

Optymalizacja systemu sygnalizacji świetlnej w oparciu o przepływowy model ruchu pojazdów.

Michał Lis

4 lipca 2019

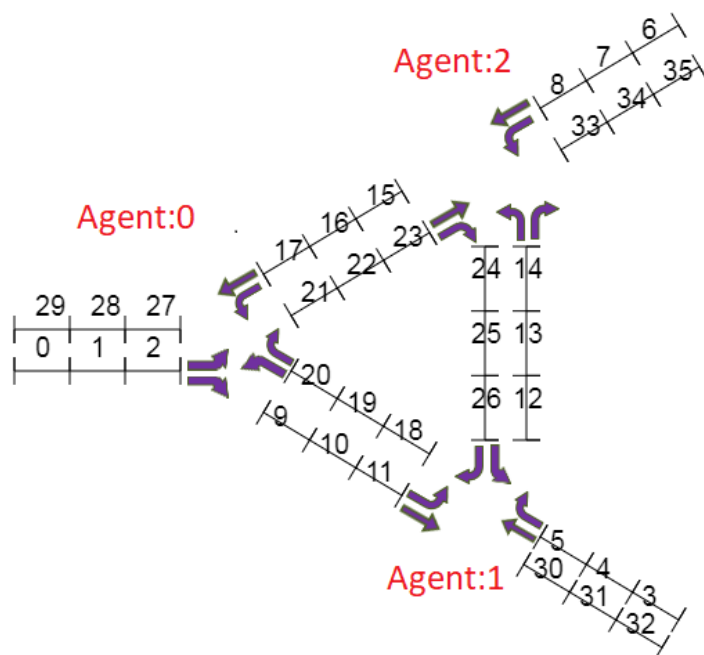
Spis treści

1	Środowiska symulacyjne i ich nauka	5
1.1	Środowisko 4	5
1.1.1	Sygnalizacje świetlne	6
1.1.2	Matematyczny zapis aktualnej sygnalizacji świetlnej	6
1.1.3	Przepływ pojazdów	6
1.1.4	Przepływ pojazdów -	9
1.2	Uczenie Środowiska 4	9
1.2.1	Podejście 1	9
1.2.2	Monitorowanie decyzji dla wybranego stanu	11
1.3	Analiza wyników uczenia	12
1.4	Monitorowanie kolejnego stanu	14

Rozdział 1

Środowiska symulacyjne i ich nauka

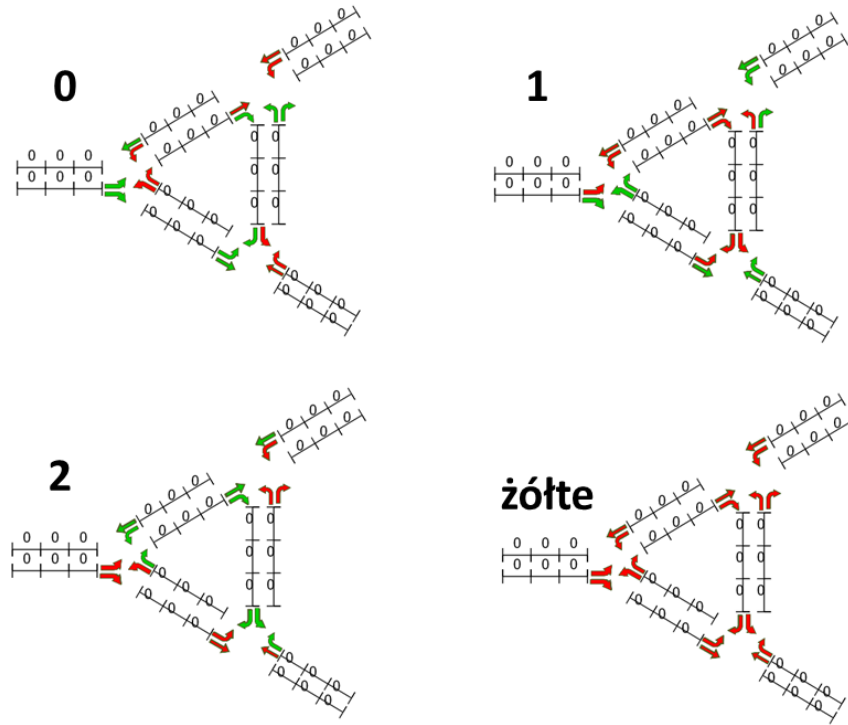
1.1 Środowisko 4



Rysunek 1.1: środowisko 4

Środowisko posiada 12 jednokierunkowych dróg. Każda droga ma 3 odcinki co daje w sumie 36 odcinków (są numerowane od 0 co widać na rysunku 1.1). W sieci dróg znajdują się 3 skrzyżowania. Do każdego z nich jest przypisany agent, który odpowiada za sterowanie sygnalizacją świetlną.

1.1.1 Sygnalizacje świetlne



Rysunek 1.2: środowisko 4 - fazy świateł

Każde skrzyżowanie posiada 4 fazy świetlne przedstawione powyżej. Fazy 0, 1 i 2 posiadają pewne zielone światła i umożliwiają ruch. Automatycznie ustawiana jest faza żółtych świateł przez 2 interwały czasowe w przypadku podjęcia akcji zmiany aktualnej fazy. Agent może także oczywiście przedłużyć aktualną fazę.

1.1.2 Możliwe do wykonania manewry

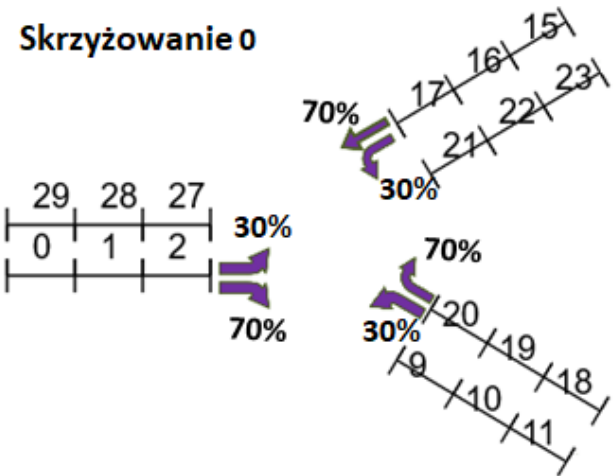
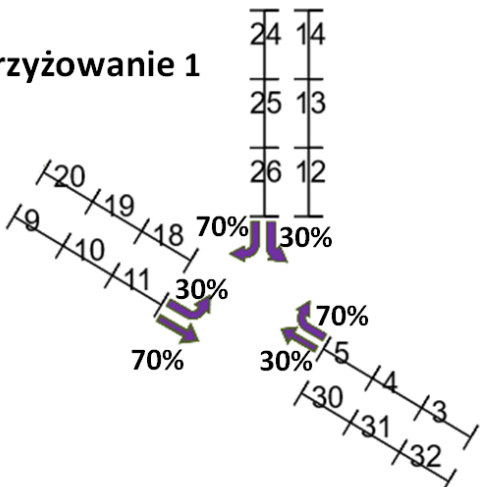
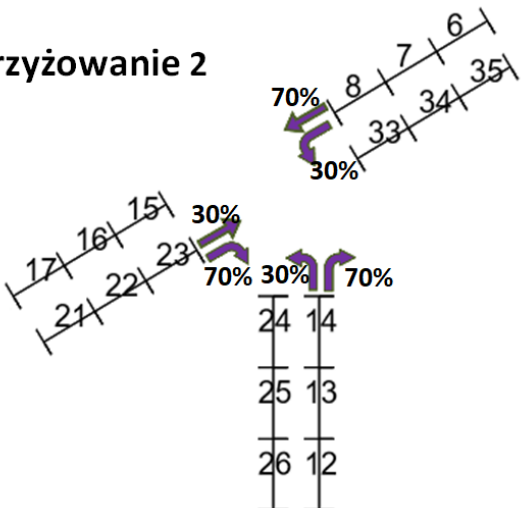
Niech manewr (m,n) będzie manewrem polegającym na bezpośrednim przejeździe z odcinka m na odcinek n . Zostanie zdefiniowana macierz sygnalizacji świetlnej S . Określa ona wykonalność dowolnego manewru.

$$S_{mn} = \begin{cases} 0 & \text{dla nieistniejącego manewru} & (1.1) \\ 0 & \text{dla manewru wstrzymanego przez czerwone światło} & (1.2) \\ 1 & \text{dla manewru zezwolonego przez zielone światło} & (1.3) \\ 1 & \text{dla istniejącego manewru bez sygnalizacji świetlnej} & (1.4) \end{cases}$$

1.1.3 Przepływ pojazdów

Pojazdy pokonują jeden odcinek podczas jednego pełnego interwału czasowego. Każde skrzyżowanie posiada dwa możliwe manewry do wykonania. Prawdopodobieństwa skrętu w prawo

to zawsze 70 procent. Jeśli nie ma skrętu w prawo - to jazda prosto ma prawdopodobieństwo 70 procent. Jazda w lewo z kolei zawsze ma pozostałe 30 procent prawdopodobieństwa. Szczegółowy rozkład wraz z wartościami macierzy P jest przedstawiony w poniższej tabeli. Pozostałe wartości macierzy P to zera.

Skrzyżowanie	P
<p>Skrzyżowanie 0</p> 	$P[2, 9] = 0,7$ $P[2, 21] = 0,3$ $P[20, 21] = 0,7$ $P[20, 27] = 0,3$ $P[17, 27] = 0,7$ $P[17, 9] = 0,3$
<p>Skrzyżowanie 1</p> 	$P[11, 30] = 0,7$ $P[11, 12] = 0,3$ $P[5, 12] = 0,7$ $P[5, 18] = 0,3$ $P[26, 18] = 0,7$ $P[26, 30] = 0,3$
<p>Skrzyżowanie 2</p> 	$P[23, 24] = 0,7$ $P[23, 33] = 0,3$ $P[14, 33] = 0,7$ $P[14, 15] = 0,3$ $P[8, 15] = 0,7$ $P[8, 24] = 0,3$

Odpyływ pojazdów z układu następuje na końcach odcinków 29,32 oraz 35. Z tego względu zostaje wprowadzona macierz rzadka $\mathbb{1}$ o wymiarach 36 na 36. Ma ona na celu usuwać

$$\mathbb{1}_{mn} = \begin{cases} 0 & \text{dla } m \neq n \vee n \in 29, 32, 35 \\ 1 & \text{dla pozostałych przypadków} \end{cases} \quad (1.5)$$

Wiersze i kolumny macierzy są numerowane od 0 - zgodnie z konwencją przyjętą w pracy. Macierz stanowa A jest określona następująco:

$$A_{ij} = \begin{cases} 0 & \text{dla } S[i, j] = 0 \\ P[i, j] & \text{dla } S[i, j] = 1 \end{cases} \quad (1.7)$$

$$\begin{cases} 1 - \delta(i) & \text{dla } i=j \end{cases} \quad (1.8)$$

$$(1.9)$$

Gdzie delta jest sumą wszystkich pozostałych liczb z kolumny i , czyli:

$$\delta(i) = \sum_{j \in \{0, \dots, 35\}, j \neq i} P[i, j]$$

1.1.4 Przepływ pojazdów -

Powyzsze przedstawienie macierzy stanowej nie zawiera w sobie jeszcze pojęcia korka. W jednym interwale czasowym może przejechać przez skrzyżowanie astronomiczna wręcz liczba pojazdów. Dodane zostanie zatem ograniczenie do maksymalnie 10 pojazdów przejeżdżających w trakcie jednego interwału czasowego. Należy sformułować funkcję, która określi przepływ z uwzględnieniem tworzenia się korka w przypadku większej liczby pojazdów. Niech i, j oznaczają rozważane odcinki wlotowe i wylotowe. Wtedy funkcja korka jest następująca:

$$f(i, j) = \begin{cases} 0 & \text{dla } S[i, j] = 0 - \text{czerwone światło} \\ P[i, j] & \text{dla } S[i, j] = 1 \wedge P[i, j]x[i] < 10 - \text{zielone światło, bez korka} \\ \frac{10}{x[i]} & \text{dla } S[i, j] = 1 \wedge P[i, j]x[i] \geq 10 - \text{zielone światło i korek} \end{cases} \quad (1.10)$$

Macierz stanowa A przedstawia się następująco:

$$A = \begin{bmatrix} 1 - S & f(1, 0) & \dots & f(35, 0) \\ f(0, 1) & 1 - S & \dots & f(35, 1) \\ f(0, 2) & f(1, 2) & \dots & f(35, 2) \\ \dots & \dots & \dots & \dots \\ f(0, 35) & f(35, 1) & \dots & 1 - S \end{bmatrix}$$

1.2 Uczenie Środowiska 4

1.2.1 Podejście 1

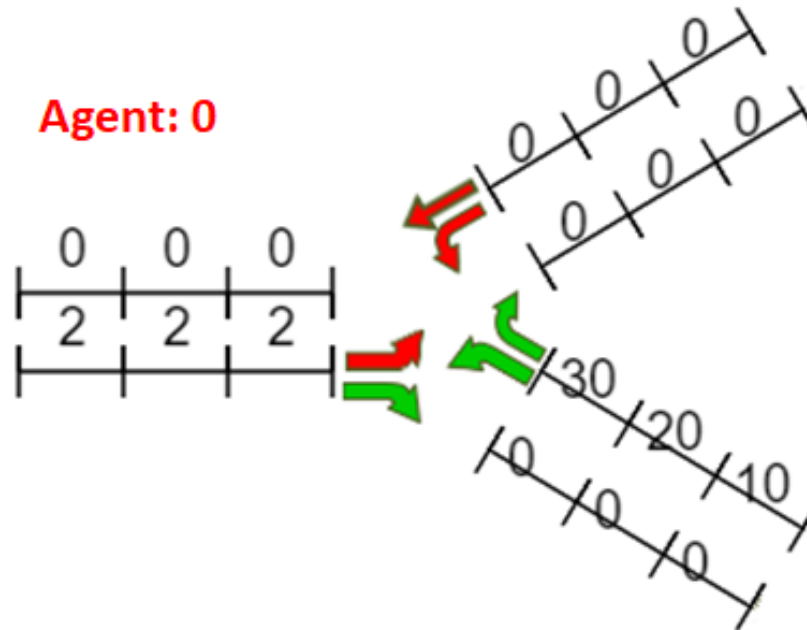
Każdy z trzech agentów jako stan przyjmuje 10 elementowy wektor. 9 elementów to ilości pojazdów na odcinkach będących przed skrzyżowaniem przypisanym do agenta. Wektor

uzupełnia wartość obecnej fazy (0,1,2 lub 'żółte'). Nagrody są przyznawane jako suma pojazdów, które przejechały przez skrzyżowanie w trakcie najbliższych 4 interwałów czasowych. Początkowo przeprowadzana jest symulacja 100 epizodów z czego każdy trwa 90 interwałów czasowych. Ma ona na celu wygenerowania danych do treningu. Do nauki agent zapamiętuje jedynie te stany, których faza to 0, 1 lub 2. Nieistotne w procesie uczenia są stany z fazą 'żółte' gdyż agent ma tylko 1 możliwą decyzję do podjęcia. Do danych zapisywane są wartości stanu oraz nagród które zostały przydzielone dla pary stan-akcja. Następnie trenowana jest sieć neuronowa przyjmująca na wejście stan - 10 elementowy wektor. Sieć na wyjściu zwraca 3-elementowy wektor określający przewidziane nagrody dla akcji podjętej w zadanym stanie. Podsumowując dla wybranego agenta:

- **Stanem** są ilości pojazdów przed skrzyżowaniem oraz aktualna faza świetlna
- **Nagrodą** w chwili t jest suma pojazdów, które przejechały przez skrzyżowanie w trakcie najbliższych 4 interwałów czasowych czyli do momentu $t+4$.
- **Dane** są generowane poprzez przeprowadzenie 100 symulacji (każda ma 90 interwałów czasowych).
- **Sieć neuronowa** na podstawie wygenerowanych danych przewiduje najlepszą akcję dla obecnego stanu
- **Końcowa symulacja** zostaje przeprowadzona wedle przewidzianych przez sieć neuronową najlepszych akcji

Algorytm jest następujący:

1. Utworzeni zostają 3 agenci dla poszczególnych skrzyżowań. Każdy z nich posiada sieć neuronową z 10 elementową warstwą wejściową i 3 elementową warstwą wyjściową.
2. Przeprowadzone zostaje 50 losowych symulacji. Agenci zapamiętują dane z każdej chwili symulacji. Na te dane składa się stan oraz 3 elementowa tablica, która pod indeksem akcji przechowuje przydzieloną nagrodę. Pozostałe wartości tej 3 elementowej tablicy są przewidywane przez sieć neuronową.
3. Następuje trening sieci neuronowej. Parametry sieci neuronowej zostały wybrane w procesie walidacji krzyżowej. Zbiór walidacyjny to 0.2 wygenerowanego zbioru w punkcie 2. Pozostała część przeznaczona jest do treningu. Warunkiem stopu jest zwiększenie się błędu dla zbioru walidacyjnego.
4. Agenci zapominają dane z przeprowadzonych 100 symulacji. Natomiast pozostawiają w pamięci wagi wytrenowanej sieci neuronowej.
5. Opcjonalnie w celu monitorowania postępów nauki przeprowadzona zostaje symulacja. Jej akcje są wybierane w sposób zachłanny. Dla aktualnego stanu zostaje zawsze wybrana akcja o najwyższej przewidzianej nagrodzie przez sieć.
6. Powrót do kroku 2

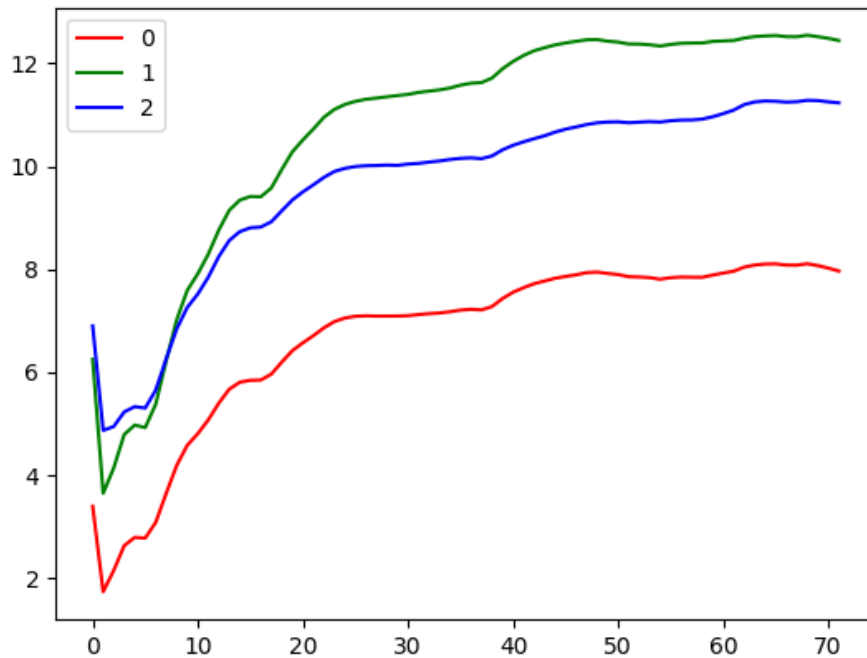


Rysunek 1.3: Monitorowany stan

1.2.2 Monitorowanie decyzji dla wybranego stanu

W tej sekcji zostanie przedstawione jak zmieniały się rekomendowane przez agenta 0 akcje dla pewnego szczególnego stanu. Ilości pojazdów przed skrzyżowaniem to $[2, 2, 2, 30, 20, 10, 0, 0, 0]$. Aktualna faza świetlna to 1. Warto zadać sobie pytanie jaka akcja jest najbardziej opłacalna dla takiego stanu. Z pewnością warto podtrzymać aktualną fazę świetlną - gdyż przed tymi światłami jest najwięcej pojazdów. By osiągnąć ten cel należy wybrać akcję 1. Porządanym zjawiskiem jest, aby agent 0 dla tego stanu przewidywał największe nagrody dla akcji 1. Po każdej sesji uczenia sieci neuronowej zapisane zostały spodziewane nagrody. Ich wykres jest poniżej (każda pełna sesja algorytmu daje podobny wykres).

Nagrody przewidziane dla akcji podjętych podczas monitorowanego stanu

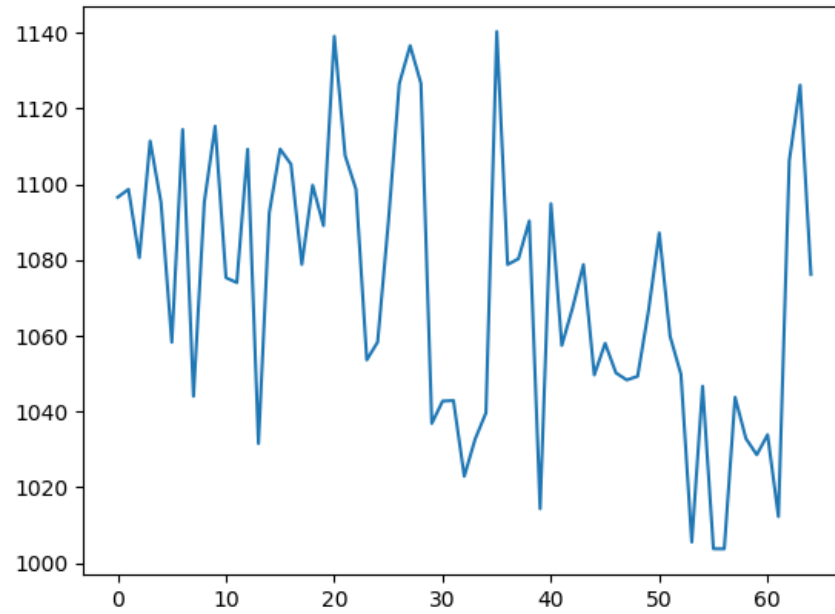


Zgodnie z przypuszczeniami z upływem czasu agent zauważa, że najlepszą akcją jest akcja 1. Jest to jak najbardziej dobry prognostyk. Niewiele gorszy wynik ma akcja 2 - zauważalnie lepszy od akcji 0. Ma to uzasadnienie w tym, że faza świetlna 2 ma zielone światło dla prawoskrętu przed najbardziej zatłoczoną drogą.

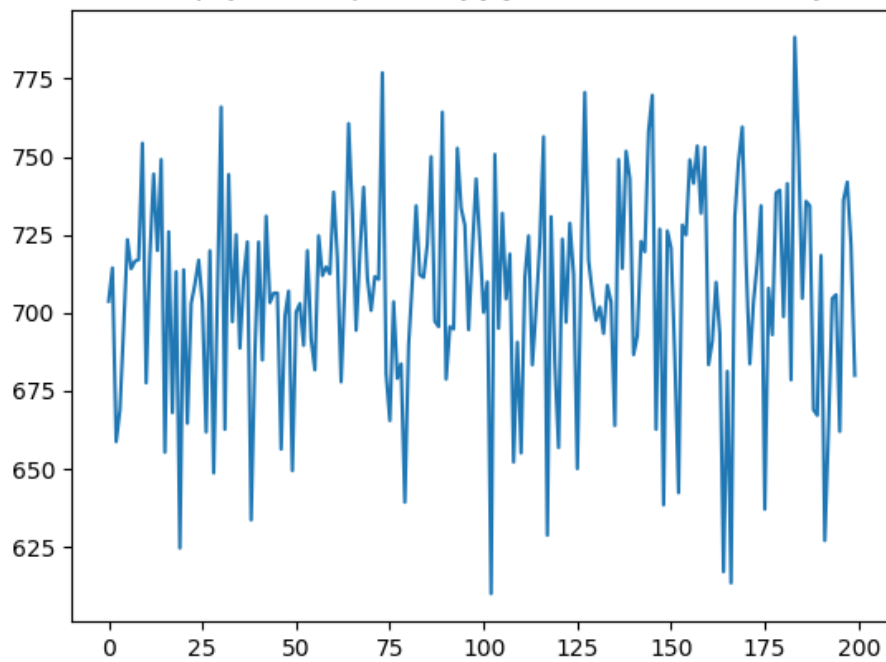
1.3 Analiza wyników uczenia

Wykres ilości pojazdów, które opuściły układ nie wygląda jakby wraz z upływem uczenia wynik się polepszał. Podobnie wygląda analogiczny wykres dla strategii losowych akcji. Należy zwrócić jednak uwagę iż średni wynik dla wyuczonej strategii to 1080. Do tego wyniku nie zbliżają się nawet najszcześniejsze rezultaty dla losowych symulacji.

Ilość pojazdów opuszczających układ - akcje wedle wyuczonej strategii

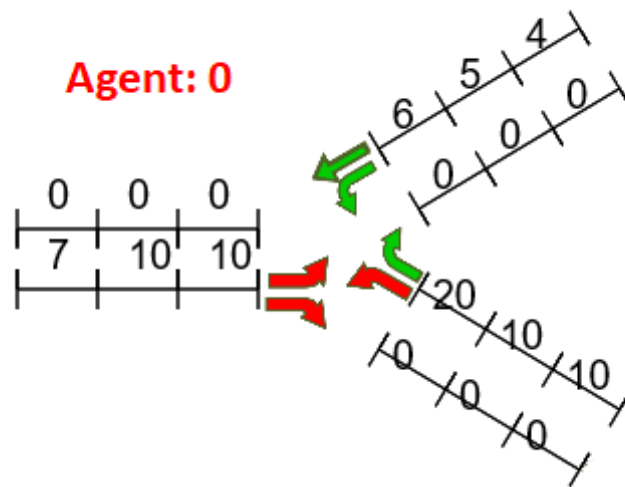


Ilość pojazdów opuszczających układ - losowe akcje



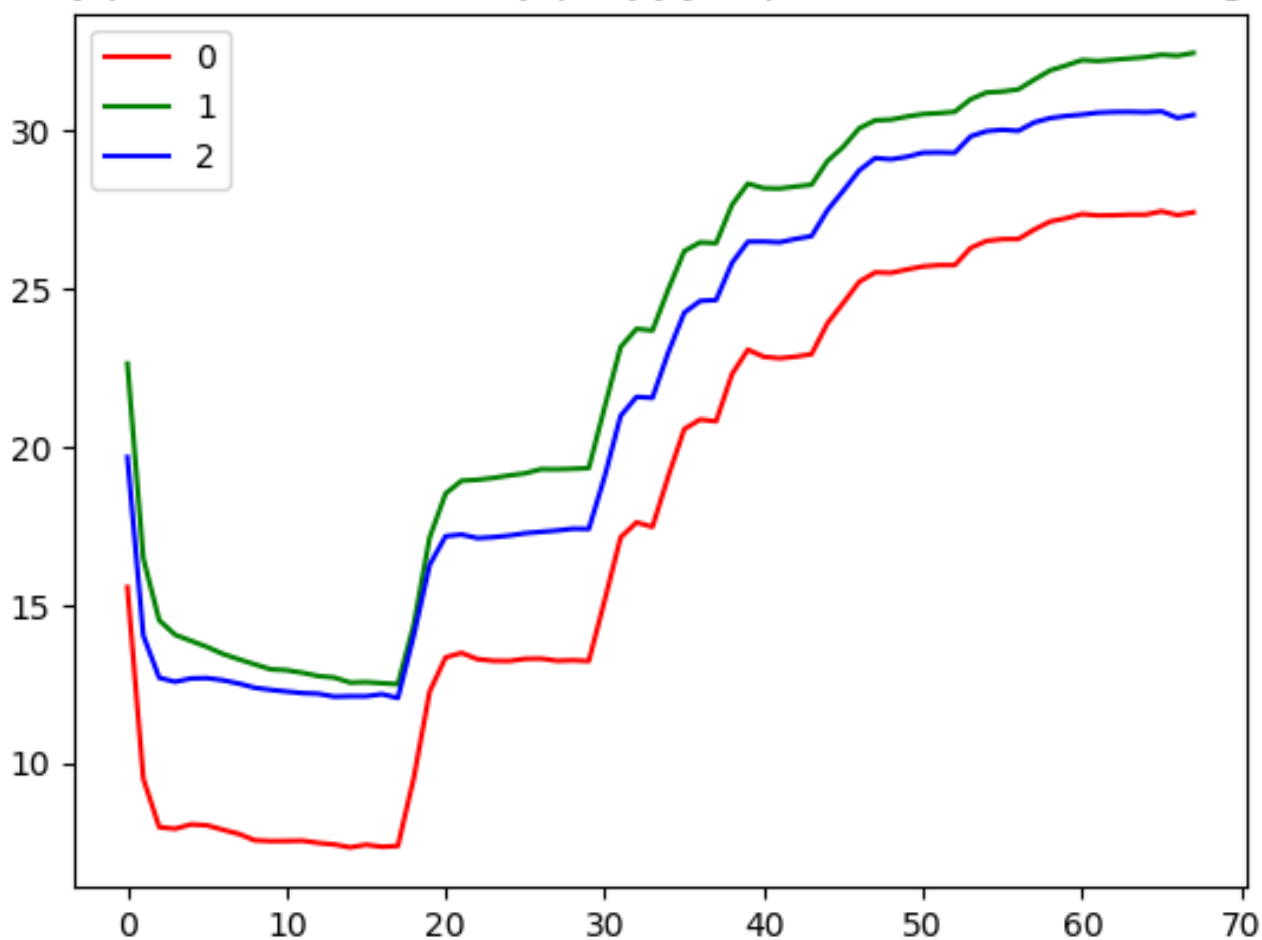
1.4 Monitorowanie kolejnego stanu

Dla poprzednio monitorowanego stanu ukazanego w (1.2.2) sieć bardzo dobrze przewidywała najlepszą akcję. Był to jednak trywialny przykład z oczywistą do przewidzenia akcją. Należy prześledzić wyniki uczenia dla innego stanu. Z pewnością trudniej agentowi będzie podjąć poprawną decyzję dla następującego stanu.

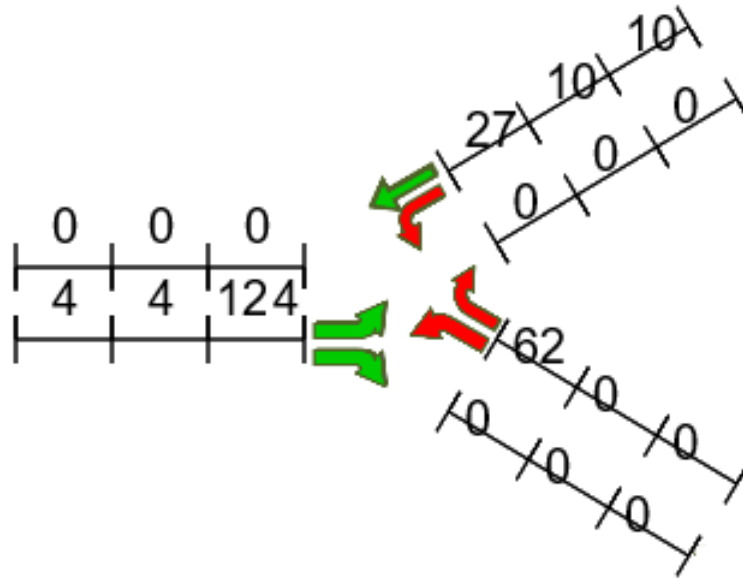


Przewidzenie najlepszej akcji (którą wydaje się być 2) nie sprawia agentowi większego kłopotu. Przeprowadzone zostało 5 pełnych sesji uczenia (każda po 10 minut) i wszystkie wyniki są bardzo podobne. Jedynie w jednej z nich były nieliczne momenty podczas których agent uznawał akcję 2 za najbardziej optymalną. Wykres przewidywanych nagród z tej sesji jest przedstawiony poniżej.

Nagrody przewidziane dla akcji podjętych podczas monitorowanego sta



Ten losowo wybrany stan, który wydaje się być wymagający w kwestii doboru najlepszej decyzji okazuje się nie sprawiać problemów dla agenta. Należy zatem prześledzić symulację, aby odnaleźć stany w których są podejmowane nieoptymalne akcje. Przykładowy stan z którym agent nie radzi sobie jest przedstawiony poniżej.



Dobrym wyborem wydaje się być podtrzymanie fazy 0. Przez conajmniej 3 najbliższe interwały czasowe faza 0 gwarantuje przepływ 30 pojazdów w trakcie pojedynczego interwału. Warto zauważyć, że 30 przejeżdżających pojazdów przez skrzyżowanie to maksymalna ilość na 1 interwał. Pomimo tego argumentu agent zazwyczaj decyduje się na akcję 1. Chociaż faza 1 także wydaje się być dobra - to karencja w postaci żółtego światła na 2 interwały czasowe zdecydowanie przechyla szalę na korzyść akcji 0 i podtrzymania aktualnej fazy.

Wykonane zostaną następujące kroki, które potencjalnie mogą pozwolić agentowi uczyć się lepiej oraz odnajdywać poprawną akcję dla monitorowanego stanu. - Epsilon-greedy learning - początkowo dane będą dalej generowane w pełni losowo. Z czasem uczenia do generowania danych używane będą także akcje dobrane wedle najwyższej oczekiwanej nagrody. - Ciągła walidacja w trakcie uczenia - Wprowadzenie możliwości ustanowienia stanu w środowisku i nauki dla niego. - Zrobić symulacje dla wszystkich 3 akcji z danego stanu no i potem niestety mogłoby się robić w chuj duże drzewko