**REINFORCEMENT LEARNING BASED DECISION SUPPORT TOOL FOR EPIDEMIC CONTROL**

Mohamed-Amine CHADI and Hajar Mousannif
Informatics System Engineering Laboratory, Cadi Ayyad University, Marrakech, Morocco
mohamedamine.chadi@ced.uca.ma, mousannif@uca.ma

**Abstract**

**Rationale:** Covid-19 is certainly one of the worst pandemics ever. In the absence of a vaccine, classical epidemiological measures such as testing in order to isolate the infected people, quarantine and social distancing are ways to reduce the growing speed of new infections as much as possible and as soon as possible, but with a cost to economic and social disruption. It is therefore a challenge to implement timely and appropriate public health interventions. **Objective:** This study investigates a reinforcement learning based approach to incrementally learn how much intensity of each public health intervention should be applied at each period in a given region. **Methods:** First we define the basic components of a reinforcement learning (RL) set up (i.e., states, reward, actions, and transition function), this represents the learning environment for the agent (i.e., an AI-Model). Then we train our agent using RL in an online fashion, using a reinforcement learning algorithm known as REINFORCE. Finally, a developed flow network, serving as an epidemiological model is used to visualize the results of the decisions taken by the agent given different epidemic and demographic state scenarios. **Main Results:** After a relatively short period of training, the agent starts taking reasonable actions allowing a balance between the public health and economic considerations. In order to test the developed tool, we ran the RL-agent on different regions (demographic scale) and recorded the output policy which was still consistent with the training performance. The flow network used to visualize the results of the simulation is considerably useful since it shows a high correlation between the simulated results and the real case scenarios. **Conclusion:** This work shows that reinforcement learning paradigm can be used to learn public health policies in complex epidemiological models. Moreover, through this experiment, we demonstrate that the developed model can be very useful if fed in with real data. **Future Work:** When treating trade-off problems (balance between two goals) like here, engineering a good reward (that encapsulates all goals) can be difficult, therefore future work might tackle this problem by investigating other techniques such as inverse reinforcement learning and Human-in-the-Loop. Also, regarding the developed epidemiological model, we aim to gather proper real data that can be used to make the training environment more realistic, as well as to apply it for network of regions instead of a single region.

**Keywords:** Covid-19, Epidemics control, Reinforcement learning, Modelling and Simulation.

1 INTRODUCTION

1.1 Background:
Epidemics of infectious diseases are an important threat to public health and global economies. As of 16 September 2020, Covid-19 has generated over 29 million of confirmed infected cases and 900 thousand deaths [1]. This been said, the development of timely precise and demographically tailored public health policies is a challenging process, as epidemics are non-linear and complex processes.

There is also intense debate about policies for limiting the damage, both to health and to the economy. On the one hand, some studies [2] suggest that quarantine and social distancing measures might be needed for as long as 18 months and will only be turned on and off alternatively during that period. Other studies on the other hand, predict that the economic harm of these measurements is estimated to be much more than that of the economic crisis of 2008 [3].

1.2 Related work:
For this matter, researchers have used till now pretty much all kinds of artificial intelligence (AI) techniques to fight the current pandemic. Ranging from detecting the disease in medical images [4], predicting next number of infections and deaths [5], to discovering new medical drugs to treat patients [6], and many other applications that can be consulted in detail here [7].

There were also various prevention strategies proposed to reduce the desolation of the pandemic, while maintaining economical balance such as age-based lockdown [8] and n-work-m-lockdown [9].
Another specific area of AI that was used is Reinforcement learning (RL) [10], which is a sub-family of machine learning that deals mainly with learning optimal decision making. In recent years, RL coupled with deep learning (i.e., Deep Reinforcement Learning) has been very successful at solving complex sequential decision-making problems in various domains such as video games [11]. Deep reinforcement learning (DRL) has proven its strength in various occasions such as playing Atari like human [12] and StarCraft II [13], playing hide and seek [14], mastering the game of go [15], and the game of DOTA [16].

Reinforcement learning is not used in video games only, but also in financial markets [17], robotics [18], healthcare domains [19], including pathologies like cancer, diabetes, anaemia, schizophrenia, epilepsy, anaesthesia, and drug discovery, just to name a few.

In the context of epidemiology and public health, learning optimal policies by formulating the problem at hand as a Markov decision process (MDP), was first introduced in [20]. Similarly, RL was used for controlling the spread of emerging infectious diseases such as the current Corona Virus pandemic [21], where an agent (trained using reinforcement learning) takes actions in different possible scenarios of a pandemic depending on the spread of disease and economic factors.

![Figure 1: illustration of the final objective of this work](image)

However, there is still a lack of effective tools to provide tailored decision support for regions (country, city, district, etc.) with different demographics, health care systems, economy and during different epidemics.

To further investigate this problem and contribute to the relatively limited state of the art, this article addresses the following points:

Section 2 is an overview and presentation of the developed epidemiological model. In section 3 we go over the methods used for the development of the epidemic’s control decision support algorithms. Section 3 lists the main findings of this study and related measurements. In section 4 we discuss the advantages and drawbacks of some techniques exploited here and ways to ameliorate the performance. Finally, section 5 summarizes the key contributions and suggests what might be future work in respect to this study.

2 METHODS

The objective of this work is to create an AI-agent whose inputs are data describing the current epidemic state as well as the demographics of the studied region and outputs are intensities/percentages of each public health measures, as illustrated in figure 1.

For this purpose, we developed an epidemiological model that is intended to help us in both computing the inputs data evolution/transitions necessary for the learning of the agent as well as in visualizing the results of the simulation.

2.1 The epidemiological Model

Below in figure 2, is the developed epidemiological model, where:

- **S1**: The transmission rate,
- **S2**: The identification rate,
- **S3**: The death rate (see definitions).

**Cur Inf**: current infections, **N**: normal, **Nxt Inf**: next infection, **U**: unknown carriers, **K**: known carriers, **R**: recovered, **D**: deaths.
We consider that a region with a normal population size ‘N’, will get in contact with an initial number of current infected people ‘Cur Inf’ and will generate $S1 \times N$ amount of new infected people ‘Nxt Inf’. The newly infected people will either be known carriers by a $S2$ rate ‘K’ or unknown carriers ‘U’. The unknown carriers ‘U’ (e.g., those that didn’t show symptoms) will represent exactly the current infected people for the next run/loop. The known carriers (e.g., those that showed symptoms or being tested and hospitalized) will either die or recover depending on the parameter $S3$.

The transmission rate: In our previous paper [22], we used modelling and simulation techniques to extract the most important demographic data of a given region (country, city, district, etc.) that contribute to the state of an epidemic in that region. In summary, we define three parameters:

1- Population density: the number of people per square kilometre.

2- Movement rate: degree of dynamism in the studied region.

3- Region’s connectedness: reflects how much attractive and important is the studied region relative to regions in its network. A region with high connectedness is usually a region that provides jobs, hospitals, schools, etc. They often tend to be the big metropolitan cities of a country.

These three parameters are impacted via public health interventions such as “travel restriction, lock-down and distance work & study”, and these are also what makes the contact rate (i.e., how many person-to-person contacts one makes in a given period of time). Note that we are not talking about the transmission rate here; a contact doesn’t necessarily imply a transmission of the disease. Actually, the transmission rate can be defined as: the contact rate * probability of transmitting the disease while making a contact without any barriers * the proportion of people that aren’t doing barriers. Barriers can be masks for example in the case of a respiratory virus such as Covid-19.

This way, the transmission rate * population size represents the number of people susceptible to receiving the disease from one infected person. Thus, if multiplied by the number of previously infected people, it will give us the number of new infections.

The identification rate: While the transmission rate enables us to calculate the next number of infections, the identification rate tells us how many among those new infected people will be known carriers or unknown carriers. Depending on the studied epidemic, the incubation period and symptoms manifestation and the ability to conduct testing will differ. Therefore, we define the identification rate as: the testing rate * incubation coefficient, where the incubation coefficient represents how long is the incubation period of a given epidemic (i.e., how long does it take for an infected person to show symptoms). It has a big impact on identifying infected people. Covid-19 for example has a relatively long incubation period (of around 10 days) as opposed to Ebola for example which will severely manifest in the infected person as quickly as 2 days at maximum. That’s one of the main reasons why Covid-19 has spread much more than other
epidemics. In the developed epidemiological model (figure 2), we consider that people that didn’t get identified will be considered as the number of current infections in the next run.

The death rate: As illustrated in the model, the previously identified ‘K’ amount of infected people will be divided into two categories, those who will recover from this disease and those who will die because of this disease. The degree of fatal danger varies from one epidemic to another (e.g., Corona-Virus and Ebola), but also the health care capacity differs from one region to another. Here we define the death rate as: Health care capacity * vulnerability coefficient. Where the vulnerability coefficient represents the proportion of people the studied epidemic is dangerous for (i.e., does it kill everyone got infected? Or only people above 70years old or maybe only people with heart disease, etc.) relative to all identified infected people.

2.2 The learning environment
The developed tool consists mainly of a Reinforcement Learning (RL) set up. In RL, an agent is interacting with its environment with a goal to maximize a reward function.

In our case, the Agent is modelling Public Health Authorities (PHA) that decides which Actions (interventions and measures) to implement given certain States of the environment.

States of the epidemic, actions allowed for the agent, transition function that describes the dynamics of the epidemic and reward function that tells the agent how good it is doing in this process, is called a Markov Decision Process (MDP).

A Markov Decision Processor MDP: = {S, A, TF, R} models an agent's sequential decision-making process, where:

- S is a set of states and includes:
  Transmission rate (S1)
  Identification rate (S2)
  Death rate (S3)

- A is a set of possible actions and includes:
  A1: Travel restriction
  A2: Lock down
  A3: Distance work & study
  A4: Use Barriers
  A5: Increase the test rate
  A6: Increase the health care capacity

- TF is the transition function:
  Given a state S, and executing action A, the agent will follow T to get to another state S’ receiving a reward R.

More about how the Transition Function (TF) was made in the next section.

- R is the reward:
  Received when action A is performed at state S. the reward function should be defined in a way to encourage the agent to learn the best trade-off actions getting the maximum out of health benefit and maximum out of economic benefit.

2.3 The Transition function
The transition function (TF) is what tells us, the relationship between actions and states (S1, S2, S3). It is basically another way of saying, how each action impacts the value of each state. It is also used to tell the agent how well it is doing in this process via a reward signal.

Formally, TF (s, a) → s’, r. Where:

s is an input vector that represents the current values of S1, S2 and S3,
a is an input vector that represents the intensities of the applied actions,
s’ in an output vector representing the new values of S1, S2 and S3, and r is the reward signal.

4) Reward function

Here we define two scores that will dictate the agent’s success or failure:

Economic score = $1 - \frac{A1 + A2 + A3 + A4 + A5 + A6}{6}$

Health score = $1 - \frac{R0 + \text{Death rate}}{2}$

Where:

$R0 = \frac{\text{current number of the class Nxt Inf}}{\text{Previous number of the class Nxt Inf}}$

The agent succeeds if: economic score > 0.3 and health score > 0.6 and thus scoring a Reward of ‘1’, and fails otherwise, thus scoring a Reward of ‘0’.

The numbers 0.3 and 0.6 represent a chosen threshold that can be changed by the user to define the trade off limits for the allowed budget for the agent to spend and the R0 and death rate that should not be exceeded.

As illustrated in figure 3, the agent controls the intensity of the 6 actions (in black) to maximize the cumulated reward. Initial states (S1, S2, S3) are defined by the user, given some studied region and epidemic’s parameters.

3 The RL-Agent

The RL-agent is a policy network that consists of a multi-output regressor, with a multi-layer perceptron (MLP) of one hidden layer of size equal to 100 units as estimator (figure 4).

The training process of the agent is to initially set the agent’s parameters randomly and launching it to interact with the environment. The agent observes the states and takes actions (i.e., choose certain intensity for each defined action). The algorithm used variant of a famously known RL-algorithm as REINFORCE[23], which is a basic RL algorithm updated using stochastic gradient descent [24] and whose pseudo-code is presented in figure 5.
4 RESULTS

4.1 Agent’s performance analysis

We launched the agent for a 20000 episodes/simulation where each episode consists of 100 interactions and plotted the smoothed average reward over each episode. As shown in figure 6, the reward function is progressively maximized and stabilized near a maximum value of 0.7, meaning that the agent is successful in about 70% out of each of the 100 interactions that will be played. The agent learned a trade-off policy that balances between health and economic considerations given different demographic and epemics states.

And below are some examples that show the mentioned learned trade off policy:

| Table 1: Examples of agent’s policy given different states |
|----------------|-----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|   S1 | S2 | S3 | A1 | A2 | A3 | A4 | A5 | A6 |
| 0.60 | 0.50 | 0.61 | 0.72 | 0.71 | 0.68 | 0.67 | 0.67 | 0.68 |
| 0.67 | 0.66 | 0.46 | 0.49 | 0.49 | 0.66 | 0.77 | 0.46 | 0.66 |
| 0.69 | 0.60 | 0.57 | 0.29 | 0.49 | 0.49 | 0.60 | 0.69 | 0.46 |

Where:
- S1, S2 and S3 are the current states (The transmission rate, the identification rate, the death rate, respectively).
- A1, A2, A3, A4, A5, and A6 are the actions applied (Travel restriction, Lock-down, Distance work & study, Barriers, Increase the test rate, Increase the health care capacity).
- S1’, S2’, S3’: next states.

As observed in these examples of table 1, the agent is minimizing the intensity of actions: Travel restriction, Lock-down, Distance work & study and maximizing on the other hand actions: Use barriers, Increase the test rate and Increase the health care capacity. This is due to the way the reward function is defined (see reward definition).

In order to maximize the reward, the agent should maximize both economic score and health score at each 100 interaction of each episode.

Economic score is at its maximum when all applied actions have minimum intensities. Health score on the other hand is at its maximum when R0 and death rate have minimum values; to have a minimum R0 (i.e., number of new infections / number of previous infections) the agent will have to get the number of new infections to as minimum as possible, so it does this by increasing moderately the barriers usage intensity and in the test rate, which in the epidemiological model, it helps dividing the number of next infections to known and unknown carriers, unknown carriers will be the new initial infections in the next run/iteration. If the initial number of infections is lower than previously, then the agent succeeded at decreasing the R0. And for the death rate, the agent could not help but maximizing the health care capacity as possible as it can.

This way the agent’s policy can be formulated as follow:

**Travel restriction, lock-down and distance work & study** should be intensified depending on how much contact rate is in the studied region. The denser and dynamic is the region, the more they should be applied but to a moderate extent of 0.4, 0.5 and 0.6, respectively.

**Barriers usage** must be as high as the transmission rate.

**Testing rate** will depend on epidemic’s incubation period; the longer it is the greater the intensity of the testing rate must be. Also, the test rate should be applied in such a way to get as little unknown carriers as possible relatively to the last time (e.g., the week before) to iteratively reduce the reproduction rate R0.

**Health care capacity** must be always as max as it can be.

Note that this is not a universal policy, changing the reward function and/or the epidemiological model will cause a considerable change in the agent’s policy, therefore good understanding of epidemics dynamics might help at approximating the simulated results to match reality.

5 **DISCUSSIONS**

5.1 **Main contributions**

a) The developed epidemiological model reflects relatively well the process of how public health interventions contribute to the evolution of the epidemic in the studied region and captures more the natural dynamics of epidemics. One example would be the fact that having more unknown carriers will generate more new infections, which shows up more when comparing epidemics with different incubation periods and ways of manifestation. Also captures some demographic properties of the studied region, such as, when travel restriction is applied, the density and dynamics of the population is reduced, yielding reduction in the contact rate and therefore in the transmission rate.

b) The developed epidemiological model can be used to manually simulate the evolution of any given epidemic in any given region. But simulating many scenarios by hand can be difficult. Therefore, we added a reinforcement learning approach that gives an AI-Agent the ability to automatically recommend an optimal policy for the given situation. The agent itself is a simple model (i.e., multi-output regressor with a MLP as estimator), which means that better learning performance can still be achieved if using other complex models.

c) The policy is meant to reduce both the reproduction rate (R0) and death rate (S3) while still being fully aware of economic considerations and given how the reward function is predefined makes the policy fully explainable.

d) The learning environment (the transition function) and the reward can be modified by the user to be adapted to whatever the desired goal is and to make importance on each component to better model states and actions impact.
e) The learning environment can be considered as a non-stochastic MDP view the fact that the transition probabilities of the epidemiological model (S1, S2 and S3) are changing each time given the applied actions and region-epidemic characteristics.

f) The whole framework (epidemiological model plus the reinforcement learning approach) is flexible, meaning that it can be used to recommend the optimal epidemic’s control policy (i.e., public health interventions) for any given region (i.e., any demographics) and for any given epidemic’s characteristics (incubation period, transmission probability, vulnerability coefficient, etc.). Therefore, the same framework/approach can be used to compute the optimal policy for any changes in (i) demographic data of the studied region (ii) data characteristics of the epidemic, and (iii) the reward definition.

5.2 Future directions
a) The epidemiological model
Scaling towards a network of regions:
In this paper, we present an epidemiological model that is meant to be applied on one region only. One way to scale this framework is to make a network of these epidemiological models in order to be able to study the spread of a given infectious disease in a whole network of regions (e.g., cities of a country, districts of a city, etc.). To cite some similar ideas and work done, Harshad et al. in their paper [25] present a quantitative way to compute lockdown decisions for individual cities or regions in a network set-up with a relatively basic variation of the SEIR model, while balancing health and economic considerations. Another related work by [26] conducts an experiment where the aim is to learn a joint policy to control the districts in a community of 11 tightly coupled districts. This experiment shows that deep reinforcement learning can be used to learn mitigation policies in complex epidemiological models with a large state space and that there can be an advantage to consider collaboration between districts when designing prevention strategies.

For this purpose, we might consider using a Multi-Agent-RL (MARL) [27], where agents represent public health authorities of each connected region, and the goal is to cooperate in order to mitigate the spread of an infectious disease while still optimizing the economic considerations.

b) The learning environment
Adding other states:
The presented learning environment (or the MDP) is composed out of 3 states meant to comprehend demographic parameters of the studied region such as density, movement, and connectedness as well as some epidemic’s properties. This been said, there exist other types of data that should be included, data reflecting statistics of the population living in the studied region such as age and gender distribution, chronic disease occurrence distribution. Adding these types of data might help approximating the results to real case scenarios as much as possible.

Adding other actions:
In our environment, we make use of 6 actions (i.e., Travel restriction, Lock-down, Distance work & study, Barriers, Increase the test rate and increase the health care capacity) considering that that no vaccination is produced yet. But these are not meant to be all existing actions allowed for the agent. In real world, some countries might use other interventions such as contact tracing as well, this helps increasing the identification rate (B), and each country might have the capacity to trace contacts at different levels.

A learned AI transition function (AI-TF):
The transition function is a function that maps actions impact on states and returned reward. Being as complex as it is, coding a real-world epidemic dynamics model manually might be very time consuming and requires highly skilled graphics experts. Therefore, considering a learning approach might be of great benefit. Perhaps the most famous paper in this respect is the recent work done by [28] where the simulated environment is a trained learning model introduced as GameGAN used to generate next frames (states) and scores of the game (reward) given actions (keyboard input) applied.

Another way is to make use of the Human in the loop techniques [29] to augment the AI-TF capabilities, especially when new epidemics show up and we know very little about the dynamics of the virus and its related disease.

This is also important if we want to anticipate the damages that might be caused by a wrong decision made by Public Health Authorities (PHA). This approach (of building a realistic generalizable model of the dynamics of epidemics) can be coupled with Model-based Reinforcement Learning [30] techniques such as Monte Carlo tree search (MCST) [31] and Deterministic Deep Policy Gradient (DDPG) [32], to enable the
AI-agent to do as much planning and look-ahead as it needs to ensure the safest and most rewarding strategies.

6 CONCLUSIONS

Defining separate optimal public health strategies (i.e., strategies that mitigate the spread of an epidemic with minimal disruption to the global economy) for each different region/demographic is a difficult and time-consuming task where time is crucial to anticipate catastrophic damages. In this respect, the present work investigates an RL based approach coupled with a general epidemiological model that helps in the making of an adaptable learning environment for AI-agents in order to derive tailored optimal public health policies.

We trained the developed agent in an online fashion against a 20000 different demographics and disease parameters, where it was stabilized in a success rate of about 70% of the cases.

The agent was able to satisfy a trade-off condition that balances between a predefined health score and economy score. Therefore, RL provides a promising tool for making state-dependent optimal strategies and also easily interpretable by decision makers. Finally, future research will focus on scaling this same approach and adding other features to it.

FUNDING:

This work was supported by a grant from the MENFPESRS “Ministère de l'Education Nationale » and the CNRST “Centre National de Recherche Scientifique et Technique” www.cnrst.ma

REFERENCES

[1] Google News Covid-19 Cases, https://news.google.com/covid19/map?hl=fr&gl=MA&ceid=MA%3Afr

[2] Neil Ferguson, Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand, DOI: 10.25561/77482, 16 March 2020

[3] World Economic Outlook, The Great Lockdown, Chapter 1, April 2020, https://www.imf.org/en/Publications/ WEO/Issues/ 2020/04/14/weo-april-2020

[4] Aboughazala, Laila & Mohammed, Kamel K. (2020). Automated Detection of Covid-19 Coronavirus Cases Using Deep Neural Networks with X-ray Images. Al-Azhar University Journal of Virus Researches and Studies. 10.21608/aujv.2020.106722.

[5] Tuli, Shreshth & Tuli, Shikhar & Tuli, Rakesh & Gill, Sukhpal Singh. (2020). Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. 10.1101/2020.05.06.20091900.

[6] Harun Al Rashid, Md & Mridul, Mayank & Mohanty, Chandana & Swayamsiddha, Swati. (2020). Application of Artificial Intelligence in COVID-19 drug repurposing. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 14. 10.1016/j.dsx.2020.06.068.

[7] Nguyen, Thanh. (2020). Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions. 10.13140/RG.2.2.36491.23846/1.

[8] Acemoglu, D., Chernozhukov, V., Werning, I. & Whinston, M. D. A multi-risk sir model with optimally targeted lockdown. Tech. Rep., National Bureau of Economic Research (2020).

[9] Karin, O. et al. Adaptive cyclic exit strategies from lockdown to suppress covid-19 and allow economic activity. medRxiv (2020).

[10] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018

[11] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, “A Survey of Deep Reinforcement Learning in Video Games,” arXiv:1912.10944 [cs], Dec. 2019

[12] Mnih, V. et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013).

[13] P. Sun, X. Sun, L. Han, J. Xiong, Q. Wang, B. Li, Y. Zheng, J. Liu, Y. Liu, H. Liu, and T. Zhang, “TStarBots: defeating the cheating level builtin AI in StarCraft II in the full game,” CoRR, vol. abs/1809.07193, 2018.

[14] Baker, B. et al. Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528 (2019).
[15] Silver, D. et al. Mastering the game of go without human knowledge. nature 550, 354–359 (2017).
[16] Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. Science 365, 885–890 (2019).
[17] Halperin, “Inverse Reinforcement Learning for Marketing,” SSRN Journal, 2017, doi: 10.2139/ssrn.3087057
[18] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” The International Journal of Robotics Research, vol. 32, no. 11, pp. 1238–1274, Sep. 2013, doi: 10.1177/0278364913495721
[19] C. Yu, J. Liu, and S. Nemati, “Reinforcement Learning in Healthcare: A Survey,” arXiv:1908.08796 [cs], Aug. 2019
[20] R. Yaesoubi and T. Cohen. Dynamic health policies for controlling the spread of emerging infections: influenza as an example. PloS one, 6(9), 2011.
[21] Ohi, Abu & Ph. D., M. & Monowar, Muhammad Mostafa & Hamid, Md. Abdul. (2020). Exploring Optimal Control of Epidemic Spread Using Reinforcement Learning. 10.21203/rs.3.rs-48613/v1.
[22] Chadi Mohamed Amine and Mousannif Hajar, Making Sense of the Current Covid 19 Situation and Suggesting a tailored Release Strategy through Modeling and Simulation, Case Study: Casablanca, Morocco, arXiv:2005.03477
[23] Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. MachLearn 8, 229–256 (1992). https://doi.org/10.1007/BF00992696
[24] Ketkar N., “Stochastic Gradient Descent,"Deep Learningwith Python. Apress.,2017. DOI: 10.1007/978-1-4842-2766-4_8
[25] Khadilkar, Harshad & Ganu, Tanuja & Seetharam, Deva. (2020). Optimising Lockdown Policies for Epidemic Control using Reinforcement Learning: An AI-Driven Control Approach Compatible with Existing Disease and Network Models. Transactions of the Indian National Academy of Engineering. 10.1007/s41403-020-00129-3.
[26] Libin, Pieter & Moonens, Arno & Verstraeten, Timothy & Perez-Sanjines, Fabian & Hens, Niel & Lemey, Philippe & Nowe, Ann. (2020). Deep reinforcement learning for large-scale epidemic control.
[27] Busoniu, Lucian & Babuska, Robert & De Schutter, Bart. (2010). Multi-agent Reinforcement Learning: An Overview. 10.1007/978-3-642-14435-6_7.
[28] S. W. Kim, Y. Zhou, J. Philion, A. Torralba and S. Fidler, "Learning to Simulate Dynamic Environments With GameGAN," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 1228-1237, doi: 10.1109/CVPR42600.2020.00131.
[29] Liang, Huanghuang & Yang, Lu & Cheng, Hong & Tu, Wenzhe & Xu, Mengjie. (2017). Human-in-the-loop reinforcement learning. 4511-4518. 10.1109/CAC.2017.8243575.
[30] Ray, Soumya & Tadepalli, Prasad. (2017). Model-Based Reinforcement Learning. 10.1007/978-1-4899-7687-1_561.
[31] Vodopivec, Tom & Samothrakis, Spyridon & Šter, Branko. (2017). On Monte Carlo Tree Search and Reinforcement Learning. Journal of Artificial Intelligence Research. 60. 881-936. 10.1613/jair.5507.
[32] Silver, David &Lever, Guy & Heess, Nicolas & Degris, Thomas & Wierstra, Daan & Riedmiller, Martin. (2014). Deterministic Policy Gradient Algorithms.