Using Reinforcement Learning for Operating Educational Campuses Safely during a Pandemic

Elizabeth Ondula and Bhaskar Krishnamachari
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089
{ondula, bkrishna}@usc.edu

Abstract

The ongoing Covid-19 pandemic has brought significant disruption to educational environments everywhere. Campuses have been forced to shut down completely to help keep students and teachers safe, affecting educational outcomes negatively. We present CampusPandemicPlanR, a reinforcement learning-based tool that could be applied to suggest to campus operators how many students to allow on campus each week. The tool aims to strike a balance between the conflicting goals of keeping students from getting infected, on the one hand, and allowing more students to come into campus to allow them to benefit from in-person classes, on the other.

Introduction

As a 2020 World Bank report titled "The COVID-19 pandemic: Shocks to education and policy responses " states "The [COVID-19] pandemic has already had profound impacts on education by closing schools almost everywhere in the planet, in the largest simultaneous shock to all education systems in our lifetimes" (Bank 2020). A number of schools have been temporarily closed and in many cases, adoption to new learning environments has affected student sentiments (Duong et al. 2020) and posed safety challenges as well as challenges to maintaining engagement and achieving learning outcomes (Khamees et al. 2020; **?**). Developing and implementing smart operational strategies under pandemic uncertainties is key to maximize overall health and safety for students while at the same time also maximizing their learning opportunities through in-person interactions wherever possible. In particular, there is a need for tools that school administrators can use to make tradeoffs between learning objectives and safety of school community members. We approach this operational decision problem by designing and developing CampusPandemicPlanR, a simulation environment for operating a campus under dynamic strategies that can be used to train reinforcement learning agents to autonomously recommend actions to the administrators. CampusPandemicPlanR is designed to assist in planning and scheduling of educational activities in the midst of a pandemic.

We model the problem of finding operational strategies for campus safety as a problem of reinforcement learning. CampusPandemicPlanR incorporates the modeling of a general school (with details of students, teachers, courses, and classrooms) and a COVID-19 transmission model. The actions of the RL agent, trained by and operating on this environment, pertain to the fraction of students from each class permitted to be on campus in the next time period (e.g., a week). These actions can be seen as a recommendation to a human operator, a school administrator, who can choose to accept or override the suggested actions The reward function is defined as a weighted combination of a penalty value for new infections and a reward for allowing in-person attendance opportunities for students. Note that there is a tradeoff between these two as allowing more students on campus during a high risk period may result in more penalties than reward, while allowing too few students during low risk period may result in low reward and low penalty. The simulator is being developed using Python and integrated with the widely-used OpenAI Gym framework to allow other researchers to extend the capability of both the environment and test various RL algorithms. To our knowledge ours is the first effort to develop an open-source simulator/RL training environment for this domain, and we believe that the effort, while currently in early stages, could lead to a useful practical tool for school/college campus administrators.

Review of Prior Work

We classify the work relevant to this paper under the following categories: i) Reinforcement Learning applied to Epidemic/Pandemic Control ii) application of reinforcement learning to educational settings, and ii) impact of pandemic on learning and campus operations.

Applying RL to Epidemic Control

In a very closely related prior work (Yanez, Hayes, and Glavin 2019), the authors have discussed the potential of using reinforcement learning for finding optimal strategies to control an epidemic outbreak, using a boarding school scenario as an example. They consider a broad formulation of the problem to train an agent to make optimal intervention actions such as school closure, vaccination and isolation. Their work focuses on the challenges associated with designing an RL environment for such problems. Our work is closely related, with key differences in the specifics: we take into account student course enrollment and classroom

space for time-tabling in our formulation and the actions in our formulation decide what fraction of students from each class to allow on campus for the next period; we also design a reward function that balances between the benefit of having students attend classes in-person and the risk of students getting infected on campus. Yanez *et al.* note that "a possible solution, that to our knowledge has not been explored in the case of epidemic control, is to share the exact environment definition in a platform such as OpenAI Gym" – such an implementation is precisely the focus of our work.

Other examples of studies that explore the application of RL to epidemic control are the works (Arango and Pelov 2020), (Uddin et al. 2020). The work (Arango and Pelov 2020) develops an OpenAI Gym environment for a general community with the action being to impose or remove a lockdown, with the goal of minimizing overshoot of available ICU beds in the community as well as time spent under lockdown. The work (Uddin et al. 2020) likewise implements an RL environment where the actions are characterized by the tuple (testing, sanitization and lockdown) and the rewards are a combination of disease spread, impact on living quality, economy and resources expended. They compare various RL algorithms and find that deep-learning based algorithms offer the best performance. These two works are not specific to an educational environment.

Applying RL to Education

Several papers in the literature (Bassen et al. 2020), (Ming and Hua 2010), (Yugay, Kyung, and Ko 2009) have discussed the application of RL to education settings for problems other than epidemic/pandemic control. The paper (Bassen et al. 2020) presents a reinforcement scheduling method adopted for scheduling online course activities to increase learning outcomes. The paper (Ming and Hua 2010) looks at course-scheduling and addressing the "curse of dimensionality" problem given the large state space for schools and universities timetables. Option-based hierarchical reinforcement learning is used for the course scheduling problem. The paper (Yugay, Kyung, and Ko 2009) addresses course scheduling in a multi-agent system, with multiple agents representing instructors and a coordinating agent. They apply reinforcement learning and heuristics to enables these agents to negotiate a timetable that takes into account a set of given hard and soft constraints.

Epidemic Control and Education

Controlling the spread of an epidemic such as COVID-19 in an educational setting can be challenging given uncertainties in the effectiveness of different intervention strategies and control measures in different regions. Although various interventions and strategies have been adopted in different regions globally, a simulation study (Di Domenico et al. 2021) finds that reopening after a lockdown may still increase cases . Agent based models developed by various studies for different campus settings (Gressman and Peck 2020), (Lv et al. 2021), (Hamer et al. 2021), (Hekmati et al. 2021) indicate that measures such as large-scale randomized testing, contract tracing, face-mask use and quarantining are effective in containing a disease. This section covered applications of reinforcement learning applied in the context of pandemic control as well as other operational activities in an educational environment.

Modeling

We consider an educational campus with n students. Each student is enrolled in one or more courses from a set of courses. We consider time to be discretized into weeks. At any time, some of the students could be infected and the rest are not. At each time step (week), a community risk value is provided and the current number (percentage) of infected students in each class is known. We treat these as forming the observable state of the system. The RL agent learns a policy that indicates the desired action that should be taken given an observation. The action to be taken is determining the percentage of students to be allowed in a particular course. After the action is taken, an infection model is applied that takes into account the external community risk and the current number of infected students to determine the number of newly infected students for the next week. Then a reward is calculated that depends upon a weighted combination of number of students allowed (more the better) and the number of students infected (fewer the better). The goal of the RL training process is to maximize this reward.

Input Data: For a given campus, we assume the following data will be provided by a school administrator:

- The set of C courses offered
- A set of students and the courses they have registered for
- Community risk value ρ_c - this is a measure of infection risk for students from the community outside the campus, obtained on a weekly basis.

State: In our model, the observed state is defined as follows. At the beginning of the n^{th} week, it is represented as the following tuple: $\langle I_1(n), I_2(n) \dots I_C(n), \rho_c(n) \rangle$, where $I_i(n)$ represents the percentage of infected students in the i^{th} course in week n and the community risk level in week n is $\rho_c(n)$. For simplicity and efficiency, we are currently discretizing the observed state space into a set of discrete levels for both the percentage of infected students in each class as well as the community risk which could be viewed as a percentage (specifically, we use 0, 1, and 2, to represent the ranges between $0 - 33\%$, $33 - 66\%$ and $66\% - 100\%$ respectively.) This could be easily modified to accommodate a more fine-grained discretization at the expense of greater storage and computational complexity for the reinforcement learning.

Action: The action that the agent takes is a proposal to the administrator for what percentage of students from each class to allow to be on campus. Again, for simplicity, we discretize the action also to 3 levels for each course corresponding to $0\%, 50\%$, and 100% respectively. If 0% for a class, this implies that class is going to be scheduled online. If 50% , we mean that half of the students from the class are allowed to attend in person while the rest will attend online, and if 100%, all students allowed

Reward: The reward is a weighted combination of two terms. One for total number of allowed students on campus

across all courses (this is a positive because it gives some educational benefit) and one for the total number of infected students (this is a negative term). The reward is calculated as follows, using a weight parameter $\alpha \in [0, 1]$

$$R(n) = \alpha \sum_{i=1}^{C} A_i(n) - (1 - \alpha) \sum_{i=1}^{C} I_i(n) \quad (1)$$

State Transition: We use a simple, approximate model loosely based on the well-known SIR model (Cooper, Mondal, and Antonopoulos 2020) to capture how many students in a classroom become infected given the initial condition and action of how many students to allow to come into the class. The model also takes into account the community risk. It is given as follows:

$$I(n+1) = min(c_1 \rho_c A_i(n)^2 + c_2 I_i(n) A_i(n), A_i(n)) \quad (2)$$

This model has two terms. The first term accounts for risk due to students who get newly infected from the external community and spread the infection in the classroom, and the second term accounts for risk of infection due to asymptomatic students from the previous week infecting students in the classroom. The minimum function ensures that the number of newly infected students is no more than the number of students allowed in the classroom. This model is approximate and conservative to some extent as it assumes on the one hand that the number of infected students allowed in the classroom is small and on the other that the number of infected students is always smaller than the number of students allowed in the classroom.

Environment: Gym campus-v0 We use OpenAI Gym (Brockman et al. 2016) (a toolkit used by many researchers for developing reinforcement learning algorithms) to develop the environment library campus-gym designed to output operational strategies that school administrators could use to make trade-offs between learning objectives and campus community health and safety. The Gym API defines an interface to integrate various reinforcement learning agent. In our current version, we have encoded the action and observations from the environment into a finite countable state and action space.

RL Agent: Q-learning is a well-known model-free, value-based algorithm (Sutton, Barto, and others 1998). We use the standard tabular Q-learning The goal is to an agent to learn by experiencing sequences of actions by estimated the value of each action any possible state. The Q-value of the action taken is the long-term expected reward under a policy. The estimated Q-values are updated after every *step* by progressively updating the difference between the current estimate and the reward obtained based on . We train a simple agent with 3000 episodes using this approach. Given a task, the agent determines an optimal policy i.e one that maximizes the total discounted expected rewards over episodes.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma max_a Q(s'_t, a'_t) - Q(s_t, a_t)] \quad (3)$$

In figure 1 we provide a high level architecture of CampusPandemicPlanR software blocks.

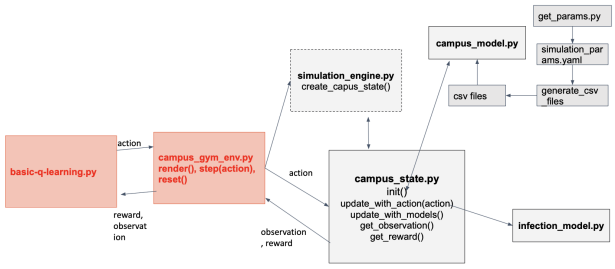


Figure 1: CampusPandemicPlanR Software Blocks

Numerical Results

The RL agent is trained over 3000 episodes, each consisting of 15 weeks of school. We show in figure 3 a visualization of the Q-table after 100, 1000 and 3000 episodes, respectively, as a heatmap. The rows in these images represent states and the columns the actions. With 4 3-value states (representing the infected percentage of students in each of 3 classes and the community risk level), we have $3^4 = 81$ states and with 3 3-value actions (representing the allowed percentage of students in each of the 3 classes), we have $3^3 = 27$ actions. We can see that the Q-table shows more states explored over episodes, and we could see certain strategies (state-action pairs) being reinforced more over time.

Figure 4 shows an increasing aggregated discounted reward value over the training episodes. This demonstrates the Q-learning in action: as more states and actions are explored, the agent finds the most rewarding combinations and its policy improves in terms of the collected rewards.

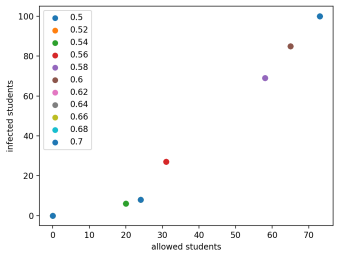
Figure 5 shows how the average reward earned by a trained agent varies over the 15 weeks. The fluctuations observed are in part due to variations in the underlying community risk which changes week by week.

Figure 2 shows the tradeoff between the number of students allowed on campus and the number that are infected, as the tuning parameter α is varied. The RL agent was trained using different values of α. We can see the tradeoff clearly – when α is high, higher number of students are allowed and more of them get infected, and when α is low, few students are allowed on campus and thus few of them get infected.

Conclusions

We have discussed in this paper how reinforcement learning could be applied to develop a decision-aid tool that can help school administrators operate a campus during a pandemic, helping them strike a balance between safety (by keeping students home) and the educational benefit by allowing them to experience in-person instruction on campus. Depending on how a weight parameter in the reward function is selected, different Pareto-optimal tradeoffs between these conflicting goals could be obtained.

This is still early work, and we have many directions for the future. These include more sophisticated and fine-



(a)

Figure 2: Tradeoff between infected and allowed students as α is varied

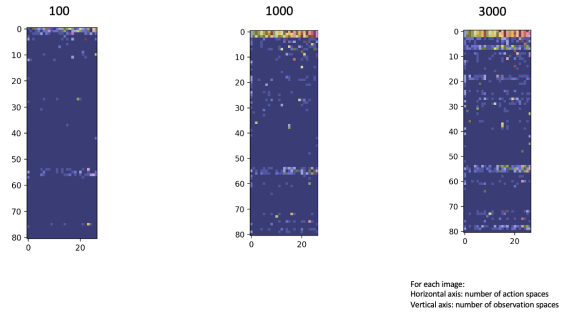


Figure 3: Q-table after 3000 training episodes

grained models for epidemic spread on campuses and making use of real campus datasets (such as in (Hekmati et al. 2021)) as well as exploring the performance of other, more sophisticated reinforcement learning algorithms for larger-scale problem instances. It would also be of interest to explore how humans interact with the RL algorithms for this domain.

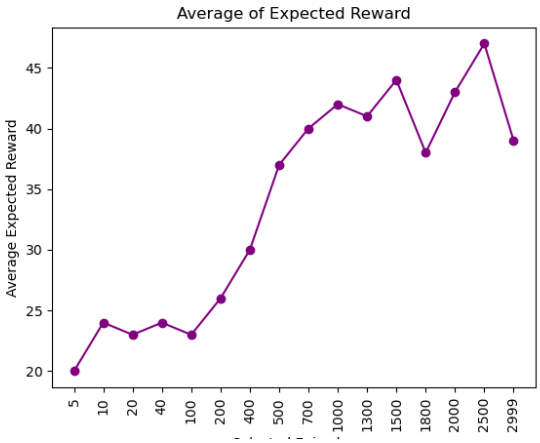


Figure 4: Episode Rewards

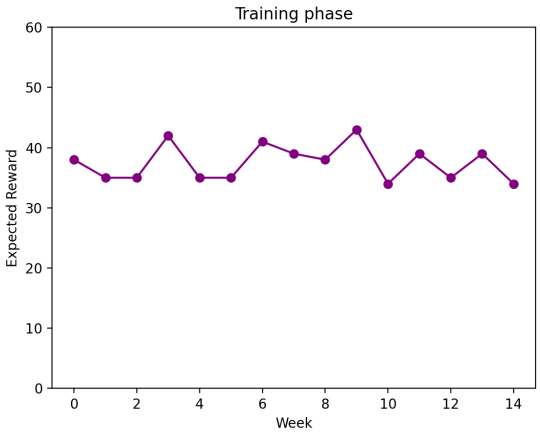


Figure 5: Average rewards by week for trained Q-learning based RL agent

References

Arango, M., and Pelov, L. 2020. Covid-19 pandemic cyclic lockdown optimization using reinforcement learning. *arXiv preprint arXiv:2009.04647*.

Bank, W. 2020. The covid-19 pandemic: Shocks to education and policy responses.

Bassen, J.; Balaji, B.; Schaarschmidt, M.; Thille, C.; Painter, J.; Zimmaro, D.; Games, A.; Fast, E.; and Mitchell, J. C. 2020. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Cooper, I.; Mondal, A.; and Antonopoulos, C. G. 2020. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals* 139:110057.

Di Domenico, L.; Pullano, G.; Sabbatini, C. E.; Boëlle, P.-Y.; and Colizza, V. 2021. Modelling safe protocols for reopening schools during the covid-19 pandemic in france. *Nature communications* 12(1):1–10.

Duong, V.; Pham, P.; Yang, T.; Wang, Y.; and Luo, J. 2020. The ivory tower lost: How college students respond differently than the general public to the covid-19 pandemic. *arXiv preprint arXiv:2004.09968*.

Gressman, P. T., and Peck, J. R. 2020. Simulating covid-19 in a university environment. *Mathematical biosciences* 328:108436.

Hamer, D. H.; White, L.; Jenkins, H. E.; Landsberg, H. N.; Klapperich, C.; Bulekova, K.; Platt, J.; Decarie, L.; Gilmore, W.; Pilkington, M.; et al. 2021. Control of covid-19 transmission on an urban university campus during a second wave of the pandemic. *medRxiv*.

Hekmati, A.; Luhar, M.; Krishnamachari, B.; and Matarić, M. 2021. Simulation-based analysis of covid-19 spread through classroom transmission on a university campus.

arXiv preprint arXiv:2104.04129.

Khamees, D.; Brown, C. A.; Arribas, M.; Murphey, A. C.; Haas, M. R.; and House, J. B. 2020. In crisis: medical students in the covid-19 pandemic. *AEM Education and Training* 4(3):284–290.

Lv, P.; Zhang, Q.; Xu, B.; Feng, R.; Li, C.; Xue, J.; Zhou, B.; and Xu, M. 2021. Agent-based campus novel coronavirus infection and control simulation. *arXiv preprint arXiv:2102.10971*.

Ming, G. F., and Hua, S. 2010. Course-scheduling algorithm of option-based hierarchical reinforcement learning. In *2010 Second International Workshop on Education Technology and Computer Science*, volume 1, 288–291. IEEE.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Uddin, M. I.; Ali Shah, S. A.; Al-Khasawneh, M. A.; Alarood, A. A.; and Alsolami, E. 2020. Optimal policy learning for covid-19 prevention using reinforcement learning. *Journal of Information Science* 0165551520959798.

Yanez, A.; Hayes, C.; and Glavin, F. 2019. Towards the control of epidemic spread: Designing reinforcement learning environments. In *AICS*, 188–199.

Yugay, O.; Kyung, L. T.; and Ko, F. I. 2009. Reinforcement learning coordination with combined heuristics in multi-agent environment for university timetabling. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, 995–1000.