_X^2$. Secondly, we have that

$$\left| \Phi_t^{\mathbf{z}}(I)^{-1} - \Phi_t^{\mathbf{z}'}(I)^{-1} \right|^2 \leq C \mathbb{E} \left[\int_0^t \|z_s - z'_s\|^2 dt \right] \tag{B.9}$$

for some positive constant C . To get (B.9) we used Gronwall's lemma, Cauchy-Schwartz inequality, and the boundedness of the coefficients, $\boldsymbol{\lambda}$, and the likelihood process $d\mathbb{Q}^{\mathbf{z}, \boldsymbol{\lambda}, Y_0}/d\mathbb{P}$. From (B.9) follows that

$$\mathbb{E} \left[\|\Psi_t(\mathbf{z}) - \Psi_t(\mathbf{z}')\|^2 \right] \leq C \mathbb{E} \left[\int_0^t \|z_s - z'_s\|^2 dt \right], \quad t \in [0, T]. \tag{B.10}$$

and hence Ψ^N is a contraction mapping for sufficiently large N (by equivalence of norms on \mathbb{R}^m). The existence of a unique fixed point of Ψ follows by the Banach fixed-point theorem for iterated mappings. Finally, (B.8) follows by plugging in $\bar{\mathbf{Z}}(\boldsymbol{\alpha})$ and $p^{\bar{\mathbf{Z}}(\boldsymbol{\alpha}), \boldsymbol{\lambda}, Y_0}$ into (B.3). \square

In light of Lemma B.1 and since Q and $\bar{\alpha}$ do not depend on the representative agent's expected total cost Y_0 , the principal's problem can be transformed to

$$W = \inf_{\alpha \in \bar{\mathbb{A}}, \lambda \in \Lambda} I(\alpha, \lambda),$$

$$I(\alpha, \lambda) := \mathbb{E}^{\mathbb{Q}^{\alpha, \lambda}} \left[\int_0^T \left(c_0(t, p_t^{\alpha, \lambda}) + f_0(t, \lambda_t) + \frac{c_\lambda}{2} \left(\lambda_t^{(S)} - \alpha_t \right)^2 \mathbb{1}_S(X_{t-}) + c_I \mathbb{1}_I(X_{t-}) \right) dt \right]. \quad (\text{B.11})$$

The measure $\mathbb{Q}^{\alpha, \lambda}$ is such that the coordinate process \mathbf{X} has transition rate matrix $\bar{Q}(t, \alpha_t, p^{\alpha, \lambda})$ under it, $p_t^{\alpha, \lambda}$ is the law of X_t under $\bar{Q}(t, \alpha_t, p^{\alpha, \lambda})$.

From the decomposition of the coordinate process,

$$X_t = X_0 + \int_0^t \bar{Q}^*(s, \alpha_s, p_s^{\alpha, \lambda}) X_{s-} ds + \mathcal{M}_t^{\alpha, \lambda}, \quad (\text{B.12})$$

we deduce the dynamic of $p^{\alpha, \lambda}$:

$$p_t^{\alpha, \lambda} = p^0 + \int_0^t \mathbb{E}^{\mathbb{Q}^{\alpha, \lambda}} [\bar{Q}^*(s, \alpha_s, p_s^{\alpha, \lambda}) X_{s-}] ds, \quad (\text{B.13})$$

where we see that the right-hand side depends on the control α_s only through the conditional expectation $\tilde{\alpha}_s := \mathbb{E}^{\mathbb{Q}^{\alpha, \lambda}}[\alpha_s \mid X_{s-} = S]$, $s \in [0, T]$. Let

$$\tilde{I}(\tilde{\alpha}, \lambda) := \int_0^T \left(c_0(t, p_t^{\alpha, \lambda}) + f_0(t, \lambda_t) + \frac{c_\lambda}{2} \left(\lambda_t^{(S)} - \tilde{\alpha}_t \right)^2 \mathbb{1}_S(X_{t-}) + c_I \mathbb{1}_I(X_{t-}) \right) dt \quad (\text{B.14})$$

and consider the deterministic control problem

$$\tilde{W} := \inf_{\tilde{\alpha} \in \tilde{\mathbb{A}}, \lambda \in \Lambda} \tilde{I}(\tilde{\alpha}, \lambda) \quad (\text{B.15})$$

where $\tilde{\mathbb{A}}$ is the collection of all measurable mappings from $[0, T]$ to A .

Lemma B.2. *We have $W = \tilde{W}$. If $(\tilde{\alpha}, \lambda)$ is a solution to the optimization problem \tilde{W} , then the predictable process α defined by $\alpha_t = \sum_{i \in \{S, I, R\}} \mathbb{1}_i(X_{t-}) \tilde{\alpha}_t^{(i)}$ together with λ is an optimal control for W . Furthermore, an optimal control exists for \tilde{W} .*

The proof readily follows by Proposition 1 and Lemma 3 of [7].

We apply the necessary part of the Pontryagin maximum principle to characterizes the solution to the optimal control problem \tilde{W} and the corresponding flow of probability measures. The Hamiltonian for the problem \tilde{W} is \tilde{H} , defined in (4.3). It is straight forward to obtain the first order optimality conditions $(\nabla_{\tilde{\alpha}}, \nabla_\lambda) \tilde{H} = 0$. After some tedious calculation we reach (4.4). This concludes the proof of Proposition 4.1. \square

References

- [1] E. BAYRAKTAR, A. BUDHIRAJA, AND A. COHEN, *A numerical scheme for a mean field game in some queueing systems based on Markov chain approximation method*, SIAM J. Control Optim., 56 (2018), pp. 4017–4044.
- [2] E. BAYRAKTAR AND A. COHEN, *Analysis of a finite state many player game using its master equation*, SIAM Journal on Control and Optimization, 56 (2018), pp. 3538–3568.

- [3] A. BENSOUSSAN, M. H. M. CHAU, AND S. C. P. YAM, *Mean field Stackelberg games: aggregation of delayed instructions*, SIAM J. Control Optim., 53 (2015), pp. 2237–2266.
- [4] R. CARMONA AND F. DELARUE, *Extensions for Volume I*, Springer International Publishing, Cham, 2018, pp. 619–680.
- [5] R. CARMONA AND M. LAURIÈRE, *Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games: II—The Finite Horizon Case*. preprint, 2019.
- [6] R. CARMONA AND P. WANG, *Finite state mean field games with major and minor players*. preprint, 2016.
- [7] R. CARMONA AND P. WANG, *Finite-State Contract Theory with a Principal and a Field of Agents*. preprint, 2018.
- [8] R. CARMONA AND P. WANG, *A probabilistic approach to extended finite state mean field games*. preprint, 2018.
- [9] A. CECCHIN AND M. FISCHER, *Probabilistic approach to finite state mean field games*, Applied Mathematics & Optimization, (2018), pp. 1–48.
- [10] P. CHAN AND R. SIRCAR, *Bertrand and Cournot mean field games*, Applied Mathematics & Optimization, 71 (2015), pp. 533–569.
- [11] A. CHARPENTIER, R. ELIE, M. LAURIÈRE, AND V. C. TRAN, *Covid-19 pandemic control: balancing detection policy and lockdown intervention under icu sustainability*, arXiv preprint [arXiv:2005.06526](https://arxiv.org/abs/2005.06526), (2020).
- [12] S. CHO, *Mean-Field Game Analysis of SIR Model with Social Distancing*. preprint, 2020.
- [13] S. E. CHOUTRI AND B. DJEHICHE, *Mean-field risk sensitive control and zero-sum games for Markov chains*, Bulletin des Sciences Mathématiques, 152 (2019), pp. 1–39.
- [14] S. E. CHOUTRI, B. DJEHICHE, AND H. TEMBINE, *Optimal control and zero-sum games for Markov chains of mean-field type*, Mathematical Control & Related Fields, 9 (2019), p. 571.
- [15] S. E. CHOUTRI AND T. HAMIDOU, *A stochastic maximum principle for Markov chains of mean-field type*, Games, 9 (2018), p. 84.
- [16] J. CVITANIĆ, D. POSSAMAÏ, AND N. TOUZI, *Dynamic programming approach to principal–agent problems*, Finance and Stochastics, 22 (2018), pp. 1–37.
- [17] B. DJEHICHE AND P. HELGESSON, *The principal-agent problem; a stochastic maximum principle approach*. preprint, 2014.
- [18] J. DONCEL, N. GAST, AND B. GAUJAL, *A Mean-Field Game Analysis of SIR Dynamics with Vaccination*. preprint, 2017.
- [19] R. ELIE, E. HUBERT, AND G. TURINICI, *Contact rate epidemic control of COVID-19: an equilibrium view*, Mathematical Modelling of Natural Phenomena, 15 (2020), p. 35.
- [20] R. ELIE, T. MASTROLIA, AND D. POSSAMAÏ, *A tale of a principal and many, many agents*, Mathematics of Operations Research, 44 (2019), pp. 440–467.
- [21] B. GAUJAL, J. DONCEL, AND N. GAST, *Vaccination in a Large Population: Mean Field Equilibrium versus Social Optimum*, in netgcoop’20, Cargèse, France, Sept. 2021.
- [22] D. A. GOMES, J. MOHR, AND R. R. SOUZA, *Discrete time, finite state space mean field games*, Journal de mathématiques pures et appliquées, 93 (2010), pp. 308–328.

- [23] ———, *Continuous time finite state mean field games*, Applied Mathematics & Optimization, 68 (2013), pp. 99–143.
- [24] D. A. GOMES, S. PATRIZI, AND V. VOSKANYAN, *On the existence of classical solutions for stationary extended mean field games*, Nonlinear Analysis: Theory, Methods & Applications, 99 (2014), pp. 49–79.
- [25] B. HOLMSTROM AND P. MILGROM, *Aggregation and linearity in the provision of intertemporal incentives*, Econometrica: Journal of the Econometric Society, (1987), pp. 303–328.
- [26] M. HUANG, R. P. MALHAMÉ, P. E. CAINES, ET AL., *Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle*, Communications in Information & Systems, 6 (2006), pp. 221–252.
- [27] E. HUBERT, T. MASTROLIA, D. POSSAMAÏ, AND X. WARIN, *Incentives, lockdown, and testing: from Thucydides’s analysis to the COVID-19 pandemic*. preprint, 2020.
- [28] E. HUBERT AND G. TURINICI, *Nash-MFG equilibrium in a SIR model with time dependent newborn vaccination*, Ricerche di Matematica, 67 (2018), pp. 227–246.
- [29] Z. KOBEISSI, *On classical solutions to the mean field game system of controls*. preprint, 2019.
- [30] V. N. KOLOKOLTSOV, *Nonlinear Markov games on a finite state space (mean-field and binary interactions)*, International Journal of Statistics and Probability, 1 (2012), pp. 77–91.
- [31] T. G. KURTZ, *Approximation of population processes*, vol. 36, SIAM, Philadelphia, PA, 1981.
- [32] L. LAGUZET, G. TURINICI, AND G. YAHIAOUI, *Equilibrium in an individual-societal SIR vaccination model in presence of discounting and finite vaccination capacity*, in New Trends in Differential Equations, Control Theory and Optimization: Proceedings of the 8th Congress of Romanian Mathematicians, World Scientific, 2016, pp. 201–214.
- [33] J.-M. LASRY AND P.-L. LIONS, *Jeux à champ moyen. i—le cas stationnaire*, Comptes Rendus Mathématique, 343 (2006), pp. 619–625.
- [34] ———, *Jeux à champ moyen. ii—horizon fini et contrôle optimal*, Comptes Rendus Mathématique, 343 (2006), pp. 679–684.
- [35] M. LAURIÈRE AND L. TANGPI, *Backward propagation of chaos*. preprint, 2019.
- [36] ———, *Convergence of large population games to mean field games with interaction through the controls*. preprint, 2020.
- [37] W. LEE, S. LIU, H. TEMBINE, W. LI, AND S. OSHER, *Controlling Propagation of epidemics via mean-field games*. preprint, 2020.
- [38] R. SALHAB, R. P. MALHAMÉ, AND J. LE NY, *A dynamic collective choice model with an advertiser*, Dynamic Games and Applications, 8 (2018), pp. 490–506.
- [39] F. SALVARANI AND G. TURINICI, *Optimal individual strategies for influenza vaccines with imperfect efficacy and durability of protection*, Mathematical Biosciences & Engineering, 15 (2018), p. 629.
- [40] Y. SANNIKOV, *A Continuous- Time Version of the Principal: Agent Problem*, The Review of Economic Studies, 75 (2008), pp. 957–984.
- [41] Y. SANNIKOV, *Contracts: The Theory of Dynamic Principal–Agent Relationships and the Continuous-Time Approach*, vol. 1 of Econometric Society Monographs, Cambridge University Press, 2013, p. 89–124.

- [42] H. TEMBINE, *COVID-19: A Data-Driven Mean-Field-Type Game Perspective*. preprint, 2020.
- [43] J. ZHANG, *Backward stochastic differential equations*, in *Backward Stochastic Differential Equations*, Springer, 2017, pp. 79–99.