Model-free estimation of the optimal treatment strategy for dynamical spatio-temporal systems

Eric B. Laber Nicholas J. Meyer John Drake Drew Kramer

1 Introduction

Sequential decision problems that evolve over both time and space cross many different disciplines of science. Some applications include infectious disease management, power grid regulation, and cyber security. Solving these problems requires a deep understanding of the complex underlying dynamical system and the ability to control the system effectively using a treatment or stimulus. If not carefully monitored, these problems could have significant impacts on global financial sectors, environmental stability, and human health. For example, infectious diseases, such as Ebola (**??**), have taken countless lives and have impacted financial markets due to public concern. Understanding these systems and having the ability to influence their impact on society is of great importance.

A critical component to managing these systems is the ability to apply treatment effectively. A treatment strategy uses up-to-date information about the system and determines where to place treatment. Typically, there are constrains that restrict the ability to apply treatments, e.g., location based availability of certain treatments or financial limitations on type and quantity. When managing these systems, experts are often concerned with certain system related outcomes. They employ treatment strategies that aim to, assuming larger is better, maximize the expected outcome. The treatment that maximizes the expected outcome is called the optimal treatment strategy. In the case of disease management, the outcome of concern might be the average number of daily infections and a quality estimate of the optimal strategy could result in many lives saved.

Estimating an optimal treatment strategy has three main challenges. One, dimension of the decision space is large and increases tremendously fast with the size of the system due to the spatial component. The large dimensions lead to expensive computational costs that restrict the ability to use existing methodology. Two, the dynamical systems that govern the evolution are often complex. This places an emphasis on robustness to model misspecification. Three, in this setting, contrary to many sequential decision problems, only a single trajectory of the system is observed. This creates challenges when estimating the optimal strategy and increases the importance of intelligent exploration and exploitation. In this paper, we present an approach for estimating optimal treatment strategies that scales will with network size, is robust to model misspecification, and produces quality results with limited data.

Many methods for estimating optimal treatment strategies focus on controlling a single unit over repeated episodes (**??**). There is often thousands of observations which can be used to estimate a strategy. Other methods, such as those focused on dynamic treatment regimes, estimate treatment strategies for many units, but they are assumed to be independent (**??**). Thus, applying treatment to one unit does not effect the outcome of another unit. In the context of controlling an epidemic, the spatial locations are not independent units and observed data is scarce. For epidemic management, **?** propose a sequential rank based treatment strategy that uses model based policy search. Their method estimates the optimal strategy by postulating a working model of the disease dynamics and using simulation optimization. When the disease dynamics model is correctly specified, the method performs well. However, their results show a significant decrease in performance when the model is misspecified.

In the following sections, we define the problem and go into detail about the method and its performance. We discuss the motivating problem of the Ebola virus in Section **??**. Section **??** defines notation and formalizes the problem. Leading up to the methods, section **??** describes features used to construct treatment strategies. In section **??**, we propose a model-based approach to estimating a treatment strategy and in section **??** we extend this approach and describe a model-free method to constructing a treatment strategy. Models for the system dynamics are discussed in section **??**. Results of simulations for toy networks are presented in section **??** and a case study for the Ebola virus is presented in section **??**. Finally, we provide concluding remarks in section **??**.

2 Ebola Virus

The Ebola virus affected more than 25,000 people in three countries of West Africa between 2013-2015. Figure **??** shows the observed outbreaks of the Ebola virus in West Africa. Using data to understand the dynamics of the virus can help control future outbreaks. **?** present an analysis of the Ebola virus by comparing many different models for the virus' spread dynamics. By comparing estimated transition models, the analysis provides insight into how the disease spreads and can aid forecasts of future outbreaks.

We build on this work and extend the modeling to include interventions. Understanding the spread dynamics is important, but it is also critical to investigate the effect of possible treatments. Effectively predicting effects of treatment could result in many lives saved. Using models from this paper, we study the impact of various treatment strategies on the Ebola virus. We use the observed data to provide realistic spread dynamics and simulate the outbreaks in West Africa under intervention.

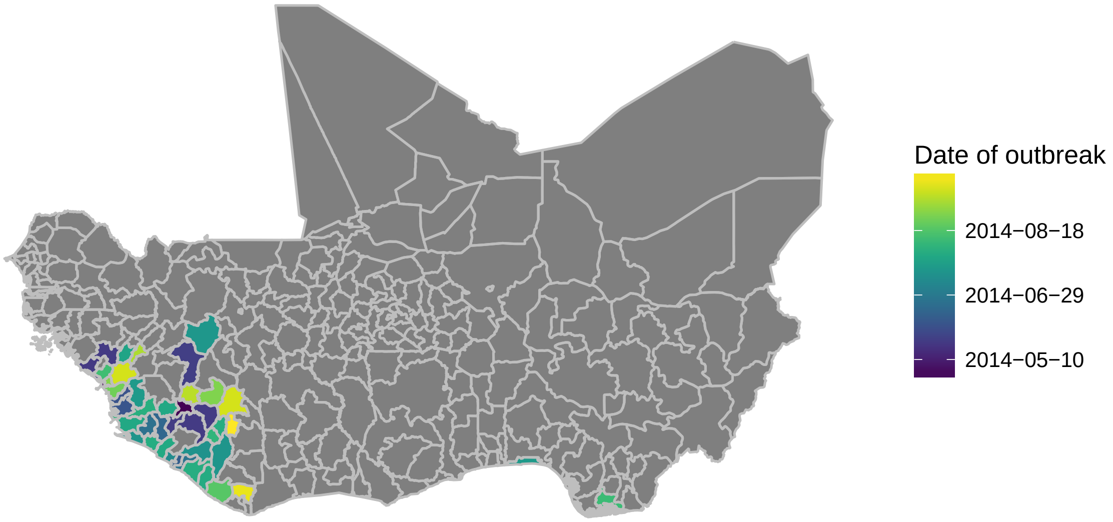

Figure 1: Observed outbreaks for West Africa with the first infections on April 26, 2014.

3 Notation and Setup

We formalize the treatment allocation process as a spatio-temporal decision problem evolving over a finite undirected network with decisions at a countably infinite set of time points. Each location in the network is a susceptible entity and when infected it may transmit the disease to adjacent locations. The state of the system at each decision point includes infection status and covariate information for each location. Treatments are binary for each location; a location can be treated or left untreated. Uninfected locations may receive a preventative treatment and infected locations may receive an active treatment. At each decision point, information about the state of the system is observed and utilized to decide which locations should receive treatment. A treatment strategy is a mapping from up-to-date state information to a distribution over all feasible treatment allocations.

The network of locations $\mathcal{L} = \{1, \ldots, L\}$ is defined by the adjacency matrix $\Omega = \{0,1\}^L$; let $\mathcal{N}_\ell = \{\ell' \neq \ell \ : \ \Omega_{\ell,\ell'} = 1\}$ be the set of neighbors of location ℓ. At time points $\mathcal{T} = \{0, 1, 2, \ldots\}$, the network is observed. Define $\boldsymbol{S}_{t,\ell} \in \mathbb{R}^p$ as the state information collected at time t for location ℓ and $\boldsymbol{S}_t = \{\boldsymbol{S}_{t,\ell}\}_{\ell \in \mathcal{L}}$. Let $A_{t,\ell} \in \{0,1\}$ be the indicator that location ℓ is treated at time t and $\boldsymbol{A}_t = \{A_{t,\ell}\}_{\ell \in \mathcal{L}}$. At each time point t, the infection status for location ℓ is $Y_{t,\ell} \in \{0,1\}$ and $\boldsymbol{Y}_t = \{Y_{t,\ell}\}_{\ell \in \mathcal{L}}$. Define $\boldsymbol{H}_t = \{\boldsymbol{S}_0, \boldsymbol{Y}_0, \boldsymbol{A}_0, \boldsymbol{S}_1, \boldsymbol{Y}_1 \ldots, \boldsymbol{S}_t, \boldsymbol{Y}_t\}$ to be the history at time t.

Define Π to be the class of all treatment strategies under consideration. Below we define what it means for a treatment strategy π to be optimal with respect to the class Π and we subsequently formalize this notion using potential outcomes. Let $\mathcal{B}_L$ to be the set of all $\{0,1\}^L$-valued random variables. A treatment strategy $\pi = \{\pi_t\}_{t \in \mathcal{T}}$ is a sequence of decision rules, one for each time point in $\mathcal{T}$. Each rule, π_t, is a map from dom $\boldsymbol{S}_t \times$ dom $\boldsymbol{Y}_t$ to $\mathcal{B}_L$. Define the outcome of the disease at time t to be $u(\boldsymbol{Y}_t)$ for some functional $u : \{0,1\}^L \to \mathbb{R}$. A strategy is optimal if it maximizes the mean of the cumulative discounted outcome $\sum_{t \geq 1} \gamma^{t-1} u(\boldsymbol{Y}_t)$.

We formalize the definition of an optimal strategy using potential outcomes. Define $\mathcal{F} =$

$\{\boldsymbol{a} : P\{\pi(\boldsymbol{s}, \boldsymbol{y}) = \boldsymbol{a}\} > 0$ for some $\pi \in \Pi$, $\boldsymbol{s}$, and $\boldsymbol{y}\}$. We use overline notation to denote past history, so $\overline{a}_t = \{a_1, \ldots, a_t\}$, and we use an asterisk superscript to denote potential outcomes. Let $W^* = \{\boldsymbol{S}_t^*(\overline{\boldsymbol{a}}_{t-1}), \boldsymbol{Y}_t^*(\overline{\boldsymbol{a}}_{t-1}) : \boldsymbol{a}_1, \ldots, \boldsymbol{a}_t \in \mathcal{F}\}_{t \in \mathcal{T}}$ be the set of possible potential outcomes when following a strategy in the class Π. The potential outcome Y_t^* under strategy π is defined as $\boldsymbol{Y}_t^*(\pi) = \sum_{\overline{\boldsymbol{a}}_{t-1}} \boldsymbol{Y}_t^*(\overline{\boldsymbol{a}}_{t-1}) \prod_{v=0}^{t-1} \mathbb{1}\left[\pi\left\{\overline{\boldsymbol{S}}_v^*(\overline{\boldsymbol{a}}_{v-1}), \overline{\boldsymbol{Y}}_v^*(\overline{\boldsymbol{a}}_{v-1})\right\} = \boldsymbol{a}_v\right]$. The optimal strategy, π^*, satisfies $\mathbb{E}\left[\sum_{t\geq 1} \gamma^{t-1} u(\boldsymbol{Y}_t(\pi^*))\right] \geq \mathbb{E}\left[\sum_{t \geq 1} \gamma^{t-1} u(\boldsymbol{Y}_t(\pi))\right]$ for all $\pi \in \Pi$. To estimate the value under a strategy, π, we make the following assumptions: (A1) sequential ignorability, $\boldsymbol{A}_t \perp W^* | \boldsymbol{H}^t$; (A2) consistency, $\boldsymbol{Y}_t = \boldsymbol{Y}_t^*(\overline{\boldsymbol{A}}_t)$ and $\boldsymbol{S}_t = \boldsymbol{S}_t^*(\overline{\boldsymbol{A}}_t)$; (A3) positivity, $P(\boldsymbol{A}_t = \boldsymbol{a}|\boldsymbol{H}_t) > \epsilon$ with probability 1 for all $\boldsymbol{a} \in \mathcal{F}$. Under the assumptions $(A1) - (A3)$, the value function, $V^\pi(\boldsymbol{s}, \boldsymbol{y}) = \mathbb{E}\left[\sum_{t \geq 1} \gamma^{t-1} u\left\{\boldsymbol{Y}_t^*(\pi)\right\} \middle| \boldsymbol{S}_0 = \boldsymbol{s}, \boldsymbol{Y}_0 = \boldsymbol{y}\right]$ has the form

$$\lim_{T \to \infty} \int \left[\sum_{t=1}^{T} \gamma^{t-1} u(\boldsymbol{y}_t)\right] \prod_{t=0}^{T-1} f_{t+1}(\boldsymbol{s}_{t+1}|\boldsymbol{h}_t, \boldsymbol{a}_t) g_{t+1}(\boldsymbol{y}_{t+1}|\boldsymbol{h}_t, \boldsymbol{a}_t) P\left\{\pi(\boldsymbol{s}_t, \boldsymbol{y}_t) = \boldsymbol{a}_t\right\} d\lambda(\overline{\boldsymbol{s}}_T, \overline{\boldsymbol{a}}_T, \overline{\boldsymbol{y}}_T) \tag{1}$$

where f_{t+1} is the conditional distribution of $\boldsymbol{S}_{t+1}$ given $\boldsymbol{H}_t$ and $\boldsymbol{A}_t$, g_{t+1} is the conditional distribution of $\boldsymbol{Y}_{t+1}$ given $\boldsymbol{H}_t$ and $\boldsymbol{A}_t$, and $\lambda(\cdot)$ is a dominating measure.

4 Features for Constructing Treatment Strategies

The quality and flexibility of the class of strategies, Π, has a large impact on the performance. In this section we discuss features that will be used in constructing treatment strategies. To produce a quality class of treatment strategies, these features must be both expressive and computationally efficient. We construct features, ψ, to describe the state of the network. At a minimum, each location in the network has two binary values: infection status $\boldsymbol{Y}_\ell$ and treatment status $\boldsymbol{A}_\ell$. For a network of size N, the dimension of the infection status and treatment status combined is 4^L. This scales tremendously fast with L and is the simplest case (i.e., no additional covariate information).

We propose a structure for ψ that scales well with network size and allows for fast updates with changes to $\boldsymbol{s}$, $\boldsymbol{y}$, or $\boldsymbol{a}$.

Summarizing the state of the network by enumerating all possible configurations is not computationally feasible. The class of strategies, Π, summarizes subnetworks that take the form of a connected path of locations. Aggregating the summaries of individual subnetworks will summarize the entire network. The ability for aggregating subnetwork summaries to describe the entire network increases as the size of the subnetworks increase.

Define $\mathcal{R}_k$ to be the set of all paths in the network of length k, such that

$$\mathcal{R}_k = \Big\{\{r_1, \ldots, r_k\} \ : \ r_i \in \mathcal{L};\ \Omega_{r_i, r_{i+1}} = 1 \text{ for } i = 1, \ldots, k-1;\ \ r_i \neq r_j \ \forall i \neq j\Big\}. \qquad (2)$$

To summarize the state of the network, we aggregate summaries of connected paths of locations. For a path $r \in \mathcal{R}_k$, define $m_r(\boldsymbol{b}) = 1 + \sum_{i=1}^{k}\sum_{j=1}^{q} \boldsymbol{b}_{i,j} 2^{qi-j}$ where $\boldsymbol{b} \in \{0,1\}^{k\times q}$ to be a unique index identifying the value of $\boldsymbol{b}$. An example of $\boldsymbol{b}$ in the case of an epidemic could be $\{\boldsymbol{s}^{\mathbb{1}}, \boldsymbol{y}, \boldsymbol{a}\}$ where $\boldsymbol{s}^{\mathbb{1}}_{i,j} = \mathbb{1}\{\boldsymbol{s}_{i,j} > 0\}$. Furthermore, $m_r(\boldsymbol{b})$ can be thought of as a decimal representation of the string of bits if $\boldsymbol{b}$ were to be flattened. Feature vector $\psi(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})$ is divided into sections $\psi(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; K) = \{\psi_1(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}), \ldots, \psi_K(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})\}$. Each component $\psi_k(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})$ for $k = 1, \ldots, K$ summarizes paths of length k. Define $\psi_{k,m}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \sum_{r\in\mathcal{R}_k} \mathbb{1}\left[m_r(\{\boldsymbol{s}^{\mathbb{1}}, \boldsymbol{y}, \boldsymbol{a}\}) = m\right]$ to be the m^{th} element of $\psi_k(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})$. Constructing features in this way results in an integer vector where each element is the number of subnetworks that has a particular sequence of covariate, infection status, and treatment status combinations for the locations in the path.

To demonstrate how the features are constructed, consider the network shown in Figure **??**. To reduce the dimensionality for the sake of this illustrative example, assume $\boldsymbol{s} = \varnothing$. Let $\boldsymbol{y} = \{0, 1, 1, 1\}$ and $\boldsymbol{a} = \{0, 1, 0, 0\}$. The set $\mathcal{R}_1$ will only have four elements, each a set with a single location.

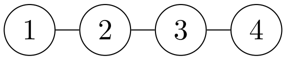

Figure 2: Example network structure to illustrate construction of features.

The first set, $\{1\}$, has index $m_{\{1\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 1$; the second set, $\{2\}$, has index $m_{\{2\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 4$; the third set, $\{3\}$, has index $m_{\{3\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 3$; the fourth set, $\{4\}$, has index $m_{\{4\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 3$. There are three paths of length 2, thus $\mathcal{R}_2$ has 3 elements, each a set with two locations. The first set, $\{1, 2\}$, has index $m_{\{1,2\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 13$; the second set, $\{2, 3\}$, has index $m_{\{2,3\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 12$; the third set, $\{3, 4\}$, has index $m_{\{3,4\}}(\{\boldsymbol{y}, \boldsymbol{a}\}) = 11$. Constructing feature vector for $K = 2$ results in the value

$$\psi(\boldsymbol{y}, \boldsymbol{a}; 2) = \{\underbrace{1, 0, 2, 1}_{\psi_1(\boldsymbol{y},\boldsymbol{a})}, \underbrace{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0}_{\psi_2(\boldsymbol{y},\boldsymbol{a})}\}. \tag{3}$$

5 Estimating the Optimal Strategy via Model-Based Policy Search

One way to estimate the optimal strategy is to estimate a working model of the system dynamics and use policy search to estimate the optimal strategy by maximizing a Monte Carlo approximation of the value function. Policy search is a technique for estimating class-optimal strategies by directly searching in the space of strategies for the one that maximizes the objective function. This approach has the desirable advantage of directly maximizing the value function. In this section, we describe a model-based approach using policy search for estimating the optimal strategy.

To construct a class of strategies for model-based policy search, we utilize the features described in section **??**. Each strategy, $\pi \in \Pi$, is indexed by a set of parameters $\boldsymbol{\theta} \in \Theta$, $\Pi = \{\pi(\,\cdot\,, \cdot\,; \theta) : \theta \in \Theta\}$. A strategy has the form $\pi(\boldsymbol{s}, \boldsymbol{y}; \theta) = \arg\max_{\boldsymbol{a}} \psi(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})^{\intercal}\theta$. Using policy search, the goal is to estimate the $\theta \in \Theta$ that will maximize the value function.

Obtaining an analytical form of the value function is computationally infeasible for more than a few steps. Thus we estimate the value function by Monte Carlo approximation using a finite horizon.

Define $V_T^\pi(\boldsymbol{s}, \boldsymbol{y}) = \mathbb{E}^\pi\left[\sum_{t=1}^T \gamma^{t-1} u(\boldsymbol{Y}_t) \middle| \boldsymbol{S}_0 = \boldsymbol{s}, \boldsymbol{Y}_0 = \boldsymbol{y}\right]$ to be a surrogate for the value function. The value function can be approximated by V_T^π; $V^\pi = \lim_{\widetilde{T}\to\infty} V_{\widetilde{T}}^\pi \approx V_T^\pi$ when T is large. To construct $\widehat{V}_T^\pi$, postulate a model for the system dynamics and estimate via Monte Carlo sampling. In the case of epidemic management, only one trajectory of the disease is observed. Thus for model identifiability, we assume that f and g are Markov, homogeneous in time, and indexed by low dimensional parameters β and η: $f_{t+1}(\,\cdot\,|\boldsymbol{h}_t, \boldsymbol{a}_t) = f(\,\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t; \beta)$ and $g_{t+1}(\,\cdot\,|\boldsymbol{h}_t, \boldsymbol{a}_t) = g(\,\cdot\,|\boldsymbol{s}_t, \boldsymbol{a}_t; \eta)$. Under these assumptions, the surrogate value function $V_T^\pi(\boldsymbol{s}, \boldsymbol{y}; \beta, \eta)$ has the form

$$\int \left[\sum_{t=1}^{T} \gamma^{t-1} u(\boldsymbol{y}_t)\right] \prod_{t=0}^{T-1} f(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \pi(\boldsymbol{s}_t); \beta) g(\boldsymbol{y}_{t+1}|\boldsymbol{s}_t, \pi(\boldsymbol{s}_t); \eta) P(\pi(\boldsymbol{s}_t) = \boldsymbol{a}_t) d\lambda(\overline{\boldsymbol{s}}_T, \overline{\boldsymbol{a}}_T, \overline{\boldsymbol{y}}_T) \qquad (4)$$

where λ is a dominating measure.

We use Thompson Sampling to balance exploration with exploitation when estimating the optimal strategy. Estimate the posterior distribution for β and η using observed data. Draw $\tilde{\beta}$ and $\tilde{\eta}$ from the estimated respective posterior distributions. Estimate the optimal strategy under the belief that $\tilde{\beta}$ and $\tilde{\eta}$ index the true underlying model. The policy search estimator of the optimal strategy is $\widehat{\pi}^*(\boldsymbol{s}, \boldsymbol{y}) = \underset{\pi\in\Pi}{\arg\max}\, V_T^\pi(\boldsymbol{s}, \boldsymbol{y}; \widetilde{\beta}, \widetilde{\eta})$. Stochasticity in $\widehat{\pi}^*$ is induced by the sampling of $\tilde{\beta}$ and $\tilde{\eta}$.

Model-based policy search is advantageous in that it directly maximizes, the value function. When f and g are correctly specified, this method will perform very well. However, when they are incorrectly specified, performance is not guaranteed. In the next section we propose a model-free estimator which estimates the optimal strategy without postulating a model for the system dynamics.

6 Estimating the Optimal Strategy via the Q-function

Utilizing a model-based approach can lead to a high-quality solution when the model is correctly specified. However, in practice, the model is likely incorrectly specified. We propose a method that does not require a working model for the system dynamics and is based on the Q-function. The Q-function for strategy π takes as input a state $\boldsymbol{s}$, infection status $\boldsymbol{y}$, and a treatment $\boldsymbol{a}$ and returns the expected total discounted outcome starting from state $\boldsymbol{s}$ and infection status $\boldsymbol{y}$, choosing treatment $\boldsymbol{a}$ initially, and following strategy π thereafter. The Q-function for strategy π is defined as

$$Q^{\pi}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \mathbb{E}^{\pi}\left[\sum_{v \geq t} \gamma^{v-t} u(\boldsymbol{Y}_v) \middle| \boldsymbol{S}_t = \boldsymbol{s}, \boldsymbol{Y}_t = \boldsymbol{y}, \boldsymbol{A}_t = \boldsymbol{a}\right] \tag{5}$$

where $\mathbb{E}^{\pi}$ is the expectation if treatments are assigned according to π and $\gamma \in [0, 1)$ is the discount factor. If the Q-function is known for the optimal strategy, π^*, then selecting treatments by maximizing Q^{π^*} is an optimal strategy. This section will focus on estimating Q^{π^*} which will be abbreviated Q^*.

6.1 Working Model of the Q-function

We assume the Q-function to have a linear form

$$Q^*(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) \triangleq Q^*(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \theta) = \tau(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})^{\intercal}\theta \tag{6}$$

where $\tau(\boldsymbol{s}, \boldsymbol{a})$ is constructed using finite horizon approximations to the Q-function. Let $\{M_j\}_{j=1}^{J}$ be a set of candidate models for the system dynamics. Define

$$Q^*_{i,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \mathbb{E}_{M_j}\left[u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}'} Q^*_{i-1,j}(\boldsymbol{S}_{t+1}, \boldsymbol{Y}_{t+1}, \boldsymbol{a}') \middle| \boldsymbol{S}_t = \boldsymbol{s}, \boldsymbol{Y}_t = \boldsymbol{y}, \boldsymbol{A}_t = \boldsymbol{a}\right]$$

for $i = 1, 2, \ldots$ and $Q^*_{0,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = 0$ for all j. For an i, j pair, $Q^*_{i,j}$ is an i-step approximation assuming the dynamics are governed by model M_j. Multiple models for the system dynamics are included to increase robustness to model misspecification. For each i, j pair, assume a linear form for the finite horizon Q-function, $Q^*_{i,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \psi(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})^\intercal \xi_{i,j}$, where ψ is defined as in section **??**. Given I and J, feature vector τ is a $1 + IJ$ length vector defined as $\{1\} \cup \{Q^*_{i,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) \ : \ i = 1, \ldots, I \text{ and } j = 1, \ldots J\}$.

6.2 Estimating the Q-function

The Q-function is indexed by the parameters θ and $\xi_{i,j}$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$. Estimating the Q-function is divided into two stages. First, construct the features τ by estimating the finite horizon Q-functions using Thompson Sampling. Then, estimate the Q-function by minimizing the squared Bellman Residual.

To estimate τ we simulate data from each model in $\{M_j\}_{j=1}^J$ using Thompson Sampling. Let β_j, η_j denote the parameters for model M_j. Then draw $\widetilde{\beta}_j, \widetilde{\eta}_j$ from the estimated posterior distributions of β_j, η_j. Simulate data from model M_j with parameters $\widetilde{\beta}_j, \widetilde{\eta}_j$. For $i = 1, \ldots, I$ and $j = 1, \ldots, J$, construct $Q^*_{i,j}$ using ridge regression to estimate $\widehat{\xi}_{i,j}$ by substituting the previous $\widehat{Q}^*_{i-1,j}$ as a plug-in estimate. Combine these models to estimate the features $\widehat{\tau}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \{1\} \cup \{\widehat{Q}^*_{i,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) \ : \ i = 1, \ldots, I \text{ and } j = 1, \ldots, J\}$ where $\widehat{Q}^*_{i,j}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \psi(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a})^\intercal \widehat{\xi}_{i,j}$.

Under the Markov assumption, the Q-function has a convenient recursive representation $Q^*(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}) = \mathbb{E}\left[u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}'} Q^*(\boldsymbol{S}_{t+1}, \boldsymbol{Y}_{t+1}, \boldsymbol{a}') | \boldsymbol{S}_t = \boldsymbol{s}, \boldsymbol{Y}_t = \boldsymbol{y}, \boldsymbol{A}_t = \boldsymbol{a}\right]$. This can be rearranged to obtain the Bellman residual (**?**)

$$\mathbb{E}\left[u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}'} Q^*(\boldsymbol{S}_{t+1}, \boldsymbol{Y}_{t+1}, \boldsymbol{a}') - Q^*(\boldsymbol{S}_t, \boldsymbol{Y}_t, \boldsymbol{A}_t)\right] = 0. \tag{7}$$

We use the Bellman residual as our estimation criterion. Using observed data $\boldsymbol{H}_T$, estimate the

Q-function for the optimal strategy, π^*, by minimizing the squared mean temporal difference error

$$\widehat{\theta} = \arg\min_{\theta} \left[\frac{1}{T} \sum_{t \geq 0}^{T-1} u(\boldsymbol{Y}_t) + \gamma \max_{\boldsymbol{a}} \widehat{\tau}(\boldsymbol{S}_{t+1}, \boldsymbol{Y}_{t+1}, \boldsymbol{a})^{\intercal}\theta - \widehat{\tau}(\boldsymbol{S}_t, \boldsymbol{Y}_t, \boldsymbol{A}_t)^{\intercal}\theta \right]^2 . \tag{8}$$

Finally, the estimator of the optimal treatment strategy is $\widehat{\pi}^*(\boldsymbol{s}, \boldsymbol{y}) = \arg\max_{\boldsymbol{a}} \widehat{Q}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \widehat{\theta})$.

7 System Dynamics Models

The dynamical system governing the evolution of the epidemic is a susceptible-infected-susceptible (SIS) model (**?**). The epidemic evolves over the network by infecting uninfected 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. Recall the distribution functions $f(\,\cdot\,|\boldsymbol{S}_t, \boldsymbol{A}_t; \beta)$ and $g(\,\cdot\,|\boldsymbol{S}_t, \boldsymbol{A}_t; \eta)$ from Section **??** which are the conditional distributions of $\boldsymbol{S}_{t+1}$ and $\boldsymbol{Y}_{t+1}$ respectively. Define $\mathcal{I}_t = \{\ell \in \mathcal{L} \;:\; Y_t = 1\}$ to be the set of infected locations at time t and $\mathcal{I}_t^c$ to be the compliment set. Conditional on time t, each element of $\boldsymbol{S}_{t+1}$ follows an auto-regressive model and each element of $\boldsymbol{Y}_{t+1}$ is Bernoulli distributed. In our simulations, we assume that there is only one covariate for each location. The transition models have the form

$$\begin{aligned} f(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \beta) = &\quad \textstyle\prod_{\ell=1}^{L} \phi\left(\frac{\boldsymbol{s}_{t+1,\ell} - \beta_0 \boldsymbol{s}_{t,\ell}}{\beta_1}\right) \\ g(\boldsymbol{y}_{t+1}|\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \eta) = &\quad \textstyle\prod_{\ell \in \mathcal{I}_t} q_\ell(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \eta)^{1-\boldsymbol{y}_{t+1,\ell}} [1 - q_\ell(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \eta)]^{\boldsymbol{y}_{t+1,\ell}} \\ &\quad \textstyle\prod_{\ell \in \mathcal{I}_t^c} p_\ell(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \eta)^{\boldsymbol{y}_{t+1,\ell}} [1 - p_\ell(\boldsymbol{s}_t, \boldsymbol{y}_t, \boldsymbol{a}_t; \eta)]^{1-\boldsymbol{y}_{t+1,\ell}} \end{aligned}$$

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

$$p_\ell(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta) = 1 - [1 - p_{\ell,0}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta)] \prod_{\ell' \in \mathcal{I}_t \cap \mathcal{N}_\ell} [1 - p_{\ell,\ell'}(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta)].$$

We define multiple models that alter the impact of treatment on the transition probabilities. Each model changes the forms of $p_{\ell,0}$, $p_{\ell,\ell'}$, and q_ℓ. These are discussed in the following subsections.

7.1 Model 1: No Resistance to Treatment

Each uninfected location receives a treatment that will reduce the probability of infection. If an infected location receives treatment, then it reduces its propensity to transmit to adjacent uninfected locations as well as increasing its own probability of recovery. The following defines this model

$$\text{logit}[p_{\ell,0}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_0 + \eta_1 a_\ell$$

$$\text{logit}[p_{\ell,\ell'}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_2 + \eta_3 a_\ell + \eta_4 a_{\ell'}$$

$$\text{logit}[q_\ell(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_5 + \eta_6 a_\ell.$$

7.2 Model 2: Resistance to Treatment Dictated by Covariates

In Model 1, covariates have no impact on the spread of the disease and are thus irrelevant for estimating an optimal treatment strategy. This model includes a covariate effect by adding resistance to treatment if a covariate is positive. If a location has a positive covariate, then it will not respond to treatment. The following defines this model.

$$\text{logit}[p_{\ell,0}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_0 + \eta_1 a_\ell \mathbb{1}_{\{s_\ell \leq 0\}}$$

$$\text{logit}[p_{\ell,\ell'}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_2 + \eta_3 a_\ell \mathbb{1}_{\{s_\ell \leq 0\}} + \eta_4 a_{\ell'} \mathbb{1}_{\{s_{\ell'} \leq 0\}}$$

$$\text{logit}[q_\ell(\boldsymbol{s},\boldsymbol{y},\boldsymbol{a};\eta)] = \eta_5 + \eta_6 a_\ell \mathbb{1}_{\{s_{\ell'} \leq 0\}}$$

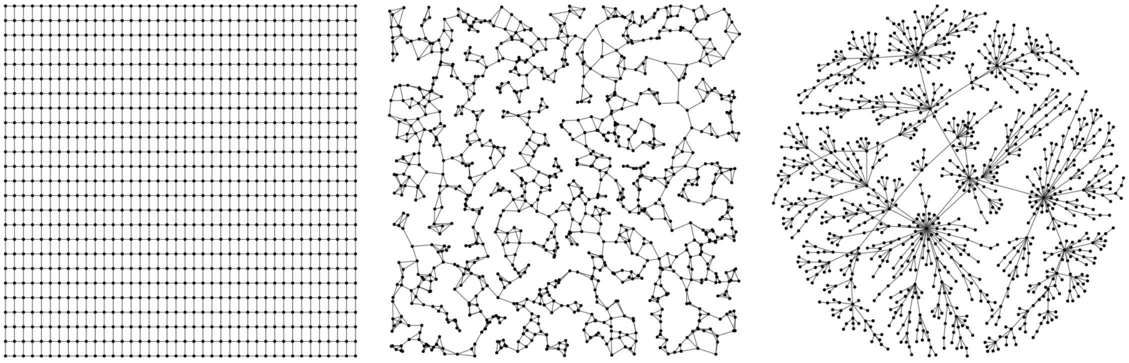

Figure 3: Figure displaying examples of the three network structures. **Left**: lattice network with 1000 locations. **Center**: random nearest neighbor network with 1000 locations. **Right**: scalefree network with 1000 nodes.

8 Simulation Experiment

8.1 Network Structures

Network structure can provide important information for constructing an optimal treatment strategy. Patterns of the locations in the network may provide information for how important a location is to treat. To demonstrate the robustness of the proposed method, simulations include three network structures: lattice structure, Barabasi structure (**?**), and random nearest neighbor structure. Examples of these networks are shown in figure **??**.

8.2 Experiment Setup

We present simulation results using multiple network structures. The generative models are tuned to have specified infection rates. To define these rates, assume a set of locations $\mathcal{L} = \{1, 2, 3, 4\}$ where location 1 is a neighbor of all other locations. Each equation will be in terms of location 1.

The rates are defined by

$$
\begin{aligned}
p_{1,0}(\{\cdots\}, \{0, \cdots\}, \mathbf{0}) =& \quad 0.01 \\
p_1(\{\cdots\}, \{0,1,1,1\}, \mathbf{0}) =& \quad 0.5 \\
p_1(\{+, \cdots\}, \{0,1,1,1\}, \{1,0,0,0\}) =& \quad 0.5 * 0.75 \\
p_1(\{\cdot, +, +, +\}, \{0,1,1,1\}, \{0,1,1,1\}) =& \quad 0.5 * 0.25 \\
q_1(\{\cdots\}, \{1, \cdots\}, \{0, \cdots\}) =& \quad 0.25 \\
q_1(\{+, \cdots\}, \{1, \cdots\}, \{1, \cdots\}) =& \quad 0.25 * 0.5
\end{aligned}
$$

where $\cdot$ represents any value and $+$ represents a positive value. These equations set the latent probability of infection with no treatment effect to 0.01; the probability of infection when 3 neighbors are infected with no treatment effect to 0.5; 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; 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; the base probability of recovery with no treatment to 0.25; the probability of not recovering to reduce by a factor of 0.5 with a treatment effect. Lastly, this experiment has only one covariate, dom $\boldsymbol{S}_{t,\ell} = \mathbb{R}$, and it follows an AR(1) process, $\boldsymbol{S}_{t,\ell} = \beta_0 \boldsymbol{S}_{t-1,\ell} + \epsilon_{t,\ell}$, where $\epsilon_{t,\ell} \overset{iid}{\sim} \mathcal{N}(0, \beta_1)$. We set $\beta_0 = 0.9$ and $\beta_1 = 1.0$. The true generative model is a linear mixture of model 1 and model 2 defined by the mixture parameter $\delta \in [0,1]$. When $\delta = 1$, the model is equivalent to model 1 and when $\delta = 0$, the model is equivalent to model 2.

For comparison, we implement four competing methods: no treatment, do not apply treatment to any locations; random, select locations with uniform probability to receive treatment; proximal, treat locations that are closest to locations of the opposite infection status; myopic, treat locations with the highest probability of being infected at the next time point under no treatment. All methods use $\lfloor L0.05 \rfloor$ treatments except for no treatment which uses zero treatments. In order to make the competing methods well defined, each method splits the treatments equally between infected and uninfected locations. If there are not enough of one type to administer all treatments, the remaining are given to the other infection type.

8.3 Experiment Results

There are 50 replications of each treatment strategy for all network and model combinations. Each simulation replication runs for 25 time points and we show the estimated mean proportion of locations infected. From the simulations results in figure **??** we can see that when the system dynamics is correctly specified, using policy search is the best performing method. However, as δ approaches 1, estimating the Q-function is increasingly favorable compared to all other methods.

9 Management of the Ebola Virus

Demonstrating our method on the Ebola virus required some changes to our setup. First, the dynamics model for Ebola was taken from **?**. With this model, called the gravity model, there is no recovery from infection. Thus $\mathrm{logit}[q_\ell(\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{a}; \eta)] = 0.0$, and the probability of infection is defined by

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

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 add stability for estimating model parameters, we force the coefficient on the gravity term and the exponent on the population product to both be positive.

To set the parameter value for the generative model, we estimate the maximum likelihood

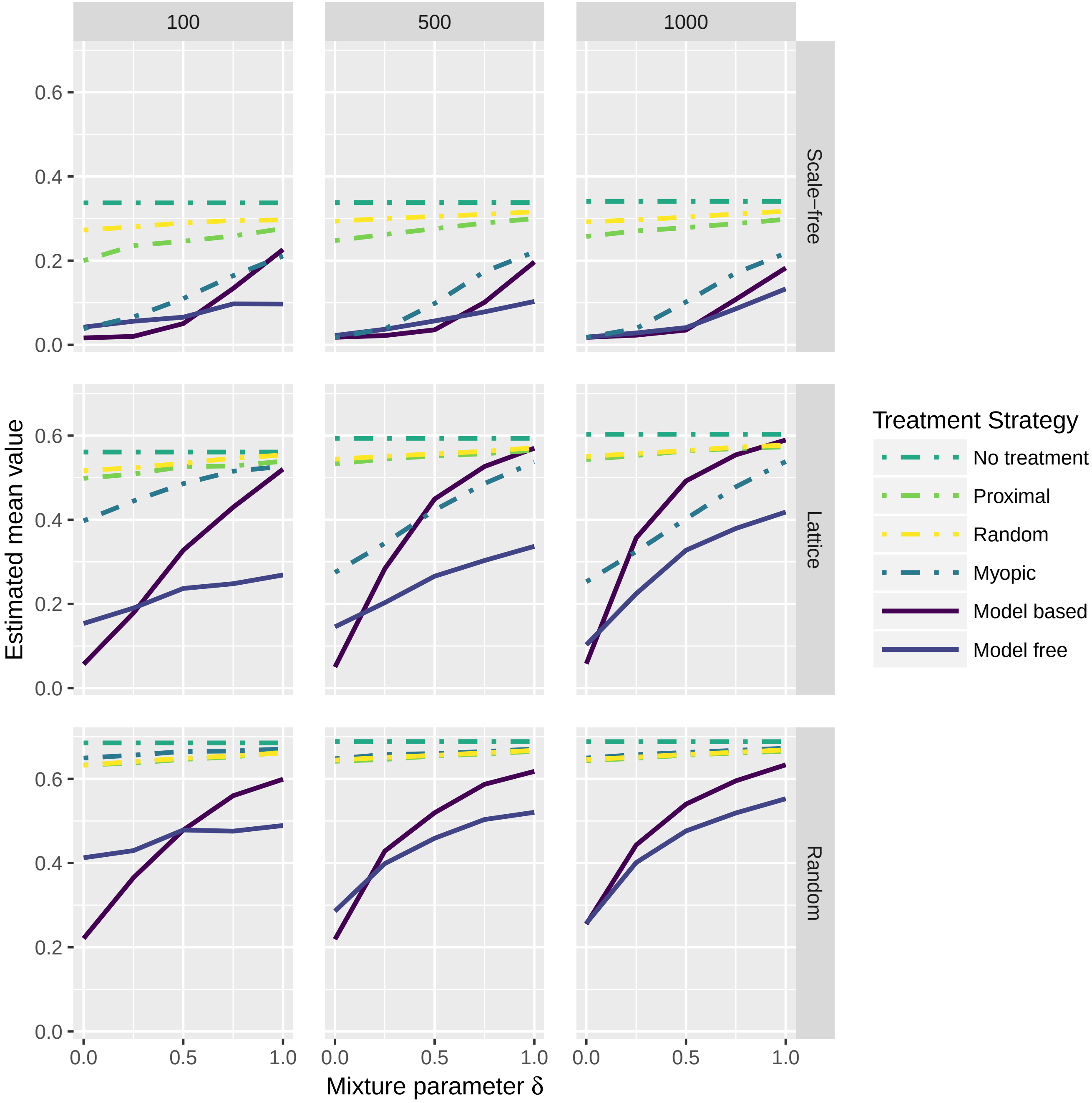
100
500
1000
Scale−free
Lattice
Random
Estimated mean value
0.0
0.2
0.4
0.6
Mixture parameter δ
0.0
0.5
1.0
Treatment Strategy
No treatment
Proximal
Random
Myopic
Model based
Model free

Figure 4

estimator $\widehat{\eta}^{MLE}$ from the observed data. From the MLE, the generative model is further tuned to have two properties. One: under no treatment, the proportion of locations infected after 25 time points 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. In both settings, the starting infection state is the first 25% of the observed infections. 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.

There is no defined network structure for the Ebola virus. To utilize the features we constructed in section **??**, we need a working network structure. The network structure for Ebola is defined as $\Omega_{\ell,\ell'}$ if regions ℓ and ℓ' share a common border.

We present results for management of the Ebola virus in table **??**. Each simulation replication started from the first 25% of observed infections and simulated 25 time points. There are 50 replications for each treatment strategy. Displayed is the estimated mean proportion of infected locations and the standard error. All methods are given a maximum of $\lfloor 290 * 0.05 \rfloor = 14$ treatments except for no treatment which uses zero treatments. All competing treatment strategies are the same as the toy simulations, except myopic has one change. For myopic, uninfected locations are still prioritized by their probability of being infected at the next time point, but infected locations are not, because there is no recovery in the Ebola simulations. Thus, infected locations are ranked by the average probability of all uninfected locations becoming infected which each probability divided by the distance between the two locations. For infected location ℓ, the priority score is proportional to $\sum_{\ell'}(1 - y_{\ell'})\frac{p'_\ell(\boldsymbol{s},\boldsymbol{y},\boldsymbol{0})}{d_{\ell,\ell'}}$. From these results, we can see that the model based policy search method out performs all other competing strategies and provides a $\sim$ 19% improvement over

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.

no treatment.

10 Conclusion