Robust Lock-Down Optimization for COVID-19 Policy Guidance

Ankit Bhardwaj[2*], Han Ching Ou[1*], Haipeng Chen[1], Shahin Jabbari[1], Milind Tambe[1], Rahul Panicker[2], Alpan Raval[2]

[1]Harvard University, hou@g.harvard.edu, {hpchen, jabbari , tambe}@seas.harvard.edu
[2] WadhwaniAI, {bhardwaj, rahul, alpan}@wadhwaniai.org
∗ Equal Contribution

Abstract

As the COVID-19 outbreak continues to pose a serious worldwide threat, numerous governments choose to establish lock-downs in order to reduce disease transmission. However, imposing the strictest possible lock-down at all times has dire economic consequences, especially in areas with widespread poverty. In fact, many countries and regions have started charting paths to ease lock-down measures. Thus, planning efficient ways to tighten and relax lock-downs is a crucial and urgent problem. We develop a reinforcement learning based approach that is (1) robust to a range of parameter settings, and (2) optimizes multiple objectives related to different aspects of public health and economy, such as hospital capacity and delay of the disease. The absence of a vaccine or a cure for COVID to date implies that the infected population cannot be reduced through pharmaceutical interventions. However, non-pharmaceutical interventions (lock-downs) can slow disease spread and keep it manageable. This work focuses on how to manage the disease spread without severe economic consequences.

Introduction

While governments are responding to the spread of COVID-19 by imposing lock-downs of varying intensity to reduce human-human contact, the situation cannot be maintained indefinitely. Each day of lock-down brings severe economic loss affecting the livelihood of billions. Thus, it is imperative to use the available resources of interventions – lock-downs, test-kits, ventilators etc., in an efficient manner. This work aims to find optimal lock-down policies based on epidemiological models and reinforcement learning.

Reinforcement learning has shown promising results on sequential decision making tasks like Go (Silver et al. 2016) and autonomous driving (Pan et al. 2017). On these tasks, training from real-world data directly is too expensive due to costly data collection process. Learning the agent model from simulations is thus necessary. However, simulations don't reflect the real world exactly due to uncertainties transferred while fitting the simulation model (Christiano et al. 2016). It can be dangerous if the policy learned is unaware of such uncertainties, which is especially true for our task.

AAAI Fall 2020 Symposium on AI for Social Good.

In addition to insufficient data and uncertainty, for our problem, it is hard to specify a single objective that one wants to achieve. It is likely that many objectives need to be met for the task to be considered successful. For example, we may want to delay the peak of infections while making sure that our hospitals are not overburdened or our economy is not affected too severely. The decision maker in this case is looking at the problem from several perspectives leading to many possible objectives that the model will be evaluated on. Thus, in this work, we incorporate multi-objective functions.

The main contributions of this work can be summarized as follows:

- We formulate the problem of lock-down implementation as a Markov Decision Process (MDP). To solve this MDP, we propose a Reinforcement Learning (RL) approach that optimizes the trade-off between health objectives and economic cost.
- We tackle the uncertainty in environment parameters that might arise from the noise in the data and the process of estimation by considering different robust approaches.
- We analyse different robust approaches including uniform sampling and adversarial sampling during the training phase. We find that there is a trade-off relation in the average-case and worst-case between RL agents with different degrees of risk aversion.
- We design different health objectives that might be of interest to decision-makers and measure our performance along these different objectives simultaneously.

With this work, we aim to address the challenging task of planning temporal resource allocation for lock-downs. The models that we use for modelling the spread of COVID are the SEIR class of epidemiological models.

Previous Work

Since as early as the 17^{th} century, when Bernoulli proposed the first mathematical epidemic model for smallpox (Bernoulli and Blower 2004), there have been numerous efforts in the modeling and control of epidemics. One important class of these models are called compartmental models. These models, as their name suggests, divide the population into different health states (compartments) and

model transitions of populations between these health states. The underlying assumption is that these compartments have homogeneously mixed populations. Susceptible-Exposed-Infected-Recovered (SEIR) family of models are compartmental models with dynamics described by ordinary differential equations. Recently, there have been advances in fitting SEIR models with machine learning techniques (Bannur et al. 2021). In this work, we use a Susceptible-Exposed-Infected-Recovered-Deceased (SEIRD) model (detailed description in subsection Epidemic Model) to model the COVID-19 data. However, the technique we propose is applicable to any of the SEIR family models.

Apart from epidemic modeling, the problem of optimizing cure and control for preventing the spread of disease is also of interest. However, most works in the computer science literature usually assume an idealistic model, such as every contact being known, no uncertainty in the disease parameters or a strong cure/isolation that guarantees the recovery of the individuals (Ball, Knock, and O'Neill 2015; Sun and Hsieh 2010; Wang 2005; Zhang and Prakash 2015; Ganesh, Massoulié, and Towsley 2005). None of these are true for most real-world diseases, such as the newly arisen COVID-19 pandemic which has no cure as of the writing of this paper. Even under most settings being ideal, a small uncertainty could have serious implications on outcomes if not handled properly. For example, the impact of curing uncertainty under perfect observation is analyzed in Hoffman and Caramanis (Hoffmann and Caramanis 2018) by providing non-constructive, algorithm-independent bounds. We aim to address the challenging setting in which there are uncertainties in most of the parameters in the model.

Robust control is a branch of control theory that has a long history. In particular, robustness toward parameter uncertainty results in a performance drop from the model toward its real-world application(Mannor et al. 2004). Numerous works (Nilim and El Ghaoui 2005, 2004; White III and Eldeib 1994) have tried to tackle such uncertainty under the robust MDP framework with different assumptions. In recent years, Reinforcement learning has demonstrated promising results on a variety of MDP problems (Silver et al. 2016; Pan et al. 2017). For applications with a high safety requirement, it is natural to combine robustness into reinforcement learning (Mihatsch and Neuneier 2002; Carpin, Chow, and Pavone 2016; Chow et al. 2017). Among these works, using an adversarial agent to adjust the environment and discover potential risk systematically has shown promising results in many real-world tasks (Pinto, Davidson, and Gupta 2017; Pattanaik et al. 2017). A recent algorithm using an adversarial framework is robust adversarial reinforcement learning (RARL) (Pinto et al. 2017) in which, two agents are trained, one protagonist and other an adversary providing attacks on input states and dynamics. In our work, similar to RARL, we use an adversarial agent to systematically search risky environmental parameters for the policy.

Another important consideration for lock-down policy makers might be to include different desirable objectives in their decision-making. Multi-objective optimization has tremendous practical importance in many real-life applications (Deb 2014). Linear combinations of Pareto optimalities are often considered and solved when the system is easy to describe (Censor 1977). Multi-Objective Reinforcement Learning however (Roijers et al. 2013; Van Moffaert and Nowé 2014), is a relatively new research area that has been actively studied only in recent years. In this work, we designed a Quality Adjusted Life Year (QALY) value function to calculate the suitable reward signal for any point of any two objectives of the QALY variant. Such a function can also be used to determine the difficulties of optimizing the two objectives simultaneously.

Notations	Definition
Health States	
S	Susceptible fraction of the population
E	Exposed fraction of the population
I	Infected fraction of the population
R	The fraction of the population that is either completely recovered or is undergoing recovery and is no longer infectious
D	Deceased fraction of the population
Transmission	
t	Current time (day)
R_o	Basic Reproductive Number
T_{inc}	The incubation time
T_{inf}	The duration for individuals being infectious
$T_{recover}$	Time for individuals to recover or quarantine in hospital
T_{fatal}	Time for a fatal infection individual to die
Intervention	
a	Action (intervention strength)
$l(a)$	Cost of the action per day
e	Lock-down effectiveness coefficient
d	Minimum duration of the intervention
T_{trans}	Transition delay after lock-down a is deployed
Objectives	
λ	Hospital Capacity
δ	Economic-Health cost weight

Table 1: The notations used across this paper.

Modelling

Epidemic Model

For modelling COVID-19, we adopt a discrete time SEIRD model (Weitz and Dushoff 2015). The SEIRD class of models is a part of compartmental models, as mentioned above. An individual can be in one of the following health states: S (a healthy individual *susceptible* to disease), E (the individual has been *exposed* and has latent disease), or I (the individual is *infected*), R (the individual is *recovering* and is no longer infectious to others) and D (the individual is *deceased*). Table 1 summarizes the symbols we use throughout this paper.

The discrete-time dynamics equations for our epidemic

model are:

$$S_{t+1} - S_t = \frac{-S_t I_t}{T_{trans}(a,e)}, \tag{1}$$

$$E_{t+1} - E_t = (\frac{S_t}{T_{trans}(a,e)} - \frac{E_t}{T_{inc}}), \tag{2}$$

$$I_{t+1} - I_t = (\frac{E_t}{T_{inc}} - \frac{I_t}{T_{inf}}), \tag{3}$$

$$R_{t+1} - R_t = \frac{I_t}{T_{recover}}, \tag{4}$$

$$D_{t+1} - D_t = \frac{R_t}{T_{fatal}}, \tag{5}$$

in which $\frac{1}{T_{inf}} = \frac{1}{T_{recover}} + \frac{1}{T_{fatal}}$ and the basic reproductive number can be obtained from $R_0 = T_{inf}/T_{trans}$. A typical SEIRD model described above starts from a population being mostly susceptible and a small fraction of infectious people. When $R_0 > 1$, each infected individual will infect more than one susceptible individual in its lifetime on average. Each susceptible individual will eventually go through the exposed, infectious to finally recovered or deceased states. A schematic diagram of the SEIRD model is shown in Figure 1. Generally in compartmental models,

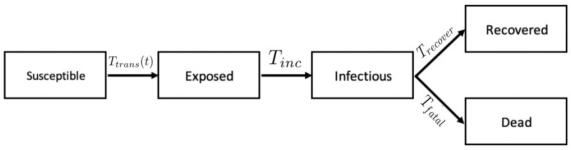

Figure 1: A schematic diagram of the SEIRD model.

T_{trans} is a constant. However, in a real-world setting, the transmission time can be reduced through the deployment of non-pharmaceutical lock-down interventions. Note that there is no direct reduction in infected population as there is no cure or vaccination available. We only consider the lock-down interventions that increase the transmission time of the virus based on their strength.

Compartmental models and their variants are commonly used in disease state forecasting and prediction. For concreteness, in this work, state populations and numerical values of the transmission parameters are based on the available data for the city of Mumbai (Group 2020). However, it is worth noting that both the intervention model and planning algorithm we propose apply for most, if not all, of the SEIR model variants.

Intervention Modelling

As there is no cure/vaccine available for COVID-19 till date, models of pharmaceutical interventions are not applicable. To manage the rapid disease spread, different lock-down policies can be considered to limit individual contacts. For example, (Ferguson et al. 2020) considered different levels of lock-down such as *Case isolated at home*, *Voluntary home quarantine*, *Social distancing of those over 70 years of age*, *Social distancing of entire population* and *Closure of schools and universities* for non-pharmaceutical interventions in British population, which all have different cost and effectiveness.

These lock-down policies should enjoy several desiderata for real-world deployment. First, each type of intervention needs to last for a minimum duration d. Second, since the lock-down has economic cost, government bodies would expect a trade-off between the total budget B spent for planning and policy deployment and public health related gains. To model such interventions, we considered lock-downs as a series of action choices. The decision maker can plan different policies in different time periods based on the limitations mentioned above.

During the lock-down period in India, the change in estimated transmission time as the effect of interventions has been observed based on (Group 2020). This corresponds to T_{trans} in the SEIRD model we proposed and the effectiveness e varies in different regions. Thus we modeled the action as extending the transmission time to different degrees and different costs per day. Such a sequence of actions forms an intervention vector $\mathbf{a}$ of length T as a planning schedule with a total cost sum.

Multi-Objective Functions

In the public health domain, governments and decision makers may want to achieve different objectives when deploying a policy. One direct objective could be eliminating the disease which can be achieved by suppressing contacts so that patients recover at a rate greater than the spread of infection. This is equivalent to minimizing the area under the infection curve. However, this is not achievable in many regions, including cities in many of the developing countries due to the huge economic cost of such strict lock-downs. Thus, we focus on economically sustainable interventions that do not reduce R_0 below 1. In epidemic theory, this means the disease cannot be eliminated within reasonable time no matter how the government plans the lock-down in these regions. Every susceptible individual will eventually go through the recovered or deceased state. In other words, although the infection curve will change, the area under the curve will remain the same.

To evaluate the effectiveness of lock-down policies under these circumstances, we use indirect objectives that are vital and achievable for sustainable interventions. For example, as there is limited hospital capacity, a patient's quality of life will likely be better when the infected population does not exceed that capacity and they can receive proper treatment. Alternatively, we may want to delay the infection to the point when we have better system preparedness, medicines, resources etc. for handling the disease. These different desired objectives can be described as a family of objective functions that are variants of the Quality Adjusted Life Year (QALY) score in our model, which is elaborated below.

QALY is a popular established metric to quantify the effectiveness of health interventions. It is often used in the public health literature (Salomon et al. 2012). It measures the effectiveness of a certain intervention by combining

quantity and quality of health improvement. Specifically, a person's life quality at any given time is mapped to $[0, 1]$, with quality 1 corresponding to perfect health while 0 corresponding to death. QALY accumulates such measurements over time as its final score. Naturally, different disease conditions lie in the range $[0, 1]$ depending on severity.

In this work, we change the time scale from years to days to adapt to the dynamics of the disease we are facing. We mainly focus mainly on two objectives, burden and delay. We define these two objective functions as:

$$O_{Burden} = \sum_t ((I(t) - \lambda)\mathbf{1}_{I(t)>\lambda} - \delta_{Burden} l(a_t)) \quad (6)$$

$$O_{Delay} = \sum_t (tI(t) - \delta_{Delay} l(a_t)). \quad (7)$$

where t refers to timestep. $a_t, I(t)$ and $l(a_t)$ refer to action, infected population fraction and cost of action at time t respectively. Also δ and λ refer to economy-health weight and hospital capacity and $\mathbf{1}$ is the indicator function. Here, we focus on optimizing a linear combination of these two objective functions, written as:

$$O_{mix}(w) = w\bar{O}_{Burden} + (1 - w)\bar{O}_{Delay}, \quad (8)$$

where the weight w ranges from 0 to 1 and $\bar{O}$ is O normalized by the absolute value of no intervention, i.e., we divide O by its absolute value in the absence of interventions.

Formulation Using MDP

Our lock-down control problem can be modeled as a Markov Decision Process (MDP) (Yang, Sun, and Narasimhan 2019). Over the last two decades, reinforcement learning (Sutton et al. 1998) has provided an effective framework for solving an MDP in both theory and application. This is especially true when the system dynamics is either complicated, unknown, or the state dimensionality is too high for classical optimal control methods. In addition, the environment parameters we estimate from real-world hospital data involve uncertainty that cannot be ignored. Thus the output policy needs to be robust to such uncertainty.

We thus consider a parameter-wise robust reinforcement learning model to solve the MDP framework. The MDP framework can be written as:

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$$

with state space $\mathcal{S}$, action space $\mathcal{A}$, and transition distribution and vector reward

$$\mathcal{P}(s'|s, a) \text{ for } s, s' \in \mathcal{S} \text{ and } a \in \mathcal{A}$$

$$\mathbf{r}(s) \in \mathcal{R}$$

and the preference weight $w \in \mathbb{R}^n$.

The states we consider are the fractions of population present in S,E,I,R and D compartments at the given time. Furthermore, we consider several discrete actions at each time step corresponding to different strengths of lock-down with different costs. It is natural to assume the strength to be monotonically increasing with the cost as otherwise the action choice will be dominated by actions with less cost but more effectiveness. For simplicity, we adopt a linear mapping for both cost and effectiveness, as:

$$T_{trans}(a, e) = (1 + ec(a))\frac{T_{inf}}{R_0} \quad (9)$$

and $c : \mathcal{A} \to [0, 1]$, in which e is the lock-down effectiveness coefficient and R_0 the basic reproduction number when there are no lock-down interventions. Both of these are estimated with data from the city of Mumbai, India. T_{trans} and T_{inf} are the transmission and infection time periods.

For the remaining tuple, the transition distribution $\mathcal{P}(s'|s, a)$ is described as the disease transmission equation 1 to 5. The total accumulated reward is exactly the objective function O in equations 6, 7. The next section describes the distribution of individual reward signals across states.

Reinforcement Learning Approach

Multiple Objectives: We have defined the state, action, transition probabilities and total reward in the MDP section. The only missing piece for a complete reinforcement learning framework is to design the reward signal at every timestep. We have designed a framework that works not only for the two example objectives we focus in this work, but on most variants of QALY.

Most variants of QALY, including our examples, are related to time and population of certain health states. We propose a function we call the QALY value $V(x, t)$ which is a function of the population x in a certain health state and time t. We focus only on I or the Infectious state in these experiments. However, the QALY value function can be generalized to a vector form to include multiple states. For controlling the hospital capacity, the function can be formulated either as a constant penalty for x exceeding the capacity or simply as a reward for x below the threshold, since the area under the infection curve is a constant, as we elaborate in section Multi-Objective Functions. We formalize this as:

$$V_{Burden}(x, t) = 1 \text{ for } x < \lambda \quad (10)$$

As for delay, we formalize this function as

$$V_{Delay}(x, t) = \frac{t}{T} \quad (11)$$

Given that the QALY value $V(x, t)$ of the objective function is defined, the reward signal at any given time t can be calculated by $r(t) = \int_0^{I(t)} V(x, t)dt$. We can thus apply the reinforcement learning approach.

One benefit of such a proposed approach is that the QALY value function of the mixed objective can be easily calculated as:

$$V_{mix}(w) = \frac{wV_{Burden}}{O_{Burden}(\textit{no action})} + \frac{(1 - w)V_{Delay}}{O_{Delay}(\textit{no action})} \quad (12)$$

Uncertainty: Another important aspect other than having a multi-objective function in the lock-down application is the

uncertainty of the parameters (e, T_{inf}, T_{inc}), which are related to the infection curve directly or indirectly. We experiment with three approaches to analyze the effect of uncertainty in a reinforcement learning setup:
(1)**Fixed RL(FRL)**: Train the RL agent using only the mean of the uncertain parameters.
(2)**Distributed RL(DRL)**: Train the RL agent using samples of uncertain parameters from the estimated range.
(3)**Adversarial RL(ARL)**: Inspired by (Pinto et al. 2017), train the RL agent with another adversarial RL agent that will maliciously pick the worst possible parameter set for the RL agent during training. Note that the worst case parameter is not trivial to find as the policy changes. The action of the adversarial RL agent is set to be the discrete uncertain parameters in the disease model.

Experiments

In this section, we describe the application of our method to a specific location – the city of Mumbai, India. In subsequent subsections, we describe how we estimate the model parameters as well as the uncertainty in these parameters. We also report the results of our method when used on the estimated parameter ranges.

Parameter and Uncertainty Estimation

We fit our SEIRD model to the time-series data from the COVID19-India API (Group 2020) for the city of Mumbai. The data is aggregated in fields called Recovered, Deceased, Hospitalized and Total_Infected, where Total_Infected = Recovered + Hospitalized + Deceased. In our SEIRD model, we fit the I compartment to Total_Infected, D compartment to Deceased and R compartment to Hospitalized + Recovered. In this sense, the R compartment in our model estimates people who are either under recovery or have recovered, and thus are no longer infectious.

We decided the initial search space for the model parameters based on the estimates given by public health experts and those cited in literature. We process the data with smoothing techniques to reduce the effect of bulk data entry. Then, we search over the parameter space for parameter sets that have a small aggregated RMSE loss between predicted numbers and actual numbers using the Hyperopt library (Bergstra, Yamins, and Cox 2013). The parameter set giving the least loss value is taken to be the best-fit parameter set for the purposes of this experiment.

We found that there are diverse parameter sets that have loss close to the best-fit parameter set. Thus, we picked all parameter sets that have a loss within a certain range of the best loss (within 10%). Among all picked parameter sets, we find the range of values taken by individual parameters. These ranges for individual parameters give us a measure of uncertainty for these parameters. We assume a uniform distribution over these ranges as our parameter distribution.

Analysis and Results

Robustness: Robustness of policy to uncertainty in parameters is an important aspect. Over the estimated uniform distribution range, we find the worst-case parameters for different methods using a fine grid-search. Then, we measure the performance of different methods on their corresponding worst-case parameters. We also find the corresponding average performance over the parameter distribution. The results are tabulated in Tables 2 and 3. As shown in these tables, the ARL helps the reinforcement learning discover risky parameters and thus performs best in its worst case scenario. For average case, however, ARL performs worse than the best method (FRL and DRL respectively). This has shown the trade-off between performance and robustness in our lock-down problem - at the cost of average performance, we can obtain better worst-case performance.

Model Performance on Burden			
Model	**Worst**	**Mean**	**Std**
Random	-3.625	-2.030	0.500
FRL	-2.385	-1.234	0.372
DRL	-2.445	-1.313	0.467
ARL	**-2.226**	-1.279	0.385

Table 2: Burden objective in equation 6.

Model Performance on Delay			
Model	**Worst**	**Mean**	**Std**
Random	-759.522	2.110	186.413
FRL	-29.486	163.403	39.201
DRL	-100.538	235.391	110.176
ARL	**9.933**	189.307	50.813

Table 3: Delay objective in equation 7.

Different objectives: We use different weights between Burden and Delay objectives and compare the results to the case when we individually focus on Delay and Burden in Table 4. The objective function we use is $(1-w) * \text{Delay} + w * \text{Burden}$ for different values of w. The aim is to maximize normalized objective for both Burden and Delay.

From Table 4, we observe that, as expected, as the weight on Burden increases, the Burden objective becomes larger for all methods in general. Similar behaviour is observed for Delay as well. When applying this method for policy guidance, we can tune w to achieve the required objectives for both Burden and Delay.

Conclusions and Future Work

We implemented reinforcement learning on the lock-down policy optimization problem for COVID-19 while considering important real-world aspects like robustness and multi-objective optimization. Robustness can be achieved by introducing an adversarial agent for parameter discovery, but at the cost of sacrificing some performance on average. For the multi-objective mixture, we study the trade-off between controlling hospital capacity and delaying the infection spread. We proposed a reward distribution framework for the reinforcement learning agent to shift from one objective to another in the lock-down problem. One point to note is that our epidemiological model (SEIRD) is a homogeneous model and is being used to optimize the policy keeping the trade-off between economy and health for the community as a whole.

Model	w	Burden	Delay	Mixed
Random	0.0	-1.074	0.675	0.675
FRL		-1.372	1.478	1.478
DRL		-1.155	1.711	1.711
ARL		-1.442	1.727	1.727
Random	0.25	-1.230	0.843	0.318
FRL		-1.017	1.154	0.611
DRL		-1.086	1.673	0.983
ARL		-0.993	1.103	0.579
Random	0.5	-1.073	0.641	-0.216
FRL		-0.622	1.207	0.293
DRL		-0.738	1.196	0.229
ARL		-0.912	1.208	0.148
Random	0.75	-1.019	0.670	-0.597
FRL		-0.550	1.090	-0.140
DRL		-0.818	1.159	-0.324
ARL		-0.703	0.932	-0.294
Random	1.0	-1.193	0.587	-1.193
FRL		-0.734	1.400	-0.734
DRL		-0.620	1.014	-0.620
ARL		-0.671	1.001	-0.671

Table 4: Model Performance for Mixed Objectives. The scores are calculated based on equation 12

The model does not discriminate between two infected individuals based on their economic contribution and neither is it capable for the same. This makes sure that we generate lockdown policy as fairly as possible.

The future direction of this work is to gather more data on both cost and effectiveness of the real-world lock-down policies on community scale so that a more complex model can be used to better estimate the real-world scenarios. For example, transmission times are known to not be homogeneous and several super-spreader events have been identified in many different spreading routes. Collecting data on such cases and modifying the model to have different transmission times for different cases of spread would give us a more holistic view of the entire scenario. Another important direction of extension would be estimating the reporting rate from other sources of data and normalizing the reported numbers to estimate parameters that are closer to the real-world.

Acknowledgements

This study is made possible by the generous support of the American People through the United States Agency for International Development (USAID). The work described in this article was implemented under the TRACETB Project, managed by WIAI under the terms of Cooperative Agreement Number 72038620CA00006. The contents of this manuscript are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

References

Ball, F. G.; Knock, E. S.; and O'Neill, P. D. 2015. Stochastic epidemic models featuring contact tracing with delays. *Mathematical biosciences* 266: 23–35.

Bannur, N.; Maheshwari, H.; Jain, S.; Shetty, S.; Merugu, S.; and Raval, A. 2021. Adaptive COVID-19 Forecasting via Bayesian Optimization. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '21. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3430984.3431047. URL https://doi.org/10.1145/3430984.3431047.

Bergstra, J.; Yamins, D.; and Cox, D. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, 115–123.

Bernoulli, D.; and Blower, S. 2004. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in medical virology* 14(5): 275–288.

Carpin, S.; Chow, Y.-L.; and Pavone, M. 2016. Risk aversion in finite Markov Decision Processes using total cost criteria and average value at risk. In *2016 ieee international conference on robotics and automation (icra)*, 335–342. IEEE.

Censor, Y. 1977. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization* 4(1): 41–59.

Chow, Y.; Ghavamzadeh, M.; Janson, L.; and Pavone, M. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research* 18(1): 6070–6120.

Christiano, P.; Shah, Z.; Mordatch, I.; Schneider, J.; Blackwell, T.; Tobin, J.; Abbeel, P.; and Zaremba, W. 2016. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518* .

Deb, K. 2014. Multi-objective optimization. In *Search methodologies*, 403–449. Springer.

Ferguson, N.; Laydon, D.; Nedjati-Gilani, G.; Imai, N.; Ainslie, K.; Baguelin, M.; Bhatia, S.; Boonyasiri, A.; Cucunubá, Z.; Cuomo-Dannenburg, G.; et al. 2020. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *Imperial College London* 10: 77482.

Ganesh, A.; Massoulié, L.; and Towsley, D. 2005. The effect of network topology on the spread of epidemics. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 2, 1455–1466. IEEE.

Group, C.-. I. O. D. O. 2020. Accessed on yyyy-mm-dd from https://api.covid19india.org/.

Hoffmann, J.; and Caramanis, C. 2018. The Cost of Uncertainty in Curing Epidemics. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 2(2): 31.

Mannor, S.; Simester, D.; Sun, P.; and Tsitsiklis, J. N. 2004. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, 72.

Mihatsch, O.; and Neuneier, R. 2002. Risk-sensitive reinforcement learning. *Machine learning* 49(2-3): 267–290.

Nilim, A.; and El Ghaoui, L. 2004. Robustness in Markov decision problems with uncertain transition matrices. In *Advances in neural information processing systems*, 839–846.

Nilim, A.; and El Ghaoui, L. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* 53(5): 780–798.

Pan, X.; You, Y.; Wang, Z.; and Lu, C. 2017. Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952* .

Pattanaik, A.; Tang, Z.; Liu, S.; Bommannan, G.; and Chowdhary, G. 2017. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632* .

Pinto, L.; Davidson, J.; and Gupta, A. 2017. Supervision via competition: Robot adversaries for learning tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 1601–1608. IEEE.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. *arXiv preprint arXiv:1703.02702* .

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48: 67–113.

Salomon, J. A.; Vos, T.; Hogan, D. R.; Gagnon, M.; Naghavi, M.; Mokdad, A.; Begum, N.; Shah, R.; Karyana, M.; Kosen, S.; et al. 2012. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *The Lancet* 380(9859): 2129–2143.

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.

Sun, C.; and Hsieh, Y.-H. 2010. Global analysis of an SEIR model with varying population size and vaccination. *Applied Mathematical Modelling* 34(10): 2685–2697.

Sutton, R. S.; et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Van Moffaert, K.; and Nowé, A. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research* 15(1): 3483–3512.

Wang, N. 2005. *Modeling and analysis of massive social networks*. Ph.D. thesis, UMD.

Weitz, J. S.; and Dushoff, J. 2015. Modeling post-death transmission of Ebola: challenges for inference and opportunities for control. *Scientific reports* 5: 8751.

White III, C. C.; and Eldeib, H. K. 1994. Markov decision processes with imprecise transition probabilities. *Operations Research* 42(4): 739–749.

Yang, R.; Sun, X.; and Narasimhan, K. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, 14636–14647.

Zhang, Y.; and Prakash, B. A. 2015. Data-aware vaccine allocation over large networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10(2): 20.