European Journal of Operational Research 304 (2023) 325–338

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastic optimization for vaccine and testing kit allocation for the COVID-19 pandemic[☆]

Lawrence Thul[a,*], Warren Powell[b]

[a] Department of Electrical Engineering, Princeton University, Princeton, NJ, USA
[b] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA

ARTICLE INFO

Article history:
Received 5 January 2021
Accepted 4 November 2021
Available online 11 November 2021

Keywords:
Decision processes
Uncertainty modeling
Reinforcement learning

ABSTRACT

We present a formal mathematical modeling framework for a multi-agent sequential decision problem during an epidemic. The problem is formulated as a collaboration between a vaccination agent and learning agent to allocate stockpiles of vaccines and tests to a set of zones under various types of uncertainty. The model is able to capture passive information processes and maintain beliefs over the uncertain state of the world. We designed a parameterized direct lookahead approximation which is robust and scalable under different scenarios, resource scarcity, and beliefs about the environment. We design a test allocation policy designed to capture the value of information and demonstrate that it outperforms other learning policies when there is an extreme shortage of resources (information is scarce). We simulate the model with two scenarios including a resource allocation problem to each state in the United States and another for the nursing homes in Nevada. The US example demonstrates the scalability of the model and the nursing home example demonstrates the robustness under extreme resource shortages.

1. Introduction

During the early months of 2020, it became evident that the SARS-CoV-2 virus was spreading through the global population at an alarming rate. The mitigation strategies in place were not sufficient to handle a crisis at this scale, devastating global economies and supply chains. After the tragic losses to life and economic damage suffered, it is imperative to reflect on the nature of the problem which was faced and how to act differently in the future.

The greatest challenge decision-makers face at the onset of an epidemic is the huge set of unknowns. There is uncertainty about the features of the disease, such as transmission rates, recovery rates, and death rates. There is uncertainty about the dynamics of the disease, such as exposure time to infection, reinfection rates, or asymptotic spreading. Once personal protective equipment is available, there is uncertainty about the effectiveness and public use. Once testing kits are available, there is uncertainty about testing accuracy and infectivity measurements in the population. Once vaccines are available, there is uncertainty about efficacy rates and public confidence. As the resources available to fight the disease are manufactured, there is uncertainty about the production rates. In the face of all the unknowns, decision-makers must act swiftly and strategically to mitigate the spread of the disease in a crucial period of time.

The epidemic problem setting has enumerable complexities associated with it. In this paper, we will focus on a subset of the problems faced by decision-makers. Specifically, we will focus on the problem of allocating vaccines throughout a region when the state of the epidemic is not known perfectly to the decision-maker. We assume that the sequential decision problem begins at the onset of vaccine production, so there will be extreme shortages of vaccines which will rollout as they are manufactured. Additionally, a limited stockpile of testing kits are also produced which implies there are a limited number of observations available to the vaccine distributors. Hence, the decision-maker must capture how valuable the observations are with respect to learning the true state of the epidemic in local zones. Fig. 1 illustrates the problem of allocating stockpiles of vaccines and testing kits to zones.

During the SARS-CoV-2 epidemic, the initial vaccine distribution strategy was to allocate vaccines proportional to the number of adults in each state as soon as they became available (Simunaci, 2020). In this paper, we design a policy using a parameterized rolling horizon stochastic optimization technique and compare it to other classes of policies. The formulation of a proper model to design allocation policies which can adapt to the non-stationary stream of data allows for more robust management of resources. In reality, there are different goals for allocating testing kits and

[☆] This work is supported by the Air Force Office of Scientific Research (Award Number FA9550-19-1-0203).

[*] Corresponding author.
E-mail address: lathul@princeton.edu (L. Thul).

https://doi.org/10.1016/j.ejor.2021.11.007
0377-2217/© 2021 Elsevier B.V. All rights reserved.

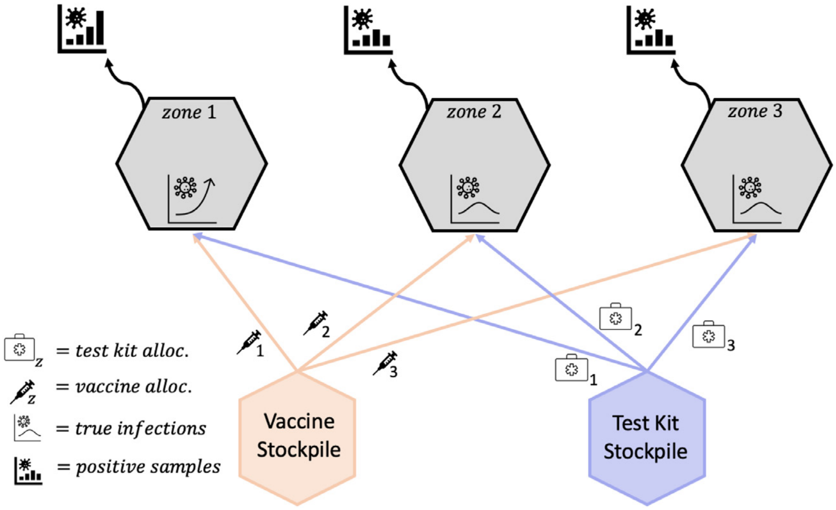


Fig. 1. Illustration for the vaccine and testing kit allocation problem to multiple zones. An allocation of vaccines and testing kits is distributed from stockpiles and sent to a given zone. Each zone also demonstrations positive observations of an underlying dynamic disease process.

vaccines, but when these problems are considered jointly the limited resources can be used more effectively. It is uncommon in the broad literature to find a multi-agent problem where there are agents which can change the state of the environment and agents which learn about the environment considered jointly.

There are many modeling and algorithmic challenges presented in this application setting. The region is partitioned into a set of zones and each zone will get individual allocations of vaccines and testing kits. This leads to the set of possible decisions becoming very high dimensional. Each zone also has sets of individuals in different states related to the epidemic. For example, a percentage of the population is infected with the disease, a percentage is susceptible to the disease, and a percentage is vaccinated or immune to the disease. From the decision-makers perspective, the true state of the infection within the population is not known perfectly, so probability distributions must be maintained as information is processed over time. This leads to state spaces over parameters of probability distributions, which grow very large and difficult to handle. The set of possible observations from each zone is a function of the number of tests allocated to it, so the observation spaces become very high dimensional as the number of zones increase. The high dimensionality of multiple aspects of the problem will lead to a limited set of approaches we can consider from the existing literature.

There are many different agents allocating resources during a pandemic at all levels (federal, state, local) of government or organizations. The framework in this paper will develop a model with hyperparameters that can be adjusted to simulate the scenario for the decision-maker. This paper will consider two scenarios to highlight the robustness and scalability of the framework. We simulate the federal allocations of vaccines and tests as separate agents in the federal government to each state. The scenario demonstrates the model and policies designed can scale to populations of hundreds of millions of people and millions of resources available. The second scenario models state-level agents allocating tests and vaccines to nursing homes in the state of Nevada. The scenario is designed to highlight the robustness of the framework to handle extreme resource shortages. Some local areas will not receive as many resources due to budgets at higher levels in the supply chain or worse outbreaks in other areas in the country. Therefore, it is imperative to ensure the framework is robust when the availability of resources is scarce. Hence, this scenario demonstrates the model is able to capture the value of information collected and the vaccine allocation policies can effectively adapt as new information streams in.

This paper makes the following contributions:

- We present the first formal multi-agent modeling extension to the unified framework for an epidemic application. We formulate a mathematical model for a multi-agent stochastic resource management problem that combines resource allocation (for the vaccines) with active learning (through testing). This model is able to capture passive information processes and perform active learning to improve the belief states by querying valuable observations.
- We propose a vaccine allocation policy which solves a parameterized direct lookahead model. The parameterization must be tuned using policy search. Furthermore, we demonstrate the necessary, but rare in the literature, search over policies across multiple communities of stochastic optimization. In our search, we tested all four classes of policies, but omitted policies with the worst performance due to space.
- We propose a test kit allocation policy by formulating a surrogate function and drawing from one-step lookahead acquisition functions from the Bayesian optimization literature. We demonstrate the utility of active learning through the test kit allocation policy when resources are extremely scarce.
- We demonstrate that under extreme resource shortages the proposed vaccination allocation and learning policies work best in conjunction compared to all other combinations of policies. The nursing home simulation highlights the power of using active learning to guide an implementation decision under resource scarcity.

The paper is organized as follows. Section 2 summarizes the literature about vaccine distribution strategies, stochastic optimization, and similar areas of research to this paper. Section 3 describes the multi-agent mathematical model using the unified framework. Section 3 is broken down into the environment agent model and controlling agent models. The controlling agent section presents the learning model, and vaccination model. Section 4 describes the formulation of policies for the vaccination agent and learning agent. Section 5 discusses the results of implementing the model on simulators designed for two different scenarios of the environment agent. Section 6 concludes and summarizes the results and contributions of the research.

2. Literature review

There have been various computational and mathematical strategies for simulation, forecasting, and control of epidemics. One of the most common ways to model a pandemic is to use compartmental models. Kermack & McKendrick (1927) creates the SIR model which is the most basic compartmental model consisting of three groups within a population: those susceptible (S) to the disease, those infected (I) with the disease, and those removed (R) from the population (from death, recovery, or immunity). Tang et al. (2020) reviews the literature about compartmental models and provides various extensions of the SIR model such as susceptible-exposed-infected-recovered (SEIR), spatial SIR models, spatiotemporal SIR models, and other possible multi-compartment extensions. Greenwood & Gordillo (2009) provides a review of the SIR model with stochastic transmission rates.

The literature regarding decision-making strategies to combat an epidemic is large and spans many disciplines. There are strategies regarding control via public policy, pharmaceutical or vaccine intervention. Köhler et al. (2020) and Morato, Pataro, da Costa, & Normey-Rico (2020) use public policy controls (e.g. social distancing/lockdowns) to mitigate the spread of infection when a vaccine is unavailable. Buhat, Duero, Felix, Rabajante, & Mamplata (2021) develops equitable testing kit allocation strategies to medical centers in the Philippines. Lin, Zhao, & Lev (2020) models a problem to decide whether a distributor will transport vaccines through a cold chain or a non-cold chain to ensure that they are still viable at administration. Ekici, Keskinocak, & Swann (2008) uses a complex spatial SEIR model in conjunction with an age-based component to decide who to feed during a flu pandemic in Georgia by setting up food distribution centers. Dai, Cho, & Zhang (2016) models an influenza supply chain in the U.S. consisting of healthcare providers, manufacturers, and distributors to ensure on-time delivery to each provider.

The allocation of vaccination and other pharmaceutical resources is the most effective method for fighting an epidemic. Duijzer, van Jaarsveld, & Dekker (2018) specifies a hybrid vaccine strategy for early intervention with low efficacy vaccines and later intervention with high efficacy vaccines. Some pharmaceutical and vaccine intervention strategies optimize the one-time allocation of resources at the beginning of an epidemic. Martin, Allen, Stamp, Jones, & Carpio (1993) applies rule-based vaccination strategies to mitigate the spread of measles on a college campus. Brandeau, Zaric, & Richter (2003) develops optimal vaccine allocation strategies across independent populations. Becker & Starczak (1997) solves a linear programming problem for allocating vaccines across a community of households. Allocation strategies for problems with small state spaces have been solved with optimal control strategies via the Hamilton-Jacobi-Bellman equations (Asano, Gross, Lenhart, & Real, 2008; Ding, Gross, Langston, Lenhart, & Real, 2007; Neilan & Lenhart, 2011; Zakary, Rachik, & Elmouki, 2017).

Bisset, Feng, Marathe, & Yardi (2009) and Porco et al. (2004) use stochastic network models to overcome the homogenous mixing issues with compartmental models. The former considers the vaccination decision at the onset of the epidemic and the latter implements a ring-vaccination policy to fight a small-pox epidemic. Zhang & Prakash (2014) formulates a graphical model to fight a pandemic with uncertainty in the transmission rates reflected through the edges in the graph. They propose and compare algorithms to allocate a limited number of vaccines by removing nodes in a network. Sélley, Besenyei, Kiss, & Simon (2015) and Watkins, Nowzari, & Pappas (2019) implements nonlinear model predictive control algorithms for optimizing compartmental epidemics in continuous time.

Intervention strategies for multi-stage optimization of vaccine allocation have been studied. Bytahtakn, des Bordes, & Kb (2018) creates a multi-stage formulation for solving a mixed integer program. Dasaklis, Rachaniotis, & Pappis (2017) proposes a linear programming model for optimizing vaccine demand in a supply chain model to control a smallpox outbreak of multiple time periods of a campaign. Nguyen & Carlson (2016) formulates a spatial SIR model for allocating vaccines among a small set of locations.

Stochastic resource allocation problems for epidemics have been studied throughout the literature. Yarmand, Ivy, Denton, & Lloyd (2014), Tanner, Sattenspiel, & Ntaimo (2008), Tanner & Ntaimo (2010) use two-stage stochastic programming formulations to allocate vaccines with uncertain parameters such as transmission rates, cost of vaccines, and mobility between regions. Cosgun & Esra Byktahtakn (2018) develops an approximate dynamic programming model using state aggregation for the dynamic allocation of resources during an AIDS epidemic. Dimitrov, Goll, Hupert, Pourbohloul, & Meyers (2011) develops an upper confidence bounding for trees algorithm for allocating antiviral drugs into a spatially distributed region with an aggregated action space. Probert et al. (2018) develops real-time forecasting models and analytic intervention strategies to mitigate the spread of a disease. Du, Sai, & Kong (2021) develops a rolling horizon scenario-based stochastic programming solution which can make decisions under uncertainty during a cholera outbreak. Han, Preciado, Nowzari, & Pappas (2015) designs an optimal resource allocation solution using geometric programming and robust optimization.

The value of learning through testing is also important if there are a limited number of testing kits to allocate. Shea, Tildesley, Runge, Fonnesbeck, & Ferrari (2014) captures the value of learning in an epidemic with the expected value of perfect information metric. Aside from epidemics, there are other problems which capture the value of learning through Bayesian optimization frameworks. We seek to perform active learning through the Bayesian optimization frameworks discussed in Frazier (2018) and Shahriari, Swersky, Wang, Adams, & De Freitas (2015). Active learning has been used for optimizing nonlinear belief models (Han & Powell, 2020). It has also been used for materials science (e.g. Packwood, 2017), engineering design (e.g. Imani & Ghoreishi, 2020), medical decision making (e.g. Wang & Powell, 2016), and drug discovery (e.g. Reyes & Powell, 2020).

There are various other stochastic optimization approaches for resource allocation problems throughout the literature. Gülpınar, Çanakoğlu, & Branke (2018) proposes an approximate dynamic programming algorithm for assigning a limited number of resources to as many tasks as possible. Creemers (2019) solves a preemptive stochastic resource constrained scheduling problem by restructuring the state space to efficiently solve a stochastic dynamic program via lookup tables. Chalabi, Epstein, McKenna, & Claxton (2008) solves a stochastic resource allocation problem with two stage stochastic programming in the healthcare setting. They calculate the expected value of perfect information to guide a learning problem for collecting more information. Li & Womer (2015) solves a stochastic resource-constrained project scheduling problem with approximate dynamic programming. They blend a rollout lookahead policy with a lookup table policy to achieve an efficient closed loop solution to their problem. Osorio, Brailsford, & Smith (2018) solves the problem of assigning donors to collection methods in a blood supply chain. They devise a stochastic integer linear programming model which couples the sample average approximation method with the epsilon-constraint algorithm. Powell (2019) designs a unified framework for stochastic optimization which ties together the strategies from over 15 different communities into four different classes of policies. Each of the stochastic optimization strategies listed in Powell (2019) can be aggregated into the four classes of policies. The four classes provide a basis for searching over all classes of policies; which is extremely rare in the literature.

Decision-making with a partially observable state of the world can be modeled as a partially observable Markov decision process (POMDP) (e.g. Cassandra, Kaelbling, & Littman, 1994). This modeling approach is widely used for problems with unobservable parameters or quantities, but it suffers from severe computational limitations (it probably cannot be applied to a problem in this paper with more than 3 or 4 zones). The ability to and the exact optimal solution is almost never possible for real world problems; in fact, the finite horizon POMDP is PSPACE-complete (e.g. Pineau et al. (2006)). Often overlooked, however, are subtle modeling assumptions that would not apply for our epidemic setting. In particular, the policy derived from the belief MDP uses the one-step transition matrix which, aside from being computationally intractable, implicitly assumes that the transition function is known to the controller. This means that the controller actually knows the dynamics of how the disease is communicated, which is not the case with COVID-19.

Pineau et al. (2006) gives a set of solutions using a point-based value iteration approach and compares it to other POMDP solvers. Ross et al. (2008) derives online planning algorithms for the POMDP problem. Hoey and Poupart (2005) formulates strategies for solving POMDPs with continuous and multi-dimensional observations spaces. Roy et al. (2005) attempts to circumvent the curse of dimensionality by finding a low dimensional belief space embedded in the high dimensional belief state to project into. There have been many algorithms developed for solving POMDPs to make the solutions tractable, but the computational complexities suggest that they would still only work for problems with small state, action, and observation spaces compared to the size of the problem discussed in this paper.

This paper captures uncertainty in the state of the epidemic while managing a limited vaccination resource (which can directly impact the environment) and a limited learning resource (which can only measure the environment through limited testing). Du et al. (2021) is the most recent and relevant research to our vaccination agent's strategy. They perform a rolling horizon policy with uncertainty around the parameters of their model. Our research differs in three major ways. Firstly, the multi-agent modeling we develop with the unified framework is different from their modeling strategy because we capture learning and vaccination through different agents. The learning agent must learn through observations with a limited number of resources. Our learning agent also learns about the probability over the compartments of the state space, instead of just the parameters of the model. Second, our controller is robust to changes to the environment model. In fact, any increasingly complex epidemic which can be tested for infections and responds to a vaccine decision could plug and play with our controller models and adaptively mitigate the spread of a virus because the environment is a black box. We demonstrate the versatility of our framework by implementing the model on scenarios with very large populations with moderate resource availability, as well as smaller populations under extreme resource shortages. Third, they present a scenario-based rolling horizon model, whereas, we present a parameterized multi-stage lookahead approximation which can be tuned to work best under different scenarios.

3. Multi-agent modeling framework

This section presents a mathematical framework extending the unified framework presented in Powell (2019) to a multi-agent setting with an epidemic application under partial observability. The standard unified framework is designed with the philosophy to model first, then solve the problem. The model consists of five components: the state variable, decision variables, exogenous information, transition function, and objective function. After the model is constructed, the problem is solved by designing policies by searching over the four classes of policies which encompass any stochastic optimization solution strategy.

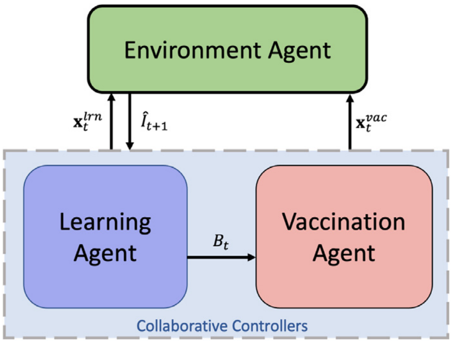


Fig. 2. Flowchart of interactions within the multi-agent model. The learning agent makes test kit allocation decisions, $\mathbf{x}_t^{lrn}$ and receives a random number of positive samples in return, $\hat{I}_{t+1}$, from the environment. The vaccination agent receives the belief state, B_t, from the learning agent model and uses the new information to make a vaccine allocation decision, $\mathbf{x}_t^{vac}$. The dynamics of the environment are directly changed by the vaccination agent.

Then, we extend the standard unified framework modeling process to a multi-agent formulation for partially observable systems. In this paper, we have an environment agent and two controlling agents. The environment agent represents the epidemic system and does not make decisions, but it can be observed through tests and impacted by vaccines. There are also two controlling agents which collaborate together to complete a joint goal of minimizing the cumulative number of new infections. There is a vaccination agent responsible for allocating a dynamic stockpile of vaccines, n_t^{vac} to a set of zones and a learning agent responsible for allocating a dynamic stockpile of testing kits, n_t^{test} to the same set of zones.

Each agent has its own model from the five components of the unified framework and its own policy function for making decisions. They can characterize their own perspectives with individual models and make decisions according to its own individual objectives using separate policies. We will demonstrate the multi-agent collaboration between a learning agent and an vaccination agent. The agents have unique abilities because they have different resources. The learning agent is responsible for constructing and maintaining a belief model describing the probability distributions over the uncertain state of the environment: the belief state, B_t. The learning agent communicates the belief state to the vaccination agent and it can utilize the new information to make the most impactful vaccine allocation decisions.

Each agent has its own model, but the actual dynamics of the environment will be represented as general functions. The learning agent will make test allocation decisions, $\mathbf{x}_t^{lrn}$, to receive samples of infected individuals, $\hat{I}_{t+1}$. Then, update the belief model, which it will communicate to the vaccination agent to inform the vaccine allocation decisions, $\mathbf{x}_t^{vac}$. The flow of information between each agent is displayed in Figure 2.

At each discrete time step there is a sequence of events that occur. For example at time t, the samples from the previous test allocation from time $t-1$ are realized, which leads to a new belief state, B_t. Then, the belief state is transferred to the vaccination agent to make a vaccine allocation decision $\mathbf{x}_t^{vac}$. The learning agent uses the knowledge from the vaccine allocation to strategically allocate the tests at time t to be distributed to collect the new samples for time $t+1$.

The following sections will present the mathematical modeling frameworks for each model. Section 3.1 presents the general model for the environment agent. Section 3.2 presents the learning agent model which is prefaced by the belief model in section 3.2.1. Section 3.3 presents the vaccination agent model.

3.1. Environment agent

The region in this problem is partitioned into a set of zones $z \in \mathcal{Z}$. Each zone within the region has a population of individuals, N_z. There is a disease present within the population which evolves according to some fixed dynamics. Throughout the time horizon T, there will be an exogenous process which will produce a stockpile of vaccines, n_t^{vac}, and testing kits, n_t^{test}, at the beginning of each time step.

The environment model is a passive agent, so it evolves through time without making decisions. However, it has its own dynamics and can be impacted by controller decisions. A passive agent only has three of the five components of the unified framework because it does not make decisions or have an objective. It has a state variable, exogenous information, and transition functions. The ground truth components of this model may be a complex simulator or the real world. For the purposes of this paper, we limit the environment state variable to be the states of the SIR model; however, other compartmental extensions are easily appended to the model. The general set of true parameters at time t are packaged into Ψ_t.

Environment State Variable The environment state variable, $\mathcal{S}_t^{env}$, represents the information the environment would need to transition to the next state from time t onward. It has the following form:

$$\mathcal{S}_t^{env} = (S_{tz}, I_{tz}, R_{tz}, \Psi_t)_{z\in\mathcal{Z}}, \tag{1}$$

where,

S_{tz} = true number of individuals susceptible to the disease in zonezat time$t$,

I_{tz} = true number of individuals infected with the disease in zonezat time$t$,

R_{tz} = true number of individuals removed (immune/vaccinated) from susceptibility the disease in zonezat time$t$,

Ψ_t = true parameters of the SIR model at time$t$,

The assumed environment state at time 0 is separated from the dynamic state because it includes latent variables which are fixed over time, given by,

$$\mathcal{S}_0^e = (S_{0z}, I_{0z}, R_{0z}, \Psi_{0z}, N_z)_{z\in\mathcal{Z}}, \tag{2}$$

where,

N_z = the population of each zone.

Environment Exogenous Information The environment has dynamically changing parameters. For example, the transmission rates, recovery rates, and vaccine efficacy are all streaming over time, but the distributions are unknown. Additionally, from the environment agent's perspective the vaccine allocations are an exogenous information process:

$$W_{t+1}^{env} = \left(\Psi_{t+1}, \mathbf{x}_t^{vac}\right) \tag{3}$$

Environment Transition Function The true transition functions are not known by the controllers and can be increasingly complex. We denote the black-box transition function as $f^{env}(\mathcal{S}_t^{env}, W_{t+1}^{env})$.

Observation Function

The final component of the environment model is the observation function which draws samples from testing centers. Assume the random variable $\hat{I}_{t+1,z}$ has a binomial distribution with parameters $(n, p) = (x_{tz}^{lrn}, p_{tz}^{test})$. The probability, p_{tz}^{test}, is the probability of randomly selecting an infected individual out of the population of individuals who received a test at time t in zone z. This probability is affected by factors such as the likelihood of going to get a test while showing symptoms, the likelihood of showing symptoms while positive and the probability of false positives and false negatives. The exact impact those factors will have on p_{tz}^{test} is not known to the controller.

The real world will almost always be more complex than any simulator of the environment, and a controlling agent would have to approximate the real world to the best of its ability. The simulator designed for this paper was made as complex and realistic as possible by including more stochasticity and complexity than the controlling agent model, and biasing the sampling to approximate dynamic human behavior and asymptomatic spread. The true simulator models used to test this model is given by Appendix 8.2.

3.2. Learning agent

This section proposes a model for the learning agent using the five components of the unified framework. The controller does not have access to the environment agent's state variable, $\mathcal{S}_t^{env}$, or the transition functions describing how they evolve, $f^{env}(\mathcal{S}_t^{env}, W_{t+1}^{env})$. The distributions over the dynamic components of the environment state variable are maintained by the following belief model.

3.2.1. Belief model

In a sequential decision problem with imperfectly known states and state transitions, the controller must maintain a belief model. This belief model contains three major components:

1) the environment model assumptions,
2) the belief state,
3) the updating equations for the belief state.

Environment Model Assumptions The states of the environment are random variables from the perspectives of the controlling agents. Eq. (1) defines the general form for the environment state variable. We assume the true parameters of the SIR model are

$$\Phi_{t+1} = \left(\bar{\beta}_z, \gamma_z, \xi\right)_{z\in\mathcal{Z}},$$

where,

$\bar{\beta}_z$ = true transmission rate at time$t$in zone$z$,

γ_z = true recovery rate in zone$z$,

ξ = vaccine efficacy.

We assume the transmission rate is a dynamically changing stochastic process in each zone, the recovery rate is fixed in each zone, and the vaccine efficacy is the same for the entire region and fixed over time. We argue these are reasonable assumptions because the transmission rates reflect human behavior within a zone over time and can be time dependent and random. The recovery rate reflects the latency between being infected and either naturally recovering or dying from the illness so it is generally fixed within each zone. It is heterogeneous between zones because the healthcare and access to hospitals may be different. We assume the vaccine technology over the time horizon is fixed so the vaccine efficacy does not change.

The assumed transition functions for the subpopulations in the environment model follow a modified version of the classic SIR compartmental model in epidemiology. The equations describe how each subpopulation within each zone interacts and evolve through the time horizon. The equations are given by,

$$S^x_{tz} = S_{tz} - \min(S_{tz}, \xi x^{vac}_{tz}), \tag{4}$$

$$S_{t+1,z} = S^x_{tz} - \frac{\beta_{t+1,z}}{N_z} I_{tz} S^x_{tz}, \tag{5}$$

$$I_{t+1,z} = (1-\gamma_z) I_{tz} + \frac{\beta_{t+1,z}}{N_z} I_{tz} S^x_{tz}, \tag{6}$$

$$R_{t+1,z} = R_{tz} + \gamma_z I_{tz} + \min(S_{tz}, \xi x^{vac}_{tz}) \tag{7}$$

where S^x_{tz} is the post-decision state of the susceptible group. The post-decision state represents the state of the susceptible subpopulation after the vaccination decision has been made, but before the system transitions to $t+1$. The post-decision state reflects the individuals effectively vaccinated between time t and $t+1$ and removed from the susceptible population. The susceptible compartment is reduced by the number of post-vaccination susceptible interacting with infected people at each step. The infected compartment gains those individuals, but individuals are removed at a rate γ_z into the removed compartment. The vaccinated individuals are moved into the removed compartment.

The transmission rates are assumed to be random perturbations in the interval (0,1) around an average β and have the following form,

$$\beta_{t+1,z} = \bar{\beta}_z + \varepsilon^{\beta}, \tag{8}$$

where, $\varepsilon^{\beta} \sim Unif(-\delta^{\beta}, +\delta^{\beta})$.

Belief State Since the learning agent cannot observe the environment perfectly at time t it must maintain a probability distribution over the entire state space: the belief state. The SIR model assumes that the total population remains constant. Therefore, the belief about the true percentage of each subpopulation in a zone has the following property,

$$S_{tz} + I_{tz} + R_{tz} = N_z. \tag{9}$$

Let $\bar{p}^S_{tz}$, $\bar{p}^I_{tz}$ and $\bar{p}^R_{tz}$ denote estimates of the percentage of each population in each subpopulation of zone z at time t. The estimates have the same property as the true parameters in Eq. (9). Hence, the most natural distribution to reflect this structure is a multinomial distribution for each zone with parameters $(N_z, \bar{p}^S_{tz}, \bar{p}^I_{tz}, \bar{p}^R_{tz})$. Specifically given by,

$$(S_{tz}, I_{tz}, R_{tz}) \sim Mult(N_z, \bar{p}^S_{tz}, \bar{p}^I_{tz}, \bar{p}^R_{tz}).$$

Furthermore, this implies the dynamic belief state for the controlling agent models is given by,

$$B_t = (\bar{p}^S_{tz}, \bar{p}^I_{tz}, \bar{p}^R_{tz})_{z\in\mathcal{Z}}. \tag{10}$$

Belief State Update The belief state updating equations take the observations, $\hat{I}_{t+1,z}$, queried from the testing centers and use them to estimate the new belief state at $t+1$. Each of the three estimates $\bar{p}^S_{t+1,z}$, $\bar{p}^I_{t+1,z}$, and $\bar{p}^R_{t+1,z}$ in each zone z will need an updating equation. We will update the belief state through a Bayesian procedure outlined in Fig. 3.

The first step to updating the model is to formulate priors through the forecasting model. To forecast, the conditional expectation of each subpopulation at $t+1$ is estimated with the dynamics we assumed in Eqs. (4–7). The closed form expectation does not exist, so we approximate it with normal distributions.

We denote variables in the forecast model with f and superscripted with the variable being forecasted from the current time t and one for the future time $t+1$ (e.g. $f^I_{t,t+1,z}$ forecasts I). We state the equations in Lemma 3.1, but leave the details to the Appendix.

Lemma 3.1. *The predictions at* $t+1$ *describe the conditional expectation of the subpopulations for each of the belief state variables passed through the transition function in Eqs. (4–7). As the size of the population gets large, the multinomial distributions will converge into normal distributions in the limit. This property allows us to approximate the conditional expectation of Eqs. (4–7) with respect to the belief state. Let* $X = \frac{\beta_{t+1,z}}{N_z^2}$ *and* $Y = I_{tz}\left(S_{tz} - x^{vac}_{tz}\right)$*, which are independent random variables. Let* $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|B_t, x^{vac}_{tz}]$*. The equations are given as follows,*

$$f^{S,x}_{ttz} = \mathbb{E}_t\left[\frac{S^x_{tz}}{N_z}\right] = \bar{\mu}^S_{tz}\Phi\left(\frac{\bar{\mu}^S_{tz}}{\bar{\sigma}^S_{tz}}\right) + \bar{\sigma}^S_{tz}\phi\left(\frac{\bar{\mu}^S_{tz}}{\bar{\sigma}^S_{tz}}\right), \tag{11}$$

$$f^S_{t,t+1,z} = \mathbb{E}_t\left[\frac{S_{t+1,z}}{N_z}\right] = f^{S,x}_{ttz} - \bar{\mu}^{XY}_{tz}\Phi\left(\frac{\bar{\mu}^{XY}_{tz}}{\bar{\sigma}^{XY}_{tz}}\right) - \bar{\sigma}^{XY}_{tz}\phi\left(\frac{\bar{\mu}^{XY}_{tz}}{\bar{\sigma}^{XY}_{tz}}\right), \tag{12}$$

$$\begin{aligned} f^I_{t,t+1,z} = \mathbb{E}_t\left[\frac{I_{t+1,z}}{N_z}\right] &= (1-\gamma_z)\bar{p}^I_{tz} + \bar{\mu}^{XY}_{tz}\Phi\left(\frac{\bar{\mu}^{XY}_{tz}}{\bar{\sigma}^{XY}_{tz}}\right) \\ &\quad + \bar{\sigma}^{XY}_{tz}\phi\left(\frac{\bar{\mu}^{XY}_{tz}}{\bar{\sigma}^{XY}_{tz}}\right), \end{aligned} \tag{13}$$

$$f^R_{t,t+1,z} = \mathbb{E}_t\left[\frac{R_{t+1,z}}{N_z}\right] = \bar{p}^R_{tz} + \gamma_z\bar{p}^I_{tz} + \bar{p}^S_{tz} - f^{S,x}_{ttz}, \tag{14}$$

where,

$$\begin{aligned} \bar{\mu}^S_{tz} &= (\bar{p}^S_{tz} - \frac{\xi}{N_z}x^{vac}_{tz}), \\ \bar{\sigma}^S_{tz} &= \sqrt{\frac{\bar{p}^S_{tz}(1-\bar{p}^S_{tz})}{N_z}}, \\ \bar{\mu}^{XY}_{tz} &= \mathbb{E}_t[X]\mathbb{E}_t[Y], \\ \bar{\sigma}^{SI}_{tz} &= \sqrt{Var_t[XY]}, \end{aligned}$$

and Φ *is the standard normal cdf and* ϕ *is the standard normal pdf. For explicit expressions and vaccination details for the moments of the random variables* X *and* Y*, see the Appendix.*

Proof. *See Appendix.* □

The sample of infected individuals from each zone are drawn from a binomial distributions with x^{lrn}_{tz} samples from each zone determined by the learning policy and unknown probability parameters. The conjugate prior for the binomial distribution is a beta distribution which effectively puts a prior distribution over the unknown parameters. The updating equation for the infected population can be seen in Lemma 3.2.

Lemma 3.2. *Let* $\hat{I}_{t+1,z}$ *be a sample drawn from a binomial distribution with* x^{lrn}_{tz} *trials. Let* $(\alpha_{tz}, \kappa_{tz})$ *be parameters of a beta distribution encoding the prior information known about* $I_{t+1,z}$*. The compound distribution produced by Bayes' Theorem is a beta-binomial distribution. The estimator for the probability of an infection,* $\bar{p}^I_{t+1,z}$ *is given by,*

$$\bar{p}^I_{t+1,z} = \frac{\hat{I}_{t+1,z} + \alpha_{tz}}{x^{lrn}_{tz} + \lambda N_z}. \tag{15}$$

$\lambda \in (0,1)$ *is a tunable weighting factor based on how much we trust the observations versus the model. Hence, the beta distribution parameter is given by,*

$$\alpha_{tz} = \lambda N_z f^I_{t+1,z}, \tag{16}$$

$$\kappa_{tz} = \lambda N_z - \alpha_{tz}, \tag{17}$$

where, $\tilde{p}^I_{t+1}$ *is computed using Eq. (13).*

Proof. *See Appendix.* □

After the tests have been administered into the population, it is possible to get an estimate of the number of infected individuals;

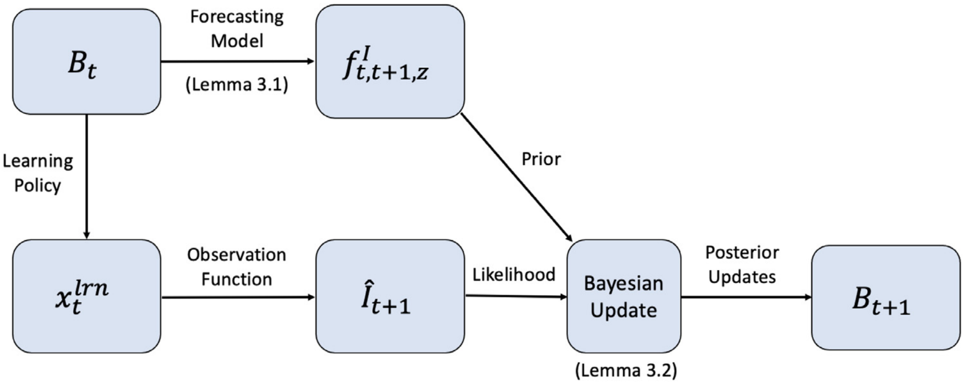


Fig. 3. Flowchart displaying the updating procedure for the belief state.

however, there are two other groups in the population: susceptible and removed. Since we only have observations of the number of infected individuals at time $t+1$, then to estimate the susceptible and removed subpopulations we will use the predictions from Lemma 3.1 and the posterior from Lemma 3.2.

Definition 3.1. Let Π_{Δ_z} be the projection operator for the set defined by,

$$\Delta_{t+1,z} = \left\{(p^S, p^R) \in [0,1]^2 : p^S + p^R = 1 - \bar{p}^I_{t+1,z}\right\}.$$

If the terms are not in the set Δ_z in definition 3.1, then they must be projected back to the nearest point. This projection operation for the susceptible and removed subspopulations is then given by,

$$(\bar{p}^S_{t+1,z}, \bar{p}^R_{t+1,z}) = \Pi_{\Delta_z}\left[\left(f^S_{t,t+1,z}, f^R_{t,t+1,z}\right)\right]. \quad (18)$$

In summary, the controlling agent updates the parameters in the belief state through the following process:

1) Make observations $\hat{I}_{t+1,z}$ for all $z \in \mathcal{Z}$,
2) Compute the belief state predictions using Lemma 3.1,
3) Compute the Bayesian update using Eq. (15) in Lemma 3.2,
4) Use Eq. (18) to update $\bar{p}^S_{t+1,z}$ and $\bar{p}^R_{t+1,z}$.

3.2.2. Learning agent model

The learning agent is responsible for the allocation of testing kits. The testing kits are used to collect information about the state of the infection in each zone. The following subsection presents the five components of the unified framework for the learning agent model.

State Variable The state variable for the learning agent model contains all information needed to update the learning agent transition functions, compute the policy, and evaluate the objective function. The learning agent state variable is given by,

$$\mathcal{S}^{lrn}_t = \left(B_t, n^{test}_t, n^{vac}_t\right),$$

where,

$$B_t = \text{the belief state from time} t,$$
$$n^{vac}_t = \text{the number of vaccines available at time} t,$$
$$n^{test}_t = \text{the number of tests available at time} t.$$

The learning agent state variable is very similar to the vaccination agent; however, it also needs the vaccine stockpile because it must be able to compute the vaccination policy to evaluate the value of collecting information. The initial state variable for the learning agent is given by,

$$\mathcal{S}^{test}_0 = \left(B_0, n^{test}_0, n^{vac}_0, \xi, (\gamma_z, \bar{\beta}_z, N_z)_{z\in\mathcal{Z}}\right).$$

Decision Variable The decision to allocate testing kits to each zone follows the same structure as the vaccine allocation decision. The test kit decision is given by the vector $\mathbf{x}^{test}_t$, and constrained by the total number of testing kits available, n^{test}_t.

$$\mathcal{X}^{lrn}_t = \{\mathbf{x}^{lrn}_t \in \mathbb{N}^{|\mathcal{Z}|} : \sum_{z\in\mathcal{Z}} x^{lrn}_{tz} \le n^{test}_t\}. \quad (19)$$

The testing kit allocation also must remain in the set of natural numbers because partial kits cannot be allocated.

Exogenous Information The exogenous information process contains all information which streams into the learning agent. The learning agent receives the random samples queried by the testing kit allocation at time t. The learning agent also receives the vaccination decision from the vaccination policy. The learning agent will receive the vaccination decision before it makes the testing kit allocation at time t, and then receive all exogenous random information between t and $t+1$. The set of all exogenous information is given by,

$$W^{lrn}_{t+1} = \left(n^{test}_{t+1}, (\hat{I}_{t+1,z}, \gamma_{t+1,z})_{z\in\mathcal{Z}}\right).$$

We omit the vaccination decision from the process to reiterate that it arrives earlier than $t+1$.

Transition Function The transition function for the learning agent describes the set of equations for updating each of the state variables. The test kit stockpile is evolving exogenously. The updating procedure for the belief model is given in section 3.2.1.3, and the explicit procedure for the components of the belief state are given by Eqs. (15) and (18).

Objective Function The joint goal of the agents is to minimize the cumulative number of new infections. Hence, the one-step cost is given by,

$$C(\mathcal{S}^{env}_t, \mathbf{x}^{vac}_t, \mathbf{x}^{lrn}_t) = \sum_{z\in\mathcal{Z}} I_{t+1,z} - I_{tz}. \quad (20)$$

The true one-step cost is not possible to evaluate online, so the expectation must be taken over the belief state. Therefore, the optimization problem for this problem becomes,

$$\min_{\pi^{lrn}\in\Pi^{lrn}} \mathbb{E}\left\{\sum_{t=0}^{T-1} C(\mathcal{S}^{env}_t, \mathbf{x}^{vac}_t, \mathbf{x}^{lrn}_t)\middle|\mathcal{S}^{lrn}_0\right\}, \quad (21)$$

where Π^{lrn} is the set of all admissible testing kit allocation policies.

3.3. Vaccination agent

The vaccination agent is responsible for making vaccine allocation decisions. The remainder of this section lays out the five components of the mathematical model for the vaccination agent: the state variable, decision variable, exogenous information, transition function, and objective function.

3.3.1. Vaccination agent model

State Variables The state variables include the information which is needed to compute the transition functions, objective function, and policy at time t. Any information which is not changing dynamically remains a latent variable defined in the initial state. The state variable for the vaccination agent's base model is defined as,

$$\mathcal{S}_t^{vac} = \left(B_t, n_t^{vac}\right) \tag{22}$$

where,

$$B_t = \text{the belief state communicated from the learning agent},$$
$$n_t^{vac} = \text{the number of vaccines available at time} t.$$

The initial state contains the initial dynamic variables and static parameters of the model, given by,

$$\mathcal{S}_0^{vac} = \left(B_0, n_0^{vac}, \xi, \delta^{\beta}, (\gamma_z, \bar{\beta}_z N_z)_{z\in\mathcal{Z}}\right).$$

Decision Variables The decision to allocate vaccines to each zone is given by a vector, $\mathbf{x}_t^{vac}$, which is constrained by the total number of vaccines available, n_t^{vac}. Hence, the vaccine decision set is given by,

$$\mathcal{X}_t^{vac} = \{\mathbf{x}_t^{vac} \in \mathbb{N}^{|\mathcal{Z}|} : \sum_{z\in\mathcal{Z}} x_{tz}^{vac} \leq n_t^{vac}\}. \tag{23}$$

Note, this set must be constrained to the natural numbers because there cannot be partial vaccines distributed.

Exogenous Information The exogenous information, W_{t+1}^{vac}, represents all information that arrives between time t and $t+1$. It is given by,

$$W_{t+1}^{vac} = \left(B_{t+1}, n_{t+1}^{vac}\right). \tag{24}$$

The vaccination agent is completely dependent on the information arriving from the learning agent.

Transition Function The entire state variable arrives exogenously, hence $\mathcal{S}_{t+1}^{vac} = W_{t+1}^{vac}$.

Objective Function The one-step contribution for the joint goal is given by Eq. (20). Hence, the optimization problem for this problem becomes,

$$\min_{\pi^{vac}\in\Pi^{vac}} \mathbb{E}\left\{\sum_{t=0}^{T-1} C(\mathcal{S}_t^{env}, \mathbf{x}_t^{vac}, \mathbf{x}_t^{lrn}) \middle| \mathcal{S}_0^{vac}\right\}, \tag{25}$$

where Π^{vac} is the set of all admissible vaccination policies.

4. Designing policies

The policy is a mapping from the state space to the decision space. At time t there are vaccination decisions (the number of vaccines to allocate to each zone) and learning decisions (the number of testing kits to allocate to each zone). In section 4.1, we illustrate two types of vaccination policies: one from the PFA class and one is a parameterized DLA policy. In section 4.2, we present a one-step lookahead learning policy for deciding which zones to allocate testing kits to.

The policy, π, is a function used to map states into decisions which we designate, $X^{\pi}(\cdot)$. There are two general strategies for designing policies for stochastic optimization: policy search and lookahead approximations. Policy search looks within a class of functions for a policy that will work best with respect to some metric. The lookahead approximation strategy approximates the value a current decision will have on the future. The two solution strategies can be organized into the four classes of policies referenced in Powell (2019). The four classes of policies are defined as:

- **Policy Function Approximations (PFA)** are analytical functions which map states to decisions. These are policy search strategies because the parameters must be tuned to perform best. Some examples of PFAs are linear parametric functions, nonparametric functions, and look-up tables
- **Cost Function Approximations (CFA)** are policies that solve a parameterized optimization model, where the parameters are designed to account for uncertainty. Some examples of CFAs are upper confidence bounding or introducing buffer stock when optimizing a supply chain.
- **Value Function Approximations (VFA)** are lookahead approximation strategies which seek to solve Bellman's optimality equation with an approximation to the optimal value function. The fields of approximate dynamic programming (Powell, 2011), reinforcement learning (Sutton & Barto, 2018), and SDDP (Birge & Louveaux, 2011).
- **Direct Lookahead Approximations (DLA)** are lookahead strategies which optimize approximate models that look directly into the future. Some examples of DLAs are model predictive control (Agachi, Cristea, Csavdari, & Szilagyi, 2016) and Monte Carlo tree search (Browne et al., 2012).

It is also possible to form hybrids between the four classes, such as parameterizing a DLA which would be a hybrid between the CFA and DLA classes. The next sections will present the best performing policies for each agent from our simulation studies in section 5.

4.1. Vaccination policies

The vaccination decision in this problem chooses how many vaccines to send to each zone, x_{tz}^{vac}. The decision space for the next set of policies is given by Eq. (23). The state space for the vaccination agent has $4|\mathcal{Z}|+1$ dimensions; hence, as $|\mathcal{Z}|$ grows finding the optimal policy becomes quickly intractable due to the curse of dimensionality. Therefore, the approximation to the optimal policy must be designed by searching through the four classes of policies to find which one works best.

The following subsections will present policies from the PFA class and the DLA class. The policy from the PFA class allocates vaccines using an analytic function of the population of each zone. The PFA policy is designed to resemble a myopic policy which would be used by decision-makers in the real world. The DLA policy is a lookahead policy which solves a parameterized lookahead model which models the future but adds parameters to be tuned in order to adjust to the simulator (or real world online).

4.1.1. PFA: population-based allocations

The proportional PFA we present was the policy used for the COVID-19 pandemic. It simply takes the proportion of the population of each zone with respect to the total population, and creates a weighting. Then, the weight is used to allocate the proportion of the vaccines available. Hence,

$$X^{PFA}(\mathcal{S}_t^{vac}) = \left\lfloor \frac{N_z}{\sum_{z\in\mathcal{Z}} N_z} n_t^{vac} \right\rfloor. \tag{26}$$

4.1.2. DLA: parameterized two-step lookahead approximation

The parameterized DLA creates an approximate model of the future to make decisions at t by looking at the impact of decisions in the future. The lookahead model consists of the five components of the unified framework; however, parts of the model have been simplified to make the problem more tractable. There are several approximations that can simplify a lookahead model such as, reducing the horizon length, discretizing states and/or decisions, sampling using Monte Carlo methods, or creating a simple policy within the lookahead model to simulate the future.

We perform multiple approximation methods for solving the base model with the lookahead model. Firstly, we truncate the horizon length to look two steps into the future. The model is still not solvable because the belief states are continuous and multidimensional. We add a set of tunable parameters, $\theta^{DLA} \in (0,1) \times \mathbb{R}^4_+$, to the lookahead model to perform various functions. The set of parameters are given by $\theta^{DLA} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$. The first element, $\theta_0 \in (0,1)$ is used to parameterize the state space to select a tunable percentile of the distribution over susceptible individuals in each zone. The second through fifth elements are used to directly parameterize multiple elements of the nonlinear quadratic program in lemma 4.1. The parameterization allows the simulator to tune the policy to find a parameterization of the multi-stage deterministic program which performs best over multiple Monte Carlo evaluations of the simulation.

The following paragraphs will sketch the lookahead model. Any variables superscripted by θ are functions of the parameterization. *Lookahead State Variable* The lookahead state variable includes the approximate state variable which will be used to model the future. The lookahead state variable chooses the θ_0 percentile of the susceptible population of each zone.

The lookahead state variable, $\tilde{S}^{vac}_{tt'}$, is denoted with a tilde and two time subscripts. The first time subscript describes the time t in the base model and the second time subscript describes the time t' approximation in the future. The lookahead state variable is not the same as the base model state variable at time t because the approximations to the belief state must be realized. The lookahead state variable at time t is given by,

$$\tilde{S}^{vac}_{tt}(\theta^{DLA}) = \left(\tilde{p}^{S,\theta}_{ttz}, \tilde{p}^{I,\theta}_{ttz}, \tilde{p}^{R,\theta}_{ttz}\right)_{z\in\mathcal{Z}},$$

where,

$$\begin{aligned}\tilde{p}^{S,\theta}_{ttz} &= \bar{p}^S_{tz} - \Phi^{-1}(\theta_0)\frac{\sigma^S_t}{N_z},\\ \tilde{p}^{I,\theta}_{ttz} &= \bar{p}^I_{tz},\\ \tilde{p}^{R,\theta}_{ttz} &= \bar{p}^R_{tz} + \Phi^{-1}(\theta_0)\frac{\sigma^S_t}{N_z}.\end{aligned}$$

Lookahead Decisions The state variable induces a tunable chance constraint on the decision set to reduce the risk of allocating more vaccines than susceptible individuals. Hence, the decision set is given by,

$$\tilde{\mathcal{X}}^{vac}_{tt'} = \{\tilde{\mathbf{x}}^{vac}_{tt'} \in \mathbb{N}^{|\mathcal{Z}|} : \sum_{z\in\mathcal{Z}} \tilde{x}^{vac}_{tt'z} \le n^{vac}_t \quad \textit{and} \quad \tilde{x}^{vac}_{tt'z} \le N_z \tilde{p}^{S,\theta}_{tt'z} \forall z \in \mathcal{Z}\}.$$

Lookahead Exogenous Information The $\tilde{\beta}_{tt'z} = \bar{\beta}_z$ for all time periods to approximate the future. $\tilde{n}^{vac}_{tt'z} = n^{vac}_{tz}$ for all time periods to approximate the future.

Lookahead Transition Functions The lookahead transition functions are much simpler than the forecasting equations in the base model because there is no longer an expectation. The decision set restricts the number of vaccines allocated to be less than the number of susceptible individuals in the lookahead state variable. Also, there is no conditional expectation over the belief state because the approximations remove the uncertainty. The forecasting equations from Lemma 3.1 simplify to,

$$\tilde{p}^{S,\theta}_{t,t'+1,z} = \left(1 - \bar{\beta}_z \tilde{p}^{I,\theta}_{tt'z}\right)\left(\tilde{p}^{S,\theta}_{tt'z} - \frac{\xi}{N_z}\tilde{x}^{vac}_{tt'z}\right), \tag{27}$$

$$\tilde{p}^{I,\theta}_{t,t'+1,z} = \left(1 - \gamma_z + \bar{\beta}_z \tilde{p}^{S,\theta}_{tt'z} - \frac{\xi \bar{\beta}_z}{N_z}\tilde{x}^{vac}_{tt'z}\right)\tilde{p}^{I,\theta}_{tt'z}, \tag{28}$$

$$\tilde{p}^{R,\theta}_{t,t'+1,z} = \tilde{p}^{R,\theta}_{tt'z} + \gamma_z \tilde{p}^{I,\theta}_{tt'z} + \frac{\xi}{N_z}\tilde{x}^{vac}_{tt'z}. \tag{29}$$

Lookahead Objective Function The joint objective function from Eq. (20) can be approximated with the lookahead model with the following equations,

$$\begin{aligned}\tilde{C}(\tilde{S}^{vac}_{tt'}, \tilde{\mathbf{x}}^{vac}_{tt'}) &= \sum_{z\in\mathcal{Z}} N_z \tilde{p}^{I,\theta}_{t,t+1,z}(\tilde{x}^{vac}_{tt'z}) - N_z \tilde{p}^{I,\theta}_{tt'z}\\ &= \sum_{z\in\mathcal{Z}} N_z\left(-\gamma_z + \bar{\beta}_z \tilde{p}^{S,\theta}_{tt'z} - \frac{\xi \bar{\beta}_z}{N_z}\tilde{x}^{vac}_{tt'z}\right)\tilde{p}^{I,\theta}_{tt'z}.\end{aligned} \tag{30}$$

The optimization problem for a two-step lookahead approximation is given by,

$$X^{DLA,2}_t = \arg\min_{\tilde{\mathbf{x}}_{tt}, \tilde{\mathbf{x}}_{t,t+1}} \tilde{C}(\tilde{S}^{cont}_{tt}, \tilde{\mathbf{x}}_{tt}) + \tilde{C}(\tilde{S}^{cont}_{t,t+1}, \tilde{\mathbf{x}}_{t,t+1}), \tag{31}$$

which is now the summation of multiple one-step costs in the future. This formulation is much more manageable than trying to optimize the multi-period objective in the base model. The policy derived from this optimization problem is given by,

Lemma 4.1. *Let* $\mathbf{x}_t = \begin{bmatrix}\tilde{\mathbf{x}}^{vac}_{tt}\\ \tilde{\mathbf{x}}^{vac}_{t,t+1}\end{bmatrix}$ *be the two stage lookahead vaccination decision vector. Then, the policy can be rewritten as a non-convex quadratic program given by,*

$$X^{DLA}(S^{vac}_t|\theta^{DLA}) = \arg\max_{\mathbf{x}_t} \quad \mathbf{x}_t^T Q^\theta \mathbf{x}_t + (\mathbf{q}^\theta)^T \mathbf{x}_t, \tag{32}$$

$$\text{subject to} \quad K_1 \mathbf{x}_t \le \mathbf{h}^\theta_1, \tag{33}$$

$$K_2 \mathbf{x}_t \le \mathbf{h}_2. \tag{34}$$

Explicit expressions for the objective function in Eq. (32) and the constraints (33) and (34) can be found in the Appendix. The matrix Q^θ *is deconstructed into block matrices and each block is parameterized by* θ_1 *and* θ_2*. The vector* $\mathbf{q}^\theta$ *is split into its first* $|\mathcal{Z}|$ *components and second* $|\mathcal{Z}|$ *components and parameterized by* θ_3 *and* θ_4 *respectively.* Q^θ *has both positive and negative eigenvalues, in general; hence it is not always positive semidefinite.*

Proof. *See Appendix.* □

The optimization problem in Eq. (31) reduces to a problem with a nonconvex quadratic objective function with linear constraints. This approximation can be solved in practice with a bilinear quadratic solver when $|\mathcal{Z}|$ is not too large ($\lesssim 100$). Additionally, $\theta^{DLA} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4) \in (0,1) \times \mathbb{R_+}^4$ which requires offline parameter tuning to find the best value. The parameterizations will effect performance and could change whether the program is convex or not.

4.2. Learning policies

The second type of decision is to allocate tests to each zone to learn about the state of the pandemic. At time t, the learning agent must decide which zones to send the n^{test}_t kits after the vaccination policy has already been made. The learning decision will impact the distribution of the random samples drawn from the environment and impact the vaccine allocation decisions in the future. The large action space will limit the feasible acquisition functions available from the literature. Many of the policies are challenging to optimize in high dimensional spaces due to the computational complexity. The restricted options will narrow down the search over learning policies. In this section, we present a learning policy designed to capture the value of information.

One-step Variance Maximization The surrogate objective is designed to optimize the estimator from Eq. (15) because the other random variables are functions of $\bar{p}^I_{t+1}$. The surrogate function is given by,

$$f^{sur}(\mathbf{x}^{lrn}_t) = \sum_{z\in\mathcal{Z}} N_z \bar{p}^I_{t+1,z}(\mathbf{x}^{lrn}_t) = \sum_{z\in\mathcal{Z}} N_z \frac{\hat{I}_{t+1} + \alpha_{tz}}{x^{lrn}_{tz} + \lambda N_z}, \tag{35}$$

where we assume $\hat{I}_{t+1,z} \sim Bin(\mathbf{x}_t^{lrn}, f_{t,t+1,z}^I)$.

Lemma 4.2. *Let the mean and variance of the surrogate function be given by,*

$$\mu(\mathbf{x}_t^{lrn}) = \sum_{z \in \mathcal{Z}} N_z f_{t,t+1,z}^I, \tag{36}$$

$$\sigma^2(\mathbf{x}_t^{lrn}) = \sum_{z \in \mathcal{Z}} \frac{x_{tz}^{lrn} \alpha_{tz} \kappa_{tz} (\lambda N_z + x_{tz}^{lrn})}{\lambda^2 (\lambda N_z + 1)}, \tag{37}$$

where α_{tz} and κ_{tz} are parameters of the beta distribution prior given by Eqs. (16) and (17). Note, the mean function is a constant function with respect to the test kit decision.

Proof. *See Appendix.* □

Lemma 4.2 reveals the mean and variance of the surrogate function. The mean function is constant; hence, we will reduce the uncertainty in the surrogate function by optimizing the sum of variances. This leads to the policy in Lemma 4.3.

Lemma 4.3. *Let Eq. (37) be the optimization problem used to produce the learning decisions via the acquisition function. The optimization problem reduces to a non-convex quadratic program with linear constraints given by,*

$$\max_{\mathbf{x}_t^{lrn}} (\mathbf{x}_t^{lrn})^T D \mathbf{x}_t^{lrn} + d^T \mathbf{x}_t^{lrn}, \tag{38}$$

$$\text{subject to} \qquad \mathbf{1}^T \mathbf{x}_t^{lrn} \leq n_t^{test}$$
$$\mathbf{x}_t^{lrn} \in \mathbb{N}^{|\mathcal{Z}|}$$

where,

$$D = diag\left(\left[\frac{\alpha_{tz}\kappa_{tz}}{\lambda^2(\lambda N_z + 1)} \right]_{z \in \mathcal{Z}} \right)$$

$$d = \left[\frac{\alpha_{tz}\kappa_{tz} N_z}{\lambda(\lambda N_z + 1)} \right]_{z \in \mathcal{Z}}.$$

Proof. *See Appendix.* □

The one-step variance maximization policy is designed to minimize the variance in the estimator at $t+1$ by choosing to allocate tests which will minimize the sum of forecasted variances. Therefore, this policy creates a surrogate function designed to capture the amount of useful information gained through testing by minimizing the forecasted uncertainty in the next time step.

4.3. Fairness in allocation

Considering this problem makes decisions allocating resources into a set of zones in a heterogeneous population, then it is important to address the problem of fairness in both testing kit allocation and vaccine allocation. In this paradigm, we modeled the problem to minimize the overall sum of infected cases, but this could lead to a spike in one area to hoard all resources because it reduces the overall cases through the horizon. While the allocation may achieve the best outcome with respect to the defined cost function, it could also create inequities with respect to access to resources during the pandemic. The real world could have unintended consequences which were not considered in the original model.

Therefore, we propose a fairness trade-off policy for each type of decision to guarantee each zone gets resources available for a percentage of the population at each time step. Then, the rest of the resources are allocated according to the policy designed to optimize the model. Let X^{opt} represent a general allocation policy used to optimize the model for minimizing the overall number of cases using n_t resources. Let $\rho \in [0, 1]$ be a tunable parameter representing the percentage of the population we guarantee will have access to the respective resources in each zone. The proportional population-based allocation is given by Eq. (26) which could be implemented for the vaccination allocations, testing kit allocations, or both. Then, the allocation policy designed to optimize the models are applied with $(1-\rho)n_t$ resources. Hence, the general fairness policies are given by,

$$X^{fair}(\cdot|\rho) = X^{opt}(\cdot|\rho) + X^{prop}(\cdot|\rho). \tag{39}$$

We can tune the model to trade-off between fairness and optimizing the model.

5. Scenario simulations

In this section, we study two scenarios to demonstrate the versatility and robustness of our multi-agent modeling framework. The first scenario models each zone as the 50 states plus Washington D.C. (51 zones) over a 22 week period starting at the onset of vaccine production. In this scenario, we use real-data from the COVID-19 pandemic for the availability of vaccines stockpiles and test capacity starting from December 14, 2020 to May 11, 2021. The second scenario models a state-level vaccination agent and test administrator working together to allocate resources to nursing homes in the state of Nevada. The state of Nevada has the highest COVID cases per 1000 nursing home residents and due to the smaller population of the state they are unlikely to receive as many resources in a proportional allocation strategy (which was actually implemented in practice). Therefore, in the nursing home scenario we demonstrate that the multi-agent modeling strategy is robust under extreme shortages. The results in this section demonstrate that the resource allocation model can scale when resources are abundant and allocated to extremely large populations and can operate under extreme testing shortages (which leads to high risk of incorrect allocations).

In the simulation studies we conducted, we simulated vaccination allocation policies across three of the four classes of policies. We test a proportional allocation PFA, a risk adjusted CFA, an unparameterized 2-step lookahead policy, and a parameterized lookahead policy. The proportional PFA can be found in section 4.1.1. The CFA is a tunable integer program which directly optimizes the one-step contribution function at each time step. The parameterized DLA performs a rolling horizon optimization with lemma 4.1, and the parameters were tuned via policy search to the optimal values. The standard two-step risk-neutral deterministic lookahead solves the lookahead model without a parameterization. This is equivalent to solving the parameterized lookahead model with parameterization $\theta^{DLA} = (0.5, 1.0, 1.0, 1.0, 1.0)$.

We also tested each of the four vaccination policies in conjunction with three different learning policies. The learning policies are an even allocation PFA, a one-step variance maximization policy, and a fairness policy. The even allocation PFA allocated the testing kits evenly across each zone. The variance maximization policy solves the one-step lookahead optimization problem in lemma 4.3 to optimize the value of information. The fairness allocation policy guarantees 30% of the kits will be allocated weighted by population size, and the rest are allocated via the variance maximization policy. All static parameters for each scenario were left to the appendix if a reader is interested for implementation details.

5.1. United States COVID-19 simulation

In the US simulation, the federal government has two agents corresponding with each other to administer a stockpile of vaccines and a stockpile of testing kits to each state (zones). The vaccination agent receives a stockpile of vaccines each week to distribute based on the results from the previous week test results.

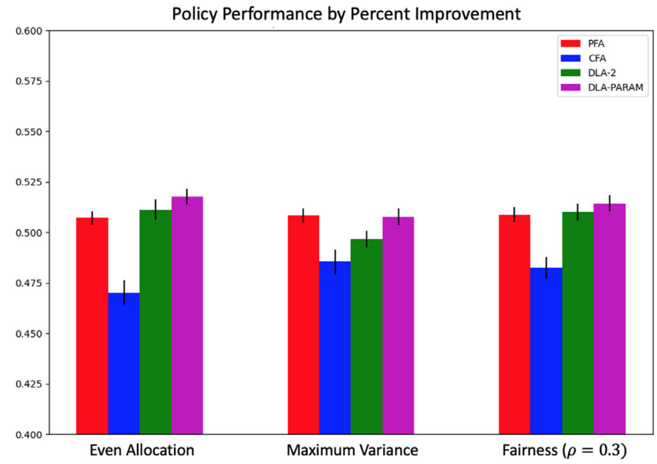


Fig. 4. Policy Performance using the percent improvement over the average number of infections with no intervention for the USA scenario. The groups of bars show each testing policy performance for each of the different vaccination policies. Each evaluation is a Monte Carlo average over 100 simulations.

The learning agent sends tests to be administered within each state to understand the state of the virus within each state. We used data from the CDC to simulate the results on a realistic availability of resources (Centers for Disease Control & Prevention, 2021) and data from the census bureau to simulate population density and sizes (US Census Bureau, 2020). The specific values of each parameter value in our model can be found in section 8.2. The transmission rates are assumed to have a constant mean generated by population densities (e.g. Martins-Filho, 2021). We assume they are constant because the public policies are generally constant over the time horizon, but there is noise added to each transmission rate process to account for dynamic and unpredictable human behavior. We assume the recovery rates have a similar structure as the transmission rates, but the constant mean values are determined based on hospital/care center density in each state (e.g. Bloom, Foroutanjazi, & Chatterjee, 2020). The vaccine efficacy is reported from CDC data based on the average over multiple clinical trials.

Fig. 4 shows the percent improvement for each combination of policies for each agent with respect to the performance of no allocation decisions. We display the performance of each vaccination policy for multiple different types of testing policies. The best policy combination for the two agents in the US scenario is to provide the vaccine administrator with the parameterized DLA policy and the test administrator with an even allocation policy. The parameterized DLA provides over a one percent improvement over the next best vaccine allocation policy under the same testing conditions, which would correspond to over half a million cases prevented. We tuned the values of each of the parameters in the lookahead model to $\theta^{DLA,*} = (0.25, 5.0, 0.2, 2.75, 0.75)$.

Fig. 5 shows the average allocations per state and the average new cases per state. The vaccination policy is the main contributor to performance for the US scenario. The next section will provide more insights into why there is not much difference in performance across each test allocation policy. In fact, we show empirically there is a critical threshold where each zone should be tested evenly versus trying to strategically allocate kits via value of information approximations.

5.2. Nursing home scenario

The alternative to the US scenario is a case where there are extreme shortages. During the height of a pandemic, there are likely to be extreme resource shortages in local areas which may not be favored for allocations at the federal level. Even if the local area is given a supply of resources, the tests are usually prioritized for symptomatic individuals and hospitals. Hence, there are scenarios where difficult decisions must be made by administrators. Consider a scenario where the state of Nevada has vaccines available for less than one percent of the nursing home residents and there is not enough testing capacity to test each of the 53 nursing homes in the state. We developed a simulation model where the infection levels in each of the nursing homes proceed independently, but there are stochastic spikes entering the nursing home which could be introduced by staff or visitors. It is imperative to try to minimize the uncertainty in the breakouts, but the testing capacity is

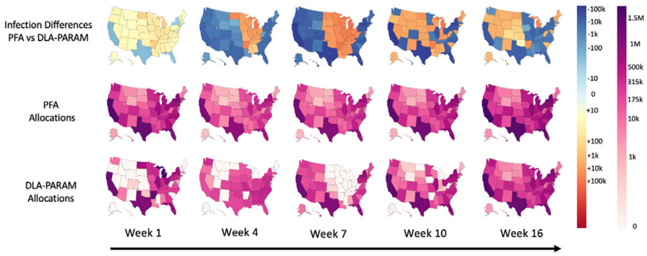


Fig. 5. Time Lapse of the differences between the Proportional PFA and the parameterized DLA policy. The top row displays the mean instantaneous difference between the infection levels for each policy simulation. The bottom two rows display the number of vaccine allocated to each zone for each of the policies.

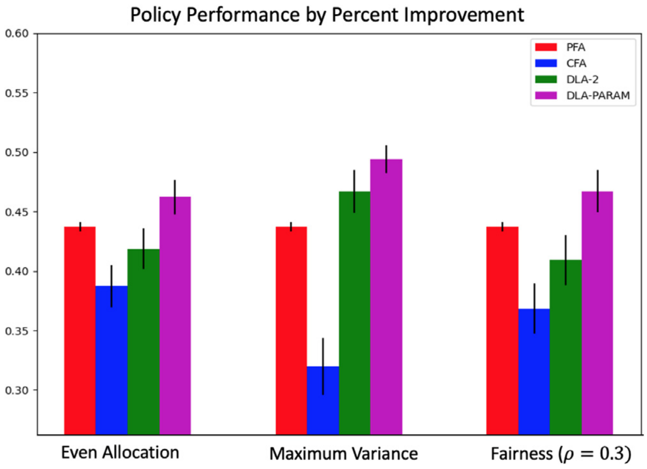


Fig. 6. Policy Performance using the percent improvement over the average number of infections with no intervention for the nursing home scenario. The groups of bars show each testing policy performance for each of the different vaccination policies. Each evaluation is a Monte Carlo average over 100 simulations.

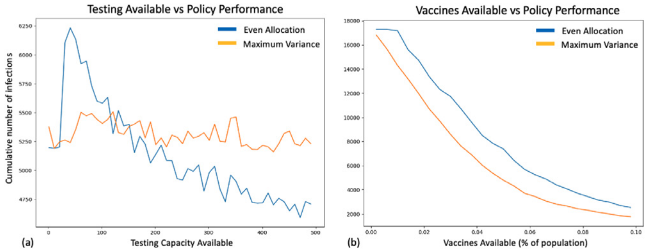


Fig. 7. We present learning policy comparisons for the even test allocation policy vs the maximum variance learning policy with the parameterized DLA vaccine allocation. (a) We vary the testing capacity versus a fixed vaccine stochastic process. (b) We vary the number of vaccines available (as a percent of the population) with a fixed testing capacity.

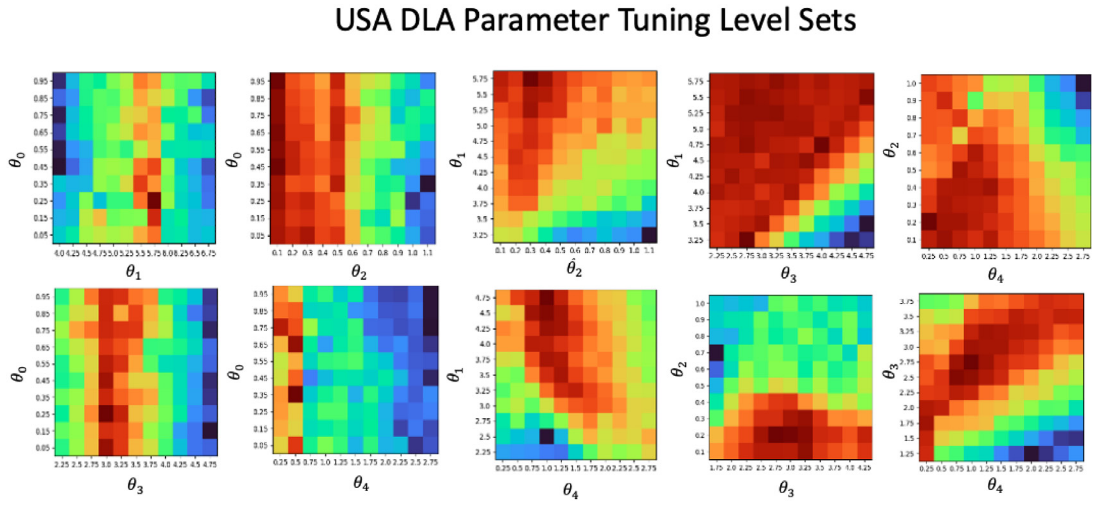


Fig. 8. Level sets for each combination of hyperparameter dimensions near the optimal values for the USA scenario.

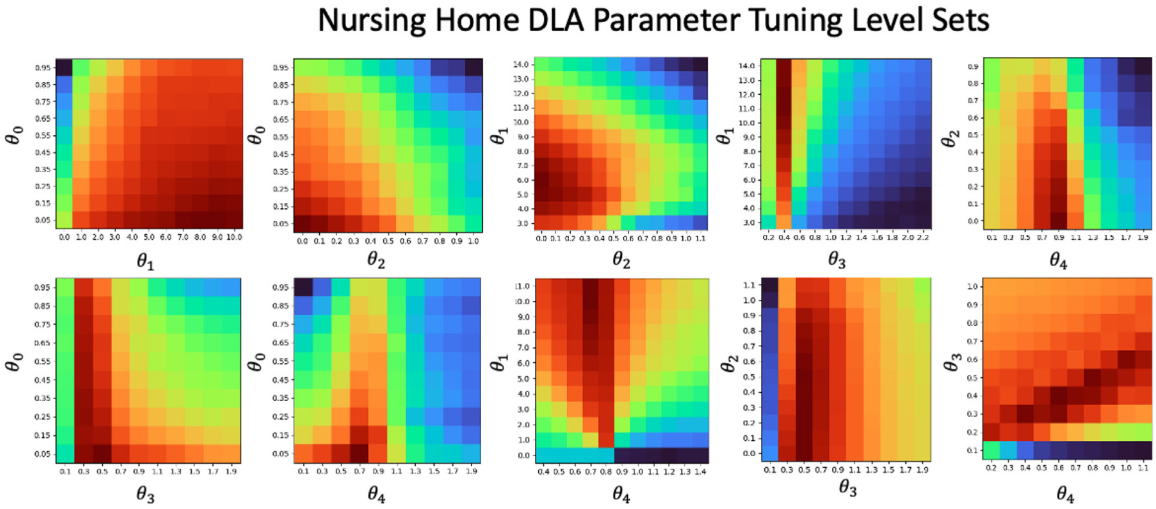


Fig. 9. Level sets for each combination of hyperparameter dimensions near the optimal values for the nursing home scenario.

under extreme shortages. We want to minimize the risk of severe outbreaks by monitoring the state of the pandemic in each nursing home, and we have to be strategic about how to allocate the testing kits.

Fig. 6 shows the infection curves for each of the different policies under severe shortages. The following plot demonstrates the

Fig. 7 demonstrates the risk of allocating evenly under a critical point of test capacity. There is a risk of severe outbreaks if every zone is not taking enough tests, whereas, there is a value to only allocating to certain zones with high variance. After the critical point it is no longer valuable to follow the maximum variance policy because there are a sufficient number of tests to collect enough valuable information from each zone.

5.3. Policy evaluation and tuning

The most critical aspect to achieving good performance for a parameterized DLA is tuning the hyperparameters of the policy. We optimized the parameters using a stochastic gradient descent method, and we present the level sets of the hyperparameter space near the optimum to show the differences. The optimal parameters for the USA problem were $\theta^{DLA,*} = (0.25, 5.0, 0.2, 2.75, 0.75)$. Figure 8 shows the level sets for each of the hyperparameters in the parameterized DLA for the USA problem.

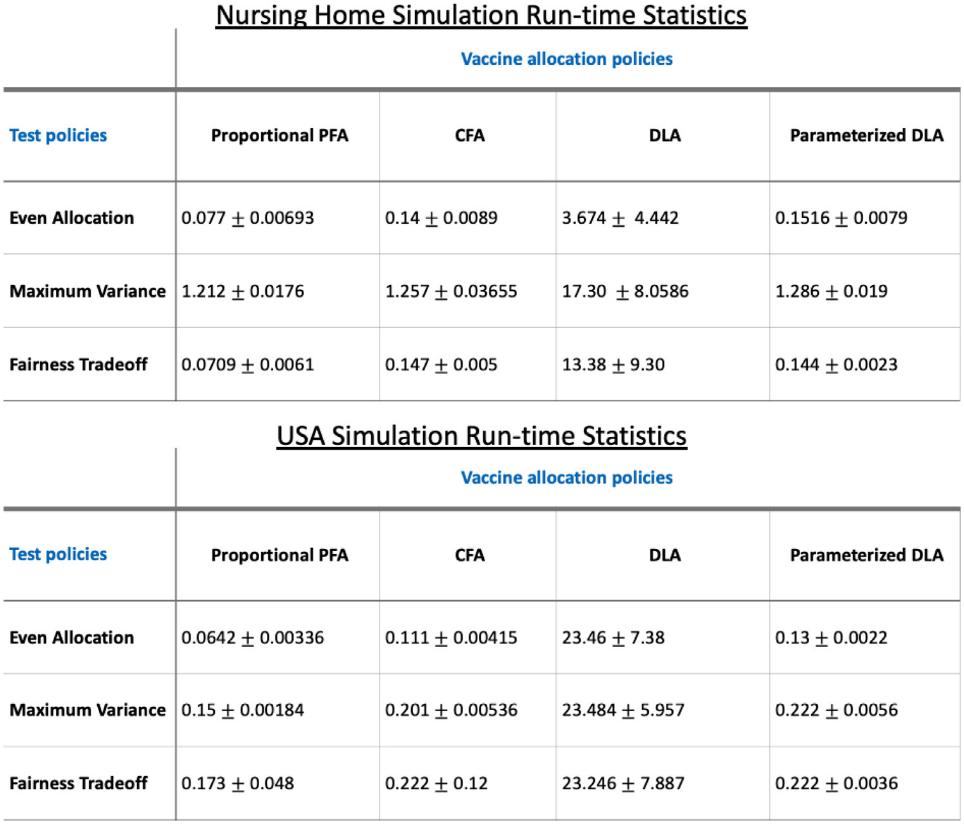
The optimal parameters for the nursing home scenario were $\theta^{DLA,*} = (0.05, 6.0, 0.0, 0.5, 0.7)$. Figure 9 shows the level sets for each of the hyperparameters in the parameterized DLA nursing home scenario.

Decision-makers must consider the runtime complexity when considering a decision-making strategy. Solving a nonlinear optimization problem takes more time because the complexity of the nonlinear solver is much larger than the complexity of an analytic function. The runtime statistics for each of the policy combinations is given in Fig. 10 below for both of the scenarios.

The unparameterized DLA takes significantly more time than the parameterized DLA. This phenomenon presents another advantage of the parameterized DLA for these specific scenarios. However, the runtimes are on the order of seconds which is negligible compared to the week long time steps for each allocation decision.

6. Conclusion

This paper contributes a multi-agent modeling extension to the unified framework for an epidemic application. We presented a formal multi-agent model for managing vaccines and tests during a



Nursing Home Simulation Run-time Statistics

	Vaccine allocation policies			
Test policies	**Proportional PFA**	**CFA**	**DLA**	**Parameterized DLA**
Even Allocation	0.077 ± 0.00693	0.14 ± 0.0089	3.674 ± 4.442	0.1516 ± 0.0079
Maximum Variance	1.212 ± 0.0176	1.257 ± 0.03655	17.30 ± 8.0586	1.286 ± 0.019
Fairness Tradeoff	0.0709 ± 0.0061	0.147 ± 0.005	13.38 ± 9.30	0.144 ± 0.0023

USA Simulation Run-time Statistics

	Vaccine allocation policies			
Test policies	**Proportional PFA**	**CFA**	**DLA**	**Parameterized DLA**
Even Allocation	0.0642 ± 0.00336	0.111 ± 0.00415	23.46 ± 7.38	0.13 ± 0.0022
Maximum Variance	0.15 ± 0.00184	0.201 ± 0.00536	23.484 ± 5.957	0.222 ± 0.0056
Fairness Tradeoff	0.173 ± 0.048	0.222 ± 0.12	23.246 ± 7.887	0.222 ± 0.0036

Fig. 10. The table shows the runtimes for each combination of policies for both of the scenarios. The runtimes are presented in seconds.

pandemic. Our work extends the unified framework for sequential decisions to the multi-agent setting for the first time. The multi-agent modeling strategy allows each agent to work with its own knowledge and adapt its policies based on the scenarios. Additionally, the unknown environment agent can easily be changed and the models for each agent do not need to be changed.

We demonstrate the robustness and scalability of the modeling strategy through two scenarios. The first scenario presents a model of COVID-19 in the USA. We collected vaccine and testing data from the CDC and used population data to construct a simulation for the agents to interact with. Then, we demonstrate the capabilities of our modeling framework to interact with the environment when there are millions of vaccines and tests to allocate to populations on the scale of hundreds of millions. The second scenario presents the state-level resource allocations to the nursing homes in the state of Nevada. The nursing home scenario shows the robustness of the model to perform under extreme resource shortages. The parameterized direct lookahead approximation can outperform policies from multiple other classes of policies, including the proportional PFA which was used to allocate vaccines during the COVID-19 pandemic.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2021.11.007

References

Agachi, P. S., Cristea, M. V., Csavdari, A. A., & Szilagyi, B. (2016). *2. Model predictive control*. De Gruyter.

Asano, E., Gross, L., Lenhart, S., & Real, L. (2008). Optimal control of vaccine distribution in a rabies metapopulation model. *Mathematical Biosciences and Engineering, 5*(1551-0018_2008_2_219), 219. https://doi.org/10.3934/mbe.2008.5.219.

Becker, N. G., & Starczak, D. N. (1997). Optimal vaccination strategies for a community of households. *Mathematical Biosciences, 139*(2), 117–132.

Birge, J. R., & Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Bisset, K. R., Feng, X., Marathe, M., & Yardi, S. (2009). Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In *Proceedings of the 2009 winter simulation conference (WSC)* (pp. 2020–2031). IEEE.

Bloom, J. A., Foroutanjazi, S., & Chatterjee, A. (2020). The impact of hospital bed density on the covid-19 case fatality rate in the united states. *The American Surgeon, 86*(7), 746–747.

Brandeau, M. L., Zaric, G. S., & Richter, A. (2003). Resource allocation for control of infectious diseases in multiple independent populations: Beyond cost-effectiveness analysis. *Journal of Health Economics, 22*, 575–598.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., … Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games, 4*(1), 1–43.

Buhat, C. A. H., Duero, J. C. C., Felix, E. F. O., Rabajante, J. F., & Mamplata, J. B. (2021). Optimal allocation of covid-19 test kits among accredited testing centers in the philippines. *Journal of Healthcare Informatics Research, 5*(1), 54–69.

Byktahtakn, I. E., des Bordes, E., & Kb, E. Y. (2018). A new epidemics–logistics model: Insights into controlling the ebola virus disease in west africa. *European Journal of Operational Research, 265*(3), 1046–1063. https://doi.org/10.1016/j.ejor.2017.08.037.

Cassandra, A. R., Kaelbling, L. P., & Littman, M. L. (1994). Acting optimally in partially observable stochastic domains. *AAAI, 94*, 1023–1028.

Centers for Disease Control and Prevention (2021). Covid data tracker. Https://covid.cdc.gov/covid-data-tracker/.

Chalabi, Z., Epstein, D., McKenna, C., & Claxton, K. (2008). Uncertainty and value of information when allocating resources within and between healthcare programmes. *European Journal of Operational Research, 191*(2), 530–539.

Cosgun, O., & Esra Byktahtakn, I. (2018). Stochastic dynamic resource allocation for HIVprevention and treatment: An approximate dynamic programming approach. *Computers and Industrial Engineering, 118*, 423–439. https://doi.org/10.1016/j.cie.2018.01.018.

Creemers, S. (2019). The preemptive stochastic resource-constrained project scheduling problem. *European Journal of Operational Research, 277*(1), 238–247.

Dai, T., Cho, S.-H., & Zhang, F. (2016). Contracting for on-time delivery in the u.s. influenza vaccine supply chain. *Manufacturing & Service Operations Management, 18*(3), 332–346. https://doi.org/10.1287/msom.2015.0574.

Dasaklis, T. K., Rachaniotis, N., & Pappis, C. (2017). Emergency supply chain management for controlling a smallpox outbreak: The case for regional mass vaccination. *International Journal of Systems Science: Operations & Logistics, 4*(1), 27–40.

Dimitrov, N. B., Goll, S., Hupert, N., Pourbohloul, B., & Meyers, L. A. (2011). Optimizing tactics for use of the us antiviral strategic national stockpile for pandemic influenza. *PloS one, 6*(1), e16094.

Ding, W., Gross, L., Langston, K., Lenhart, S., & Real, L. (2007). Rabies in raccoons: Optimal control for a discrete time model on a spatial grid. *Journal of Biological Dynamics*. https://doi.org/10.1080/17513750701605515.

Du, M., Sai, A., & Kong, N. (2021). A data-driven optimization approach for multi--period resource allocation in cholera outbreak control. *European Journal of Operational Research, 291*(3), 1106–1116.

Duijzer, L. E., van Jaarsveld, W., & Dekker, R. (2018). The benefits of combining early aspecific vaccination with later specific vaccination. *European Journal of Operational Research, 271*(2), 606–619.

Ekici, A., Keskinocak, P., & Swann, J. L. (2008). Pandemic influenza response. In *2008 winter simulation conference* (pp. 1592–1600). IEEE.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv:1807.02811,.

Greenwood, P. E., & Gordillo, L. F. (2009). Stochastic epidemic modeling. In *Mathematical and statistical estimation approaches in epidemiology* (pp. 31–52). Springer.

Gülpınar, N., Çanakoğlu, E., & Branke, J. (2018). Heuristics for the stochastic dynamic task-resource allocation problem with retry opportunities. *European Journal of Operational Research, 266*(1), 291–303.

Han, S., Preciado, V. M., Nowzari, C., & Pappas, G. J. (2015). Data-driven network resource allocation for controlling spreading processes. *IEEE Transactions on Network Science and Engineering, 2*(4), 127–138.

Han, W., & Powell, W. B. (2020). Optimal online learning for nonlinear belief models using discrete priors. *Operations Research, 68*(5), 1538–1556. https://doi.org/10.1287/opre.2019.1921.

Imani, M., & Ghoreishi, S. F. (2020). Bayesian optimization objective-based experimental design. *American Control Conference*, 3405–3411.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, 115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118.

Köhler, J., Schwenkel, L., Koch, A., Berberich, J., Pauli, P., & Allgöwer, F. (2020). Robust and optimal predictive control of the covid-19 outbreak. *Annual Reviews in Control.*

Li, H., & Womer, N. K. (2015). Solving stochastic resource-constrained project scheduling problems by closed-loop approximate dynamic programming. *European Journal of Operational Research, 246*(1), 20–33.

Lin, Q., Zhao, Q., & Lev, B. (2020). Cold chain transportation decision in the vaccine supply chain. *European Journal of Operational Research, 283*(1), 182– 195. https://doi.org/10.1016/j.ejor.2019.11.005.

Martin, C., Allen, L., Stamp, M., Jones, M., & Carpio, R. (1993). A model for the optimal control of a measles epidemic. In *Computation and control III: Proceedings of the third Bozeman conference, Bozeman, Montana, August 5–11, 1992* (pp. 265–283). https://doi.org/10.1007/978-1-4612-0321-6_20.

Martins-Filho, P. R. (2021). Relationship between population density and covid-19 incidence and mortality estimates: A county-level analysis. *Journal of Infection and Public Health, 14*(8), 1087–1088. https://doi.org/10.1016/j.jiph.2021.06.018.

Morato, M. M., Pataro, I. M., da Costa, M. V. A., & Normey-Rico, J. E. (2020). A parametrized nonlinear predictive control strategy for relaxing covid-19 social distancing measures in brazil. *ISA Transactions.*

Neilan, R. M., & Lenhart, S. (2011). Optimal vaccine distribution in a spatiotemporal epidemic model with an application to rabies and raccoons. *Journal of Mathematical Analysis and Applications, 378*(2), 603– 619. https://doi.org/10.1016/j.jmaa.2010.12.035.

Nguyen, C., & Carlson, J. M. (2016). Optimizing real-time vaccine allocation in a stochastic sir model. *PLoS One, 11*(4), e0152950.

Osorio, A. F., Brailsford, S. C., & Smith, H. K. (2018). Whole blood or apheresis donations? A multi-objective stochastic optimization approach. *European Journal of Operational Research, 266*(1), 193–204. https://doi.org/10.1016/j.ejor.2017.09.005.

Packwood, D. (2017). *Bayesian optimization for materials science*. Springer.

Porco, T. C., Holbrook, K. A., Fernyak, S. E., Portnoy, D. L., Reiter, R., & Aragón, T. J. (2004). Logistics of community smallpox control through contact tracing and ring vaccination: A stochastic network model. *BMC Public Health, 4*(1), 1–20.

Powell, W. (2019). A unified framework for stochastic optimization. *European journal of operational research, 275*(3), 795–821. https://doi.org/10.1016/j.ejor.2018.07.014.

Powell, W. B. (2011). *Approximate dynamic programming: Solving the curses of dimensionality*. Draft

Probert, W. J., Jewell, C. P., Werkman, M., Fonnesbeck, C. J., Goto, Y., Runge, M. C., … Ferrari, M. J., et al. (2018). Real-time decision-making during emergency disease outbreaks. *PLoS Computational Biology, 14*(7), e1006202.

Reyes, K., & Powell, W. B. (2020). Optimal learning for sequential decisions in laboratory experimentation. arXiv:2004.05417,.

Sélley, F., Besenyei, Á., Kiss, I. Z., & Simon, P. L. (2015). Dynamic control of modern, network-based epidemic models. *SIAM Journal on Applied Dynamical Systems, 14*(1), 168–187.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE, 104*(1), 148–175.

Shea, K., Tildesley, M. J., Runge, M. C., Fonnesbeck, C. J., & Ferrari, M. J. (2014). Adaptive management and the value of information: Learning via intervention in epidemiology. *PLoS Biology, 12*(10), e1001970.

Simunaci, L. (2020). Pro-rata vaccine distribution is fair, equitable.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, L., Zhou, Y., Wang, L., Purkayastha, S., Zhang, L., He, J., … Song, P. X.-K. (2020). A review of multi-compartment infectious disease models. *International Statistical Review, 88*(2), 462–513.

Tanner, M. W., & Ntaimo, L. (2010). Iis branch-and-cut for joint chance-constrained stochastic programs and application to optimal vaccine allocation. *European Journal of Operational Research, 207*(1), 290–296.

Tanner, M. W., Sattenspiel, L., & Ntaimo, L. (2008). Finding optimal vaccination strategies under parameter uncertainty using stochastic programming. *Mathematical Biosciences, 215*(2), 144–151. https://doi.org/10.1016/j.mbs.2008.07.006.

US Census Bureau (2020). 2020 census data. https://data.census.gov/cedsci/.

Wang, Y., & Powell, W. (2016). An optimal learning method for developing personalized treatment regimes. arXiv:1607.01462,.

Watkins, N. J., Nowzari, C., & Pappas, G. J. (2019). Robust economic model predictive control of continuous-time epidemic processes. *IEEE Transactions on Automatic Control, 65*(3), 1116–1131.

Yarmand, H., Ivy, J. S., Denton, B., & Lloyd, A. L. (2014). Optimal two-phase vaccine allocation to geographically different regions under uncertainty. *European Journal of Operational Research, 233*(1), 208–219.

Zakary, O., Rachik, M., & Elmouki, I. (2017). On the analysis of a multi-regions discrete sir epidemic model: An optimal control approach. *International Journal of Dynamic Control, 5*, 917–930. https://doi.org/10.1007/s40435-016-0233-2.

Zhang, Y., & Prakash, B. A. (2014). Scalable vaccine distribution in large graphs given uncertain data. In *Proceedings of the 23rd ACM international conference on information and knowledge management* (pp. 1719–1728). https://doi.org/10.1145/2661829.2662088.