Supplementary materials for this article are available at https://doi.org/10.1007/s13253-023-00551-4.

Deep Spatial Q-Learning for Infectious Disease Control

Zhishuai LIU, Jesse CLIFTON, Eric B. LABER, John DRAKE, and Ethan X. FANG

Infectious diseases are a cause of humanitarian and economic crises across the world. In developing regions, a severe epidemic can result in the collapse of healthcare infrastructure or even the failure of an affected state. The most recent 2013–2015 outbreak of Ebola virus disease in West Africa is an example of such an epidemic. The economic, infrastructural, and human costs of this outbreak provide strong motivation for the examination of adaptive treatment strategies that allocate resources in response to and anticipation of the evolution of an epidemic. We formalize adaptive management of an emerging infectious disease spreading across a set of locations as a treatment regime that maps up-to-date information on the epidemic to a subset of locations identified as high-priority for treatment. An optimal treatment regime in this context is defined as maximizing the expectation of a pre-specified cumulative utility measure, e.g., the number of disease-free individuals or the estimated reduction in morbidity or mortality relative to a baseline intervention strategy. Because the disease dynamics are not known at the beginning of an outbreak, an optimal treatment regime must be estimated online, i.e., as data accumulate; thus, an effective estimation algorithm must balance choosing interventions that lead to information gain and thereby model improvement with interventions that appear to be optimal under the current estimated model. We develop a novel model-free algorithm for the online management of an infectious disease spreading over a finite set of locations and an indefinite or infinite time horizon. The proposed algorithm balances exploration and exploitation using a semi-parametric variant of Thompson sampling. We also introduce a graph neural network-based estimator in order to improve the performance of this class of algorithms. Simulations, including those mimicking the spread of the 2013–2015 Ebola outbreak, suggest that an adaptive treatment strategy has the potential to significantly reduce mortality relative to *ad hoc* management strategies.

Supplementary materials accompanying this paper appear online.

Key Words: Infectious diseases; Reinforcement learning; Graph neural networks .

Liu and Clifton implemented and evaluated algorithms and wrote initial drafts of the manuscript. Fang and Laber directed research and edited drafts. Drake led scientific development related to infectious disease modeling and application to Ebola.

Z. Liu, Department of Statistical Science, Duke University, Durham, NC, USA.
J. Clifton, Department of Statistics, NC State University, Raleigh, NC, USA.
E. B. Laber (✉), Department of Statistical Science, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA (E-mail: *eric.laber@duke.edu*).
J. Drake, School of Ecology, University of Georgia, Athens, GA, USA.
E. X. Fang, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-023-00551-4

Published online: 08 July 2023

1. INTRODUCTION

Infectious diseases are a persistent and serious threat to public health worldwide (Mathers 2008; Lozano et al. 2013; Bloom and Cadarette 2019). Despite technological advances and increasingly vigilant biosurveillance, global rates of infectious diseases are not decreasing (Smith et al. 2014). An effective real-time intervention strategy for an emerging infectious disease could have significant benefits, including reduction of mortality, morbidity, and healthcare costs. Consequently, the development of such a strategy is a priority for public health and security policy-makers (Cecchine and Moore 2006). We formalize such an intervention system as a treatment regime that maps the current status of the epidemic to a subset of locations identified as high-priority for treatment. An optimal treatment regime maximizes the mean of a pre-specified cumulative utility measure, e.g., the number of disease-free individuals throughout the epidemic.

Our goal is to find an optimal treatment regime for the management of emerging infectious diseases that, given the current outbreak information at different locations and resource constraints, identifies which locations should be prioritized for treatment. This problem is complicated by the following three issues: (i) spillover effects make the number of possible interventions an exponential function of the number of treatment units; i.e., treatment at one location can affect outcomes at other locations, so one must consider the joint treatment allocation across all locations. (ii) Disease dynamics are unknown at the time of the outbreak, so one must balance choosing interventions that lead to a significant information gain and subsequently an improved disease dynamic model, with choosing interventions that appear to be optimal based on current model estimates. (iii) Resource constraints impose additional restrictions on how and where interventions can be applied. One approach to estimating an optimal treatment regime is to posit a model for the disease dynamics and then to use simulation-based optimization to estimate an optimal treatment regime (Carr and Roberts 2010; Kasaie and Kelton 2013; Nowzari et al. 2015; Hu et al. 2017; Laber et al. 2018a; Kompella et al. 2020; Guan et al. 2022). If the posited model is low-dimensional and accurately reflects the disease process, this approach can be particularly effective early in the epidemic when data are scarce. However, such methods can perform poorly if the posited model is misspecified. An alternative is to construct a semi-parametric estimator of the optimal treatment regime that does not require a correctly specified dynamic model; examples of such estimators in non-spatiotemporal domains include regression-based estimators (Murphy 2005; Henderson et al. 2010; Almirall et al. 2010; Zhao et al. 2011; Chakraborty and Moodie 2013; Schulte et al. 2014; Moodie et al. 2014; Kosorok and Moodie 2015; Laber et al. 2017; Ertefaie et al. 2021) and direct-search estimators (Orellana et al. 2010; Rubin and van der Laan 2012; Zhang et al. 2012; Zhao et al. 2012; Zhang et al. 2013, 2015; Zhao et al. 2015; Zhou et al. 2017; Liu et al. 2018; Pan and Zhao 2020). Thus, a natural approach is to apply a parametric simulation-optimization approach during the early stages of an epidemic, and subsequently migrate to a semi-parametric estimator as data accumulates. Our goal is to develop a class of online semi-parametric estimators that can be used in such a strategy. The class estimators that we propose is based on fitted Q-iteration (FQI; Watkins 1989; Maei et al. 2010; Ernst et al. 2005; Ertefaie 2014; Riedmiller 2005) and Thompson sampling (Thompson 1933). A key challenge associated with extending reinforcement learning algo-

rithms to spatio-temporal decision problems is using spatial information in a statistically and computationally efficient manner. To this end, we propose an automated (i.e., data-driven) feature construction algorithm based on graph neural networks (Yan et al. 2006; Cai et al. 2018; Fey and Lenssen 2019; Ma et al. 2020) that succinctly summarizes local information which is then used in FQI. Our method is a variant of single-agent deep reinforcement learning, which has achieved great empirical successes in the past decade (Riedmiller 2005; Mnih et al. 2015, 2016; Arulkumaran et al. 2017). We note that our setup is related to but distinct from the cooperative multi-agent reinforcement learning problem (Sunehag et al. 2017; Wang et al. 2020; Hernandez-Leal et al. 2019). Whereas the cooperative multi-agent setting, involves a series agents learning locally, the optimal treatment regime in our setting is centralized with treatments being coordinated jointly across locations.

This work is motivated by our involvement in a retrospective study of the 2013–2015 outbreak of Ebola virus disease in West Africa (Kramer et al. 2016a; Li et al. 2017). The Ebola outbreak resulted in more than 10,000 deaths and the near-total collapse of healthcare infrastructure in affected areas (WHO Ebola Response Team 2014; Hamel and Slutsker 2015). We consider the daily allocation of treatments across 290 contiguous geopolitical regions. Our goal is to learn an optimal treatment regime that can be used to control future outbreaks by studying if and how the spread of the 2013–2015 outbreak could have been better controlled through adaptive treatment allocations subject to resource constraints. Both simulation and real data analysis results indicate that management strategies based on the proposed method can lead to significant reductions in the spread of the disease over *ad hoc* allocation strategies.

The rest of this paper is organized as follows. In Sect. 2, we review the 2013–2015 outbreak of Ebola virus disease. In Sect. 3, we define an optimal treatment regime under the framework of potential outcomes when the data-generating model forms a Markov decision process. In Sect. 4, we define a spatial FQI with graph embeddings and a semi-parametric variant of Thompson sampling. In Sect. 5, we review model-based policy search (Laber et al. 2018a) for spatio-temporal problems. We illustrate the proposed methods using a suite of simulation experiments in Sect. 6 and a simulation of the spread of Ebola in West Africa in Sect. 7. Open problems are discussed in Sect. 8.

2. EBOLA VIRUS

Ebola Virus Disease (EVD) is an acute hemorrhagic illness caused by a handful of viruses collectively known as the ebolaviruses. The 2013-2015 West Africa Ebola epidemic, caused by the *Zaire ebolavirus*, originated in the Guéckédou Prefecture of Guinea, from which it spread to neighboring Liberia and Sierra Leone. A major outbreak resulted in more than 28,000 cases ensued, ignited small outbreaks in Nigeria, Mali, and the United States. Patients with EVD may exhibit a range of symptoms, including fever, muscular pain, vomiting, diarrhea, rash, organ failure, and death (Feldmann and Geisbert 2011). The overall case fatality rate of the West Africa epidemic exceeds 39%. Person-to-person transmission of Ebola is typically by exposure to infected body fluids. Although infectious cases are typically symptomatic, Ebola is difficult to contain without adequate and quickly

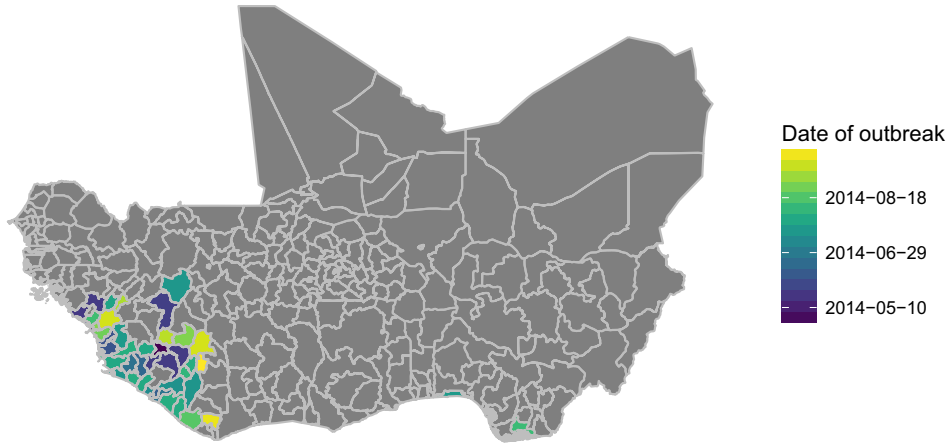

Figure 1. Observed outbreaks for West Africa with the first infections on April 26, 2014.

implemented infection control procedures. Additionally, the social disruption caused by Ebola outbreaks can make the scope of incipient outbreaks challenging to determine.

Several models for the spread of Ebola in West Africa have been constructed. Rainsch et al. (2015) fit regression models to weekly data on the incidence of infection. They found that the case counts, population data, and distances between affected and unaffected areas were significant predictors of transmission. Merler et al. (2015) developed an agent-based simulator that accounts for the early decline of spread in terms of the increasing availability of Ebola treatment units, safe burials, and distribution of household protection kits. Finally, Kramer et al. (2016a) fit a coarse-grained gravity model to understand how the spread of the infection to new areas was affected by the attributes of donor and recipient regions. Their model considered only the first infection in a region to be of interest, and thus focused on the spread path. They found that the spread was best explained by the distances between source and recipient locations, population density, and border closures among neighboring countries (Figs. 1, 2 and 3).

3. MODEL

We consider a decision process evolving in discrete time, $t = 0, 1, \ldots$, and across a finite set of locations $\mathcal{L} = \{1, \ldots, L\}$. In our application to EVD, the time points correspond to days, and the locations are geopolitical units. We assume that at each location ℓ and each time point t, the following sequence of events transpires:

(i) A set of measurements is taken; these, together with past measurements, are summarized as the current state of the location, $\boldsymbol{S}_t^\ell \in \mathcal{S}_0 \subseteq \mathbb{R}^m$.

(ii) The decision maker selects a binary treatment $A_t^\ell \in \mathcal{A}_0 = \{0, 1\}$. Thus, without any restrictions, there are 2^L possible treatment allocations at each time point.

(iii) We observe an outcome $Y_t \in \mathcal{Y}_0$.

Let $\mathcal{S} = \mathcal{S}_0^L$, $\mathcal{A} = \mathcal{A}_0^L$, and $\mathcal{Y} = \mathcal{Y}_0^L$. Define $\boldsymbol{S}_t = \left(\boldsymbol{S}_t^1, \boldsymbol{S}_t^2, ..., \boldsymbol{S}_t^L\right)^\top \in \mathcal{S} \subseteq \mathbb{R}^{m \times L}$, $\boldsymbol{A}_t = \left(A_t^1, A_t^2, \cdots, A_t^L\right)^\top \in \mathcal{A} = \{0, 1\}^L$, and $\boldsymbol{Y}_t = \left(Y_t^1, Y_t^2, \cdots, Y_t^L\right)^\top \in \mathcal{Y}$. We assume that $\boldsymbol{Y}_t$ is contained in $\boldsymbol{S}_{t+1}$ (this is without any loss of generality as the state can always be expanded to ensure that this is the case). Let $2^{\mathcal{A}}$ denote the power set of binary vectors $\mathcal{A}$ of length L. We assume that there exists a function $\psi : \mathcal{S} \to 2^{\mathcal{A}} \setminus \emptyset$ such that $\psi(\boldsymbol{s})$ denotes the set of feasible treatments when the state is $\boldsymbol{s}$. Further, we assume that there exists a function $u_0 : \mathcal{Y}_0 \to \mathbb{R}$ such that $U_t^\ell = u_0(\boldsymbol{Y}_t^\ell)$ is the utility with location ℓ and time t, and the total utility at time t is

$$u(\boldsymbol{Y}_t) = \sum_{\ell=1}^{L} U_t^\ell.$$

The assumption of an additive utility model is a mild restriction in application as common measures are aggregated across locations in this way, e.g., number of infected locations etc.

A treatment regime in this context is a map $\pi : \mathcal{S} \to \mathcal{A}$ which satisfies $\pi(\boldsymbol{s}) \in \psi(\boldsymbol{s})$ for all $\boldsymbol{s} \in \mathcal{S}$. Define $\pi^\ell(\boldsymbol{s})$ as the ℓ-th element of $\pi(\boldsymbol{s})$. A decision maker following π applies treatment $\pi(\boldsymbol{s}_t)$, i.e., applies $\pi^\ell(\boldsymbol{s}_t)$ to location $\ell \in \mathcal{L}$, if presented with $\boldsymbol{S}_t = \boldsymbol{s}_t$ at time t. An optimal treatment regime maximizes the mean discounted utility if used to select treatments at each time point. We formalize this definition using potential outcomes (Rubin 1974; Robins 1986, 1987; Splawa-Neyman et al. 1990; Tsiatis et al. 2019). Let $\overline{\boldsymbol{a}}_t = (\boldsymbol{a}_0, \ldots, \boldsymbol{a}_t)$ and $\overline{\boldsymbol{s}}_t = (\boldsymbol{s}_0, \ldots, \boldsymbol{s}_t)$ denote the history up to time t. Let $\boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1})$ be the potential state under treatment sequence $\overline{\boldsymbol{a}}_{t-1}$ where $\boldsymbol{S}_0(\overline{\boldsymbol{a}}_{-1}) \equiv \boldsymbol{S}_0$, and let $\boldsymbol{Y}_t^*(\overline{\boldsymbol{a}}_t)$ be the potential outcome under treatment sequence $\overline{\boldsymbol{a}}_t$. The potential state at time t under a regime π is $\boldsymbol{S}_t^*(\pi) = \sum_{\overline{\boldsymbol{a}}_{t-1}} \boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1}) \prod_{v=0}^{t-1} \mathbb{1}_{\pi\{\boldsymbol{S}_v^*(\overline{\boldsymbol{a}}_{v-1})\}=a_v}$, where $\mathbb{1}_\nu$ is an indicator that evaluates to one if the clause ν is true and zero otherwise. Similarly, the potential outcome at time t is $\boldsymbol{Y}_t^*(\pi) = \sum_{\overline{\boldsymbol{a}}_t} \boldsymbol{Y}_t^*(\overline{\boldsymbol{a}}_t) \prod_{v=0}^{t} \mathbb{1}_{\pi\{\boldsymbol{S}_v^*(\overline{\boldsymbol{a}}_{v-1})\}=a_v}$. Letting $\gamma \in (0, 1)$ be a fixed discount factor, the value of treatment regime π is $V(\pi) = \mathbb{E}\left[\sum_{t \geqslant 0} \gamma^t u\left\{\boldsymbol{Y}_t^*(\pi)\right\}\right]$ so that the optimal regime, π^{opt}, satisfies $V(\pi^{\text{opt}}) \geqslant V(\pi)$ for all π. Note that our framework can be extended to other measures of cumulative utility, e.g., average utility or total utility over a finite horizon (see Linn et al. 2017; Wang et al. 2018; Rowland et al. 2019, for additional discussion in non-spatial applications).

Let $\boldsymbol{W}^* = \left\{\boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1}), Y_t^*(\overline{\boldsymbol{a}}_t) : \overline{\boldsymbol{a}}_t \in \{0, 1\}^{L \times (t+1)}\right\}_{t \geqslant 0}$ denote the set of potential states and outcomes. To identify π^{opt} in terms of the data generating model, we impose a series of assumptions which are standard in the dynamic treatment regimes literature (Murphy 2003; Robins 2004).

Assumption 1. We assume that for all t:
(A1) Consistency: $\boldsymbol{S}_t = \boldsymbol{S}_t^*(\overline{\boldsymbol{A}}_{t-1})$ and $\boldsymbol{Y}_t = \boldsymbol{Y}_t^*(\overline{\boldsymbol{A}}_t)$.
(A2) Positivity: $P(\boldsymbol{A}_t = \boldsymbol{a}_t | \overline{\boldsymbol{S}}_t = \overline{\boldsymbol{s}}_t, \overline{\boldsymbol{A}}_{t-1} = \overline{\boldsymbol{a}}_{t-1}) > 0$ for all $\overline{\boldsymbol{s}}_t, \overline{\boldsymbol{a}}_{t-1}$ with $\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)$.
(A3) Strong ignorability: $\boldsymbol{A}_t \perp \boldsymbol{W}^* | \overline{\boldsymbol{S}}_t, \overline{\boldsymbol{A}}_{t-1}$.

In the context of online estimation, where treatment assignment is under the control of the decision maker, (A2) and (A3) can be ensured by construction. We note that these assumptions may not hold if the decision maker deviates from the recommendation of the

estimated regime. We discuss this point further in Sect. 8. Hereafter, we assume that (A1)-(A3) hold implicitly.

In addition, it is standard in the context of dynamic treatment regimes to assume that there are independent replicates, e.g., patients in a study, that make the optimal regime nonparametrically identifiable. However, because of spatial interference (Karwa and Airoldi 2018; Tec et al. 2022; Forastiere et al. 2021), one cannot treat the locations as independent and consequently additional structure must be imposed on the model to identify π^{opt} (Hudgens and Halloran 2008; Laber et al. 2018a). We assume that the decision process, possibly transformed, is Markov and, under this assumption, impose a semi-parametric model on the conditional mean discounted utility given state and treatment. These modeling assumptions are standard in problems with an infinite or indefinite time horizon (Powell 2007; Szepesvári 2010; Hernández-Lerma and Lasserre 2012; Puterman 2014; Sutton and Barto 2018). In particular, we assume that the states $\boldsymbol{S}_t$ have been constructed so that the induced decision process is Markov, i.e., $P\left(\boldsymbol{S}_{t+1} \in \mathcal{B} \middle| \overline{\boldsymbol{S}}_t, \overline{\boldsymbol{A}}_t\right) = P\left(\boldsymbol{S}_{t+1} \in \mathcal{B} \middle| \boldsymbol{S}_t, \boldsymbol{A}_t\right)$ with probability one for any (measurable) set $\mathcal{B} \subseteq \mathcal{S}$, and this probability does not depend on the time t.

4. SPATIAL FITTED-Q ITERATION

Under Assumption 1, the optimal treatment regime can be characterized using a recursive regression equation known as the Bellman optimality equation (Bellman 1957; Maei et al. 2010; Hernández-Lerma and Lasserre 2012; Puterman 2014; Ertefaie and Strawderman 2018). For any $\boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{a} \in \psi(\boldsymbol{s})$, define

$$Q(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}\left[u(Y_t) + \sum_{k \geqslant 1} \gamma^k u\left\{Y^*_{t+k}(\pi^{\text{opt}})\right\} \middle| \boldsymbol{S}_t = \boldsymbol{s}, \boldsymbol{A}_t = \boldsymbol{a}\right],$$

so that $Q(\boldsymbol{s}, \boldsymbol{a})$ is the expected cumulative discounted utility starting state $\boldsymbol{S}_t = \boldsymbol{s}$, applying treatment $\boldsymbol{A}_t = \boldsymbol{a}$, and then following the optimal regime thereafter. It can be shown that $\pi^{\text{opt}}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a} \in \psi(\boldsymbol{s})} Q(\boldsymbol{s}, \boldsymbol{a})$ (Bertsekas et al. 1995). Furthermore, the Q-function $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ satisfies

$$Q(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}\left\{u\left(\boldsymbol{Y}_t\right) + \gamma \max_{\boldsymbol{a}_{t+1} \in \psi(\boldsymbol{S}_{t+1})} Q\left(\boldsymbol{S}_{t+1}, \boldsymbol{a}_{t+1}\right) \middle| \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right\}, \tag{1}$$

which, importantly, identifies the Q-function in terms of the data generating model. One approach to constructing an estimator of $Q(\boldsymbol{s}, \boldsymbol{a})$, and subsequently π^{opt}, is to use (1) to construct an estimating function to which $Q(\boldsymbol{s}, \boldsymbol{a})$ is a unique solution. For example, one might posit a linear model of the form $Q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\beta}) = \phi(\boldsymbol{s}, \boldsymbol{a})^{\intercal}\boldsymbol{\beta}$, where $\phi(\boldsymbol{s}, \boldsymbol{a})$ is a known feature vector, and $\boldsymbol{\beta}$ is a vector of unknown coefficients; in this case, if there exists $\boldsymbol{\beta}^*$ such that $Q(\boldsymbol{s}, \boldsymbol{a}) \equiv Q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\beta}^*)$, then it follows under mild conditions that $\boldsymbol{\beta}^*$ is a solution to

$$\mathbb{E}\left[\left\{u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}_{t+1} \in \psi(\boldsymbol{S}_{t+1})} Q\left(\boldsymbol{S}_{t+1}, \boldsymbol{a}_{t+1}; \boldsymbol{\beta}\right) - Q\left(\boldsymbol{S}_t, \boldsymbol{A}_t; \boldsymbol{\beta}\right)\right\} \phi(\boldsymbol{S}_t, \boldsymbol{A}_t)\right] = 0,$$

which can be used to construct an estimator $\widehat{\boldsymbol{\beta}}_t$ of $\boldsymbol{\beta}^*$ given data observed through time t (see Ertefaie and Strawderman 2018; Luckett et al. 2020; Saghafian 2021).

The estimating equation approach to constructing an estimator of $Q(\boldsymbol{s}, \boldsymbol{a})$ is elegant due to its simplicity and direct connection to the Bellman optimality equations. However, it is not without drawbacks. One such drawback is that the form of the estimating equation requires limiting the class of models for $Q(\boldsymbol{s}, \boldsymbol{a})$ to those which are amenable to root-solving via standard software, such as smooth parametric models, or writing bespoke root-finding algorithms. A second, potentially more important drawback, is that the estimating equation formulation does not permit interactive model-building in the same way typical supervised learning does. This is critical when the model is being used to inform public health decisions in a pandemic. Thus, instead of considering the estimating equation characterization to construct an estimator, we use the fact that $Q(\boldsymbol{s}, \boldsymbol{a})$ is the fixed point of the so-called Bellman optimality operator to express the optimal policy as the limit of a series of regressions.

Define the Bellman optimality operator B, which acts on functions $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, as $(\mathrm{B}w)(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}\Big\{u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}_{t+1} \in \psi(\boldsymbol{S}_{t+1})} w(\boldsymbol{S}_{t+1}, \boldsymbol{a}_{t+1}) \mid \boldsymbol{S}_t = \boldsymbol{s}, \boldsymbol{A}_t = \boldsymbol{a}\Big\}$. The Bellman optimality equation is thus succinctly expressed as $Q = \mathrm{B}Q$. Moreover, Q is the unique fixed point of the operator B, which is a contraction in the sup-norm. The associated fixed-point iteration algorithm known as *value iteration* is given by the update $Q_k \leftarrow BQ_{k-1}$ for $k \geqslant 1$, where the initial value, Q_0 is initialized arbitrarily; under mild regularity conditions, Q_k converges geometrically to Q (Bertsekas and Tsitsiklis 1996). This characterization leads to an iterative algorithm for the semi-parametric estimation of the optimal Q-function given observations from the underlying decision process.

4.1. FITTED Q-ITERATION

Fitted Q-iteration is a regression-based approximation of the value iteration algorithm (Ernst et al. 2005; Riedmiller 2005; Busoniu et al. 2010). We first describe the algorithm without exploiting any underlying spatial structure. Let $\mathcal{Q}$ denote the posited class of the Q-functions. At time t, let $\widehat{Q}_{t,0} \equiv 0$, and for $k \geqslant 1$, let

$$\widehat{Q}_{t,k} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{v=0}^{t-1} \left\{ u\left(\boldsymbol{Y}_v\right) + \gamma \max_{\boldsymbol{a}_{v+1} \in \psi(\boldsymbol{S}_{v+1})} \widehat{Q}_{t,k-1}\left(\boldsymbol{S}_{v+1}, \boldsymbol{a}_{v+1}\right) - Q(\boldsymbol{S}_v, \boldsymbol{A}_v) \right\}^2, \quad (2)$$

which can be viewed as an application of approximate value iteration in which the empirical distribution has been used in place of the true expectation. The estimated optimal regime is $\widehat{\pi}_t(\boldsymbol{s}) = \operatorname{argmax}_{\boldsymbol{a} \in \psi(\boldsymbol{s})} \widehat{Q}_{t,K}(\boldsymbol{s}, \boldsymbol{a})$, where K is the desired number of iterations. Note that the least squares estimator in (2) can be constructed using essentially any regression estimator, e.g., trees, neural networks, Gaussian Process models, and so on. Furthermore, these models can be built interactively at each step to avoid severe misspecification.

When the process under study is a disease spreading across a network, the number of locations may far exceed the number of time points, e.g., if L are individuals in social network. Furthermore, the effects of treatment are likely to be local in that the greatest impact will be on those to whom treatment is applied and their close contacts. With this in mind, we now show how the generic fitted Q-iteration algorithm can be extended to

pool information across locations and thus increase efficiency. We do this by showing that the Q-functions defined in the preceding section, $\{Q_k\}_{k\geqslant 0}$, can be expressed as sums of 'location-specific Q-functions' which can be estimated using local data thereby increasing the number of observations in (2) from t to $t \times L$.

The first (non-trivial) Q-function is

$$\begin{aligned} Q_1(\boldsymbol{s},\boldsymbol{a}) &= \mathbb{E}\left\{u(\boldsymbol{Y}_t)\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right\} \\ &= \sum_{\ell=1}^{L}\mathbb{E}\left\{U_t^{\ell}\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right\} \\ &= \sum_{\ell=1}^{L} q_1^{\ell}(\boldsymbol{s},\boldsymbol{a}), \end{aligned}$$

where $q_1^{\ell}(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}(U_t^{\ell}|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a})$ is the first local Q-function. Define $\pi_1^{\text{opt}}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})} Q_1(\boldsymbol{s},\boldsymbol{a})$ and $\nu_1^{\ell}(\boldsymbol{s}) = q_1^{\ell}\left\{\boldsymbol{s},\pi_1^{\text{opt}}(\boldsymbol{s})\right\}$. The second Q-function is

$$\begin{aligned} Q_2(\boldsymbol{s},\boldsymbol{a}) &= \mathbb{E}\left\{u(\boldsymbol{Y}_t)+\gamma\max_{\boldsymbol{a}_{t+1}\in\psi(\boldsymbol{S}_{t+1})} Q_1(\boldsymbol{S}_{t+1},\boldsymbol{a}_{t+1})\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right\} \\ &= \sum_{\ell=1}^{L}\mathbb{E}\left\{U_t^{\ell}+\gamma\nu_1^{\ell}(\boldsymbol{S}_{t+1})\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right\} = \sum_{\ell=1}^{L} q_2^{\ell}(\boldsymbol{s},\boldsymbol{a}), \end{aligned}$$

where $q_2^{\ell}(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\left\{U_t^{\ell}+\gamma\nu_1^{\ell}(\boldsymbol{S}_{t+1})\right\}$. For $k\geqslant 1$, define $\pi_k^{\text{opt}}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})} Q_k(\boldsymbol{s},\boldsymbol{a})$ and $\nu_k^{\ell}(\boldsymbol{s}) = q_k^{\ell}\left\{\boldsymbol{s},\pi_k^{\text{opt}}(\boldsymbol{s})\right\}$, and let $q_{k+1}^{\ell}(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}\left\{U_t^{\ell}+\gamma\nu_k^{\ell}(\boldsymbol{S}_{t+1})\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right\}$, so we have

$$Q_{k+1}(\boldsymbol{s},\boldsymbol{a}) = \sum_{\ell=1}^{L} q_{k+1}^{\ell}(\boldsymbol{s},\boldsymbol{a}),$$

which is a sum over local Q-functions $q_{k+1}^{\ell}(\boldsymbol{s},\boldsymbol{a})$'s. We note that the Q-function in equation (1) corresponds to Q_{∞} here.

Let $\boldsymbol{\Phi}$ be a class of maps from $\mathcal{S}\times\mathcal{A}$ into $\mathbb{R}^J$ such that for any $\boldsymbol{\phi}\in\boldsymbol{\Phi}$, we have $\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{a}) = \left\{\boldsymbol{\phi}^1(\boldsymbol{s},\boldsymbol{a}),\ldots,\boldsymbol{\phi}^L(\boldsymbol{s},\boldsymbol{a})\right\}$, where $\boldsymbol{\phi}^{\ell}(\boldsymbol{s},\boldsymbol{a})\in\mathbb{R}^J$ is a feature vector for location ℓ constructed from $(\boldsymbol{s},\boldsymbol{a})$. Let $\mathfrak{Q}$ be a class of maps from $\mathbb{R}^J$ into $\mathbb{R}$; choices for $\boldsymbol{\Phi}$ and $\mathfrak{Q}$ are discussed in the next section. We posit working models for $q_k^{\ell}(\boldsymbol{s},\boldsymbol{a})$ of the form $q_k^{\ell}(\boldsymbol{s},\boldsymbol{a}) = q_k\left\{\boldsymbol{\phi}^{\ell}(\boldsymbol{s},\boldsymbol{a})\right\}$, where $q_k\in\mathfrak{Q}$ and $\boldsymbol{\phi}\in\boldsymbol{\Phi}$. The fitted Q-iteration algorithm with this class of models is thus comprised of the following steps. At time t, the first (non-trivial) iteration of the fitted-Q algorithm is

$$(\widehat{\boldsymbol{\phi}}_{t,1},\widehat{q}_{t,1}) = \underset{(\boldsymbol{\phi},q)\in\boldsymbol{\Phi}\times\mathfrak{Q}}{\text{argmin}}\sum_{v=1}^{t}\sum_{\ell=1}^{L}\left[U_v^{\ell}-q\left\{\boldsymbol{\phi}^{\ell}(\boldsymbol{S}_v,\boldsymbol{A}_v)\right\}\right]^2,$$

so that $\widehat{q}^{\ell}_{t,1}(\boldsymbol{s},\boldsymbol{a}) = \widehat{q}_{t,1}\left\{\widehat{\boldsymbol{\phi}}^{\ell}_{t,1}(\boldsymbol{s},\boldsymbol{a})\right\}$. Define $\widehat{\pi}_{t,1}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})}\sum_{\ell=1}^{L}\widehat{q}^{\ell}_{t,1}(\boldsymbol{s},\boldsymbol{a})$ and $\widehat{v}^{\ell}_{t,1}(\boldsymbol{s}) = \widehat{q}^{\ell}_{t,1}\left\{\boldsymbol{s},\widehat{\pi}_{t,1}(\boldsymbol{s})\right\}$. For $k \geqslant 2$, the k-th iteration of the fitted-Q algorithm is

$$(\widehat{\phi}_{t,k},\widehat{q}_{t,k}) = \underset{(\boldsymbol{\phi},q)\in\boldsymbol{\Phi}\times\Omega}{\text{argmin}}\sum_{v=0}^{t-1}\sum_{\ell=1}^{L}\left[U^{\ell}_{v}+\gamma\widehat{v}^{\ell}_{t,k-1}(\boldsymbol{S}_{v+1})-q\left\{\boldsymbol{\phi}^{\ell}(\boldsymbol{S}_{v},\boldsymbol{A}_{v})\right\}\right]^{2},$$

so that $\widehat{q}^{\ell}_{t,k}(\boldsymbol{s},\boldsymbol{a}) = \widehat{q}_{t,k}\left\{\widehat{\boldsymbol{\phi}}^{\ell}_{t,k}(\boldsymbol{s},\boldsymbol{a})\right\}$, $\widehat{\pi}_{t,k}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})}\sum_{\ell=1}^{L}\widehat{q}^{\ell}_{t,k}(\boldsymbol{s},\boldsymbol{a})$, and $\widehat{v}^{\ell}_{t,k}(\boldsymbol{s}) = \widehat{q}^{\ell}_{t,k}\left\{\boldsymbol{s},\widehat{\pi}_{t,k}(\boldsymbol{s})\right\}$. After K iterations, estimated optimal regime is $\widehat{\pi}_{t}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})}\sum_{\ell=1}^{L}\widehat{q}^{\ell}_{t,K}(\boldsymbol{s},\boldsymbol{a})$.

4.1.1. Spatial Fitted Q-Iteration with Graph Embeddings

Graph embeddings have been widely studied and used to great empirical success in analyzing structured data such as text (e.g., Yan et al. 2006; Mikolov et al. 2013; Cai et al. 2018). They have also been used to estimate optimal policies in single-stage decision problems on networks (Ma et al. 2020). We consider a supervised approach in which the embeddings are adaptive, i.e., data-driven, to improve the quality of the estimated Q-functions. To the best of our knowledge, the construction used here is new and may be of independent interest.

Our goal is to construct a class of features $\boldsymbol{\Phi} : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^{L\times J}$ so that each $\boldsymbol{\phi}\in\boldsymbol{\Phi}$ creates a vector of local-summaries of $(\boldsymbol{s},\boldsymbol{a})$ $\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{a}) = \left\{\boldsymbol{\phi}^{1}(\boldsymbol{s},\boldsymbol{a}),\ldots,\boldsymbol{\phi}^{L}(\boldsymbol{s},\boldsymbol{a})\right\}$, one for each location, that is amenable to estimation as described in the preceding section. Recall that $\boldsymbol{s}\in\mathbb{R}^{m}$. Let $\mathcal{H}$ be a class of functions mapping $\mathbb{R}^{m+1}$ into $\mathbb{R}^{J}$, and let $\mathcal{G}$ be a class of functions mapping from $\mathbb{R}^{J}\times\mathbb{R}^{J}$ into $\mathbb{R}^{J}$. In our implementation, we consider each of these to be feed-forward neural networks (Bebis and Georgiopoulos 1994), though other choices are possible. For given $h\in\mathcal{H}$ and $g\in\mathcal{G}$, we construct the embedding of location ℓ at time t as follows. Define $f^{(1)} : \mathbb{R}^{m+1}\to\mathbb{R}^{J}$ as $f^{(1)}(\boldsymbol{b};g,h) = h(\boldsymbol{b})$ for all $\boldsymbol{b}\in\mathbb{R}^{m+1}$; thus, $f^{(1)}(\boldsymbol{s}^{\ell},a^{\ell};g,h) = h(\boldsymbol{s}^{\ell},a^{\ell})$ is a summary of the state and treatment at location ℓ. Let $f^{(2)} : \mathbb{R}^{J}\times\mathbb{R}^{J}\to\mathbb{R}^{J}$ be

$$f^{(2)}\left(\boldsymbol{b}^{1},\boldsymbol{b}^{2};g,h\right) = \frac{1}{2}g\left\{h(\boldsymbol{b}^{1}),\,h(\boldsymbol{b}^{2})\right\}+\frac{1}{2}g\left\{h(\boldsymbol{b}^{2}),h(\boldsymbol{b}^{1})\right\},$$

for all $\boldsymbol{b}^{1},\boldsymbol{b}^{2}\in\mathbb{R}^{m+1}$. Recursively, for $k\geqslant 2$, define $f^{(k)} : \bigotimes^{k}\mathbb{R}^{J}\to\mathbb{R}^{J}$ by

$$f^{(k)}(\boldsymbol{b}^{1},\ldots,\boldsymbol{b}^{k};g,h) = \frac{1}{k}\sum_{j=1}^{k}g\left\{h(\boldsymbol{b}^{j}),\;f^{(k-1)}(\boldsymbol{b}^{-j};g,h)\right\},$$

where $\boldsymbol{b}^{-j}$ is $\boldsymbol{b}^{1},\cdots,\boldsymbol{b}^{k}$ excluding $\boldsymbol{b}^{j}$. We consider all permutations to make the embedding invariant to the order in which locations are processed. Let $N^{\ell}\subseteq\{1,\ldots,L\}$ be a neighorhood of location ℓ, e.g., all m-order neighbors or the locations within some pre-specified distance. Given $h\in\mathcal{H}$, $g\in\mathcal{G}$, let

$$\phi^{\ell}(\boldsymbol{s},\boldsymbol{a};g,h) = g\left(h(\boldsymbol{s}^{\ell},a^{\ell}),\,f^{(|N^{\ell}|)}\left[\left\{(\boldsymbol{s}^{j},a^{j})\right\}_{j\in N^{\ell}}\right]\right).$$

If the set $\mathcal{N}^\ell$ is large, computing all possible permutations of its elements may be computationally infeasible. In this case, one can sample permutations uniformly at random. The collection of feature maps is thus

$$\mathbf{\Phi} = \left\{ \boldsymbol{\phi} = \left\{ \boldsymbol{\phi}^1(\cdot; g, h), \ldots, \boldsymbol{\phi}^L(\cdot; g, h) \right\} \, : \, g \in \mathcal{G}, \, h \in \mathcal{H} \right\}.$$

This construction can be understood as comprising the following steps: (i) computing a basis expansion $h \in \mathcal{H}$; (ii) recursively applying a binary operation, $g \in \mathcal{G}$, to this basis expansion along all orderings of elements in a given neighborhood of ℓ and taking the average of the results; and (iii) applying this binary operation to the features at ℓ and this average. This construction has a number of advantages. First, it can process a neighbor set of any size. Second, the learned feature vector of location ℓ is invariant to the ordering of its neighbors in $\mathcal{N}^\ell$. Finally, it considers the interaction of all neighbor pairs through the binary operation. We note that there are alternative constructions such as message passing graph neural networks which possess the first two properties (Lee et al. 2019; Fey and Lenssen 2019).

4.2. Generalized Bootstrap and Thompson Sampling

Methodologically, online sequential decision problems reside at the intersection of two disciplines: (i) sequential optimal design which focuses on choosing treatments to maximize information gain and thereby create high-quality estimated models (Chernoff 1972; Atwood 1973; Lai and Wei 1982; Bartroff et al. 2012); and (ii) dynamic programming which seeks to efficiently derive an optimal treatment regime using the estimated models (Bellman 1957; Powell 2007; Busoniu et al. 2010; Sutton and Barto 2018). As the primary goal is maximizing the discounted cumulative reward, the goal of many algorithms for online decision making is to experiment, i.e., deviate from the current estimated optimal treatment regime, only if and when such experimentation is likely to produce information that pays dividends in terms of long-term performance. The need to judiciously balance experimentation for information gain and optimizing for immediate utility gain is known as the *exploration-exploitation* trade-off in computer science. A common strategy to ensure sufficient information gain is to force exploration either through randomization or by maximizing at each step an objective function that includes an additional term for information gain (Pronzato 2000; Auer 2000; Russo and Van Roy 2014; Lattimore and Szepesvári 2020).

We consider a randomized treatment allocation strategy that can be viewed as a semi-parametric variant of the celebrated Thompson sampling. This work is based on resampling or purturbation in (non-spatial) decision problems (e.g., see Eckles and Kaptein 2014; Fortunato et al. 2017; Plappert et al. 2017; Osband et al. 2019). Thompson sampling (Thompson 1933) was originally proposed as a Bayesian approach to exploration wherein at each time point one draws a model from the posterior given current data and then selects the optimal regime assuming the selected model is correct (Scott 2010; Agrawal and Goyal 2011; Kaufmann et al. 2012; Agrawal and Goyal 2013; Korda et al. 2013; Gopalan et al. 2014; Hu et al. 2017). However, Thompson sampling cannot be directly applied in this form as we have not specified a model for the complete system dynamics; instead we use the more

general concept of a confidence distribution in which the sampling distribution is used as a kind of surrogate for the posterior and used to compute probabilities (confidence levels) of interest (Xie and Singh 2013). We use the fact that $\widehat{Q}_0, \widehat{Q}_1, \ldots, \widehat{Q}_K$ are M-estimators and thus use a multiplier bootstrap appropriate for Markov processes to estimate the sampling distributions of these estimators (Jin et al. 2001; Chatterjee and Bose 2005; Minnier et al. 2011). In particular, let $W_{t,0,1}, \ldots, W_{t,t-1,L}$ denote i.i.d. random variables each with unit mean and unit variance, and define

$$\widehat{q}^B_{t,k} = \underset{q\in\mathcal{Q}}{\operatorname{argmin}} \sum_{v=0}^{t-1}\sum_{\ell=1}^{L} W_{t,v,\ell}\left\{U^\ell_v + \gamma\,\hat{v}^\ell_{t,k-1}(\boldsymbol{s}_{v+1}) - q(\boldsymbol{\phi}^\ell(\boldsymbol{s}_v,\boldsymbol{a}_v))\right\}^2,$$

where $\widehat{q}^B_{t,0} \equiv 0$. Let $\widehat{Q}^B_{t,k} = \sum_{\ell=1}^{L}\widehat{q}^B_{t,k}$. The sampling distribution of $\widehat{Q}_{t,k}$ is thus estimated by the conditional distribution of $\widehat{Q}^B_{t,k}$ obtained via repeated draws of the weights $W_{t,0,1}, \ldots, W_{t,t-1,L}$. This induces a sampling distribution over the estimated regime $\widehat{\pi}_t(\boldsymbol{s}_t) = \operatorname{argmax}_{\boldsymbol{a}_t\in\psi(\boldsymbol{s}_t)}\widehat{Q}_{t,k}(\boldsymbol{s}_t,\boldsymbol{a}_t)$. Let $\widehat{Q}^B_{t,k}$ denote the bootstrap analog of $\widehat{Q}_{t,k}$. Let $\widehat{P}^B_t$ be the probability taken with respect to the bootstrap distribution at time t. Define the confidence that treatment $\boldsymbol{a}_t$ is optimal in the current state as $\Omega_t(\boldsymbol{a}_t) = \widehat{P}^B_t\left\{\boldsymbol{a}_t = \operatorname{argmax}_{\boldsymbol{a}_t\in\,\psi(\boldsymbol{S}_t)}\widehat{Q}^B_{t,k}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right\}$. At each time t we draw $\boldsymbol{A}_t$ so that $P(\boldsymbol{A}_t = \boldsymbol{a}_t) = \Omega_t(\boldsymbol{a}_t)$. Implementation of this algorithm does not, however, require multiple draws of the weights. Rather, one need only draw a single set of weights $W_{t,0,1}, \ldots, W_{t,t-1,L}$, compute $\widehat{Q}^B_{t,k}$ using this set of weights, and subsequently $\boldsymbol{A}_t = \operatorname{argmax}_{\boldsymbol{a}_t\in\psi(\boldsymbol{S}_t)}\widehat{Q}^B_{t,k}(\boldsymbol{s}_t,\boldsymbol{a}_t)$; i.e., one does not actually have to estimate the sampling distribution of $\widehat{Q}_{t,k}$.

4.3. OPTIMIZATION

Finding a maximizer of the estimated Q-function involves searching a combinatorially-large space, which is intractable for moderately large L. In order to approximate $\operatorname{argmax}_{\boldsymbol{a}\in\psi(\boldsymbol{s})}\widehat{Q}_{t,k}(\boldsymbol{s},\boldsymbol{a})$, we maximize a quadratic approximation, yielding a tractable binary quadratic program. In particular, we introduce the approximation

$$\widetilde{q}^\ell_{t,k}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\xi}^\ell) \triangleq \xi^\ell_0(\boldsymbol{s}) + \sum_{i\in\mathcal{N}^\ell}\boldsymbol{\xi}^\ell_{1,i}(\boldsymbol{s})a^i + \sum_{i,j\in\mathcal{N}^\ell}\boldsymbol{\xi}^\ell_{2,i,j}(\boldsymbol{s})a^i a^j.$$

Each each time t, iteration k, and location ℓ, we compute

$$\widehat{\boldsymbol{\xi}}^\ell_{t,k} = \widehat{\boldsymbol{\xi}}^\ell_{t,k}(\boldsymbol{S}_t) = \arg\min_{\boldsymbol{\xi}^\ell}\sum_{\boldsymbol{a}_i}\left\{\widetilde{q}^\ell_{t,k}(\boldsymbol{S}_t,\boldsymbol{a}_i;\boldsymbol{\xi}^\ell) - \widehat{q}^\ell_{t,k}(\boldsymbol{S}_t,\boldsymbol{a}_i)\right\}^2,$$

where $\boldsymbol{a}_i, i = 1, \ldots, I$ is a collection of allocations in $\mathcal{A}$; in our experiments we generate $\boldsymbol{a}_i$ uniformly at random. For a given treatment budget ρ, the approximate maximizer of

$\widehat{Q}_{t,k}(\boldsymbol{s},\cdot)$ is given by the solution to

$$\begin{aligned} \max_{\boldsymbol{a}\in\{0,1\}^L} \quad & \sum_{\ell=1}^{L} \widehat{q}^{\ell}_{t,k}\left(\boldsymbol{S}_t, \boldsymbol{a}; \widehat{\boldsymbol{\xi}}^{\ell}_{t,k}\right) \\ \text{subject to} \quad & \sum_{\ell=1}^{L} a^{\ell} = \rho. \end{aligned} \tag{3}$$

Problem (3) is a binary quadratic program which can be solved efficiently with off-the-shelf solvers such as Gurobi (Bixby 2007).

5. ESTIMATING THE OPTIMAL STRATEGY VIA POLICY-SEARCH

The primary alternative to Q-learning that we consider in our simulation experiments is model-based policy-search. Thus, we briefly review it here. In model-based policy-search, one estimates a model for the underlying system dynamics, e.g., the disease model, and then uses the fitted model to identify the optimal strategy within a pre-specified class via Monte Carlo (Laber et al. 2018a).

For the purpose of illustration, we consider a class of candidate strategies of the form $\Pi = \{\pi(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ with $\boldsymbol{\Theta}\subset\mathbb{R}^p$ compact and

$$\pi(\boldsymbol{s};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{a}} \varphi(\boldsymbol{s},\boldsymbol{a})^{\intercal}\boldsymbol{\theta},$$

where φ is a known feature vector.

Computing the value function, $V(\pi)$, in closed form is not possible in general. Thus, we estimate the value function using Monte Carlo approximation applied over a finite horizon. Let $V^{\pi}_{T}(\boldsymbol{s}) = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{T}\gamma^t u(\boldsymbol{Y}_t)\middle|\boldsymbol{S}_0=\boldsymbol{s}\right\}$ be a surrogate for the value function; it follows under mild moment conditions that $V(\pi)\approx V_T(\pi)$ when T is large (Bertsekas 2007). To construct as estimator $\widehat{V}_T(\pi)$ of $V_T(\pi)$, we postulate a model for the system dynamics. As the process is assumed to be Markov and homogeneous, it is completely determined by the transition kernel. We posit a parametric model for this kernel, $\kappa(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\boldsymbol{\beta})$, which is indexed by $\boldsymbol{\beta}\in\mathfrak{B}\subseteq\mathbb{R}^q$; thus, under this model with parameters $\boldsymbol{\beta}$ we have

$$P\left(\boldsymbol{S}_{t+1}\in\mathcal{B}\middle|\boldsymbol{S}_t=\boldsymbol{s},\boldsymbol{A}_t=\boldsymbol{a}\right) = \int \kappa(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\beta)\mathrm{d}\lambda(\boldsymbol{s}'),$$

where λ is a dominating measure. The T-step value function for strategy π under parameter vector $\boldsymbol{\beta}$, starting from state $\boldsymbol{s}$ is thus equal to

$$\begin{aligned} V_T(\boldsymbol{s},\pi;\boldsymbol{\beta}) = \int \left[\sum_{t=0}^{T-1}\gamma^t u(\boldsymbol{s}_{t+1})\right]\kappa(\boldsymbol{s}_T|\boldsymbol{s}_{T-1},\boldsymbol{a}_{T-1};\beta) \\ \left[\prod_{t=0}^{T-1} P\left\{\pi(\boldsymbol{s}_t)=\boldsymbol{a}_t\right\}\kappa(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{a}_{t-1};\beta)\right]\mathrm{d}\lambda(\overline{\boldsymbol{s}}_T,\overline{\boldsymbol{a}}_{T-1}), \end{aligned} \tag{4}$$

where λ is a dominating measure, $f(\boldsymbol{s}_0|\boldsymbol{s}_{-1}, \boldsymbol{a}_{-1}, \boldsymbol{y}_{-1})$ is taken to be a point mass at $\boldsymbol{s}$, and we have used the fact that $\boldsymbol{Y}_t$ is $\boldsymbol{S}_{t+1}$ measurable so that we can write $u(\boldsymbol{Y}_t) = u(\boldsymbol{S}_{t+1})$.

We use Thompson Sampling to balance exploration and exploitation when estimating the optimal strategy. This yields the following algorithm which executes at each time point: (i) obtain a posterior distribution for β using the observed data (at time $t = 0$ this is taken to be the prior); (ii) draw $\widetilde{\beta}$ from the posterior; and (iii) estimate the optimal strategy under the belief that $\widetilde{\beta}$ indexs the true underlying model. The policy search estimator of the optimal strategy at each time point is thus $\widehat{\pi}^{\text{opt}}_{PS}(\boldsymbol{s}) = \underset{\pi \in \Pi}{\text{argmax}}\ V^{\pi}_{T}(\boldsymbol{s}; \widetilde{\beta})$. Stochasticity in $\widehat{\pi}^{\text{opt}}_{PS}$ is induced by the sampling variability of $\tilde{\beta}$.

When the model is not severely misspecified, the model-based policy-search based on a low-dimensional dynamics can often yield better estimates of the optimal policy than model-free approaches when data are scarce, i.e., early in the decision process. However, as we illustrate in the next section, model-based methods can be highly sensitive to misspecification.

6. SIMULATION EXPERIMENTS

In our experiments, we consider a replicating agent spreading over a network according to a susceptible-infected-susceptible (SIS) model (Weiss and Dishon 1971); this model was chosen in part because it allows for correct specification for the model-based estimators that we use as a baseline for comparison with our proposed method. Under the SIS model, each location transitions among the three states: susceptible, infected, and susceptible. Infection spreads from infected to susceptible locations. As soon as a location has been infected, it has the potential to recover from the disease. Once a location recovers, it immediately becomes susceptible to the disease again (see Keeling and Eames 2005, for discussion about epidemic models and references).

We let $Y^{\ell}_t \in \{0, 1\}$ denote the infection status of location ℓ at time t, i.e., $Y^{\ell}_t = 1$ if location ℓ is infected at time t and zero otherwise. Define $\mathcal{I}_t = \{\ell \in \mathcal{L} : Y^{\ell}_t = 1\}$ to be the set of infected locations at time t and $\mathcal{I}^c_t$ its complement. With each location is an associated covariate $x^{\ell}_t \in \mathbb{R}$. The state at time t is thus $\boldsymbol{S}_t = (\boldsymbol{X}_t, \boldsymbol{Y}_{t-1})$. The evolution of the state is governed by the following models

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \nu) &= \prod_{\ell=1}^{L} \phi\left(\frac{\boldsymbol{x}^{\ell}_{t+1} - \nu_0 \boldsymbol{x}^{\ell}_t}{\nu_1}\right), \\
b(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{y}_{t-1}, \boldsymbol{a}_t; \boldsymbol{\eta}) &= \left[\prod_{\ell \in \mathcal{I}_{t-1}} q_{\ell}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\eta})^{1-y^{\ell}_t}[1 - q_{\ell}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\eta})]^{y^{\ell}_t}\right] \cdot \\
&\quad \left[\prod_{\ell \in \mathcal{I}^c_{t-1}} p_{\ell}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\eta})^{y^{\ell}_t}[1 - p_{\ell}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\eta})]^{1-y^{\ell}_t}\right],
\end{aligned} \tag{5}
$$

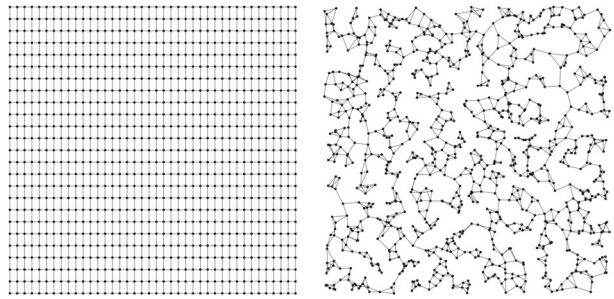

Figure 2. Instances of the two network structures used in the simulation examples. Left: lattice network with 1000 locations. Right: random three-nearest-neighbor network with 1000 locations.

where $\phi(\cdot)$ is the probability density function for the standard normal distribution,

$$p_\ell(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \boldsymbol{\eta}) = 1 - \{1 - p_{\ell,0}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta})\} \cdot \prod_{\ell' \in \mathcal{I}_{t-1} \cap \mathcal{N}^\ell} \{1 - p_{\ell,\ell'}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta})\}, \tag{6}$$

and

$$\begin{aligned} \operatorname{logit}\left\{p_{\ell,0}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta})\right\} &= \eta_0 + \eta_1 a^\ell, \\ \operatorname{logit}\left\{p_{\ell,\ell'}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta})\right\} &= \eta_2 + \eta_3 a^\ell + \eta_4 a^{\ell'}, \\ \operatorname{logit}\left\{q_\ell(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta})\right\} &= \eta_5 + \eta_6 a^\ell. \end{aligned} \tag{7}$$

Thus, the model is indexed by $\boldsymbol{\beta} = (\nu_0, \nu_1, \eta_0, \ldots, \eta_6)^\intercal$. It can be seen that under this model treating an uninfected location reduces the probability that it becomes infected, while treating an infected location has the dual effect of reducing its likelihood of transmitting infection to adjacent uninfected location and accelerating its recovery.

To study the effects of model misspecification, we introduce a contamination model g^{Contam} for the conditional infection probabilities. For each $\epsilon \in [0, 1]$, we define a contaminated model b^ϵ whose infection probabilities are

$$b^\epsilon(\cdot \mid \boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\eta}) \triangleq (1 - \epsilon) b(\cdot \mid \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{y}; \boldsymbol{\eta}) + \epsilon g^{\text{Contam}}(\cdot \mid \boldsymbol{s}, \boldsymbol{a}).$$

Our contamination model g^{Contam} is based on a "shield-state" variant of the SIS model in which infection probabilities are mediated by the covariate x (in particular, the indicator $\mathbb{1}\{x \leq 0\}$), modified such that the transmissions of infection to a location from its neighbors are no longer independent. More details of the contamination model are in the Supplemental Materials. Thus, in the case of full contamination ($\epsilon = 1$), the SIS model defined in equations (5)–(7) is severely misspecified. Meanwhile, a logistic regression model with sufficiently expressive features φ^ℓ is approximately correctly specified in each case, though with high variance.

6.1. EXPERIMENT SETUP

Parameters indexing the generative models are tuned to have specified infection rates using a network of size four. Thus, in tuning these parameters we have $\mathcal{L} = \{1, 2, 3, 4\}$ and

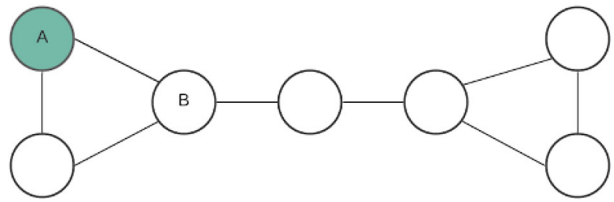

Figure 3. An instance of the "lookahead" network structure, in which the location labeled A is infected. A myopic strategy with a budget of $\tau = 1$ might treat location A in order to minimize the expected number of infections at the next step, a non-myopic strategy would treat location B as it has more neighbors and therefore would cause more infections if infected. In our experiments, a lookahead network of size L consists of $L/7$ repetitions of the structure.

we assume that location 1 is a neighbor of all other locations. Each equation is given in terms of location 1. The rates are defined as

$$\begin{aligned} p_{1,0}(\{\cdots\}, \{0, \cdots\}, \mathbf{0}) &= 0.01, \\ p_1(\{\cdots\}, \{0, 1, 1, 1\}, \mathbf{0}) &= 0.5, \\ p_1(\{+, \cdots\}, \{0, 1, 1, 1\}, \{1, 0, 0, 0\}) &= 0.375, \\ p_1(\{\cdot, +, +, +\}, \{0, 1, 1, 1\}, \{0, 1, 1, 1\}) &= 0.125, \\ q_1(\{\cdots\}, \{1, \cdots\}, \{0, \cdots\}) &= 0.25, \\ q_1(\{+, \cdots\}, \{1, \cdots\}, \{1, \cdots\}) &= 0.125, \end{aligned}$$

where $\cdot$ represents any value and $+$ represents a positive value. In particular, these equations set the latent probability of infection without treatment to 0.01; the probability of infection when 3 neighbors are infected without treatment to 0.5; preventative treatments to reduce the probability of infection by a factor of 0.75 when three neighbors are infected and none of which are treated; active treatments to reduce the probability of infecting a neighbor by a factor of 0.25 assuming only three infected neighbors and all of which are treated; the base probability of recovery with no treatment to 0.25; and the probability of not recovering to decrease by a factor of 0.5 with a treatment.

We present simulation results from rolling out each of the strategies under consideration for $T = 25$ time steps on lattice, random nearest neighbor, and custom network structures with $L = 100$ locations, with contamination parameters $\epsilon \in \{0, 0.5, 1\}$. We provide more details in the Supplemental Materials.

6.2. FEATURE CONSTRUCTION FOR MODEL-FREE ESTIMATION IN SIS ENVIRONMENTS

We consider classes of Q-functions which are linear in (1) the raw features at each location (i.e., infection status, treatment status, and any covariates); (2) handcrafted features designed to efficiently encode information about each locations' neighbors; and (3) learned features using the graph neural network as described in Sect. 4.1.1. We describe the construction of the handcrafted features here; details on the graph neural network implementation such as the number of neurons, tuning procedures, etc., can be found in the Supplemental Materials.

In order to pool information across individual locations in the network, we construct features at each location, which contain both that location's state and treatment status as well as its neighbors' state. In our case, we define the binary covariate as $\iota^\ell \triangleq \mathbb{1}\left\{x^\ell \leq 0\right\}$; this is informed by the shield-state variant of the SIS model mentioned above, in which infection probabilities are mediated by the value of ι^ℓ. We summarize the features of a location's neighbors using binary encodings. Let $\mathcal{N}_k^\ell$ be the set of paths of length k beginning with location ℓ, and for each $r_k^\ell \in \mathcal{N}_k^\ell$ with covariates $\{(\iota^{\ell^{(1)}}, a^{\ell^{(1)}}, y^{\ell^{(1)}}), \ldots, (\iota^{\ell^{(k)}}, a^{\ell^{(k)}}, y^{\ell^{(k)}})\}$, let $b(r_k^\ell) = \sum_{i=1}^k \left(2^{3(i-1)}\iota^{\ell^{(i)}} + 2^{3(i-1)+1}a^{\ell^{(i)}} + 2^{3(i-1)+2}y^{\ell^{(i)}}\right) + 1$. Then, writing the j^{th} basis vector in $\mathbb{R}^{2^{3k}}$ as e_j, the handcrafted feature of location ℓ is given by $\sum_{r_k^\ell \in \mathcal{N}_k^\ell} e_{b(r_k^\ell)}$; that is the vector of counts of each feature combination on each path of length k from a location ℓ.

Denoting $\cup$ as the vector concatenation, we write the feature function for each location ℓ as

$$\begin{aligned}\varphi_k^\ell : \{0,1\}^L \times \{0,1\}^L \times \{0,1\}^L &\longrightarrow \mathbb{R}^{2^3+2^{3k}} \\ (\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{y}) &\mapsto \left[\iota^\ell \; a^\ell \; y^\ell\right] \cup \left(\sum_{r_k^\ell \in \mathcal{N}_k^\ell} e_{b(r_k^\ell)}\right),\end{aligned}$$

where $\left[\iota^\ell \; a^\ell \; y^\ell\right]$ is the binary vector of length 8 corresponding to $(\iota^\ell, a^\ell, y^\ell)$, and $\cup$ means vector concatenation. For example, the first-order neighbor feature vector is of length 16 and the second-order neighbor feature vector is of length 72.

6.3. Experiment Results

We compare the following learning algorithms:

- Model-free estimators of Q_1, using either a graph neural net work architecture, a linear architecture with raw features, or a linear architecture with the hand crafted features described in Sect. 6.2. We refer to these as `M-g`, `M-r`, and `M-h`, respectively ("`M`" represents myopic);
- Model-free estimators of Q_2, which we referred to as `F-g-2`, `F-r-2`, and `F-h-2` ("`F`" represents FQI since these entail applying two steps of spatial FQI);
- Model-free estimators of Q_3, which we referred to as `F-g-3`, `F-r-3`, and `F-h-3`;
- The policy search (`PS`) algorithm of Laber et al. (2018a) (Sect. 5), using Bayesian optimization (implemented in the BayesOpt package in Python (Nogueira 2018)) to carry out the requisite optimizations.

Additionally, we compute the average performance of the random strategy, which chooses a subset of ρ locations to treat. We also estimate the performance of two "oracle" strategies. The first is the model-based policy search learning algorithm described above, except in this case rollouts are conducted with the exact (and correctly specified) model, rather than an

estimate (OPS). The second oracle strategy conducts one step of FQI using the true infection probability (TP). (See the Supplemental Materials for more details on these algorithms.)

Let Γ be a learning algorithm, i.e., $\Gamma : \text{dom } \boldsymbol{H}_t \to \mathcal{A}$, where $\boldsymbol{H}_t$ is a history of observations. Define the utility to be the total number of infected locations at time t, i.e., $u(\boldsymbol{y}_t) = \sum_{\ell=1}^{L} y_t^{\ell}$. Then let $V(\Gamma; \beta)$ be the expected cumulative discounted utility over $T = 25$ time steps incurred by learning algorithm Γ when the generative model is given by $\boldsymbol{\beta}$:

$$V(\Gamma; \boldsymbol{\beta}) \triangleq \mathbb{E}_{\Gamma, \boldsymbol{\beta}} \left[\sum_{t=0}^{25} \gamma^t u(\boldsymbol{Y}_t) \right].$$

Figure 4 displays the normalized mean infection counts under each learning algorithm and each SIS environment considered. There are several features of these results worth noting:

- For the myopic and FQI learning algorithms, the hand-crafted features and the neural network features improve performance significantly relative to using the raw features. This result is anticipated by the literature emphasizing the importance of feature construction in reinforcement learning (Song et al. 2016). It is encouraging that the neural network features perform better than the handcrafted features based on the structure of the true generative model.

- Given the structure of these networks, FQI learning algorithms with handcrafted features or neural network architecture outperform or have similar performance to their myopic counterparts. Because of the dense connection among the locations in the lattice and nearest neighbor network, FQI learning algorithms with neural network architecture have a significant advantage compared with the random strategy. As the connections in the custome network are very spare, the decrease in normalized mean infection is relatively small. It is worth noting that the FQI with neural network architecture is consistently among the best.

- As expected, model-based policy-search performs well when the model is correctly specified or moderately misspecified, but its performance deteriorates significantly as the degree of misspecification increases (brittleness to model misspecification is also shown in Rose et al. (2019)).

- FQI learning algorithm with neural network architecture is the most stable method in that it is robust to the network type and model contamination. It is the only strategy that outperforms the random strategy in all settings. Under the scenarios with $\epsilon = 1$, its performance is similar to the oracle myopic strategy and is better than the oracle policy search strategy.

More results can be found in the Supplemental Materials.

7. MANAGEMENT OF THE EBOLA VIRUS

Demonstrating our method with Ebola virus disease requires some changes to our setup. First, the dynamic model for Ebola is taken from Kramer et al. (2016a). This model is called

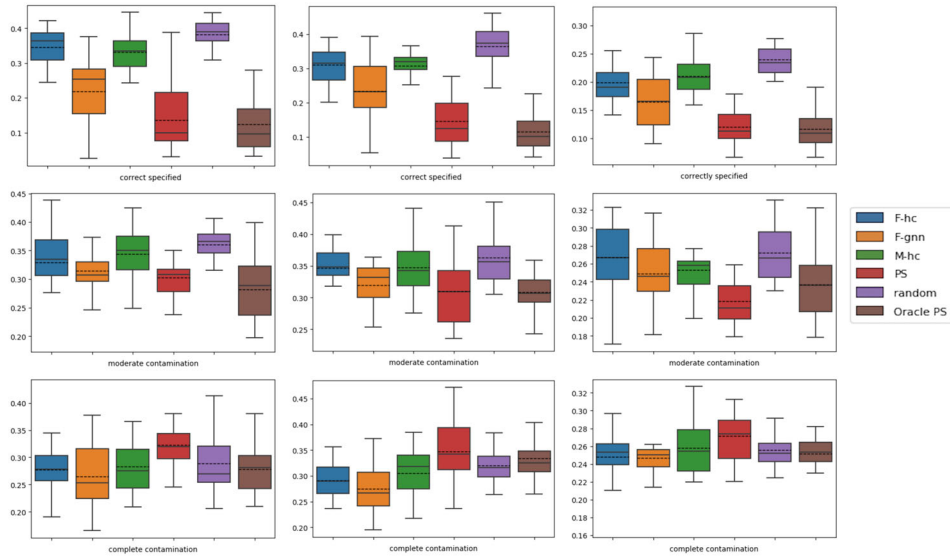

Figure 4. Normalized mean infection counts under different learning algorithms in the SIS environment, for lattice (left), nearest neighbor (center), and lookahead (right) network structures with $L = 100$ and different levels of contamination. The solid line in each box represents the sample median, and the dotted line represents the sample mean. F-hc: Q_2 with handcrafted features; F-gnn: Q_2 with graphical neural network architecture; M-hc: myopic algorithm with handcrafted features; random: random allocation policy; PS: policy search algorithm; Oracle PS: oracle policy search algorithm.

the gravity model, and the state-information is constant and given by the population of each location n^ℓ as well as the distances between each location $d^{\ell,\ell'}$. We take the neighbors $\mathcal{N}^\ell$ of each location ℓ to be the four locations ℓ' with the smallest values of $d_{\ell,\ell'}/n_\ell n_{\ell'}$. Under the gravity model there is no recovery from infection, and the probability of transmission of infection from an infected location ℓ' to an uninfected location ℓ is defined by $(\forall \ell \in \{1, \ldots, L\})(\forall \ell' \in \mathcal{N}^\ell)$

$$\text{logit}[p_{\ell,\ell'}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta)] = \eta_0 - e^{\eta_1} \frac{d^{\ell,\ell'}}{(n^\ell n^{\ell'})^{e^{\eta_2}}} + \eta_3 a^\ell + \eta_4 a^{\ell'}.$$

This model acquires its name from the second term, know as the gravity term. The numerator is the distance between two locations and it is normalized by the product of the populations in each location in the denominator. To stabilize the estimation of the model, we force the coefficient on the gravity term and the exponent on the population product to both be positive.

We fit the gravity model with no treatment terms to the observed infection data from the 2013–2015 Ebola epidemic to obtain $\widehat{\eta}_0^{\text{MLE}}, \widehat{\eta}_1^{\text{MLE}}, \widehat{\eta}_2^{\text{MLE}}$, and tuned the generative model used in our experiments to meet two conditions. In the first, we took π_1 to be the policy that applies no treatment and considered models with parameters of the form $\eta(\tau_1) = \{\tau_1 \widehat{\eta}_0^{\text{MLE}}, \log(\tau_1) + \widehat{\eta}_1^{\text{MLE}}, \widehat{\eta}_2^{\text{MLE}}, 0.0, 0.0\}$. We then choose $\widehat{\tau}_1$ such that $\mathbb{E}_{\{\pi_1,\eta(\tau)\}}\left(\sum_{\ell=1}^{L} \boldsymbol{Y}_{25}^\ell\right) \approx 0.7\, L$ (where $\mathbb{E}_{\{\pi,\eta\}}$ refers to the expectation taken with respect to trajectories in which the dynamics are given by the gravity model with parameter η and

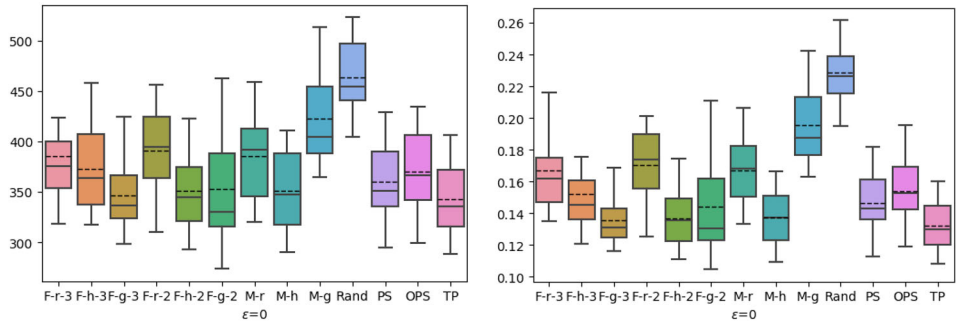

Figure 5. Cumulative discounted utility (left) and the normalized mean infection counts (right) under different learning algorithms in the Ebola environment. The solid line represents the sample median, and the dotted line represents the sample mean.

strategy π is followed). In the second, we took π_2 to be the greedy policy under the true generative model; π_3 to be the uniformly random policy with budget $\lfloor 0.15L \rfloor$; and, considering models with parameters of the form $\eta(\tau_2) = \{\widehat{\tau}_1\widehat{\eta}_0^{\text{MLE}}, \log(\widehat{\tau}_1) + \widehat{\eta}_1^{\text{MLE}}, \widehat{\eta}_2^{\text{MLE}}, \tau_2, \tau_2\}$, chose $\widehat{\tau}_2$ to minimize $\mathbb{E}_{\{\pi_3,\eta(\tau_2)\}}\left(\sum_{\ell=1}^{L} \boldsymbol{Y}_{25}^{\ell}\right) - \mathbb{E}_{\{\pi_2,\eta(\tau_2)\}}\left(\sum_{\ell=1}^{L} \boldsymbol{Y}_{25}^{\ell}\right)$ over a grid of values between 0 and 10.

7.1. FEATURE CONSTRUCTION FOR MODEL-FREE ESTIMATION IN GRAVITY ENVIRONMENT

As before, in addition to the graph neural net-based policies described above, we use a linear Q-function architecture with hand-crafted features. We append each location's population and action and infection status, i.e., $\left(n^{\ell}, a^{\ell}, y^{\ell}\right)$, to each of its neighbor's features. We also include the distances of a location to each of its neighbors. Thus the feature vector at location ℓ and time t is (using $\cup$ to denote vector concatenation):

$$\varphi^{\ell}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{y}_t) = \left[n^{\ell}\ a_t^{\ell}\ y_t^{\ell}\right] \cup \left(\bigcup_{\ell' \in \mathcal{N}^{\ell}} \left[n^{\ell'}\ a_t^{\ell'}\ y_t^{\ell'}\ d^{\ell,\ell'}\right] \right).$$

7.2. RESULTS

Figure 5 displays the results for each of the learning algorithms described above (in the case of policy search, assuming correct specification of the model). We find that Q_3 with graphical neural network architecture outperforms other learning algorithms. Q_2 with handcrafted features, Q_2 with neural network architecture, and Q_1 with handcrafted features have similar performance, which are very close to the oracle TP strategy. Interestingly, FQIs with graphical neural network architecture generate better performance than the model based policy search algorithm with correct specified infection model. We conclude that our methods can generate effective real-time intervention strategies, which lead to significant reductions in the spread of EVD compared to random allocation strategies. The learned strategies can provide adequate and in time implemented infection control procedures from incipient outbreaks.

8. CONCLUSION

We develop a semiparametric (model-free) approach to the online control of an emerging infectious disease. In simulation experiments, this approach provided better control of an infectious disease than *ad-hoc* strategies and was robust to certain kinds of model misspecification compared to the model based policy search approach.

There are a number of important and interesting open problems associated with spatio-temporal decision-making. Disease surveillance data can be noisy, sparse, or incomplete. Extending the proposed methods to accommodate such data, perhaps by generalizing the underlying Markov decision process framework to a partially observable Markov decision process (see, e.g., Ross et al. 2008), could potentially improve solution quality. While we have proposed an online methodology, the updates to the policy estimator require significant computation time. While this is acceptable for settings where decisions operate on a scale of days, it would be desirable for decision support systems operating on a finer time scale or deployed on CPU-limited devices (e.g., mobile phones) to develop estimators that can be updated in linear time.

A set of critical open problems involves incorporating decision strategies such as those presented here into a broader decision-support system. Estimated optimal treatment strategies for human diseases are intended to *inform*, not dictate, decision-making. Thus, communication and visualization tools are needed to assist decision makers in using data-driven management strategies like the one proposed here. Additionally, if the recommendations of the estimated strategy are only one input into the decision-making of the relevant actors—in particular, if decision makers sometimes deviate from the recommendations on the basis of some additional information—then the sequential decision-making model developed in Sect. 3 no longer adequately models the decision process. (For instance, the assumption of strong ignorability (A3) may be violated.) The development of tools for reinforcement learning and causal inference, which more fully account for the fact that reinforcement-learning based decision support tools will likely be only one input into the final decisions made regarding, say, the control of an epidemic, remains an essential and fascinating unsolved problem.

One potential drawback of the proposed method is its black-box nature. It is difficult to delineate the information used to construct the graph embeddings and the structure of interference. The development of interpretable methods that provide additional insights into the learned optimal strategy is an important area for future work.

[Received November 2022. Revised May 2023. Accepted May 2023.]

REFERENCES

Agrawal S, Goyal N (2011) Analysis of thompson sampling for the multi-armed bandit problem. arXiv preprint arXiv:1111.1797

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. ICML 3:127–135

Almirall D, Ten Have T, Murphy SA (2010) Structural nested mean models for assessing time-varying effect moderation. Biometrics 66:131–139

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866

Atwood CL (1973) Sequences converging to d-optimal designs of experiments. The Annals of Statistics 342–352

Auer P (2000) Using upper confidence bounds for online learning. In: Proceedings 41st annual symposium on foundations of computer science, pp 270–279. IEEE

Bartroff J, Lai TL, Shih M-C (2012) Sequential experimentation in clinical trials: design and analysis, vol 298. Springer Science & Business Media, Berlin

Bebis G, Georgiopoulos M (1994) Feed-forward neural networks. IEEE Potentials 13:27–31

Bellman R (1957) Dynamic programming, 1st edn. Princeton University Press, Princeton

Bertsekas DP (2007) Dynamic programming and optimal control, vol II. Athena Scientific, Nashua

Bertsekas DP, Bertsekas DP, Bertsekas DP, Bertsekas DP (1995) Dynamic programming and optimal control, vol 1. Athena Scientific, Belmont

Bertsekas DP, Tsitsiklis JN (1996) Neuro-dynamic programming, 1st edn. Athena Scientific, Nashua

Bixby B (2007) The gurobi optimizer. Transp Res Part B 41:159–178

Bloom DE, Cadarette D (2019) Infectious disease threats in the twenty-first century: strengthening the global response. Front Immunol 10:549

Busoniu L, Babuska R, De Schutter B, Ernst D (2010) Reinforcement learning and dynamic programming using function approximators, vol 39. CRC Press, Boca Raton

Cai H, Zheng VW, Chang KC-C (2018) A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans Knowl Data Eng 30:1616–1637

Carr S, Roberts S (2010) Planning for infectious disease outbreaks: a geographic disease spread, clinic location, and resource allocation simulation. In: Proceedings of the 2010 winter simulation conference, pp. 2171–2184. IEEE

Cecchine G, Moore M (2006) Infectious disease and national security: strategic information needs. Rand Corporation, Santa Monica

Chakraborty B, Moodie E (2013) Statistical methods for dynamic treatment regimes. Springer, Berlin

Chatterjee S, Bose A et al (2005) Generalized bootstrap for estimating equations. Ann Stat 33:414–436

Chernoff H (1972) Sequential analysis and optimal design. Vol. 8. SIAM, Philadelphia

Eckles D, Kaptein M (2014) Thompson sampling with the online bootstrap. arXiv preprint arXiv:1410.4009

Ernst D, Geurts P, Wehenkel L (2005) Tree-based batch mode reinforcement learning. J Mach Learn Res 6:503–556

Ertefaie A (2014) Constructing dynamic treatment regimes in infinite-horizon settings. arXiv preprint arXiv:1406.0764

Ertefaie A, McKay JR, Oslin D, Strawderman RL (2021) Robust q-learning. J Am Stat Assoc 116:368–381

Ertefaie A, Strawderman RL (2018) Constructing dynamic treatment regimes over indefinite time horizons. Biometrika 105:963–977

Feldmann H, Geisbert TW (2011) Ebola haemorrhagic fever. Lancet 377:849–862

Fey M, Lenssen JE (2019) Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428

Forastiere L, Airoldi EM, Mealli F (2021) Identification and estimation of treatment and interference effects in observational studies on networks. J Am Stat Assoc 116:901–918

Fortunato M, Azar MG, Piot B, Menick J, Osband I, Graves A, Mnih V, Munos R, Hassabis D, Pietquin O, et al (2017) Noisy networks for exploration. arXiv preprint arXiv:1706.10295

Gopalan A, Mannor S, Mansour Y (2014) Thompson sampling for complex online problems. ICML 14:100–108

Guan Q, Reich BJ, Laber EB (2022) A spatiotemporal recommendation engine for malaria control. Biostatistics 3:1023–1038

Hamel MJ, Slutsker L (2015) Ebola: the hidden toll. Lancet Infect Dis 15:756–757

Henderson R, Ansell P, Alshibani D (2010) Regret-regression for optimal dynamic treatment regimes. Biometrics 66:1192–1201

Hernandez-Leal P, Kartal B, Taylor ME (2019) A survey and critique of multiagent deep reinforcement learning. Auton Agents Multi-Agent Syst 33:750–797

Hernández-Lerma O, Lasserre JB (2012) Discrete-time Markov control processes: basic optimality criteria, vol 30. Springer Science & Business Media, Berlin

Hu T, Laber E, Meyer N, Pacifici K, Drake J (2017) Note on thompson sampling for large decision problems. Under review 1:1–10

Hudgens MG, Halloran ME (2008) Toward causal inference with interference. J Am Stat Assoc 103:832–842

Jin Z, Ying Z, Wei L-J (2001) A simple resampling method by perturbing the minimand. Biometrika 88:381–390

Karwa V, Airoldi EM (2018) A systematic investigation of classical causal inference strategies under mis-specification due to network interference. arXiv preprint arXiv:1810.08259

Kasaie P, Kelton WD (2013) Simulation optimization for allocation of epidemic-control resources. IIE Trans Healthc Syst Eng 3:78–93

Kaufmann E, Korda N, Munos R (2012) Thompson sampling: an asymptotically optimal finite-time analysis. In: International conference on algorithmic learning theory, pp. 199–213. Berlin, Heidelberg: Springer

Keeling MJ, Eames KT (2005) Networks and epidemic models. J R Soc Interface 2:295–307

Kompella V, Capobianco R, Jong S, Browne J, Fox S, Meyers L, Wurman P, Stone P (2020) Reinforcement learning for optimization of covid-19 mitigation policies. arXiv preprint arXiv:2010.10560

Korda N, Kaufmann E, Munos R (2013) Thompson sampling for 1-dimensional exponential family bandits. Adv Neural Inf Process Syst 26:1448–1456

Kosorok MR, Moodie EE (2015) Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine. (Vol. 21). SIAM, Philadelphia

Kramer AM, Pulliam JT, Alexander LW, Park AW, Rohani P, Drake JM (2016) Spatial spread of the West Africa Ebola epidemic. R Soc Open Sci 3:160294

Laber E, Rose E, Davidian M, Tsiatis A (2017) Q-learning. Wiley StatsRef. https://doi.org/10.1002/9781118445112.stat07998

Laber EB, Meyer NJ, Reich BJ, Pacifici K, Collazo JA, Drake JM (2018) Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. J Roy Stat Soc Ser C (Appl Stat) 67:743–789

Lai TL, Wei CZ et al (1982) Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Ann Stat 10:154–166

Lattimore T, Szepesvári C (2020) Bandit algorithms. Cambridge University Press

Lee J, Lee Y, Kim J, Kosiorek A, Choi S, Teh YW (2019) Set transformer: a framework for attention-based permutation-invariant neural networks. In: International conference on machine learning, pp 3744–3753. PMLR

Li S-L, Bjørnstad ON, Ferrari MJ, Mummah R, Runge MC, Fonnesbeck CJ, Tildesley MJ, Probert WJM, Shea K (2017) Essential information: Uncertainty and optimal control of ebola outbreaks. In: Proceedings of the National Academy of sciences

Linn KA, Laber EB, Stefanski LA (2017) Interactive q-learning for quantiles. J Am Stat Assoc 112:638–649

Liu Y, Wang Y, Kosorok MR, Zhao Y, Zeng D (2018) Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. Stat Med 37:3776–3788

Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY et al (2013) Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. Lancet 380:2095–2128

Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR (2020) Estimating dynamic treatment regimes in mobile health using v-learning. J Am Stat Assoc 115:692–706

Ma Y, Wang Y, Tresp V (2020) Causal inference under networked interference. arXiv preprint arXiv:2002.08506

Maei HR, Szepesvári C, Bhatnagar S, Sutton RS (2010) Toward off-policy learning control with function approximation. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 719–726

Mathers C (2008) The global burden of disease: 2004 update. World Health Organization, Geneva

Merler S, Ajelli M, Fumanelli L, Gomes MFC, Piontti AP, Rossi L, Chao DL, Longini IM Jr, Halloran ME, Vespignani A (2015) Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. Lancet Infect Dis 15:204–211

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Minnier J, Tian L, Cai T (2011) A perturbation method for inference on regularized regression estimates. J Am Stat Assoc 106:1371–1382

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: International conference on machine learning, pp 1928–1937. PMLR

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518:529–533

Moodie EE, Dean N, Sun YR (2014) Q-learning: flexible learning about useful utilities. Stat Biosci 6:223–243

Murphy SA (2003) Optimal dynamic treatment regimes. J R Stat Soc Ser B (Stat Methodol) 65:331–355

Murphy SA (2005) A generalization error for q-learning. J Mach Learn Res 6:1073–1097

Nogueira FMF Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python, 2014. Accessd 6 Jan 2022

Nowzari C, Preciado VM, Pappas GJ (2015) Optimal resource allocation for control of networked epidemic models. IEEE Trans Control Netw Syst 4:159–169

Orellana L, Rotnitzky A, Robins JM (2010) Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. Int J Biostat 6. https://doi.org/10.2202/1557-4679.1200

Osband I, Van Roy B, Russo DJ, Wen Z et al (2019) Deep exploration via randomized value functions. J Mach Learn Res 20:1–62

Pan Y, Zhao Y-Q (2020) Improved doubly robust estimation in learning optimal individualized treatment rules. J Am Stat Assoc 116:283–294

Plappert M, Houthooft R, Dhariwal P, Sidor S, Chen RY, Chen X, Asfour T, Abbeel P, Andrychowicz M (2017) Parameter space noise for exploration. arXiv preprint arXiv:1706.01905

Powell WB (2007) Approximate dynamic programming: solving the curses of dimensionality, vol 703. John Wiley & Sons, Hoboken

Pronzato L (2000) Adaptive optimization and d-optimum experimental design. Ann Stat 28:1743–1761

Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, Hoboken

Rainsch G, Shanker MB, Wellman M, Merlin T, Meltzer MI (2015) Regional spread of Ebola virus, West Africa, 2014. Emerg Infect Dis J 21:444

Riedmiller M (2005) Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In: European conference on machine learning, pp 317–328. Springer

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. Math Model 7:1393–1512

Robins JM (1987) Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect". Comput Math Appl 14:923–945

Robins JM (2004) Optimal structural nested models for optimal sequential decisions. In: Proceedings of the second seattle symposium in biostatistics, pp 189–326. Springer

Rose EJ, Laber EB, Davidian M, Tsiatis AA, Zhao Y-Q, Kosorok MR (2019) Sample size calculations for smarts. arXiv preprint arXiv:1906.06646

Ross S, Pineau J, Paquet S, Chaib-Draa B (2008) Online planning algorithms for POMDPs. J Artif Intell Res 32:663–704

Rowland M, Dadashi R, Kumar S, Munos R, Bellemare MG, Dabney W (2019) Statistics and samples in distributional reinforcement learning. arXiv preprint arXiv:1902.08102

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688

Rubin DB, van der Laan MJ (2012) Statistical issues and limitations in personalized medicine research with clinical trials. Int J Biostat 8:18

Russo D, Van Roy B (2014) Learning to optimize via information-directed sampling. Adv Neural Inf Process Syst, 27: 1583–1591

Saghafian S (2021) Ambiguous dynamic treatment regimes: a reinforcement learning approach. arXiv preprint arXiv:2112.04571

Schulte PJ, Tsiatis AA, Laber EB, Davidian M (2014) Q-and a-learning methods for estimating optimal dynamic treatment regimes. Stat Sci Rev J Inst Math Stat 29:640

Scott SL (2010) A modern Bayesian look at the multi-armed bandit. Appl Stoch Model Bus Ind 26:639–658

Smith KF, Goldberg M, Rosenthal S, Carlson L, Chen J, Chen C, Ramachandran S (2014) Global rise in human infectious disease outbreaks. J R Soc Interface 11:20140950

Song Z, Parr RE, Liao X, Carin L (2016) Linear feature encoding for reinforcement learning. Adv Neural Inf Process Syst 29

Splawa-Neyman J, Dabrowska D, Speed T et al (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci 5:465–472

Sunehag P, Lever G, Gruslys A, Czarnecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K, et al (2017) Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, Cambridge

Szepesvári C (2010) Algorithms for reinforcement learning. Synth Lect Artif Intell Mach Learn 4:1–103

Tec M, Scott J, Zigler C (2022) Weather2vec: representation learning for causal inference with non-local confounding in air pollution and climate studies. arXiv preprint arXiv:2209.12316

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25:285–294

Tsiatis AA, Davidian M, Holloway ST, Laber EB (2019) Dynamic treatment regimes: statistical methods for precision medicine. CRC Press, Boca Raton

Wang L, Zhou Y, Song R, Sherwood B (2018) Quantile-optimal treatment regimes. J Am Stat Assoc 113:1243–1254

Wang Y, Xu T, Niu X, Tan C, Chen E, Xiong H (2020) STMARL: a spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control. IEEE Trans Mob Comput 21:2228–2242

Watkins CJCH (1989) Learning from delayed rewards. PhD thesis, King's College, Cambridge

Weiss GH, Dishon M (1971) On the asymptotic behavior of the stochastic and deterministic models of an epidemic. Math Biosci 11:261–265

WHO Ebola Response Team (2014) Ebola virus disease in West Africa-the first 9 months of the epidemic and forward projections. N Engl J Med 2014:1481–1495

Xie M-G, Singh K (2013) Confidence distribution, the frequentist distribution estimator of a parameter: a review. Int Stat Rev 81:3–39

Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin S (2006) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29:40–51

Zhang B, Tsiatis AA, Laber EB, Davidian M (2012) A robust method for estimating optimal treatment regimes. Biometrics 68:1010–1018

Zhang B, Tsiatis AA, Laber EB, Davidian M (2013) Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika 100:681–694

Zhang Y, Laber EB, Tsiatis A, Davidian M (2015) Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics 71:895–904

Zhao Y, Zeng D, Rush AJ, Kosorok MR (2012) Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 107:1106–1118

Zhao Y, Zeng D, Socinski MA, Kosorok MR (2011) Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics 67:1422–1433

Zhao Y-Q, Zeng D, Laber EB, Kosorok MR (2015) New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc 110:583–598

Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR (2017) Residual weighted learning for estimating individualized treatment rules. J Am Stat Assoc 112:169–187