Using Reinforcement Learning for Multi-Objective Cluster-Level NPI Optimization

Xueqiao Peng  
Xueqiao Peng  
The Ohio State University  
Columbus, Ohio, USA

Jiaqi Xu  
Jiaqi Xu  
The Ohio State University  
Columbus, Ohio, USA

Xi Chen  
Xi Chen  
The Ohio State University  
Columbus, Ohio, USA

Dinh Song An Nguyen  
Dinh Song An Nguyen  
The Ohio State University  
Columbus, Ohio, USA

Andrew Perrault  
Andrew Perrault  
The Ohio State University  
Columbus, Ohio, USA

ABSTRACT
Non-pharmaceutical interventions (NPIs) play a critical role in the defense against emerging pathogens. Among these interventions, familiar measures such as travel bans, event cancellations, social distancing, curfews, and lockdowns have become integral components of our response strategy. Contact tracing is especially widely adopted. However, the optimization of contact tracing involves navigating various trade-offs, including the simultaneous goals of minimizing virus transmission and reducing costs. Reinforcement learning (RL) techniques provide a promising avenue to model intricate decision-making processes and optimize policies to achieve specific objectives, but even modern deep RL techniques struggle in the high dimensional partially observable problem setting presented by contact tracing. We propose a novel RL approach to optimize a multi-objective infectious disease control policy that combines supervised learning with RL, allowing us to capitalize on the strengths of both techniques. Through extensive experimentation and evaluation, we show that our optimized policy surpasses the performance of five benchmark policies.

KEYWORDS
reinforcement learning, machine learning, contact tracing, public health

ACM Reference Format:

1 INTRODUCTION
The COVID-19 pandemic has highlighted the crucial role of non-pharmaceutical interventions (NPIs) in effectively managing the spread of infectious diseases. The implementation of NPIs requires careful consideration of multiple objectives, including the prevention of viral transmission and the reduction of costs associated with quarantine measures. Contact tracing has emerged as a widely adopted policy within the realm of NPIs and has been extensively studied in the context of COVID-19 [7, 8, 11, 21].

Nevertheless, optimizing NPIs remains a challenging open problem in many settings for several reasons. First, the objective is inherently multi-objective—intensified control efforts lead to higher costs. In addition, sensing actions, such as testing, may be included in all but the earliest stages of an infectious disease crisis. These have their own costs and constraints associated with them. Secondly, inferring the probability that an individual is difficult for infections that do substantial transmission asymptomatically, such as SARS-CoV-2. This inference problem is perhaps surprisingly high dimensional, as we show it is dependent on the symptom status and test results of all individuals in the same cluster due to the transmission heterogeneity.

![Figure 1: Illustration of our approach. We combine a infection probability decoder that uses supervised learning with a reinforcement learning-based policy.](image)

In this work, our goal is to develop a generic approach for cluster-level optimization of NPIs. To tackle this challenge, we propose a novel approach that integrates convolutional neural networks (CNN) and reinforcement learning (RL) model[5, 20] (Fig. 1). The CNN is used to solve the high dimensional infection inference problem and uses a novel representation of the symptom and test state of the entire cluster as input, allowing a single CNN to be trained for all cluster sizes. The RL agent takes the CNN output and other features as its state and selects an action for each individual (including quarantine and testing) and aims to maximize a multi-objective reward function. This reward function includes a penalty.
for days where an individual is infectious but not isolated, a penalty for days where they are quarantined but not infectious, as well as a cost for any control action that is taken (e.g., test cost). As a case study, we have developed a branching process-based SARS-CoV-2 virus simulator, where we evaluate the effectiveness of our method. In this work, we focus on optimization only—in the longer term, we aim to use the results of optimization to automatically discover simple, implementable policies.

This paper makes the following contributions:

- We propose a novel RL approach for finding optimal contact tracing policies. Our approach combines a supervised learning model with an RL model, leveraging the strengths of both techniques to optimize the desired objectives. The resulting agent can be trained and deployed simultaneously across all cluster sizes.
- We show the existence of a theoretically simple, yet optimal, threshold type policy for contact tracing in the setting where no sensing actions are available. Running this policy requires supervised learning only.
- We develop a simple branching process-based model for SARS-CoV-2 and compare our policies with baselines. We show that we achieve better rewards across a range of objective parameters.

Related work. We identify two main thrusts of work that optimize contact tracing and NPIs: network and branching process. Network models represent connections between individuals as edges in a possibly dynamic contact graph [4, 9, 12, 15, 16]. These approaches can leverage network structure in their decisions but make the strong assumption that the entire contact network is known. The closest existing approach to ours is RLGN [12], which formulates the problem as a sequential decision-making task within a temporal graph process. These approaches often consider a fixed budget of interventions rather than a multi-objective reward function. In contrast, branching processes are used, resulting in a cluster-based, tree-structured view of contagion [10, 13, 17]. These approaches have the advantage of aligning more closely with the information available to public health decision-makers in many practical settings (but allow for less expressive policies). All of these models are agent-based in the sense that they model individuals rather than subpopulations—because contact tracing decisions depend on the specific time that certain events happen for individuals (e.g., exposure, symptoms), the additional detail that agent-based models provide is valuable for modeling and optimization.

2 \hspace{1cm} \textbf{BRANCHING PROCESS ENVIRONMENT}

We take a branching process-based view of an infectious disease crisis (Fig. 2). We track two generations of potential individuals: the seed case and their contacts. We assume that interventions begin after a reporting and tracing delay. At that point, day \( t_{\text{start}} \) (\( t_{\text{start}} = 3 \) in Fig. 2), we observe the symptom history for each agent up to day \( t \) and must decide which action to take for each agent (e.g., quarantine, test). On day \( t \), we observe the symptom state of each agent plus the results of any sensing actions (defined below) we have taken up to day \( t \) and must decide what action to take for each agent on day \( t \). The simulation proceeds for a fixed period of time until \( T \).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{An agent-based branching process model. The diagram depicts standard contact tracing for an example seed case with six contacts.}
\end{figure}

In Fig. 2, we present an application of contact tracing policy in the branching process framework. The seed case remains infectious for two days without exhibiting symptoms, followed by one day with symptoms, before entering isolation. In this example, all six contacts were exposed on the same day. Contacts 1 and 4 are infected and show symptoms on day 2 and day 3, respectively. All contacts are asked for quarantine if their infection probability is higher than a threshold. Contact 3 and contact 5 serve quarantine on day 3. Contact 2 and contact 6 start quarantining on day 4.

In an infectious disease crisis, we can use whatever data is available to construct such a branching process model. Many of the required components are distributions that are often estimated by epidemiologists in the early stages of an outbreak. We describe distributions we used to simulate SARS-CoV-2 and their sources in Tab. 1. Components that are not known can be filled in conservatively or sensitivity analysis can be performed. In some cases, distributional estimates can be shared across diseases—for example, POLYMOD [14] provides contact distributions for the US and Western European settings for both droplet and physical contact. The superspreading dynamics of infection can be impactful because it is often that most transmission is driven by a small number of seed cases, and this concentration can be exploited by control policies [17]. Nevertheless, superspreading dynamics are often poorly understood, especially early in a crisis and greater understanding would benefit approaches such as this paper’s.

We define the objective function as

\[
\text{Obj} = \frac{\left( -S_1 - \alpha_2 \times S_2 - \alpha_3 \times S_3 \right)}{\text{cluster size}} \tag{1}
\]

where

- \( S_1 \) is the count of transmission days where an infected individual is not quarantined,
- \( S_2 \) is the count of days where a quarantined individual is not infected, and \( \alpha_2 \) (which we assume is in \([0, 1]\)) is the weight for this term,
- \( S_3 \) is the sum of the action costs (e.g., test cost) and \( \alpha_3 \) is the weight for this term, and
- \text{cluster size} normalizes the objectives to a score per individual.

In summary, the objective function seeks to minimize the number of transmission days (i.e., days where an individual is infectious...
We show that the optimization problem from the previous section can be formulated as a partially observable Markov decision process (POMDP). However, solving this POMDP directly is wildly intractable. Some hope arrives from the result that, under a simplified model that contains only sensing-type actions, the POMDP can be solved optimally if the probability that an individual is infectious can be estimated—itself a challenging problem due to the high computational complexity of the best-known algorithms.

Motivated by this conclusion, we formulate our solution approach: we use a convolutional neural network (CNN) to estimate the probability of infectiousness for each individual in a cluster, and this output, along with cluster-wide statistics, serves as the state for the RL agent.

### Table 1: Parameters of the SARS-CoV-2 branching process model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Assumed value</th>
<th>Details and references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incubation time</td>
<td>Log-normal: Log mean 1.57 days and log std 0.65 days</td>
<td>Mean: 5.94 days. Bi et al. [2]</td>
</tr>
<tr>
<td>Duration of infectious period</td>
<td>5 days after onset if symptomatic</td>
<td>Bi et al. [2]</td>
</tr>
<tr>
<td>Probability that an infected individual shows symptoms</td>
<td>0.8</td>
<td>Buitrago-Garcia et al. [3]</td>
</tr>
<tr>
<td>Probability of symptoms without infectiousness</td>
<td>0.01 per day</td>
<td>Perrault et al. [17]</td>
</tr>
<tr>
<td>Probability of asymptomatic infection</td>
<td>0.2</td>
<td>Buitrago-Garcia et al. [3]</td>
</tr>
<tr>
<td>Probability of highly transmissible infectiousness multiplier for highly transmissible individuals</td>
<td>24.4</td>
<td>Perrault et al. [17]</td>
</tr>
<tr>
<td>Test parameters</td>
<td>TP = 0.86, FP = 0.66</td>
<td>Besutti et al. [1]</td>
</tr>
<tr>
<td>TN = 0.14, FN = 0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delays</td>
<td>Observation Delay = 3 days</td>
<td>Assumed</td>
</tr>
<tr>
<td></td>
<td>Test Result Delay = 1 day</td>
<td></td>
</tr>
</tbody>
</table>

3.1 POMDP Formulation

We define a POMDP [6] as \( \langle S, A, R, P, \Omega, O, y, S_0 \rangle \), where \( S \) and \( A \) represent the state and action spaces, respectively, \( R : S \times A \to \mathbb{R} \) is the reward function, \( P : S \times A \to \Delta S \) is the transition function, \( \Omega \) is the observation state, \( O : S \times A \to \Delta \Omega \) is the observation probabilities, \( y \in [0, 1] \) is the discount factor, and \( S_0 : \Delta S \) is the distribution of initial states.

We briefly describe how to interpret the control problem of the previous section as a POMDP. We define the state space as containing all of the relevant information required to simulate the cluster, including whether the seed case is highly transmissible, whether each contact of a seed case will become infected, whether they will show symptoms and if so, on what day. This simulator data cannot be observed directly—instead we must rely on receiving action-dependent observations. We define the action space as the set of daily quarantine and sensing actions that are available for each individual in the cluster. For instance, in our experiments, we consider five actions: no quarantine and no test, quarantine and no test, and no quarantine, test and quarantine, and test and quarantine only if positive. If we have \( N \) individuals in the cluster, we have an action space of size \( |A|^N \). For observations, we receive two types of information from each individual in each timestep: symptom information and test results. We receive test results only when a sensing-type action is taken and these results are noisy (Tab 1). Similarly, we always observe symptoms if they are present, but both infectiousness without symptoms and symptoms without infectiousness are possible. The resulting observation space size is \( 4^N \).

In principle, solving the POMDP formulation results in the optimal control policy. In practice, exact solving is not possible due to the high computational complexity of the best-known algorithms. A particular source of difficulty is the problem of calculating the posterior probability of infection for each individual given the observations. A key challenge is that the variation in infectiousness of the seed case causes the posterior probability of infection for each
We could use RL directly to solve the POMDP using the observation information as the state. Indeed, we show that this is somewhat effective if we leverage the state representation we develop in the next section. However, as we know the unobserved infectious state for each agent in simulation, we hypothesize that using a supervised learning model to predict $p_{\text{inf}}$ and using this as input to the RL algorithm will lead to better objective values compared to pure RL (and in the experiments, we see that the improvement is often substantial). Another option for estimating $p_{\text{inf}}$ would be to use an algorithm for approximate probabilistic inference such as Markov chain Monte Carlo, but doing so is challenging due to the high dimensional discrete observation space where most observations have zero probability for a given state of infectiousness.

A key question for applying supervised learning is how to represent the observation space. We have two desiderata. First, we would like the representation to not vary with cluster size. We can also achieve this property in the RL agent, resulting in an agent that simultaneously be deployed across all cluster sizes, which makes both training and deployment simpler. Second, there is an advantage to using a representation that inherently accounts for the symmetries that arise due to the ordering of individuals. i.e., if we permute the order of individuals in an observation, it should not affect $p_{\text{inf}}$ for each individual.

After testing several representations that satisfy these properties, we arrive at the $7 \times T$ matrix shown in Fig. 3, where $T$ is the simulation length (in our experiments, $T = 30$). This is an egocentric representation of the observation—it is from the perspective of a particular contact and contains all information gathered so far. We train the supervised learning model $f$ to produce output dimension $[0, 1]^T$, i.e., for every day of the simulation, what is the probability that the agent will be infectious given the observation using simulation outputs where the infectiousness of each individual is provided.

The representation contains the following information. The first row is 1 for each day after (inclusive) that the individual shows symptoms. The second row is a binary indicator of whether this particular contact and contains all information gathered so far. We can also use the above proof technique to derive the optimal policy for any menu of non-sensing actions. A useful generalization is when the quarantine action has a cost and a failure rate.

### 3.2 Optimal Policy Without Sensing Actions

We first consider a simplified POMDP where the only actions available are a quarantine action and no quarantine action. We show that, if the posterior probability of infection can be calculated exactly, the optimal policy has a threshold-type form: if the posterior probability of infection is above a threshold, quarantine and otherwise do not. We show this initially for a costless quarantine action with 100% efficiency as this is what we use in experiments (Thm. 1). We then generalize the result to any menu of non-sensing actions because the expected reward of each action can be exactly calculated given the posterior probability of infection (Thm. 2). We remark that these results provide additional context to the findings of Perrault et al. [17] by defining the class of optimal risk-based policies.

Let $p_{\text{inf}}$ represent the posterior probability of infection for an individual given the observations so far.

**Theorem 1.** With a costless quarantine action that is always successful and a null action, the objective function of Eq. 1, the optimal policy is to quarantine if $p_{\text{inf}} > \frac{\alpha_2}{1 + \alpha_2}$ and take the null action otherwise.

**Proof.** Because we have access to the exact posterior probability of infection, we can calculate the expected objective value for each action exactly:

$$E[r] = \begin{cases} -\alpha_2 \cdot (1 - p_{\text{inf}}) & \text{if quarantined} \\ -p_{\text{inf}} & \text{if not quarantined.} \end{cases}$$

(2)

We can then show that if $p_{\text{inf}} > \frac{\alpha_2}{1 + \alpha_2}$, the quarantine action has higher expected reward. □

We can use the above proof technique to derive the optimal policy for any menu of non-sensing actions. A useful generalization is when the quarantine action has a cost and a failure rate.

**Theorem 2.** With a quarantine action with success rate $0 \leq \beta \leq 1$ and cost 1 and a null action, the optimal policy is to quarantine if $p_{\text{inf}} > \frac{\alpha_2 - \beta \cdot \alpha_1}{(1 + \alpha_2) \cdot \beta}$ and otherwise do not.

These results highlight the importance of the posterior probability of infection. We next dedicate our attention to producing useful estimates of $p_{\text{inf}}$.

### 3.3 Supervised Learning

We could use RL directly to solve the POMDP using the observation information as the state. Indeed, we show that this is somewhat effective if we leverage the state representation we develop in the next section. However, as we know the unobserved infectious state for each agent in simulation, we hypothesize that using a supervised learning model to predict $p_{\text{inf}}$ and using this as input to the RL algorithm will lead to better objective values compared to pure RL (and in the experiments, we see that the improvement is often substantial). Another option for estimating $p_{\text{inf}}$ would be to use an algorithm for approximate probabilistic inference such as Markov chain Monte Carlo.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\sim$</th>
<th>$\sim$</th>
<th>$\sim$</th>
<th>$\sim$</th>
<th>$\sim$</th>
<th>$\sim$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3:** The observation representation used for supervised learning, shown on a cluster of size 10 after observing the outcome of day 2.
3.4 Reinforcement Learning

To make RL effective, we develop a compact state representation that includes supervised learning outputs. As with supervised learning, we want the representation to have the same size for all clusters and to naturally encode permutation invariance. The representation we use is a 7 × 3 matrix shown in Fig. 4. As with the supervised learning representation, it is egocentric and time-specific.

The first and second rows represent the $p_{inf}$ outputs from supervised learning for the last three days and next three days, respectively. The third row indicates whether the individual exhibited symptoms for each day in the past three days. The fourth row is an indicator for if this individual was tested for each of the past three days. The fifth row denotes the test results with a one day delay. The sixth row is the cluster size. The last row indicates the number of tests conducted in the cluster in the past three days.

Training the RL algorithm is straightforward. First, we train the supervised learning predictor from data collected from the simulator. In our experiments, we use a fixed but stochastic control policy to collect this data. This has the advantage that a single supervised learning training run can serve as input to an arbitrary number of RL training. If the optimal policies are dramatically different than the data collection policy, an addition run of supervised learning can be performed with the current RL policy to increase its accuracy.

Once the supervised learning predictor is trained, we train RL with Proximal Policy Optimization (PPO) [19]. In our experiments, we use six different policy initializations, train each for 800000 environment interactions and pick the best based on 100 evaluation runs. All training is performed on a single core, using Intel i5-8259U @2.3GHz with 8GB of RAM, and a single RL training run takes 20 minutes.

4 EXPERIMENTS

We compare different control policies in the branching process environment we construct for SARS-CoV-2. We consider a set of five control actions for each individual for each day: null action, quarantine, test but don’t quarantine, quarantine but don’t test, and test and quarantine only if results are positive. We assume that there is no failure rate for actions, and all actions that include a test cost 1 and others are costless. For $\alpha_2$, we use small values of 0.01 and 0.02 as typical SARS-CoV-2 contact tracing policies accept a large number of quarantine days for non-infectious individuals. For $\alpha_3$, we use values of 0.001, 0.005, 0.01, 0.02, 0.03 and 0.2. We sample cluster size from a uniform distribution on (2, 40). The model code is available online (https://github.com/XueqiaoPeng/CovidRL).

4.1 Supervised Learning Model

We experiment with a variety of supervised learning model architectures (Tab. 2) to find one that achieves a high AUC across cluster sizes. We find that CNNs are generally most effective and compare different kernels and layer structures. In single layer architectures, we find that larger 2D convolutions tend to achieve higher AUC. We then found that a single convolution layer followed by a linear layer performs just as well as deeper architectures—this setup of a (5, 2) 2D convolution followed by a linear layer is what we use in the experiments below.

Table 2: We find that two-layer architectures using a 2D convolution followed by a linear layer achieve performance on par with larger models.

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv1d (5,2)</td>
<td>0.798</td>
<td>0.807</td>
<td>0.823</td>
<td>0.830</td>
</tr>
<tr>
<td>Conv1d (5,3)</td>
<td>0.814</td>
<td>0.830</td>
<td>0.835</td>
<td>0.839</td>
</tr>
<tr>
<td>Conv2d (5,2)</td>
<td>0.800</td>
<td>0.814</td>
<td>0.827</td>
<td>0.830</td>
</tr>
<tr>
<td>Conv2d (5,3)</td>
<td>0.832</td>
<td>0.820</td>
<td>0.838</td>
<td>0.840</td>
</tr>
<tr>
<td>Conv2d (5,4)</td>
<td>0.858</td>
<td>0.849</td>
<td>0.843</td>
<td>0.859</td>
</tr>
<tr>
<td>Conv2d (5,5)</td>
<td>0.864</td>
<td>0.895</td>
<td>0.893</td>
<td>0.893</td>
</tr>
<tr>
<td>2 Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv1d (5,2)</td>
<td>0.824</td>
<td>0.830</td>
<td>0.833</td>
<td>0.840</td>
</tr>
<tr>
<td>Conv1d (5,3)</td>
<td>0.883</td>
<td>0.903</td>
<td>0.898</td>
<td>0.897</td>
</tr>
<tr>
<td>Conv2d (5,2)</td>
<td>0.955</td>
<td>0.960</td>
<td>0.947</td>
<td>0.961</td>
</tr>
<tr>
<td>Conv2d (5,3)</td>
<td>0.951</td>
<td>0.960</td>
<td>0.940</td>
<td>0.964</td>
</tr>
<tr>
<td>3 Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv1d (5,3)</td>
<td>0.958</td>
<td>0.957</td>
<td>0.950</td>
<td>0.961</td>
</tr>
<tr>
<td>4 Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv1d (4,3)</td>
<td>0.958</td>
<td>0.958</td>
<td>0.953</td>
<td>0.965</td>
</tr>
</tbody>
</table>

4.2 Benchmark Policies

We compare the RLSL approach we propose to several baselines.

- **Threshold** is the threshold-type policy suggested in Sec. 3.2. It does not use test actions. This policy turns out to be highly conservative and results in long quarantine duration for all contacts for the tested $\alpha_2$ values.
### Table 3: RSL achieves higher objective values (higher is better) than baselines across all tested $a_2$ and $a_3$.

<table>
<thead>
<tr>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_3 = 0.01$</td>
<td>$a_3 = 0.01$</td>
<td>$a_3 = 0.005$</td>
<td>$a_3 = 0.01$</td>
<td>$a_3 = 0.005$</td>
<td>$a_3 = 0.01$</td>
<td>$a_3 = 0.005$</td>
<td>$a_2 = 0.01$</td>
<td>$a_2 = 0.01$</td>
<td>$a_2 = 0.01$</td>
<td>$a_2 = 0.01$</td>
<td>$a_2 = 0.01$</td>
</tr>
<tr>
<td>RSL (Ones)</td>
<td>-3.77 ± 0.25</td>
<td>-6.08 ± 0.18</td>
<td>-10.27 ± 0.15</td>
<td>-30.13 ± 0.48</td>
<td>-44.22 ± 0.84</td>
<td>-46.60 ± 1.47</td>
<td>-110.92 ± 1.54</td>
<td>-4.01 ± 0.21</td>
<td>-17.64 ± 0.32</td>
<td>-25.39 ± 0.48</td>
<td>-49.26 ± 0.86</td>
</tr>
<tr>
<td>Threshold</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
<td>-21.79 ± 0.20</td>
</tr>
<tr>
<td>Quarantine</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
<td>-97.18 ± 0.97</td>
</tr>
<tr>
<td>14 Days</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
<td>-258.98 ± 18.53</td>
</tr>
<tr>
<td>No Quarantine</td>
<td>1.04 ± 0.17</td>
<td>2.60 ± 0.11</td>
<td>-0.95 ± 0.02</td>
<td>2.44 ± 0.14</td>
<td>-0.97 ± 0.11</td>
<td>2.29 ± 0.12</td>
<td>-0.99 ± 0.10</td>
<td>2.00 ± 0.15</td>
<td>-0.97 ± 0.11</td>
<td>2.29 ± 0.12</td>
<td>-0.99 ± 0.10</td>
</tr>
<tr>
<td>$S_1$</td>
<td>-2.597 ± 0.282</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$S_2$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$S_3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 4: $S_1$, $S_2$ and $S_3$ per individual compared across different cluster sizes (lower is better), using $a_2 = 0.01$ and $a_3 = 0.01$. Even relatively conservative strategies such as 14-day quarantine from exposure fail to isolate some infections in our simulation. RSL can benefit substantially from the additional information available in large clusters resulting in strong performance with low test costs.

### Table 5: In cases where test costs are higher, RSL produces polices that test too often, resulting in lower performance than RSL models with only quarantine actions—we discuss potential fixes.

**4.3 Analysis**

Our experimental results report the average objective value and standard error taken over 10 random clusters (Tab. 3). We find that RSL and Threshold achieve better performance than baselines in all cases. However, our current methods for RSL struggle relative to Threshold when tests are expensive. Our experimental results could be broadened by including more $a_3$ values and more analysis as to where the RSL policies gain their advantage (but see discussion of Tab. 5 below for some insights).

Focusing on the setting of $a_1 = 0.01$ and $a_2 = 0.01$, we report objective values broken out by component and by cluster size as measured per individual (Tab. 4). Here we can get an intuitive grasp of what is happening in the different policies. Threshold aggressively quarantines, resulting in $S_2 \approx 16 - 20$, i.e., 16–20 days of quarantine without infection per contact, for the tested $a_3$ values. This is able to drive $S_1$ to a low value, resulting in an average objective value of $\approx -21.79$. Recall that $S_1$ is much more highly (100 times) higher than $S_2$ in this setting. Symptom-based and 14-day quarantine reduce $S_2$ by a factor of 8 to 100, but this causes $S_1$ to be roughly 150 to 200 times higher. By leveraging tests, RSL can reduce $S_2$ by a factor of 2–3 and $S_1$ by a factor of 0.8–3.5.

In the ablation study (Tab. 5), we gain a more detailed view into the operation of the RSL policy. We see that the introduction of the SL outputs to the RL state results in better performance in all tested scenarios compared to RL Only, which uses the state representation of Fig. 4 without the first two rows.

We can observe limitations of the supervised infectiousness prediction model in Tab. 4, where the $S_2$ cost does not decrease as cluster size increases—from Thm. 1, we can conclude that if $P_{inf}$ is correct, the ratio of $S_1$ to $S_2$ should not depend on cluster size for Threshold. There are several possible causes of this issue. First, the SL model outputs might be misclassified, as is often the case for neural networks trained on highly imbalanced data. This issue could be fixed with post-hoc calibration such as Platt scaling [18]. In this instance, a more sophisticated calibration could be employed with separate calibration parameters per cluster size, if necessary. Second, it may be the case that the SL model outputs are wrong for reasons other than calibration. For example, it may receive insufficient relevant training data as it is trained on data produced from a random policy and not Threshold or RSL. It is also possible that we performed insufficient architecture search.

We also see that RSL (No Test) often performs better than RSL as test costs increase. This suggests that RSL is not finding a true
optimal policy. This could likely be address by using a wider range of initialization values for RSL— for example, initializing some seeds to policies that test very little (the initialization we use for RSL and RL Only tests heavily). This observation has a silver lining: RL (No Test) can achieve much stronger performance than baselines even without tests. This implies that RL (No Test) is able to correct for the errors in Threshold to find a policy closer to what is suggested by Thm. 1.

5 DISCUSSION AND FUTURE WORK

This work aims to develop a generic multi-objective optimization approach for cluster-level optimization of NPIs. We formulate this problem for RL in a branching process environment. We present initial results that demonstrate the potential of our approach—in a branching process model of SARS-CoV-2, we can achieve substantially higher objective values than baseline policies. The resulting policies can be applied across all cluster sizes and do not take much time to train on consumer hardware. The policies we propose are able to heavily exploit superspreading dynamics.

Our vision for an infectious disease crisis is that a canonical probabilistic model of the disease is constructed and updated throughout the crisis. The model can be constructed from estimates of key disease parameters that are made from various sources throughout a crisis and can reflect uncertainty in these estimates. We advocate that superspreading dynamics be given substantial attention in these early stages due to the substantial influence on interventions that we find it can have. Using this canonical model, a branching process environment can be constructed and optimized against as we propose in this paper. We do not consider uncertainty in the parameters of this model, but it is possible to do so with existing techniques and leads to different RL algorithmic choices depending on the form of the uncertainty and the desired objective.

A key disadvantage of our approach as presented is the complexity of the resulting policies. For instance, to execute our RLSL policy requires training and drawing outputs from two neural networks. In contrast, policies that were employed in the SARS-CoV-2 pandemic consisted of short lists of rules. We believe that this is not an inherent weakness of our approach—we can leverage interpretable ML and RL techniques to “distill” the RLSL policies into, say, low-depth decision trees, allowing them to be applied at scale with low logistical cost. There will be some decrease in quality, but we suspect still substantial advantage over baselines.

An area for future study is cost and benefit of taking a cluster- rather than individual-level view of policy application. This imposes additional logistical costs and the benefit is dependent on the degree of cluster-level transmission heterogeneity that is present. This trade-off is not well understood and is a critical area for future work.

REFERENCES


