Факултет по математика и информатика

Катедра „Компютърна информатика”

ДИПЛОМНА РАБОТА

на тема

„Решение на задача за reinforcement обучение на обект с непрекъснати състояния с невробиологичен симулатор NEST“

**Дипломант:** Борислав Стоянов Марков  
**Магистърска програма:** **Изкуствен Интелект**  
**Факултетен номер:** **0MI3400048**

**Научен ръководител**: Проф. Петя Копринкова-Христова,

Институт по Информационни и Комуникационни

Технологии (ИИКТ), БАН

София: 2023г.

# 1. Съдържание

[1. Съдържание 2](#_Toc132886765)

[2. Увод 3](#_Toc132886766)

[Дипломната работа ще включва кратък обзор в областта на Spike Timing Neural Networks, описание на теоретичната постановка, код на Python с използване на библиотеката NEST Simulator и анализ на резултатите. В процеса на разработка на магистърската теза ще се изпробват различни параметри на биологично подобните неврони и решението ще бъде илюстрирано с подходящи визуализации и графики, съпътстващи обучителния процес. 3](#_Toc132886767)

[3. Средата „Cart Pole“[4] в Gym 3](#_Toc132886768)

[4. Въведение в невробиологичните симулации 5](#_Toc132886769)

[4.1 Основи на невробиологията 5](#_Toc132886770)

[4.2 Математически апарат за моделиране на невроните 6](#_Toc132886771)

[4.3 Невробиологичен симулатор NEST 8](#_Toc132886772)

[5. Подход за решаване на задачата 10](#_Toc132886773)

[5.1 Темпорални методи (TD) 14](#_Toc132886774)

[5.2 Контрол на моториката и реинфорсмънт обучението 15](#_Toc132886775)

[5.3 Допаминът и ролята му в TD реинфорсмънт методите за обучение 17](#_Toc132886776)

[5.4 Актьор-критика архитектура за обучение на моторния апарат 21](#_Toc132886777)

[5.5 Победителят печели всичко 22](#_Toc132886778)

[5.6 Обучение с импулсно-времево зависима пластичност(STDP) 24](#_Toc132886779)

[5.7 Обща постановка за решаване на задачата 27](#_Toc132886780)

[6. Реализация на проекта 35](#_Toc132886781)

[6.2 Експериментална част 36](#_Toc132886782)

[6.2.1 Обучение на CartPole 38](#_Toc132886783)

[6.2.2 Обучение при Frozen Lake 3x3 с хлъзгане 42](#_Toc132886784)

[6.2.3 Обучение при Frozen Lake 4x4 без хлъзгане 43](#_Toc132886785)

[6.3 Параметри на постановката и анализ на резултатите 43](#_Toc132886786)

[6. Идеи за бъдещо развитие и подобрения 45](#_Toc132886787)

[7. Източници и използвана литература 46](#_Toc132886788)

[Приложения 47](#_Toc132886789)

[1. Сорс код (Source code) 47](#_Toc132886790)

# 2. Увод

Невробиологията все повече набира скорост в света на изкуствения интелект. Има все повече изследвания на функционирането на нервни клетки, които са довели до създаването на биологично обоснованите spike timing модели на невроните, както и много знания за структурната организация и функционирането на мозъка на бозайниците при вземане на решения. Доказано е, че много от решенията се вземат по метода на поощрението и наказанието (Reinforcement Learning).

Целта на дипломната работа е да се разработи модел на биологично обоснована (spike timing) невронна мрежа посредством библиотеката NEST Simulator, която е в състояние да решава оптимизационната задача за вземане на решения за обект с непрекъснати състояния от пакета Gym.

Дипломната работа ще включва кратък обзор в областта на Spike Timing Neural Networks, описание на теоретичната постановка, код на Python с използване на библиотеката NEST Simulator и анализ на резултатите. В процеса на разработка на магистърската теза ще се изпробват различни параметри на биологично подобните неврони и решението ще бъде илюстрирано с подходящи визуализации и графики, съпътстващи обучителния процес.

# 3. Средата „Cart Pole“[4] в Gym

Средата Cart Pole е част от подпакета за класически контрол в Gym. Това са много опростени игри със случайно начално състояние и непрекъснат вектор на състоянието. Контролът на агента е дискретен с малък брой действия, обикновено две. Този пакет е разработен за обучителни цели по метода на поощрението и наказанието (Reinforcement Learning). Целта е да се сравняват различни решения при една и съща среда. На следващата фигура е показана примерна визуализация на тази среда. Имаме количка, която може да се движи наляво и надясно и закрепен прът, с две степени на свобода.

|  |
| --- |
|  |

Фигура 3.1 – Примерна визуализация на Cart Pole.

При движение налявно и надясно прътът се балансира. При изпускане на баланса епизодът приключва. Тъй като целта е да се запази баланс за по-дълго, поощрение се дава на всяка стъпка в която пръта не се счита за паднал.

Основна информация за средата са дадени в следващата таблица.

|  |  |
| --- | --- |
| Пространство на действията | Discrete(4) |
| Вектор на наблюдението | (4,) |
| Максимален вектор на наблюдението | [4.8,inf,0.42,inf] |
| Минимален вектор на наблюдението | [-4.8,-inf,-0.42,-inf] |
| Импортиране в Питон | gym.make("CartPole-v1") |

Таблица 3.1. Общи параметри за средата

Агентът може да се придвижва в две посоки, всяка с код от 0 и 1 включително, а именно: наляво-0, надясно-1.

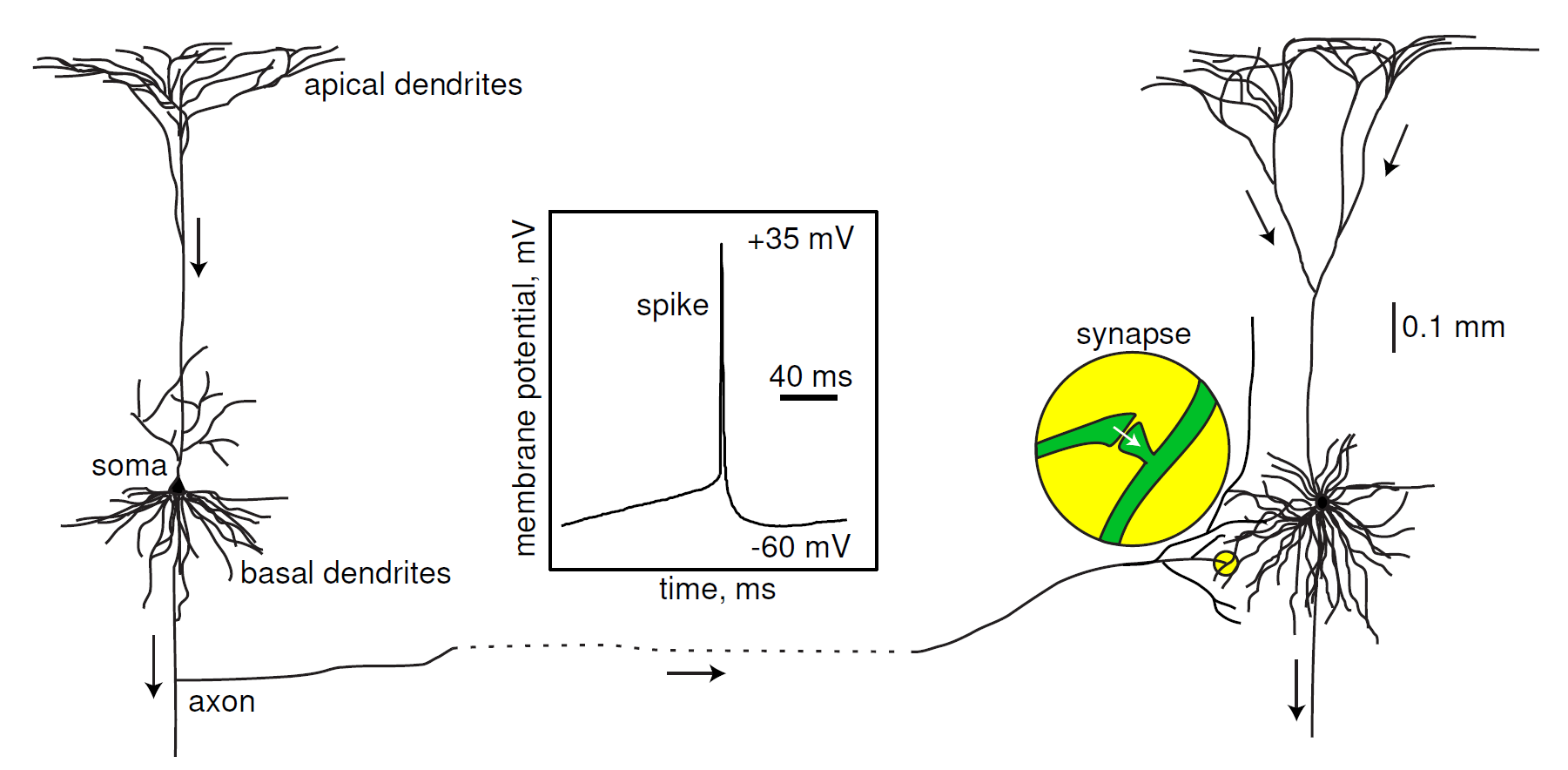
# 4. Въведение в невробиологичните симулации

Невробиологията е дисциплина, която може да засяга един или друг аспект свързан с работата на нервната система при живите организми. За нашите цели се фокусираме повече върху математическия апарат, отколкото върху биологичната основа. Най-забележимият аспект на нервната система при човека и при други живи организми е начинът на вземане на решения и способността да се обучават с поощрение и наказание. Връзката между реинфорсмънт обучението (Reinforcement learning) и новробиологичните науки се крие в химично имплементираната награда – допаминът ([1] глава 15). Допаминът пренася темпоралната грешка (TD error) до структурите в мозъка, които са отговорни за вземане на решение.

## 4.1 Основи на невробиологията

Невроните са основните компоненти на нервната система. Това са клетки специализирани в пренасяне и обработка на информация посредством електрохимически и химически сигнали [1]. Невроните се състоят от тяло, дендрити и един аксон. Дендритите са разклонения от тялото, чрез които клетката се свързва с аксони от други неврони или са сензори, в случай на сензорни неврони.

Невронът събира импулси от много входове и когато сумарно тези входове преминат някаква граница, невронът изстрелва потенциал (action potential) или тъй наречения импулс (spike). Това е и фундаменталният начин на комуникация между невроните ([2], Ch.2). Изходният сигнал на неврона са електрически импулси, пътуващи по аксона, наречени спайкове (spikes). След произвеждане на импулс, напрежението на мембраната се връща до равновесен потенциал на мембраната.



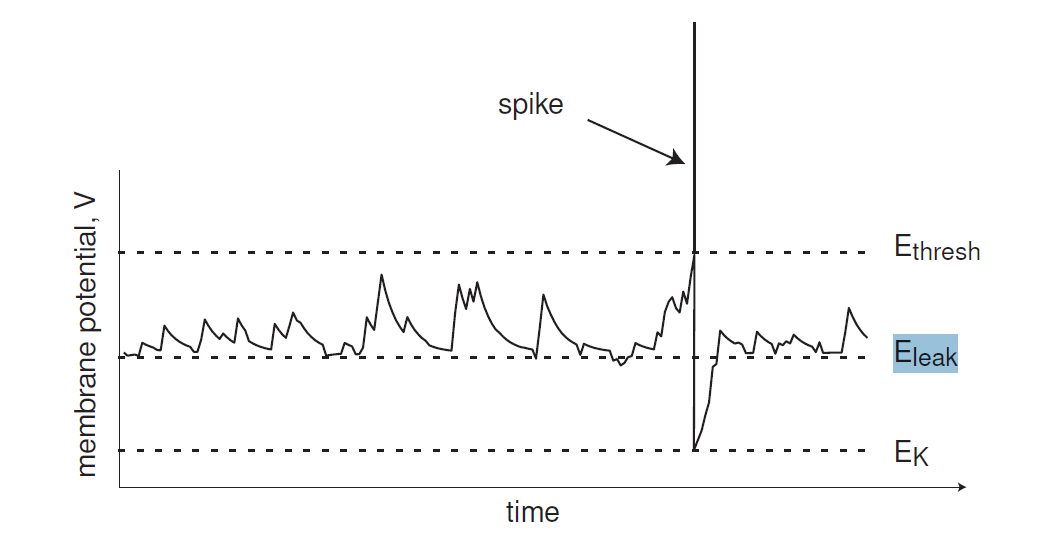
Фиг.4.1.1. Рисунка на два свързани неврони и ин-витро записан спайк  
Адаптирано от [3], Глава 1.1

## 4.2 Математически апарат за моделиране на невроните

Към момента невроните се моделират по много различни начини и има описани десетки видове диференциални уравнения на различни неврони. Един от първите математически формализми на неврони е на Ходжкин и Хъксли, описан през петдесетте години на миналия век. Уравнението се оказва доста сложно за решаване на практически задачи и затова по-късно са предложени опростени модели. Integrate-and-fire са фамилия модели от които най-популярният към момента е leaky integrate-and-fire неврон. Това е идеализация на неврон с утечки (по закона на Ом) , който е суматор на токове.

|  |  |
| --- | --- |
|  | (4.2.1) |

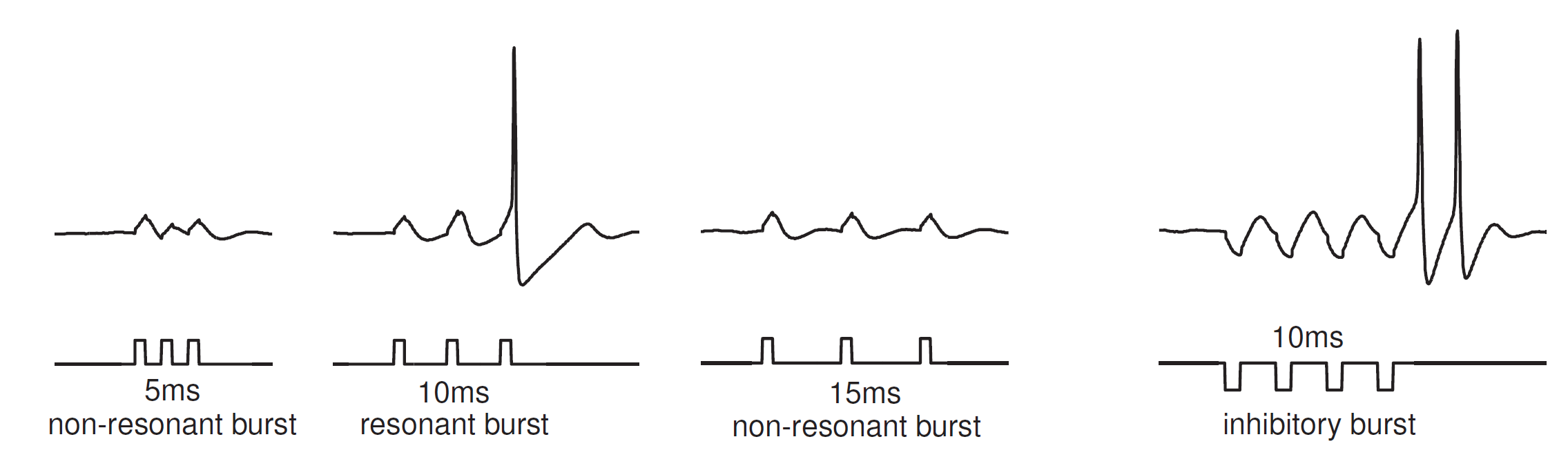
Тук формализма е взет от [3], C е капацитетът на мембраната на неврона, V е мембранният потенциал, gleak e проводимостта на клетъчната мембрана, Eleak e равновесният потенциал на мембраната. На следващата фигура се онагледява действието на неврона.



Фиг. 4.2.1 Потенциална диаграма на суматорен неврон с утечка (Leaky integrate-and-fire neuron). Адаптация от [3].

На фигурата се вижда, че при напрежение надвишаващо Ethresh се произвежда токов импулс (spike) и напрежението се връща до стойност Ek. При липса на входни токове или шум се забелязва как напрежението намалява експоненциално, стремейки се към Eleak.

Тук си струва да споменем и един феномен, както и трудностите свързани с предизвикване или непредизвикване на импулси от невроните. Например Изикевич[3] показва диаграма са еднакви по амплитуда импулси, но с различен период на излъчване и резултатите са изненадващи.



Фиг. 4.2.2 Резонансен отговор на мезенцефалния V неврон на мозъчния ствол на плъх към импулси инжектиран ток с период от 10 ms. Адаптация от [3].

Неврона се стимулира с три токови импулса. Когато честотата на стимулация е висока (5 милисекунден период), имитирайки силен вход, неврона дори не предизвиква спайк. Въпреки това, стимулация с по-ниска честота (10 милисекунден период), който резонира с честотата на подпраговите трептения на неврона предизвиква пикова реакция, независимо дали стимулацията е възбуждаща или инхибираща. Стимулация с още по-ниска честота (период от 15 ms) не може отново да предизвика пикова реакция. По този начин невронът е чувствителен само към входовете с резонансна честота. Също импулсите, приложени към кортикален пирамидален неврон, предизвикват отговор само в първия случай (малък период), но не и в останалите случаи.

## 4.3 Невробиологичен симулатор NEST

NEST е симулатор за невронни мрежи основани на спайкове (spiking neural network models или SNN) и може да послужи за: модели за обработка на информация, модели на мрежова динамика, модели на обучение и синаптична пластичност [5]. Симулаторът представлява библиотека на Python и може да се вгради в по-голямо приложение. NEST предоставя над 50 невронни модела, много от които са публикувани. Изборът варира от прости неврони с интегриране и запалване със синапси, базирани на ток или проводимост, през моделите на Изикевич или AdEx до моделите на Ходжкин-Хъксли. Структурата на програмата е например:

|  |
| --- |
| 1. Създаване на невронните групи и други устройства (напр. генератори на шум или входен ток) 2. Свързване на групите неврони в желаната структура със съответните тегла на връзките 3. Поставяне на виртуални измервателни уреди и свързване с интересуващите ни групи неврони (напр. волтметри, детектори на импулси); 4. Симулиране с метода nest.Simulate(t), който проиграва една евентуална симулация за време t милисекунди. |

Таблица 4.3.1 Структура на програма използваща NEST simulator

Невронната мрежа в NEST се състои от два основни типа елементи: възли и връзки. Възлите са неврони, устройства или подмрежи. Устройствата се използват за стимулиране на неврони или за запис от тях. Възлите могат да бъдат подредени с пространствена структура за изграждане на мрежи, разпределени в пространството.

Новите възли се създават с командата Create(), която приема като аргументи името на модела на желания тип възел и по желание броя на възлите, които трябва да бъдат създадени, и инициализиращите параметри. Функцията връща NodeCollection от манипулатори към новите възли, които можете да присвоите на променлива за по-късна употреба. NodeCollection е компактно представяне на манипулаторите на възлите, които са цели числа, наречени идентификатори. Много функции на NEST очакват или връщат NodeCollection. По този начин е лесно да се прилагат функции към големи набори от възли с едно извикване на функция.

Можем да разгледаме примерен код на програма, състояща се от един неврон, на който подаваме постоянен ток от 376 [pA] и замерваме напрежението на мембраната, както и генерираните импулси.

|  |
| --- |
| import nest  import matplotlib.pyplot as plt  import nest  nest.ResetKernel()  neuron = nest.Create("iaf\_psc\_alpha", {"I\_e":376.0})  voltmeter = nest.Create("voltmeter")  spikerecorder = nest.Create("spike\_recorder")  nest.Connect(voltmeter, neuron)  nest.Connect(neuron, spikerecorder)  nest.Simulate(1000.0)  plt.rcParams["figure.figsize"] = (5,2)  nest.voltage\_trace.from\_device(voltmeter)  plt.show()  nest.raster\_plot.from\_device(spikerecorder, hist=False, title="spikerecorder")  plt.show() |

Таблица 4.3.2 Примерна програма използваща NEST simulator с един неврон и симулация от 1000 милисекунди

Редът, в който са зададени аргументите на Connect(), отразява потока от събития: ако невронът генерира спайкове, той изпраща събитие до записващото устройство за спайкове. Обратно, волтметърът периодично изпраща заявки до неврона, за да поиска неговия мембранен потенциал в този момент. Това може да се разглежда като перфектен електрод, забит в неврона.

Изходът от програмата са две графики изобразяващи скоковете в напрежението и генерираните импулси.

|  |  |
| --- | --- |
|  |  |

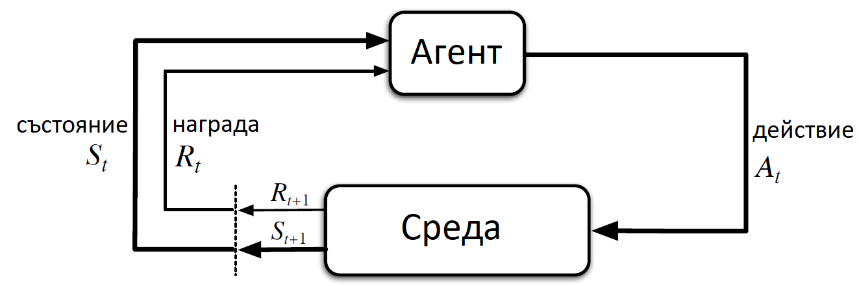
Фигура 4.3.1 Напрежение на мембраната и импулси от неврона

При по-голям ток, импулсите ще следват по-бързо, при по-малък ток на неврона импулсите ще намалеят или ще изчезнат.

# Подход за решаване на задачата

Обучението с поощрение и наказание(Reinforcement learning) е да научите какво да правите как да съпоставяте ситуациите към действията, така че да се максимизира цифров сигнал за получаване на награда. На обучаемия агент не се казва какви действия трябва да направи, но вместо това трябва да се открие кои действия носят най-голяма награда, като се изпробват такива евентуални действия. В най-интересните и трудни случаи действията могат да засегнат не само непосредствената награда, но също и следващата ситуация и чрез това всички последващи награди. Тези двe характеристики - търсене по принцип проба-грешка и отложено възнаграждение - са двете най-важни отличителни черти на реинфорсмънт обучението. То е различно от обучението с учител, едно от най-изучаваните и най-актуалните изследвания в областта на машинното самообучение. Обучението с учител е такова, че се обучава от зададено, предварително маркирано обучаемо множество данни, предоставено от външен източник. Реинфорсмънт обучението също е различно и от машинно самообучение без учител, което обикновено е свързано с намиране на структура, скрита в колекции от немаркирани данни. Термините обучение с учител и обучение без учителне могат да класифицират изчерпателно парадигмите на машинното самообучение. Някои могат са си мислят, че реинфорсмънт обучението е обучение без учител, тъй като не разчита на примери за правилно поведение, но реинфорсмънт обучението се опитва да увеличи максимално сигнала за награда, вместо да се опитва да намери скрита структура.

CartPole като задача за Reinforcement Learning е формулирана като агент и среда, взаимодействащи си чрез марковски процес. След всяко действие на агентът At средата ни отговаря с ново състояние St+1 и се дава съответната награда Rt+1.

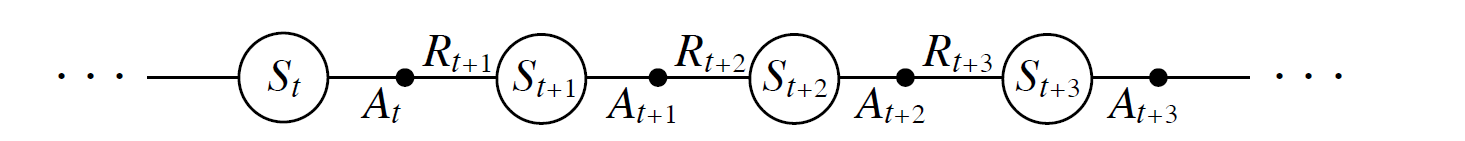


Фиг. 5.1. Взаимодействие Агент-Среда при марковски процес на решението, адаптирано от [1]Глава 3.1

Целта на агента е да максимизира кумулативната награда в дългосрочен план. Формално поредицата от получаваните наградите след стъпка на времето t означаваме с . Генерално се иска да се максимизира очакваната награда, като награда за някаква поредица от действия, означаваме с Gt. В най-опростеният случай кумулативната наградата е сумата от награди за всяка стъпка:

Където T е крайната стъпка. Този подход е в случай, когато имаме край на някаква поредица от взаимодействия на Агент-Среда, която поредица ще наричаме **епизод**. Епизодът приключва с успех или неуспех в зависимост от играта. Епизодите са независими един от друг и всеки епизод стартира независимо от това дали предшестващият го е приключил с успех или неуспех.

На следващата фигура е представено схематично как изглежда един епизод на взаимодействие между агент и среда разгънат по време, нарастващо от ляво надясно. Кръговете са състоянията, определяни от средата и бидейки в едно състояние, агентът трябва да определи следващото действие A (action).



Фиг.5.1.2 Примерен епизод, вж. [1] глава 6.4

Времето за приключване на епизод, T, е случайна променлива и варира от епизод до епизод. За случаите при постоянно вървящи процеси и формулировката за Gt става неподходяща ,тъй като функцията, която се опитваме да максимизираме става безкрайност. Поради това в теорията на реинфорсмънт обучението се въвежда фактор:

Където γ е параметър, и се нарича отстъпка. Смисълът е, че награда, получена k стъпки в бъдещето ще си струва толкова колкото и награда получена веднага умножена с γk-1.

Всички алгоритми за реинфорсмънт обучение включват част за апроксимация на функции-стойност(value functions) – функции на състоянието или на двойка състояние-действие, която ни казва колко добре е агентът да се намира в дадено състояние (или до колко е добро дадено действие, когато агентът е в определено състояние). Терминът „до колко е добро“ се свежда до очакваната награда. Тези функции-стойност са дефинирани от определено поведение на агента, като определено поведението ще наричаме политика(policy). По формално политиката е съответствие между състояние и вероятност да се избере дадено действие от агента. Ако агентът следва политика π във време t , тогава π(a|s) е вероятността At=a при St=s. Знакът „|“ във функцията π(a|s) определя вероятностно разпределение върху за всяко . Реинфорсмънт обучението определя как да се измени политиката на агента в зависимост от резултата от натрупания опит.

Функцията-стойност за дадено състояние s при дадена политика π отбелязваме като и е очакваната награда, когато стартираме от състояние s и следваме политиката π. За марковски процес (MDP) дефинираме формално като:

|  |  |
| --- | --- |
|  | (5.1.1) |

Където означава очакваната стойност от случайна величина, при агент който следва политика π и t е някаква стъпка. Стойността на терминалното състояние е винаги със стойност нула. Наричаме vπ функция стойност (state-value function) за политика π.

Аналогично се дефинира стойността, при избиране на действие a от състояние s, следвайки политика π ,отбелязвана като qπ (s,a):

|  |  |
| --- | --- |
|  | (5.1.1) |

Наричаме qπ (s,a) функция действие-стойност (action-value function) за политика π. Функциите vπ и qπ биват апроксимирани на базата на опита получен от агента.

Решаването на задача с реинфорсмънт обучение се свежда до намиране на политика, която постига висока награда в дългосрочен план. За крайни MDP можем точно да дефинираме оптимална политика по следния начин. Стойностните функции определят частично подреждане върху политиките. Една политиката е дефинирана като по-добра от или равна на друга политика, ако нейната очаквана възвръщаемост е по-голяма от или равно на възвръщаемостта от тази на втората за всички състояния. По-формално π ≥ π’ ако и само ако за всички . Винаги има поне една политика, която е по-добра или равна на всички останали политики. Тя е оптимална политика. Въпреки, че може да има повече от една оптимална политика означаваме всички оптимални политики с . Всички те споделят една и съща функция-стойност v\*,дефинирана като

за всички .

Оптималните политики също споделят една и съща действие-стойност функция, отбелязвана като q\* , и дефинирана като:

за всички и . Така можем да изразим q\* чрез v\* както следва:

Състоянието на средата Sза задачата CartPole е вектор от състояния на отделни компоненти, които са реални числа. Така множеството е безкрайно. В случая бихме могли да ползваме таблични методи и да квантуваме множеството като приведем такова едно множество в изброимо множество, чрез подходяща дискретизация, но за текущата работа подходът ще е различен.

За да минимизираме някоя от функциите-стойност трябва да направим апроксимация чрез параметри, които можем да обучаваме. Можем да представим функцията v(s) като функция с вектор от настройваеми параметри и ще записваме за апроксимирана стойност, тръгвайки от състояние s и даден вектор с параметри **w**. Функцията може да бъде функция изчислена с многослойна невронна мрежа от тип SNN, и ще бъде разгледана такава реализация в следващите глави.

## 5.1 Темпорални методи (TD)

Методите с темпорална грешка (TD)[1] и методите Монте Карло използват опит за решаване на проблема с апроксимацията на функцията за стойност. Чрез дадени известен опит следвайки политика π, и двата метода актуализират оценката си за V от за нетерминалните състояния St, възникващи в конкретният опит. Общо казано Монте Карло методи изчакват, докато оценката на дадено състояние след посещението стане известно, след което използвайте това връщане като цел за V(St). Прост метод Монте Карло за всяко посещение, подходящ за нестационарни среди е:

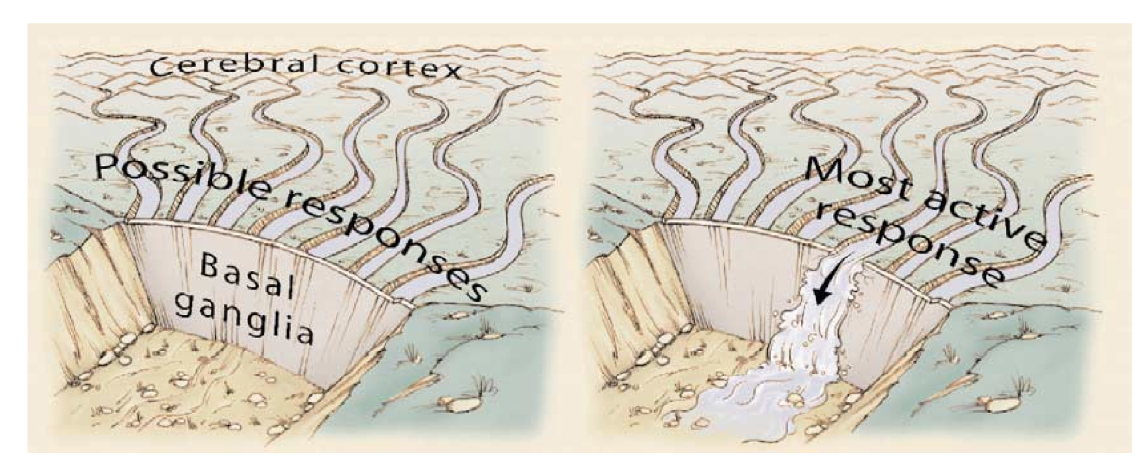
където Gt e очакваната стойност след стъпка t, α е някакъв коефициент на обучение. Докато Монте-Карло методите трябва да изчакат до края на епизода за да определят увеличението на V(St), то TD методите трябва да изчакат само до следващата стъпка. Във време t+1 вече може да се направи актуализация на на апроксимираната функция, знаейки наградата Rt+1 и имайки апроксимация на V(St+1). Най-простата форма на TD може да направи промяната

веднага след преминаване в състояние St+1 и получена награда Rt+1. Такъв TD метод се нарича TD(0) или едностъпков TD, защото е частен случай на TD(λ) и n-стъпковите методи TD. Предстои ни да разгледаме как ще се представи апрокцимацията на функциите-стойност за отделните състояния в областта на невронауката и как ще бъдат извършвани оптимизацията за постигане на максимална награда в дългосрочен план, но за целта трябва да разгледаме някои положения от моторният контрол и връзката с реинфорсмънт обучението.

## 5.2 Контрол на моториката и реинфорсмънт обучението

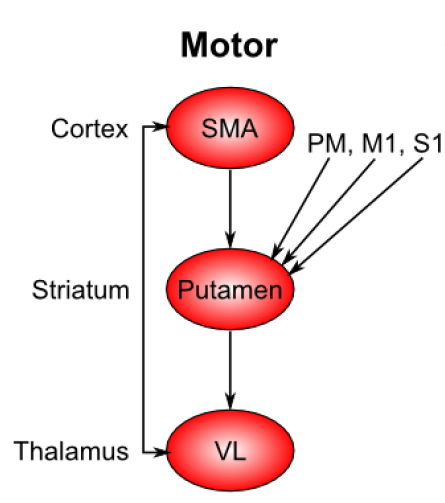
Основите на познанието са изградени върху сетивно-моторната верига – обработка на сензорни входове за определяне кое двигателно действие да извършите след това. Това е най-основната функция на всяка нервна система. Човешкият мозъкът има огромен брой такива вериги, обхващащи еволюционната времева скала от най-примитивните рефлекси в периферната нервна система, до най-абстрактните и неразгадаеми планове, като например решението за прилагане към и посещаване на висше училище, което вероятно включва най-високите нива на обработка в префронталния кора (PFC).

На подкорно ниво малкият мозък(cerebellum) и базалните ганглии (basal ganglia) са двете основни контролни области, всяка от които има специално адаптирани механизми за обучение, които се различават от общите кортикални механизми за обучение. Базалните ганглии са специализирани за учене от сигнали за награда/наказание, в сравнение с очакванията за награда/наказание и след това това наученото оформя избора на действие, който организмът ще направи при различни обстоятелства. Именно това е и реинфорсмънт обучение. Малкият мозък(cerebellum) е специализиран да се учи от грешки, по-специално грешки от сензорните резултати, свързани с двигателните действия, по отношение на очакванията за тези сензорни резултати свързани с тези двигателни действия. По този начин малкият мозък може да усъвършенства изпълнението на даден двигателен план, за да го направите по-точен, ефективен и добре координиран.



Фиг. 5.2.1 Ролята на базалните ганглии в избора на действие. Cerebral cortex - мозъчна кора, Basal ganglia – базални ганглии, Possible responses – възможни отговори, Most active response – най-активен отговор. Много възможни реакции се считат възможни и базалните ганглии избират тази реакция с най-голяма активност да бъде изпълнена. Фигурата е репродуцирана от Gazzaniga et al (2002)(виж [2] фиг. 7.1 ).

Базалните ганглии изпълняват своята функция за избор на действие в широк диапазон от фронтални кортикални области, благодарение на последователност от паралелни вериги на свързаност. На следващата фигура е дадена схема за избора на действие при моторните действия.



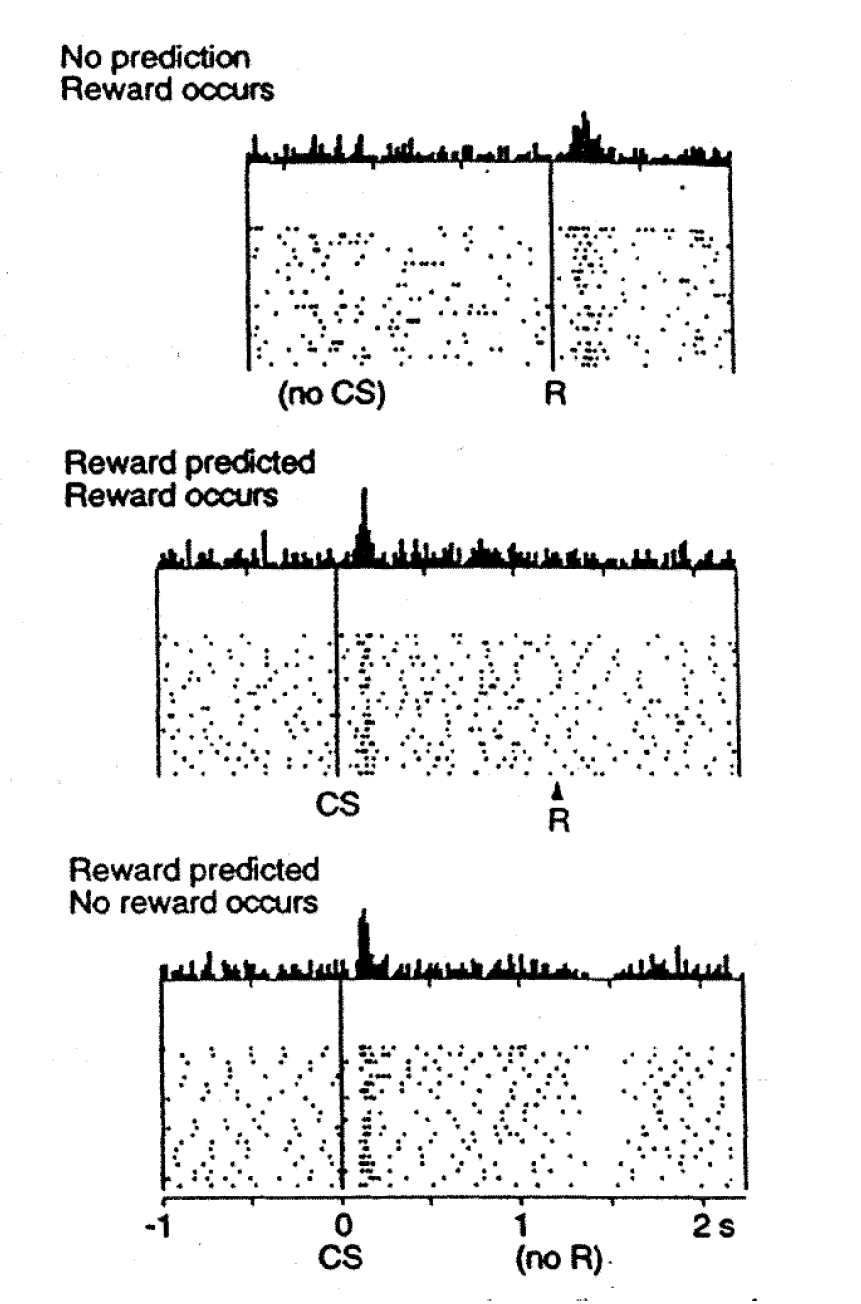
Фиг.5.2.2 Двигателна верига: SMA = допълнителна двигателна зона – свързаният стриатум (putamen) също получава от премоторна мозъчна кора (PM) и първична моторна (M1) и соматосензорна (S1) области – всичко необходимо за правилно контекстуализиране двигателни действия. Адаптирано от [2], фиг.7.2

Тук няма да разглеждаме сложните свързаности в базалните ганглии и как функционират, но ще се спомене това, което е важно за решаване на текущата задача. Базалните ганглии се учат да избират действията, които носят награда.

Разделението на работата между фронталния кортекс и базалните ганглии е такова, че фронталният кортекс комбинира много различни възможни действия, с обогатени модели на свързаност от други кортикални области, осигуряващи обобщения на високо ниво на текущата среда, които след това активират набор от различни възможни действия, и след това базалните ганглии избират най-доброто (най-вероятно да бъде наградено) от тези действия, за да го изпълнят действително. В по-широк смисъл, фронталният кортекс е размит творчески тип, с милион идеи, но не способност да се фокусира върху реалния свят и му е трудно да стесни нещата до точката на действителност и да направи каквото и да е: нещо като мечтател. Междувременно базалните ганглии са истински тип поемане на отговорност и може да вземе трудните решения и да свърши нещата[2]. Вероятно това е причината там да са две отделни системи (фронтален кортекс и базални ганглии), които въпреки това работят много тясно заедно, за решаване на цялостния проблем за избор на действие.

## 5.3 Допаминът и ролята му в TD реинфорсмънт методите за обучение

Нека разгледаме един от важните експерименти, доказващи връзката на определени звена в мозъка и реинфорсмънт обучението. На следващата фигура е представен експеримент с условен стимул и награда.



Фиг.5.2.1 Характерни модели на неврално задействане на допаминергичните неврони във вентралната тегментална област (VTA) и substantia nigra pars compacta (SNc), в проста задача за кондициониране (Schultz et al, 1997).

Преди кондициониране, когато бъде получена награда, допаминовите неврони предизвикват изблик на активност (горен панел – хистограмата отгоре показва сумата на невронни спайкове в текущия момент). След като животното се научи да свързва условен стимул (CS) (напр. тон) с наградата, допаминовите невроните се изместват към началото на CS, а не към самата награда. Ако награда бъде задържана след CS, има спад или пауза в освобождаването на допамин, което показва, че е имало някаква прогноза за наградата и когато тя не успее да бъде постигната, има отрицателна грешка в прогнозата. Този общ модел на възбуждане на невронни групи при различни условия е в голяма степен съвместим с модели за реинфорсмънт обучение, базирани на грешка при прогнозиране на възнаграждението. Възпроизведено от Schultz et al, 1997.

От изчислителна гледна точка най-простият модел на грешка при прогнозиране на възнаграждението е кондиционирането на Rescorla-Wagner модел (Rescorla и Wagner 1972) и е просто разликата между действителната награда и очакваната награда:

където δ („делта“) е грешката при прогнозиране на наградата, r е действително получената сума на наградата и е очакваното възнаграждение, което се изчислява като претеглена сума върху входните стимули x с тегла w. Теглата се адаптират, за да се опитат да предвидят точно действителните стойности на възнаграждението и всъщност тази делта стойност уточнява посоката, в която трябва да се променят теглата:

Това е идентично с правилото за делта обучение, включително претеглената зависимост от активността на стимула x – иска се промяна само на теглата за стимули, които действително присъстват (т.е. ненулеви x).

Когато предвиждането на наградата е правилно, тогава действителната стойност на наградата се анулира от прогнозата, както е показано във втория панел. Това правило предсказва точно и другите случаи, показани на фигурата (грешки при прогнозиране на положителни и отрицателни награди). Това, което моделът на Rescorla-Wagner не успява да улови, е изстрелването на допамин до началото на условния стимул(CS) във втори панел на фигурата. Малко по-сложният модел, представен в 5.1 известен като времеви разлики (TD) улавя това започване на стимула (CS), като въвежда време в уравнението (както и името предполага). В сравнение с Rescorla-Wagner, TD просто добавя един допълнителен термин към делта уравнението, представляващо бъдещите стойности на възнагражденията, които може да дойдат по-късно във времето:

където f представлява бъдещите награди и сега очакваната награда трябва да се опита да предвиди както текущата награда r, така и тази бъдеща награда f. В проста задача за кондициониране, където CS надеждно прогнозира последваща награда, началото на CS води до увеличаване на тази f стойност, защото след като CS пристигне, има голяма вероятност за награда в близко бъдеще. Освен това самото f не е предвидимо, тъй като началото на условния стимул (CS) не е предсказан от по-ранно събитие (и ако беше, тогава тази по-ранна реплика би била истинския стимул и щеше да стимулира допаминовия синтез). Следователно, очакването на r-шапка не може да анулира стойността f, и настъпва допаминов изблик. Въпреки че тази f стойност обяснява CS-началото на допаминово изстрелване, тя повдига въпроса как може системата да знае какви награди предстоят в бъдеще? Като всичко, свързано с бъдещето, ти по същество просто трябва да се познае, като се използва миналото като отправна точка по възможно най-добрия начин. TD прави това, като се опитва за налагане на последователност в оценките на възнаграждението във времето. На практика оценката в момент t се използва за обучение на оценка във време t-1 и т.н., за да се поддържа всичко възможно най-последователно във времето и в съответствие с действителните награди, които се получават във времето.

Всичко това може да бъде получено по много задоволителен начин чрез указване на нещо, известно като стойностна функция, за която бе споменато по-рано в (5.1.1). Тук ще въведем обаче друго по-удобно означение V(t), замествайки състоянието като параметър с параметърът за време t. Отново V(t) е сбор от всички настоящи и бъдещи награди, като бъдещите награди са намалени с „гама“ фактор, който дава интуитивната представа, че наградите по-нататък в бъдещето струват по-малко от тези, които ще настъпят по-рано.

Може да се премахне безкрайността и да се замени с рекурсивната зависимост:

И понеже не знаем нищо със сигурно, заместваме съответната функция със апроксимация, означена със „шапка“ отгоре:

Така че това уравнение ни казва каква трябва да бъде нашата оценка в настоящия момент t, по отношение на бъдещата оценка в момент t+1. След това изваждаме V-шапка от двете страни, което ни дава израз, който е различен начин за изразяване на горното равенство – че разликата между тези членове трябва да бъде равна на нула:

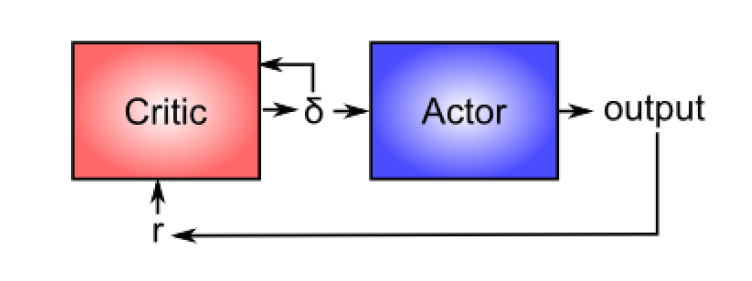
Това е математически означава, че темпоралните методи(TD) се опитват да поддържат оценките последователни във времето по такъв начин, че техните разлики трябва да са нула. Но тъй като апроксимираме нашите оценки с V-шапка, тази разлика няма да бъде нула, а всъщност там където не е нула е степента, в която има грешка при прогнозиране на наградата:

Ако сравните това с уравнението с f в него по-горе, можете да видите, че:

и иначе всичко останало е същото, освен че изяснихме зависимостта от времето на всички променливи и нашето очакване е „очакване за стойност“ вместо за възнаграждение (заменяйки r-hat с V-hat). Също така, както при Rescorla-Wagner, делта стойността тук води до изучаване на очакванията за стойност. Правилото за обучение TD може да се използва за обяснение на голям брой различни феномени на кондициониране и съответствието му с активиране на допаминови неврони в мозъка. То представлява истински триумф на подхода за изчислително моделиране за разбиране (и прогнозиране) на мозъчната функция.

## 5.4 Актьор-критика архитектура за обучение на моторния апарат

Актьор-критика са фамилия алгоритми, които научават както политики на поведение π, така и функциите стойност V. „Актьорът“ е компонент, който научава политикa, а „критикaта“ е компонентът, който научава каквато и политика да се следва в момента от актьора, да „критикува“ актьора избор на действие. Критиката използва TD алгоритъм, за да научи функцията-стойност на състоянието за текущата политика на актьора π. Функция-стойност позволява на критиката да критикува действието на актьора за направените избори чрез изпращане на TD грешки, до актьора.



Фиг.5.4.1 Основна структура на архитектурата актьор-критика за управление на двигателни действия. Фигурата е взета от [2] Фиг.7.6

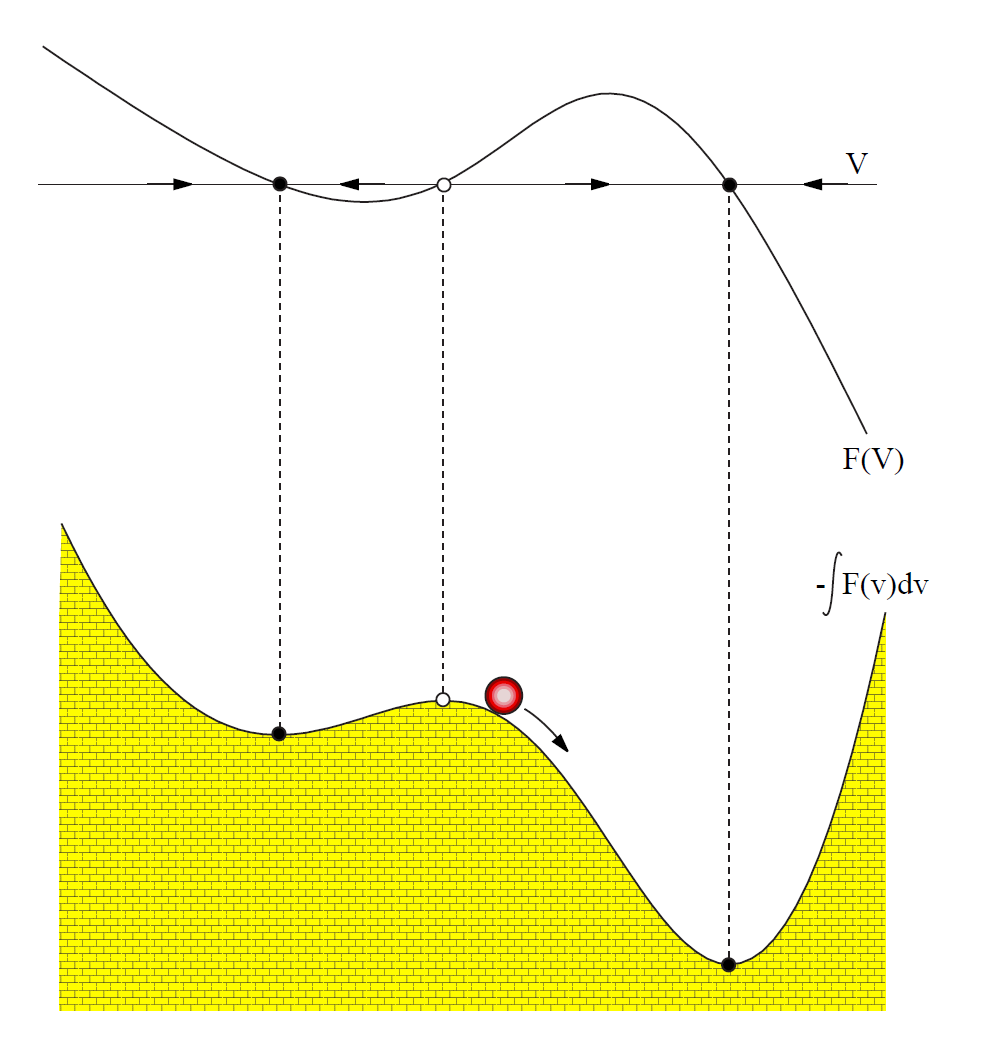
Критиката е отговорна за обработката на сигнала за възнаграждение (r), превръщайки го в грешка при прогноза на възнаграждението (δ), които са подходящи за стимулиране на обучението и на двете, съответно критиката и актьора. Актьорът е отговорен за генерирането на двигателните действия при съответния сензорен вход и не го прави обработвайте награда или очакване за награда директно. Това е ефективно разделение на работата е и от съществено значение за обучението и за трансформиране на наградите в грешки при прогнозиране на наградата, в противен случай системата ще се научи прекалено много само и единствено на прости задачи, които тя би следвало да е овладяла отдавна.

Сега, когато имате по-добра представа за това как работи допаминът, можем да преразгледаме неговата роля в обучението в базалните ганглии. От изчислителна гледна точка, ключовата идея е разграничение между актьора и критиката, където се предполага, че наградите произтичат поне отчасти от правилно изпълнение от актьора. Базалните ганглии са актьорът в този случай, а допаминовият сигнал е полученото на изхода на критиката, която след това служи като тренировъчен сигнал за актьора (и за критиката също). Сигналът за грешка при прогнозиране на възнаграждението, произведен от допаминовата система, е добър тренировъчен сигнал, тъй като стимулира по-силно учене в началото на процеса на придобиване на умения, когато наградите са по-непредсказуеми, и намалява силата на обучение, тъй като умението се усъвършенства и по този начин наградите са по-предвидими. Ако системата вместо това научи директно въз основа на само на наградата(а не на очакваната награда) то тя ще продължи да учи умения, които отдавна са усвоени и това вероятно би довело до редица лоши последствия (синаптичните тегла стават все по-силни, смущения с друго по-ново обучение и т.н.).

За да се продължи с излагане на теоретичната част и даване на алгоритъм за решаване на задачата ще изложим няколко механизма, участващи в решението и са стандартни подходи за невронни мрежи базирани на биологични неврони и ще продължим по-късно с апроксимирация на функцията V и минимизиране на грешката δ .

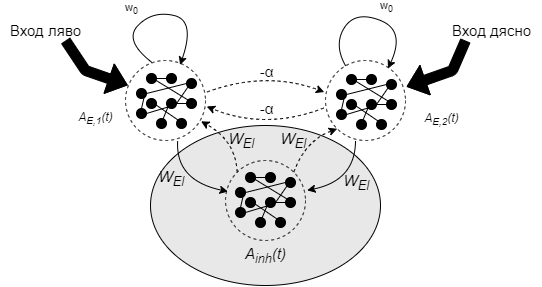
## 5.5 Победителят печели всичко

При динамичните системи изборът на различни действия при различни параметри понякога може да се окаже проблем, тъй като невронните групи навлизат в устойчиво равновесно състояние, и не могат да бъдат изместени от него. Това би се изродило в нашият случай като например агентът да отива само надясно да кажем. За тези положения при динамични системи от един неврон споменава Изикевич в [3] в глава 3.2.6. Наистина за системи с повече неврони теоретичното изчисление на практика е невъзможно и единствено симулациите са водещото, тъй като няма добре установена математическа теория за това. Ще разгледаме механичната интерпретация според [3] на устойчиво и неустойчиво равновесно състояние на следващата фигура.



Фиг. 5.2.1 Механична интерпретация на стабилен и нестабилен еквилибриум. Вж. [3].

Топката на фигурата няма маса (без инерция) и се движи към възможно най-ниската точка със скорост пропорционална на наклона. Механичната диаграма се променя по време на симулацията. Искаме да има 2 различни устойчиви положения (по едно за двете възможни действия на агента) тръгвайки от една нестабилна точка. За целта ще използваме WTA (Winner Takes All) схема, което представлява начин на свързване на невронни групи и като цяло е известен подход в невронните мрежи. WTA има един изход от K възможни. За WTA са необходими възбуждане и потискане на невроните. Всеки изход е представен от група неврони с положителна връзка (excitatory neurons). Входният сигнал възбужда невроните от всяка група, но всеки от тях потиска всички останали и така се състезава с тях. След време само една от невронните групи се оказва най-силна.

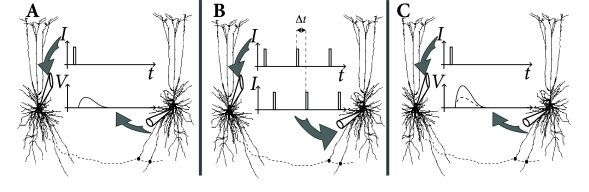


Фиг. 5.2.2 Примерен WTA с два възможни избора и ефективно потискане. Две популации с възбудими неврони взаимодействат с обща група от потискащи неврони. Адаптирано от [6] глава 16.3.

На фигура 5.2.2 са показани как са свързани две невронни групи. Всяка от групите AE,1 AE,2 действа усилващо на обща група Ainh, която пък от своя страна действа потискащо на АЕ,1 и АЕ,2. Групите АЕ действат самоусилващо с някакво тегло w0. Изборът на действие при WTA става като след някакви милисекунди на симулация, достатъчна да се възбудят невроните, преброим спайковете генерирани във всяка от групите АЕ . Избираме действието отговарящо на групата с най-много спайкове.

## 5.6 Обучение с импулсно-времево зависима пластичност(STDP)

Като цяло се приема, че обучението при бозайницие има механизъм на промяна на синаптичната ефективност, както е и при по-простите организми. Исторически теоретичните обосновки се базират на постулата на Хеб (Hebb, 1949), че последователното активиране на два свързани неврона усилва връзката помежду им. Тази промяна наричаме Хебианова пластичност и тя зависи от от пресинаптична активност („pre-synaptic“) и постсинаптична активност („post-synaptic“).



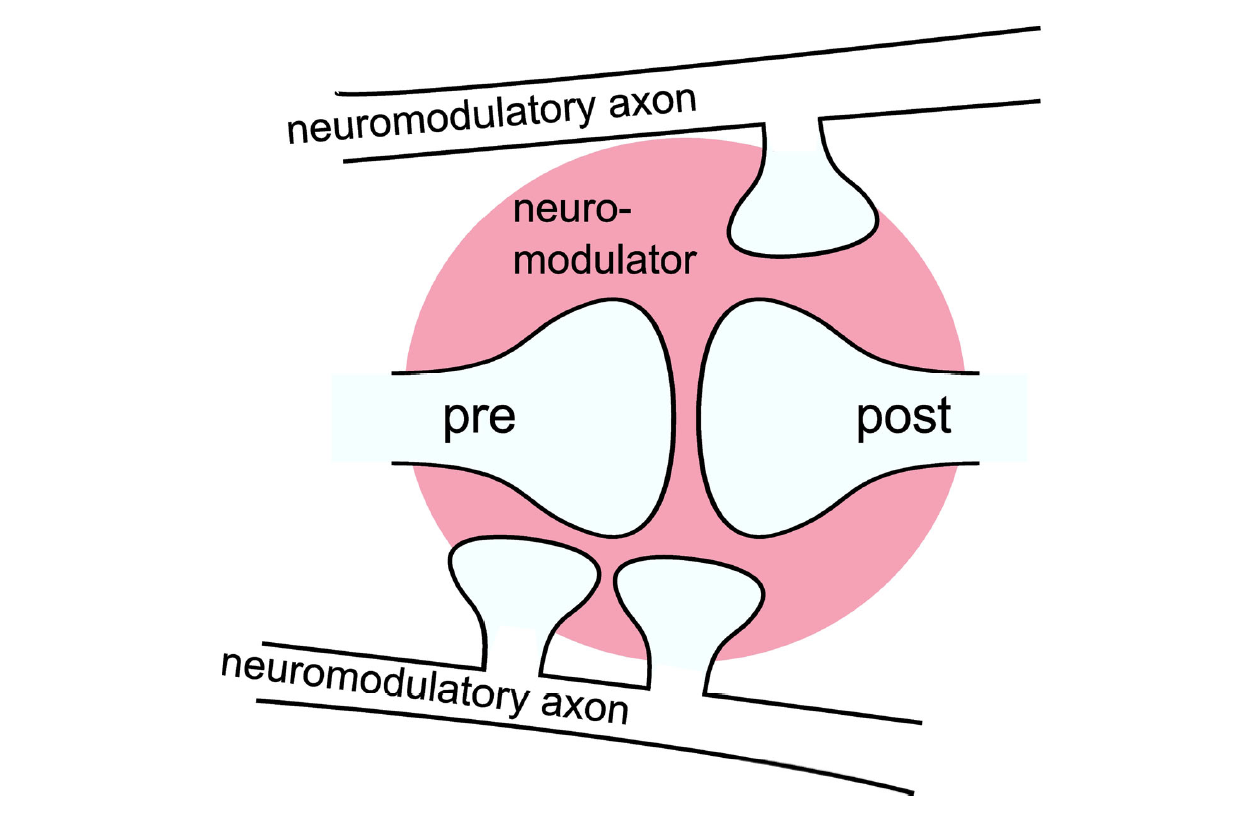
Фиг.5.3.1. Импулсно-времево зависима пластичност (Spike Timing Dependent Plasticity - STDP) **А**. Вътреклетъчни електроди измерват активността на два свързани неврона(аксоните са пунктираната линия). Подава се тестов токов импулс (I) на пресинаптичния неврон и се отчита активност в постсинаптичния неврон (V) . **B.** По време на активиран протокол за синаптична пластичност на двата неврона са подадени токови импулси в точно определено време. **C.** След извеждане на невроните от протокола на синаптична зависимост отново подаваме тестов токов импулс (I) на пресинаптичния неврон и се отчита повишена активност в постсинаптичния неврон (V) (с пунктир е отчетената активност преди сдвояването, плътната линия е след сдвояването. Фигурата е взета от [6], фиг.19.4.

На фигура 5.3.1 са дадения разяснения за STDP като механизъм. Да означим времето на пресинаптичния импулс като tpre и времето на постсинаптичния импулс като tpost. Промяната на теглото на пластичния синапс е зависима от времето . В най-опростен вид промяната на теглата се дава с формулите (вж.[6] глава 19.2.2):

|  |  |
| --- | --- |
| *във време при* | (5.3.1) |
| *във време при* | (5.3.2) |

Тук А+(w)и А-(w)представляват силата на промяната на теглата, τ+ и τ- са константи а w е теглото на синаптичната връзка. Сигнали, които имат далечни по време импулси, допринасят много малко към обучението заради експоненциално-намаляващата зависимост, дадена с интервала .

Активирането на пластична активност може да бъде с външен невромодулатор, който се излива извън клетките. Има такъв биологичен механизъм в бозаиниците и други животни. „Изливането“ на невромодулатор около синапсите се нарича обемно пoдаване (volume transmission). Самото активиране на протокола за обучение е обяснено нагледно на следващата фиг.5.3.2.



Фиг.5.3.2 Обемно подаване на невротрансмитер. „neuromodulatory axon“ – невромодулаторни аксони, „pre“ – аксон от предхождащ неврон, „post” – аксон от последващ неврон. Областта в розово представлява областта на модулираните синапси, за простота е даден само един синапс. Фигурата е взета от [7].

Така обучението на пластичните синапси се контролира посредством друга група от невромодулиращи неврони. Подобен механизъм за активиране на протокола STDP е заложен в Симулатора NEST и се нарича обемен трансмитер (volume\_transmitter). С този механизъм обучението е не само на база на пресинаптичните и постсинаптичните неврони, но и на трети невромодулаторен сигнал. Моделът в симулатора NEST, който представя обучаемите синапси с механизъм STDP и поддържа обемен трансмитер, като трети сигнал се нарича „**stdp\_dopamine\_synapse**“. Параметрите по подразбиране на 'stdp\_dopamine\_synapse' могат да бъдат отпечатани с кода: nest.GetDefaults('stdp\_dopamine\_synapse').

|  |
| --- |
| {'A\_minus': 1.5,  'A\_plus': 1.0,  'b': 0.0,  'c': 0.0,  'delay': 1.0,  'has\_delay': True,  'n': 0.0,  'num\_connections': 0,  'receptor\_type': 0,  'requires\_symmetric': False,  'synapse\_model': 'stdp\_dopamine\_synapse',  'synapse\_modelid': 30,  'tau\_c': 1000.0,  'tau\_n': 200.0,  'tau\_plus': 20.0,  'vt': -1,  'Wmax': 200.0,  'Wmin': 0.0,  'weight': 1.0,  'weight\_recorder': ()} |

Таблица 5.3.1 Параметри по подразбиране за модела 'stdp\_dopamine\_synapse'

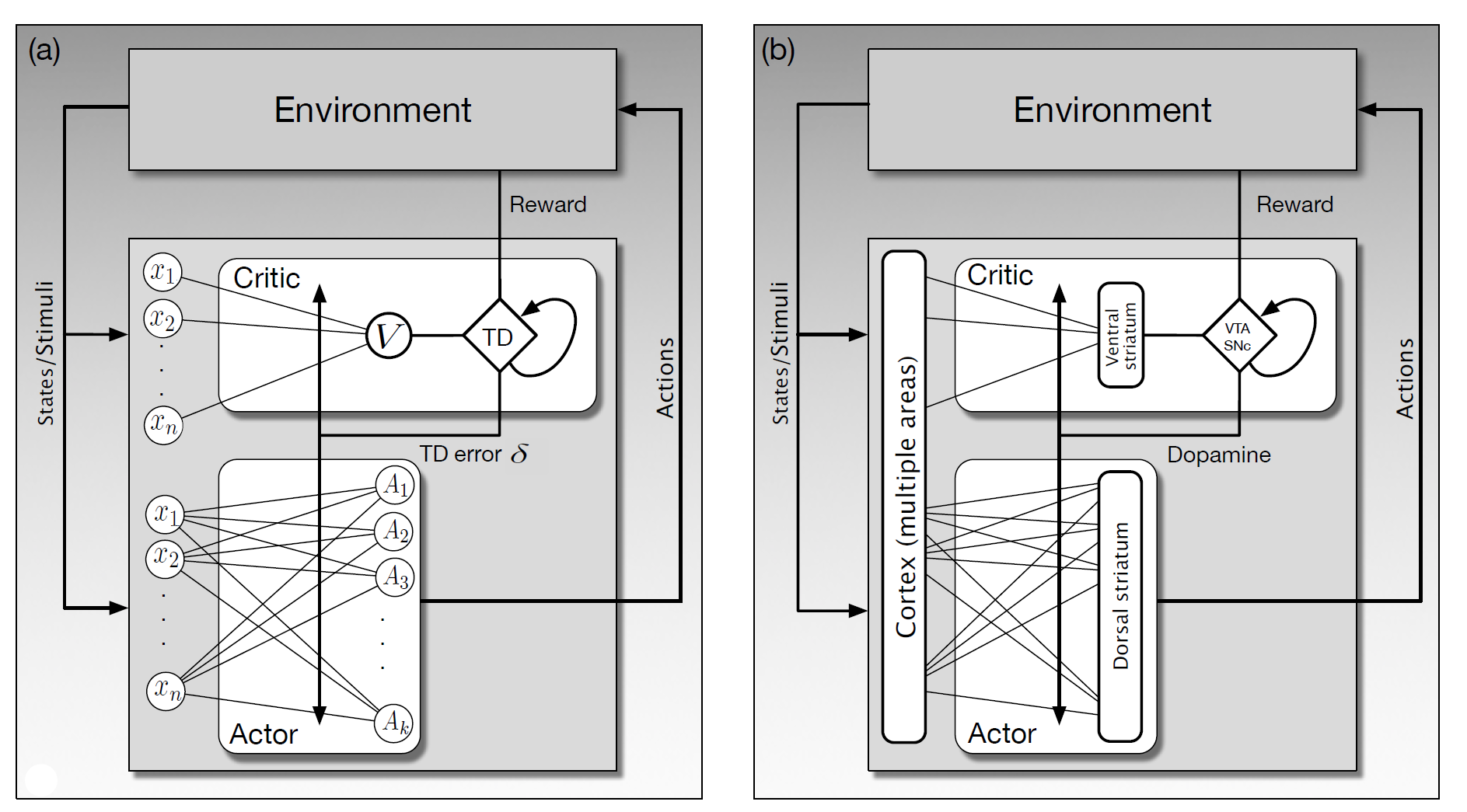
Значенията на параметрите могат да бъдат намерени от [5] на страницата за модели. Ще изброим по-важните, които ни засягат.

|  |  |
| --- | --- |
| Име на параметъра | Значение |
| A\_plus | Коефициент на обучение, когато пресинаптичният импулс изпреварва по време постсинаптичния импулс за два свързани неврона. |
| A\_minus | Коефициент на обучение, когато постсинаптичният импулс изпреварва пресинаптичния по време импулс за два свързани неврона. |
| tau\_plus | STDP времева константа от уравнение (5.3.1) |
| tau\_c | Времева константа на „следата“ на обучението (eligibility trace) |
| tau\_n | Времева константа на следата на допамина (dopaminergic trace) |
| b | Допаминова базова концентрация |
| Wmin | Минимални тегла на обучаемите синапси |
| Wmax | Максимални тегла на обучаемите синапси |

Таблица 5.3.2 По-важните параметри за допаминови синапси и техните описания

## 5.7 Обща постановка за решаване на задачата

Най-забележимата допирна точка на обучението по метода на поощрение-наказание и невронауката е дълбоката химическа връзка на допамина, заложена при бозайниците. В глави 5.3 и 5.4 бяхме разгледали как допаминът отговаря за преноса на времевата грешка (TD) до съответните структури на мозъка, където се извършва обучение и се взема решение за по-нататъчно действие. За методите основани на времевата грешка във време t тя е . Да припомним, че алгоритмите актьор-критика научават, както политиката за актьора, така и функциите за очаквана награда. „Актьорът“ е компонент, който научава политиката на действие, а „критиката“ е компонент, който научава да „критикува“ текущо следваната политика от „актьора“. Критиката използва времевата грешка (TD) за да апроксимира функция за състояние-стойност за текущата политика (vπ(s)). Смята се, че две структури в стриатума от мозъка на бозайниците отговарят за актор и критика, това са Dorsal striatum и Ventral striatum (вж.[1] глава 15.7).

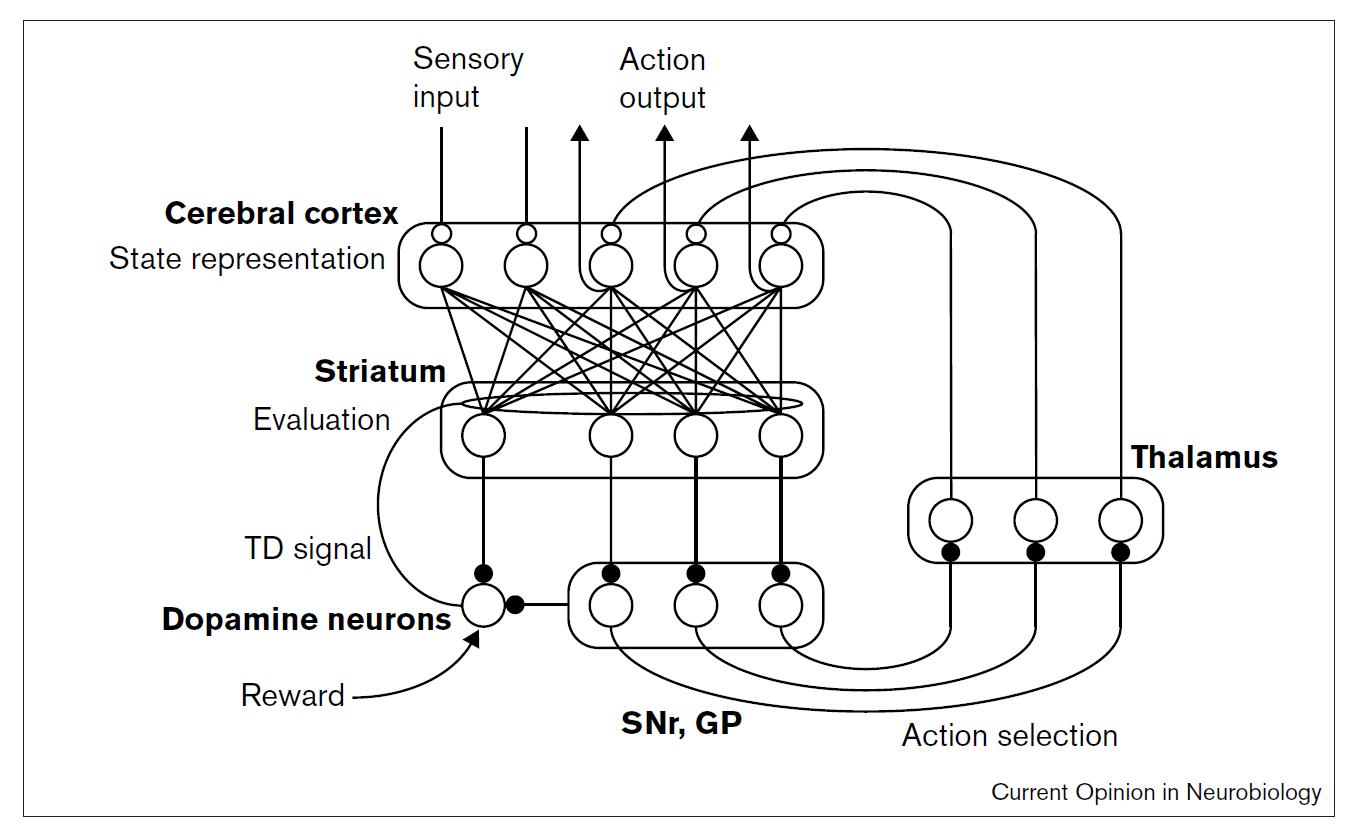
**

Фиг. 5.7.1 Актьор-Критика с невронна мрежа и хипотетична невронна имплементация. а) Актьор-критика като изкуствена невронна мрежа. Актьорът променя политиката спрямо TD грешката δ, който получава от критиката. Критиката създава грешката TD от сигнала за награда Reward. Актьорът няма директен достъп до Reward сигнала. Критиката няма директен достъп до действието. b) Хипотетична невро-имплементация на актьор критика. Актьорът и компонентът научаващ функцията за стойност са съответно в вентралната и дорсалната части на стриатума. Времевата грешката (TD) δ се предава от допаминът, генериран от VTA. Фигурата е копирана от [1] фиг 15.5.

Аналогично за текущият проблем ще използваме комбинация от актьор-критика и алгоритъм, научаващ функцията v(s). По схемата от фиг. 5.7.1 (b) ще използваме компонентът критика за да научи функцията v(s), а компонентът за актьор ще съдържа готовото решение за политика , кодирано в синапсите на връзките.

За апроксимация на v(s) от компонентът за критика ще се използва алгоритъм подобен на SARSA, с известна адаптация. Теорията за реинфорсмънт обучение и алгоритмите помагат за съгласуването на грешката при предвиждане на възнаграждението с конвенционалната представа, че допаминът сигнализира за награда. Алгоритми, които се обсъждат тук, функционират като сигнал за подсилване, което означава, че то е основният двигател на ученето. Например δ е критичният фактор в TD модела на класическо откриване на зависимости и е подсилващият сигнал за научаване както на функцията-стойност, така и на политиката на на актьора. Различни форми на δ сигнала, които са директно зависими от конкретно действие, са подсилващи сигнали за Q-обучение и Sarsa. Сигналът за награда Rt е решаващ компонент на δt-1, но не е пълният определящ фактор за неговото подсилване в тези алгоритми. Допълният член дава зависимост от по-високо ниво за δt-1 и дори да има награда в някакъв момент (Rt ≠ 0) , времевата грешка TD все още може да е 0, ако наградата е напълно предвидена.

Преди да преминем към методите за апроксимиране на търсената функция Vt, ще разгледаме още едно мнение по въпроса как са свързани звената в базалните ганглии. Кенджи Дойа [8] представя мненията на неврообщността в своята обзорна статия и ще използваме диаграмата от статията за се коментира по нея.



Фиг. 5.7.2 Схематична диаграма на кортико-базалния ганглиен цикъл за контрол на моторните функции и евентуалните роли в реинфорсмънт обучението. Фигурата е взета от [8] фиг.2.

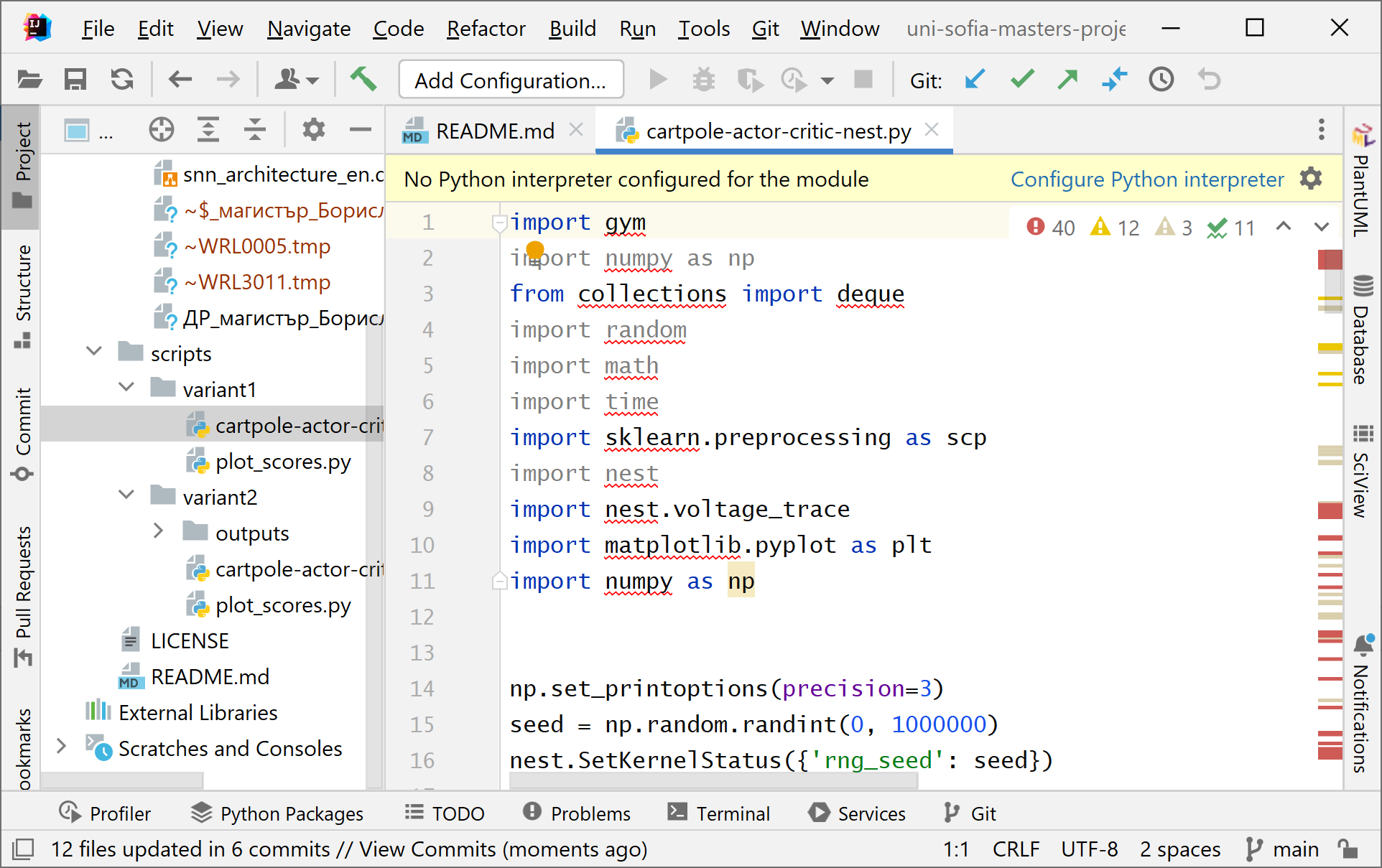
Невроните в стриатума предсказват бъдеща награда за текущото състояние и действията на кандидат-действията. Грешка в прогнозата на бъдеща награда, т.е. TD грешката, е кодирана в активността на допаминовите неврони и се използва за обучение в кортико-стриаталните синапси. Едно от действията-кандидати е избрано в SNr и GP в резултат на състезание на прогнозирани бъдещи награди. Директният и индиректен пътища в globus pallidus са пропуснати за простота на диаграмата и няма да ги коментираме, защото много биха усложнили текущата задача. Запълнените и празните кръгове означават съответно инхибиторни и възбуждащи синапси. Тази диаграма много прилича на Фиг.5.7.1 от книгата на Р.Сътън и А.Барто по начина на свързване и представяне на TD грешката.

## 5.8 Подход за решаване с варианти

Тъй като за решаване на такъв тип задача не може да се подходи директно с решение, можем в следащите глави при реализацията на проекта да бъдат предложени варианти и доближаване до решението по итеративен път. Ще бъде следван основният подход от високо ниво, обяснен в предната точка. Ще има експерименти с различни мрежи, които обаче ще следват принципите за актьор-критика и обучение с времева грешка TD(0). За отделните експерименти ще бъдат разделени на подексперименти.

# 6. Реализация на проекта

Проектът е реализиран като github публичен проект и може да се разгледа и през браузър (виж Приложения). За да се пусне локално се изисква инсталация на Python, конкретно тук използваме “Python 3.11.0” заедно с Conda (независим от езика мениджър на пакети и система за управление на средата). Използвана е операционна система Линукс – Ubuntu. Връзка към сорс кода е качен в гитхъб (Вж. Приложение 1) и е неразделна част от този документ. Структурата на приложението е дадена на фигура 6.1. Използваната среда за текстообработка и работа с git е IntelliJ .



Фигура 6.1. Обща структура на проекта

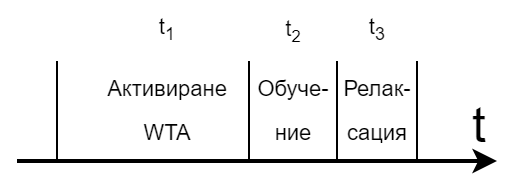
Подробни инструкции на са дадени в README.md файла.

В централната папка има папка „scriptс“ и в нея имаме подпапки за вариантите. Всеки вариант има скрипт “cartpole-actor-critic-nest.py” на програмния език Python. С него се стартира процеса на обучение. По време на обучение резултатите от точките (поощрението) се записват във файл „scripts/variantX/outputs/scores.txt“ за последваща визуализация. Скриптът „scripts/variantX/plot\_scores.py“ ще ни визуализира картинка с резултатите след текущото обучение. Процесът на обучение и работа на вече обученият агент не са разделени. За край на обучение се приема момента, когато средно аритметичната награда от последните SOLVED\_HISTORY\_SCORES\_LEN=10 епизода е над SOLVED\_SCORE =195.

## 6.2 Експериментална част

Ще бъдат разгледани два варианта за евентуално решение. Като начало се дават общите неща и за двата варианта.

Основното при такъв тип симулация е как ще се извършва времеделенето и симулацията. Има вариант при който симулацията върви непрекъснато и невронната мрежа получава въздействие от средата чрез външен интерфейс. Тук в тази дипломна работа е избран по-прост начин, а именно чрез цикъл в който се редуват обучение и въздействие.



Фиг.6.2.1 Времеделене при симулация

На фигура 6.2.1 е показано как става това. В главният цикъл на програмата в който се управлява посоката на агента имаме за всяка стъпка тези три времена, които се редуват за всяка стъпка от дадения експеримент. По-надолу на фиг. 6.2.2 е обяснено по-подробно а времената са споменати в стъпки 5, 8, и 9 от фигурата. Времето t1= 40ms (в кода означено като константа STEP) е времето в което се активира кръгът Winner Take All за избор на едно от четирите действия. На самите неврони им трябва някакво техническо време да се установи кой ще спечели състезанието и да може ефективно да потисне останалите. Това време може да е от порядъка на 40ms до към 400ms. Опитът показа, че по-големи стойности не променят резултата, а по-малки не дават сигурен резултат. Времето t2=20 (в кода означено като LEARN\_TIME) е времето в което се обучават допаминовите връзки, но само при положителна награда. Времето t3=50ms (в кода съответно REST\_TIME) е времето в което се успокояват невроните от WTA за да се върнат в изходна позиция готови за ново възбуждане през следващият цикъл. Ще аргументирам защо времената са подредени в тази последователност. Интервалът t2 не трябва да припокрива t1 защото действието още не е взето от WTA и наградата още не е дадена от средата, съответно няма какво да обучаваме все още. Докато действа t2 невроните от WTA трябва да са във възбудено състояние, за да е ефективно обучението на допаминовите синапси по закона на Хеб, а именно че възбудените неврони по едно и също време усилват връзката си.

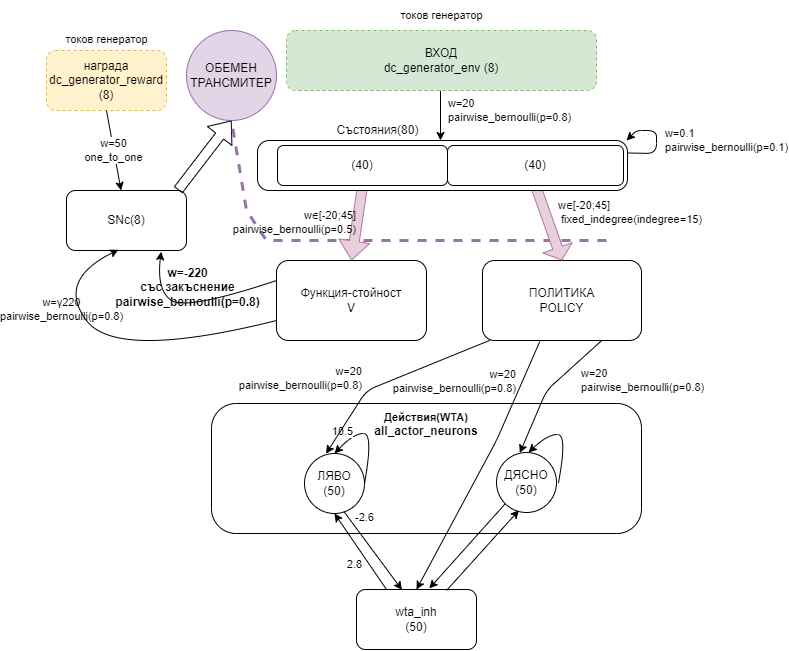


Фиг.6.2.2 Опростена блок-схема на обучението на агента и за двата варианта

### 6.2.1 Вариант 1 за решение на CartPole

Тъй като активността на невроните се моделира със spike timing, се налага преминаването от спайкове към числени стойности на изхода на всяка група неврони. Входът от околната среда ще се моделира от 8 токови генератора. Ще разгледаме по-подробно как се преобразуват числените стойности в токов сигнал малко по-надолу. След това ще имаме състоянието на средата представено от две групи по 40 неврона, общо 80 неврона. Ще имаме една група за функцията-стойност V от 40 неврона и една група за функцията на политиката π от 40 неврона. Свързването е представено на Фигура 6.2.1.1.

При подаване на състояние от средата се активират различни нива на токовите генератори на входа и се активира само определената група неврони от Състояния (STATES), отговаряща за това състояние.



Фиг.6.2.1.1 Диаграма на свързване на невронните групи за вариант 1

Типът свързване на STATES със V е по двойки тип Бернули (pairwise\_bernoulli) и има случаен елемент при такъв тип свързване. При свързване pairwise\_bernoulli за всяка двойка неврони от групи A и B връзка ще бъде направена с вероятност „p“. Пример за такъв тип свързване се дава със следния код.

|  |
| --- |
| n, m, p = 10, 12, 0.2  A = nest.Create('iaf\_psc\_alpha', n)  B = nest.Create('iaf\_psc\_alpha', m)  conn\_spec\_dict = {'rule': 'pairwise\_bernoulli', 'p': p}  nest.Connect(A, B, conn\_spec\_dict) |

Таблица 6.2.1.1 Примерен код за свързване по двойки тип Бернули.

При увеличаване на бройката неврони във всяка от групите би се увеличил и тока в приемащия неврон, затова свързване тип „всеки със всеки“ не се препоръчва. Връзките между STATE и V са с допаминови синапси, първоначално с равномерно разпределени случайни тегла в интервала [-20;+45] , които се обучават посредством пластични синапси (вж. 5.3) (STDP, spike-timing dependent plasticity – Markram et al., 1997; Bi and Poo, 1998, 2001). Формулите за промяна на теглата в опростен вид са (5.3.1) и (5.3.2). Връзката обхваща само едната половина на STATE, т.е. 40 неврона и с код на Python се изразява като STATE[40:].

Свързването на STATES със POLICY e фиксиран брой входящи връзки(fixed\_indegree), при което също има случаен елемент. За такъв тип свързване между две групи А и B, всеки неврон от приемащата група B има фиксиран брой входящи връзки. По този начин при нарастване на бройката неврони в групата A сумарните токове в групата B няма да нарастнат. Примерен код ще дадем в следващата таблица:

|  |
| --- |
| A = nest.Create('iaf\_psc\_alpha', 5)  B = nest.Create('iaf\_psc\_alpha', 3)  conn\_spec\_dict = {'rule': 'fixed\_indegree', 'indegree': 2}  syn\_spec\_dict = {'weight': [[1.2, -3.5],[0.4, -0.2],[0.6, 2.2]]}  nest.Connect(A, B, conn\_spec\_dict, syn\_spec\_dict) |

Таблица 6.2.1.2 Примерен код за свързване тип фиксиран брой входящи връзки.

Връзките между „STATE“ и „POLICY“ са с допаминови синапси, първоначално със случайни тегла в интервала [-20;+45] , които се обучават посредством пластични синапси (вж. 5.3) (STDP). Връзката обхваща само втората половина на „STATE“, т.е. 40 неврона и с код на Python се изразява като STATE[0:40].

Звеното „STATE“ е свързано към невронните групи за действия (WTA) с тип pairwise\_bernoulli (p=0.8). Всяка група от WTA се състои от 50 неврона и отговаря съответно на действията на агента, наречени „action\_left” и „action\_right“. Връзките между „States“ и „Actions“ не са обучаеми и не се променят.

Групата „V“ са свързани с друга невронна група, наречена „SNc“ от 8 неврона и се състои от връзки с фиксирани тегла, които не се обучават. Връзките са на две групи, от които едната е със закъснение, равна по време на една стъпка, за да формира очакваната грешка δ, както е уточнено в 5.3. При липса на грешка, т.е. δ=0 няма да има активиране на обемния трансмитер за подаване на допамин и няма да има обучение на допаминовите групи. При грешка различна от нула, т.е. δ≠0 ще има активиране на SNc и съответно ще има подаване на допамин към всички допаминови връзки и ще има обучение на теглата на допаминовите връзки.

Разглеждайки гореизложената схема на свързване като Атьор-критика имаме, че актьорът са невронните групи POLICY + WTA, докато критиката е V+SNc. Условието актьорът да няма директен достъп до Reward сигнала и критиката да няма директен достъп до действието е изпълнено.

Ще разгледаме преобразуването на входният сигнал от обръжаващата средата до такъв сигнал, който е подходящ да захрани подобна симулационна невробиологична мрежа. Входът от средата или тъй нареченото векторно пространство на наблюдението (Observation Space) се представя от вектор, състоящ се от 4 стойности.

|  |  |  |  |
| --- | --- | --- | --- |
| индекс | Вид наблюдение | Минимална ст-т- | Максимална ст-т |
| 0 | Позиция на количката | -4.8 | +4.8 |
| 1 | Скорост на количката | -∞ | +∞ |
| 2 | Ъгъл на балансираното рамо | -0.418 rad | +0.418 rad |
| 3 | Ъглова скорост на балансираното рамо | -∞ | +∞ |

Таблица 6.2.1.3 Входни параметри от средата на CartPole

Ако съпоставим четири токови генератора на четирите входни стойности и ги скалираме с някакъв параметър решението би могло да работи частично. Трудно бихме се справили с отрицателните стойности, защото такива отрицателни токови импулси биха неутрализирали други положителни токови импулси. Също е трудно да се справим с безкрайността (∞) при скоростта на количката и ъгловата скорост на балансираното рамо. За целта можем да разделим обхвата на всеки един от параметрите на два интервала, обхващащи положителната и отрицателната част от дефиниционната област. Ще въведем и подходящо скалиране, така че изходящите стойности да са във фиксиран интервал и така ще избегнем безкрайността от дефиниционната област. С код на Python това става лесно с помощта на MinMaxScaler (от пакета scikit-learn) и функция, разделяща всеки параметър на два параметъра. Когато позицията на количката е отрицателна стойност, то тя бива разделена на две стойности (0, abs(s(0))). За целта използваме следната функция, дадена в таблица.

|  |
| --- |
| def transform\_state(s):  transformed = np.array([  abs(s[0]) if s[0] > 0 else 0,  abs(s[0]) if s[0] < 0 else 0,  abs(s[1]) if s[1] > 0 else 0,  abs(s[1]) if s[1] < 0 else 0,  abs(s[2]) if s[2] > 0 else 0,  abs(s[2]) if s[2] < 0 else 0,  abs(s[3]) if s[3] > 0 else 0,  abs(s[3]) if s[3] < 0 else 0])  return transformed |

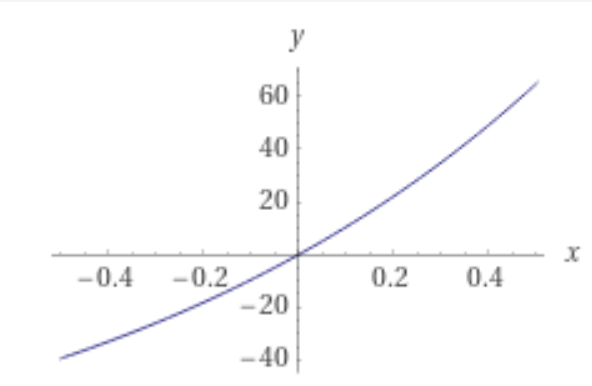
Таблица 6.2.1.4 Преобразуване на входните параметри и преминаване от пространство с размерност 4 към пространство от размерност 8

Разбира се, вероятно има и по кратък начин да се направи подобно преобразуване, но за нагледност и простота, кода е оставен по този начин. След такова преминаване към размерност 8 на входните параметри, имаме скалиране.

|  |
| --- |
| scaler = scp.MinMaxScaler(feature\_range=(0.01, 1), copy=True, clip=True)  scaler.fit([[0, 0, 0, 0, 0, 0, 0, 0], [+1.5, +1.5, +1.5, +1.5, +0.13, +0.13, +2.1, +2.1]])  ...  new\_transformed\_state\_scaled = scaler.transform(new\_transformed\_state.reshape(1, -1)).reshape(-1) |

Таблица 6.2.1.5 Преобразуване на входните параметри към интервал с фиксирана стойност [0.01;1]

Скалирането е с фиксиран интервал на изходящите стойности от 0.01 до 1. Входните минимални и масимални стойности са зададени с фиксирани стойности и са получени на базата на експерименти по емпиричен начин. Така например скоростта на количката не ни интересува ако е по-голяма от 1.5 или по-малка от -1.5. Допълнително изходящите стойности се преобразуват през експоненциална фунцкия , като идеята е по-ниските стойности да влияят по-слабо и отивайки към по-високи стойности, влиянието да се засилва. Факторът 100 е подбран емпирично.



Фиг. 6.2.1.2 Допълнително експоненциално преобразуване на входа

Така можем да дадем примерни входове от средата и техния еквивалент подаван като токове към мрежата.

|  |  |  |  |
| --- | --- | --- | --- |
| Вход от средата, размерност 4 | Трансформация към размерност 8 | Трансформация със скалиране | Трансформация [pA] |
| [0.033,  0.016,  0.048,  0.003] | [0.033, 0.0,  0.016, 0.0,  0.048, 0.0,  0.003, 0.0] | [0.032, 0.01,  0.021, 0.01,  0.375, 0.01,  0.012, 0.01] | [3.214, 1.005,  2.095, 1.005,  45.484, 1.005,  1.163, 1.005] |
| [0.117,  1.38,  -0.069,  -1.999] | [0.117, 0.0,  1.38, 0.0,  0.0, 0.069,  0.0, 1.999] | [0.087,0.01,  0.921, 0.01,  0.01, 0.538,  0.01, 0.952] | [9.097, 1.005,  151.112, 1.005,  1.005, 71.209,  1.005, 159.201] |

Таблица 6.2.1.6 Тристъпково преобразуване на входните параметри от средата CartPole във формат удобен за захранване на токови генератори

Така скалиран тока става между 1.005 и 171.82 пико ампера. Допълнително може да се регулира приноса на всеки един токов генератор с теглото на връзката от INPUT към STATE, която към момента е w=20.

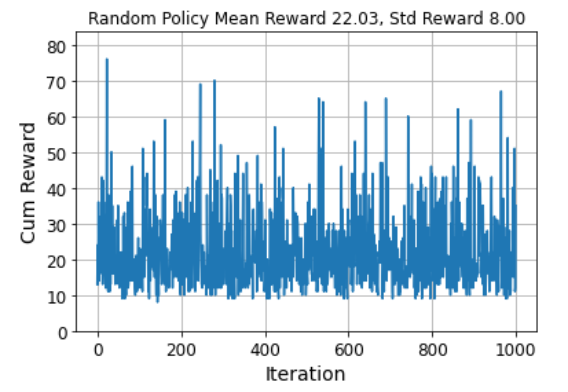
Тук е редно да спомена, че не бихме могли да се справим с отрицателна награда без съществено да променим постановката, тъй като обемният трансмитер на допамин работи на базата на генерирани спайкове, които винаги са положително число (няма как да генерираме отрицателни спайкове). Всички тези тегла и стойности биха могли да се сторят на читателя подбрани на случаен принцип и е интересно какво би се случило ако са заложени различни хиперпараметри. Достатъчно е коефициентите да са такива, че да предизвикват спайкове в невроните в мрежата и от там нататък се активира WTA. Разликата ще бъде, че вместо да имаме състезаващи се много на брой спайкове във WTA ще имаме състезание от по-малко на брой спайкове и в края на краищата се взема само максималното, така че тези константи не са толкова важни стига на диаграмите да се вижда че има ясно изразени спайкове във всички нива на мрежата.

Нека да разгледаме резултатите от първият вариант. Пуснали сме скрипта първоначално със 100 епизода. Като резултат имаме получената награда за всеки епизод, или иначе казано колко дълго сме успели да балансираме рамото без да падне.

|  |  |
| --- | --- |
| (А) | (Б) |
| (В) | (Г) |
| (Д) | (Е) |

Фиг. 6.2.1.3 Резултати от обучение, вариант 1 за 100 епизода, А,Б,В,Г,Д – Спайкове и хистограма на различни невронни групи за последният епизод, E – резултат от обучение и получени точки

Вижда се, че действително има епизоди, които са успели да балансират рамото над 80 стъпки (Фиг. 6.2.1.3,Е). За да преценим дали е случайност трябва да разгледаме някаква отправна точка, например при случайно избиране на действие, какво бихме постигнали. Ще разгледаме резултатите за една такава случайна политика от [11].



Фиг. 6.2.1.4 Отправна точка със случайна политика на агента за 1000 сепизода, Kопирано от [11].

Средната стойност на кумулативната награда от отправната точка (baseline) е 22, докато във вариант 1, разгледан тук е около 30. Това очначава, че този вариант не може да обучи агент да реши напълно задачата, а можем да считаме, че е малко по-добър от агент със хаотична политика. Какво би могло да се подобри. Например наградата, давана от средата е винаги 1.0, което е подвеждащо до някаква степен, дори балансираното рамо да падне в следващата стъпка, наградата указва на агента, че се е справил успешно. За целта ще преправим функцията на наградата и ще пуснем вариант 1 още веднъж. За функция на наградата ще използваме . Такава функция ни дава награда в интервала от 0 до 10.0 и зависи изцяло от ъгъла на балансираното рамо.



Фиг. 6.2.1.5 Резултати от обучение при преправена функция на наградата, вариант 1 за 100 епизода

Колкото по-изправено е балансираното рамо, толкова е по-голяма наградата. Резултатът не е много по-различен от преди и можем да считаме, че няма подобрение. Промяна на други хиперпараметри, като например γ =0.93;0.95; 0.98;1.0 също не дава подобрение.

|  |  |
| --- | --- |
| γ=1.0 | γ=0.98 |
| γ=0.95 | γ=0.93 |

Фиг. 6.2.1.6 Резултати от обучение на вариант 1 при промяна на γ

### 6.2.2 Вариант 2 за решение на CartPole

Текущата мрежа и начин на свързване от вариант 1 не могат да бъдат подобрени с изброените методи в горната подточка, поради което се налага да се помисли за друг начин на свързване и подаване на входния сигнал. За целта съм разработил втори скрипт, спазващ принципите на свързване и разграничаване на актьор-критика, и също зависещ от темпоралната грешка TD(0).

Скриптът от вариант 2 е направен да приема определени параметри от командния ред, за да не се налага всеки път да се променя сорс кода. Опциите на командния ред са както следва:

1. “-о” – избор на изходна директория за резултати
2. “-c” – избор на изтриване на изходната директория, “true” или “false”
3. “-n” – максимален брой епизоди
4. “-v” – отпечатване на подробни съобщения в терминала.

Пускането от операционна система Линукс става по следния начин:

|  |
| --- |
| python cartpole-actor-critic nest.py -n 100 -v |

Таблица. 6.2.2.1 Пускане на скрипта за обучение от папка “variant2” за 100 епизода с подробно отпечатване на логовете (verbose)

Обучението завършва успешно и резултатът се отпечатва на терминала.

|  |
| --- |
| Episode 20 finished after 10 timesteps  SOLVED  [[0.029 0.01 0. ]  [0.014 0.402 0. ]  [0.829 2.906 0. ]]  ====== all\_states === all\_actions ===  source target synapse model weight delay  -------- -------- --------------- -------- -------  1 181 dopa\_synapse 0.09612 1.000  1 271 dopa\_synapse -0.004897 1.000  1 361 dopa\_synapse -0.007254 1.000  ................. |

Таблица 6.2.2.2 Край на обучението на агента за вариант 2

Като резултат имаме и картинки в директорията “scripts/variant2/outputs”. Сега нека да обърнем вниманието към логиката в скрипта.

Компонентът критика ще научи функцията V(s), а компонентът за актьор ще съдържа готовото решение, кодирано в синапсите на връзките. Тъй като активността на невроните се моделира със spike timing, се налага преминаването от спайкове към числени стойности на изхода на всяка група неврони.

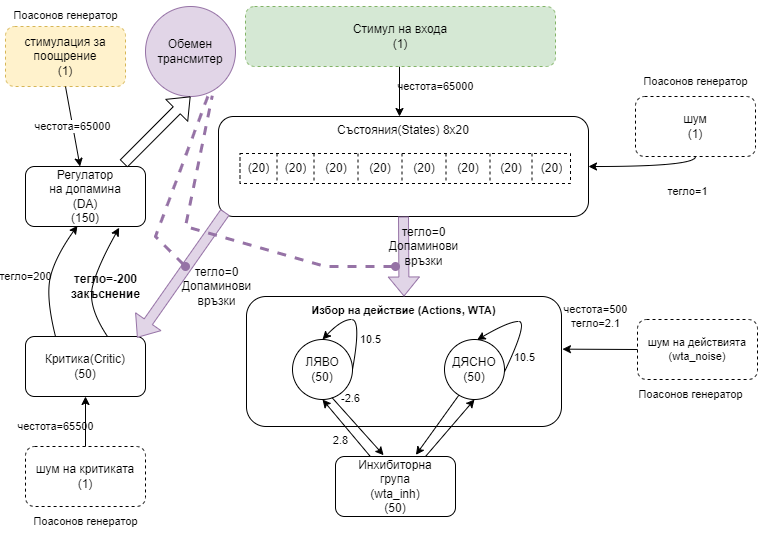
Аналогично на вариант 1 имаме преобразуване на входния сигнал от обръжаващата средата до такъв сигнал, който е подходящ да захрани подобна симулационна невробиологична мрежа. Входът от средата или тъй нареченото векторно пространство на наблюдението (Observation Space) се представя от вектор, състоящ се от 4 стойности и се преобразува до вектор от 8 стойност

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| индекс | Вид наблюдение | Допустими ст-ти | Отрязване на стойностите  (clipping) | Скалиране и разделяне на 2 стойности | Допълнително преобразуване |
| 0 | Позиция на количката | [-4.8; +4.8] | [-1.5; +1.5] | [0.01;1]  [0.01;1] | , |
| 1 | Скорост на количката | [-∞;+∞] | [-1.5; +1.5] | [0.01;1]  [0.01;1] |  |
| 2 | Ъгъл на балансираното рамо | [-0.418;  +0.418] | -0.13;+0.13 | [0.01;1]  [0.01;1] |  |
| 3 | Ъглова скорост на балансираното рамо | [-∞;+∞] | -1.5; +1.5 | [0.01;1]  [0.01;1] |  |

Таблица 6.2.2.1 Входни параметри от средата на CartPole и преобразуване до размерност 8

За всяка от 8 размерности на вече преобразувания входен вектор се създава отделна група от по 20 неврона, разположени таблично и се нарича „States“. Тази група ще представлява са възможните състояния на средата и се състои от 8\*20=160 неврона. При преместване на агента на различна позиция се активират по-малко или повече определените групи неврони отговаряща за всяка една компонента на 8-размерния вектор. Активацията става посредством генератор на поасонов шум с определена честота, наречен „стимул“. Така на всяка стъпка от всеки епизод, всяка компонента от „States” бива захранена с различно тегло от връзката си със стимула.

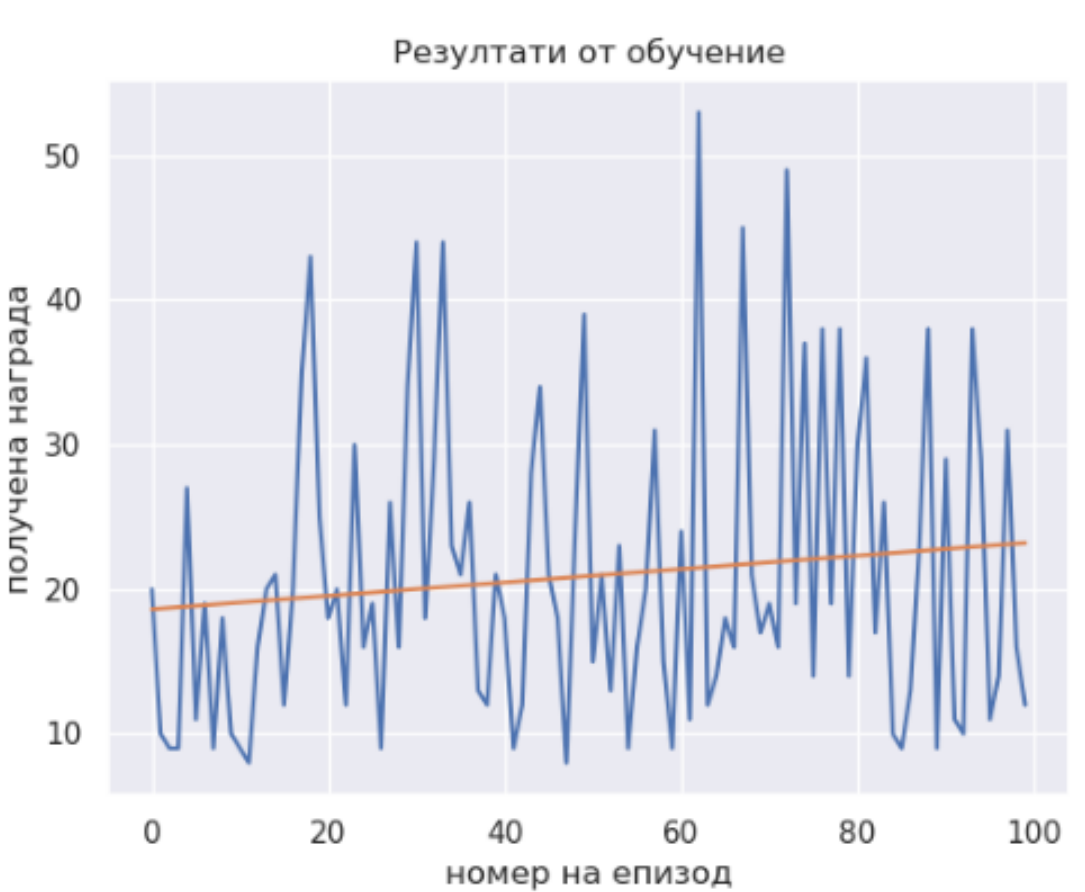
„States“ са свързани към WTA схема, с 2 възможни състояния като свързването е „всеки с всеки“ (в NEST - „all\_to\_all“). Всяка група от WTA се състои от 50 неврона и отговаря съответно на действията на агента, наречени „Actions”, от 0 до 1 включително, а именно: наляво-0, надясно-1. Връзките между „States“ и „Actions“ са с допаминови синапси, първоначално с тегла 0.0, които се обучават посредством пластични синапси (STDP , вж. 5.6). Диаграмата на свързване е дадена на следващата фигура.



Фиг.6.2.2.1 Диаграма на свързване на невронните групи за вариант 2

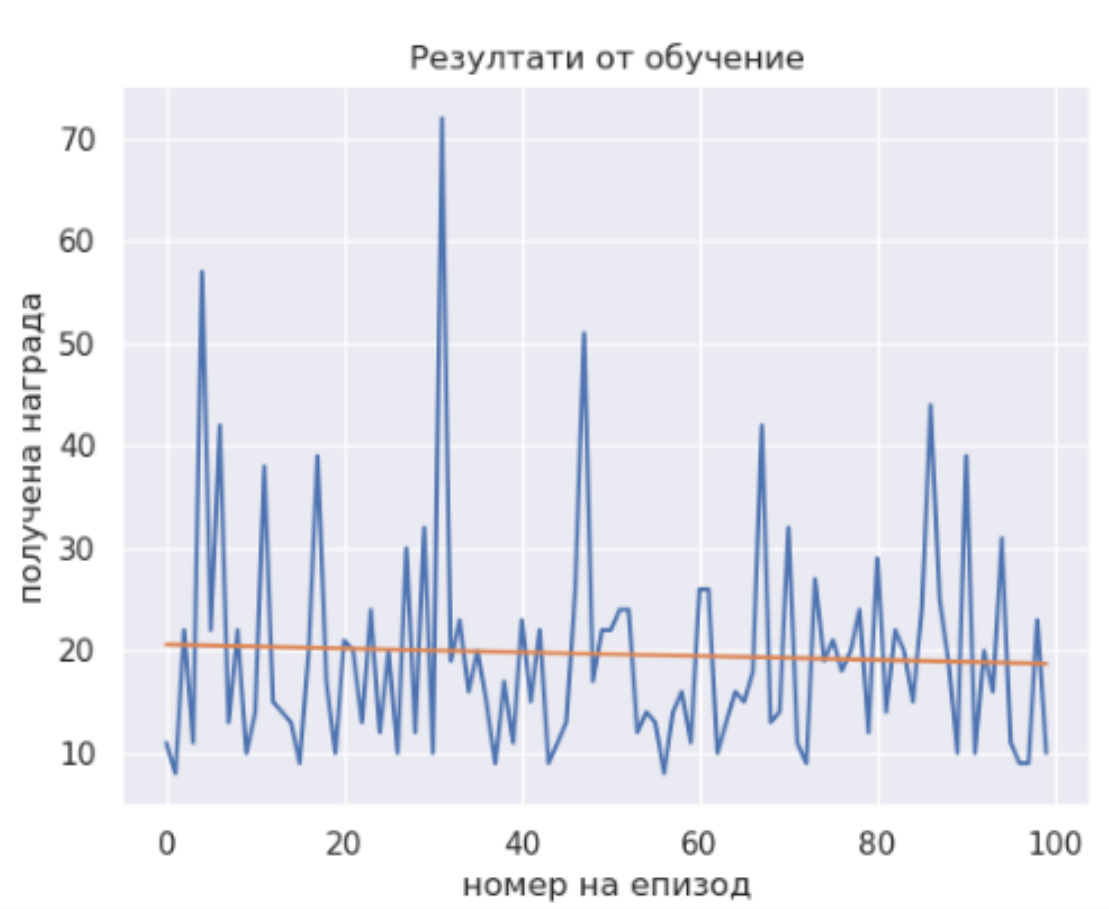
„States“ са свързани с друга невронна група, наречена „Critic“ от 50 неврона също с допаминови връзки. Тази група представя „v\*“ функцията от уравнението на Белман. „Critic“ е свързана с друга група от 150 неврона, отговарящи за нивото на допамина, условно наречена „DA“. Наградата от средата ще се формира като сигнал от поасонов шумогенератор с определена честота пропорционална на наградата. Този вход е наречен „Reward Stimulus“. В синаптичните връзки между „States“ и „Actions“ е заложено решението, защото в процеса на обучение на „Critic“, успешните ходове на агента са предпоставка за усилваща връзка от дадено състояние на средата, към определена посока, например „Ляво“. „Reward Stimulus“ постъпва през групата „DA“. Така „DA“ отчита очакваната награда, а не абсолютната награда, което съответства на TD грешката от (5.7). Допаминовите синапси в симулатора NEST получават нивото на допамина, генерирано от DA посредством устройствотo volume transmitter, описан в 5.6. Тъй като средата CartPole дава награда на всеки ход от всеки експеримент ще имаме обучение винаги, когато наградата се различава от очакваната награда. Тук е редно да спомена, че не бихме могли да се справим с отрицателна награда без съществено да променим постановката, тъй като обемният трансмитер на допамин работи на базата на генерирани спайкове, които винаги са положително число (няма как да генерираме отрицателни спайкове).

След пускане на скрипт вариант 2 наблюдаваме следните резултати.



Фиг. 6.2.2.2 Резултати от обучение на вариант 2 при γ=0.95

Вариантът не дава очакваните резултати за решение и ще анализираме каква може да е причината. Наградата се дава на всяка стъпка независимо дали е в правилната посока. Т.е. агента може да третира като успешен всеки ход. Значи погледа в бъдеще се контролира от параметъра γ и бихме могли да увеличим този параметър на 0.999, защото стойности близки или равни на 1 означават, че наградите в бъдеще са толкова ценни, колкото и тези получени сега.



Фиг. 6.2.2.3 Резултати от обучение на вариант 2 при γ=0.999

Отново резултатите не дават надежда за решение и ще преминем към следващата промяна. Ще се промени начина на оценка. Наградата от средата ще е пропорционална на ъгъла на балансираното рамо. Колкото по изправено е рамото, толкова по-висока ще е наградата. При наклон на рамото след определен ъгъл наградата се променя на 0. Както споменахме при вариант 1, фунцкията е следната: . Имаме допълнителна константа с която можем да регулираме скалирането на наградата REWARD\_SCALING=2.5.

|  |  |
| --- | --- |
| (А) | (Б) |
| (В) | (Г) |
| (Д) | (Е) |

Фиг. 6.2.2.4 Резултати от обучение, вариант 2 за 250 епизода, А,Б,В,Г – Спайкове и хистограма на различни невронни групи за последния епизод, Д,E – резултати от обучение и получени точки за две поредни стартирания

Виждаме, че има значително подобрение (Фиг.6.2.2.4-Е) и средните точки от епизод надминават 100 точки. Можем да считаме, че задачата е решена с това ново подобрение във вариант 2.

Тренирането на подобна задача отнема на DELL XPS, Intel I9 32GB RAM с виртуална машина VirtualBox и Ubuntu около 1 час. Намалявайки невроните в някои от групите може да ускори изчисленията, но не значително.

## 6.3 Параметри на постановката и анализ на резултатите

За допаминовите синапси са зададени следните параметри:

|  |
| --- |
| tau\_c = 200.0 # Time constant of eligibility trace  tau\_n = 5.0 # Time constant of dopaminergic trace  tau\_plus = 20.  # Connect states to actions  nest.CopyModel('stdp\_dopamine\_synapse', 'dopa\_synapse', {  'vt': vol\_trans.get('global\_id'), 'A\_plus': 0.05, 'A\_minus': 0.05, "tau\_plus": tau\_plus,  'Wmin': -100., 'Wmax': 100., 'b': 0., 'tau\_n': tau\_n, 'tau\_c': tau\_c}) |

Таблица 6.3.1 Копиране на модела „stdp\_dopamine\_synapse“ и подмяна на параметрите

На скоростта на обучение можем да повлияем като скалираме поощрението от средата. Към момента това става с емпиричната формула:

|  |
| --- |
| REWARD\_SCALING = 2.5 \* 20 / NUM\_STATE\_NEURONS  ...  nest.SetStatus(nest.GetConnections(reward\_stimulus, DA\_neurons), {'weight': float(reward) \* REWARD\_SCALING}) |

Таблица 6.3.2 Скалиране на наградата и влияние на скоростта на обучение

Всъщност скоростта на обучението зависи до колко ще усилим генерираният шум от поасоновия генератор „reward\_stimulus“ към допаминовите звена „DA\_neurons“ и получената стойност се получава експериментално.

Синапсите имат един механизъм за следа (eligibility trace) в обучението и той може да ни помогне много в текущата задача, но може и да доведе до объркване. Тази следа може да спомогне да проследи през какви стъпки е минал агента стигайки до конкретно решение. Това регулираме с константата „tau\_c“ когато копираме модела „stdp\_dopamine\_synapse“.

Друг полезен механизъм е регулиране на базовата концентрация на допамин „b” от таблица 6.3.3. Давайки положителна стойност, например 0.01 е все едно имаме награда от средата на всяка стъпка, без да има реална награда. Вътрешният механизъм е заложен в STDP обучението на невроните. Реализирането на ε-greedy политика π от алгоритъма става посредством поасонов шумогенератор „wta\_noise“ с честота „WTA\_NOISE\_RATE“ показан на Фиг.6.2.2.1 долу вдясно. Давайки по-голяма честота, ще засилим силата на случайния сигнал за вземането на решения в „Actions”. Давайки по-малка честота ще имаме по-малко шум и съответно няма да изпробваме нови комбинации от действия а ще следваме само наученото, което може и да не е оптимално.

Как влияе времеделенето при експерименталната част? Времеделенето при вариант 1 и вариант 2 са еднакви като значение но с различни стойности. Първото време определено с константа STEP влияе до колко е сигурно активирането на WTA кръга. Ако е малко и недостатъчно, агентът ще избира винаги едно и също действие. Ако е много голямо, тогава ще чакаме много при изпълнение на експериментите. Второто време влияе до толкова до колкото да се обучат допаминовите синапси. Ако е много голямо, тогава стойностите бързо ще се наситят до Wmax на повечето синапси. Ако е много малко, тогава обучението ще е много дълго. То допринася към коефициента за обучение, който не е в явен вид.

Хиперпараметрите на обучението на подобна система са много на брой, над 20 в случая и изпробването на всички комбинации не е възможно, а и не е цел на настоящия труд. Достатъчно беше да се покаже поне един работещ вариант и през какви стъпки, със съответните обяснения, се преминава.

# 6. Идеи за бъдещо развитие и подобрения

Като идеи за бъдещо развитие мога да посоча подобряване на обучението и премахване на времеделенето от Фиг.6.2.2. Това времеделене е сложено заради техническа трудност да обучим само синапсите, които искаме.

Друг вариант за подобрение е да се експериментира с хиперпараметрите от вариант 1 и вариант 2 и търсене на подобрение.

Може да се експереминентира и с различна по структура мрежа с включване на Go и NoGo гейтове от базалните ганглии, които умишлено съм пропуснал поради високата сложност и липса на достатъчно източници и литература с работещи решения.

# 7. Източници и използвана литература

[1] Sutton R., Barto A. (2018), Reinforcement Learning: An Introduction, The MIT Press, <http://www.incompleteideas.net/book/the-book-2nd.html>

[2] O’Reilly R. et al. (2020), Computational Cognitive Neuroscience, Open Textbook, freely available, <https://compcogneuro.org/>

[3] Izhikevich E (2005), Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting, The MIT Press, (<https://www.izhikevich.org/publications/dsn.pdf>)

[4] CartPole, OpenGym, <https://www.gymlibrary.dev/environments/classic_control/cart_pole/>

[5] NEST simulator, <https://nest-simulator.readthedocs.io/en/v3.3/index.html>

[6] Gerstner W, Kistler M, Naud R, Paninski L. (2014), Neuronal Dynamics, Cambridge university press, <https://neuronaldynamics.epfl.ch/online/index.html>

[7] Potjans W., Morrison A., Diesmann M. (2010), Enabling functional neural circuit simulations with distributed computing of neuromodulated plasticity, Frontiers in COMPUTATIONAL NEUROSCIENCE.

[8] Doya, K. (2000), Complementary roles of basal ganglia and cerebellum in

learning and motor control, Current Opinion in Neurobiology 2000, 10:732–739, Elsevier Science Ltd

[9] Koprinkova-Hristova, P.(2020), Brain-Inspired Spike Timing Model of Dynamic Visual Information Perception and Decision Making with STDP and Reinforcement Learning, Springer Nature Switzerland AG

[10] Florian, R. (2007), Reinforcement Learning Through Modulation of Spike-Timing-Dependent Synaptic Plasticity

[11] Jian Xu, How to Beat the CartPole Game in 5 Lines, <https://towardsdatascience.com/how-to-beat-the-cartpole-game-in-5-lines-5ab4e738c93f>

# Приложения

## 1. Сорс код (Source code)

<https://github.com/borkox/uni-sofia-masters-project>