UCLA Samueli School of Engineering

Stage-Wise School Reopening using Reinforcement Learning

ECE 238 Final Project Presentation
Mohan Srinidhi Acharya, Terri Tsai

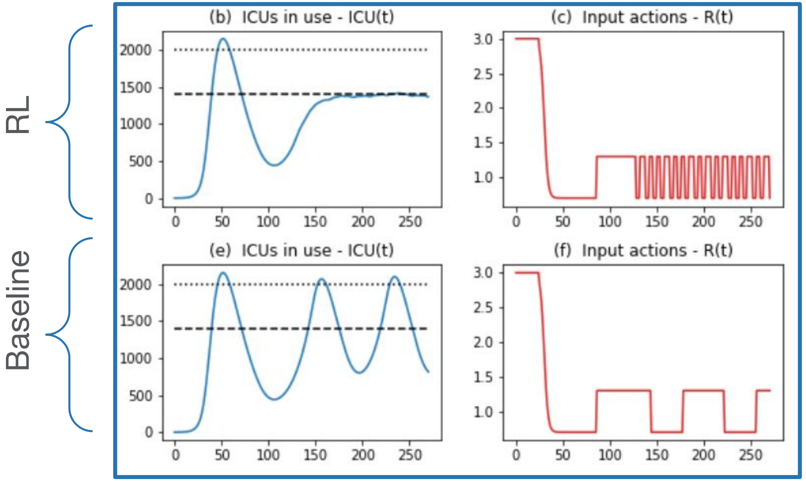
Introduction

- COVID-19
  - is unprecedented
  - has delayed transmission time, lag in onset symptoms, hard to predict
- Lockdown
  - Pros: keep community healthier
  - Cons: economic, mental health, productivity decline…

How to best regulate school reopening at different degrees?

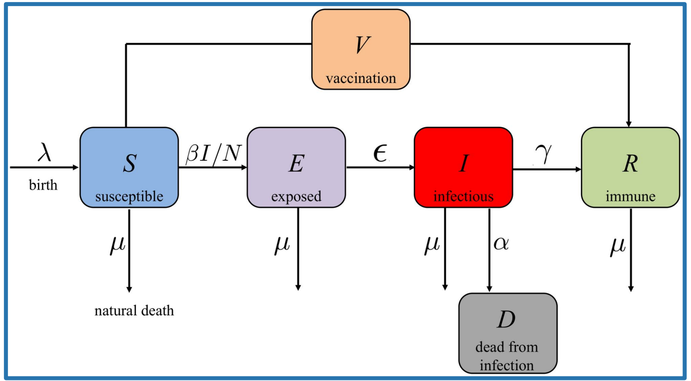
- Use reinforcement learning
- Reinforcement learning: agent learns best actions to take in an environment that will maximize cumulative reward
- RL is fitting for this problem; we want our current decision to be good for the current time and for the future time.

Related Works



Mauricio Arango and Lyudmil Pelov. Covid-19 pandemic cyclic lockdown optimization using reinforcement learning, CoRR, abs/2009.04647.2020.

- Lockdown/reopen (2 actions) represented by R (effective reproduction number)
- Agent learns optimized cyclic lockdowns in day increments
- Goal: maximize opening while avoid exceeding a threshold



Zhe Xu, Bo Wu, and Ufuk Topcu. Control strategies for covid-19 epidemic with vaccination, shield immunity and quarantine. A metric temporal logic approach. PLOS ONE, 16, 03 2021.

- SEIR that models vaccinations

- Multiple levels of reopening actions (2~5) represented by R (effective reproduction number)
- Agent learns optimized cyclic lockdowns in day, week, biweekly increments
- Goal: maximize opening while avoid exceeding a threshold
- Include vaccination considerations

Problem Formulation

State:
continuous, 12-dimension:
(S, E, I, R, D, n[7], n[6], n[5], n[4], n[3], n[2], n[1]), normalized

Action:
discrete {0, 1, 2};
representing R = {0.7, 1.2, 1.5}

$$\beta = \gamma R$$

Reward:

$$\text{reward}_1(t) = \begin{cases} 0, & \text{if } action = 0 \\ \alpha_1, & \text{if } action = 1 \\ \alpha_2, & \text{if } action = 2 \end{cases}$$

$$\text{reward}_2(t) = \begin{cases} 0, & \text{if } m_t < m_{threshold} \\ \frac{-\alpha_1}{m_{threshold}} m_t, & \text{if } m_t \geq m_{threshold} \end{cases}$$

$$\text{reward}_{total}(t) = \text{reward}_1(t) + \alpha_{linear}\text{reward}_2(t)$$

State transition:

SEIR differential equations

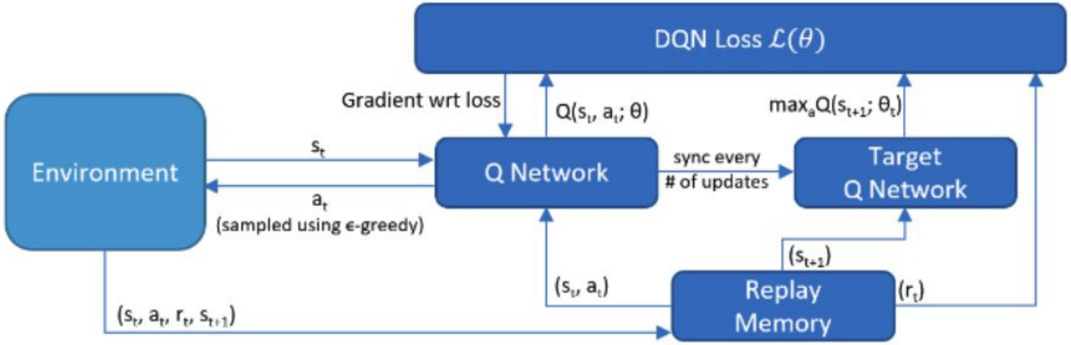
$$\left\{ \begin{aligned} S_{t+1} &\leftarrow S_t - \frac{\beta I_t}{N} S_t - V \\ E_{t+1} &\leftarrow E_t - \epsilon E_t + \frac{\beta I_t}{N} S_t \\ I_{t+1} &\leftarrow I_t + \epsilon E_t - (\gamma + \alpha) I_t \\ R_{t+1} &\leftarrow R_t + \gamma I_t + V \\ D_{t+1} &\leftarrow R_t + \alpha I_t \end{aligned} \right.$$

$$\underbrace{(n^{[7]}_{t+1}, n^{[6]}_{t+1}, n^{[5]}_{t+1}, n^{[4]}_{t+1}, n^{3]}_{t+1}, n^{[2]}_{t+1}, n^{[1]}_{t+1})} \leftarrow (n^{[6]}_t, n^{[5]}_t, n^{[4]}_t, n^{[3]}_t, n^{[2]}_t, n^{[1]}_t, \epsilon E_t)$$

tracking the last 7 days of new cases

# of new cases at current time t

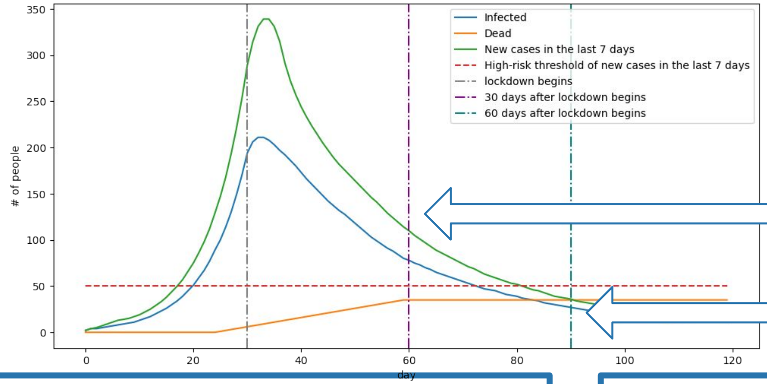
Proposed Solution / Algorithm



- DQN [Off Policy] with Experience Replay and Target Network.
- Neural Network (Function approximator) replaces Q-Table.
- Uses MSE loss function.
- Takes considerable time to converge.

1. **Experience Replay**
   - Memory bank that stores states,actions,rewards created.
   - Sample random experiences.
   - Sequential dependency eliminated.
2. **Target Network**
   - DQN updates entire network each step.
   - Target always moving, hence backprop error not consistent.
   - Target Network copy of Neural Network but only copied every N time steps.

Performance Metric / Initialization



initialization where new cases in the last 7 days is above threshold

below threshold

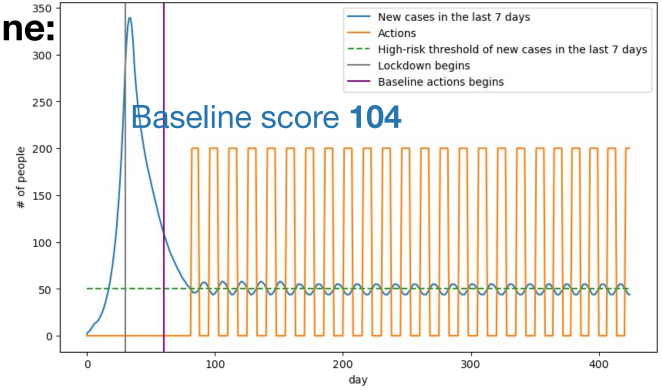
Metric Score Function:

$$\text{score}_1(t) = \begin{cases} 0, & \text{if } action = 0 \\ 1, & \text{if } action = 1 \\ 2, & \text{if } action = 2 \end{cases}$$

$$\text{score}_2(t) = \begin{cases} 0, & \text{if } m_t < m_{threshold} \\ \frac{-m_t}{m_{threshold}}, & \text{if } m_t \geq m_{threshold} \end{cases}$$

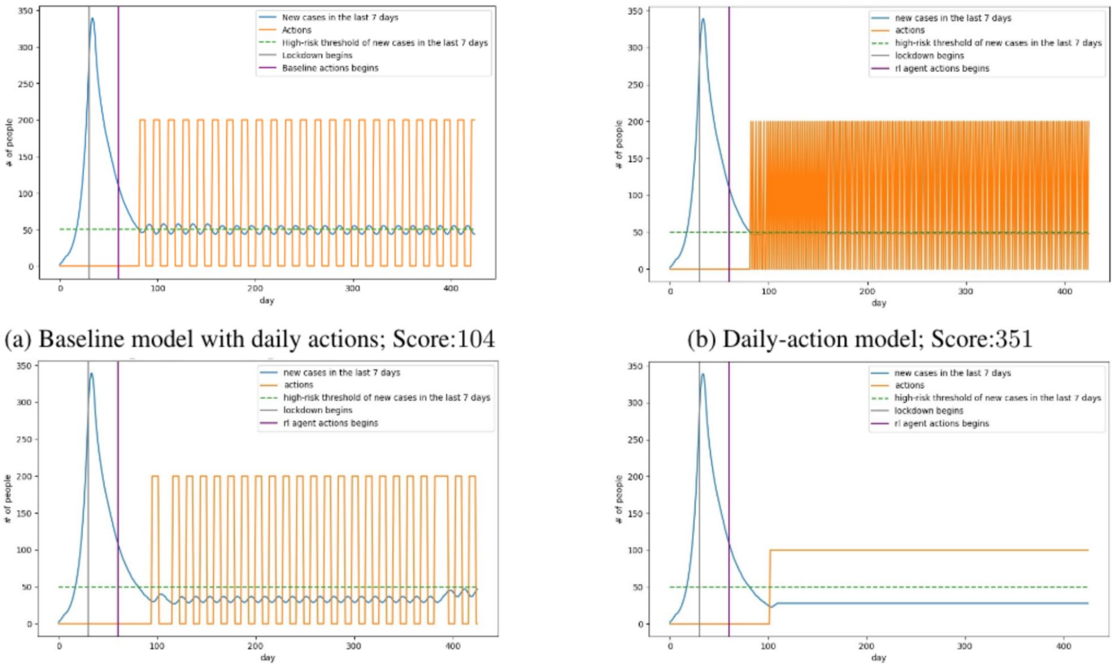
$$\text{score} = round\left(\sum_{t=1}^{365} \text{score}_1(t) + \text{score}_2(t)\right)$$

Baseline:



Baseline score **104**

Results - Action Frequency



(a) Baseline model with daily actions; Score:104

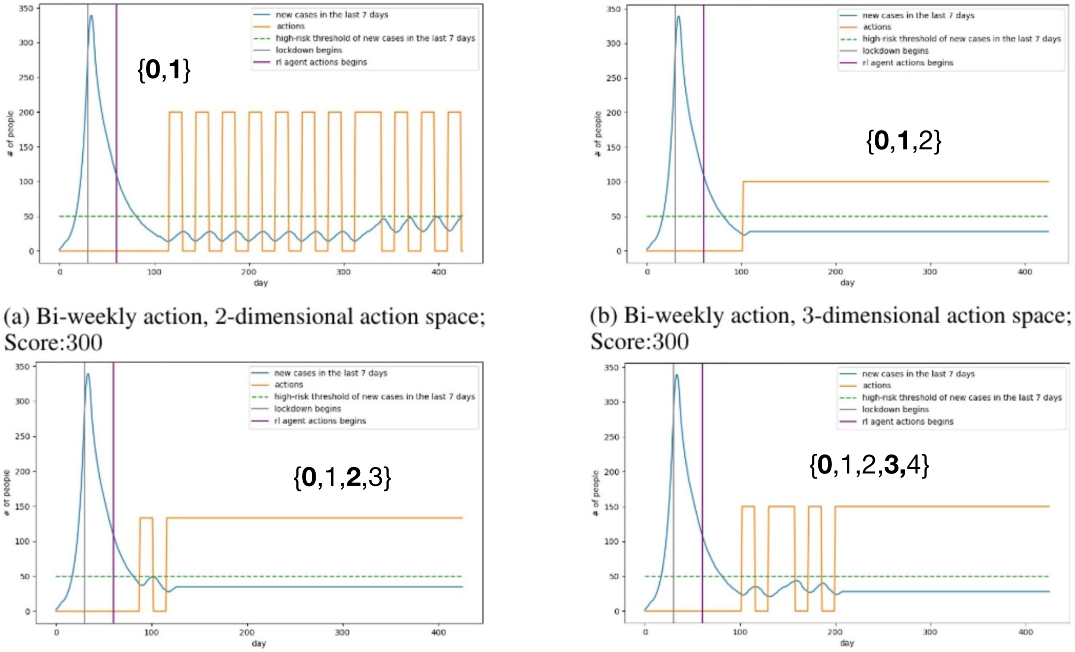
(b) Daily-action model; Score:351

(c) Weekly-action model; Score:303

(d) Bi-weekly-action model; Score:291

- Daily Action model not feasible despite high metric score.
- Weekly Action model more practical but lower metric score due to frequent shutdowns.
- Bi-Weekly Action model most preferable due to practical considerations.
- Conservative policy (Prefers Action 1 over Action 2) in order to keep a check on new cases.

Results - Action Space Size



(a) Bi-weekly action, 2-dimensional action space; Score:300

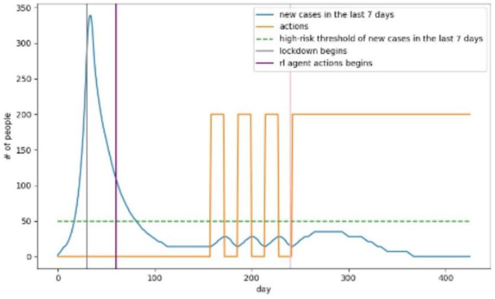
(b) Bi-weekly action, 3-dimensional action space; Score:300

(c) Bi-weekly action, 4-dimensional action space; Score:400

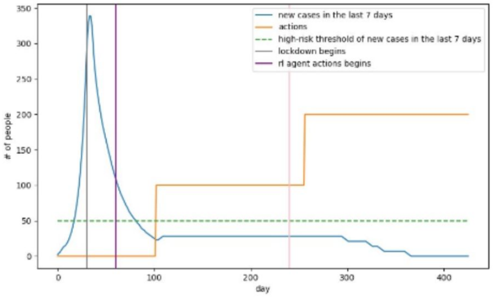
(d) Bi-weekly action, 5-dimensional action space; Score:390

- 2-stage action space allows agent to choose between fully reopen and complete shutdown.
- 3-stage action space allows half-open option that agent chooses most of the time.
- 4-stage and 5-stage action [4-stage picks action 2, 5-stage picks action 3] space allows for higher diversity by allowing varying degrees of openness.
- Greater action space size generally results in a higher metric score.

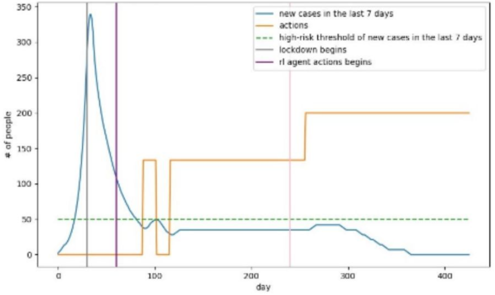
Results - Vaccination Inclusion



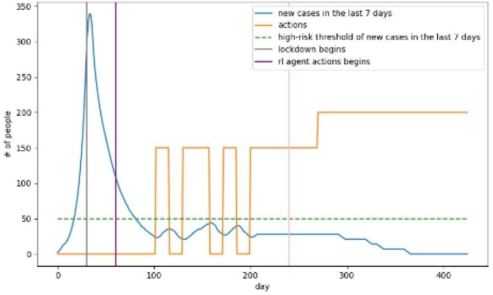
(a) Bi-weekly action, 2-dimensional action space, with vaccination; Score:193



(b) Bi-weekly action, 3-dimensional action space, with vaccination;; Score:461



(c) Bi-weekly action, 4-dimensional action space, with vaccination; Score:785



(d) Bi-weekly action, 5-dimensional action space, with vaccination; Score:969

- Vaccination introduced on day 240. (180 days after simulation starts)
- Effect of vaccination clearly seen as number of new cases decrease.
- For all action space sizes, RL agent picks least restrictive action (Action 2), full reopening as vaccination takes shape.
- Higher Action space sizes result in larger metric score as more freedom allowed.

Conclusion

- Agent with larger action space options or higher action frequency tends to score higher. (Not necessarily the optimal policy)
- Good idea to include a larger action space, even if agent picks only a few select actions.
- Five stages of reopening preferred, no concerns over inconvenience since agent only picks subset of actions. The agent still picks good actions.
- Actions designed to be taken Weekly and Bi-Weekly.
- **Future Work** : Actions taken after arbitrary number of days after looking at past progress/ feasibility of frequency of action change.
- **Future Work** : Move from DQN to Actor-Critic methods to check effect on policy.

Questions ?