Abstract

Emerging infectious diseases are a cause of humanitarian and economic crises across the world. In developing regions, a serious 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, infrastructure, 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., 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 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 varient of the Q-learning 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 model improvement with optimization based on current model estimates by solving a randomly perturbed estimating equation for the optimal treatment regime at each time point. Simulations 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.

1 Introduction

Emerging infectious diseases are a persistent and serious threat to public health across the world [Mathers, 2008, Lozano et al., 2013]; furthermore, 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 a tremendous positive 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 individuals 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 over the course of the epidemic.

Estimating an optimal treatment regime for the management of an emerging infectious disease is complicated as: (i) spillover effects make the number of possible interventions an exponential function of the population size at each time point; (ii) disease dynamics are unknown at the time of outbreak so one must balance choosing interventions that lead to large information gain and subsequently an improved disease dynamics model with choosing interventions that appear to be optimal based on current model estimates; and (iii) evolving resource constraints impose additional restrictions on how and where interventions can be applied. One approach to estimating an an optimal treatment regime is to postulate a model for the disease dynamics and then to use simulation-based optimization to estimate an optimal treatment regime [Laber et al., 2016, Hu et al., 2017]. If the postulated model is low-dimensional or scientific knowledge can be used to construct informative priors, then this approach can be particularly effective early in the epidemic when data are scarce. However, such methods can perform poorly if the postulated model is misspecified. An alternative is to construct a semi-parametric estimator of the optimal treatment

regime that does not require a correctly specified 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] and direct-search estimators [Orellana et al., 2010, Zhang et al., 2012, Zhao et al., 2012, Zhang et al., 2013, 2015, Zhao et al., 2015, Zhou et al., 2017]. Thus, a natural approach is to apply a parametric simulation-optimization approach during early stages of an epidemic and subsequently switch to a semi-parametric estimator as data accumulates. We develop an online semi-parametric estimator of the optimal treatment regime for spatiotemporal decision problems based on Q-learning [Watkins, 1989, Maei et al., 2010, Ertefaie, 2014] and Thompson sampling [Thompson, 1933]. In a series of simulation experiments, applying model-based simulation optimization at early stages of an epidemic and then switching to Q-learning with Thompson sampling when sufficient data are available leads to improved epidemic control when the disease dynamics model is misspecified.

This work is motivated by our involvement in the study the 2013-2015 outbreak of Ebola virus disease in West Africa [Kramer et al., 2016a, Li et al., 2017] which 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]. To inform management of future outbreaks, it is vital to study if and how the spread of the 2013-2015 outbreak could have been better controlled through adaptive treatment allocation. We consider the daily allocation of treatments across 290 contiguous geopolitical regions over a 200 day period; our results indicate that a principled, i.e., data-driven, management strategy can lead to significant reductions in the spread of the disease over *ad hoc* allocation strategies.

In Section 2, we review the 2013-2015 outbreak of Ebola Virus disease. In Section 3, we define an optimal treatment regime and develop Q-learning with Thompson sampling. In Section 3.2,

we propose data-driven feature construction for spatiotemporal treatment regimes. We illustrate the proposed methods using a suite of simulation experiments in Section 4 and a simulation of the spread of Ebola in West Africa in Section 5. Open problems are discussed in Section 6.

2 Ebola Virus

Ebola Virus Disease (EVD) is an acute hemorrhagic illness caused by infection by one of a handful of viruses collectively known as ebolaviruses. The 2013-2016 West Africa Ebola epidemic, caused by the *Zaire ebolavirus*, originated in the Guéckédou Prefecture of Guinea, from which it spead to neighboring Liberia and Sierra Leone. A major outbreak resulting in more than 28,000 cases ensued, igniting small outbreaks in Nigeria, Mali, and the United States. Ebola virus disease may express in a range of symtoms, including fever, muscular pain, vomiting, diarrhea, rash, organ failure, and death Feldmann and Geisbert [2011]. The overall case fatality rate for the West Africa epidemic exceeded 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 if adequate infection control procedures cannot be quickly implemented in health care settings. Additionally, the social disruption caused by Ebola outbreaks – including the need to relinquish traditional burial practices, stigmatization, and fear of government authorities and actions – can make the scope of incipient outbreaks difficult to determine. Outbreaks that spread widely are especially difficult to manage as they encompass heterogeneous populations. Several models for the spread of Ebola in West Africa have been constructed. Gabriel Rainisch et al. [2015] fitted regression models to weekly data on the incidence of infection, finding effects of case counts, population data, and distances between affected and nonaffected areas all to be significant predictors of transmission. Merler et al. [2015] developed an agent-based simulation that both accounted for the early spread of decline in terms of the increasing availability of Ebola treatment units, safe burials, and distribution of

household protection kits. Finally, Kramer et al. [2016b] 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, focusing on the path of spread. They found spread to be best explained by the distance between source and recipient locations, population density and border closures among neighboring countries.

3 Notation and Setup

We consider a decision process evolving in discrete time, $t = 0, 1, \ldots$, and across a finite set of locations $\mathcal{L} = \{0, 1, \ldots, L\}$. In our application to Ebola, the time points correspond to days and the locations are geopolitical units; however, in other applications time points need not be fixed in calendar time. We assume that at each location, ℓ, and each time point, t:

- a set of measurements is taken, these along with past measurements, treatments and outcomes are summarized into what is termed the current state of the location and is denoted $\boldsymbol{S}_t^\ell \in \mathbb{R}^p$;
- the decision maker can select a binary treatment $A_t^\ell \in \{0, 1\}$ so that without any additional restrictions there are 2^L possible treatment allocations at each time point;
- an outcome $Y_t^\ell \in \mathcal{Y}$ is measured.

Define $\boldsymbol{S}_t = \left\{\boldsymbol{S}_t^\ell\right\}_{\ell \in \mathcal{L}}$, $\boldsymbol{A}_t = \left\{A_t^\ell\right\}_{\ell \in \mathcal{L}}$, and $\boldsymbol{Y}_t = \left\{\boldsymbol{Y}_t^\ell\right\}_{\ell \in \mathcal{L}}$. We assume that there exists a function $\psi : \mathbb{R}^p \to 2^{\{0,1\}^L}$ such that $\psi(\boldsymbol{s}) \neq \emptyset$ denotes the set of feasible treatments when the current state is $\boldsymbol{s}$. In addition, we assume that there exists a function $u : \mathcal{Y}^L \to \mathbb{R}$ so that $u(\boldsymbol{y}_t)$ measures the utility of outcome $\boldsymbol{Y}_t = \boldsymbol{y}_t$.

A treatment regime in this context is a map $\pi : \mathbb{R}^p \to \{0, 1\}^L$ which satisfies $\pi(\boldsymbol{s}_t) \in \phi(\boldsymbol{s}_t)$ for all $\boldsymbol{s}_t \in \mathbb{R}^p$; a decision maker following π would apply treatment $\pi(\boldsymbol{s}_t)$, i.e., apply treatment $\{\pi(\boldsymbol{s}_t)\}_\ell$ to location $\ell \in \mathcal{L}$, if presented with $\boldsymbol{S}_t = \boldsymbol{s}_t$ at time t. An optimal treatment regime maximizes

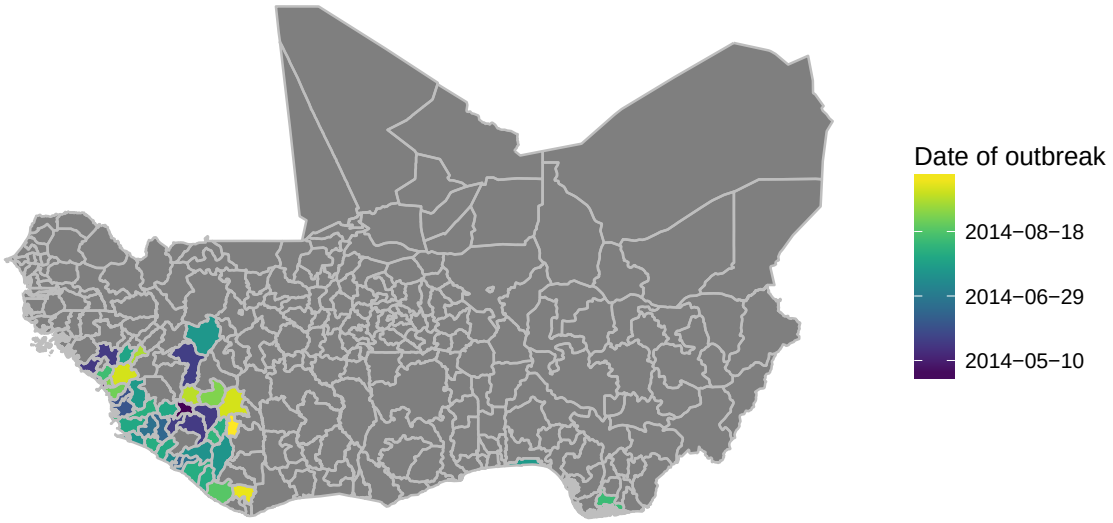

Figure 1: Observed outbreaks for West Africa with the first infections on April 26, 2014.

the mean discounted utility if applied to select treatments at each time point. We formalize this definition using potential outcomes [Rubin, 1974, Robins, 1986, 1987, Splawa-Neyman et al., 1990].

Let an overbar denote history, e.g., $\overline{\boldsymbol{a}}_t = (\boldsymbol{a}_0, \ldots, \boldsymbol{a}_t)$ and $\overline{\boldsymbol{s}}_t = (\boldsymbol{s}_0, \ldots, \boldsymbol{s}_t)$. Define $\boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1})$ to be the potential state under treatment sequence $\overline{\boldsymbol{a}}_{t-1}$ where $\boldsymbol{S}_0(\overline{\boldsymbol{a}}_{-1}) \equiv \boldsymbol{S}_0$; define $\boldsymbol{Y}_t^*(\overline{\boldsymbol{a}}_t)$ to be the potential outcome under treatment sequence $\overline{\boldsymbol{a}}_t$. The potential state at time t under a regime π is $\boldsymbol{S}_t^*(\pi) = \sum_{\overline{\boldsymbol{a}}_{t-1}} \boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1}) \prod_{v=0}^{t-1} 1_{\pi\{\boldsymbol{S}_v^*(\overline{\boldsymbol{a}}_{v-1})\}=a_v}$; similarly the potential outcome at time t is $Y_t^*(\pi) = \sum_{\overline{\boldsymbol{a}}_t} Y_t^*(\overline{\boldsymbol{a}}_t) \prod_{v=0}^{t} 1_{\pi\{\boldsymbol{S}_v^*(\overline{\boldsymbol{a}}_{v-1})\}=a_v}$. Let $\gamma \in (0,1)$ be a fixed discount factor; the value of a regime, π, is defined as $V(\pi) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t u\left\{Y_t^*(\pi)\right\}\right]$ so that the optimal regime, π^{opt}, satisfies $V(\pi^{\text{opt}}) \geq V(\pi)$ for all π.

To identify π^{opt} in terms of the data generating model we make a series of assumptions which are standard in the dynamic treatment regimes literature [Murphy, 2003, Robins, 2004, Schulte et al., 2014] . Let $\boldsymbol{W}^* = \left\{\boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1}), Y_t^*(\overline{\boldsymbol{a}}_t) : \overline{\boldsymbol{a}}_t \in \{0,1\}^{L\times(t+1)}\right\}_{t\geq 0}$ denote the set of potential states and outcomes. We assume that the following hold 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 $\overline{\boldsymbol{s}}_t, \overline{\boldsymbol{a}}_{t-1}$ and $\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)$; and (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, both (A2) and (A3) can be ensured by construction. Furthermore, while we stated (A1) as an assumption, there is some debate as to whether this should instead be taken as an axiom of causal inference [Pearl, 2010]. In addition to (A1)-(A3), 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. Because of spatial interference, 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., 2016]. We assume that the decision process, suitably transformed, is Markov and 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 [Sutton and Barto, 1998, Powell, 2007, Szepesvári, 2010, Puterman, 2014].

3.1 Q-Learning with Thompson Sampling

We assume that the states $\boldsymbol{S}^t$ have been constructed so that the induced decision process is stationary and Markov so that $P\left(\boldsymbol{S}_{t+1} \in \mathcal{B} \middle| \overline{\boldsymbol{S}}_t, \overline{\boldsymbol{A}}_t\right) = P\left(\boldsymbol{S}_{t+1} \in \mathcal{B} \middle| \boldsymbol{S}_t, \boldsymbol{A}_t\right)$ with probability one for any (measurable) set $\mathcal{B} \subseteq \mathbb{R}^p$; furthermore, we assume that the state has be constructed so that the utility $u(\boldsymbol{Y}_t)$ can be written as a function of the triple $(\boldsymbol{S}_t, \boldsymbol{A}_t, \boldsymbol{S}_{t+1})$. Under these assumptions, the optimal treatment regime can characterized using a recursive regression equation known as the Bellman optimality equation [Bellman, 1957, Maei et al., 2010, Puterman, 2014, Ertefaie, 2014]. Define $Q : \mathbb{R}^p \times \{0,1\}^L \rightarrow \mathbb{R}$ as

$$Q(\boldsymbol{s}_t, \boldsymbol{a}_t) \triangleq \mathbb{E}\left[\sum_{k \geq 0} \gamma^k u\left\{\boldsymbol{Y}^*_{t+k}(\pi^{\text{opt}})\right\} \middle| \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right],$$

so that $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the mean discounted utility of being in state $\boldsymbol{s}_t$ taking action $\boldsymbol{a}_t$ and subsequently following the optimal regime. Under (A1)-(A3) it can be shown [Bertsekas et al., 1995] that $\pi^{\text{opt}}(\boldsymbol{s}_t) = \arg\max_{\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)} Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and furthermore that $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ satisfies

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

Consequently, for any function $g : \mathbb{R}^p \times \{0,1\}^L \rightarrow \mathbb{R}^d$ it follows that

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

The preceding equation can be used as the basis of an estimating equation for $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ [Maei et al., 2010, Maei, 2011, Ertefaie, 2014, Luckett et al., 2016].

Let $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta)$ be a class of postulated models for $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ indexed by $\theta \in \Theta$. In our applications, we consider linear models of the form $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta) = \boldsymbol{b}(\boldsymbol{s}_t, \boldsymbol{a}_t)^\intercal \theta$, where $\boldsymbol{b}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ are basis functions; we describe data-driven construction of these features in Section 3.2. We assume that $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta)$ is continuously differentiable in θ for all $(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and we set $g(\boldsymbol{s}_t, \boldsymbol{a}_t) = \nabla_\theta Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta)$. Intuition for this choice comes from viewing (1) as the normal equations for least squares estimation of the regression of $u(Y_t) + \gamma \max_{\boldsymbol{a}_{t+1} \in \Psi(\boldsymbol{S}_{t+1})} Q(\boldsymbol{S}_{t+1}, \boldsymbol{a}_{t+1})$ on $(\boldsymbol{S}_t, \boldsymbol{A}_t)$ using the class of models $\{Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta) : \theta \in \Theta\}$ [see Baird et al., 1995, Maei, 2011, for additional insights]. Define

$$\Lambda_t(\theta) = \sum_{v=0}^{t-1} \left\{ u(Y_v) + \gamma \max_{\boldsymbol{a}_{v+1} \in \psi(\boldsymbol{S}_{v+1})} Q(\boldsymbol{S}_{v+1}, \boldsymbol{a}_{v+1}; \theta) - Q(\boldsymbol{S}_v, \boldsymbol{A}_v; \theta) \right\} \nabla Q(\boldsymbol{S}_v, \boldsymbol{A}_v; \theta),$$

and let $\widehat{\theta}_t$ be a solution to $\Lambda_t(\theta) = 0$. Thus, $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \widehat{\theta}_t)$ is the estimator of $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and subsequently the estimated optimal treatment regime is $\widehat{\pi}_t(\boldsymbol{s}_t) = \arg\max_{\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)} Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \widehat{\theta}_t)$. However, consistent estimation the optimal regime requires indefinite exploration of the space of candidate treatment regimes [Robbins, 1952, Berry and Fristedt, 1985, Kaelbling et al., 1996, Sutton and Barto, 1998]; selecting actions according to $\widehat{\pi}_t$ would make treatment selection a deterministic function $\overline{\boldsymbol{S}}_t, \overline{\boldsymbol{A}}_{t-1}$ thereby violating the positivity assumption (A2) and potentially precluding consistent estimation (Robins [2004], Schulte et al. [2014] see Laber et al. [2016] for a demonstrative example).

There are several ways one could force exploration into online estimation of an optimal treatment regime. In the reinforcement literature, a common approach is to soften or relax the maximization operator [Sutton and Barto, 1998]. Soft-max approaches select an action $\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)$ with probability proportional to a non-decreasing function of $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \widehat{\theta}_t)$, e.g., $P\left(\boldsymbol{A}_t = \boldsymbol{a}_t \middle| \boldsymbol{S}_t = \boldsymbol{s}_t\right\} \propto$

$\exp\left\{\alpha_t Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \widehat{\theta}_t)\right\}$ where $\{\alpha_t\}_{t\geq 0}$ is a sequence of positive tuning parameters. Closely related are ϵ-greedy methods which select $\boldsymbol{A}_t = \arg\max_{\boldsymbol{a}_t \in \psi(\boldsymbol{s}_t)} Q(\boldsymbol{S}_t, \boldsymbol{a}_t; \widehat{\theta}_t)$ with probability $(1-\epsilon)$ and select an action uniformly at random from the set $\psi(\boldsymbol{S}_t)$ with probability ϵ, where $\epsilon \in (0,1)$ is a tuning parameter. However, in spatio-temporal applications like the ones considered here, the size of the set $\psi(\boldsymbol{S}_t)$ is exponential in the number of locations making such sampling computationally intractable; furthermore, in the context of an emerging epidemic, uniform exploration of the treatment space may not be ethical.

Thompson sampling [Thompson, 1933] was originally proposed as a Bayesian approach to forcing 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; in addition, a fully Bayesian approach for large spatio-temporal decision problems can be computationally burdensome [Laber et al., 2016]. Instead, we use an approximate Thompson sampling algorithm wherein we: (i) perturb the estimating equation $\Lambda_t(\theta) = 0$; (ii) use the root of the perturbed equation, say $\widetilde{\theta}_t$, as an approximate draw from the sampling distribution of $\widehat{\theta}_t$; and (iii) select action $\boldsymbol{A}_t = \arg\max_{\boldsymbol{a}_t \in \psi(\boldsymbol{S}_t)} Q\left(\boldsymbol{S}_t, \boldsymbol{a}_t; \widetilde{\theta}_t\right)$. Let θ^* denote the true parameter value so that $Q(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta^*)$. It can be seen that $\Lambda_t(\theta^*)$ is a sum of martingale differences, thus we use a perturbation based on the generalized bootstrap [Jin et al., 2001, Chatterjee et al., 2005, Minnier et al., 2011]. For each t, let $W_{t,0}, \ldots, W_{t,t-1}$ be independent and identically distributed exponential random variables with scale parameter one and define the perturbed objective function

$$\widetilde{\Lambda}_t(\theta) = \sum_{v=0}^{t-1} W_{t,v} \left\{ u(Y_v) + \gamma \max_{\boldsymbol{a}_{v+1} \in \psi(\boldsymbol{S}_{v+1})} Q(\boldsymbol{S}_{v+1}, \boldsymbol{a}_{v+1}; \theta) - Q(\boldsymbol{S}_v, \boldsymbol{A}_v; \theta) \right\} \nabla Q(\boldsymbol{S}_v, \boldsymbol{A}_v; \theta),$$

so that $\widetilde{\theta}_t$ is a root of $\widetilde{\Lambda}_t(\theta)$ and the selected action at time t is $\boldsymbol{A}_t = \arg\max_{\boldsymbol{a}_t \in \psi(\boldsymbol{S}_t)} Q(\boldsymbol{S}_t, \boldsymbol{a}_t; \widetilde{\theta}_t)$.

Remark 3.1. The proposed estimator can be viewed as an approximate and semiparametric variant of Thompson sampling in the following sense. The sampling distribution of the perturbed estimator $\widetilde{\theta}_t$ is designed to mimic the sampling distribution of $\widehat{\theta}_t$ which in turn can be viewed as a surrogate for a posterior distribution of θ^*.

3.2 Feature Construction

An appropriate class of models for the Q-function depends on the domain of application; however, the Q-function is a complex mathematical object and its form may not be easy to directly elicit from domain experts. In the context of emerging infectious diseases, there are often theory-based models of disease spread that can be used to derive a class of models for the Q-function.

Let $\mathcal{M} = \{M_j\}_{j=1}^J$ denote a class of candidate parametric models of disease spread so that each $M \in \mathcal{M}$ specifies a distribution over the distribution of $\boldsymbol{S}_1$ and the conditional distribution of $\boldsymbol{S}_{t+1}$ given $\boldsymbol{S}_t$ and $\boldsymbol{A}_t$ is indexed by $\beta_M \in \mathbb{R}^{p_M}$. For each $M \in \mathcal{M}$ and $\beta_M \in \mathbb{R}^{p_M}$, let $\mathbb{E}_{M,\beta_M}$ denote expectation with respect to model M under parameters β_M and let $\pi^{\text{opt}}_{M,\beta_M}$ denote the optimal treatment regime under model M and parameters β_M. Define the Q-function under model M and parameters β_M as

$$Q_{M,\beta_M}(\boldsymbol{s}_t, \boldsymbol{a}_t) \triangleq \mathbb{E}_{M,\beta_M}\left[\sum_{k\geq 0} \gamma^k u\left\{\boldsymbol{Y}^*_{t+k}(\pi^{\text{opt}}_{M,\beta_M})\right\} \middle| \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right]. \tag{2}$$

One can approximate $Q_{M,\beta_M}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ by simulating data under M and β_M and solving the estimating equation (1) over a flexible class of models; in our application, we used a nonlinear basis expansion (see the Supplemental Materials for details.)

Given the history at time t, let $\widehat{\beta}_{M,t}$ be an estimator of the parameters indexing $M \in \mathcal{M}$, e.g.,

based on maximum likelihood, and let $\widehat{Q}_{M,\widehat{\beta}_{M,t},t}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ be an be an estimator of $Q_{M,\widehat{\beta}_{M,t}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. At each time t, we consider models for $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ of the form $Q(\boldsymbol{s}_t, \boldsymbol{a}_t; \theta) = \theta_0 + \sum_{j=1}^{J} \theta_j \widehat{Q}_{M,\widehat{\beta}_{M,t}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. If any of the postulated disease dynamics models are correctly specified, than the preceding model will also be correctly specified. It is clear that there are many possible ways in which the preceding model could be expanded, e.g., directly adding basis functions of $\boldsymbol{s}_t$ or taking a non-linear function of the model-based Q-functions. However, we do not consider such extensions here.

4 Simulation experiments

We evaluate the finite sample performance of Q-learning with Thompson Sampling in suite of simulated examples of disease spread over nodes in a network. Thus, we assume that each location is associated with a set of neighbors which might be defined, for example, as sharing a border or being sufficiently close according to some distance measure; we assume that the definition of a neighbor is symmetric so that if location ℓ is a neighbor of location ℓ' then ℓ' is neighbor of ℓ. Let $\Omega \in \{0,1\}^{L \times L}$ denote the adjacency matrix encoding this neighborhood relation, i.e., $\Omega_{\ell,\ell'} = 1$ if ℓ and ℓ' are neighbors and zero otherwise.

To explore the impact of model-misspecification and to facilitate comparison with a model-based estimated optimal treatment regime, we consider disease models that are representable as a mixture of two susceptible-infected-susceptible (SIS) models [Weiss and Dishon, 1971]. Define the outcome $Y_\ell^t \in \{0,1\}$ to be an indicator that location ℓ is infected at time t and define the utility function $u(\boldsymbol{y}_t) = \sum_{\ell=1}^{L} y_t^\ell$ to be the number of infected locations at time t. The disease dynamics are determined by: (i) the distribution of the initial state $\boldsymbol{S}_1$; (ii) the conditional distribution of the state $\boldsymbol{S}_{t+1}$ given $(\boldsymbol{S}_t, \boldsymbol{Y}_t, \boldsymbol{A}_t)$; and (iii) the conditional distribution of the outcome $\boldsymbol{Y}_{t+1}$ given $(\boldsymbol{S}_t, \boldsymbol{Y}_t, \boldsymbol{A}_t)$. In our examples, the state measured at location ℓ at time t is univariate and real-valued. Define $\mathcal{I}_t = \left\{\ell \in \mathcal{L} : Y_t^\ell = 1\right\}$ to be the set of infected locations at time t, define

$\text{logit}(u) = \log\{u/(1-u)\}$ to be the logit function, and let $\phi(u)$ denote the probability density function of a standard normal random variable. We initialize the state to be identically zero at outbreak, i.e., $\boldsymbol{S}_1 \equiv 0$ and the evolution of the state follows an AR(1) process so that the conditional density of $\boldsymbol{S}_{t+1}$ given $(\boldsymbol{S}_t, \boldsymbol{Y}_t, \boldsymbol{A}_t) = (\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t)$ is

$$f\left(\boldsymbol{s}_{t+1}\middle|\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \zeta, \tau\right) = \prod_{\ell=1}^{L} \tau^{-1}\phi\left(\frac{s_{t+1,\ell} - \zeta s_{t,\ell}}{\tau}\right),$$

where $\tau > 0$ and $\zeta \in \mathbb{R}$. The outcome distribution is generated as follows: $Z_t \sim \text{Bernouilli}(\omega)$ and

$$\begin{aligned} &P\left(\boldsymbol{Y}_{t+1} = \boldsymbol{y}_{t+1}\middle|\boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{A} = \boldsymbol{a}_t, Z_t = z_t; \varsigma\right\} \\ &\qquad = \prod_{\ell \in \mathcal{I}_t} q_\ell\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)^{1-y_{t+1,\ell}} \left\{1 - q_\ell\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\}^{y_{t+1,\ell}} \\ &\qquad\qquad \times \prod_{\ell \in \mathcal{I}_t^c} p_\ell\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)^{y_{t+1,\ell}} \left\{1 - p_\ell\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\}^{1-y_{t+1,\ell}}, \end{aligned}$$

where $p_\ell\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right) = 1 - \left\{1 - p_{\ell,0}\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\} \prod_{\ell' \,:\, \Omega_{\ell,\ell'}=1} \left\{1 - p_{\ell,\ell'}\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\}$ and

$$\begin{aligned} \text{logit}\left\{p_{\ell,0}\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\} &= \varsigma_0 + \varsigma_1 a_\ell 1_{z_t s_t^\ell \le 0} \\ \text{logit}\left\{p_{\ell,\ell'}\left(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma\right)\right\} &= \varsigma_2 + \varsigma_3 a_\ell 1_{z_t s_t^\ell \le 0} + \varsigma_4 a_{\ell'} 1_{z_t s_t^{\ell'} \le 0} \\ \text{logit}\left\{q_\ell(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t, z_t; \varsigma)\right\} &= \varsigma_5 + \varsigma_6 1_{z_t s_t^\ell \le 0}. \end{aligned}$$

Thus, the preceding model is indexed by the ten parameters $\zeta, \tau, \omega, \varsigma = (\varsigma_0, \varsigma_1, \ldots, \varsigma_6)$. When $\omega = 0$ the dynamics depend only on which locations are infected at each time point whereas when $\omega = 1$ the state of each location is a measure of its resistance to treatment; values of $0 < \omega < 1$ correspond to mixtures over these two models.

To study the effects of different spatial topologies we consider the spread over three different

network types: a lattice, Barabasi-Albert network [Barabási and Albert, 1999], and a random three-nearest-neighbors network; see Figure 2 for examples of these networks. We consider networks of size $L = 100, 500$, and 1000. For each network type and value of ω, we tune the remaining parameter values to attain a fixed rate of infection after X time points under no treatment; a description of this tuning procedure and the parameter values are provided in the Supplemental Materials.

In our implementation we used the features as described in Section 3.2 with Q-functions corresponding to models with $\omega = 0$ and $\omega = 1$; thus, the model is misspecified except at these two extremes. For comparison, we implemented the following five competing methods: (i) no treatment, do not apply treatment to any locations; (ii) random, select locations to be treated uniformly at random; (iii) proximal, treat locations that are closest to locations of the opposite infection status; (iv) myopic, treat locations with the highest estimated probability of being infected at the next time point based on a correctly specified disease model; and (v) model-based policy search as described in Laber et al. [2016] assuming that $\omega \equiv 0$ which first estimates disease dynamics and then uses simulation optimization (see the Supplemental Materials for a a description). The results based on 50 Monte Replications are displayed in Figure (3). As anticipated the model-based method performs well under a correctly specified model ($\omega = 0$) but this performance deteriorates as the model becomes increasingly misspecified. Q-learning with Thompson Sampling consistently performs well across the range of models, networks, and number of locations.

5 Simulation of Ebola virus disease in West Africa

RESUME HERE

Kramer et al. [2016a] modeled the spread of the Ebola Virus using multiple models for the system dynamics. The model chosen to best fit the observed infections is called the gravity model. With this model, there is no recovery from infection. Thus $\text{logit}[q_\ell(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta)] = 0.0$, and the

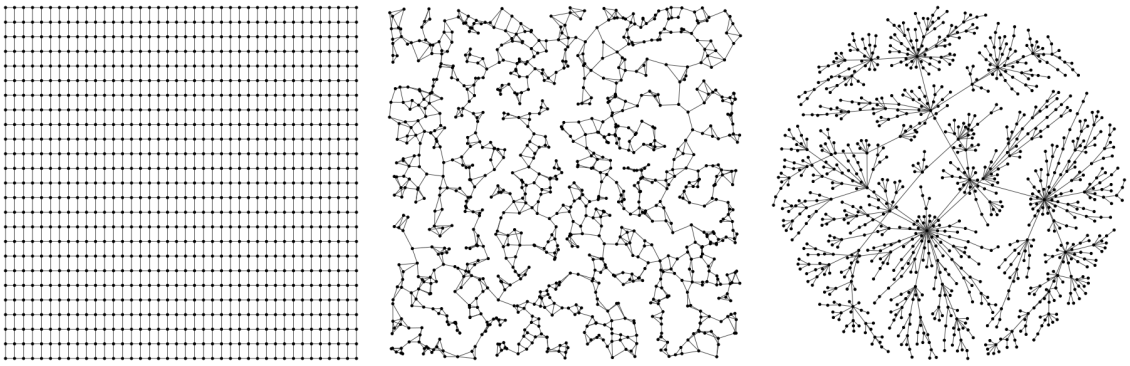

Figure 2: Instances of the three network structures used in the simulation examples. **Left**: lattice network with 1000 locations. **Center**: random three-nearest-neighbor network with 1000 locations. **Right**: Barabasi-Albert network with 1000 locations.

probability of infection is defined by

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

To set the parameter value for the generative model, we estimate the maximum likelihood estimator $\widehat{\eta}^{MLE}$ from the observed data. Then the generative model is tuned to have two properties. One: under no treatment, the proportion of locations infected after 25 time points when starting at the observed infections is 70%. Two: when treating all locations, there should be a 95% reduction in new infections after 25 time points as compared to no treatment. For condition one, the infection rate is tuned by setting α such that $\mathbb{E}^{\pi_1}[\mathbf{1}^\intercal \boldsymbol{Y}_{25}] = 0.7L$ where the generative model is $\{\alpha\widehat{\eta}_0^{MLE}, \log(\alpha) + \eta_1^{MLE}, \eta_2^{MLE}, 0.0, 0.0\}$ and π_1 applies no treatment. For condition two, the treatment effects are tuned by setting β such that $\mathbb{E}^{\pi_2}[\mathbf{1}^\intercal(\boldsymbol{Y}_{25} - \boldsymbol{Y}_0)] = 0.05\mathbb{E}^{\pi_1}[\mathbf{1}^\intercal(\boldsymbol{Y}_{25} - \boldsymbol{Y}_0)]$ where the generative model is $\{\alpha\widehat{\eta}_0^{MLE}, \log(\alpha) + \eta_1^{MLE}, \eta_2^{MLE}, \beta, \beta\}$ and π_2 applies treatment to all locations.

We present results for management of the Ebola virus in table 1. Each simulation replication started from the observed infections and simulated 25 time points. There are 50 replications for each treatment strategy. From these results we can see that the model based policy search method out performs all other competing strategies.

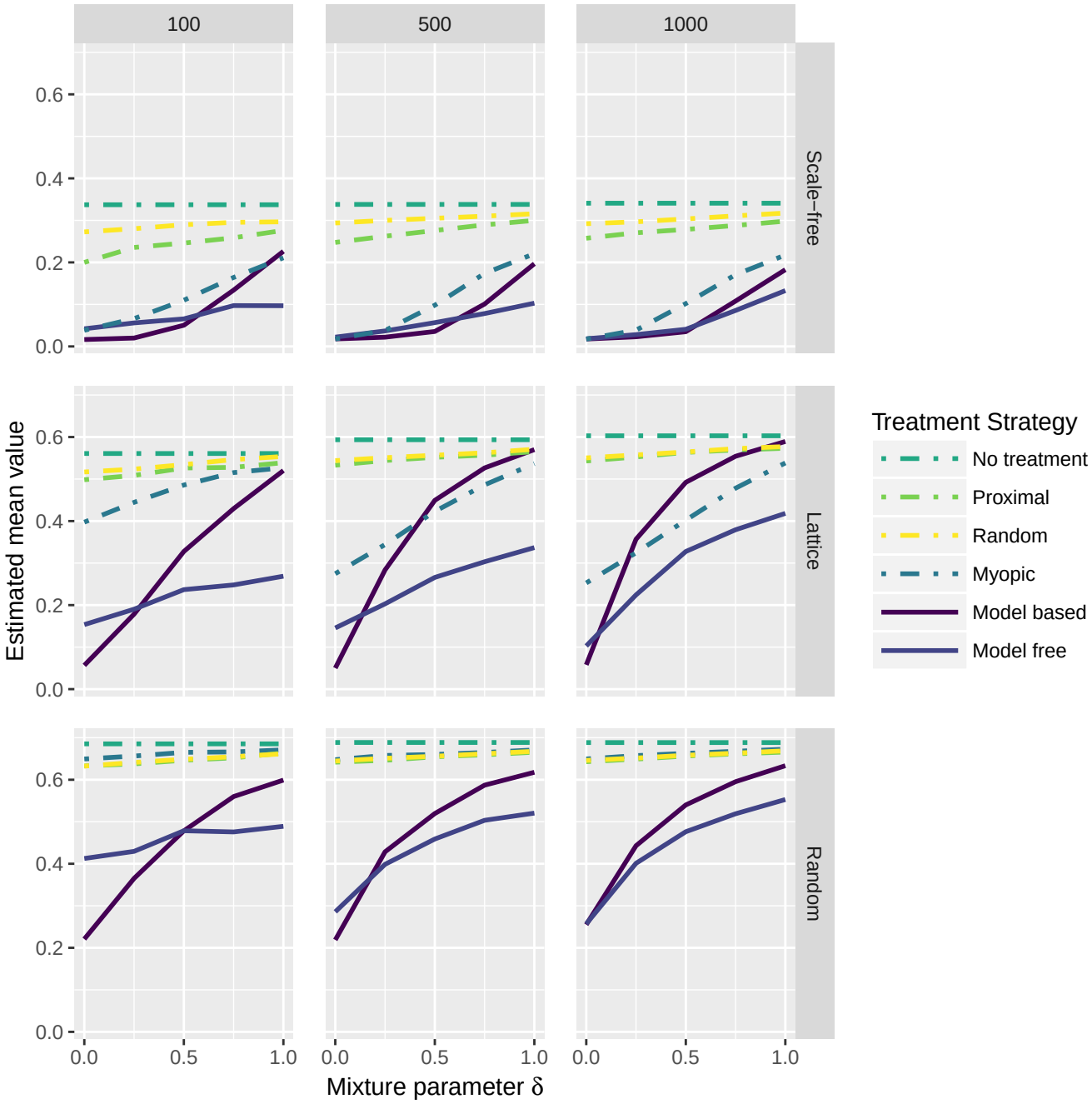

Figure 3

None	Random	Proximal	Myopic	Model based
0.6919 (0.0038)	0.6403 (0.0040)	0.6114 (0.0040)	0.5819 (0.0041)	0.5025 (0.0042)

Table 1: Simulation results for the management of the Ebola Virus.

6 Conclusion

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135, 2013.

Daniel Almirall, Thomas Ten Have, and Susan A Murphy. Structural nested mean models for assessing time-varying effect moderation. *Biometrics*, 66(1):131–139, 2010.

Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pages 30–37, 1995.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.509. URL `http://science.sciencemag.org/content/286/5439/509`.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957. URL `http://books.google.com/books?id=fyVtp3EMxasC&pg=PR5&dq=dynamic+programming+richard+e+bellman&client=firefox-a#v=onepage&q=dynamic%20programming%20richard%20e%20bellman&f=false`.

Donald A Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Springer, 1985.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Gary Cecchine and Melinda Moore. *Infectious disease and national security: strategic information needs*. Rand Corporation, 2006.

Bibhas Chakraborty and EE Moodie. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

Snigdhansu Chatterjee, Arup Bose, et al. Generalized bootstrap for estimating equations. *The Annals of Statistics*, 33(1):414–436, 2005.

Ashkan Ertefaie. Constructing dynamic treatment regimes in infinite-horizon settings. *arXiv preprint arXiv:1406.0764*, 2014.

Heinz Feldmann and Thomas W Geisbert. Ebola haemorrhagic fever. *Lancet*, 377(9768):849–862, 5 March 2011. URL `http://dx.doi.org/10.1016/S0140-6736(10)60667-8`.

Gabriel Rainisch, Manjunath B. Shankar, Michael Wellman, Toby Merlin, and Martin I. Meltzer. Regional spread of ebola virus, west africa, 2014. *Emerging Infectious Disease journal*, 21(3): 444, 2015. URL `http://wwwnc.cdc.gov/eid/article/21/3/14-1845`.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *ICML*, volume 14, pages 100–108, 2014.

Mary J Hamel and Laurence Slutsker. Ebola: the hidden toll. *The Lancet Infectious Diseases*, 15 (7):756–757, 2015.

Robin Henderson, Phil Ansell, and Deyadeen Alshibani. Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 66(4):1192–1201, 2010.

T. Hu, E.B. Laber, N.J. Meyer, K. Pacifici, and J. Drake. Note on thompson sampling for large decision problems. *Under review*, 1(2):1–10, 2017.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 2008.

Zhezhen Jin, Zhiliang Ying, and Lee-Jen Wei. A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390, 2001.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

Michael R Kosorok and Erica EM Moodie. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.

Andrew M. Kramer, J. Tomlin Pulliam, Laura W. Alexander, Andrew W. Park, Pejman Rohani, and John M. Drake. Spatial spread of the west africa ebola epidemic. *Royal Society Open Science*, 3(8), 2016a. doi: 10.1098/rsos.160294. URL `http://rsos.royalsocietypublishing.org/content/3/8/160294`.

Andrew M Kramer, J Tomlin Pulliam, Laura W Alexander, Andrew W Park, Pejman Rohani, and John M Drake. Spatial spread of the west africa ebola epidemic. *Royal Society Open Science*, 3 (8), 2016b. URL `http://dx.doi.org/10.1098/rsos.160294`.

E.B. Laber, N.J. Meyer, B.R. Reich, J.K. Pacifici, J.A. Collazo, and J. Drake. On-line estimation of

an optimal treatment allocation strategy for the control of white-nose syndrome in bats. *Under review*, 11(2):1–25, 2016.

E.B. Laber, E.J. Rose, M. Davidian, and A.A. Tsiatis. Q-learning. *Wiley StatsRef*, doi: 10.1002/9781118445112.stat07998, 2017.

Shou-Li Li, Ottar N. Bjrnstad, Matthew J. Ferrari, Riley Mummah, Michael C. Runge, Christopher J. Fonnesbeck, Michael J. Tildesley, William J. M. Probert, and Katriona Shea. Essential information: Uncertainty and optimal control of ebola outbreaks. *Proceedings of the National Academy of Sciences*, 2017. doi: 10.1073/pnas.1617482114. URL `http://www.pnas.org/content/early/2017/05/10/1617482114.abstract`.

Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. 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. *The Lancet*, 380(9859):2095–2128, 2013.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *arXiv preprint arXiv:1611.03531*, 2016.

Hamid R Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 719–726, 2010.

Hamid Reza Maei. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

Colin Mathers. *The global burden of disease: 2004 update*. World Health Organization, 2008.

Stefano Merler, Marco Ajelli, Laura Fumanelli, Marcelo F C Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis L Chao, Ira M Longini, Jr, M Elizabeth Halloran, and Alessandro Vespignani. 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(2):204–211, 2015. URL `http://www.sciencedirect.com/science/article/pii/S1473309914710746`.

Jessica Minnier, Lu Tian, and Tianxi Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382, 2011.

Erica EM Moodie, Nema Dean, and Yue Ru Sun. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243, 2014.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6 (Jul):1073–1097, 2005.

Liliana Orellana, Andrea Rotnitzky, and James M Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The international journal of biostatistics*, 6(2), 2010.

Judea Pearl. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872–875, 2010.

Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9-12):1393–1512, 1986.

James M Robins. Addendum to a new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640, 2014.

Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Katherine F Smith, Michael Goldberg, Samantha Rosenthal, Lynn Carlson, Jane Chen, Cici Chen, and Sohini Ramachandran. Global rise in human infectious disease outbreaks. *Journal of the Royal Society Interface*, 11(101):20140950, 2014.

Jerzy Splawa-Neyman, DM Dabrowska, TP Speed, et al. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

R.S. Sutton and A.G. Barto. *Reinforcment Learning: An Introduction*. The MIT Press, 1998.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989. URL `http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf`.

George H. Weiss and Menachem Dishon. On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences*, 11(34):261 – 265, 1971. ISSN 0025-5564. doi: http://dx.doi.org/10.1016/0025-5564(71)90087-3. URL `http://www.sciencedirect.com/science/article/pii/0025556471900873`.

WHO Ebola Response Team. Ebola virus disease in west africathe first 9 months of the epidemic and forward projections. *N Engl J Med*, 2014(371):1481–1495, 2014.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Yichi Zhang, Eric B Laber, Anastasios Tsiatis, and Marie Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904, 2015.

Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.