

UNIVERZA V LJUBLJANI  
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matjaž Pogačnik

**Spodbujevano učenje na impulznih  
nevronskih mrežah**

DIPLOMSKO DELO

UNIVERZITETNI ŠTUDIJSKI PROGRAM  
PRVE STOPNJE  
RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: prof. dr. Zoran Bosnić

Ljubljana, 2026

To delo je ponujeno pod licenco *Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji 2.5 Slovenija* (ali novejšo različico). To pomeni, da se tako besedilo, slike, grafi in druge sestavine dela kot tudi rezultati diplomskega dela lahko prosto distribuirajo, reproducirajo, uporabljajo, priobčujejo javnosti in predelujejo, pod pogojem, da se jasno in vidno navede avtorja in naslov tega dela in da se v primeru spremembe, preoblikovanja ali uporabe tega dela v svojem delu, lahko distribuira predelava le pod licenco, ki je enaka tej. Podrobnosti licence so dostopne na spletni strani [creativecommons.si](http://creativecommons.si) ali na Inštitutu za intelektualno lastnino, Streliška 1, 1000 Ljubljana.



Izvorna koda diplomskega dela, njeni rezultati in v ta namen razvita programska oprema je ponujena pod licenco GNU General Public License, različica 3 (ali novejša). To pomeni, da se lahko prosto distribuira in/ali predeluje pod njenimi pogoji. Podrobnosti licence so dostopne na spletni strani <http://www.gnu.org/licenses/>.

*Besedilo je oblikovano z urejevalnikom besedil L<sup>A</sup>T<sub>E</sub>X.*

**Kandidat:** Matjaž Pogačnik

**Naslov:** Spodbujevano učenje na impulznih nevronske mrežah

**Vrsta naloge:** Diplomski naloga na univerzitetnem programu prve stopnje  
Računalništvo in informatika

**Mentor:** prof. dr. Zoran Bosnić

**Opis:**

Kandidat naj preuči principe impulznih nevronske mrež in pristope spodbujevanega učenja, primernih za takšne modele. Razvije in implementira izbran algoritem učenja ter ga eksperimentalno ovrednoti na izbranem problemskem okolju. Rezultate naj analizira in primerja z obstoječimi pristopi.



# Kazalo

Povzetek

Abstract

<b>1</b>	<b>Uvod</b>	<b>1</b>
<b>2</b>	<b>Pregled področja in sorodnih del</b>	<b>5</b>
<b>3</b>	<b>Modeliranje nevronov in sinaps</b>	<b>9</b>
3.1	STDP Sinaptični model . . . . .	19
<b>4</b>	<b>Spodbujevano učenje na impulznih nevronske mrežah</b>	<b>23</b>
4.1	R-STDP učenje . . . . .	23
4.2	TD učenje in model akter-kritik . . . . .	34
<b>5</b>	<b>Implementacija in uporabljena orodja</b>	<b>51</b>
<b>6</b>	<b>Možne razširitve</b>	<b>53</b>
6.1	Izognoitveno obnašanje ( <i>angl. aversive behaviour</i> ) . . . . .	53
6.2	Rekurenčne povezave . . . . .	54
<b>7</b>	<b>Zaključek</b>	<b>57</b>
<b>8</b>	<b>Ekstra</b>	<b>59</b>
	<b>Viri</b>	<b>61</b>



# Seznam uporabljenih kratic

kratica	angleško	slovensko
<b>SNN</b>	Spiking neural network	Impulzna nevronska mreža
<b>R-STDP</b>	Reward modulated spike timing dependent plasticity	Sinaptična plastičnost odvisna od nagrajevanja in časovne razporeditve impulzov
<b>TD</b>	Temporal difference	Časovna razlika





# Povzetek

**Naslov:** Spodbujevano učenje na impulznih nevronske mrežah

**Avtor:** Matjaž Pogačnik

V tem diplomskem delu obravnavamo spodbujevano učenje na impulznih nevronske mrežah, tipu nevronske mreže, ki se obnašajo podobno kot človeški možgani. Predstavimo in rešimo nekatere izzive pri učenju impulznih mrež ter razvijemo inovativne rešitve, ki upoštevajo tako zahtevnost simulacije kot biološko smiselno. Razvijemo sistem, v katerem s prilagoditvijo klasične R-STDP sinapse omogočimo učinkovito učenje in dodeljevanje zaslug preteklim odločitvam, pri tem pa ohranjamo osnovni princip R-STDP, kjer sinapse kodirajo vpliv pretekle aktivnosti nevronov na izbrane akcije, brez uvedbe negativnih nagrad ali nerealističnih mehanizmov.

Razviti sistem nato razširimo v impulzno nevronske akter-kritik sistem, ki lahko rešuje tudi probleme z zakasnjnimi nagradami, pri čemer se pričakovana nagrada prenaša iz cilja nazaj v prejšnja stanja. Sistem najprej ovrednotimo na poenostavljenih, nato na bolj kompleksnih problemih, kot sta igra Pong in problem mrežnega sveta (gridworld), kjer se mreža uči optimalnih strategij in propagacije nagrade v diskretnih okoljih. V igri Pong rezultati razvitega sistema kažejo na postopno daljše sekvence igranja brez zgrešitve žogice, medtem ko se v nalogi mrežnega sveta postopno izboljšuje strategija in krajša pot do cilja.

**Ključne besede:** impulzne nevronske mreže, spodbujevano učenje, R-STDP učenje, TD učenje.



# Abstract

**Title:** Reinforcement learning on spiking neural networks

**Author:** Matjaž Pogačnik

In this diploma thesis, we study reinforcement learning in spiking neural networks, a type of neural network that behaves similarly to the human brain. We address key challenges related to learning in spiking networks and develop innovative solutions that consider both the computational complexity of simulation and biological plausibility. By modifying the classical R-STDP synapse, we develop a system that enables efficient learning and credit assignment to past decisions while preserving the core principle of R-STDP, where synapses encode the influence of past neural activity on selected actions, without introducing negative rewards or biologically implausible mechanisms.

The developed system is then extended into a spiking neural actor-critic system capable of solving problems with delayed rewards, where the expected reward is propagated from the goal state back to preceding states. The system is evaluated on simplified tasks and on more complex problems, such as the game Pong and the gridworld task, in which the network learns optimal strategies and reward propagation in discrete environments. In Pong, the results show progressively longer sequences of play without missing the ball, while in the gridworld task the strategy gradually improves and the path to the goal becomes shorter.

**Keywords:** spiking neural networks, reinforcement learning, R-STDP learning, TD learning.



# Poglavje 1

## Uvod

Impulzne nevronske mreže predstavljajo razred nevronskih mrež, katerih delovanje temelji na diskretnih impulzih in izraziti časovni dinamiki. Za razliko od klasičnih umetnih nevronskih mrež, ki informacije obdelujejo z zveznimi aktivacijami v diskretnih slojih, impulzne nevronske mreže temeljijo na asinhronih dogodkih v času, kar jih približa dejanskemu delovanju bioloških možganov. Njihova ključna prednost zato ni zgolj v potencialni računski učinkovitosti, temveč predvsem v možnosti modeliranja znanih bioloških mehanizmov, kot so časovna integracija signalov, sinaptična plastičnost in nevromodulacija.

Raziskovanje impulznih nevronskih mrež je zato smiselno ne le z vidika umetne inteligence, temveč tudi z vidika računske nevroznanosti. Takšni modeli omogočajo preučevanje, kako lahko iz lokalnih pravil učenja, osnovanih na aktivnosti posameznih nevronov in sinaps, ter globalnih nagradnih signalov vznikne smiselno vedenje. S tem se umetna inteligenca približa razumevanju učenja v vedenjskem smislu, kjer sistem ne optimizira vnaprej znane funkcije, temveč se skozi interakcijo z okoljem postopoma prilagaja in oblikuje strategije delovanja.

Posebej pomemben okvir za takšno učenje predstavlja spodbujevano učenje, ki je v bioloških sistemih tesno povezano z delovanjem dopaminskega sistema. Medtem ko so algoritmi spodbujevanega učenja v klasičnem strojnem

učenju dobro uveljavljeni, njihova neposredna uporaba v impulznih nevronskih mrežah ni mogoča zaradi drugačne narave signalov, izrazite časovne odvisnosti in zahtev po biološki verjetnosti. To odpira vprašanje, kako zasnovati učne mehanizme, ki so hkrati učinkoviti pri reševanju nalog in skladni z znanimi procesi v možganih.

Osrednji problem, ki ga obravnavamo v tem delu, je učenje v impulznih nevronskih mrežah v okoljih s takojšnjimi in zakasnjjenimi nagradami. Enostavnejši mehanizmi, kot je sinaptična plastičnost, odvisna od nagrajevanja in časovne razporeditve impulzov (angl. R-STDP), omogočajo učenje v primerih, kjer nagrada sledi akciji neposredno, vendar odpovejo pri nalogah, kjer je nagrada časovno oddaljena. S tem se pojavi problem časovne dodelitve zaslug (angl. credit assignment), ki predstavlja enega ključnih izzivov spodbujevanega učenja v biološko vernih modelih.

V tej diplomski nalogi se tega problema lotimo postopno. Najprej v poglavju 3 predstavimo in ovrednotimo različne modele nevronov in sinaps, ki služijo kot temelj za nadaljnji razvoj učnih mehanizmov. Na tej osnovi v poglavju 4.1 razvijemo sistem spodbujevanega učenja, ki temelji izključno na impulznih nevronskih mrežah. S prilagoditvijo klasične R-STDP sinapse omogočimo učinkovitejše dodeljevanje zaslug preteklim odločitvam, pri tem pa ohranimo osnovni princip lokalnega učenja. Pomembna značilnost pristopa je uporaba izključno nenegativnih nagradnih signalov, saj v bioloških dopaminskih sistemih negativni dopamin ne obstaja; znižanje dopaminske aktivnosti predstavlja odsotnost ali zmanjšanje pričakovane nagrade, ne pa ločen negativni signal. To razlikuje naš pristop od številnih obstoječih metod, ki uporabljajo eksplicitne negativne nagrade.

Ker opisani pristop še vedno ne omogoča učinkovitega reševanja nalog z daljšo časovno odvisnostjo nagrad, v poglavju 4.2 razviti sistem razširimo z učenjem na podlagi časovne razlike (angl. temporal-difference learning, TD) v impulzno nevronske arhitekture akter–kritik, navdihnjeno z dopaminskimi vezji bazalnih ganglijev. V takšnem sistemu akter skrbi za izbiro akcij, medtem ko kritik ocenjuje pričakovano prihodnjo nagrado posameznih stanj in

generira učni signal za prilagajanje sinaps. Ta razširitev omogoča propagacijo pričakovane nagrade iz ciljnih stanj nazaj v prejšnja stanja ter s tem učinkovito učenje v okoljih z zakasnjnimi nagradami. Delovanje razširjenega sistema ovrednotimo na problemu mrežnega sveta (gridworld), kjer se sistem postopno uči boljših strategij in krajših poti do cilja.

V ta namen v poglavju 3 najprej predstavimo in ovrednotimo različne modele nevronov in sinaps ter njihove lastnosti. Nato v poglavju 4.1 razvijemo sistem, ki temelji na prilagojeni R-STDP sinapsi in omogoča učinkovitejše dodeljevanje zaslug preteklim odločitvam v nalogah s takojšnjimi nagradami, kar prikažemo na nalogi igranja igre Pong. Ker takšen pristop še ne omogoča učenja v okoljih z daljšo časovno odvisnostjo nagrad, v poglavju 4.2 sistem razširimo z učenjem na podlagi časovne razlike (TD) v impulzno nevronske arhitekture akter–kritik, navdihnjeno z dopaminskimi vezji bazalnih ganglijev. Razširjeni sistem omogoča propagacijo pričakovane nagrade v prejšnja stanja in je ovrednoten na problemu mrežnega sveta, kjer se agent postopno uči učinkovitejše strategije in krajše poti do cilja.





## Poglavje 2

# Pregled področja in sorodnih del

Raziskave impulznih nevronske mreže so se v zadnjih desetletjih razvijale predvsem na presečišču umetne inteligence, nevroznanosti in nevromorfnega inženirstva. Osrednji izziv na tem področju predstavlja učenje, saj klasični gradientni pristopi, ki se uporabljajo pri umetnih nevronske mrežah, niso neposredno uporabni zaradi diskretne narave impulzov in nelinearne časovne dinamike.

Eden temeljnih pristopov k učenju v impulznih nevronske mrežah je sinaptična plastičnost, odvisna od časovne razlike med impulzi pre- in postsinaptičnih nevronov (angl. spike-timing-dependent plasticity, STDP). Razširitve tega mehanizma z globalnim nagradnim signalom, kot je R-STDP, omogočajo uporabo impulznih nevronske mreže v okviru spodbujevanega učenja. Takšni pristopi so uspešni pri nalogah s takojšnjimi nagradami, vendar se soočajo z omejitvami pri reševanju problemov z zakasnjjenimi nagradami, kjer je potrebna časovna propagacija učnega signala.

Na področju spodbujevanega učenja obstaja obsežna literatura, ki obravnava različne naloge, kot so navigacija, vodenje robotov in igranje iger. Dela, kot je Dobrevski M, Skočaj D 2021, obravnavajo uporabo spodbujevanega učenja v robotskih sistemih, kjer agent uči strategije na podlagi interakcije z

realnim, šumnim okoljem. Podobni pristopi so uporabljeni tudi v simuliranih okoljih, kjer se raziskujejo lastnosti učnih algoritmov, na primer pri problemu vozička s palico Svete A 2020 ali pri igranju iger Šutar M 2023. V teh delih so uporabljeni predvsem klasični modeli nevronske mreže ali simbolni opisi stanj, časovna dinamika pa ni eksplicitno modelirana na ravni posameznih dogodkov.

Impulzne nevronske mreže se od rekurentnih arhitektur, kot so LSTM ali GRU, razlikujejo po tem, da čas ni implicitno kodiran v stanju mreže, temveč je neposredno prisoten v obliki časovnih zamikov med impulzi. To omogoča naravno obdelavo dogodkovno vodenih in asinhronih signalov, hkrati pa odpira možnost energetsko učinkovite implementacije na nevromorfni strojni opremi. Pri tem računanje ne temelji na zaporednem množenju matrik, temveč na redkih dogodkih, kjer se ob pojavu impulza posodobi le del mreže.

Posebno zanimiva smer raziskav je povezovanje impulznih nevronske mreže z učenjem na podlagi časovne razlike. V delu Wunderlich T, et al. 2019 je predstavljena uporaba TD učenja na trdo-ožičeni (nevromorfni) impulzni nevronske mreži, kjer je končna naloga igranje igre Pong. Trdo-ožičena implementacija pomeni, da je mreža realizirana na specializirani strojni opremi, kjer so nevroni in sinapse fizično implementirani, kar omogoča visoko energetsko učinkovitost, hkrati pa omejuje fleksibilnost modela.

Biološko bolj neposredna implementacija TD učenja je arhitektura akter-kritik, ki je navdihnjena z delovanjem bazalnih ganglijev in dopaminskega sistema v možganih Wiebke P, et al. 2011. V tem okviru kritik ocenjuje vrednost stanj in generira dopaminski signal, ki predstavlja napako napovedi nagrade, medtem ko akter na tej osnovi prilagaja strategijo izbire akcij. Takšen nagradni sistem je tesno povezan z eksperimentalnimi ugotovitvami o delovanju dopaminskih nevronov, ki kodirajo razliko med pričakovano in dejansko nagrado.

V primerjavi z obstoječimi deli se ta diplomska naloga osredotoča na enotno impulzno nevronske arhitekturo, ki združuje lokalna pravila učenja, globalne nagradne signale in biološko smiselno implementacijo brez uporabe

---

eksplicitnih negativnih nagrad. Poudarek je na časovni dodelitvi zaslug in postopni razširitvi osnovnega R-STDP pristopa v arhitekturo akter–kritik, ki omogoča učenje v okoljih z zakasnjnimi nagradami.

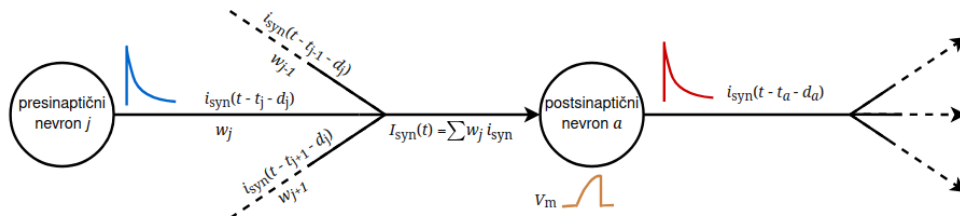


## Poglavje 3

# Modeliranje nevronov in sinaps

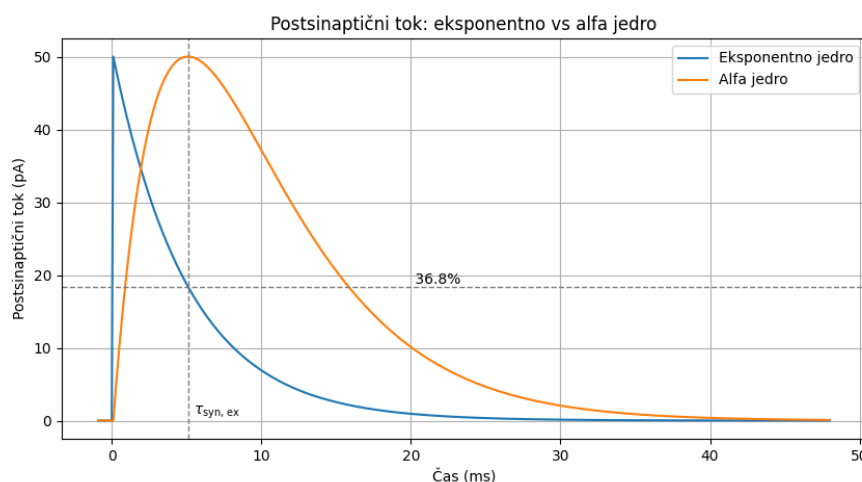
Impulzne nevronske mreže so določene z modelom nevrona in modelom sinapse, ki povezuje nevrone. Obstaja veliko modelov, v nadaljevanju pa bosta predstavljena in primerjana dva modela nevronov glede na njuno uporabnost pri spodbujanem učenju na impulznih nevronskih mrežah. Predstavljen bo tudi model sinapse, primeren za spodbujano učenje, ki bo uporabljen v sistemih razvitih v nadaljevanju.

Nevronski modeli opisujejo električne lastnosti celične membrane nevrona v možganih. Nevroni bodo prek sinaps sprejemali izhodne (postsinaptične) tokove nevronov, s katerimi so povezani, in skozi čas glede na utež sinapse posodabljali svoj membranski potencial. Tok, ki prek sinapse s pozitivno utežjo od nevrona na začetku sinapse (presinaptičnega nevrona) prihaja do nevrona na koncu (postsinaptičnega nevrona), povzroči zvišanje membranskega potenciala. Ko membranski potencial nevrona preseže vrednost  $V_{th}$ , se sproži impulz, pri čemer ta nevron sprosti svoj postsinaptični tok na sinapso. Ta sistem je prikazan na sliki 3.1, kjer so predstavljeni postsinaptični tokovi presinaptičnih nevronov  $i_{syn}(t - t_j - d_j)$ . Indeks  $j$  označuje posamezne presinaptične nevrone,  $t$  predstavlja trenutni čas,  $t_j$  čas sprožitve akcijskega potenciala presinaptičnega nevrona,  $d_j$  pa zakasnitev signala zaradi njegovega prenosa prek sinapse do postsinaptičnega nevrona.



Slika 3.1: Prikaz pre- in postsinaptičnih razmerij ter posameznih postsinaptičnih tokov.

Vrednost membranskega potenciala se ne glede na vhodne tokove skozi čas zmanjšuje glede na uhajalsko prevodnost  $g_L$ . Takim nevronske modelom pravimo tokovno gnani modeli uhajajočega integrirajočega nevrona (*angl. leaky integrate-and-fire model* ali *leaky IAF*). Po sprožitvi impulza se velikost postsinaptičnega toka spreminja po krivulji, ki jo določa izbrano jedro modela in predstavlja obliko toka po impulzu. V nadaljevanju bomo predstavili dva pristopa: enostavnejši model z eksponentno oblikovanim postsinaptičnim tokom in kompleksnejši model z alfa oblikovanimi postsinaptičnimi tokovi. Obe obliki sta pri enakih parametrih prikazani na sliki 3.2 in podrobneje, skupaj s parametri, opisani v nadaljevanju. Označenih je tudi 36,8% maksimalne vrednosti, ki jo, po izračunih v poglavju 3.0.1, eksponentno jedro doseže ravno pri  $t = \tau_{\text{syn, ex}}$ , času, ko alfa jedro doseže ravno maksimalno vrednost.



Slika 3.2: Postsinaptični tok modela z eksponentnim in alfa jedrom pri sinaptični uteži  $w = 50$  in  $\tau_{syn, ex} = 5$  ms. Eksponentno jedro povzroči takojšen skok postsinaptičnega toka ob sprožitvi impulza in eksponentno odtekanje, medtem ko alfa jedro modelira postopen dvig toka do maksimuma in nato počasnejši upad, kar bolj odraža časovno dinamiko bioloških sinaps.

Membranski potencial leaky IAF nevrona (*LIF nevron*) se spreminja glede na ravnovesje med kapacitivnostjo in uhajanjem prek membranske prevodnosti, vhodne tokove  $I_{syn}$  ter zunanji šum  $I_e$ . Model membrane, ki določa, kako se spreminja membranski potencial, je definiran z naslednjimi parametri.

- $E_L$  — **mirujoči membranski potencial**  
Električni potencial, na katerega se membranski potencial relaksira v odsotnosti vhodnih tokov;
- $C_m$  — **membranska kapacitivnost**  
Kapacitivnost membrane, ki določa, kako hitro se membranski potencial odziva na vhodne tokove;
- $\tau_m$  — **membranska časovna konstanta**  
Čas, v katerem membrana pasivno integrira tok; definiran kot razmerje

med kapacitivnostjo  $C_m$  in uhajalsko prevodnostjo  $g_L$  (*leakage conductance*). Konstanto  $\tau_m$  lahko definiramo tudi kot produkt med kapacitivnostjo in uporom membrane  $\tau_m = C_m R_m = \frac{C_m}{g_L}$ ;

- $t_{ref}$  — **refrakcijsko obdobje**

Čas, v katerem se nevron po sprožitvi akcijskega potenciala ne more ponovno prožiti;

- $V_{th}$  — **prag proženja**

Membranski potencial, pri katerem nevron sproži akcijski potencial;

- $V_{min}$  — **spodnja meja membranskega potenciala**

Absolutna spodnja meja za membranski potencial;

- $I_e$  — **zunanji konstantni tok**

Dodani tok, ki modelira stalni zunanji šum.

Membranski potencial  $V_m$  pri tokovno gnanem modelu uhajajočega integrirajočega nevrona je v odvisnosti od prej navedenih parametrov opisan z diferencialno enačbo

$$\frac{dV_m}{dt} = -\frac{V_m - E_L}{\tau_m} + \frac{I_{syn} + I_e}{C_m}. \quad (3.1)$$

Prvi člen na desni strani enačbe opisuje pasivno uhajanje membranskega potenciala proti mirujoči vrednosti  $E_L$  s časovno konstanto  $\tau_m$ , ki določa hitrost relaksacije membrane. Drugi člen predstavlja vpliv vhodnih tokov, kjer enako kot doslej  $I_{syn}$  označuje skupni sinaptični tok, ki ga ustvarjajo vsi presinaptični nevroni,  $I_e$  pa zunanji konstantni tok oziroma šum. Membranska kapacitivnost  $C_m$  določa, kako močno posamezni tokovi vplivajo na spremembo membranskega potenciala. Skupni tok  $I_{syn}$ , ki ga postsinaptični nevron prejme prek vseh sinaps, lahko razdelimo na dve komponenti glede na vrednosti uteži sinaps. Če je utež sinapse pozitivna, bo membranski potencial glede na postsinaptični tok naraščal proti pragu proženja. Pravimo, da je takšna povezava vzbujajoča. Obratno, če je utež sinapse negativna, bo membranski potencial padal. Takšni povezavi pravimo inhibitorna povezava.



Notacijo  $I_{\text{syn}}$  in  $i_{\text{syn}}$  bomo v nadaljevanju razdelili na  $i_{\text{syn, ex}}$  in  $i_{\text{syn, in}}$ , skupno zapisano z  $i_{\text{syn, X}}$ , kjer je  $X \in \{\text{ex, in}\}$ ,  $I_{\text{syn}}$  pa bomo zapisali kot vsoto  $I_{\text{syn, ex}}$  in  $I_{\text{syn, in}}$ . Tak splošnejši zapis omogoča uporabo različnih sinaptičnih časovnih konstant, definiranih v nadaljevanju, za vzbujaajoče in inhibitorne povezave.  $I_{\text{syn}}$  tako zapišemo kot

$$I_{\text{syn}}(t) = I_{\text{syn, ex}}(t) + I_{\text{syn, in}}(t),$$

kjer

$$I_{\text{syn, X}}(t) = \sum_j w_j \sum_k i_{\text{syn, X}}(t - t_j^k - d_j),$$

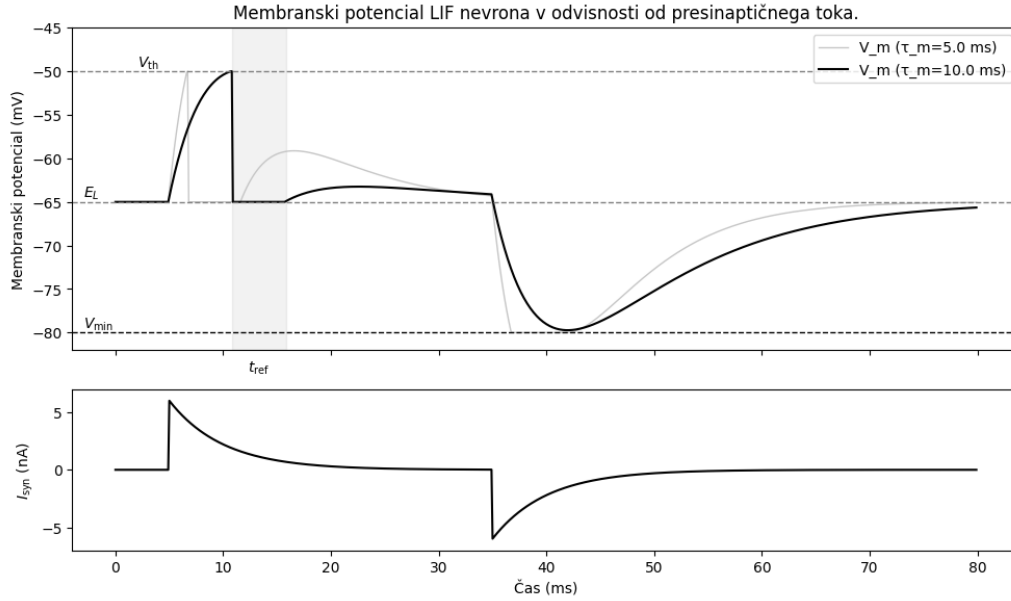
kjer  $j$  teče po vzbujaajočih ( $X = \text{ex}$ ) in inhibitornih ( $X = \text{in}$ ) sinapsah z utežmi  $w_j$  do presinaptičnih nevronov,  $k$  pa po časih impulzov nevrona  $j$ .  $d_j$  predstavlja zakasnitev zaradi potovanja signala po sinapsi do nevrona  $j$ .  $i_{\text{syn, X}}(t - t_j^k - d_j)$  predstavlja postsinaptični tok nevrona  $j$ .

Postsinaptični tokovi so ne glede na jedro določeni z naslednjima parametroma.

- $\tau_{\text{syn, ex}}$  — **sinaptična časovna konstanta (vzbujajoča)**  
Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju. Pri modelu z alfa-jedrom predstavlja čas dviga alfa-funkcije; pri eksponentnem jedru pa čas padca eksponentne funkcije, pri kateri je čas dviga neskončno majhen;
- $\tau_{\text{syn, in}}$  — **sinaptična časovna konstanta (inhibitorna)**  
Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju, vendar za inhibitorne sinapse.

Opisana dinamika je prikazana na sliki 3.3, kjer je membranski potencial leaky IAF nevrona (*LIF nevrona*) prikazan kot odziv na skupni presinaptični tok dveh nevronov: nevrona z vzbujaajočo in nevrona z inhibitorno povezavo,

oba z eksponentno oblikovanim postsinaptičnim tokom. Na grafu je ravnino, kako  $\tau_m$  vpliva na hitrost spremembe membranskega potenciala ter na odtekanje potenciala proti mirujočem  $E_L$  po impulzu in med refrakcijskim obdobjem.



Slika 3.3: Sprememba membranskega potenciala  $V_m$  LIF nevrona (*leaky IAF*) kot funkcija presinaptičnega toka  $I_{syn}$ , ki ga tvorita dva presinaptična nevrona z eksponentnim postsinaptičnim tokom: vzbujaajoči nevron z utežjo  $w_{ex} = 6$  in inhibitorni nevron z utežjo  $w_{in} = -6$ . Ostali parametri so  $\tau_{syn, ex} = \tau_{syn, in} = 5$  ms, zakasnitev  $d_j = 0$  ms,  $V_{th} = -50$  mV,  $E_L = -65$  mV,  $V_{min} = -80$  mV,  $t_{ref} = 5$  ms,  $I_e = 0$  nA in  $C_m = 1$  nF. Prikazana je krivulja za  $\tau_m = 10$  ms in  $\tau_m = 5$  ms za primerjavo vpliva časovne konstante membrane na dinamiko. Vzbujaajoči nevron se proži ob  $t = 5$  ms, inhibitorni pa ob  $t = 35$  ms.

### 3.0.1 Model z eksponentnim jedrom

V simulatorju NEST je model z eksponentnim jedrom (*iaf\_psc\_exp*) definiran s sistemom diferencialnih enačb prvega reda Tsodyks, Uziel in Markram 2000. Postsinaptični tok  $i(t)$  se spreminja po sistemu

$$\frac{dx}{dt} = \frac{z}{\tau_{rec}} - ux\delta(t - t_{sp}) \quad (3.2)$$

$$\frac{di}{dt} = -\frac{i}{\tau_{syn, X}} + ux\delta(t - t_{sp}) \quad (3.3)$$

$$\frac{dz}{dt} = \frac{i}{\tau_{syn, X}} - \frac{z}{\tau_{rec}}, \quad (3.4)$$

kjer  $t_{sp}$  predstavlja čas presinaptičnega impulza,  $\tau_{rec}$  čas povrnitve sinaptičnih virov,  $u$  delež sinaptičnih virov, porabljenih pri impulzu, in  $\delta(t - t_{sp})$  delta porazdelitev za instantne posodobitve ob impulzih.

Preverimo, ali postsinaptični tok po impulzu res sledi preprosti eksponentni funkciji. Če opazujemo samo spreminjanje  $i(t)$  skozi čas brez novih impulzov, velja  $\delta(t - t_{sp}) = 0$  in se diferencialna enačba za  $i$  poenostavi v

$$\frac{di}{dt} = -\frac{i}{\tau_{syn, X}}. \quad (3.5)$$

Rešitev te diferencialne enačbe je tako

$$i(t) = i_0 e^{-t/\tau_{syn, X}}, \quad (3.6)$$

kjer vidimo, da je jedro res eksponentna funkcija z začetkom v  $i_0$ . Skok potenciala po impulzu je določen z utežjo sinapse  $w$ , zato je  $i_0 = w$ , postsinaptični tok pa je določen z vrednostjo  $\tau_{syn, X}$ .

Zdaj lahko izračunamo količino naboja, ki ga po impulzu prenesemo po sinapsi. Ta nam bo koristila pri primerjavi in izbiri ustreznega modela za sisteme, razvite v nadaljevanju (poglavje 3.0.3). Količina naboja  $q$ , ki se po sprožitvi impulza prenese prek sinapse, je definirana kot ploščina pod krivuljo postsinaptičnega toka

$$q = \int_0^\infty i_{syn, X}(t) dt = \tau_{syn, X} \cdot w.$$

Na sliki 3.2 je poleg parametra  $\tau_{\text{syn, ex}}$  označenih tudi 36,8% maksimalnega postsinaptičnega toka. Pri eksponentnem jedru bo namreč postsinaptični tok to vrednost dosegel natanko pri  $\tau_{\text{syn, ex}}$ , kar smo izračunali po naslednji enačbi.

$$i_{\text{syn, ex}}(t) = we^{-\frac{t}{\tau_{\text{syn, ex}}}}$$

$$i_{\text{syn, ex}}(\tau_{\text{syn, ex}}) = we^{-1} \approx 0.3679w.$$

### 3.0.2 Model z alfa jedrom

Model z alfa jedrom je kompleksnejši in biološko bolj realističen model postsinaptičnih tokov. V simulatorju NEST je postsinaptični tok modela z alfa jedrom (*iaf\_psc\_alpha*) definiran kot

$$i_{\text{syn, X}}(t) = \frac{e}{\tau_{\text{syn, X}}} te^{-\frac{t}{\tau_{\text{syn, X}}}} \Theta(t),$$

kjer je  $\Theta(t)$  enotina stopnica, ki zagotavlja, da je postsinaptični tok ničeln za čase pred sprožitvijo impulza. Postsinaptični tok je normaliziran tako, da doseže enotski maksimum ob času  $t = \tau_{\text{syn, X}}$ .

$$i_{\text{syn, X}}(t = \tau_{\text{syn, X}}) = 1.$$

Enačba opisuje časovni potek postsinaptičnega toka po sprožitvi akcijskega potenciala presinaptičnega nevrona. Faktor  $t$  povzroči postopen dvig toka po impulzu, medtem ko eksponentni člen določa njegovo kasnejše zmanjševanje. Posledično postsinaptični tok ne doseže maksimuma takoj ob sprožitvi impulza, temveč šele ob času  $\tau_{\text{syn, X}}$ , kar modelira hitrost odpiranja in zapiranja ionskih kanalov v bioloških sinapsah.

Parameter  $\tau_{\text{syn, X}}$  določa časovno skalo dinamike sinapse: večja vrednost povzroči počasnejši dvig in daljše trajanje postsinaptičnega toka, manjša pa hitrejši in krajši odziv. Takšna oblika postsinaptičnega toka vodi v bolj zglajeno časovno integracijo vhodnih impulzov.

Enako kot pri eksponentnem jedru skupni naboj  $q$ , ki ga prenese postsinaptični tok pri alfa jedru, definiramo kot ploščino pod krivuljo postsin-

naptičnega toka.

$$q = \int_0^{\infty} i_{\text{syn, X}}(t) dt = e\tau_{\text{syn, X}}.$$

### 3.0.3 Izbira modela nevrona

V sistemih, ki jih bomo implementirali v nadaljevanju, skušamo pri modeliranju mehanizmov v človeških možganih uporabiti čim manj poenostavitev ali posplošitev. Za to je bolj primeren model nevrona z alfa jedrom, ki ima biološko bolj realistično obliko postsinaptičnega toka. V nadaljevanju sta kljub temu uporabljena oba modela, saj se zaradi različnih oblik postsinaptičnega toka za spodbujevano učenje bolje obnese model z eksponentnim jedrom.

Za nas je najpomembnejša razlika v količini prenesenega naboja  $q$ . Kot bo opisano v poglavju 4.1, to namreč vpliva na to, koliko lahko zunanji šum vpliva na frekvenco impulzov. Količina prenesenega naboja  $q_{\text{alfa}}$  je pri alfa jedru večja od prenesenega naboja pri eksponentnem jedru  $q_{\text{exp}}$  za faktor  $\frac{q_{\text{alfa}}}{q_{\text{exp}}} = e$ . To razliko lahko prilagodimo z nižjimi vrednostmi uteži sinaps, razlika v vplivu na frekvenco impulzov pa je posledica različno dolgega časovnega intervala, v katerem je postsinaptični tok blizu maksimalne vrednosti. Pri alfa jedru je ta interval večji kot pri eksponentnem jedru, zaradi česar bodo zaporedni postsinaptični impulzi skozi čas precej bolj prekrivni. Pri integriranju različnih postsinaptičnih tokov skozi čas tako pride do učinka nizkoprepustnega filtra, ki ublaži nenadne spremembe v amplitudi skupnega toka na vhodu v postsinaptični nevron. Če se nevron proži z določeno stalno frekvenco, bo ob dodanem šumu varianca v frekvenci impulzov pri alfa jedru manjša kot pri eksponentnem.

Primerjamo varianco frekvence pri obeh jedrih v času 5000 ms preko petih postsinaptičnih nevronov, v katere neodvisno injiciramo Poissonov šum, saj je ta biološko najbolj realističen. Ta namreč neposredno predstavlja impulze nevronov.

$$P(k \text{ impulzov v } \Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \quad k = 0, 1, 2, \dots \quad (3.7)$$

Da dosežemo čim bolj enako osnovno frekvenco impulzov postsinaptičnih nevronov pri obeh jedrih, je utež sinapse med nevroni z alfa jedrom  $w_{\text{alfa}}$  za faktor  $e$  manjša od uteži sinaps do nevronov z eksponentnim jedrom  $w_{\text{eksp}}$ . Vsi parametri simulacije so navedeni v tabeli 3.0.3. Iz rezultatov simulacije, prikazanih v tabeli 3.0.3, pričakovano opazimo večjo varianco pri eksponentnem jedru.

Parameter	Vrednost
Število postsinaptičnih nevronov	5
Trajanje simulacije	5000 ms
$C_m$	250.0 pF
$\tau_m$	20.0 ms
$E_L$	0.0 mV
$V_{\text{th}}$	20.0 mV
$V_{\text{reset}}$	0.0 mV
$t_{\text{ref}}$	2.0 ms
$\tau_{\text{syn,ex}}$	5.0 ms
$w_{\text{eksp}}$	25.0
$w_{\text{alfa}}$	$25.0 / e \approx 9.20$
Frekvenca Poissonovega šuma	8000 Hz na nevron

Tabela 3.1: Parametri simulacije uporabljeni pri primerjavi modelov nevronov.

Jedro	Povprečje (ms)	Varianca (ms <sup>2</sup> )
Exponentno	$7.846 \pm 0.021$	<b><math>0.402 \pm 0.028</math></b>
Alfa	$7.800 \pm 0.023$	<b><math>0.270 \pm 0.006</math></b>

Tabela 3.2: Povzetek statistike medimpulznih intervalov nevronov z alfa in eksponentnim jedrom. Povprečje in standardni odklon sta izračunana na vseh postsinaptičnih nevronih.

### 3.1 STDP Sinaptični model

V sistemih, ki bodo implementirani v tej nalogi bomo uporabljali sinapso s plastičnostjo, odvisno od nagrade in časovne razporeditve impulzov (*angl. reward-modulated-spike-timing-dependent plasticity* ali *R-STDP*). Časovna razporeditev impulzov (*STDP*) prilagaja sinaptične uteži glede na relativni čas impulzov pre- in postsinaptičnih nevronov. V svoji klasični obliki STDP uresničuje Hebbov princip:

“Nevroni, ki se skupaj prožijo, se povežejo.”

Če se presinaptični nevron sproži **pred** postsinaptičnim ( $\Delta t > 0$ ), se sinapsa **okrepi** (potencira). Če se presinaptični nevron sproži **po** postsinaptičnem ( $\Delta t \leq 0$ ), se sinapsa **oslabi** (depresira).

Matematično je to opisano s funkcijo okna STDP:

$$\text{STDP}(\Delta t) = \begin{cases} A_+ e^{-|\Delta t|/\tau_+}, & \text{če } \Delta t > 0 \text{ (presinaptični pred postsinaptičnim)} \\ A_- e^{-|\Delta t|/\tau_-}, & \text{če } \Delta t \leq 0 \text{ (postsinaptični pred presinaptičnim)} \end{cases}$$

kjer so:

- $A_+$  in  $A_-$  multiplikatorja za potenciranje in depresijo,
- $\tau_+$  in  $\tau_-$  časovne konstante, ki določajo okno vpliva časovnih razlik.

#### 3.1.1 Dopaminska modulacija

V vseh sistemih, razvitih v tej diplomski nalogi, kot osnovo uporabljamo model R-STDP sinapse, kot ga definira Izhikevich, E. M. 2007. V tem modelu odvisnost od nagrade vpeljemo z nevromodulatorjem dopaminom. Pogosto se koncentracija dopamina obravnava kot neposreden odraz količine nagrade oziroma vrednosti stanja, v katerem se nahaja agent, vendar dopamin modulira le plastičnost sinaps oziroma učenje. Dopamin je potreben tudi pri učenju izogibanja neželenim stanjem, kot je opisano v poglavju ???. Pri nevromodulirani STDP dopaminsko koncentracijo predstavlja vrednost  $n$ , ki

neposredno modulira velikost in smer sinaptične plastičnosti, tj. velikost in predznak posodobitve uteži povezave. Sinaptična dinamika je opisana z enačbami po Potjans, Morrison in Diesmann 2010:

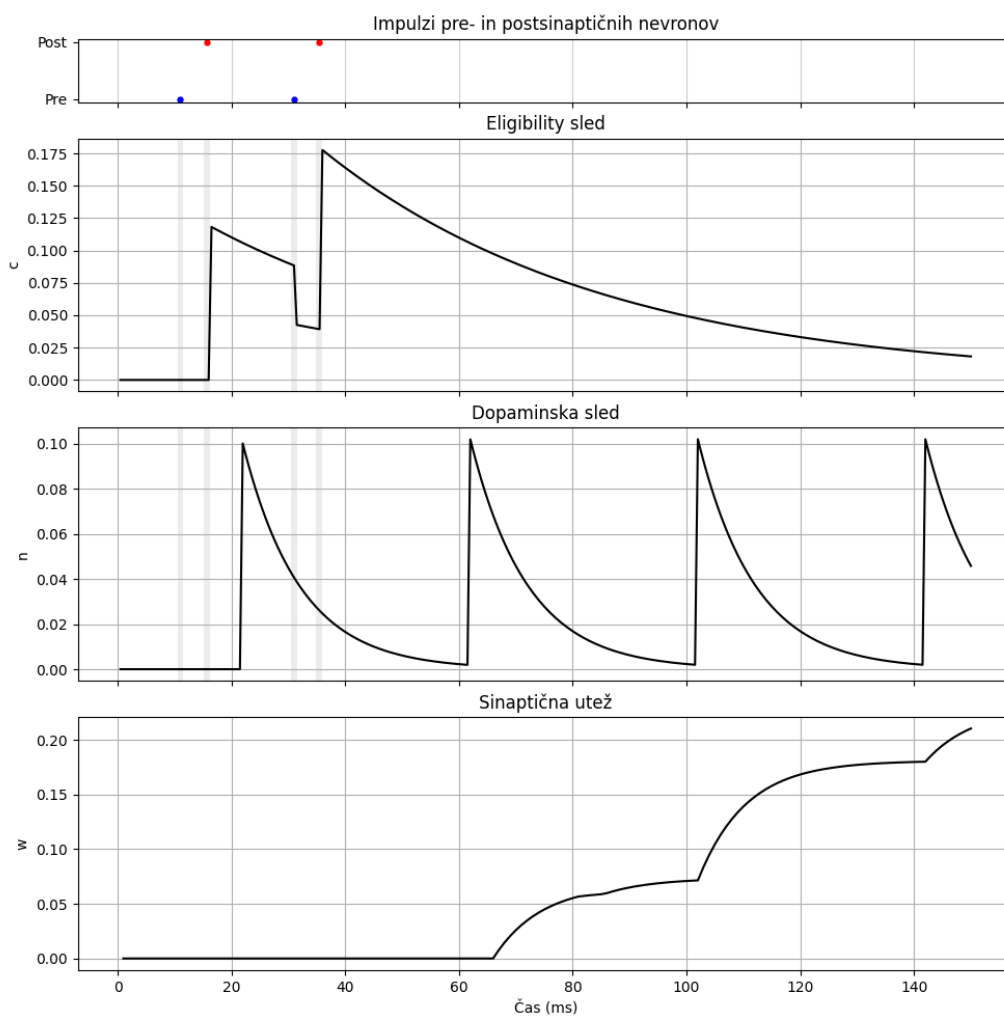
$$\begin{aligned}\dot{w} &= c(n - b) \\ \dot{c} &= -\frac{c}{\tau_c} + \text{STDP}(\Delta t) \delta(t - s_{\text{pre/post}}) C_1 \\ \dot{n} &= -\frac{n}{\tau_n} + \frac{\delta(t - s_n)}{\tau_n} C_2,\end{aligned}$$

kjer so:

- $w$  — sinaptična utež,
- $c$  — *eligibility* sled, ki spremlja pare sproženih pre- in postsinaptičnih nevronov ter aproksimira odgovornost sinapse za proženje postsinaptičnega nevrona kot posledico proženja presinaptičnega nevrona.
- $n$  — dopaminska koncentracija/sled,
- $b$  — bazalna dopaminska koncentracija,
- $s_{\text{pre/post}}$  — čas pre- ali post-sinaptičnega impulza,
- $s_n$  — čas impulzov dopaminskih nevronov,
- $C_1, C_2$  — konstante,
- $\tau_c, \tau_n$  — časovne konstante odtekanja *eligibility* in dopaminskih sledi.

Slika 3.4 prikazuje spreminjanje *eligibility* sledi  $c$  in uteži sinapse  $w$  v odvisnosti od proženja pre- in postsinaptičnega nevrona ter dopaminske sledi  $n$ . Dopaminski nevroni, ki določajo dopaminsko sled, se prožijo na 40ms.





Slika 3.4: *Eligibility* sled  $c$ , dopaminska sled  $n$  in spreminjanje sinaptične uteži pri presinaptičnih impulzih pri  $[10.0, 30.0]$  ms in postsinaptičnih impulzih pri  $[12.0, 32.0]$  ms, simulirane v času 150 ms pri R-STDP sinapsi s  $\tau_c = 50.0$  ms,  $\tau_{c,\text{delay}} = 50.0$  ms,  $\tau_n = 10.0$  ms,  $\tau_{\text{plus}} = 10.0$  ms,  $b = 0.0$ ,  $A_{\text{plus}} = 0.2$ ,  $A_{\text{minus}} = 0.2$  in sinaptično zakasnitvijo 0.5 ms.



## Poglavje 4

# Spodbujevano učenje na impulznih nevronskih mrežah

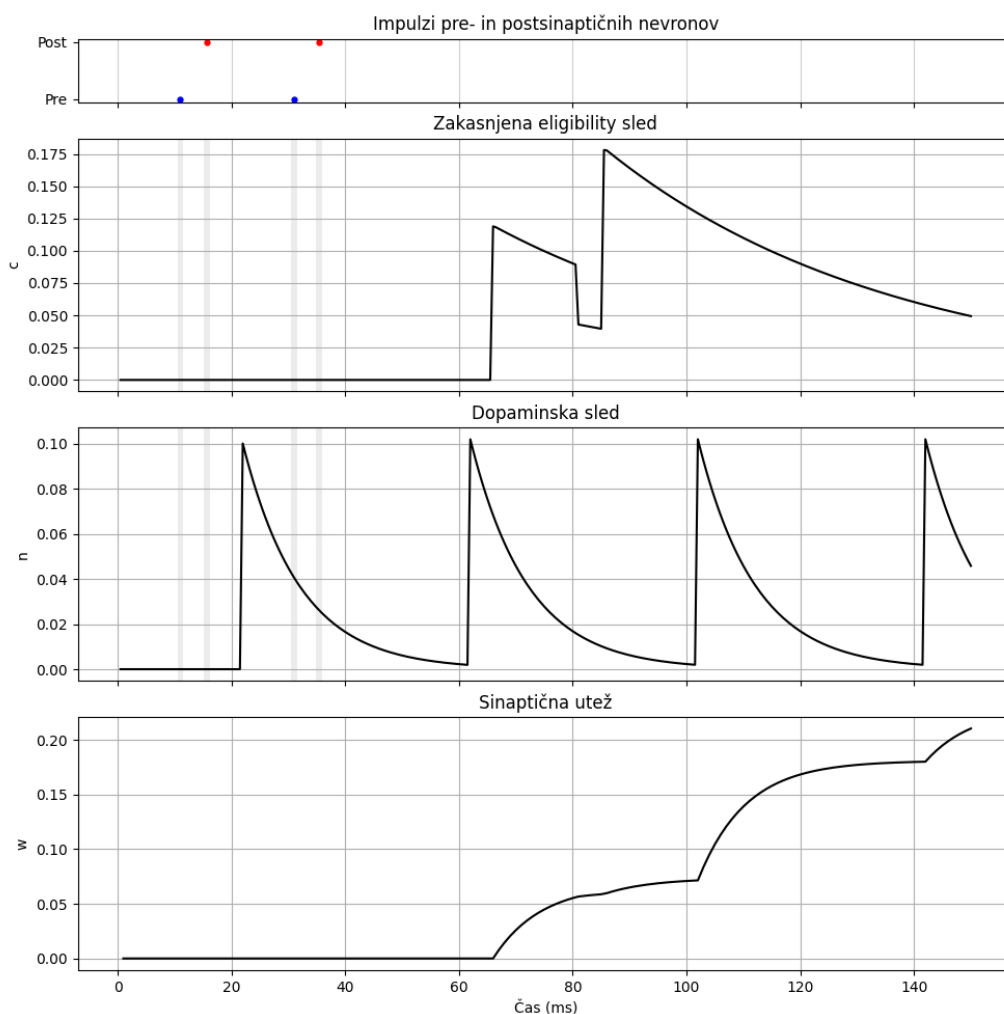
Imamo klasičnega agenta spodbujevanega učenja, ki prejema informacije o zunanjem okolju prek stimulacije vhodnih nevronov, nato pa kot odziv na trenutno stanje izbere akcijo, ki vpliva na okolje. Če se znajde v nagrajenem stanju, agenta nagradimo. S pomočjo nagrajevanja in interakcije z okoljem se agent nauči akcij, ki v določenem stanju privedejo do nagrade.

### 4.1 R-STDP učenje

R-STDP učenje temelji na krepitvi povezav, ki so bile odgovorne za pravilno akcijo agenta v določenem stanju. To dosežemo tako, da prek vseh povezav zvišamo koncentracijo dopamina, pri čimer se najbolj okrepijo tiste povezave, ki so povzročile največ kavzalnih parov pre- in postsinaptičnih impulzov. Te povezave so namreč imele najvišjo vrednost *eligibility* sledi *c*. Agent v trenutnem stanju izbere akcijo, morebitna nagrada pa je voljo šele ob prihodu v naslednje stanje. Zato želimo posodobiti povezave, ki so bile aktivne v prejšnjem stanju in so bile odgovorne za akcijo, ki nas je pripeljala v naslednje stanje.

Naš agent je za začetek sestavljen iz  $N_{in}$  vhodnih nevronov, ki predstavljajo možna stanja in so povezani z  $N_a$  nevroni na izhodu. Vhod in izhod sta povezana po režimu *all-to-all*, kjer so vsi vhodni nevroni povezani z izhodnimi nevroni. Ob prihodu v določeno stanje ustrezni vhodni nevron stimuliramo tako, da ta se ta za čas 200 ms proži s frekvenco 100 Hz. Akcijo izberemo na koncu intervala glede na aktivnost izhodnih nevronov, ki predstavljajo možne akcije. Med njimi izberemo nevron, ki se je v trenutnem stanju največkrat prožil. Če vstopimo v nagrajeno stanje, bomo  $N_{dopa}$  dopaminskih nevronov stimulirali s tokom 600 pA. Dopaminski nevroni ob impulzu enakomerno projicirajo dopamin med vse povezave med vhodnimi in izhodnimi nevroni.

Pri zvišani koncentraciji dopamina ob prihodu v nagrajeno stanje bi lahko poleg zelenih povezav posodabljali že povezave, ki so aktivne v novem stanju. Da se temu izognemo, bomo onemogočili posodabljanje povezav, za katere je nagrada prišla prehitro. R-STDP sinapso bomo zato prilagodili tako, da bomo celotno *eligibility* sled v času premaknili za vrednost  $\tau_{c,delay}$  in s tem onemogočili posodabljanje zaradi nagrad, ki so prispele v času krajšem od  $\tau_{c,delay}$  po aktivnosti sinapse. Dinamika prilagojene sinapse je prikazana na sliki 4.1.



Slika 4.1: *Eligibility* sled  $c$ , dopaminska sled  $n$  in posodabljanje sinaptične uteži pri presinaptičnih impulzih pri  $[10.0, 30.0]$  ms in postsinaptičnih impulzih pri  $[12.0, 32.0]$  ms, simulirane v času 150 ms pri R-STDP sinapsi s  $\tau_c = 50.0$  ms,  $\tau_{c,\text{delay}} = 50.0$  ms,  $\tau_n = 10.0$  ms,  $\tau_{\text{plus}} = 10.0$  ms,  $b = 0.0$ ,  $A_{\text{plus}} = 0.2$ ,  $A_{\text{minus}} = 0.2$  in sinaptično zakasnitvijo 0.5 ms.

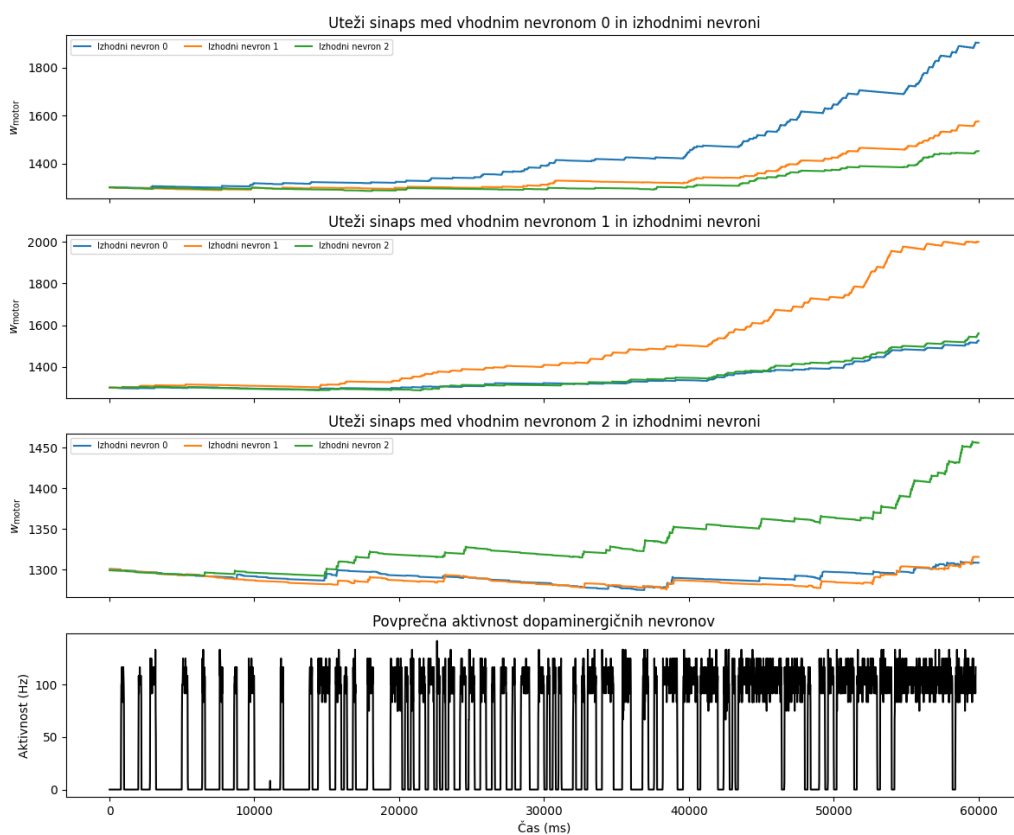
Nagrada, ki jo predstavlja aktivnost dopaminskih nevronov, bo vedno večja ali enaka 0, kar pomeni, da se bodo sinapse skozi čas le krepile. Povezave, odgovorne za izbiro določene akcije v določenem stanju, morajo zato med seboj tekmovati za prevlado. Pri tem moramo omogočiti dovolj veliko

varianco frekvence impulzov izhodnih nevronov, predvsem v začetni fazi, ko imajo vse povezave približno enake uteži. V nasprotnem primeru bodo vse povezave posodobljene za približno enako vrednost glede na R-STDP. Varianco med impulzi pri enakih povezavah dosežemo z zunanjim Poissonovim šumom. Naš agent bo uporabljal model nevrona z eksponentnim jedrom, saj Poissonov šum v tem primeru povzroči večjo varianco izhodnih nevronov kot model z alfa jedrom, kar smo pokazali v poglavju 3.0.3. V začetni fazi bodo tako akcije večinoma izbrane naključno, ob majhnem številu izhodnih impulzov pa bo razlika variance relativno večja kot pri višji aktivnosti izhodnih nevronov. Tako bo v kasnejših fazah učenja izbira akcije vse manj odvisna od šuma.

Za začetek preverimo R-STDP učenje in obnašanje razvitega sistema na preprosti nalogi s tremi stanji. Prehod v vsako stanje je naključen, v vsakem stanju pa je nagrajena le ena izbira akcije. V stanju 0, ki ga predstavlja vhodni nevron 0, je nagrajena akcija 0 (izhodni nevron 0), v stanju 1 akcija 1, v stanju 2 pa akcija 2. Na sliki 4.1 je razvidna prevlada pravih sinaps, višanje divergence v sinapsah skozi čas ter rast povprečne nagrade tekom učenja. V simulaciji uporabljamo privzete NEST parametre za nevrone tipa *iaf-psc-exp* ter zakasnjene dopaminsko modulirane sinapse. Ostali parametri sistema so navedeni v tabeli 4.1.

Simbol	Pomen	Vrednost
$t_{\text{SIM}}$	Trajanje simulacije	60000 ms
	Minimalna utež sinaps	
$w_{\text{motor, min}}$	med vhodnimi in izhodnimi nevroni	500
	Maksimalna utež sinaps	
$w_{\text{motor, max}}$	med vhodnimi in izhodnimi nevroni	2000
$\tau_c$		5 ms
$\tau_{c,\text{delay}}$		200 ms
$\tau_n$	Odtekanje dopaminske sledi	10 ms
$\tau_+ = \tau_-$	Pozitivna STDP konstanta	20 ms
$b$	Bazalna dopaminska koncentracija	0.1
$A_+$	Pozitivni STDP multiplikator	0.7
$A_-$	Negativni STDP multiplikator	0.3
$d$	Zakasnitev sinaps	0.5 ms
$\lambda_{\text{motor}}$	Povprečna hitrost Poissonovega šuma	1000 Hz
$w_{\text{Poisson}}$	Uteži sinaps Poissonovega šuma	100
	Začetne uteži sinaps	
$w_{\text{init, motor}}$	med vhodnimi in izhodnimi nevroni	$w_{\text{init, motor}} \sim \mathcal{N}(1300, 1)$

Tabela 4.1: Parametri simulacije R-STDP učenja na preprostem problemu.



Slika 4.2: Prikaz uteži sinaps med vhodnimi in izhodnimi nevroni ter povprečne aktivnosti dopaminergičnih nevronov med simulacijo R-STDP učenja na preprostem problemu.

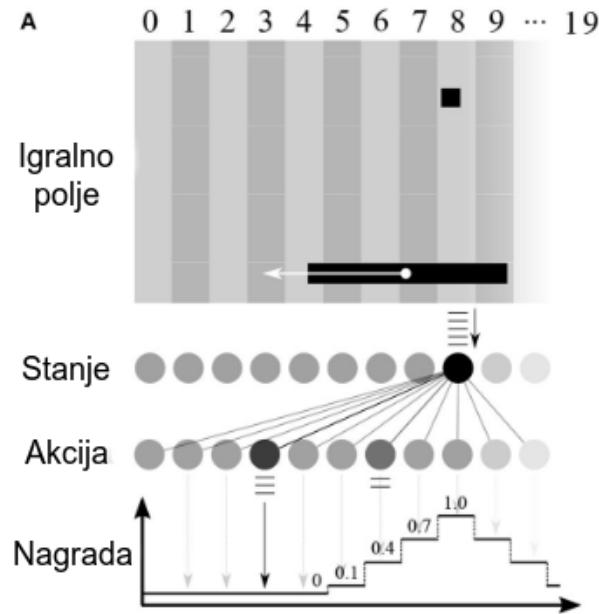


### 4.1.1 Igra Pong

V nadaljevanju bomo R-STDP učenje predstavili na agentu, ki igra igro *Pong*. R-STDP učenje je kratkovidno, saj se bomo naučili akcij le, če nagrada sledi takoj, ne pa tudi, če je nagrada zakasnjena. Za zakasnjene nagrade uporabljamo TD (*angl. Temporal Difference*) učenje, ki ga bomo implementirali v poglavju ???. Igra Pong v osnovi zahteva veliko predvidevanja, vendar lahko igranje poenostavimo v obliko, ki jo je mogoče osvojiti z R-STDP učenjem. Igro bomo definirali tako, da ima žogica stalno hitrost, določeno smer in pozicijo v x, y ravnini. Na levi strani igrišča bo naš agent premikal platformo v vertikalni smeri, na desni strani pa je stena, od katere se žogica prožno odbije. Če bi od agenta zahtevali predvidevanje, bi morali stanja definirati kot kartezični produkt x,y pozicije žogice, njene smeri in y pozicije platforme. Problem bomo poenostavili v problem sledenja žogici, kot v delu Wunderlich T, et al. 2019, kjer agent izbira zeleno ciljno točko platforme. Predvidevanje zato ni potrebno. Tako stanja kot akcije agenta so diskretizirane možne y pozicije žogice. Stanje je nagrajeno s stimulacijo dopaminskih nevronov s tokom  $I_R$ , ki je sorazmeren razliki med nagrado  $R_b$ , izračunani glede na oddaljenost zelene pozicije  $j$  od trenutne y pozicije žogice  $j$ , in povprečno nagrado  $\bar{R}_i$  v iteraciji  $i$ . Takšna konfiguracija je prikazana na sliki 4.1.1. S pomočjo povprečne nagrade omejimo krepitev sinaps, če te ne izboljšajo trenutne politike.

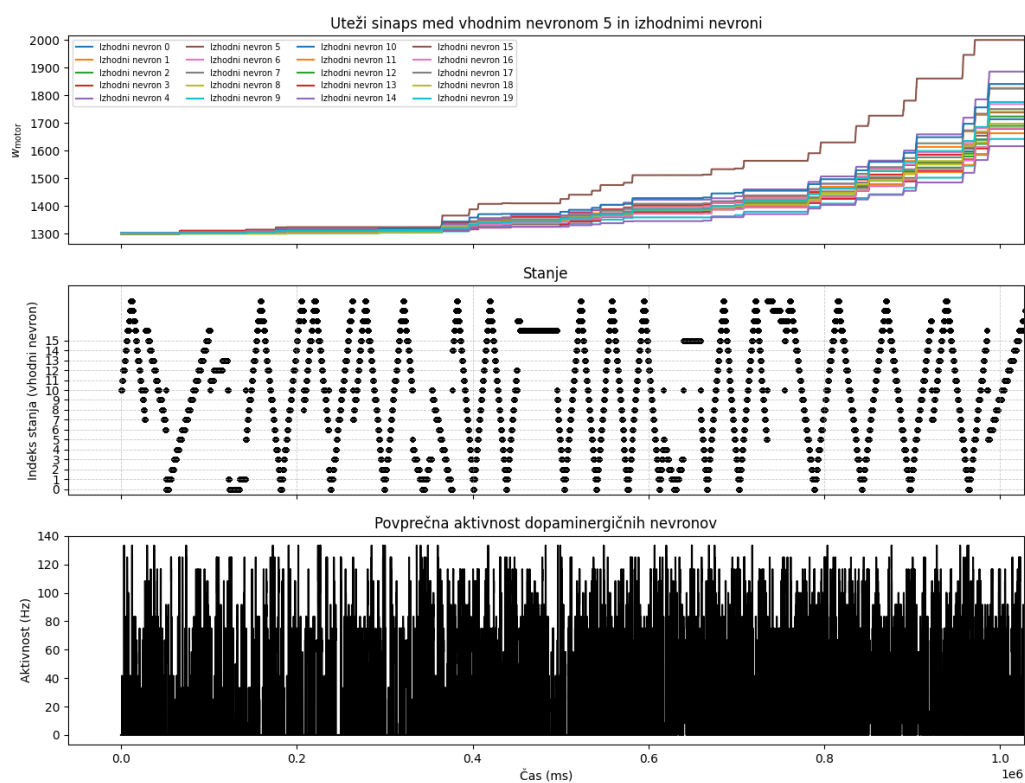
$$R_b = \begin{cases} 1 - |j - k| \cdot 0.3 & \text{if } |j - k| \leq 3, \\ 0 & \text{otherwise.} \end{cases} \quad (4.1)$$

$$I_R = \max(R_b - \bar{R}_i, 0) \cdot 600 \text{ pA} \quad (4.2)$$



Slika 4.3: Grafična predstavitev agenta in okolja (Wunderlich T, et al. 2019).

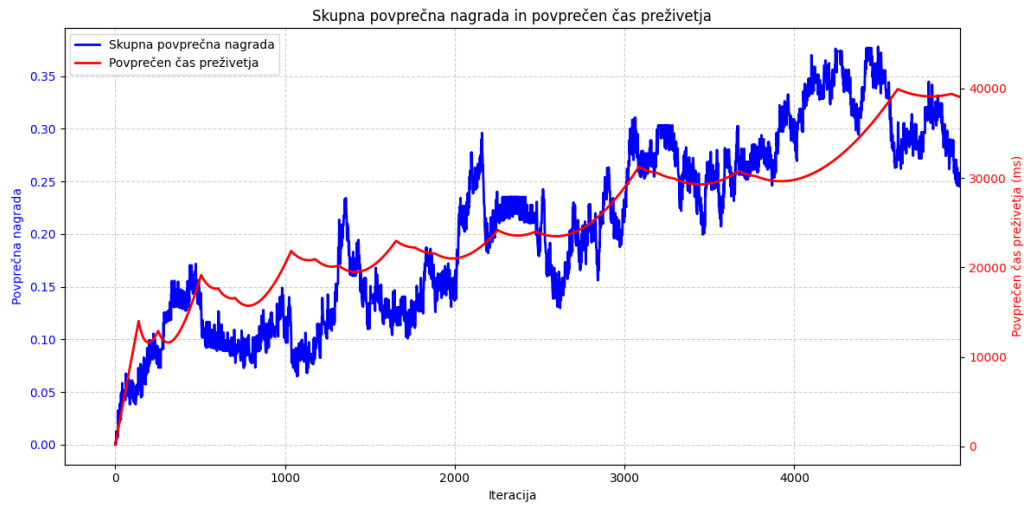
Pričakujemo, da bodo v posameznih stanjih prevladale sinapse, ki iz vhodnega nevrona vodijo do akcij okrog izhodnega nevrona, ki predstavlja isto  $y$  pozicijo, kot jo ima takrat žogica. Polje smo po  $y$  osi diskretizirali na 20 stanj. Po simulaciji, ki je trajala 6000 ms, na sliki 4.1.1 vidimo graf povezav med vhodnim nevrom, ki predstavlja pozicijo  $y = 5$  in 20 izhodnimi nevroni, kjer med učenjem prevladuje izhodni nevron 5, sledi pa mu izhodni nevron 4. Za simulacijo smo uporabili enake parametre kot pri primeru R-STDP učenja na preprostem problemu.



Slika 4.4: Graf uteži povezav med vhodnim nevronom 5 in izhodnimi nevroni ter povprečna aktivnost dopaminergičnih nevronov med 6000 ms simulacijo igranja igre Pong.

## Rezultati

Učenje spremljamo s povprečnim časom preživetja, ki predstavlja čas, ki je minil od zadnje zgrešitve žogice ali začetka igre. Pričakujemo, da bo oblika krivulje skupne povprečne nagrade sledila krivulji povprečnega časa preživetja, saj bo agent ob uspešnem sledenju žogici prejemal višje in pogostejše nagrade. To lahko potrdimo iz slike 4.1.1. Kvaliteto učenja bi lahko pri poljubnem problemu spremljali zgolj s skupno povprečno nagrado.



Slika 4.5: Skupna povprečna nagrada  $\bar{R}_i$  in povprečen čas preživetja tekom 6000 iteracij po 200 ms.

R-STDP učenje je v tej obliki učinkovito le pri nagradah, ki niso oddaljene, oziroma drugače povedano, agent se ne bo naučil potencialne poti skozi različna nenagrajena stanja, da bi prišel do končne nagrade. Primer problema z oddaljeno nagrado je iskanje poti do nagrade v mreži, kjer se agent lahko premika v sosednja polja levo, desno, gor in dol. Pri trenutni implementaciji se bo agent naučil prehoda le iz stanj, ki so neposredno ob nagrajenem stanju.

Pri učenju bomo agenta nagradili, ko preide v končno stanje, nato pa ga po-

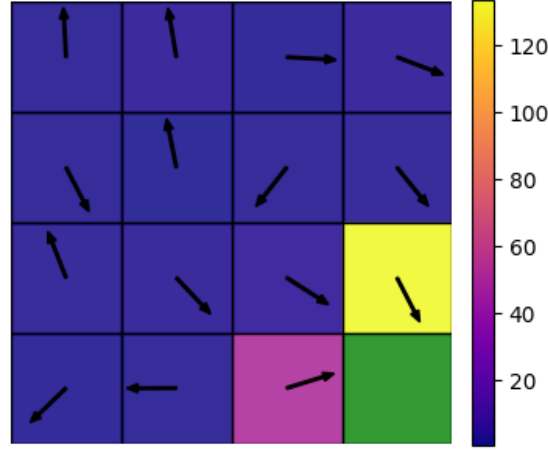
stavili v naključno stanje. V vsakem stanju  $i$  bomo trenutno politiko agenta prikazali s puščicami, katerih smer določa normaliziran vektor  $\hat{x}_i$ , ki predstavlja preferenco akcije glede na razlike v utežeh sinaps.

$$\begin{aligned}\vec{x}_i &= \sum_{j=0}^3 w_{ij} \cdot \vec{d}_j, \\ L_i &= \|\vec{x}_i\|, \\ \hat{x}_i &= \begin{cases} \frac{\vec{x}_i}{L_i} & \text{if } L_i > 0 \\ 0 & \text{otherwise} \end{cases},\end{aligned}$$

kjer je  $w_{ij}$  utež sinapse iz vhodnega nevrona  $i$  do izhodnega nevrona  $j$  in  $\vec{d}_j$  smerni vektor, ki predstavlja akcijo izhodnega nevrona  $j$

$$\vec{d}_0 = (0, 1), \quad \vec{d}_1 = (0, -1), \quad \vec{d}_2 = (-1, 0), \quad \vec{d}_3 = (1, 0).$$

Za prikaz “samozavesti” pri izbiri akcije v stanju  $i$  kot rezultata učenja bomo polja ustrezno obarvali glede na maksimalno razliko med utežmi med vhodnim nevronom  $i$  in vsakim od izhodnih nevronov. Slika 4.6 prikazuje rezultat učenja po 500 iteracijah, kar potrjuje, da se agent ni sposoben naučiti poti do nagrade iz poljubnega stanja, temveč le iz stanj neposredno ob nagradi.



Slika 4.6: Prikaz politike po 500 iteracijah po 200 ms. Končno stanje je obarvano z zeleno.

## 4.2 TD učenje in model akter-kritik

Časovno razlikovalno učenje (angl. Temporal Difference Learning, TD) je metoda spodbujevanega učenja, ki posodablja oceno vrednosti stanj ali parov stanje–akcija sproti, med interakcijo z okoljem.

Osnovna posodobitvena enačba za vrednostno funkcijo stanja

$$V(s_t) \leftarrow V(s_t) + \alpha \delta_t,$$

kjer je  $\alpha$  hitrost učenja, TD-napaka  $\delta_t$  pa je definirana kot

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t).$$

V izrazu je  $r_{t+1}$  nagrada ob prehodu iz stanja  $s_t$  v stanje  $s_{t+1}$ , faktor  $\gamma \in [0, 1]$  pa določa relativno težo prihodnjih nagrad. TD-napaka predstavlja razliko med izboljšano napovedjo vrednosti in prejšnjo oceno.

### 4.2.1 Model akter-kritik

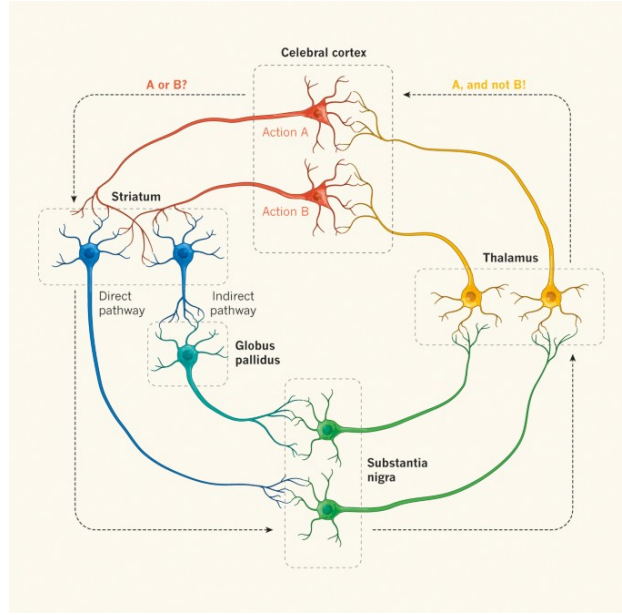
TD učenje bomo implementirali z modelom akter-kritik (*angl. actor-critic*), po zgledu Wiebke P, et al. 2011, na nalogi z mrežo. Model akter-kritik je sestavljen iz dveh delov: akterja, dopaminsko moduliranega RSTDP dela, kot smo ga že implementirali, in pa kritika, ki ocenjuje vrednost trenutnega stanja. Celoten model je navdihnjen po dopaminskem sistemu, prisotnem v človeških možganih, natančneje v bazalnih ganglijah.

Bazalni gangliji so skupina jeder v možganih, ki igrajo ključno vlogo pri nadzoru gibanja, učenju akcij in odločanju ter realizirajo obliko TD učenja. Model akter-kritik je poenostavitev in abstrakcija resničnih mehanizmov v možganih, ki to omogočajo. V bazalnih ganglijah, kot so prikazani na sliki 4.7, pri tem sodeluje več skupin nevronov in povezav, ki so v modelu akter-kritik, kot ga predstavlja Wiebke P, et al. 2011, logično združeni v *korteks*, ki predstavlja vhodne nevrone, *motorične nevrone*, ki predstavljajo izhodne nevrone in možne akcije, ter skupine nevronov kritika: *striatum*, *ventralni pallidum* in dopaminergične nevrone. *Substantia nigra* in *talamus* bazalnih ganglijev sta funkcionalno združena v dopaminergične nevrone, ki projicirajo dopamin do povezav med vhodom in striatumom ter vhodom in izhodnimi motoričnimi nevroni. Tako v bazalnih ganglijah kot v modelu akter-kritik, kot ga predstavlja Wiebke P, et al., razlikujemo dve glavni poti: direktno in indirektno pot, ki vodita iz nevronov *striatuma* do dopaminergičnih nevronov. Direktna pot je zakasnjena inhibitorna pot, ki poteka neposredno iz striatuma do dopaminergičnih nevronov, indirektna pot pa je inhibitorna do nevronov *ventralnega palliduma*, posebne skupine nevronov, ki inhibira aktivnost dopaminergičnih nevronov. *Ventralni pallidum* se nahaja ventralno od *globusa pallidusa*, prikazanega na klasičnem diagramu bazalnih ganglijev, in je povezan s pričakovanjem nagrade in odločanjem, zato Wiebke P, et al. za skupino nevronov na indirektni poti verjetno izbere to poimenovanje.

Ob prisotnosti osnovne, od 0 različne frekvence nevronov *ventralnega palliduma* bo indirektna pot imela na dopaminergične nevrone vzbujajoč učinek. Indirektna in direktna povezava na dopaminergične nevrone delujeta konkurenčno. Indirektna pot ima minimalen zamik in aktivnost striatuma v trenutnem stanju neposredno preslika v povišano aktivnost dopaminergičnih nevronov. Hkrati v času nahajanja v trenutnem stanju direktna povezava inhibira dopaminergične nevrone sorazmerno z aktivnostjo striatuma, kakršna je bila ta v prejšnjem stanju, zaradi zakasnitve. Indirektna in direktna povezava skupaj računata TD-napako  $\delta_t$ , ki v trenutnem stanju glede na izračunan približek vrednosti trenutnega stanja okrepi sinapse prejšnjega stanja, odgovorne za izbiro akcije, ki nas je pripeljala v trenutno stanje. Ta mehanizem je prikazan na sliki 4.9. Povprečna teža sinaps med vhodnim nevronom  $i$  in striatumom predstavlja pričakovano nagrado in približek vrednosti stanja  $i$ . Ob prehodu iz stanja z visoko povprečno utežjo sinaps do striatuma v stanje z nizko bo direktna povezava prevladala in bodo dopaminergični nevroni inhibirani, in obratno. Če se premaknemo v stanje s približno enako povprečno utežjo povezave do striatuma, se bosta direktna in indirektna povezava izničili, dopaminergični nevroni pa se bodo prožili s frekvenco, ki jo določa zunanji Poissonov šum.

Skupino vhodnih nevronov oziroma korteks bo predstavljalo  $N_{in}$  nevronov, ki bodo povezani z  $N_a$  motoričnimi oziroma izhodnimi nevroni. V akterju bomo zaradi razlogov, navedenih v poglavju 4.1, uporabljali model nevrona z eksponentnim jedrom, v kritiku pa biološko bolj realistične nevrone z alfa jedrom. Ker bodo vhodni nevroni akterju in kritiku skupni, bomo med vhodnimi in izhodnimi nevroni dodali dodatni nivo  $N_{in}$  vhodnih motoričnih nevronov, ki so z vhodnimi nevroni preko statičnih povezav z utežmi  $w_{in \rightarrow in, motor}$  povezani po režimu *one-to-one*, torej en vhodni nevron z enim nevronom vmesnega nivoja. Vmesni nivo nam bo omogočil prilagajanje frekvence in šuma, potrebnega za R-STDP učenje v akterju, ločeno od kritika, kar nam bo olajšalo iskanje

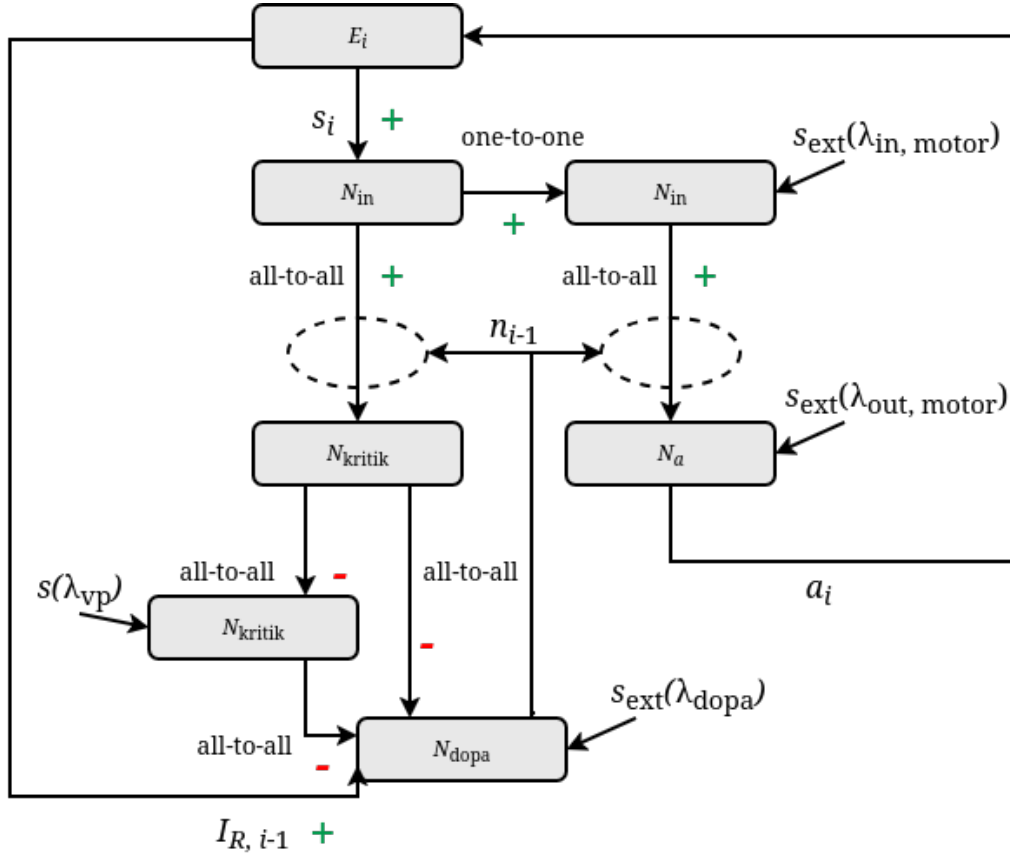




Slika 4.7: Diagram skupin nevronov bazalnih ganglijev.

ustreznih hiperparametrov. Vmesni nevroni so po režimu *all-to-all* (vsak vhodni nevron je povezan z vsakim izhodnim) povezani z izhodnimi nevroni prek zakasnenih R-STDP sinaps z normalno porazdeljenimi utežmi. Prav tako so vhodni nevroni po režimu *all-to-all* povezani z  $N_{\text{kritik}}$  nevroni striatuma prek zakasnenih R-STDP sinaps z normalno porazdeljenimi utežmi. Striatum je preko statičnih inhibitornih povezav z utežmi  $w_{\text{str} \rightarrow \text{vp}}$  povezan z  $N_{\text{kritik}}$  nevroni ventralnega palliduma in preko statičnih inhibitornih povezav z utežmi  $w_{\text{str} \rightarrow \text{dopa}}$  ter zakasnitvijo  $d_{\text{dir}}$  z  $N_{\text{dopa}}$  dopaminergičnimi nevroni. Ventralni pallidum je z dopaminergičnimi nevroni prav tako povezan preko statičnih inhibitornih povezav, ki pa niso zakasnjene in imajo uteži  $w_{\text{vp} \rightarrow \text{dopa}}$ . Vse povezave kritika so povezane po režimu *all-to-all*. V nevrone vmesnega nivoja med vhodnimi in izhodnimi nevroni injiciramo Poissonov šum s povprečno hitrostjo  $\lambda_{\text{in, motor}}$ , v izhodne nevrone pa Poissonov šum s hitrostjo  $\lambda_{\text{out, motor}}$ . Poissonov šum prav tako injiciramo v nevrone ventralnega palliduma s povprečno hitrostjo  $\lambda_{\text{vp}}$  in v dopaminergične nevrone s povprečno hitrostjo  $\lambda_{\text{dopa}}$ . Generatorji Poissonovega šuma so s posameznimi skupinami

nevronov povezani prek statičnih povezav z utežmi  $w_{\text{ext}, \text{in}}$ ,  $w_{\text{ext}, \text{out}}$ ,  $w_{\text{ext}, \text{vp}}$  in  $w_{\text{ext}, \text{dopa}}$ . Hiperparametri R-STDP sinaps in modelov nevronov so skupaj z ostalimi hiperparametri navedeni v poglavju 4.2.2. Opisani model je prikazan na sliki 4.8.



Slika 4.8: Prikaz implementiranega aktor-kritik sistema.

Za razliko od izvirnega sistema, predstavljenega v ??, bomo za model sinapse uporabili našo zakasnjeno R-STDP sinapso, kot smo jo razvili v poglavjih 3.1 in 4.1, ki poleg presinaptičnih impulzov upošteva tudi post-sinaptične po pravilu STDP. Tako bomo lahko kot akter uporabili R-STDP sistem, ki smo ga razvili v poglavju 4.1. Naša implementacija se bo razlikovala tudi v načinu izbire akcije, saj za izbrano akcijo izberemo tisto, katere pripadajoči izhodni nevron ima najvišje število impulzov v trenutnem sta-

nju. V izvirnem sistemu je akcija izbrana glede na prvi impulz pripadajočega izhodnega nevrona, ki se je sprožil kot rezultat stimulacije v trenutnem stanju. Pri tem načinu moramo po prvem impulzu v čim krajšem časovnem intervalu inhibirati vse ostale izhodne nevrone. Tako bomo bolj neposredno okrepili povezave, odgovorne za izbrano aktivnost, saj smo z inhibicijo izničili *eligibility* sled sinaps do ostalih izhodnih nevronov, poleg tega tudi ne bo potrebe po tekmovanju sinaps, vendar moramo zato preveriti impulze izhodnih nevronov v vsakem koraku simulatorja. V primeru simulatorja NEST je to vsakih 0,1 ms, kar pa je problematično, saj simulator teče v C++ zaledju, ki ga zapustimo takoj, ko prekinemo simulacijo. Tako je bistvena razlika med tem, ali 100-krat pošljemo ukaz `nest.Simulate(0.1)` ali enkrat `nest.Simulate(10)`. Naš sistem bo zaradi hitrosti simulacije po številu nevronov manjši od izvirnega sistema.

### 4.2.2 Izbira parametrov

Parametri so bili izbrani eksperimentalno, brez oziranja na biološko točnost. Pri spreminjanju velikosti posameznih skupin nevronov moramo pri izbiri parametrov paziti, da ohranjamo osnovno frekvenco dopaminergičnih nevronov ter ravnovesje med inhibicijo in vzbujanjem zaradi direktne in indirektna povezave. Zmanjšanje frekvence, ki jo izvaja plast nevronov med vhodnimi in izhodnimi nevroni, mora biti dovolj veliko, da šum pri osnovnih utežeh sinaps med srednjo plastjo in izhodnimi nevroni omogoči učenje, kot je razloženo v poglavju 4.1. V tabelah 4.2.2, 4.2.2, 4.2.2, 4.2.2, 4.2.2 in 4.2.2 so navedene konstante in parametri implementiranega modela. Parametri, ki niso prikazani v tabelah, imajo privzete vrednosti NEST simulatorja.

Simbol	Pomen	Vrednost
POLL_TIME	Čas simulacije na iteracijo	200
$f(s_{in,i})$	frekvenca stimulacije vhodnega nevrona $i$	100 Hz

Tabela 4.2: Parametri simulacije

Simbol	Pomen	Vrednost
<b>Parametri skupin nevronov kritika</b>		
tip	Tip modela nevrone	<i>iaf_psc_alpha</i>
$C_{m,in}$	Membranska kapacitivnost	250.0 pF
$\tau_{m,in}$	Časovna konstanta membrane	10.0 ms
$V_{reset,in}$	Potencial ponastavitve	0.0 mV
$V_{th,in}$	Prag proženja	20.0 mV
$t_{ref,in}$	Refraktorna doba	0.5 ms
$\tau_{syn,ex,in} = \tau_{syn,in,in}$	Ekscitatorna in inhibitorna sinaptična konstanta	2 ms
$\tau_{-,a}$	Negativna STDP konstanta	20.0 ms
$V_{m,in}$	Začetni membranski potencial	0.0 mV
$E_{L,in}$	Mirovalni potencial	0.0 mV
<b>Parametri motoričnih nevronov</b>		
tip	Tip modela nevrone	<i>iaf_psc_exp</i>
$C_{m,a}$	Membranska kapacitivnost motornih nevronov	250.0 pF
$\tau_{m,a}$	Časovna konstanta membrane	10.0 ms
$V_{reset,a}$	Potencial ponastavitve	0.0 mV
$V_{th,a}$	Prag proženja	20.0 mV
$t_{ref,a}$	Refraktorna doba	0.1 ms
$\tau_{syn,ex,a} = \tau_{syn,in,a}$	Ekscitatorna in inhibitorna sinaptična konstanta	2 ms
$\tau_{-,a}$	Negativna STDP konstanta	20.0 ms
$V_{m,a}$	Začetni membranski potencial	0.0 mV
$E_{L,a}$	Mirovalni potencial	0.0 mV

Tabela 4.3: Parametri nevronov

Simbol	Pomen	Vrednost
<b>Parametri sinaps med vhodnimi in vhodnimi motoričnimi nevroni</b>		
tip	Tip sinapse	Privzeta konstantna NEST sinapsa
$w_{\text{in} \rightarrow \text{in, motor}}$	Uteži sinaps med vhodnimi in vhodnimi motoričnimi nevroni	120
<b>Parametri sinaps med vhodnimi in izhodnimi motoričnimi nevroni</b>		
tip	Tip sinapse	Zakasnjena dopaminsko modulirana STDP sinapsa
$\tau_c$	Odtekanje <i>eligibility</i> sledi	5 ms
$\tau_{c, \text{delay}}$	Zakasnitev sledi $c$	200 ms
$\tau_n$	Odtekanje dopaminske sledi	10 ms
$\tau_+$	Pozitivna STDP konstanta	20 ms
$b$	Bazalna dopaminska koncentracija	0.1
$A_+$	Pozitivni STDP multiplikator	1.5
$A_-$	Negativni STDP multiplikator	1.0
$W_{\text{min},a}$	Minimalna utež	500
$W_{\text{max},a}$	Maksimalna utež	4000
$w_{\text{in, motor} \rightarrow a}$	Začetne uteži sinaps med vhodnimi in izhodnimi motoričnimi nevroni	$\mathcal{N}(1300, 1)$

Tabela 4.4: Parametri sinaps med vhodnimi in motoričnimi nevroni

Simbol	Pomen	Vrednost
<b>Parametri sinaps med vhodnimi nevroni in striatumom</b>		
tip	Tip sinapse	Zakasnjena dopaminsko modulirana STDP sinapsa
$\tau_c$	Odtekanje <i>eligibility</i> sledi	5 ms
$\tau_{c, \text{delay}}$	Zakasnitev sledi $c$	200 ms
$\tau_n$	Odtekanje dopaminske sledi	10 ms
$\tau_+$	Pozitivna STDP konstanta	20 ms
$b$	Bazalna dopaminska koncentracija	0.1
$A_+$	Pozitivni STDP multiplikator	1.5
$A_-$	Negativni STDP multiplikator	1.0
$W_{\min, str}$	Minimalna utež	150
$W_{\max, str}$	Maksimalna utež	1000
$w_{\text{in} \rightarrow \text{str}}$	Začetne uteži sinaps med vhodnimi in striatum nevroni	$\mathcal{N}(150, 8)$

Tabela 4.5: Parametri sinaps med vhodom in striatumom

Simbol	Pomen	Vrednost
tip	Tip sinapse	Privzeta konstantna NEST sinapsa
$w_{\text{str} \rightarrow \text{vp}}$	Uteži sinps med striatumom in ventral pallidumom	-50
$w_{\text{str} \rightarrow \text{dopa}}$	Uteži sinps med striatumom in dopaminergičnimi nevroni	-55
$w_{\text{vp} \rightarrow \text{dopa}}$	Uteži sinps med ventral pallidumom in dopaminergičnimi nevroni	-65
$d_{\text{dir}}$	Zakasnitev sinaps direktne povezave	200 ms

Tabela 4.6: Parametri sinaps kritika



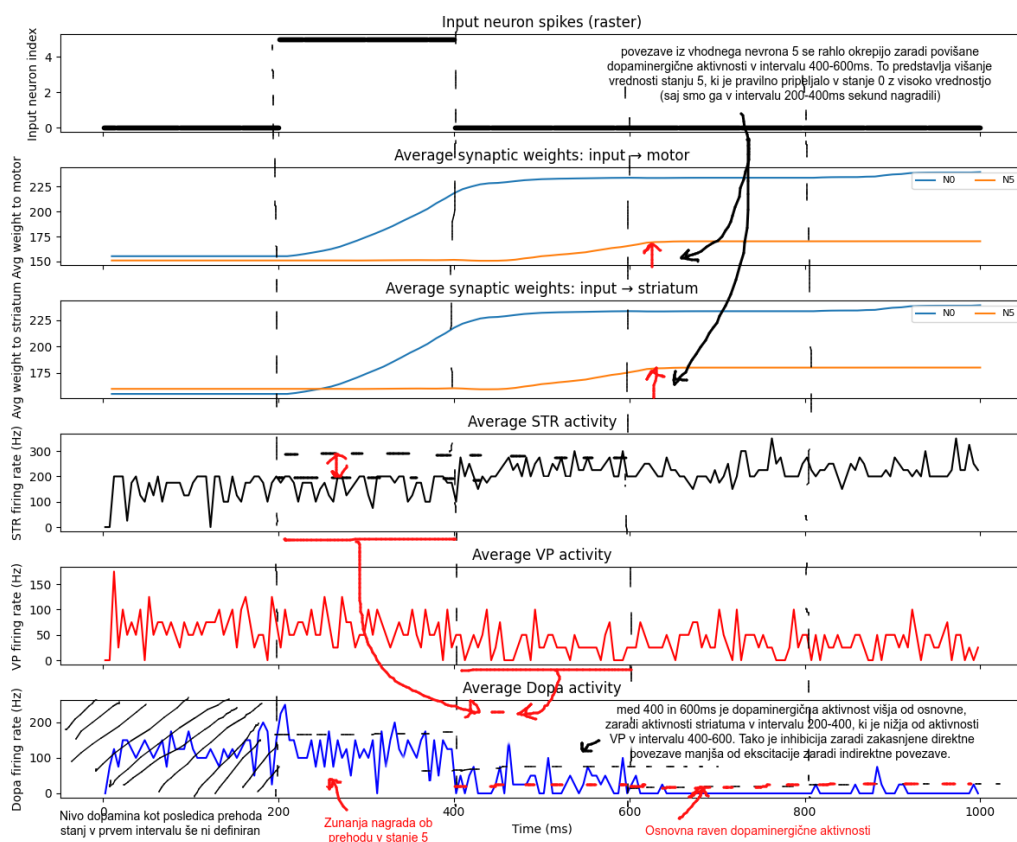
Simbol	Pomen	Vrednost
$\lambda_{vp}$	Povprečna hitrost šumnih impulzov nevronov ventral palliduma	5200
$\lambda_{dopa}$	Povprečna hitrost šumnih impulzov dopaminergičnih nevronov	4000
$\lambda_{in, motor}$	Povprečna hitrost šumnih impulzov vhodnih motoričnih nevronov	100 Hz
$\lambda_{out, motor}$	Povprečna hitrost šumnih impulzov izhodnih motoričnih nevronov	100 Hz
$w_{ext, in}$	Uteži statičnih povezav med generatorjem Poissonovega šuma in vhodnimi motoričnimi nevroni	50
$w_{ext, out}$	Uteži statičnih povezav med generatorjem Poissonovega šuma in izhodnimi motoričnimi nevroni	50
$w_{ext, vp}$	Uteži statičnih povezav med generatorjem Poissonovega šuma in nevroni ventralnega palliduma	50
$w_{ext, dopa}$	Uteži statičnih povezav med generatorjem Poissonovega šuma in dopaminergičnimi nevroni	50

Tabela 4.7: Parametri generatorjev šuma

### 4.2.3 Učenje

Stanja bomo razdelili v intervale dolžine 200 ms, ob prehodu med stanji pa bomo vhodne nevrone stimulirali enako kot v poglavju 4.1. Pri prehodu med

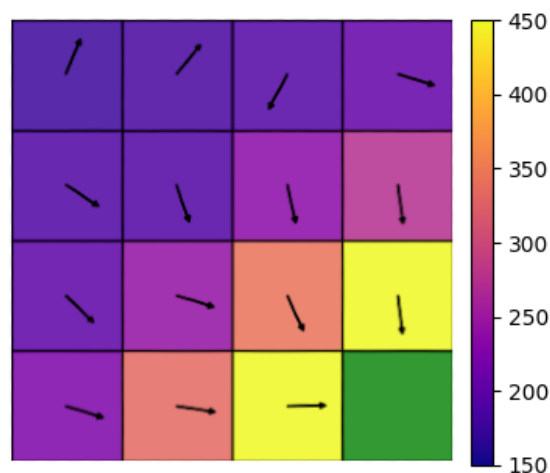
stanji se ob dovolj visoki koncentraciji dopamina lahko okrepijo tudi sinapse do izhodnih nevronov, ki so povezane z vhodnimi nevroni prejšnjega stanja, saj lahko *eligibility* sledi ostanje večje od 0 tudi preko več prehodov stanj. To v osnovi ni napačno in je posledica uporabe RSTDP sinapse, vendar bomo za učinkovitejše učenje med prehodi stanj prekinili stimulacijo 50 ms pred stimulacijo novega stanja, saj so za našo nalogo stanja med seboj neodvisna. Za določeno stanje ni pomembno, v katerem stanju smo bili prej. Mehanizme, opisane v prejšnjem poglavju, najprej preverimo na sekvenci prehodov iz stanja 0 v nagrajeno stanje 5 in nazaj v stanje 0, kot je prikazano na sliki 4.9. Ob prehodu v nagrajeno stanje se bodo uteži povezav med vhodnimi nevroni, ki predstavljajo stanje 0 in striatumom okrepile. To neposredno predstavlja višjo vrednost stanja 0 kot posledico višje pričakovane nagrade. Zunanja nagrada, ki smo jo dovedli ob prehodu v nagrajeno stanje 5, ne pomeni, da v stanju 5 pričakujemo visoko nagrado. Pričakovana nagrada se namreč zviša le ob prehodu v stanje z višjo vrednostjo ali ob prisotnosti zunanje nagrade in je odvisna od akcije, ki jo izvedemo v tem stanju. Ob prehodu iz stanja 5 nazaj v stanje 0 preidemo iz stanja z osnovnimi utežmi do striatuma v stanje 0, ki ima zdaj okrepljene uteži do striatuma. To predstavlja prehod v stanje z višjo vrednostjo. Posledica tega je, v primerjavi z osnovno frekvenco dopaminergičnih nevronov, povišana dopaminergična aktivnost, ki povzroči sorazmerno povišanje uteži sinaps do striatuma v stanju 5. V nadaljevanju ostajamo v stanju 0, kjer se ob prehodu iz stanja 0 v stanje 0 vrnemo k osnovni dopaminergični frekvenci.



Slika 4.9: Prikaz sinaptičnih uteži med vhodnimi nevroni stanj 0 in 5 do izhodnih nevronov ter nevronov striatuma, povprečne aktivnosti nevronov striatuma, povprečne aktivnosti nevronov ventralnega palliduma ter povprečne dopaminergične aktivnosti tekom sekvence prehodov med stanji 0 in 5.

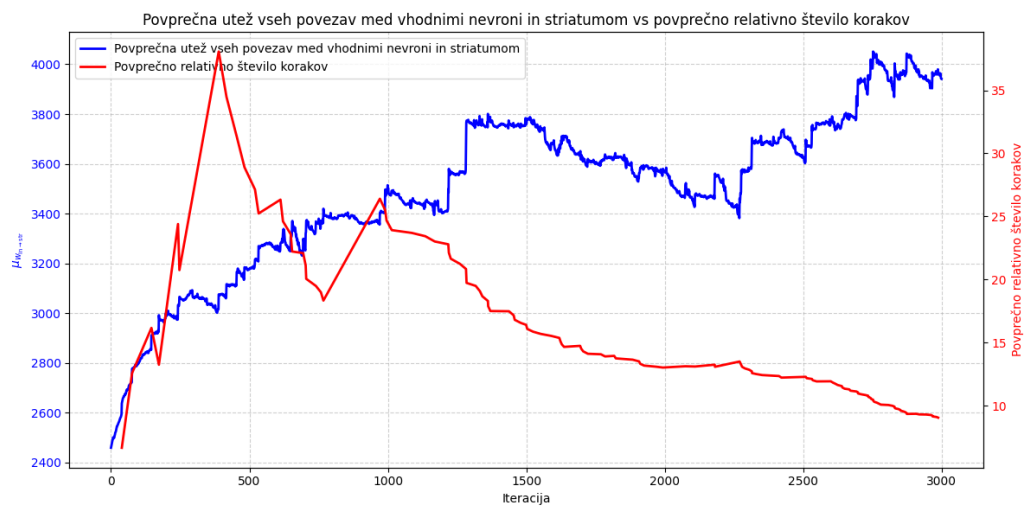
#### 4.2.4 Rezultati

Naučeno politiko bomo prikazali podobno kot v poglavju 4.1, vendar bomo “samozavest” izbire akcije v določenem stanju prikazali skupaj s povprečno utežjo povezav med vhodnimi nevroni pripadajočega stanja in striatumom. Rezultat učenja na 4x4 mreži po 3000 iteracijah je prikazan na sliki ???. Izbira akcije v posameznem polju je razvidna iz smeri puščice, kjer vidimo, da se je agent naučil skoraj optimalne navigacije do cilja iz poljubnega stanja. Pričakovano imajo stanja neposredno ob nagrajenem stanju najvišjo pričakovano vrednost, dlje kot se oddaljimo od tega stanja, nižja je pričakovana nagrada. Stanje 0, ki je najdlje od cilja, ima pri trenutni izbiri parametrov minimalno pričakovano nagrado. Propagiranje pričakovane nagrade od končnega stanja lahko pospešimo tako, da povišamo amplitudo posodobitve povezav do striatuma, vendar bomo s tem zvišali tudi rast motoričnih sinaps. Te bodo rasle prehitro, zato tekmovanje sinaps, kot je opisano v poglavju 4.1, ne bo tako učinkovito. Z drugimi besedami, stanja bomo nagrajevali prehitro.



Slika 4.10: Rezultat učenja modela na 4x4 mreži tekom 3000 iteracij po 200 ms.

Podobno kot pri R-STDP bomo učenje spremljali s povprečno nagrado, vendar tokrat k nagradi ne bo prispevala le zunanja nagrada, temveč tudi pričakovane nagrade. Pričakujemo, da bodo pričakovane nagrade med učenjem v vseh stanjih naraščale. To preverimo z neposrednejšo evalvacijo učenja na mreži, kjer po vsaki ponastavitvi stanja (ko dosežemo cilj) štejemo korake, dokler ponovno ne pridemo v ciljno stanje. Ker so različna stanja, v katera naključno postavimo agenta, različno oddaljena od cilja, bomo število korakov delili z manhattansko razdaljo do cilja. Tako dobimo “relativne korake”. Povprečno število relativnih korakov glede na povprečno utež vseh povezav med vhodnimi nevroni in nevroni striatuma skozi 3000 iteracij je prikazano na grafu 4.11, kjer vidimo, da povprečno število korakov, ki jih agent potrebuje, da pride do cilja, res pada sorazmerno z rastjo povprečne pričakovane nagrade.



Slika 4.11: Povprečna utež sinaps med vsemi vhodnimi nevroni in striatumom tekom 3000 iteracij po 200 ms.

## Poglavje 5

# Implementacija in uporabljena orodja

Rešitve so implementirane v jeziku Python, kjer za simulacijo uporabljamo simulator NEST, katerega zaledje je implementirano v C++. Za simulator, ki ga uporabljamo prek Pythonovega vmesnika, je v okviru tega diplomskega dela implementiran tudi modul, prav tako v C++. V okviru te naloge uporaba impulznih nevronske mreže zunaj simuliranega okolja, na trdo ožičenih nevronske čipih ali na robotih, ni zajeta, zato posebna oprema za ta namen ni bila uporabljena. Sistemi, razviti v diplomski nalogi, so poleg medsebojne primerjave ovrednoteni tudi z drugimi trenutno obstoječimi implementacijami spodbujevanega učenja na impulznih nevronske mrežah.





## Poglavje 6

### Možne razširitve

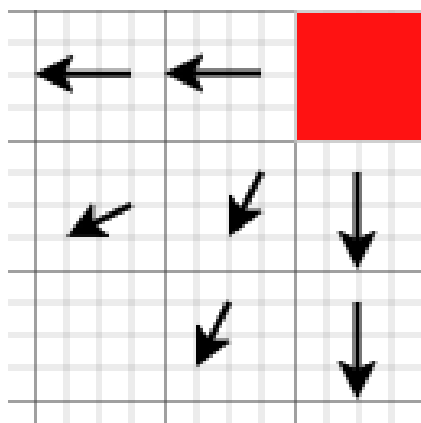
#### 6.1 Izognitveno obnašanje (*angl. aversive behaviour*)

V večini del, ki se ukvarjajo s spodbujevanim učenjem, se izogibanje stanjem, za katere želimo, da se jih agent izogiba, doseže s pomočjo negativne nagrade. Negativna nagrada v enačbi sinapse R-STDP obrne predznak posodobitve. Tako so sinapse, odgovorne za vstop v neželeno stanje, negativno posodobljene. V človeških možganih negativnega dopamina ni. Pojavi se ideja, da je izogibanje negativnim stanjem prav tako posledica učenja, kjer je nivo dopamina  $n > 0$ . Dopamin namreč predstavlja učenje, ne nujno nagrade. Negativno nagrado bi tako lahko predstavili s posebnim vhodom, ki predstavlja nek negativen stimulus ali “bolečino”, ki jo tekom učenja želimo zmanjšati. Zopet lahko uporabimo načela R-STDP in TD učenja, kjer zmanjšanje nivoja bolečine predstavlja nagrado. Trenutnemu akter-kritik sistemu bi dodali še eno kopijo kritika, ki računa časovno razliko nivoja bolečine in deluje na dopaminergične nevrone, ki so skupni obema kritikoma. Oba kritika tako delujeta konkurenčno. Ob prehodu iz stanja z visokim nivojem negativnega stimulusa v stanje z nizkim, dopaminergične nevrone vzbudimo, v obratnem primeru pa inhibiramo. V primeru enakega nivoja dovedenega negativnega stimulusa kritik negativne nagrade ne vpliva na dopaminergične nevrone.

Sistem se do negativnega stimulusa v tem primeru, kljub uporabi časovne razlike, obnaša kratkovidno.

Nagrajene bodo samo povezave, ki so nas vodile stran od bolečine, ker pa je negativen stimulus vedno doveden samo iz zunanosti sistema, bodo nagrajene povezave le v stanja neposredno ob negativnem stanju. Striatum za razliko od tega nivo dovedene nagrade napoveduje sam. Če želimo okrog negativnega stanja negativno označiti tudi stanja, ki nas potencialno vodijo vanj, bi morali v sistem dodati še skupino nevronov, ki stanja asociirajo z negativnim stimulusom in ga tako napovedujejo.

Pričakujemo, da bi oba kritika med seboj tekmovala za nagrajevanje tako akcij, ki vodijo bližje nagradi, kot tudi tistih, ki vodijo stran od negativnega stanja.



Slika 6.1: Pričakovana politika ob kritiku negativnih stanj (brez kritika nagrajenih stanj)

## 6.2 Rekurenčne povezave

Velika predpostavka sistemov, razvitih v tej diplomski nalogi, je, da rekurenčnih povezav ni. Tako so stanja časovno med seboj skoraj popolnoma neodvisna. Če dodamo več vmesnih nivojev in rekurenčne povezave, bodo stanja med seboj postala časovno odvisna. Pravzaprav stanja ne moremo

več definirati samo z aktivnostjo vhodnih nevronov, saj v vsakem trenutku stanje vsebuje tudi informacijo iz nevronov, ki so se prožili arbitrarno v preteklosti in nosijo informacijo o nekem prejšnjem stanju. V primeru našega akter-kritik sistema bi tako v vsakem trenutku  $t$  kritik računal časovno razliko med dvema neskončno kratkima stanjema  $s_t$  in  $s_{t-d}$ , kjer je  $d$  zakasnitev direktne povezave. Kljub temu pričakujemo, da rezultat ne bi bil drugačen saj bi ob prisotnosti 200 ms stimulacije, ki je do zdaj predstavljala stanje, v tem intervalu vseeno prevladala nevronska aktivnost, ki je neposredna posledica stimulacije vhodnih nevronov.

V doseganjih eksperimentih pravilna akcija določenega stanja ni bila odvisna od akcij, ki so nas pripeljale v to stanje, oziroma od zgodovine stanj. V primeru sprehajanja po mreži bomo končno stanje nagradili ne glede na to, iz katerega stanja vstopimo v nagrajeno stanje. Pričakujemo, da bi rekurenčne povezave predstavljale prednost pri nalogah, kjer je zgodovina stanj pomembna, oziroma kjer je nagrada stanja odvisna od prejšnjih stanj. Če bi v primeru sprehajanja po mreži premik v končno stanje iz stanja nad njim pripeljal do nagrade, prehod iz stanja levo pa ne, bi lahko tako končno stanje obravnavali kot dve različni stanji, glede na prehod. Sistem z rekurenčnimi povezavami bi kljub informaciji samo o polju interno predstavljal stanja odvisna tudi od prejšnjih premikov.

Rekurenčne povezave pa predstavljajo tudi dodaten izziv. V primeru našega akter-kritik sistema bi bila na primer potrebna redefinicija trenutnega načina izbire akcij, saj sta lahko dva izhodna nevrona povezana med seboj in se bosta vedno prožila skupaj. Rešitev bi lahko bila dopuščanje izbire več akcij hkrati, kjer takšno situacijo "kaznujemo".



## Poglavje 7

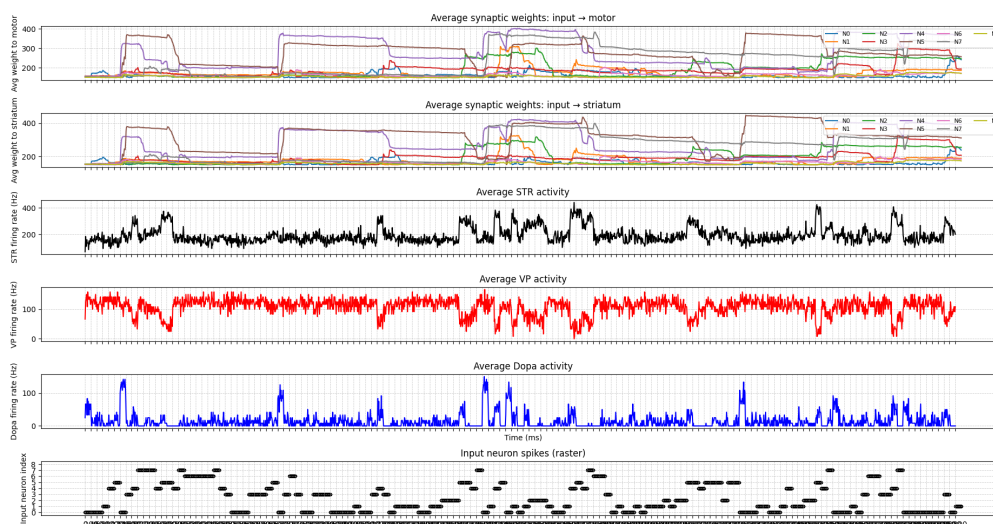
### Zaključek

Namen te diplomske naloge je bil predstaviti in rešiti nekatere izzive pri učenju impulznih nevronske mreže. V nalogi razvijemo inovativne rešitve, ki upoštevajo tako zahtevnost simulacije, kot tudi smiselnost z vidika nevrologije in resničnih mehanizmov v možganih. Dodatno smo se izognili vpeljavi negativne nagrade oziroma negativne koncentracije dopamina, saj to v resničnih možganih ni mogoče. Tako smo razvili R-STDP sistem, ki temelji zgolj na tekmovanju med sinapsami, za probleme, kjer se želimo določenim stanjem izogibati, pa predlagamo razširitev, ki ne uporablja negativne koncentracije dopamina. Od biološko realističnega sistema se najbolj oddaljimo pri iskanju parametrov sistema. Parametre smo iskali le z vidika doseganja zelenih mehanizmov, ne pa tudi skladnosti z vrednostmi, izmerjenimi v resničnih možganih. To si dopuščamo tudi zato, ker se način proženja nevronov, prenos signalov po sinapsah ter konfiguracije nevronov, kot so v bazalnih ganglijah zdijo, bolj skupni različnim živalskim vrstam kot parametri nevronov in sinaps. Center, odgovoren za sluh in modulacijo glasilk, je na primer po strukturi pri človeku in netopirju podoben, vendar pri netopirju ti centri očitno delujejo na precej višji frekvenci. Nadaljnje iskanje hiperparametrov bi najverjetneje lahko privedlo do boljših rezultatov, kot so bili doseženi v tej diplomski nalogi. Pri sistemih brez rekurenčnih povezav in brez dodatnih popolno povezanih plasti nevronov, kakršni so sistemi implementirani v tej

nalogi, višanje števila nevronov v posamezni skupini ne bi nujno privedlo do boljših rezultatov, vendar to ni bilo preverjeno zaradi računske zahtevnosti.

# Poglavje 8

## Ekstra



Slika 8.1: Obnašanje sistema tekom učenja na 3x3 mreži. Polja so oštevilčena od leve proti desni od zgoraj navzdol. Cilj se nahaja na polju 8. Povezave vhoda do striatuma stanj 5 in 7 so pričakovano najvišje, sledi pa jim 4, ki neposredno vodi v 5 in 7

Iz zgornje kolekcije grafov izberi izseke, ki predstavljajo ključne situacije med učenjem opisane mehanizme ocenjevanja nagrade in učenja.





# Viri

- Deleva A (2015). “TD learning in Monte Carlo tree search : masters thesis”. Magistrska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- Dobrevski M, Skočaj D (2021). “Deep reinforcement learning for map-less goal-driven robot navigation”. V: *International Journal of Advanced Robotic Systems*. 2021 18.1. DOI: 10.1177/1729881421992621.
- Izhikevich, E. M. (2007). “Solving the distal reward problem through linkage of STDP and dopamine signaling”. V: *Cerebral cortex (New York, N.Y. : 1991)* 17.10. DOI: 10.1093/cercor/bh1152.
- Potjans, Wiebke, Abigail Morrison in Markus Diesmann (2010). “Enabling Functional Neural Circuit Simulations with Distributed Computing of Neuromodulated Plasticity”. V: *Frontiers in Computational Neuroscience* Volume 4 - 2010. ISSN: 1662-5188. DOI: 10.3389/fncom.2010.00141. URL: <https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2010.00141>.
- Šutar M (2023). “Uporaba predvidevanja akcij nasprotnika pri učenju inteligentnega agenta”. Diplomaska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- Svete A (2020). “Posplošitev problema vozička s palico na zahtevnejše domene”. Diplomaska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

- T, Štromajer (2022). “Using machine learning to train a shepherd dog”. Magistrska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- Tsodyks, tMisha, Asher Uziel in Henry Markram (2000). “t Synchrony Generation in Recurrent Networks with Frequency-Dependent Synapses”. V: *Journal of Neuroscience* 20.1, RC50–RC50. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.20-01-j0003.2000. eprint: <https://www.jneurosci.org/content/20/1/RC50.full.pdf>. URL: <https://www.jneurosci.org/content/20/1/RC50>.
- Wiebke P, et al. (2011). “An Imperfect Dopaminergic Error Signal Can Drive Temporal-Difference Learning”. V: *PLoS computational biology* 7.5. DOI: 10.1371/journal.pcbi.1001133.
- Wunderlich T, et al. (2019). “Demonstrating Advantages of Neuromorphic Computation: A Pilot Study”. V: *Frontiers in neuroscience* 13.260. DOI: 10.3389/fnins.2019.00260.