Capstone Project Proposal:
<Policy optimization in Epidemic>

<Mai Le Xuan Anh>

Project Summary

Abstract

In a stochastic partially-observable environment, an agent faces the dilemma of exploration and exploitation to make a decision in order to optimize a particular reward function. Policy optimization maps a certain state of environment to a decision. Various attempts have been made in devising such policy using well-known techniques such as Partially Observable Markov Decision Process(POMDP) or Monte Carlo Tree Search (MCTS). We evaluate these techniques in the context of the Epidemic scenario. This paper proposes an algorithm to approximate an optimal policy in the Epidemic scenario.

Introduction, Specific Aims, And Background

Pandemics have been a threat to public health and human society. The ability to control pandemics to reduce their economic and social consequences is a challenge due to the difficulties in simulate such scenario in real life situation. Therefore, many models of pandemics have been proposed to resemble the actual scenario with a set of interventions to control the pandemics. These models allow us to devise algorithm to approximate an optimal policy in minimizing the consequences of the epidemics.

The Epidemic scenario is mentioned in [1] using the model of Coevolving Graphical Discrete Dynamical System (CGDDS). We rewrite the definition of CGDDS for sake of record.

Such Epidemic scenario is represented by symbol S over a given domain D of state values and a given domain L of label values, is a pair (G, F), whose components are as follows:

- Graph G(V, E): Let the vertex set $V = \{v_1, v_2, ...v_n\}$ represent the set of n $\geq$ 1 agents (individuals). For each vertex v_i, let vector s_i denote its states $s_i = (s_i^1, s_i^2, ...s_i^k) \in D = (D_1 \times D_2 \times \ldots \times D_k)$, where k is the number of states of vertex v_i. Intuitively, the states comprise the agents health state, behavioral state (e.g., level of fear, risk aversion, etc.), and static demographic attributes. Let the edge set $E = \{e_1, e_2, \ldots, e_m\} \subset (V \times V)$ represent the contacts between agents. For any edge $e \in E$, let vector l_e denote its labels $l_e = (l_e^1, l_e^2, \ldots, l_e^h) \in L = (L_1 \times L_2 \times \ldots \times L_h)$, where h denotes the number of labels. In our social contact network, the edge labels include the contact duration and the contact type (home, school, work, shopping, or others).

- Functions $F = (f, g^V, g^E)$, where f is a set of local transition functions; g^V is a set of vertex modification functions; and g^E is a set of edge modification functions. For each vertex v_i, let $f_i : D \times D^{V_i} \times L^{E_i} \rightarrow D$ be its local state transition function, where V_i and E_i are the neighboring vertices and edges of v_i. Normally, V_i are vertices adjacent to v_i and E_i are edges incident on v_i. The f_i function corresponds to the propagation process that changes the states of an agent based on

(i) the current states of the agent

(ii) the states of all its neighboring agents

(iii) the current labels on the contact edges with its neighboring agents

These variables determine a distribution over D; then a state is chosen from the distribution as the output of f_i. So f_i is a random function. Let $g^V = \{g_1^V, g_2^V, \dots g_{k_V}^V\}$ be a set of k_V vertex modification functions, where each $g_j^V : D^V \times L^E \to D^V$ directly changes states of vertices based on the current state of the whole graph. We assume V is constant. Let $g^E = \{g_1^E, g_2^E, \dots g_{k_E}^E\}$ be a set of k_E edge modification functions, where each $g_j^E : D^V \times L^E \to L^{V \times V}$ changes the set of edges and the edge labels based on the current state of the whole graph.

In [1], D is defined using SEIR model which includes only 4 possible states of an agent: susceptible, exposed, infectious, and removed state. For g^V, this corresponds to the set of pharmaceutical interventions (**PIs**, e.g., antiviral, vaccination). For g^E, this corresponds to the set of nonpharmaceutical interventions (**NPIs**, e.g., school closure, quarantine and social distancing) that change the graph structure.

The reward function $R(t)$ at each time step t is the number of agents (nodes) with removed state. An optimal policy would try to minimize this function in a maximum number of time steps. For example, given a simulation of Epidemics in 200 days, $R(200)$ should be minimized.

Goals And Potential Impact

Given a CGDDS with n agents, then every time step, the CGDDS simulation system has to update in $O(n)$ nodes and $O(n^2)$ edges.

Our goal is to devise and assess the performance of our algorithm in a smaller scale of CGDDS described in [1]. For testing convenience, we limit ourselves in 1000 agents. Even though this may not reflect the actual scale of a pandemic, we believe that once we figure such algorithm, we can generalize and increase the scale of our graph model using High Performance Computing (HPC).

We expect to modified the techniques in POMDP and MCTS to improve our algorithm. The impact of the research could be significant in controlling future pandemics as well as other "infectious" phenomenons such as wild fire.

Methodology

Our method will evaluate the performance of existing framework in approximating optimal policy for Epidemics scenario. The two prominent methods that we are going to look at are POMDP and MCTS which are carefully described in [2] and [3] respectively.

POMDP is the generalization of Markov Decision process where the state-transition function is stochastic and we may only observe part of the environment.

Formally, a POMDP consists of:

- $|S|$ states $S = \{1, \dots, |S|\}$ of the world;
- $|A|$ actions (or controls) $U = \{1, \dots |A|\}$ available to the policy;
- $|Y|$ observations $Y = \{1, \dots, |Y|\}$;
- a (possibly stochastic) reward $r(i) \in R$ for each state $i \in S$.

Each action $u \in U$ determines a stochastic matrix $[q(j|i,u)]$ where $i = 1 \dots |S|, j = 1 \dots |S|$, such that $q(j|i,u)$ denotes the probability of making a transition from state $i \in S$ to state $j \in S$ given action $u \in U$. For each state i, an observation $y \in Y$ is generated independently with probability $\nu(y|i)$. The distributions $q(j|i,u)$ and $\nu(y|i)$, along with a description of the rewards, constitutes the model of the POMDP in figure 1.

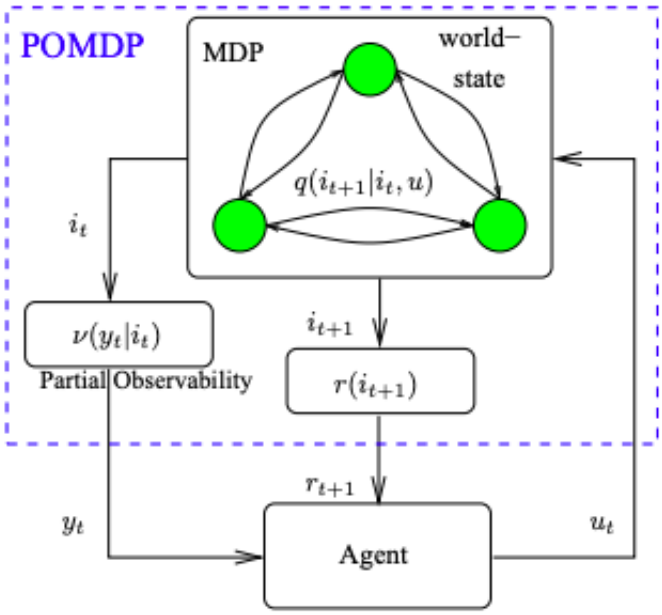

Figure 1: Diagram of the world perspective of POMDP, and the stochastic process $\nu(y_t, i_t)$ mapping the current state i_t to an observation y_t, thus hiding the true state information.

The application of MCTS in large-scale POMDP will be the second stage of the research which is described in [3]. This approach employs techniques that reduce the effect of curse of dimensionality as well as increase the running time of the algorithm.

Budget And Justification

Budget may be required if we need to get permission from the authors of [1] to attain access to their large scale CGDDS system which would allow us to test the performance of algorithm in a large scale environment.

References

References Cited

[1] S. D. X. F. Y. M. Keith Bisset, Jiangzhuo Chen and M. Marathe, "Indemics: An interactive high-performance computing framework for data-intensive epidemic modeling," in *ACM Transactions on Modeling and Computer Simulation, Vol. 24, No. 1, Article 4.* ACM, 2014.

[2] D. Aberdeen, "A (revised) survey of approximate methods for solving partially observable markov decision," 2003.

[3] D. Silver and J. Veness, "Monte-carlo planning in large pomdps," in *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010.