COMBAT EPIDEMICS WITH REINFORCEMENT LEARNING

Yuxi Li (yuxili@gmail.com)
June 9, 2020

ABSTRACT

The coronavirus disease 2019 will have a huge impact on the economy and the society worldwide. It will last a long time and the next pandemic will come surely. It is pressing and critical to make informed and intelligent decisions. We argue that reinforcement learning is a promising framework to combat epidemics; and epidemics are a fruitful application area for reinforcement learning to make substantial real life impact. We present a simplistic model, with discussions and potential extensions. We hope to draw more attention from communities of reinforcement learning and artificial intelligence, epidemiology and public health, and economics to combat epidemics together.

1 INTRODUCTION

Sustained severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) outbreak has resulted in titanic confirmed cases and deaths of coronavirus disease 2019 (COVID-19). The World Health Organization (WHO) declared on March 11, 2020 the SARS-CoV-2 outbreak as a pandemic, the most severe level of epidemics. It will have a huge impact on the economy and the society worldwide. Worse yet, more and more evidence show that it will likely stay for a long time (Kissler et al., 2020), before effective vaccine and treatment become available, and it is costly to establish herd immunity so that most of a population gain immunity and indirectly protect those not immune.[1] It is thus pressing and critical for countries to respond to the pandemic in an informed and intelligent way. Recently, Shapiro and Rothman (2020) warn that "Perhaps the most essential lesson ... is this: the next pandemic will surely come. The only question is when. AHSs and others must be better prepared for that moment." Here AHSs stands for academic health systems. Actually, much earlier, experts suggested to enforce global monitoring of disease outbreak preparedness to prevent a potential pandemic, e.g., Harvard Global Health Institute (2018).

COVID-19 has drawn much attention in the research communities. There are growing number of articles published in leading journals like New England Journal of Medicine (NEJM),[2] Lancet,[3] Journal of American Medical Association (JAMA),[4] Science,[5] Nature, as well as in preprint platforms like arXiv, bioRxiv and medRxiv, contributed by researchers from epidemiology, public health, economics,[6][7] and artificial intelligence (AI) and machine learning (Bullock et al., 2020).

In this article, we focus on the control of epidemics. Most previous works study the transmission of an infectious disease and intervention measures with an epidemiological transmission model, either a classical susceptible-infectious-recovered (SIR) model or its extension SEIR model with the addition of an exposed compartment or some variants. Usually a simulator is built based on such a transmission model, and various manually designed intervention measures are implemented in the simulator to study morbidity and mortality. Most papers from epidemiology and public health about controlling epidemics follow this way. Such study provides guidance on policy making for public

[1] https://hub.jhu.edu/2020/04/30/herd-immunity-covid-19-coronavirus/

[2] https://www.nejm.org/coronavirus

[3] https://www.thelancet.com/coronavirus

[4] https://jamanetwork.com/journals/jama/pages/coronavirus-alert

[5] https://www.sciencemag.org/collections/coronavirus

[6] https://www.brookings.edu/topic/coronavirus-covid19/

[7] https://voxeu.org/pages/covid-19-page

health and economy. However, it is time-consuming to design strategies manually, which are usually in fact sub-optimal. One potential refinement is to automate the evaluation and optimization of intervention strategies. Moreover, there are investigations about both an epidemic and the economy, e.g. Alvarez et al. (2020), Charpentier et al. (2020), Eichenbaum et al. (2020), Lin et al. (2010), and an edited book, Baldwin and di Mauro (2020). Usually, simplified assumptions are necessary to combine a transmission model with an economic model. One potential refinement is thus to relax assumptions and consider practical scenarios. When working with simulation, it is straightforward to consider realistic factors, like an optimal strategy of additional hospital facilities and heterogeneity at individual level, which may be nontrivial to be incorporated in a model. To this end, we introduce the framework of reinforcement learning (RL) (Sutton and Barto, 2018; Szepesvári, 2010) to combat an epidemic and its impact on the economy. RL deals with sequential decision making, by automating previously manually designed strategies, and it is straightforward to consider practical scenarios, esp. with simulation.

In the following, we present a brief review of related work, background about epidemiological model and RL, a simplistic RL formulation to combat an epidemic and its impact on economy, and discussions and extensions.

2 A BRIEF REVIEW

RL has made significant progress recently. Remarkable examples are AlphaGo (Silver et al., 2016) and AlphaGo Zero (Silver et al., 2017) for computer Go, AlphaZero (Silver et al., 2018) for two-player perfect information games and AlphaStar (Vinyals et al., 2019) for StarCraft, a multi-player imperfect information game. Besides games, RL also has various successful applications, e.g., in recommendation (Agarwal et al., 2016; Gauci et al., 2019; Ie et al., 2019), robotics (Peng et al., 2020), ride-sharing (Tang et al., 2019), chemical synthesis (Segler et al., 2018), and drug design (Zhavoronkov et al., 2019). There are also efforts applying RL to healthcare, e.g., dynamic treatment regimes (Shortreed et al., 2011; Zhang and Bareinboim, 2019), weaning of mechanical ventilation in ICU (Prasad et al., 2017), mobile health (Liao et al., 2020; Menictas et al., 2019), laboratory tests (Cheng et al., 2019), and individualized sepsis treatment (Komorowski et al., 2018). Moreover, RL has been making achievements in automatic machine learning (AutoML), e.g., neural architecture search (Zoph and Le, 2017), algorithm design (Li and Malik, 2017; Xu et al., 2018), combinatorial optimization (Chen and Tian, 2019; Lu et al., 2020), and data augmentation in computer vision (Cubuk et al., 2019). Note, there are discussions about issues with applying (deep) RL, e.g., Dulac-Arnold et al. (2019); Henderson et al. (2018); Gottesman et al. (2018; 2019; 2020).

Several papers study SEIR models at population level. Hellewell et al. (2020) study the effectiveness of contact tracing and isolation with simulation based on a stochastic transmission model, considering scenarios with various initial cases, basic reproduction number R_0, the delay from symptom onset to isolation, contract tracing probability, the proportion of asymptomatic transmission, and the proportion of subclinical infections. Kucharski et al. (2020) study the combination of a stochastic transmission model with multiple datasets to estimate the dynamics of transmission during January and February 2020 in Wuhan, and evaluate the potential occurrence of sustained spread in other locations if cases were introduced. Kissler et al. (2020) combine viral, environmental, and immunologic factors like the degree of seasonal variation in transmission, the duration of immunity, the degree of cross-immunity between SARS-CoV-2 and other coronaviruses, and the intensity and timing of intervention measures to study the transmission dynamics. The authors identify that the status of overwhelming of critical care capacity is critical for the success of social distancing, and that intermittent or prolonged social distancing may be necessary into 2022 to avoid exceeding healthcare capacity. Prem et al. (2020) and Liu et al. (2020) investigate SEIR transmission models with age- and location-specific social mixing, with different ways to estimate contact matrices. Leung et al. (2020) study first-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning.

There are a couple of papers studying SEIR models at individual level, e.g., Ferguson et al. (2020) and Koo et al. (2020). Ferguson et al. (2020) study the effect of non-pharmaceutical interventions (NPIs) for controlling a pandemic in the settings of UK and US. This study influenced the policy making in UK, US and other countries.[8] Individual-level simulation makes investigations at fine granularities; however, it may require costly computation.

There are investigations about both an epidemic and the economy, e.g. Alvarez et al. (2020), Charpentier et al. (2020), Eichenbaum et al. (2020), Lin et al. (2010), and an edited book, Baldwin and di Mauro (2020). Alvarez et al. (2020) study an optimal lockdown intensity and duration policy considering the fatalities of an epidemic and the cost of the lockdown, by a SIR epidemiology model and a linear economy model. The authors embed the fraction of lockdown into the SIR model and the objective function, formulate a Bellman-Hamilton-Jacobi equation, discretize it, and solve it with value iteration. The approach is simplified by considering only the fraction of lockdown. Alvarez et al. (2020) share similarity with Lin et al. (2010), which apply optimal control to an SIR model, using a decision variable to reduce the rate of contact, which is not directly achievable, whereas a lockdown ratio may be directly achievable. Charpentier et al. (2020) propose a variant of SIR model with compartments of susceptible, infected non-detected, infected detected, recovered non-detected, recovered detected, hospitalized, hospitalized in ICU, and dead. Charpentier et al. (2020) study optimal decision with lockdown and detection, with experiments under conditions of 50% increase of ICU capacity in two months among others. Alvarez et al. (2020), Charpentier et al. (2020) and Lin et al. (2010) normalize population to 1, which is convenient for theoretical analysis; however, this may make it inconvenient to consider some realistic factors, e.g., additional hospital facility, not mentioning different, customized NPIs for different regions in a city. Eichenbaum et al. (2020) propose SIR-macro model to consider both an epidemic and an economy, considering factors like consumption, production, medical preparedness, treatment, vaccination, etc. To build the model in a parsimonious way, some simplified assumptions are made, e.g., the absence of heterogeneity in consumption and production. One strength of RL is that it can incorporate realistic factors in a straightforward way.

3 BACKGROUND

We present background about epidemiological model and RL.

3.1 SIR MODEL

In the classic SIR model, a population is categorized into S, I, and R compartments, which respectively stands for susceptible, infectious, and recovered or dead individuals. An important extension to SIR is SEIR model, where E stands for exposed compartment, for individuals exposed to an infectious disease but not infectious yet. In the SIR model, susceptible individuals may acquire the infectious disease at a certain rate when they contact an infectious person, become infectious, and later either recover or decease. In SEIR, susceptible individuals enter the exposed state first, then become infectious. The SIR/SEIR model or their variants also specify transmission dynamics among compartments. See e.g., Ferguson et al. (2020); Hellewell et al. (2020); Kissler et al. (2020); Koo et al. (2020); Leung et al. (2020); Prem et al. (2020); Wu et al. (2020).

In an SIR model with lockdown, the dynamics follows. We have N, S, I, R, and D, denoting the numbers of total, susceptible, infectious, recovered, and dead people, respectively. This model is a variant of the standard SIR model, with the addition of the death compartment. We deploy an abstract intervention measure, lockdown, and use $L(t)$ to denote the ratio of those susceptible and infectious who are locked down. We assume $L(t)$ is the effective lockdown ratio. We recover the classic SIR model when the lockdown ratio $L(t)$ is always 0. We assume recovered people are permanently immune, i.e., they will not be infected anymore. As a result, it is not necessary to lock down recovered people. We assume perfect testing and tracing, so that recovered people will not be locked down. Here β is the transmission rate or contact rate. The force of infection is given by $\beta S(1-L(t))I(1-L(t))/N$. We denote the mean recovery rate as γ, in 1/days; and denote the case fatality ratio (CFR) as ϕ. We can assume a proper birth rate to make N a constant, and new births are in the compartment of susceptible.

$$\frac{dS}{dt} = -\beta \frac{S(1-L(t))I(1-L(t))}{N}$$

[8] https://www.nature.com/articles/d41586-020-01003-6

$$\frac{dI}{dt} = \beta \frac{S(1 - L(t))I(1 - L(t))}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I(1 - \phi)$$

$$\frac{dD}{dt} = \gamma I \phi$$

3.2 RL

In the following, we briefly introduce RL. An RL agent interacts with an environment over time. At each time step t, the agent receives a state s_t in a state space $\mathcal{S}$, and selects an action a_t from an action space $\mathcal{A}$, following a policy $\pi(a_t|s_t)$, which is the agent's behavior, i.e., a mapping from state s_t to actions a_t. The agent receives a scalar reward r_{t+1}, and transitions to the next state s_{t+1}, according to the environment model, which refers to the state transition probability $\mathcal{P}(s_{t+1}|s_t, a_t)$ and the reward function $\mathcal{R}(r_{t+1}|s_t, a_t)$. The return is the discounted accumulated reward with the discount factor $\gamma \in (0, 1]$, $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$.[9] The agent aims to maximize the expectation of such long-term return from each state, or each state action pair. The problem is set up in discrete state and action spaces. It is not hard to extend it to continuous spaces. In partially observable environments, an agent cannot observe states fully, but has observations. As a result, we replace state s_t with observation o_t in previous discussion.

When an RL problem satisfies the Markov property, i.e., the future depends only on the current state and action, but not on the past, it is formulated as a Markov decision process (MDP), defined by $\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}$. When the system model is available, we may use dynamic programming methods: policy evaluation to calculate value/action value function for a policy, value iteration or policy iteration for finding an optimal policy. When there is no model, we resort to RL methods, like Q-learning and policy gradient, and their variants. RL also works when the model is available. RL framework is widely applicable. An RL environment can be an MDP, a partially observable MDP (POMDP), a multi-armed bandit, a game, etc.

4 RL Formulation: A Simplistic Model

In this section, we formulate the control of an epidemic and its impact on the economy in RL framework. We start with a simplistic yet illustrative model, influenced by Alvarez et al. (2020). It is straightforward to build a simulator based on an SIR/SEIR model and intervention measures, and find an optimal policy/intervention measure with RL.

4.1 A simplistic model

To formulate a problem in the RL framework, we need to define state, action, and reward. By simulating the interaction between an agent (the decision maker) and an environment (the population and the economy in the society), we can collect training data and find an optimal policy (sequential intervention measures) with some RL algorithm.

It is desirable to have a compact yet informative state representation. In a simplistic version, we represent a state as a vector, consisting of the numbers of susceptible, infectious, and recovered.

An action is an intervention measure. We use lockdown ratios for the action space, following Alvarez et al. (2020).

We define a reward function in the form of: production - healthcare cost - death cost. In particular, we have: # of working people * w - # of newly recovered * healthcare cost - # of newly dead * fatality cost. We may regard w as wage or willingness to pay.

4.2 Parameters Setting

In the following, we discuss parameters setting to implement the simplistic model.

[9] In epidemiology literature, the recovery rate is denoted as γ, the same as discount factor in RL. We need to differentiate them.

We set the recovery rate γ as one over infectious days. We set infectious days as 5, as in Kissler et al. (2020), thus $\gamma = 1/5$. In Alvarez et al. (2020), $\gamma = 1/18$. We set $\beta = R_0\gamma$, as in Kissler et al. (2020). Here R_0 is the basic reproduction number. We set R_0 as 2.2, following Prem et al. (2020). We set CFR $\phi = 0.05$. We will vary these parameters for sensitive analysis.

We use 8 actions for each state, from 0.0 to 0.7. We assume 30% of workforce (e.g., health-care personnel, police, staff for basic utilities, and other essential government staff) are working even during the extreme control measures.

As in Alvarez et al. (2020), one way to interpret w is GDP per capita, let say $65,000 USD; and the shadow cost of each life lost is 20 times annual GDP per capita, or about $1.3 million. The healthcare cost is around 1/10 of GDP per capita; see Bartsch et al. (2020). When we formulate in a daily basis, assuming 250 working days, and normalizing w as 1, we have healthcare cost as 25 and fatality cost as 50,000. The reward function then is: # of working people * 1 - # of newly recovered * 25 - # of newly dead * 50000.

We have worked on toy problems, with 30 people, with tabular state, action, value, and policy representations. The results are expected — RL finds better policies than policies with fixed lockdown ratios.

5 Discussion/Extensions

Our goal is to integrate RL with an epidemiological transmission model, combat an epidemic and its impact on the economy, by finding an optimal intervention policy. We first discuss if the model presented in Section 4 is too simple. We then present potential extensions. Note, these extensions may be complementary to each other, e.g., we may combine extensions about RL with those about transmission model. We also discuss if simulation is sufficient for combating epidemics.

5.1 Is the model too simple?

The presented RL + SIR model is apparently very simple. However, it has its power and beauty.

The lockdown ratio is abstract, so it generalizes well to various scenarios, so that when interventions can achieve an optimal lockdown ratio, i.e., to implement an optimal action in RL, then the realized policy can achieve an optimal objective. Different countries, states/provinces, counties, towns may have different technical, geographical, societal, cultural, and political conditions, so they may have different ways to implement intervention measures to achieve a certain lockdown ratio.

For example: we may have the following non-pharmaceutical interventions (NPIs): isolation of cases, home quarantine of susceptible people, home quarantine of elderly people, schools and universities closure, bars and restaurants closure, sports closure, conferences closure, gatherings ban. Each of them can achieve certain level of lockdown ratio, and a decision maker can choose some combination of them to achieve a target lockdown ratio. When we treat each of these NPIs as an action, we need some additional selecting criteria, since some of them may have equivalent effect on lockdown ratio. For example, if closing schools or closing bars and restaurants can achieve the same lockdown ratio, we need to set different priorities for education and drinking/dinning to select one NPI over another.

Reducing the effective reproduction number R_t is essential in containing an epidemic. It is desirable to control R_t directly, e.g, to add a constraint of $R_t < 1$ in the optimization. However, it may not be feasible to implement it, since we need to estimate R_t in the first place. It is relatively straightforward to implement a lockdown ratio, which reduces R_t. Continuing the above example, different NPIs may achieve the same lockdown ratio, as a result, having the same effect on reducing R_t.

5.2 Extension 1: RL

It is necessary to deploy function approximation for large scale problems. One way is to deploy neural networks for the representation with the simplistic model. We may start with representing state with numbers of S, I, R in a single day, to model realistic number of people, like one million.

We should study a state representation considering history, e.g., numbers of S, I, R and actions taken, during, say last 14 days. A RNN, like LSTM, is a great tool for this. See Li (2017) for an overview about deep RL.[10]

When we deploy neural networks, one major issue is interpretability. We may find intuition of the resulting policy/value. An optimal value/policy should behave reasonably, with "smooth" functions. We can visually check value and policy (to some degree) and compare with manually designed policies, like fixed lockdown ratios, or some rule-based policies (e.g., from the literature). This partially provides explainability, although when using neural networks, we may not be able to explain exactly/clearly why a particular action is selected. We may derive an intuitive policy during post-processing.

Our goal is by nature multi-objective optimization, namely, simultaneously optimizing the infection rate and the economic impact. It is thus desirable to borrow ideas from recent achievements in this direction, see e.g., Szepesvári (2020a).

There are usually uncertainty in parameter estimation, e.g., for R_0, CFR, etc. As a result, there are uncertainties in the transmission model for the virus and the transition model for RL. We can vary parameters for sensitivity analysis and consider heterogeneity of these parameters, e.g., for R_0 in Donnat and Holmes (2020). See a recent talk about model misspecification in RL (Szepesvári, 2020b).

It is helpful to incorporate risk management tools like conditional Value-at-Risk (CVaR) (Yu et al., 2009) and extreme value theory (EVT) for fat tails (Cirillo and Taleb, 2020). We may deploy meta-learning and few-shot learning to make the learned policy robust to parameter uncertainty.

The lockdown ratio is continuous by nature. Other NPIs may also be continuous. It is helpful to consider algorithms like soft actor-critic (Haarnoja et al., 2018) and TD3 (Fujimoto et al., 2018).

Note, the numbers of S, E, I, R should be partially observable in practice, due to testing capacity, willingness to test, etc., although (most if not all) SIR/SEIR models treat them as fully observable in simulation. However, one issue of regarding S, E, I, R as latent variables is we probably do not have the true values, but only those reported. Thus we may follow the tradition in epidemiology/public health and use SIR/SEIR models.

5.3 Extension 2: Transmission model

There are realistic factors of a pathogen like SARS-CoV-2 to consider when building a transmission model. It is desirable to study the heterogeneity of the basic reproduction number R_0, see e.g., Donnat and Holmes (2020), and bias in estimation of case fatality ratio (CFR), see e.g., Angelopoulos et al. (2020). An incubation period is the time between exposed to a pathogen and onset of symptoms. Infection has a duration, and individual infectiousness is variable. The serial interval is the time between successive cases in a sequence of transmissions. Asymptomatic or subclinical people may be infectious, and people may become infectious after recovered, for SARS-CoV-2. Infection usually comes with seasonal variation, e.g., strong during winter while weak during summer. Immunity usually has a duration, e.g., 40 weeks. Cross-immunity has various degrees, e.g., people immune to other betacoronavirus like HCov-OC43 and HCov-HKU1, which cause influenza, may be immune to SARS-CoV-2, but with different degrees. For more information, see literature in epidemiology/public health modelling, like Ferguson et al. (2020); Hellewell et al. (2020); Kissler et al. (2020); Koo et al. (2020); Leung et al. (2020); Prem et al. (2020); Wu et al. (2020), and epidemiological and clinical characteristics of the virus and the disease, like Arons et al. (2020); Guan et al. (2020); Richardson et al. (2020); Wu and McGoogan (2020); Young et al. (2020).

[10] We may want to know the numbers of S, I, R, rather than just percentages/ratios to the population, when we represent states, reward and policy, then we need function approximation. We may use linear function approximation to represent states with numbers of S, I, R in a single day, for potentially better interpretability. Considering history may help uncover some latent pattern in the data; in some sense, it is about partial observability. Linear function approximation may not be easy to represent a history. And neural networks like RNN appear as a "natural" choice. Note that neural networks are "magic" in some sense. However, if it is a convenient tool to make our model work (empirically), it is desirable to consider it.

5.4 Extension 3: Age- and location-specific SEIR model

Contact patterns of different ages, e.g., young, adult, and elderly, and locations, e.g., home, schools, workplace, and community, affect person-to-person transmission. During an on-going outbreak, such contact patterns will change from normal conditions, due to physical distancing measures, behavioural habit changes, etc. It is desirable to incorporate such population mixing into the simulator and study its effect on performance of an optimal policy. Check Prem et al. (2020) and Liu et al. (2020) for more details.

The following is from Prem et al. (2020), after fixing typos. The population is divided according to age into 5-year bands until age 70 years, and a category of aged 75 and older, resulting in 16 age groups. $C_{i,j}$ describe the contacts of age group j made by age group i. The age-specific mixing patterns of individuals in age group i, $C_{i,j}$, alter their likelihood of being exposed to the virus given a certain number of infected individuals in the population. $\kappa = 1 - \exp(-1/d_L)$ is the daily probability of an exposed individual becoming infectious, with d_L being the average incubation period. $\gamma = 1 - \exp(-1/d_l)$ is the daily probability that an infected individual recovers, with the average duration of infection as d_l. An infected individual in an age group can be clinical (I^c) or subclinical (I^{sc}). ρ_i refers to the probability that an individual is symptomatic or clinical. 1-ρ_i denotes the probability of an infected case being asymptomatic or subclinical. When $\rho_i = 0$ for all i, the model simplifies to a standard SEIR model. The force of infection for age group i at time t is given by $\beta \sum_{j=1}^{n} C_{i,j} I^c_{j,t} + \alpha\beta \sum_{j=1}^{n} C_{i,j} I^{sc}_{j,t}$, where β is the transmission rate and α is the proportion of transmission that resulted from a subclinical individual.

$$S_{i,t+1} = S_{i,t} - \beta S_{i,t} \sum_{j=1}^{n} C_{i,j} I^c_{j,t} - \alpha\beta S_{i,t} \sum_{j=1}^{n} C_{i,j} I^{sc}_{j,t}$$

$$E_{i,t+1} = \beta S_{i,t} \sum_{j=1}^{n} C_{i,j} I^c_{j,t} + \alpha\beta S_{i,t} \sum_{j=1}^{n} C_{i,j} I^{sc}_{j,t} - \kappa E_{i,t}$$

$$I^c_{i,t+1} = \rho_i \kappa E_{i,t} - \gamma I^c_{i,t}$$

$$I^{sc}_{i,t+1} = (1 - \rho_i)\kappa E_{i,t} - \gamma I^{sc}_{i,t}$$

$$R_{i,t+1} = R_{i,t} + \gamma I^c_{i,t} + \gamma I^{sc}_{i,t}$$

5.5 Extension 4: Additional hospital facility, realistic interventions

Kissler et al. (2020) and Li et al. (2020) identify that the success of social distancing hinges on the critical care capacity. We may add available Intensive Care Unit (ICU) beds into state representation. When it becomes severe, it may be necessary to set up additional ICU beds. Chen et al. (2020) show the effect of setting up a new hospital facility in China. We may treat more hospital facility as additional actions. It is straightforward in RL to modify such state and action definition to incorporate realistic, important factors.

We may consider realistic intervention measures. For example, we may implement NPIs in the simulator, evaluate them and find an optimal intervention policy out of them. However, as we discussed in 5.1, we need to consider their equivalence in achieving the same lockdown ratio, and the same effect of reducing the effective reproduction number R_t.

Aggressive testing and contact tracing are critical for controlling the pandemic, like in China, Singapore and South Korea. Testing is essential for investigating the percentage of population who have gained immunity. Testing is also critical for identifying those with asymptomatic infection.

5.6 Extension 5: Deeper integration with economic models

Economists deploy concepts like competitive equilibrium to investigate the tradeoff among production, consumption, and in our context, an epidemic. It is desirable to delve into such literature, e.g., Eichenbaum et al. (2020) and an edited book, Baldwin and di Mauro (2020), to make our model more reasonable in the sense of economics. Our presented simplistic model was inspired by Alvarez et al. (2020) from the economics community.

5.7 Extension 6: Individual level simulation

A population level simulator can illustrate population level characteristics of an epidemic; while at the same time, it may lose some details or be inconvenient to implement some detailed behaviours. An individual level simulator can capture such details, and implement intervention measures at fine-granularity. It is possible for an individual level simulator to consider at the granularity of individuals with details in many factors: 1) demographics, like age, gender, occupation, immunity, comorbidity, etc.; 2) healthcare system, like hospitals, their conditions w.r.t. locations, number of doctors/nurses, numbers of ICU beds/ventilators, etc.; 3) location, like homes, schools, workplace, restaurants, grocery stores, etc.; 4) transportation systems, like transportation types such as bus, subway, car, bicycle, and walking, and routes and roads; 5) economy, like company types such as IT, finance, manufacturing, services, utilities, etc., and supply chain system; 6) societal systems, like WHO, the Centers for Disease Control and Prevention (CDC), government, etc. Being flexible, an individual level simulator may encounter complexity and computation burden.

5.8 Is simulation sufficient?

SIR/SEIR models or variants are the “mainstream” approach in modelling study for epidemics in epidemiology and public health. A well-known problem for simulation is model misspecification and the issue of simulation to reality.

In many COVID-19 research papers, parameters were roughly estimated, or even borrowed from previous similar viruses in the beginning of the outbreak, with sensitivity analysis by varying parameters. Even now, some key parameters, like the basic reproduction number R_0 and case fatality ratio (CFR), are still rough estimates; see e.g. Donnat and Holmes (2020) and Angelopoulos et al. (2020) respectively.

Even so, simulation may be the best we can do to model an epidemic, esp. for the current one caused by a new virus SARS-CoV-2. Real life simulation is too complex for an epidemic, with too many relevant factors about individuals, virus, economy, and society. The difficulties for building a high-fidelity simulator for epidemics, besides technical factors like transmission model/parameters and economic factors, there are also social, cultural, and political aspects, which may not be easy to quantify. As a result, it is expected that policy makers may not take insights from academic research literally. Researchers are supposed to avoid providing misinformation, and warn the readers of limitations of the research study.

Considering the difficulties discussed above, a feasible plan is to focus on technical factors, at least in the early stages; in particular, we base our research on the state of the art of transmission models in epidemiology and public health. We may employ various techniques like sensitivity analysis, risk management and meta-learning to address the concerns about parameter uncertainty and model robustness. With a new, severe epidemics like the current one caused by SARS-CoV-2, some insights are urgently needed. Simulation can provide such insights; and RL is expected to have excellent performance in a simulation environment.

Acknowledgement

We thank Csaba Szepesvári for insightful comments. All errors are our own.

References

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S., and Slivkins, A. (2016). Making contextual decisions with low technical debt. *ArXiv*.

Alvarez, F., Argente, D., and Lippi, F. (2020). A simple planning problem for COVID-19 lockdown. *Working Paper*.

Angelopoulos, A. N., Pathak, R., Varma, R., and Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the covid-19 case fatality rate. *ArXiv*.

Arons, M., Hatfield, K., Reddy, S., Kimball, A., James, A., Jacobs, J., Taylor, J., Spicer, K., Bardossy, A., Oakley, L., Tanwar, S., Dyal, J., Harney, J., Chisty, Z., Bell, J., Methner, M., Paul, P., Carlson, C., McLaughlin, H., Thornburg, N., Tong, S., Tamin, A., Tao, Y., Uehara, A., Harcourt, J., Clark, S., Brostrom-Smith, C., Page, L., Kay, M., Lewis, J., Montgomery, P., Stone, N., Clark, T., Honein, M., Duchin, J., Jernigan, J., and for the Public Health?Seattle and King County and CDC COVID-19 Investigation Team (2020). Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *NEJM*.

Baldwin, R. and di Mauro, B. W. (2020). *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever It Takes*. CEPR Press.

Bartsch, S. M., Ferguson, M. C., McKinnell, J. A., OShea, K. J., Wedlock, P. T., Siegmund, S. S., , and Lee, B. Y. (2020). The potential health care costs and resource use associated with covid-19 in the united states. *Healthcare Affairs*, 39(6):1–7.

Bullock, J., Luccioni, A., Pham, K. H., Lam, C. S. N., and Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against covid-19. *ArXiv*.

Charpentier, A., Elie, R., Lauriere, M., and Tran, V. C. (2020). COVID-19 pandemic control: balancing detection policy and lockdown intervention under ICU sustainability. *ArXiv*.

Chen, S., Zhang, Z., Yang, J., Wang, J., Zhai, X., Barnighausen, T., and Wang, C. (2020). Fangcang shelter hospitals: a novel concept for responding to public health emergencies. *Lancet*, 395:1305?14.

Chen, X. and Tian, Y. (2019). Learning to perform local rewriting for combinatorial optimization. In *NeurIPS*.

Cheng, L.-F., Prasad, N., and Engelhardt, B. E. (2019). An optimal policy for patient laboratory tests in intensive care units. In *Pacific Symposium on Biocomputing (PSB)*.

Cirillo, P. and Taleb, N. N. (2020). Tail risk of contagious diseases. *Nature Physics*.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019). Autoaugment: Learning augmentation policies from data. In *CVPR*.

Donnat, C. and Holmes, S. (2020). Modeling the heterogeneity in COVID-19's reproductive number and its impact on predictive scenarios. *ArXiv*.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. In *ICML 2019 Workshop on Reinforcement Learning for Real Life (RL4RealLife)*.

Eichenbaum, M. S., Rebelo, S., and Trabandt, M. (2020). The macroeconomics of epidemics. *Working Paper*.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., van Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P. G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., and Ghani, A. C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London Report.

Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *ICML*.

Gauci, J., Conti, E., Liang, Y., Virochsiri, K., He, Y., Kaden, Z., Narayanan, V., Ye, X., and Chen, Z. (2019). Horizon: Facebook's open source applied reinforcement learning platform. In *ICML 2019 RL4RealLife Workshop*.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. (2020). Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *ICML*.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25:14–18.

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., Yao, J., Lage, I., Mosch, C., wei H. Lehman, L., Komorowski, M., Faisal, A., Celi, L. A., Sontag, D., and Doshi-Velez, F. (2018). Evaluating reinforcement learning algorithms in observational health settings. *ArXiv*.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., Li, S., lin Wang, J., Liang, Z., Peng, Y., Wei, L., Liu, Y., hua Hu, Y., Peng, P., ming Wang, J., Liu, J., Chen, Z., Li, G., Zheng, Z., Qiu, S., Luo, J., Ye, C., Zhu, S., Zhong, N., and for the China Medical Treatment Expert Group for Covid-19 (2020). Clinical characteristics of coronavirus disease 2019 in China. *NEJM*.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*.

Harvard Global Health Institute (2018). *Global Monitoring of Disease Outbreak Preparedness: Preventing the Next Pandemic*. Harvard University, Cambridge, MA.

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Funk, S., and Eggo, R. M. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob Health*, 8:e488–96.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *AAAI*.

Ie, E., wei Hsu, C., Mladenov, M., Narvekar, S., Wang, J. C., Wu, R., Jain, V., and Boutilier, C. (2019). Recsim — a configurable recommender systems environment. In *ICML 2019 RL4RealLife Workshop*.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., and Lipsitch, M. (2020). Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720.

Koo, J. R., Cook, A. R., Park, M., Sun, Y., Sun, H., Lim, J. T., Tam, C., and Dickens, B. L. (2020). Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *Lancet Infectious Disease*.

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., and on behalf of the Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infectious Disease*.

Leung, K., Wu, J. T., Liu, D., and Leung, G. M. (2020). First-wave covid-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet*.

Li, K. and Malik, J. (2017). Learning to optimize. In *ICLR*.

Li, R., Rivers, C., Tan, Q., Murray, M. B., Toner, E., and Lipsitch, M. (2020). Estimated demand for us hospital inpatient and intensive care unit beds for patients with COVID-19 based on comparisons with Wuhan and Guangzhou, China. *JAMA Network Open*, 3(5).

Li, Y. (2017). Deep Reinforcement Learning: An Overview. *ArXiv*.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

Lin, F., Muthuraman, K., and Lawley, M. (2010). An optimal control theory approach to non-pharmaceutical interventions. *BMC Infectious Diseases 2010, 10:32*, 10(32).

Liu, Y., Gu, Z., Xia, S., Shi, B., Zhou, X.-N., Shi, Y., and Liu, J. (2020). What are the underlying transmission patterns of covid-19 outbreak? ? an age-specific social contact characterization. *Lancet EClinicalMedicine*.

Lu, H., Zhang, X., and Yang, S. (2020). A learning-based iterative method for solving vehicle routing problems. In *ICLR*.

Menictas, M., Rabbi, M., Klasnja, P., and Murphy, S. (2019). Artificial intelligence decision-making in mobile health. *The Biochemist*, 41(5):20–24.

Peng, X. B., Coumans, E., Zhang, T., Lee, T.-W., Tan, J., and Levine, S. (2020). Learning agile robotic locomotion skills by imitating animals. *ArXiv*.

Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. (2017). A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *UAI*.

Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Jit, M., and Klepac, P. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health*.

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., and the Northwell COVID-19 Research Consortium (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. *JAMA*.

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555:604–610.

Shapiro, S. D. and Rothman, P. B. (2020). How academic health systems can move forward once covid-19 wanes. *JAMA*.

Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. (2011). Informing sequential clinical decision-making through reinforcement learning: an empirical study. *MLJ*, 84:109–136.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Nature*, 362:1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550:354–359.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd Edition)*. MIT Press.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool.

Szepesvári, C. (2020a). Constrained mdps and the reward hypothesis. `https://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html`.

Szepesvári, C. (2020b). Model misspecification in reinforcement learning. `https://www.youtube.com/watch?v=JuHs9yPPKwA`.

Tang, X., Qin, Z., Zhang, F., Wang, Z., Xu, Z., Ma, Y., Zhu, H., and Ye, J. (2019). A deep value-network based approach for multi-driver order dispatching. In *KDD*.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wunsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575.

Wu, J. T., Leung, K., and Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*, 395(10225):689–97.

Wu, Z. and McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china. *JAMA*, 323(13):1239–1242.

Xu, Z., van Hasselt, H., and Silver, D. (2018). Meta-gradient reinforcement learning. In *NeurIPS*.

Young, B. E., Ong, S. W. X., Kalimuddin, S., Low, J. G., Tan, S. Y., Loh, J., Ng, O.-T., Marimuthu, K., Ang, L. W., Mak, T. M., Lau, S. K., Anderson, D. E., Chan, K. S., Tan, T. Y., Ng, T. Y., Cui, L., Said, Z., Kurupatham, L., Chen, M. I.-C., Chan, M., Vasoo, S., Wang, L.-F., Tan, B. H., Lin, R. T. P., Lee, V. J. M., Leo, Y.-S., Lye, D. C., and for the Singapore 2019 Novel Coronavirus Outbreak Research Team (2020). Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *JAMA*, 323(15):1488–1494.

Yu, Y.-L., Li, Y., Szepesvári, C., and Schuurmans, D. (2009). A general projection property for distribution families. In *NIPS*.

Zhang, J. and Bareinboim, E. (2019). Near-optimal reinforcement learning in dynamic treatment regimes. In *NeurIPS*.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., Xing, L., Guo, T., , and Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37:1038–1040.

Zoph, B. and Le, Q. V. (2017). Neural architecture search with reinforcement learning. In *ICLR*.