CS-456 Mini-Project Report: Deep Q-learning for Epidemic Mitigation

Chen Zeyi (zeyi.chen@epfl.ch) Wu Haoran (haoran.wu@epfl.ch)

This project aims to find an optimal decision-making policy regarding the mitigation of an epidemic process across Switzerland with the aids from an artificial agent trained through deep Q-learning (DQN). The rest of this report is structured as follows. Section 1 studies the raw behavior of epidemics without any mitigation initiatives. Section 2 includes the simulation results and evaluations regarding Prof Russo's confinement-only policy. Section 3 presents the results produced by a binary-action DQN agent with different exploration policies, and the efficiency is monitored by certain evaluation procedure. Section 4 extends the action space to be more complicated with multiple actions involved. A wrap-up result analysis is then provided in Section 5.

1 Preliminary Studies

In this preliminary section, we examine the intervention-free behavior of the epidemic model. We refer the reader to a schematic representatino of the variable flow and a map of Switzerland in the guidance document.

Question 1.a) Epidemic Model Behavior without mitigation

The action preprocessor was initialized to be null before an episode of 30 weeks was simulated. **The associated results are plotted in Figure 1**, where time is measured in weeks and all the variables share the y axis scaling. It can be inferred from the plots that

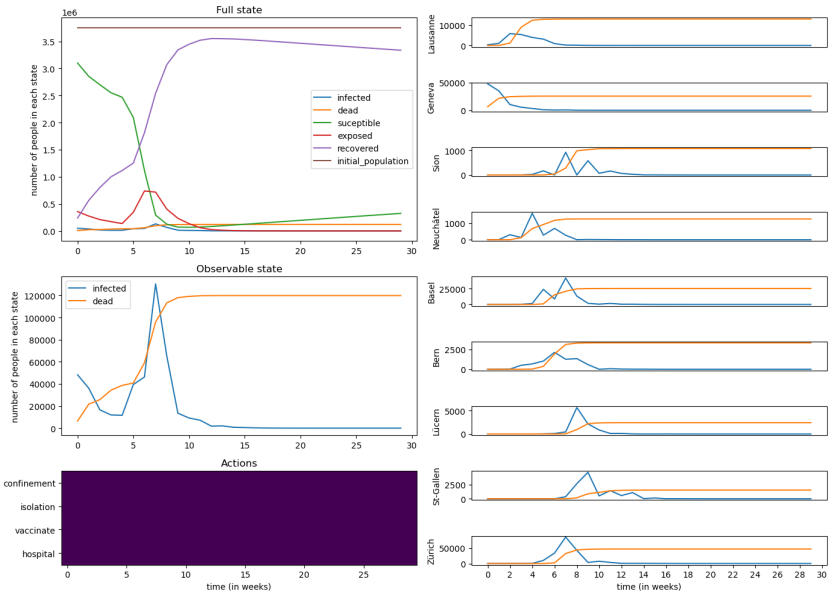


Figure 1: Plots for Question 1. Top left: the plot of variables $s^{[w]}_{\text{total}}, e^{[w]}_{\text{total}}, i^{[w]}_{\text{total}}, r^{[w]}_{\text{total}}, d^{[w]}_{\text{total}}$ over time; Middle left: the plot of variables $i^{[w]}_{\text{total}}, d^{[w]}_{\text{total}}$ over time; Right: the set of plots of variables $i^{[w]}_{\text{city}}, d^{[w]}_{\text{city}}$ over time.

Discussion on the evolution of the variables in Figure 1:

i. From the *Full state* plot, one could notice a sharp decrease in "susceptible" and a simultaneous huge increase in "recover" around the 7th and 8th week, while the other variables, especially "infected", remain at a relatively steady level . The inverse relation between "susceptible" and "recovered" is possibly because that the initial susceptible persons were being exposed thus diagnosed to be infected, yet the number of infected was balanced by the soaring recovered population without large fluctuations (by the way state variables flow). A minor group of the infected patients died of the virus during the period.

ii. From the *Observable state* plot, it can be inferred that **"infected" measured the infected population observed in that particular week excluding those who had recovered, while "dead" as a monotone curve measured the accumulative deaths over the weeks since it is irrecoverable**. A rapid increase in the number of "infected" is witnessed after an initial short reduction (consistent with *Full state* plot), correspondingly followed by a record of growing deaths out of the infection. By the 12th week the infected people dropped nearly to zero and the death rate remained still at around 120,000 persons per week.

iii. According to the city plots and the map of Switzerland, the virus from the source city (which is Geneva in this case) would initially attack accessible cities (Lausanne and Neuchatel) and then spread over the country. Consistent with the "Observable state" plot, infection peaks came in around week 7 in most cities and the number of deaths grew accordingly. Situations tended to become stable after the 15th week.

2 Professor Russo's Policy

Question 2.a) Implementation of Pr. Russo's Policy

Under Prof Russo's policy, a 4-week confinement will be imposed when the number of the infected people at the end of the previous week surpasses 20,000 (See the codes for the implementation. Figure 2 gives the associated plots regarding the policy where time is measured in weeks and all the variables share the y axis scaling.

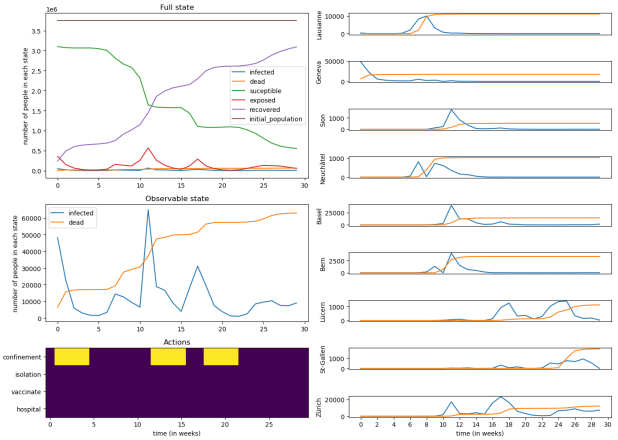


Figure 2: Plots for Question 2a. Top left: the plot of variables $s^{[w]}_{\text{total}}, e^{[w]}_{\text{total}}, i^{[w]}_{\text{total}}, r^{[w]}_{\text{total}}, d^{[w]}_{\text{total}}$ over time; Middle left: the plot of variables $i^{[w]}_{\text{total}}, d^{[w]}_{\text{total}}$ over time; Bottom left: the plot of the actions taken by Russo's policy; Right: the set of plots of variables $i^{[w]}_{\text{city}}, d^{[w]}_{\text{city}}$ over time.

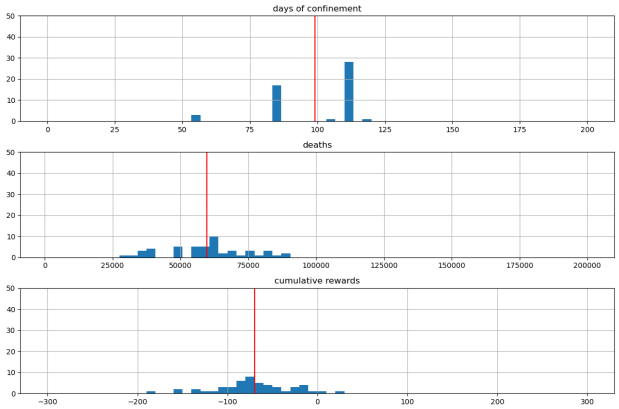


Figure 3: Histograms for Question 2b. Corresponding variables from top to bottom: $N_{\text{confinement}}$, N_{deaths} and $R_{\text{cumulative}}$.

Discussion on how the epidemic simulation responds to Pr. Russo's Policy:

i. Compared to the one in Figure 1, the "Full state" plot in Figure 2 indicates a slow-down for both the reduction in the number of susceptible persons and the rise in the number of the recovered, especially when there is a confinement. The reason is that people are decreasingly exposed to the virus under confinement (as witnessed in the plot) thus the susceptible individuals are less vulnerable to infection. As an another consequence, the "infected" and "dead" curves remain at an even lower level than in Figure 1.

ii. From the *Observable state* plot one can observe that total number of the infected and the dead persons was largely ameliorated under Prof Russo's policy across the examined thirty weeks, the highest of which dropped from 12,000 to 60,000 approximately. Sensitive to the policy, the infection rate fell down rapidly each time the confinement is initiated, which temporarily bridled the weekly number of deaths during the period.

iii. The cities' plots present a delay of the peak of infection among the population, which may spring from the reduced exposures to the virus. The rise of death numbers were also put off accordingly, in comparison to Figure 1.

Question 2.b) Evaluation of Pr. Russo's Policy

Fifty evaluation procedures were applied with $N_{\text{confinement}}$, $R_{\text{cumulative}}$ and N_{deaths} recorded, the means of which are 99.12 (days), -70 and 59739.6 (persons) in respective. The histograms included in Figure 3 demonstrate

the distributions of the three variables. It is worth-noting that the variances for the variables are large, which implies the instability of Prof Russo's policy.

3 A Deep Q-learning approach

Settings: The samples in the observation space were scaled by 100 followed by a $(\cdot)^{\frac{1}{4}}$ function, and the *ADAM* optimizer was employed by our DQN agent. The hyperparameters used in this section are presented in Table 1 as suggested. Note that the input layer has 126 ($:= 2 \times 7 \times \#$ cities) dimensions while the output layer is 2-dimensional indicating the q-values for the binary actions.

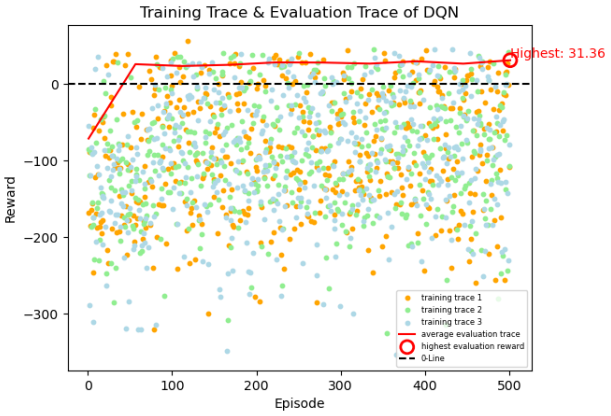
hyperparameter	Value
neural network architecture	A 3 hidden layer fully connected neural net with layers of size **input size**, 64, 32, 16, **output size**
activations	ReLU activations after each layers (except the output which is purely linear)
target update rate	(fully) update the target network every 5 episodes
training length	train for 500 episodes
learning rate	$5 \cdot 10^{-3}$
discount factor γ	0.9
buffer size	20000
batch size	2048

Table 1: Training hyperparameters.

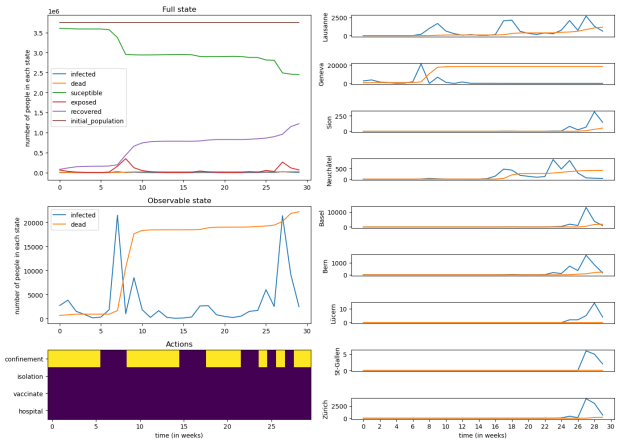
3.1 Deep Q-Learning with a Binary Action Space

Question 3.a) DQN with Fixed Exploration

An ϵ-greedy DQN agent $\pi_{\text{FE-DQN}}$ with $\epsilon = 0.7$ was implemented, trained and evaluated, during which the training trace and the evaluation trace were recorded. **A visualization of the reward evolution in both the training and evaluation traces is given in Figure 4a. It seems that $\pi_{\text{FE-DQN}}$ is learning a meaningful policy since the reward exhibits an increasing trend over the episodes and tends to converge at around 30.** Three example episodes under the optimal empirical policy $\pi^*_{\text{FE-DQN}}$ were tracked **with one of the resulting dynamic plotted in Figure 4b**.



(a) Visualization for the training trace (scatters) and the average evaluation trace (line) of the fixed-exploration DQN $\pi_{\text{FE-DQN}}$



(b) Dynamic plots induced the optimal fixed-exploration DQN $\pi^*_{\text{FE-DQN}}$

Figure 4: Plots for the DQN with Fixed Exploration (Question 3.a)

Interpretation of the policy $\pi^*_{\text{FE-DQN}}$:
As witnessed in *Observable state* Figure 4b, a confinement would be given timely whenever the infected population upsurged and performed effectively in pulling the number back to the baseline. This is achieved indirectly by reducing the exposure of the virus to the susceptible, as noticed in *Full state* where the "suceptible" number stopped dropping each time a confinement was imposed. The control over the infected number also resulted in less deaths.

Question 3.b) DQN with Decreasing Exploration

We consider another DQN agent $\pi_{\text{DE-DQN}}$ where the exploration probability ϵ depends on the time:

$$\epsilon(t) = \max\left(\frac{\epsilon_0(T_{\max} - t)}{T_{\max}}, \epsilon_{\min}\right) \tag{1}$$

where t is the episode number, $T_{\max} = 500$, $\epsilon_0 = 0.7$ and $\epsilon_{\min} = 0.2$. Like in the setting of fixed epsilon, **a visualization for the corresponding training and evaluation traces are plotted in Figure 5**. We plot both traces for $\pi_{\text{FE-DQN}}$ and $\pi_{\text{DE-DQN}}$ together in Figure 6 for comparison purposes.

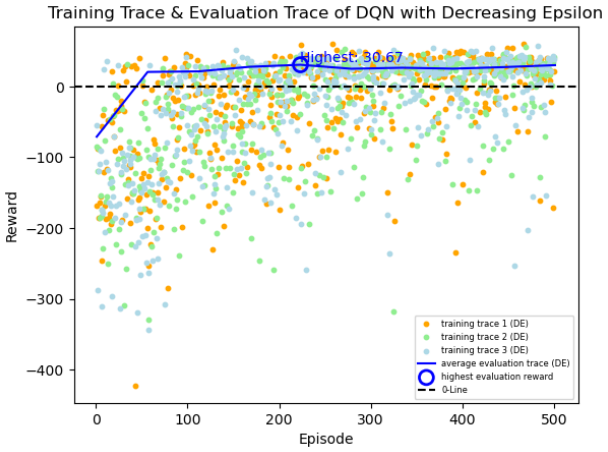


Figure 5: Visualization for the training trace (scatters) and the average evaluation trace (line) of the decreasing-exploration DQN $\pi_{\text{DE-DQN}}$

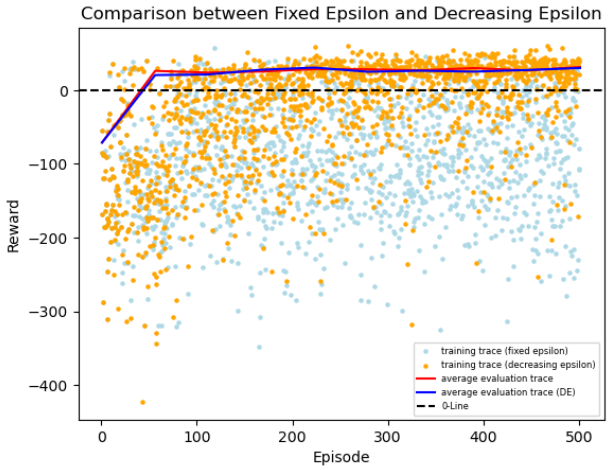


Figure 6: A comparison between the traces of $\pi_{\text{FE-DQN}}$ and $\pi_{\text{DE-DQN}}$

Discussion on comparing $\pi_{\text{FE-DQN}}$ and $\pi_{\text{DE-DQN}}$:
The empirical evidence shows that, in this particular case, the two agents perform comparably well with a high similarity in the shape and the values along the evaluation trace lines. This can be explained by the fact that $\epsilon(t)$ is close to ϵ_0 when t is small thus abundant exploration was guaranteed for both agents at the initial stage, during which the near-optimal policy could have been found. Yet, in later episodes, the training samples of $\pi_{\text{DE-DQN}}$ have their rewards more concentrated above zero for more biases on the exploitation.

Question 3.c) Evaluating $\pi^*_{\text{FE-DQN}}$ against Pr. Russo's policy

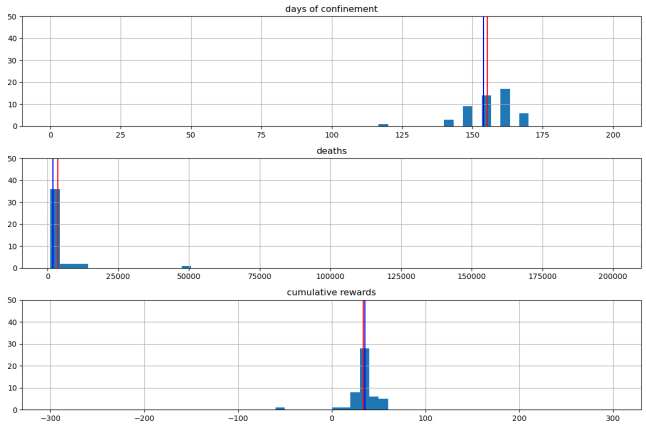


Figure 7: Histograms for Question 3c under π^*_{FE-DQN}. Corresponding variables from top to bottom: $N_{\text{confinement}}$, N_{deaths} and $R_{\text{cumulative}}$. The red and the blue vertical lines represent mean and median respectively.

The histogram in Figure 7 was generated using the policy $\pi^*_{\text{FE-DQN}}$. **In contrast to Figure 3 for Prof Russo's policy, the mean reward turned to positive and the mean death was largely reduced by around 60,000, which means that $\pi^*_{\text{FE-DQN}}$ prevails over Prof Russo's policy**. The cost for such an advantage is the increased number of days under confinement.

4 Policies Towards More Complex Actioin Spaces

Settings: In this section, all agents were trained with decreasing exploration following the scheme (1). Analogue to Section 3, the samples in the observation space were scaled by 100 followed by a $(\cdot)^{\frac{1}{4}}$ function and *ADAM* optimizer was employed. All hyperparameters other than *training length* and *learning rate*, which will be specified in later experiments, followed Table 1. The sizes of the input layer and the output layer depend on the action manners.

4.1 Toggle-Action-Space Multi-Action Agent

During the training in this part, a null action and four toggle actions are provided when the agent makes decisions. Note that the input layer has 378 ($:= (2+4) \times 7 \times \#$ cities)dimensions while the output layer is 5-dimensional indicating the q-values for the five actions at the corresponding state.

Question 4.1.a) (Theory) Action space design

There are two main reasons for using the toggle action space instead of directly computing $Q(s, a)$ for each action:

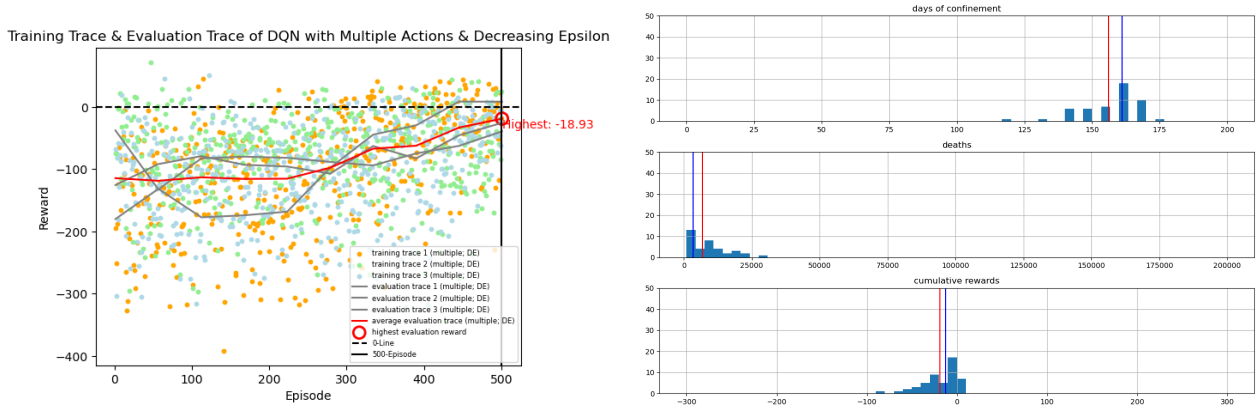
i. **(On network architecture)** Such a toggle action space reduces the number of the output neurons form 8 to 5 thus decreases the amount of weights and biases that need to be learned and stored in the network, which leads to lower complexity and memory requirements.

ii. **(On training)** Toggled actions control the agent's behavior by enabling smoother transitions between different action choices by partially modifying the strategy rather than completely switching actions, which promotes learning stability during the training procedure.

Question 4.1.b) Toggle-action-space multi-action policy training

The toggled action space was implemented and the observation space was updated through the observation preprocessor, on which the DQN agent was trained accordingly. Our selection for the learning rate was 1×10^{-3} which proved to give stable learning procedures and appropriate convergence behaviors during the 500 training episodes.

Remark: We would like to remark on why we prefer $\text{lr} = 1 \times 10^{-3}$ than the suggested 1×10^{-5} for the back-propagation process. Figure 8 presents the traced data when the agent learnt at a rate of 1×10^{-5} and the associated evaluation histogram for the obtained optimal policy. The grey lines in Figure 8a represent



(a) Visualization for the training trace (scatters) and the average evaluation trace (red line) for DQN with $lr = 1 \times 10^{-5}$.

(b) The evaluation histogram for the obtained optimal policy after learning with $lr = 1 \times 10^{-5}$

Figure 8: Plots for the DQN agent with $lr = 1 \times 10^{-5}$

the three training traces before averaged, whose tracks significantly vary. The highly distinguished decisions witnessed in the three traces imply that the learning could be extremely unstable, and 500 episodes may not be sufficient for the process to converge. Besides, the empirical rewards in Figure 8b are most negative, which substantiates that the agent has not learnt successfully. One approach to resolve the problem is to train longer. Indeed, the trace lines become rather consistent after 650 episodes of training when we extended each training period to 1000 episodes as shown in Figure 9, and the scatters become concentrated above zero. Another way is to raise the initial learning rate as we did. We inclined to the latter solution to guarantee the fairness in further

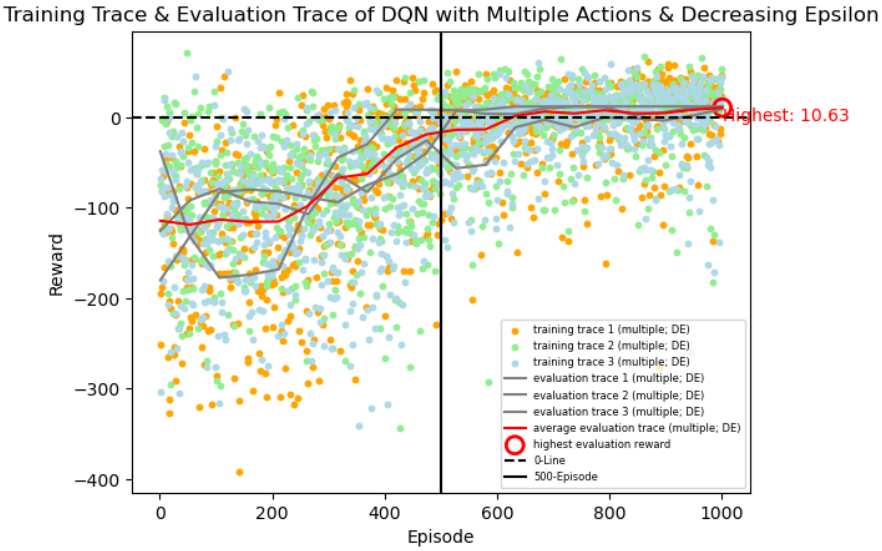
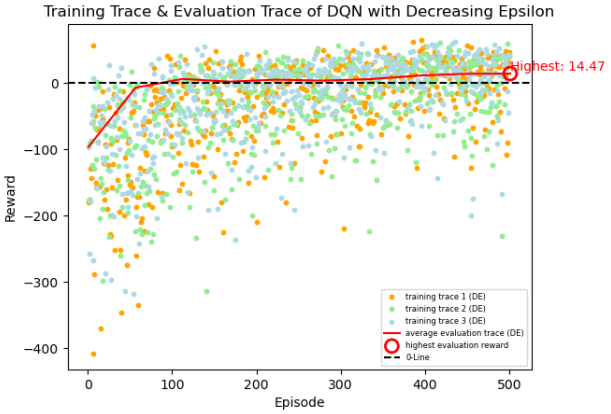


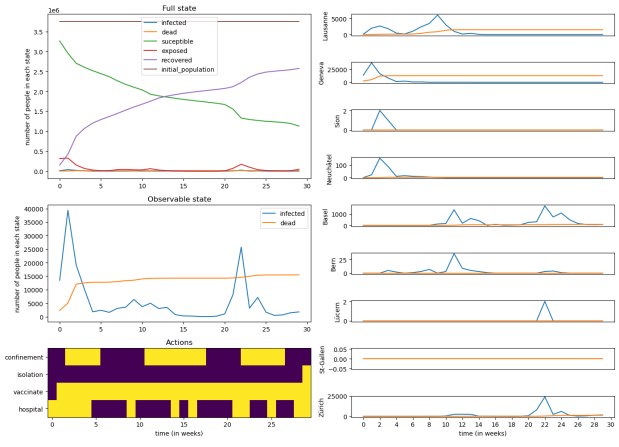
Figure 9: Visualization for the training trace (scatters) and the average evaluation trace (red line) for DQN with $lr = 1 \times 10^{-5}$ for 1000 training episodes.

comparisons with other policies.

With lr $= 1 \times 10^{-3}$, it was found that **the learning was much stabilized and properly converged** when the training length is 500 as in Table 1. See the associated training and evaluation traces in Figure 10.



(a) Visualization for the training trace (scatters) and the average evaluation trace (line) of the decreasing-exploration DQN π_{toggle}.



(b) Dynamic plots induced by the optimal toggled policy π^*_{toggle}

Figure 10: Plots for the DQN on the toggle-action-space (Question 4.1.b)

Interpretation of the policy π^*_{toggle}:
At the initial stage of vaccinating, the infection number proliferated as people recovered without it were getting infected. Confinements were imposed to cool down the number of the infected whenever there was a spike, and the mortality was controlled at a decent level with hospital beds constantly added. A nationwide isolation would be executed to slow down the propagation between cities in response to simultaneous high infections observed in several regions, like during the last few weeks in Figure 10b (see the last four city plots).

Question 4.1.c) Toggle-action-space multi-action policy evaluation

The evaluation histogram for π^*_{toggle} is given in Figure 11. Under π^*_{toggle}, the empirical distributions of $N_{\text{confinement}}$, N_{deaths} and $R_{\text{cumulative}}$ have more outliers than the ones in Figure 7 for the binary-action $\pi^*_{\text{FE-DQN}}$, thus medians would be more informative than means. The large variation in those variables may spring from a more complicated set of actions. Compared to $\pi^*_{\text{FE-DQN}}$, there seems to be a compromise in death number to achieve a slightly better median reward with fewer confinement days (See Section 5 for detailed statistics).

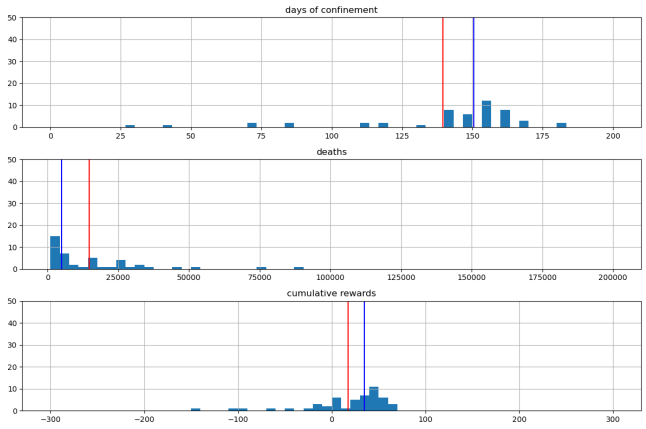


Figure 11: Evaluation Histogram for π^*_{toggle}

Question 4.1.d) Assumption made by toggle-action-space policy

The toggle-action-space policy makes two essential assumptions on the action space:

i. The action space is discrete (or stricter, binary). The toggled actions would be over-simplified when confronted with continuous action spaces. For example, when the dosage is considered in vaccination, the action is no longer discrete can cannot be measured by toggling.

ii. The actions in the action space are independent, i.e. not mutually affected. When the actions are sequential (e.g. vaccination and booster) or mutually exclusive (e.g. lock-down and vaccination), the toggled policy would possibly result in infeasible action sets.

4.2 Factorized Q-values, multi-action agent

A multi-agent based on factorized Q-values is implemented and trained in this section, according to the fashion

$$Q(\mathbf{a}^{[w]}, s) = Q(a^{[w]}_{conf} \cup a^{[w]}_{\text{isol}} \cup a^{[w]}_{\text{hosp}} \cup a^{[w]}_{\text{vacc}}, s) = \sum_{\mathfrak{d} \in \text{decisions}} Q(a_{\mathfrak{d}}, s). \tag{2}$$

Learning from the experience in Section 4.1, we again adopt 1×10^{-3} as the learning rate. The observation space is no longer extended thus the input size returns to 126 and the output size is 8 ($:= 4 \times 2$), while there would be $16 = 2^4$ actual action sets. It is remarkable that in our implementation for the action space, the actions were translated by a decimal-binary converter such that a one-to-one correspondence was constructed between integers and binary quadruples, e.g. e.g.

$$a = 10 \Leftrightarrow \{\text{confinement} = 1, \text{isolation} = 0, \text{hospitality} = 1, \text{vaccination} = 0\}.$$

Question 4.2.a) Multi-action factorized Q-values policy training

The evaluation trace and training trace gathered from the multi-action factorized Q-value agent are plotted in Figure 12 and a comparison with the toggle-action-space policy is given in Figure 13. **Generally speaking, the agent was learning well since the curve grows steadily with an acceptable final reward.** Refer to Figure 14 for an arbitrary episode of dynamics induced by π^*_{factor}.

Interpretation of the policy π^*_{factor}:
Unlike π^*_{toggle}, π^*_{factor} seldom imposes vaccination but relies more on confining and adding hospital beds to control the number of infected and the dead. Occasionally, a short-term isolation would be announced when there were multiple peaks of infection occurring in different cities at the same period (like the week 6 and week 7 in the city plots of Figure 14). As a consequence, the infection rebounds more frequently than under π^*_{toggle} as spotted in *Observable state* since people without vaccination were more likely to be re-infected. **The policy does achieve acceptable death rate and rewards yet continuous extra hospitalization without vaccination may not be realistic.**

Question 4.2.b) Multi-action factorized Q-values policy evaluation

Figure 15 gives the evaluation histogram associated with π^*_{factor}. In contrast to Figure 11 for π^*_{toggle}, it turned out that the distributions are more centralized without many outliers, which implies that the death number

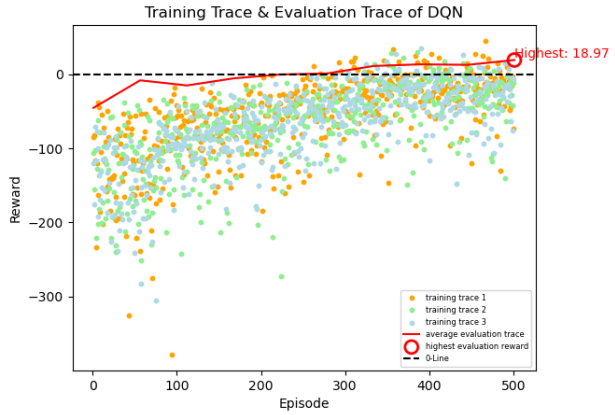


Figure 12: Visualization for the training trace (scatters) and the average evaluation trace (line) of toggle-action-space DQN π_{toggle}

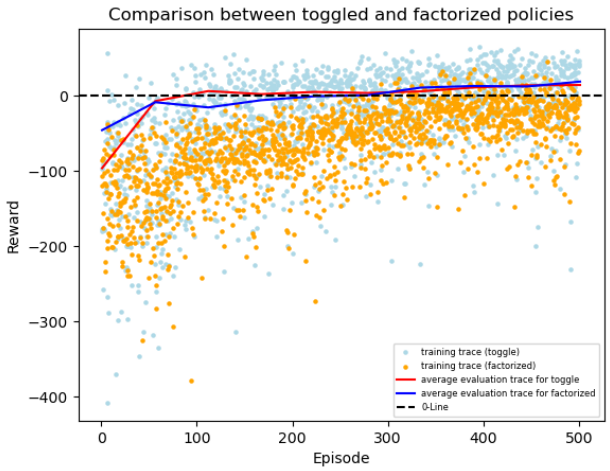


Figure 13: A comparison between the traces of π_{toggle} and π_{factor}

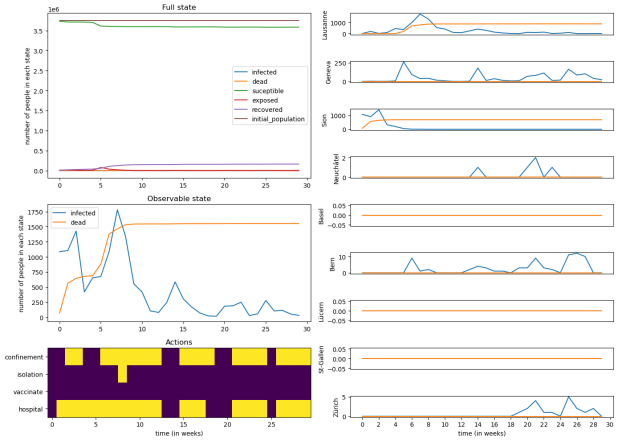


Figure 14: Dynamic plots induced by the optimal factorized policy π^*_{factor}.

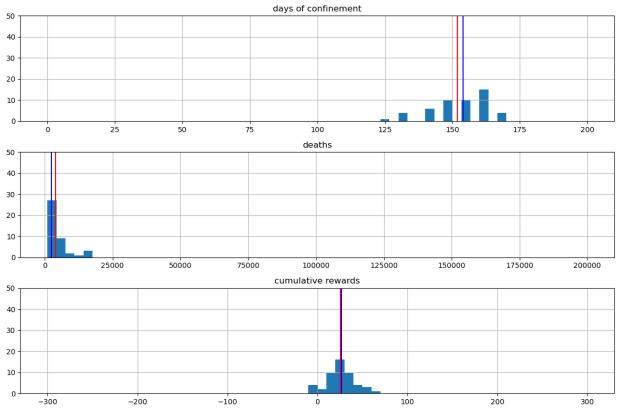


Figure 15: Evaluation Histogram for π_{factor}

was steadily controlled at a low level and the rewards were more likely positive. Confinement day counts are around 150, analogue to the toggled case.

Question 4.2.c) Assumption made by Factorized-Q-values

A key assumption for the factorized Q-values is the independence between actions, i.e. the q-values out of individual actions would not be intervened during interactions. Otherwise, the q-values from different actions may not be summable and the calculation in (2) may not hold. An action space where the factorized Q-values are not suitable is, for example, the one containing actions "mask wearing" and "confinement", in which the q-value of the former would be useless when the latter is initiated thus the two actions cannot have their q-values factorized.

5 Wrapping Up

Question 5.a) (Result analysis) Comparing the training behaviors

From Figure 16, one may conclude that each within a 500-episode training, single-action DQNs outperformed the multi-action policies (toggled and factorized) from a reward perspective, as the former found high-level rewards at an early stage yet the latter climbed up slowly to some lower final rewards. This may due to the concision of the binary action space which eased the exploration procedures, while the multi-action agents suffered from balancing exploration and exploitation among complicated action combinations. However, this conclusion only applies to the 500 episodes' training, since the evaluation curves for both toggled and factorized policies demonstrated growing tendencies, thus have the potential to win over the binary-action policies when given more training time.

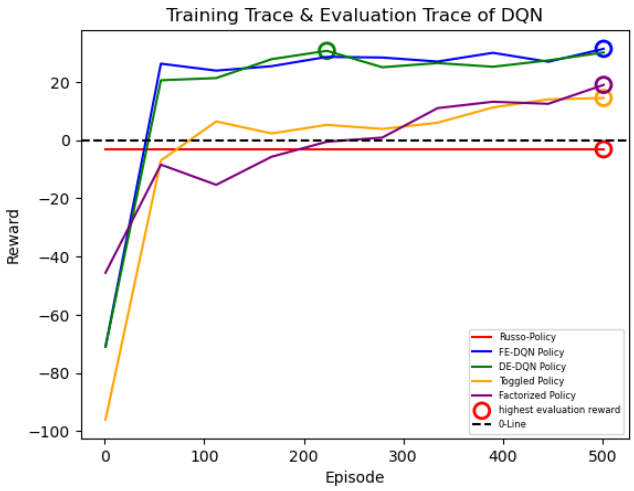


Figure 16: Evaluation curves for different policies

Question 5.b) (Result analysis) Comparing policies

As mentioned in Section 4, median was also monitored besides mean to avoid misinformation carried by extreme outliers. Table 2 ad Table 3 give the results of the six metrics induced by the optimal policies in different types.

Policy	avg$[N_{\text{confinement}}]$	avg$[N_{\text{isolation}}]$	avg$[N_{\text{vaccination}}]$	avg$[N_{\text{hospital}}]$	avg$[N_{\text{deaths}}]$	avg$[R_{\text{cumulative}}]$
π_{Russo}	**99.1**	-	-	-	59739.6	-70.0
$\pi_{\text{FE-DQN}}$	159.88	-	-	-	**2899.4**	**32.8**
$\pi_{\text{DE-DQN}}$	159.0	-	-	-	3130.0	32.1
π_{toggle}	139.6	1.82	63.0	**37.9**	14643.0	17.8
π_{factor}	151.9	**1.68**	**1.26**	66.5	3899.7	25.7

Table 2: Empirical means of monitored variables under different policies (optimal values in bold)

Policy	med$[N_{\text{confinement}}]$	med$[N_{\text{isolation}}]$	med$[N_{\text{vaccination}}]$	med$[N_{\text{hospital}}]$	med$[N_{\text{deaths}}]$	med$[R_{\text{cumulative}}]$
π_{Russo}	**112.0**	-	-	-	60942.5	-71.9
$\pi_{\text{FE-DQN}}$	161.0	-	-	-	1732.0	32.9
$\pi_{\text{DE-DQN}}$	161.0	-	-	-	**1139.5**	33.9
π_{toggle}	150.5	0.0	0.0	**31.5**	4913.5	**34.9**
π_{factor}	154.0	0.0	0.0	63.0	2366.5	26.6

Table 3: Empirical medians of monitored variables under different policies

Discussion: Prof Russo's policy led to the least number of confinement days, yet sacrificing on significantly large amounts of deaths and cumulative rewards. From mean perspective, $\pi_{\text{FE-DQN}}$ proved to achieve the best death number and the cumulative reward, respectively on which $\pi_{\text{DE-DQN}}$ and π_{toggle} were the winners when median is considered. Between the two multi-action agents, π_{toggle} performed slightly better in isolation and vaccination number while π_{toggle} required less extra beds in hospital.

Question 5.c) (Interpretability) Q-values

The heat-map of the evolution of all Q-values with time for $\pi_{\text{FE-DQN}}$ and π_{factor} are plotted in Figure 17 and Figure 18 in respective. The policies are rather interpretable as the decision with a higher q-value is made between every binary-action pair, and factorized q-value is maximized when each element has the best value.

Question 5.d) Cumulative reward and the number of actions

As shown in Question 5.a), the cumulative reward is not necessarily increasing and optimized with the increasing number of actions, of which the reasons are as follows:

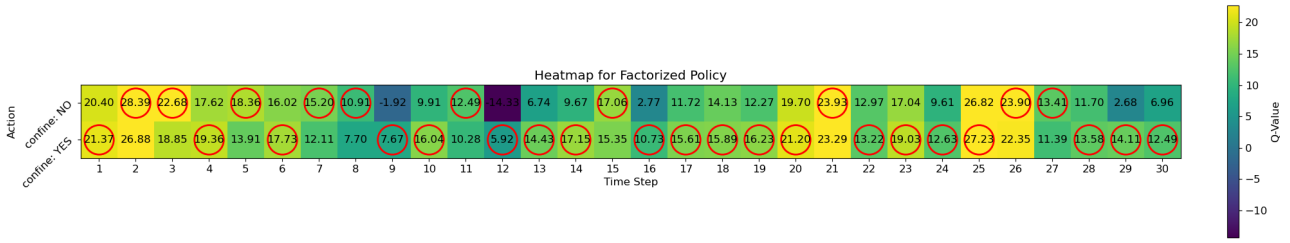
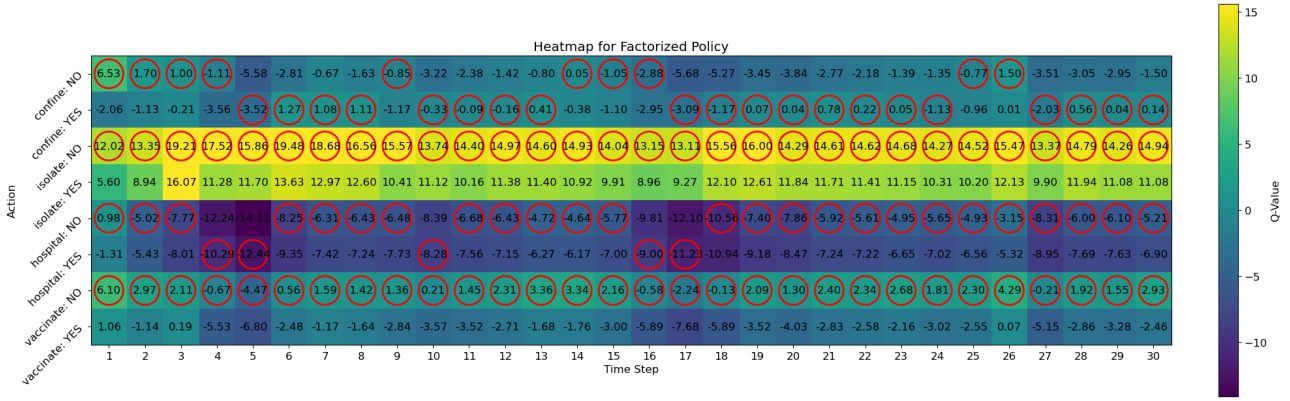
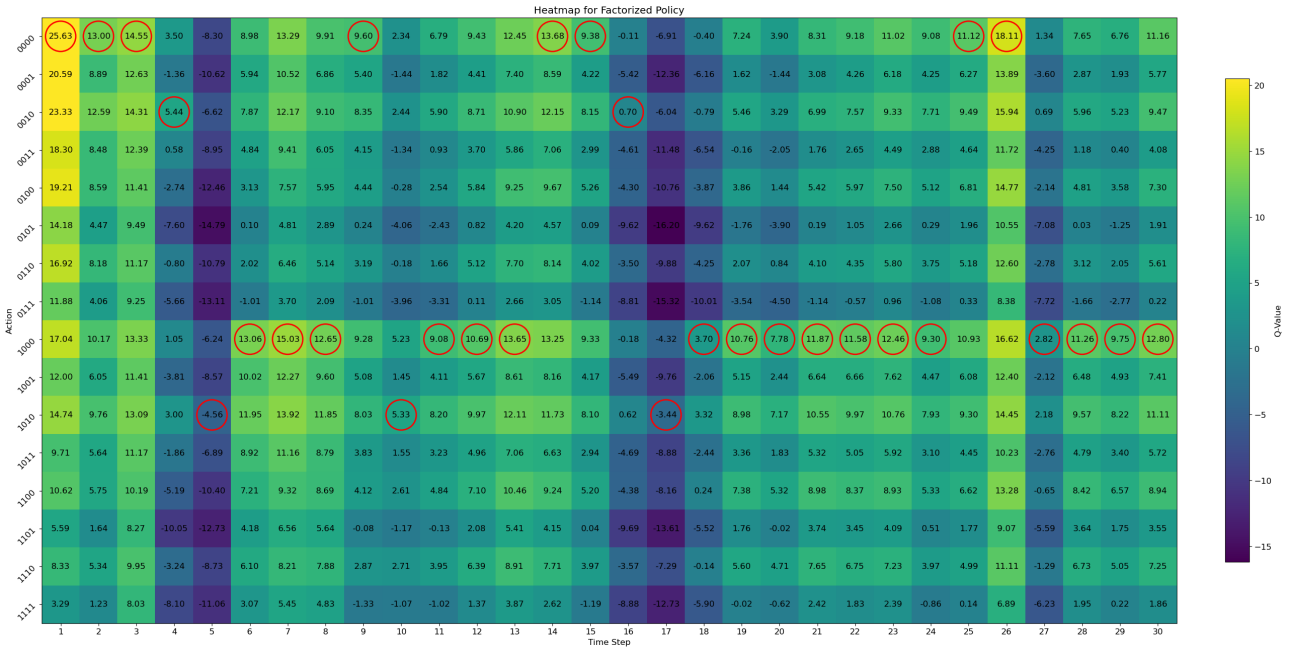


Figure 17: Heat-map for Q-value evolution of $\pi_{\text{FE-DQN}}$



(a) Heat-map for **individual action's** Q-value evolution of π_{factor}.



(b) Heat-map for **factorized** Q-value evolution of π_{factor}. Note that the action sets are represented by quadruples on the y-axis with Boolean values whose indices correspond to "confine", "isolate", "hospital" and "vaccinate" respectively; e.g. "1010" means (confine, not isolate, hospitalize, not vaccinate).

Figure 18: Heat-maps for π_{factor}. Note that in each step, the selected factorized q-value in Figure 18b is exactly the sum of the selected individual q-values in Figure 18a.

i. Action costs would be increased when more actions are activated and announced, according to (3).

$$R^{[w]} = R_{\text{c}} - \mathcal{C}(\mathbf{a}^{[w]}) - D \cdot \Delta d^{[w]}_{\text{total}} \tag{3}$$

ii. The number of deaths per week $\Delta d^{[w]}_{\text{total}}$ may not drop under multiple actions when the policy is not optimal, e.g. the toggle-action-space policy with $\text{lr} = 1 \times 10^{-5}$ as shown in Figure 8.