Machine Learning-Powered Mitigation Policy Optimization in Epidemiological Models

Jayaraman J. Thiagarajan [1] **Rushil Anirudh** [1] **Peer-Timo Bremer** [1] **Timothy Germann** [2] **Sara Del Valle** [2] **Frederick Streitz** [1]

Abstract

A crucial aspect of managing a public health crisis is to effectively balance prevention and mitigation strategies, while taking their socio-economic impact into account. In particular, determining the influence of different non-pharmaceutical interventions (NPIs) on the effective use of public resources is an important problem, given the uncertainties on when a vaccine will be made available. In this paper, we propose a new approach for obtaining optimal policy recommendations based on epidemiological models, which can characterize the disease progression under different interventions, and a look-ahead reward optimization strategy to choose the suitable NPI at different stages of an epidemic. Given the time delay inherent in any epidemiological model and the exponential nature especially of an unmanaged epidemic, we find that such a look-ahead strategy infers non-trivial policies that adhere well to the constraints specified. Using two different epidemiological models, namely SEIR and EpiCast, we evaluate the proposed algorithm to determine the optimal NPI policy, under a constraint on the number of daily new cases and the primary reward being the absence of restrictions.

1. Introduction

One of the key challenges in managing a public health crisis is the allocation of scarce resources and how to balance the cost-benefits of mitigation strategies. This is especially true for non-pharmaceutical interventions (NPIs), which potentially impact a large number of otherwise not affected populations (Ferguson et al., 2020; Morato et al., 2020). In order to asses the impact of optimal NPIs, one can parameterize epidemiological models to represent disease progression at different levels of interventions, and measure the predicted number of new infections (Ghamizi et al., 2020). Conceptually, this can be used to formulate an optimization problem in the context of non-linear control, wherein users can assign a cost/reward to different levels of interventions and specify constraints on the peak response (Alvarez et al., 2020; Yaesoubi & Cohen, 2016). However, epidemiological models pose a number of unique challenges in both formulating the problem as well as solving it.

The first challenge is to find a meaningful formulation of what can be considered optimal under which constraints (Khadilkar et al., 2020; Libin et al., 2020). An intuitive objective would be to minimize harm, i.e. aim to minimize the number of infected or potential casualties. However, this directly leads to a trivial solution of applying maximal NPIs (e.g. lockdown) at all times, which by design will lead to the fewest number of infections. However, this disregards any socio-economic costs, i.e. lost business activity or negative health effects due to inactivity. In practice, balancing the different aspects explicitly is challenging and requires not only in-depth economic studies (Guerrieri et al., 2020) but also knowledge of when vaccines might become available or what role herd immunity will play as an exit strategy (Lurie et al., 2020), etc. Instead, we use a common simplification which reformulates the problem in terms of health care management with the primary constraint being the number of daily new cases, and a reward assignment process that encourages lessened restrictions. This reflects the desire to guarantee adequate care for all sick given the finite healthcare resources, and assumes that in isolation applying any NPI will incur a cost to be avoided. As will be discussed in more detail below, our approach allows policy makers to assign relative rewards to different NPIs reflecting the differences between, for example, closing all non-essential businesses vs. preventing large gatherings.

The second challenge is the time delay inherent in any epidemiological model and the exponential nature of an unmanaged epidemic (Liu, 2020; Germann et al., 2019). Due to effects such as incubation times or asymptomatic infections, interventions taken (or removed) today may show a significant effect only days or weeks later. Simultaneously,

[1]Lawrence Livermore National Laboratory [2]Los Alamos National Laboratory. Correspondence to: Jayaraman J. Thiagarajan <jjayaram@llnl.gov>.

Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022.

allowing an exponential growth even within a given constraint may result in unavoidable future violations despite maximal NPIs. Both of these issues emphasize the need for a substantial look-ahead window to consider not only the current state but a long term forecast. Ideally, this window should be "infinite" to avoid any possibility of unexpected problems due to delayed dynamics. However, this formulation makes the optimization expensive especially for more complex models. Here, we present a greedy approach based on a two-level look-ahead strategy able to produce optimal policy recommendations for a variety of different scenarios. Furthermore, the optimization incorporates practical considerations, such as, the limited frequencies at which public policies can be changed, or the variable cost of different interventions. Finally, the two-level optimization provides an intuitive trade-off between short term gains and potential future costs beyond the total reward inherent in the system. For example, the optimal policy might place relatively stringent early NPIs to avoid predicted peaks months later. However, given the uncertainty of long term predictions one may prefer a more optimistic choice early as long as there is sufficient lead time to prevent peaks later even if this produces an overall lower reward. Using empirical studies with two popular epidemiological models, namely SEIR and agent-based EpiCast, we demonstrate the effectiveness of our approach.

2. Background: Epidemiological Models

2.1. SEIR Model

We consider a compartmental SEIR model from which one can obtain trajectories of the epidemic, given the current state and epidemic-specific parameters. An SEIR model divides the population into *Susceptible*, *Exposed*, *Infected* and *Recovered* compartments and can be described in terms of ordinary differential equations (ODEs) (He et al., 2020). While exposed refers to the latent infected but not yet infectious population, recovered contains the population that is no longer infectious (also referred as removed). While this has been popularly used to model influenza epidemics (Mills et al., 2004), there have been several existing efforts that have utilized this model with great success in the case of the recent COVID-19 epidemic (López & Rodo, 2020; Roda et al., 2020; Yang et al., 2020). Formally, an SEIR model is described as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta S(t) I(t) \\ \frac{dE}{dt} &= \beta S(t) I(t) - \sigma E(t) \\ \frac{dI}{dt} &= \sigma E(t) - \gamma I(t) \\ \frac{dR}{dt} &= \gamma I(t). \end{aligned} \tag{1}$$

Here, the the rate of changes in the compartments are parameterized by infectious rate β, incubation rate σ and the recovery rate γ. The severity of an epidemic is characterized by the *basic reproduction number* that quantifies the number of secondary infections from an individual in an entirely susceptible population as $R_0 = \frac{\beta}{\gamma}$. Following common parameter settings assumed by COVID-19 studies in different countries (López & Rodo, 2020; Yang et al., 2020; He et al., 2020), we set $\gamma = 0.1$, $\sigma = 0.2$. Interestingly, even in this simple model, one can introduce the effects of different NPI choices through changes in the infectious rate β. While there are existing works that attempt to estimate the current value of β using additional data sources (e.g. mobility), in order to better fit the observed trajectory (Soures et al., 2020), our focus is on choosing the optimal policy of intervention. In particular, we define the set of NPI choices through corresponding β values $\mathcal{N} := \{0.25, 0.3, 0.5, 0.7, 0.8, 0.9\}$, where higher β implies lower restrictions.

2.2. Agent-based EpiCast Model

EpiCast is an individual-based model, with daily contacts between people in household, workplace, school, neighborhood, and community settings. The primary data source is U.S. Census demographics at the tract level (the $\sim 65,000$ tracts are subsets of the ~ 3000 counties, with typically a few thousand people in each tract), and Census tract-to-tract workerflow data (i.e., how many people live in tract A and work in tract B). This is used to construct a model population with tract-level age and household size demographics, and realistic daily workflow pattern, which captures most of the short-range mobility. In addition, occasional long-distance travel is possible. A 12-hour timestep is used, so (unless on travel) individuals spend the night-time at home and day-time at school or workplace, if they belong to one (and they are open). Additional details are provided in the Supporting Information of (Germann et al., 2006). In the original model (Germann et al., 2006; Halloran et al., 2008), the individual age- and context-specific contact rates that account for the duration and closeness of interactions between pairs of individuals in different settings (home, school, workplace, neighborhood, community, etc.) were uniform across the US. In a recent school dismissal study (Germann et al., 2019), different communities were allowed to close their schools at different times, depending upon the current local disease incidence. In adapting this model to COVID-19, these local policies have been extended to all community mitigation measures: school dismissal, workplace closure, shelter-in-place, and other social distancing.

3. Problem Formulation

We first provide a formal definition of the policy optimization problem. Without loss of generality, let us denote the

current state of an epidemic, as described by a model E, as $\mathrm{X}[t]$, where t is the time step. Here, X can refer to factors of interest such as – the set of susceptible, exposed, infected and removed populations in an SEIR model (Hethcote & Van den Driessche, 1991) or a more complex, fine-grained internal states of an agent-based system(Germann et al., 2006). One can evaluate the future states by evaluating the epidemiological model $\mathrm{X}[t+1:t+k] = \mathrm{E}(X[t]; \mathrm{n}[t], k)$, where $\mathrm{n}[t]$ denotes the NPI applied at time-step t. In our formulation, we assume that each $n[t] \in \mathcal{N}$, where $\mathcal{N}$ is a set of discrete NPI choices ordered by severity, i.e. $\mathcal{N}(i)$ is more restrictive with a smaller reward than $\mathcal{N}(i+1)$. Note that in practice, the NPI choices will typically be represented by some combination of model parameters depending on the model. For example, in case of an SEIR model, $\mathcal{N}$ represents a set of infection rates β which correspond to more or less stringent NPIs. We refer to a sequence of NPIs, $\mathcal{P} = \{n[1], n[2], \cdots, n[T]\}$, as a *policy*. With $\mathcal{C} \in \mathbb{R}^{|\mathcal{N}|}$ denoting the reward for adopting each of the NPI choices in $\mathcal{N}$, one can compute the reward for the policy as $\mathcal{R} = r[1] + r[2] + \cdots + r[T]\}, \forall r[t] \in \mathcal{C}$. Here, the symbol $|.|$ is the cardinality of a set.

The design objective of not overloading the healthcare system is specified by a threshold on the maximum number of new daily cases, denoted by τ. Note that, other constraints could be used, i.e. the number of currently active cases, predicted number of patients requiring the ICU, etc. Finally, we consider policies in sets of d days both to reflect the fact that public policy cannot realistically be changed on a day to day basis as well as to make the optimization more tractable. The goal is to maximize $\mathcal{R}$ while not violating τ anytime during the course of the epidemic. While a few studies have been recently proposed in the literature to utilize reinforcement learning approaches with a tractable epidemiological model to optimize for mitigation policies (Libin et al., 2020; Khadilkar et al., 2020; Ghamizi et al., 2020; Liu, 2020), our goal is to deal with complex models, such as the agent-based EpiCast considered in our study, which is computationally expensive to be repeatedly evaluated. Furthermore, we are interested in designing a scalable optimization approach that can be rapidly executed for a wide-variety of constraints and specifications, which is known to be a bottleneck for episodic-training based reinforcement learing methods.

4. Approach

We now describe the new two-level optimization used to solve the problem described above, followed by a discussion on dealing with models that are computationally expensive to be optimized directly.

The key novelty of our approach for solving the above optimization problem is to split the reward using a two-level scheme with a finite time horizon that enables an efficient early termination of a greedy search. In particular, policy choices are made every d days (*frequency*) based on combining a short-term and a long-term reward:

Short-term reward is computed from the current day for k days forward and a policy choice is given its full reward if the constraint is met for all k days and no reward if it is predicted to violate the constraint at any point. This is equivalent to a brute force search of all policy choices and the straight-forward check of the constraint, both restricted to a short-term forecast that can be computed efficiently. Note that typically $k > d$ which represents a first restriction on the search space, as in principle, one can initially relax the policy and subsequently switch to more a stringent NPI at a later stage, and thereby prevent the constraint violation. In practice, the short-term forecast represents a hard cutoff on response times, for example, to mobilize additional ICUs and provides a guaranteed (within the accuracy of the model) window of time for interventions.

Long-term reward is evaluated for an additional k_s days (on top of k days from short-term reward computation) using a potentially different NPI. Conceptually, the long-term reward reflects an optimistic choice to relax NPIs for the next d days even if they are predicted to become problematic within $k + k_s$ days as long as tightening restrictions after k days can correct for any violations. In other words, the search explores the NPI from the first k days as well as potentially more stringent NPIs (in the next k_s days) that can help control the disease propagation. However, unlike the short-term reward which is a constant for all k days (zero if the constraint is violated at any point), long-term rewards are assigned proportional to the number of days the selected policy remains within the given constraint. Therefore, a more relaxed policy that violates the constraint at some point can eventually accumulate a higher reward than another one that stays restrictive during the entire forecasting period. Similar to the short-term reward, the limit to k_s days reduces the total simulation time necessary for the optimization. Furthermore, we reduce the number of explored scenarios by only considering more restrictive NPIs. Assuming a more relaxed NPI is acceptable past the initial k days, it will always be explored in the next outer loop in $d < k$ days.

An additional benefit of providing users the ability to choose these time scales is that it balances the level of risk with the uncertainties inherent in forecasting and enables one to compensate for external factors. For example, if we expect new treatments to become available, choosing a smaller k_s results in more aggressive policy choices while still considering the need to correct for problems if this hope does not materialize. We use different set of NPI choices and corresponding reward assignments for SEIR and Epicast models, which are provided in the next section. Intuitively, higher reward is assigned to relaxed policy choice, taking

Algorithm 1 Mitigation policy optimization.

input : Epidemiological model E;
Initial state $X[0]$; look-ahead parameters k, k_s; threshold τ; frequency of NPI change d; Time steps T;
Set of NPI choices $\mathcal{N}$; Set of rewards for each of the NPIs $\mathcal{C}$.
output : Policy $\mathcal{P}$, Reward $\mathcal{R}$
for $t \leftarrow 1$ **to** T **by** d **do**
  for $i \leftarrow 1$ **to** $|\mathcal{N}|$ **do**
    $\bar{\mathrm{X}}[1:k] = E(\mathrm{X}[t-1]; \mathcal{N}(i), k)$; /* Run E for k steps with $\mathcal{N}(i)$ */
    Using $\bar{\mathrm{X}}[1:k]$, obtain the counts of new infected cases at each time step $N_c[1:k]$;
    if $\max(N_c[1:k]) > \tau$ **then**
      break
    end
    $\text{reward}_{\text{short}}[i] = \mathcal{C}(i)$*k; /* Compute short-term reward */
    for $j \leftarrow i$ **to** 1 **by** -1 **do**
      $\tilde{\mathrm{X}}[1:k_s] = E(\bar{\mathrm{X}}[k]; \mathcal{N}(j), k_s)$; /* Run E for k_s steps with $\mathcal{N}(j)$ */
      Using $\tilde{\mathrm{X}}[1:k_s]$, obtain the counts of new infected cases at each time step $\tilde{N}_c[1:k_s]$;
      /* Find the first index where the threshold is violated */
      for $ind \leftarrow 1$ **to** k_s **do**
        if $\tilde{N}_c[ind] > \tau$ **then**
          break
        end
      end
      $f[j] = (ind-1) * \mathcal{C}(j)$; /* long-term reward when one NPI switch is allowed */
    end
    $\text{reward}_{\text{long}}[i] = \max(f)$
    $\text{reward}_{\text{net}}[i] = \text{reward}_{\text{short}}[i] + \text{reward}_{\text{long}}[i]$
  end
  $n = \arg\max(\text{reward}_{\text{net}})$; /* use NPI with the highest net score */
  $\mathcal{P}[t:t+d] = \mathcal{N}(n)$;
  $\mathcal{R}[t:t+d] = \mathcal{C}(n)$;
  $\mathrm{X}[t:t+d] = E(\mathrm{X}[t-1]; \mathcal{N}(n), d)$; /* Evaluate E for d steps with the chosen NPI */
end

into account longer term societal impact.

4.1. Surrogate Model

At the core of our approach is the need to obtain look-ahead estimates of the pandemic state given the current state $\mathrm{X}[t]$ and the NPI $n[t]$, which in turn requires evaluation of the epidemiological model E. In cases where this evaluation is computationally expensive, it is beneficial to build a machine learned surrogate model that predicts the future states given the current state and the NPI choice. Formally, we build the surrogate to produce $k-$step predictions, $\mathrm{X}[t+1:t+K] = \hat{F}(X[t]; n[t])$, where k is the look-ahead parameter. Note that, since the ODE-solver for SEIR is highly efficient, we do not use a surrogate in that case. However, the agent-based EpiCast model is computationally expensive and hence we build a machine learning surrogate for faster evaluation. The details of the surrogate model are provided in Section 5.2.

4.2. Algorithm

A detailed description of our optimization is given by Algorithm 1. Given an epidemiological model E (or a surrogate model in the case of EpiCast), its initial state $\mathrm{X}[0]$ and the threshold τ as inputs, the policy optimization assumes that NPIs can be changed every d time steps. The outer loop runs once every d days (frequency of policy change) and an inner loop that iterates through the NPI choices. For every NPI, we first forecast the state of the infection for k days and the corresponding short-term reward is estimated. Note that, the short-term reward is assigned either fully for all k days, if the constraint is never violated, or none at all.

Next, in order to compute the long-term reward, we consider the subset of stricter policies (referred by index j) compared to the choice (referred by index i) made in the first k days. For each j, we forecast the infection state after k_s days with the initial state set to $\bar{\mathrm{X}}[k]$. The long-term reward for switching to policy index j, *i.e.*, $f[j]$, is thus computed as the

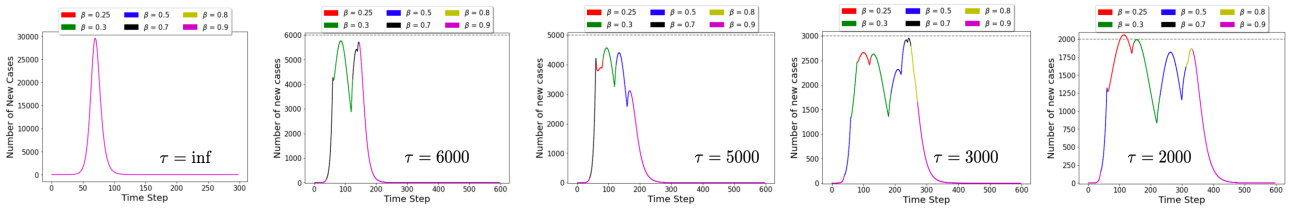

Figure 1. β **Optimization for SEIR.** Policies inferred using the proposed approach with an SEIR model. The NPI choices comprised of different settings for the infectious rate β, which is known to be strongly linked to interventions. We show the policies obtained for different values of the threshold τ.

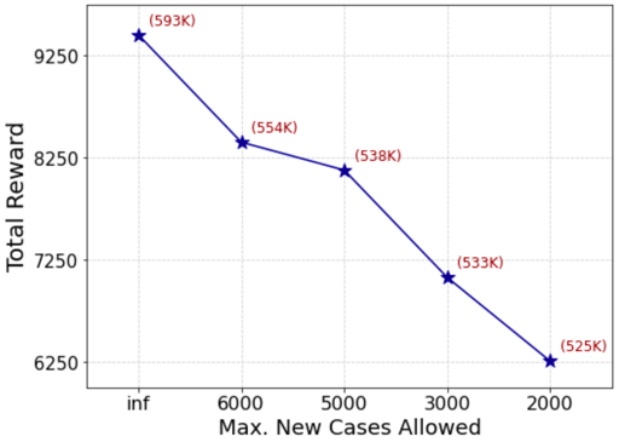

Figure 2. **Effect of τ on the optimal policy.** For each case, we show the total reward and the cumulative number of infections (in parentheses). We find that as τ becomes lower, we can obtain more conservative policies in terms of infections, but comes at the price of diminished rewards.

total reward accumulated over all days between k and $k+k_s$ when the constraint is not violated (number of infections is lesser than τ). We repeat this for all j's, and assign the long-term reward for policy $\mathcal{N}(i)$ as the maximum over all $f[j]$. Finally, the short-term and long-term rewards are combined to rank the NPI choices for the current interval.

5. Case Studies

In this section, we use the proposed optimization algorithm to determine the NPI policy using both SEIR and EpiCast models. All our empirical studies are carried out using different scenarios, and we employ a simple reward assignment for each of the NPI choices. In both the models, we set the frequency of changing the NPI $d = 14$ days. As described earlier, while SEIR reflects the effects of NPI coarsely through the change in infectious rate β, EpiCast allows more fine-grained characterization of school or business restrictions.

5.1. Optimizing β Switches in SEIR

In the case of SEIR, we consider the scenario where we are at the beginning of a pandemic, i.e $I(0) = 0$, $E(0) = 1$, $R(0) = 0$. Since the four compartments in the SEIR model add up to the total population, the susceptible compartment $S(0) = N - 1$. In this study, the population N is set to $3e6$. As discussed earlier, following existing work on COVID-19, we set the incubation rate $\sigma = 0.2$ and recovery rate $\gamma = 0.1$. Given the set of NPI choices in the form of β values (infectious rate) $\mathcal{N} = \{0.25, 0.3, 0.5, 0.7, 0.8, 0.9\}$ and the corresponding reward assignments $\mathcal{C} \coloneqq \{1, 3, 6, 8, 12, 15\}$. The higher the infectious rate β, more severe the infection is or less restrictive the interventions are.

The naïve policy of using the most restrictive NPI will provide the highest reward, when there is no constraint on the limit on the number of new cases each day, i.e., the threshold $\tau = inf$. As showed in Figure 1, this naïve policy peaks within the first 75 days and the number of new cases reaches as high as $30K$. This unrestricted policy can naturally lead to significant overheads to the healthcare system. To circumvent this, we perform the optimization in Algorithm 1, by placing an upper limit on the number of cases, $\tau = \{6000, 5000, 3000, 2000\}$. As one might expect, the policy selection over the entire period is highly non-trivial, given the combinatorial nature of this optimization process. In our algorithm, we set the look-ahead parameters $k = 21$ days and $k_s = 35$ days, i.e., a total of 8 weeks.

In Figure 1, we illustrate the policies inferred using our approach for different values of τ. Interestingly, in all cases, our greedy optimization produces effective policies that meet the constraint. As one might expect, the policies become more complex as τ becomes lower. We make two key observations from the results. First, due to the constraint τ, the epidemic takes a much longer duration to flatten; for example at $\tau = 3000$ it takes about 350 time steps when compared to ~ 100 time steps in the case of $\tau = inf$. Second, our optimization balances short-term and long-term rewards, thus producing policies that are not overly conservative. For example, unless the constraint τ becomes very strong (say 2000), the most restrictive NPI ($\beta = 0.25$) is rarely necessary. This effectively balances between the

exponential growth of the epidemic and societal impacts of total inactivity through highly restrictive interventions.

Since the duration for flattening is much longer with our policies, due to the use of less restrictive NPIs or no intervention ($\beta = 0.9$), it is important to study the total number of cases across the entire duration. This will indicate if the relaxations recommended by the inferred policy eventually leads to more (cumulative) cases. To study this trade-off, Figure 2 shows the total reward and the cumulative number of cases achieved using the different policies. Interestingly, with only a little compromise in the total reward, our policy achieves significant reduction in the total number of cases (indicated in parentheses near each marker). For example, using $\tau = 6000$ results in $\sim$ 40K lesser cases in total, when compared to $\tau = inf$. However, as the constraint on τ becomes more severe, i.e. 3000 or 2000, we see effects of herd immunity as sufficient number of cases still appear.

5.2. Optimizing Phase Switches in EpiCast

In this section, we describe how our optimization approach can be used with the agent-based EpiCast model. When compared to the SEIR model, EpiCast can model the impact of fine-grained interventions (e.g., 2 days of school closure every week vs full closure) and hence can contain more complex dynamics, particularly when there is an interplay between different intervention choices. This improved modeling comes at the price of high computational complexity as well as challenges in adjusting policies on the fly.

To make the problem tractable epidemiologists have chosen to split the different parameter choices into a set of discrete policies, i.e. a given school or work schedule, and a number of continuous parameters. Furthermore, as these simulations require complex workflows and large scale parallel computing it is practically infeasible to change either policy or parameters during an individual run thus leaving parameter matching as well as optimization to an outer loop solution.

- **Parameters of EpiCast**: probability of transmission between individuals, percentage of asymptomatic infections, and relative infectiousness;
- **State specification**: number of currently infected individuals, number of recovered (and assumed immune) patients, and the level of compliance to non-pharmaceutical interventions like masks.

We use two sets of scenarios with respect to school and industry operations, that roughly match the phases outlined in (The United States Government, 2020).

School Closure NPIs: (a) Phase 0, with only essential businesses open and no schools (P0); (b) Phase 1 which corresponds to slightly more businesses and less stringent distancing measures but with schools still universally closed (P1) (c) Phase 2 with even less restrictions on businesses and schools potentially opening – 5 days of school (P2a), 3 days of school (P2c), 2 split cohorts each with 2 days of school (P2d); and 1 day of school (P2e).

Industry Closure NPIs: (a-b) 5 day workweek (P1q, P2q); (c-d) 3 day work week (P1r, P2r); (e-f) two split cohorts working 5 out of 10 days (P1v, P2v). Note that, in phase 1 we assume that only a subset of the businesses to be open, while in phase 2 the restrictions are relaxed and more businesses are operational.

Our goal is to develop a framework for policy makers that given a set of (potentially complex) scenarios expressed through EpiCast would provide suggestions on optimal policy choices given certain constraints. Since the computational costs make a direct policy optimization as in the SEIR model is infeasible, we propose to first build surrogate models for the different policies each able to emulate EpiCast across a wide range of parameter settings.

Experiment Design for EpiCast. While building surrogate models is typical in scientific problems (Koziel & Leifsson, 2013), the quality of the surrogates relies directly on the experiment design used to generate the dataset. To create the necessary training data for these models, we introduce an iterative approach aimed at reducing the computational costs as well as improving model fits. The challenge in fitting a surrogate models to all phases of the disease is that large portions of the parameter space are invalid, in particular with respect to the initially infected and recovered populations, i.e. a large number of currently infected without anyone recovered. While it is possible to execute EpiCast with any parameter setting, using unrealistic combinations wastes compute resources and likely affects the quality of the surrogate. However, it is unclear *a* priori which parameter configurations may be valid or how to sample from this space. Instead, we first create simulation ensembles resembling realistic early outbreaks, i.e. few infected and recovered individuals and simulate the corresponding disease progression until outbreak has passed (typically 360 days or longer). We then analyze all intermediate epidemic states and create additional simulations using these as starting conditions while varying the remaining disease parameters. In our experience two iterations of this process, meaning three sets of simulation ensembles, are sufficient to cover the parameter space densely enough for surrogate modeling.

Surrogate Model. Given the dataset generated using the experiment design described earlier, we build a machine-learned surrogate that can map the initial pandemic state $\mathrm{X}[0]$ to predict the future trajectory of the states. Since predicting long-term trends can be challenging with EpiCast, we resort to estimating the states for the next 21 time steps.

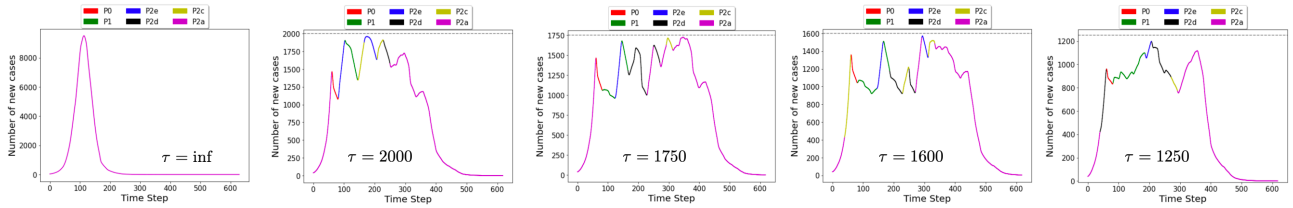

Figure 3. **School opening policy optimization** - Policies inferred using our approach with the EpiCast model. The NPI choices comprised of different phases of school opening. We show the policies obtained for varying values of the threshold τ.

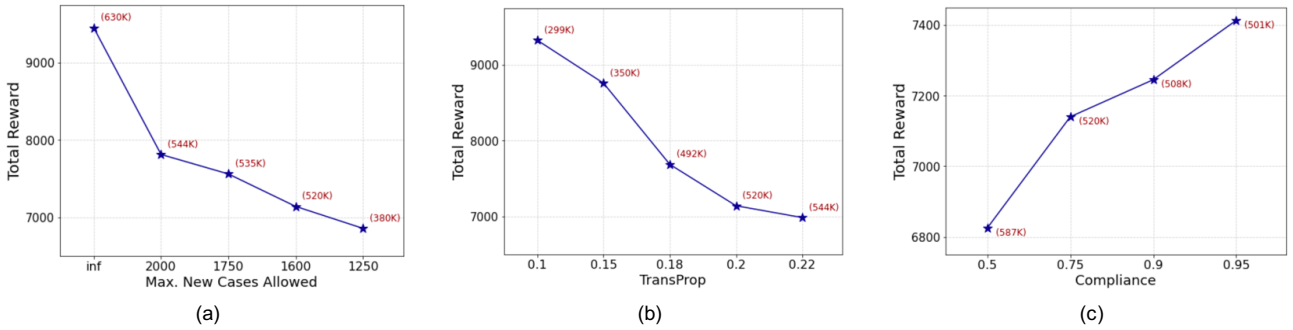

Figure 4. **Analysis**. (a) Effect of τ on the cumulative infections and achieveable reward; (b) The impact of the transmission probability parameter on the policy. We find that higher TransProp leads to a spike in the total infections, while also reducing the reward significantly; (c) For a given TransProp, the Compliance parameter reveals a positive impact on the quality of the policy.

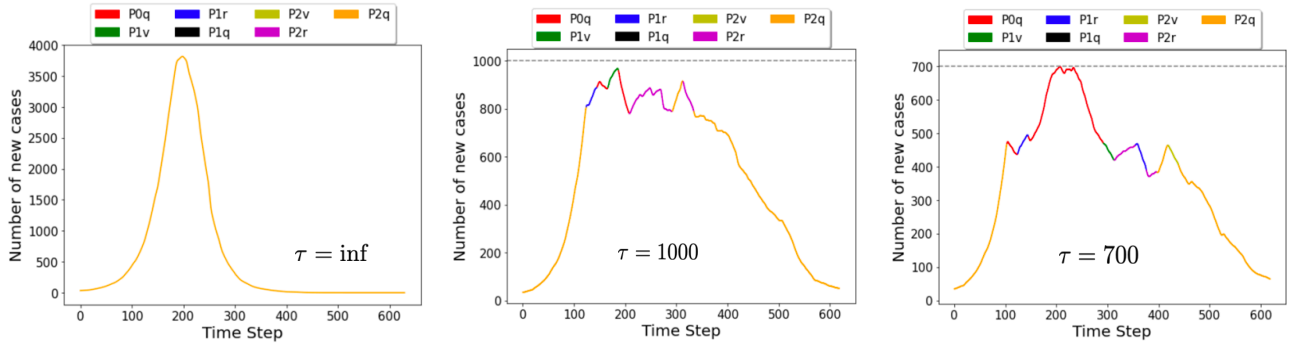

Figure 5. **Industry opening policy optimization** - Policies inferred using our approach with the EpiCast model. The NPI choices comprised of different schedules for industry opening. We show the policies obtained for varying values of the threshold τ

We create this dataset using sliding windows on the EpiCast samples and design a surrogate F that takes as input the current state (6-dimensional input) and outputs the trajectory curves for number of infected and removed cases.

Following the recent surrogate modeling literature (Anirudh et al., 2020), we first constructed a low-dimensional latent space to better capture the structure in the curves. We explored the use of a simplified formulation with principal component analysis (PCA) on the curves (concatenated) as well as a more sophisticated multi-variate sequence-to-sequence models (autoencoders). We find that both these strategies are capable of accurately representing the short-term dynamics, as measured using reconstruction error of held-out validation data. Note that, we constructed a latent space of 5 dimensions for each of the NPI choices considered (different school closing schedules). This pre-training step reformulates the surrogate modeling problem as predicting into the latent space, in lieu of the curves directly.

The surrogate model was implemented as a fully connected network with 5 hidden layers (configuration $[64, 128, 256, 64, 32]$) and ReLU activations. We trained the model with the MSE objective and an ℓ_2 regularizer on the weights using the Adam optimizer with learning rate 0.001. We also compared the performance of this model against a random forests regressor containing 100 trees and found the fully connected network to be marginally better –

R-squared statistics of 0.94 and 0.93 respectively.

(i) Policies for school closure. In this study, we considered a scenario where the total population for a region of interest was set to $1e7$ and the initial state of the pandemic was set to the following values:

- Number of infected: 250; Number of removed: 25K
- TransProp: 0.2; Asymptomatic ratio: 0.3
- Relative infectiousness: 0.9; Compliance: 0.75

The set of NPI choices comprised of 6 different settings under phases 0,1 and 2 ($\mathcal{N}$ = [P0,P1,P2e, P2d, P2c, P2a]) with the corresponding rewards $\mathcal{C} = [1, 3, 6, 8, 12, 15]$. Similar to the previous experiment, we set the frequency of NPI change at $d = 14$ days and used the look-ahead parameters $k = 21$ and $k_s = 35$ days respectively. We ran the proposed algorithm with the pre-trained surrogates for each of the 6 NPI settings and computed the optimal mitigation policy for a duration of 600 time-steps. The most relaxed policy P2a corresponds to the case of $\tau = inf$ and achieves the maximum reward. However, as showed in Figure 3, one can obtain more realistic policies by placing a constraint on the number of new cases.

In particular, as we reduce τ from 2000 to 1250 we observe that the policy rolls back to a very restricted setting (P0 or P1) for longer periods of time, indicating that the exponential nature of the epidemic requires some rather stringent constraints especially early on to get to a more controlled state of the epidemic. Furthermore, all results suggest that the primary effect of the restrictions is to broaden the peak until the epidemic has run its course and even the most relaxed policy shows an steep decline in cases. Interestingly, as illustrated in Figure 4(a), the optimal policy at $\tau = 2000$ achieves a significant reduction in the total number of infections (544K) when compared to the $\tau = \infty$ case (630K) for $\approx 15\%$ drop in the total reward. This clearly shows the efficacy of our look-ahead optimization in inferring non-trivial policies, while taking into account the complex interplay between the different NPI choices.

Impact of TransProp: It is well-known to epidemiologists that the probability of transmission between individuals is a critical parameter in controlling the trajectory of infections. Hence, we studied the impact of this parameter on the policies inferred using our algorithm. For this purpose, we varied the TransProp parameter between 0.1 and 0.22 while fixing the rest at the settings specified earlier. Note that, for this analysis ,we fixed the threshold $\tau = 1600$. From the result in Figure 4(b), it is apparent that the efficacy of the policy becomes increasingly inferior as TransProp grows – increase in the total number of cases as well as a significant reduction in the total reward.

Effect of Compliance: Another important aspect of mitigating epidemics such as COVID-19 is the public compliance to the mandates enforced and EpiCast includes that as one of its input parameters. We studied its effect on the policies inferred by varying compliance between 0.5 and 0.95 (higher value implies more compliant), while fixing the rest of the parameters as specified earlier. Similar to the previous experiment, we fixed $\tau = 1600$. As showed in Figure 4(c), improved compliance does lead to better policies both in reduced total number of infections and increased reward. However, given the relatively high transmission probability of 0.2, there is a limit beyond which the reward could not be increased even with a high compliance.

(ii) Policies for industry closure. In this experiment, we considered NPI choices pertinent to business opening and followed the protocol from the school opening study for running the policy optimization. Given $\mathcal{N}$ =[P0,P1v,P1r,P1q,P2v,P2r,P2q] and the corresponding reward assignments $\mathcal{C} = [1, 3, 6, 8, 12, 15, 18]$, we obtained the policies in Figure 5. First, we find that for the same initial settings as before, the trajectory for the infection growth is less severe when only businesses are open (schools closed) when compared to the case where schools are also open. Even when we aggressively reduced the threshold τ to 700, we could find an optimal policy, which however required a brief period of returning to phase 0 (total closure).

6. Conclusions

In this paper, we proposed an epidemic mitigation policy optimization algorithm that systematically accounts for short- and long-term effects of non-pharmaceutical interventions, via sophisticated epidemiological models. Our approach provides a principled, yet scalable, way to navigate through the combinatorial choices of interventions. Using case studies with SEIR and agent-based EpiCast models, we demonstrated the efficacy of our greedy algorithm in determining non-trivial policies (high reward) that satisfy the constraint on the number of infections. This can be used by policy makers and epidemiologists to gain critical insights into the interplay between different NPIs and current state of the epidemic.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

References

Alvarez, F. E., Argente, D., and Lippi, F. A simple planning problem for covid-19 lockdown. Technical report, National Bureau of Economic Research, 2020.

Anirudh, R., Thiagarajan, J. J., Bremer, P.-T., and Spears, B. K. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceedings of the National Academy of Sciences*, 117(18):9741–9746, 2020.

Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., et al. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. *Imperial College London*, 10:77482, 2020.

Germann, T. C., Kadau, K., Longini, I. M., and Macken, C. A. Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006. doi: 10.1073/pnas.0601266103. URL https://www.pnas.org/content/103/15/5935.

Germann, T. C., Gao, H., Gambhir, M., Plummer, A., Biggerstaff, M., Reed, C., and Uzicanin, A. School dismissal as a pandemic influenza response: When, where and for how long? *Epidemics*, 28:100348, 2019. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2019.100348. URL http://www.sciencedirect.com/science/article/pii/S1755436518301749.

Ghamizi, S., Rwemalika, R., Cordy, M., Veiber, L., Bissyandé, T. F., Papadakis, M., Klein, J., and Le Traon, Y. Data-driven simulation and optimization for covid-19 exit strategies. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3434–3442, 2020.

Guerrieri, V., Lorenzoni, G., Straub, L., and Werning, I. Macroeconomic implications of covid-19: Can negative supply shocks cause demand shortages? Technical report, National Bureau of Economic Research, 2020.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini, I. M., Cummings, D. A. T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C. A., Burke, D. S., and Cooley, P. Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, 105 (12):4639–4644, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0706849105. URL https://www.pnas.org/content/105/12/4639.

He, S., Peng, Y., and Sun, K. Seir modeling of the covid-19 and its dynamics. *Nonlinear Dynamics*, pp. 1–14, 2020.

Hethcote, H. W. and Van den Driessche, P. Some epidemiological models with nonlinear incidence. *Journal of Mathematical Biology*, 29(3):271–287, 1991.

Khadilkar, H., Ganu, T., and Seetharam, D. 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*, pp. 1–4, 2020.

Koziel, S. and Leifsson, L. *Surrogate-based modeling and optimization*. Springer, 2013.

Libin, P., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P., and Nowé, A. Deep reinforcement learning for large-scale epidemic control. *arXiv preprint arXiv:2003.13676*, 2020.

Liu, C. A microscopic epidemic model and pandemic prediction using multi-agent reinforcement learning. *arXiv preprint arXiv:2004.12959*, 2020.

López, L. and Rodo, X. A modified seir model to predict the covid-19 outbreak in spain and italy: simulating control scenarios and multi-scale epidemics. *Available at SSRN 3576802*, 2020.

Lurie, N., Saville, M., Hatchett, R., and Halton, J. Developing covid-19 vaccines at pandemic speed. *New England Journal of Medicine*, 382(21):1969–1973, 2020.

Mills, C. E., Robins, J. M., and Lipsitch, M. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

Morato, M. M., Bastos, S. B., Cajueiro, D. O., and Normey-Rico, J. E. An optimal predictive control strategy for covid-19 (sars-cov-2) social distancing policies in brazil. *arXiv preprint arXiv:2005.10797*, 2020.

Roda, W. C., Varughese, M. B., Han, D., and Li, M. Y. Why is it difficult to accurately predict the covid-19 epidemic? *Infectious Disease Modelling*, 2020.

Soures, N., Chambers, D., Carmichael, Z., Daram, A., Shah, D. P., Clark, K., Potter, L., and Kudithipudi, D. Sirnet: Understanding social distancing measures with hybrid neural network model for covid-19 infectious spread. *arXiv preprint arXiv:2004.10376*, 2020.

The United States Government. Opening Up America Again. https://www.whitehouse.gov/openingamerica/, 2020.

Yaesoubi, R. and Cohen, T. Identifying cost-effective dynamic policies to control epidemics. *Statistics in medicine*, 35(28):5189–5209, 2016.

Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3):165, 2020.