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

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Kandidat: Matjaž Pogačnik
Naslov: Spodbujevano učenje na impulznih nevronskih mrežah
Vrsta naloge: Diplomska naloga na univerzitetnem programu prve stopnje Računalništvo in informatika
Mentor: prof. dr. Zoran Bosnić

Opis:
Kandidat naj preuči principe impulznih nevronskih 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

1 Uvod **1**

2 Pregled področja in sorodnih del **5**

3 Modeliranje nevronov in sinaps **9**

3.1 Model z eksponentnim jedrom 15

3.2 Model z alfa jedrom 16

3.3 Izbira modela nevrona 17

3.4 R-STDP Sinaptični model 19

4 Spodbujevano učenje na impulznih nevronskih mrežah **23**

4.1 R-STDP učenje 23

4.2 TD učenje in model akter-kritik 36

5 Možne razširitve **53**

5.1 Izognitveno obnašanje (*angl. aversive behaviour*) 53

5.2 Rekurenčne povezave 54

6 Zaključek **57**

Viri **59**

Seznam uporabljenih kratic

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

Povzetek

Naslov: Spodbujevano učenje na impulznih nevronskih mrežah

Avtor: Matjaž Pogačnik

V tem diplomskem delu obravnavamo spodbujevano učenje na impulznih nevronskih mrežah, tipu nevronskih mrež, 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 smiselnost. 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 nevronski akter–kritik sistem, ki lahko rešuje tudi probleme z zakasnjenimi 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 zakasnjenimi 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 nevronsko arhitekturo 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 zakasnjenimi 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 nevronsko arhitekturo 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 nevronskih mrež 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 nevronskih mrežah, niso neposredno uporabni zaradi diskretne narave impulzov in nelinearne časovne dinamike.

Eden temeljnih pristopov k učenju v impulznih nevronskih 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 nevronskih mrež v okviru spodbujevanega učenja. Takšni pristopi so uspešni pri nalogah s takojšnjimi nagradami, vendar se soočajo z omejitvami pri reševanju problemov z zakasnjenimi 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 nevronskih mrež 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 nevronskih mrež z učenjem na podlagi časovne razlike. V delu Wunderlich T, et al. 2019 je predstavljena uporaba TD učenja na trdo-ožičeni (nevromorfni) impulzni nevronski 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 nevronsko arhitekturo, ki združuje lokalna pravila učenja, globalne nagradne signale in biološko smiselno implementacijo brez uporabe

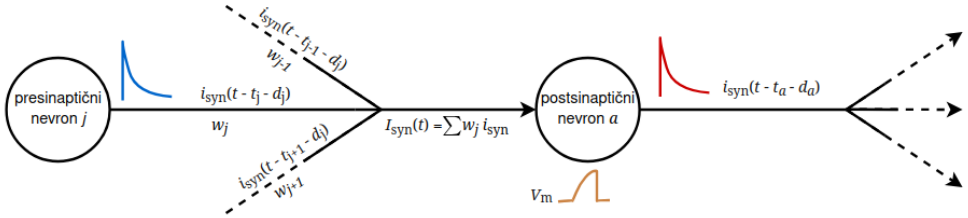
eksplicitnih negativnih nagrad. Poudarek je na časovni dodelitvi zaslug in postopni razširitvi osnovnega R-STDP pristopa v arhitekturo akter–kritik, ki omogoča učenje v okoljih z zakasnjenimi nagradami.

Poglavje 3

Modeliranje nevronov in sinaps

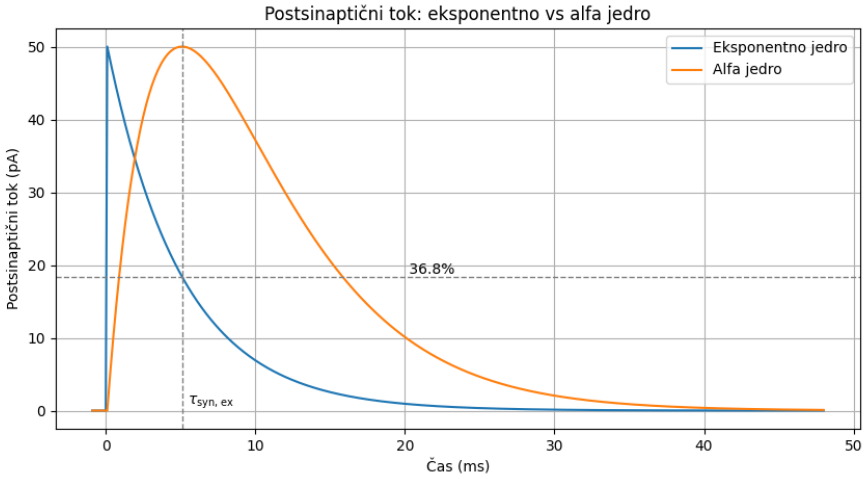
Impulzne nevronske mreže so določene z modelom nevrona in modelom sinapse, ki povezuje nevrone. Obstaja veliko modelov, v nadaljevanju pa bosta predstavljena in primerjana dva modela nevronov glede na njuno uporabnost pri spodbujevanem učenju na impulznih nevronskih mrežah. Predstavljen bo tudi model sinapse, primeren za spodbujevano 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 $i_{\text{syn}}(t-t_j-d_j)$. 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 nevronskim 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 3.1, eksponentno jedro doseže ravno pri $t = \tau_{\text{syn, ex}}$, času, ko alpha jedro doseže ravno maksimalno vrednost.



Slika 3.2: Postsinaptični tok modela z eksponentnim in alfa jedrom pri sinaptični uteži $w = 50$ in $\tau_{\text{syn, ex}} = 5$ 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 bolje 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;

- τ_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 τ_m lahko definiramo tudi kot produkt med kapacitivnostjo in uporom membrane $\tau_m = C_m R_m = \frac{C_m}{g_L}$;

- t_{ref} — **refrakcijskio 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

$$\frac{dV_m}{dt} = -\frac{V_m - E_L}{\tau_m} + \frac{I_{\text{syn}} + I_e}{C_m}. \tag{3.1}$$

Prvi člen na desni strani enačbe opisuje pasivno uhajanje membranskega potenciala proti mirujoči vrednosti E_L s časovno konstanto τ_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_{syn} in i_{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_{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 vzbujajoče in inhibitorne povezave. I_{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 vzbujajočih (X = ex) in inhibitornih (X = 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.

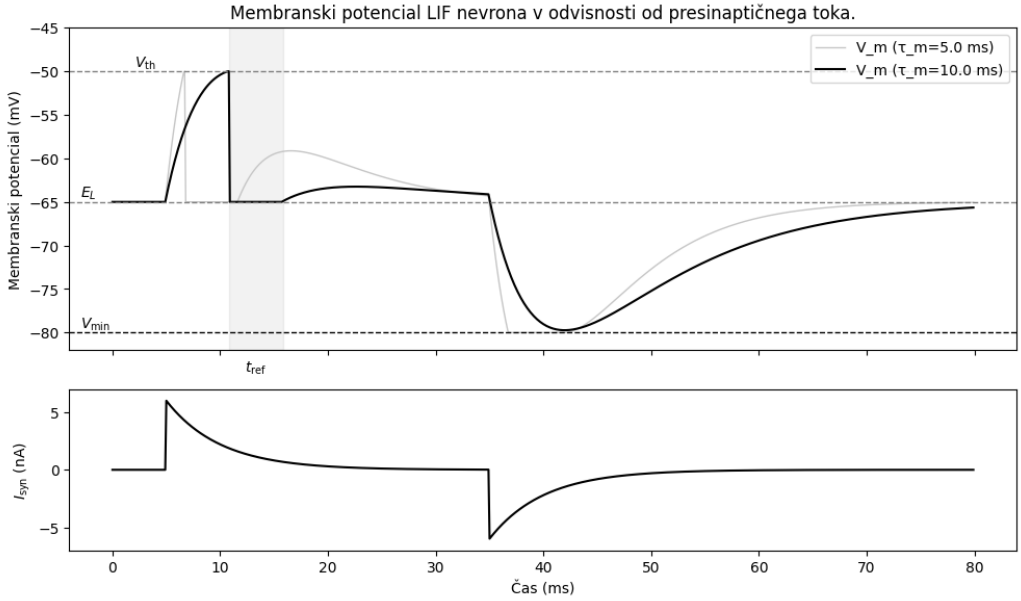
Postsinaptični tokovi so ne glede na jedro določeni z naslednjima parametroma.

- $\tau_{\text{syn,ex}}$ — **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 vzbujajočo in nevrona z inhibitorno povezavo,

oba z eksponentno oblikovanim postsinaptičnim tokom. Na grafu je ravidno, kako τ_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: vzbujajoči nevron z utežjo $w_{\text{ex}} = 6$ in inhibitorni nevron z utežjo $w_{\text{in}} = -6$. Ostali parametri so $\tau_{\text{syn, ex}} = \tau_{\text{syn, in}} = 5$ ms, zakasnitev $d_j = 0$ ms, $V_{\text{th}} = -50$ mV, $E_L = -65$ mV, $V_{\text{min}} = -80$ mV, $t_{\text{ref}} = 5$ 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. Vzbujajoči nevron se proži ob $t = 5$ ms, inhibitorni pa ob $t = 35$ ms.

3.1 Model z eksponentnim jedrom

V simulatorju NEST, ki ga bomo uporabljali za implementacijo kasnejših sistemov, 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

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

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

$$\frac{dz}{dt} = \frac{i}{\tau_{\text{syn, X}}} - \frac{z}{\tau_{rec}}, \tag{3.4}$$

kjer t_{sp} predstavlja čas presinaptičnega impulza, τ_{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

$$\frac{di}{dt} = -\frac{i}{\tau_{\text{syn, X}}}. \tag{3.5}$$

Rešitev te diferencialne enačbe je tako

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

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 $\tau_{\text{syn, X}}$.

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 3.3). Količina naboja q, ki se po sprožitvi impulza prenese prek sinapse, je definirana kot ploščina pod krivuljo

postsinaptičnega toka

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

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.

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

3.2 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 postsinaptičnega toka.

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

3.3 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_{alfa} je pri alfa jedru večja od prenesenega naboja pri eksponentnem jedru q_{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 prek petih

postsinaptičnih nevronov, v katere neodvisno injiciramo Poissonov šum, saj je ta biološko najbolj realističen. Ta namreč neposredno predstavlja impulze nevronov.

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

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

Parameter	**Vrednost**
Število postsinaptičnih nevronov	5
Trajanje simulacije	5000 ms
C_m	250.0 pF
τ_m	20.0 ms
E_L	0.0 mV
V_{th}	20.0 mV
V_{reset}	0.0 mV
t_{ref}	2.0 ms
$\tau_{\text{syn,ex}}$	5.0 ms
w_{eksp}	25.0
w_{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^2)
Eksponentno	7.846 ± 0.021	$\mathbf{0.402 \pm 0.028}$
Alfa	7.800 ± 0.023	$\mathbf{0.270 \pm 0.006}$

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

3.4 R-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."

Za spremljanje "skupnega proženja" parov nevronov bomo definirali vrednost STDP, ki predstavlja velikost potencialne posodobitve uteži sinapse med parom nevronov glede na čas impulza pre- in postsinaptičnega nevrona. V impulznih nevronskih mrežah to omogoča označevanje ali krepitev sinaps parov nevronov, ki so bili odgovorni za proženje izhodnih nevronov ob stimulaciji določenih vhodnih nevronov. Temu pravimo tudi dodeljevanje zaslug (*credit assignment*), ki je eden temeljnih izzivov pri učenju impulznih nevronskih mrež. Pri klasičnih nevronskih mrežah bi v ta namen uporabili gradient, pri impulznih nevronskih mrežah pa to spremlja vrednost STDP, ki je pozitivna, če se presinaptični nevron sproži pred postsinaptičnim ($\Delta t > 0$) in negativna, če se presinaptični nevron sproži po postsinaptičnem ($\Delta t \leq 0$). Tako STDP služi kot groba aproksimacija gradienta. 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 sta A_+ in A_- multiplikatorja za potenciranje in depresijo, τ_+ in τ_- pa časovne konstante, ki določajo okno vpliva časovnih razlik.

Pare nevronov lahko posodabljamo izključno na podlagi vrednosti STDP, s katero krepimo ali slabimo posamezne sinapse, običajno pa to vrednost uporabljamo v kombinaciji z nagrado. Količina nagrade oziroma koncentracija nevromodulatorja bo za celotno impulzno nevronsko mrežo ali skupino

nevronov modulirala amplitudo učenja oziroma posodabljanja povezav. Povezavo med parom nevronov, bomo tako posodobili glede na vrednost STDP in prisotno količino nagrade v danem trenutku.

3.4.1 Dopaminska modulacija

V vseh sistemih, razvitih v tej diplomski nalogi, kot osnovo za sinapso, ki poleg vrednosti STDP upošteva tudi nagrado, uporabljamo R-STDP model sinapse (Izhikevich, E. M. 2007), kjer odvisnost od nagrade vpeljemo prek nevromodulatorja dopamina. Pogosto se koncentracija dopamina obravnava kot neposreden odraz količine nagrade oziroma vrednosti stanja, v katerem se agent nahaja, vendar ta modulira le plastičnost sinaps oziroma učenje. Namesto realizacije negativne nagrade z negativno koncentracijo dopamina, kar je pogost pristop, je biološko smiselneje izogibanje neželenim stanjem doseči z učenjem akcij, ki nas vodijo stran od teh stanj, ob pozitivni koncentraciji dopamina. Tak sistem je opisan v poglavju 5.1.

Pri R-STDP sinapsi vrednost STDP spremljamo prek sledi upravičenosti *eligibility trace* c, ki se za posamezen par nevronov ob proženju zviša glede na vrednost STDP, nato pa skozi čas odteka glede na parameter τ_c. To nam omogoča pripisovanje odgovornosti tudi ob zakasnjenih nagradah, glede na razliko med časom aktivnosti para navronov in časom dovedene nagrade. Tako bodo sinapse, ki so bile aktivne ravno pred zvišanjem koncentracije dopamina obravnavane (glede na vrednost STDP), kot odgovorne za aktivnost izhodnih nevronov, ki je potencialno povzročila zvišanje koncentracije dopamina oziroma nagrado. Sinapse, aktivne dlje v preteklost, bodo ustrezno posodobljene manj. Takemu učenju pravimo učenje s tremi dejavniki (*three-factor learning*), ki označi sinapse za potencialno spremembo in ohranja informacijo o pretekli aktivnosti pre- (prvi faktor) in postsinaptičnega nevrona (drugi faktor), dokler ne prispe tretji faktor, v tem primeru dopaminski signal n, ki predstavlja koncentracijo dopamina. Ta mehanizem olajša časovno dodelitev zaslug (*temporal credit assignment*), saj omogoča povezovanje hitre aktivnosti nevronov s počasnejšimi vedenjskimi odzivi, s

čimer sinapse prejmejo ustrezno posodobitev, tudi če je nagrada ali signal zakasnjen.

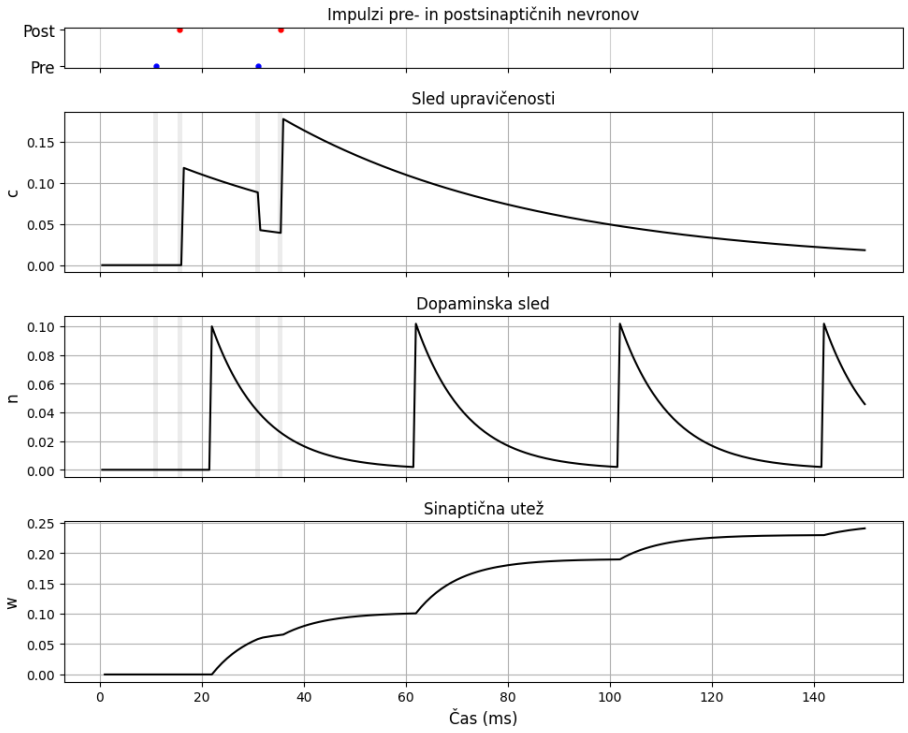
V primerjavi s klasičnimi TD (*temporal difference - časovo razlikovalnimi*) algoritmi v ne-spiking nevronskih mrežah, kjer se sledi upravičenosti uporabljajo za posodobitev stanj ali akcij v diskretnih časovnih korakih, je v spiking neural networks sled c neposredno vezana na dejanske sprožitve impulzov in se lahko sproži ob vsaki interakciji med pre- in postsinaptičnim nevronom. Tako sled uprabičenosti predstavlja lokalno, biološko interpretabilno komponento učenja, ki je skladna z eksperimentalnimi dokazi o treh faktorjih učenja v možganih.

Dopaminsko koncentracijo predstavlja dopaminska sled n, ki jo lahko definiramo kot vrednost, ki spremlja aktivnost posebnih dopaminergičnih nevronov. Ob zvišani aktivnosti dopaminergičnih nevronov se bo dopaminska sled zvišala, ob odsotnosti aktivnosti, pa bo odtekala glede na parameter τ_n. Dopaminska sled je uporabljena za neposredno modulacijo velikost in smeri sinaptične plastičnosti, tj. velikosti in predznaka posodobitve uteži povezave. Sinaptična dinamika R-STDP sinapse je opisana z enačbami po Potjans, Morrison in Diesmann 2010:

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

Enačbe opisujejo dinamiko treh ključnih količin: sinaptične uteži w, sledi upravičenosti c in dopaminske sledi n. Sinaptična utež w se neposredno posodablja kot produkt trenutne sledi upravičenosti c in odstopanja dopaminske koncentracije n od bazalne vrednosti b. Sled c spremlja pare sproženih pre- in postsinaptičnih nevronov ter aproksimira odgovornost posamezne sinapse za proženje postsinaptičnega nevrona kot posledico aktivnosti presinaptičnega nevrona. Njena dinamika vključuje eksponentno odtekanje s časovno konstanto τ_c ter impulzne spremembe ob vsakem sproženju nevronov, modulirane s konstanto C_1. Dopaminska sled n pa se eksponentno zmanjša z lastno

časovno konstanto τ_n in se poveča ob sprožitvah dopaminskih nevronov, ki se pojavijo ob časih s_n in so skalirani s konstanto C_2. 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: Sled upravičenosti 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, $\tau_n = 10.0$ 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. S slik lahko vidimo pozitivno sled upravičenosti zaradi proženja presinaptičnega nevrona pred postsinaptičnim. Ob prisotnosti neničelne dopaminske sledi n ob proženju dopaminergičnega nevrona vsakih 40 ms, se sinaptična utež w okrepi.

Poglavje 4

Spodbujevano učenje na impulznih nevronskih mrežah

V delu uporabljamo 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 prek posodabljanja povezav, ki so bile aktivne v prejšnjem stanju in so bile odgovorne za akcijo, ki nas je pripeljala v naslednje stanje.

4.1 R-STDP učenje

R-STDP (*Reward modulated spike timing dependent plasticity*) 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 impulzov postsinaptičnih nevronov kot direktna posledica proženja presinaptičnih nevronov. Take povezave imajo ob času prihoda nagrade najvišjo vrednost sledi upravičenosti.

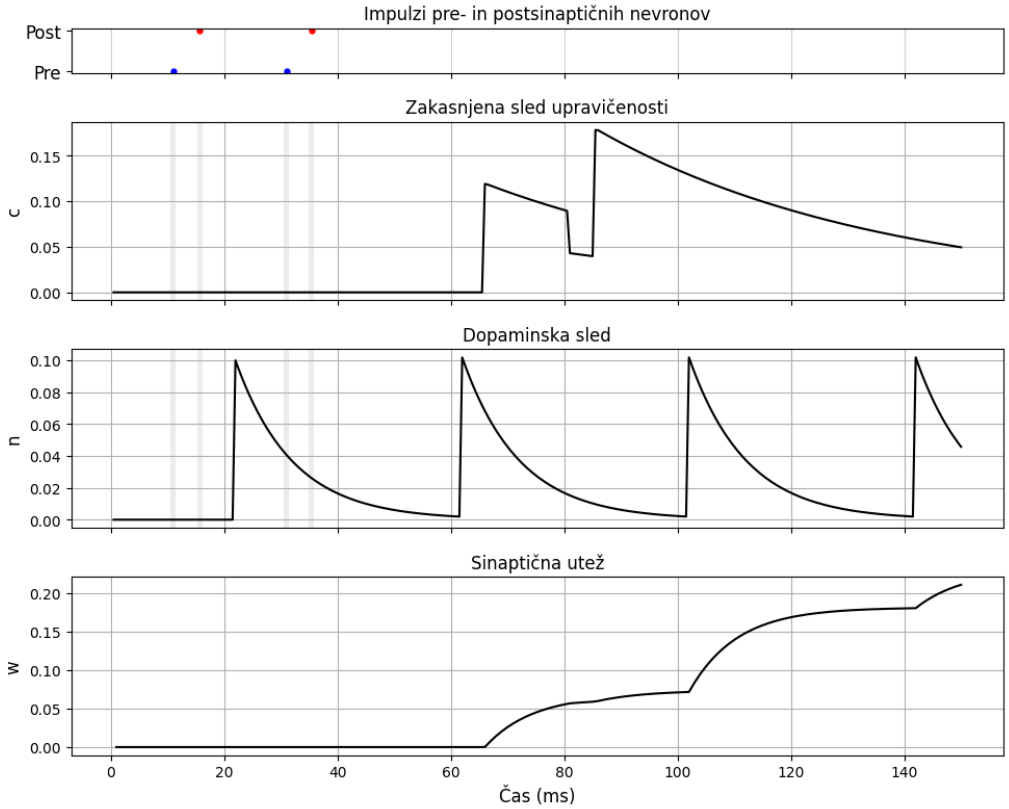
Naš agent je za začetek sestavljen iz N_{in} vhodnih nevronov, ki predsta-

vljajo 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 želenih 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 $\tau_{c,\text{delay}}$ in s tem onemogočili posodabljanje zaradi nagrad, ki so prispele v času krajšem od $\tau_{c,\text{delay}}$ po aktivnosti sinapse. Dinamika prilagojene sinapse je prikazana na sliki 4.1.

Nagrada, ki jo določa aktivnost dopaminskih nevronov, bo vedno večja ali enaka 0, kar pomeni, da se bodo sinapse s časom le krepile. Povezave, odgovorne za izbiro določene akcije v določenem stanju, morajo zato med seboj tekmovati za prevlado. Ob prisotnosti nagrade moramo povezave, ki so odgovorne za izbiro pravilne akcije, napram ostalim povezavam okrepiti dovolj, da bodo v prihodnje prevladale nad drugimi akcijami in povečale verjetnost izbire pravilne akcije. Razlika v amplitudi posodobitev posameznih povezav je v našem primeru, kjer je koncentracija dopamina za vse povezave v danem trenutku enaka, določena izključno s sledmi upravičenosti.

Sled upravičenosti bo višja za povezave z višjimi utežmi, saj take povezave povzročajo več impulzov postsinaptičnega nevrona kot neposredna posledica proženja presinaptičnih nevronov. Pri nižjih vrednostih uteži, ko so vse povezave približno enake, pa dovolj veliko razliko v sledi upravičenosti dosežemo prek variance v frekvenci impulzov. Ta povzroči, da nekatere si-



Slika 4.1: Sled upravičenosti c, dopaminska sled n in posodabljanje sinaptične uteži pri presinaptičnih impulzih pri $[10.0, 30.0]$ ms in postsinaptičnih inpulzih 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. S slik lahko vidimo pozitivno sled upravičenosti zaradi proženja presinaptičnega nevrona pred postsinaptičnim, vendar zamaknjeno za $tau_{c,\text{delay}} = 50$ ms. Ob prisotnosti neničelne dopaminske sledi n ob proženju dopaminergičnega nevrona vsakih 40 ms, se sinaptična utež w okrepi, vendar šele po zakasnitvi glede na impulze pre- in postsinaptičnega nevrona. Tako nagrada, ki je prispela prehitro ne povzroči okrepitve sinapse.

napse po naključju prispevajo k akciji nekoliko bolj ali manj. To pomeni, da se vrednosti sledi upravičenosti razlikujejo med sinapsami, posledično pa lahko dopaminska modulacija posodobi sinapse različno in vzpostavi razlikovanje med njimi, tudi če imajo vse povezave enake začetne uteži. Pri nižjih vrednostih uteži, kjer deterministični prispevek sinaps še ni dominanten, naključna varianca impulzov služi kot mehanizem, da "razbije simetrijo" in omogoči, da se sledi upravičenosti razlikujejo, dokler uteži niso dovolj velike, da prevladujejo deterministični prispevki.

Pogosto se razlikovanje med sinapsami doseže z dopuščanjem negativih nagrad, ki sinapse, ki so bile odgovorne za napačno izbrano akcijo negativno posodobijo. V našem sistemu negativnih nagrad, ki zahtevajo negativno koncentracijo dopamina ne bomo dopuščali, saj v naravi negativna koncentracija dopamina ni mogoča. Varianco v frekvenci impulzov bomo tako dosegli s Poissonovim šumom. V diferencialni enačbi, ki določa spreminjanje V_m pri nižjih vrednostih uteži nad I_{syn} prevladuje zunanji konstantni tok I_e, ki ga pri nas povzroča generator Poissonovega šuma. Tako ima, dokler ne prevladuje deterministični prispevek okrepljenih povezav, Poissonov šum večji vpliv na V_m in je tako varianca v frekvenci impulzov večja.

To služi tudi kot mehanizem za raziskovanje (*exploration*), ki bo pri nižjih utežeh, ko imajo vse povezave približno enake uteži omogočil naključno izbiranje akcije. Tekom učenja, ko se sinapse okrepijo, bo agent akcije izbiral v glavnem glede na prevladujoče sinapse, kar predstavlja izkoriščanje (*exploitation*) naučene politike (verjetnosti izbire posamezne akcije v določenem stanju). Naš sistem bo tako tekom učenja postopoma višal razmerje med izkoriščanjem in raziskovanjem.

Za model nevrona, ki ga bomo uporabili v nadaljevanju, bomo izbrali model nevrona z eksponentnim jedrom, saj Poissonov šum v tem primeru povzroči večjo varianco izhodnih nevronov kot model z alfa jedrom. To smo pokazali v poglavju 3.3.

Za uspešno učenje je glavni izziv izbira pravilnih hiperparametrov sistema, ki omogoča, da se sinapse diferecirajo dovolj hitro. Ker ne dopuščamo

negativnih nagrad bodo vse povezave s časom rasle, kar postopoma zmanjšuje verjetnost izbire naključne akcije. Tako se ob neustrezni izbiri parametrov lahko zgodi, da se vse povezave približno enakomerno krepijo, dokler raziskovanja praktično ni več in bo izbira akcije odvisna od naključno prevladujoče poti od vhodnih nevronov do izhodnih. Pravilne hiperparametre smo v nadaljevanju iskali eksperimentalno, kjer so najbolj vplivni hiperparameteri frekvenca in utež povezave iz generatorja Poissonovega šuma do izhodnih nevronov, parametra A_+ in A_- ter parameter τ_c.

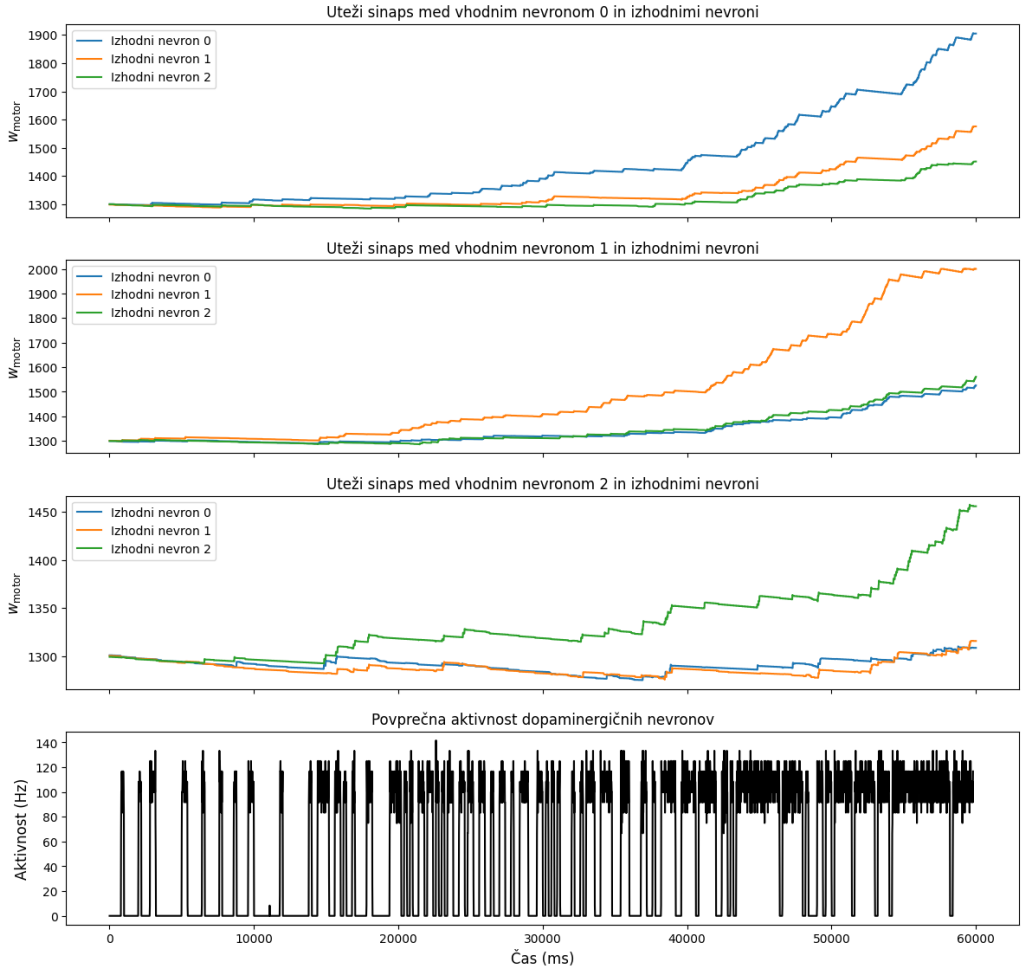
4.1.1 Simulator NEST

V nadaljevanju bomo vse razvite sisteme implementirali s pomočjo simulatorja NEST. Zaledje in posamezne komponente simulatorja so implemntirani v C++, kar nam omogoča učinkovito simulacijo impulznih nevronskih mrež. Za namen uporabe zakasnjene sinapse R-STDP, ki smo jo predstavili v prejšnjih poglavjih, moramo za integracijo z ostalimi komponentami simulatorja napisati poseben C++ modul.

Za začetek bomo R-STDP učenje implementirali na preprosti nalogi s tremi stanji, oštevilčenimi z 0, 1 in 2. V vsakem stanju lahko podobno agent izbere akcijo 0, 1 ali 2. Prehod v stanje po izbiri akcije je nakjučno (neodvisen od izbire akcije), v vsakem stanju pa je nagrajena le ena izbira akcije. V stanju 0, ki ga predstavlja stimulacija vhodnega nevrona 0, bomo kot nagrajeno akcijo izbrali akcijo 0, ki jo predstavlja izhodni nevron 0, v stanju 1 akcijo 1, v stanju 2 pa akcijo 2. Na sliki **??** je razvidna prevlada pravilnih sinaps, višanje divergence v sinapsah skozi čas ter rast frekvence nagrad tekom učenja. V simulaciji uporabljamo privzete parametre NEST za nevrone z eksponentnim jedrom (*iaf_psc_exp*) ter zakasnjene dopaminsko modulirane sinapse. Ostali parametri sistema so navedeni v tabeli **??**.

Simbol	**Pomen**	**Vrednost**
t_{SIM}	Trajanje simulacije	60000 ms
$w_{\text{motor, min}}$	Minimalna utež sinaps med vhodnimi in izhodnimi nevroni	500
$w_{\text{motor, max}}$	Maksimalna utež sinaps med vhodnimi in izhodnimi nevroni	2000
τ_c		5 ms
$\tau_{c,\text{delay}}$		200 ms
τ_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
λ_{motor}	Povprečna hitrost Poissonovega šuma	1000 Hz
w_{Poisson}	Uteži sinaps Poissonovega šuma	100
$w_{\text{init, motor}}$	Začetne uteži sinaps 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.



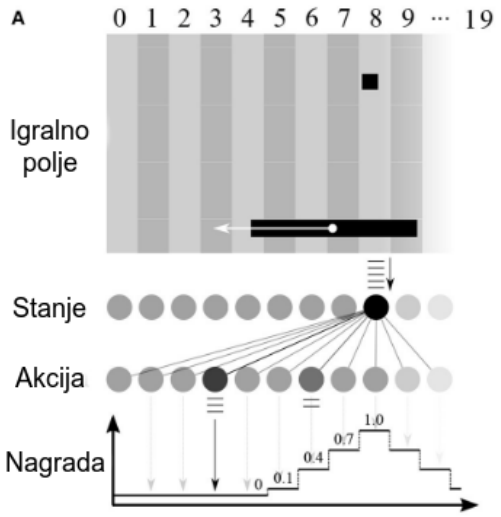
Slika 4.2: Prikaz uteži sinaps med vhodnimi in izhodnimi nevroni ter povprečne aktivnosti dopaminergičnih nevronov med simulacijo R-STDP učenja na preprostem problemu. Preko uteži vidimo, da sinapse med vhodnimi nevroni in izhodnimi nevroni, ki predstavljajo pravilno akcijo skozi čas prevladajo nad ostalimi akcijami. Posledično se frekvenca izbire pravilne akcije v posameznem stanju viša, kar je razvidno iz grafa povprečne aktivnosti dopaminergičnih nevronov, ki postaja vse "gostejši".

4.1.2 Igra Pong

V nadaljevanju bomo R-STDP učenje predstavili na kompleksnejšem problemu: igri *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 lahko uporabimo TD (*angl. Temporal Difference*) učenje, ki ga implementiramo v poglavju **??**, za zdaj pa bomo igranje igre Pong, ki sicer zahteva predvidevanje in ima zakasnjeno nagrado, preslikali na problem s takojšnjimi nagradami. Igro bomo definirali tako, da ima žogica stalno hitrost, določeno smer in pozicijo v (x, y) ravnini. Na eni strani igrišča bo naš agent premikal platformo v horizontalni smeri, na drugi pa je stena, od katere se žogica prožno odbije. Takšna konfiguracija je prikazana na sliki **??**. Če bi od agenta zahtevali predvidevanje, bi morali stanja definirati kot kartezični produkt (x, y) pozicije žogice, njene smeri in x pozicije platforme. Problem bomo poenostavili v problem sledenja žogici, kot v delu Wunderlich T, et al. 2019, kjer agent izbira želeno ciljno točko platforme. Predvidevanje zato ni potrebno. Tako stanja kot akcije agenta so diskretizirane možne x 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 želene pozicije j od trenutne x pozicije žogice k, in povprečno nagrado $\bar{R}_i$ v iteraciji i. S pomočjo povprečne nagrade omejimo krepitev sinaps, če te ne izboljšajo trenutne politike.

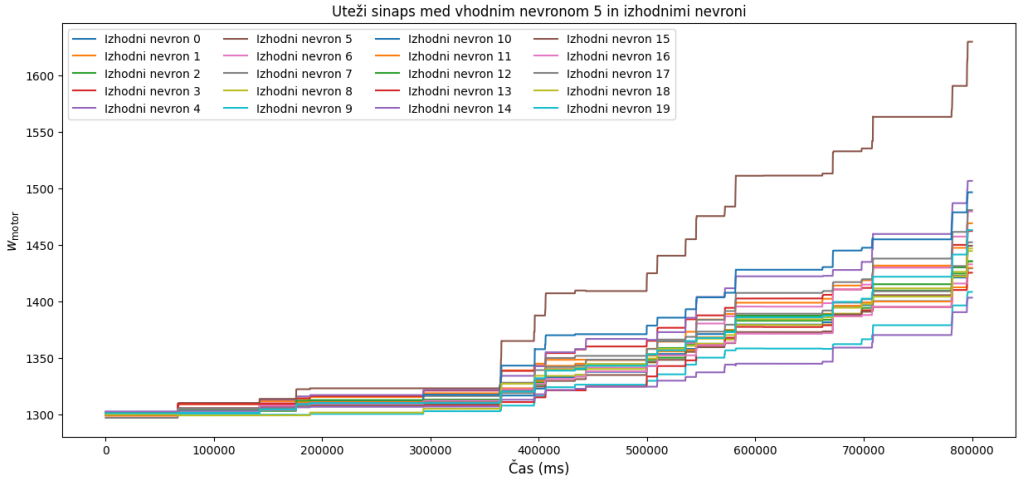
$$R_b = \begin{cases} 1 - |j - k| \cdot 0.3 & \text{if } |j - k| \leq 3, \\ 0 & \text{sicer.} \end{cases} \tag{4.1}$$

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



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 x pozicijo, kot jo ima takrat žogica. Polje smo po x osi diskretizirali na 20 stanj. Po simulaciji, ki je trajala 4000 iteracij (pozicij žogice) po 200 ms, na sliki **??** vidimo graf povezav med vhodnim nevronom, ki predstavlja pozicijo $x = 5$ 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.

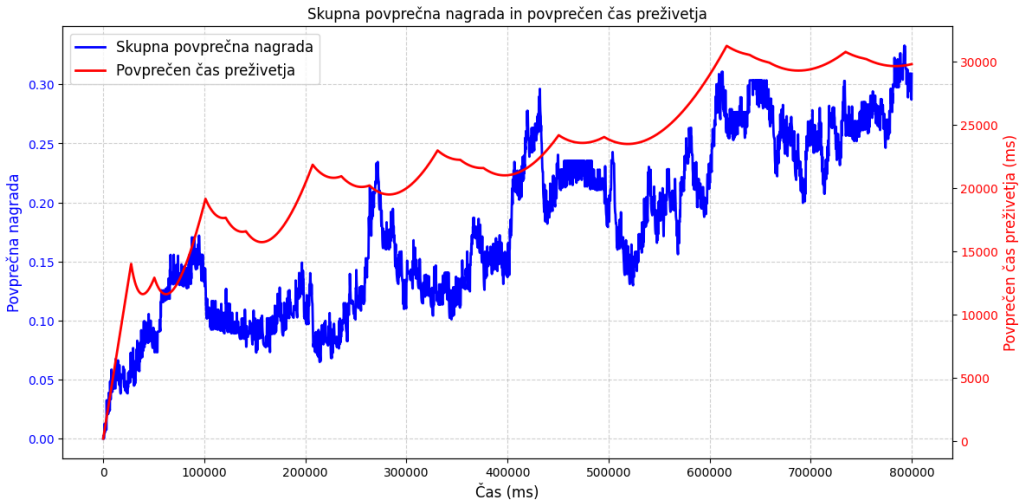


Slika 4.4: Graf uteži povezav med vhodnim nevronom 5 in izhodnimi nevroni tekom simulacije 4000 iteracij (pozicij žogice) po 200 ms igranja igre Pong. Tekom učenja pričakovano prevladuje povezava do izhodnega nevrona 5, kar predstavlja sledenje žogici.

Rezultati

Učenje spremljamo s povprečnim časom preživetja, ki predstavlja čas, ki je minil od zadnje zgrešitve žogice ali začetka igre. Povprečni čas preživetja je izračunan na podlagi zaporednih časov med posameznimi zgrešenimi odboji žogice. Ob vsaki iteraciji se čas od zadnjega zgrešenega udarca poveča za dolžino koraka simulacije (200 ms), ob zgrešenem udarcu pa se ponastavi na nič. Tako dobimo zaporedje časov preživetja, ki predstavlja, kako dolgo je agent uspešno odbijal žogico med dvema zaporednima napakama. Povprečni čas preživetja je nato določen kot kumulativno povprečje teh vrednosti skozi čas, kar pomeni, da pri vsaki časovni točki upoštevamo povprečje vseh do tedaj izmerjenih časov preživetja. 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. Pri skaliranih vrednostih povprečne nagrade in povprečnega časa preživetja lahko na sliki **??** vidimo, da imata krivulji podobno obliko in da se skozi čas obe višata.

Ker povprečna nagrada neposredno odraža izbire pravilnih akcij v določenem stanju, lahko, pri pravilno določeni funkciji nagrajevanja, za poljuben problem kvaliteto učenja spremljamo zgolj s skupno povprečno nagrado.



Slika 4.5: Skupna povprečna nagrada $\bar{R}_i$ in povprečen čas preživetja tekom 4000 iteracij po 200 ms. Pri usklajenem razponu obeh vrednosti je razvidno, da imata obe krivulji približno enako obliko in da obe skozi čas naraščata.

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 postavili v naključno stanje. Končno politiko agenta bomo vizualizirali tako, da bomo v vsakem polju mreže i prikazali preferirano smer. Verjetnost izbire posamezne akcije aproksimiramo s povprečno utežjo povezave od vhodnega do izhodnega nevrona. Preferirano smer za posamezno stanje i bomo pri-

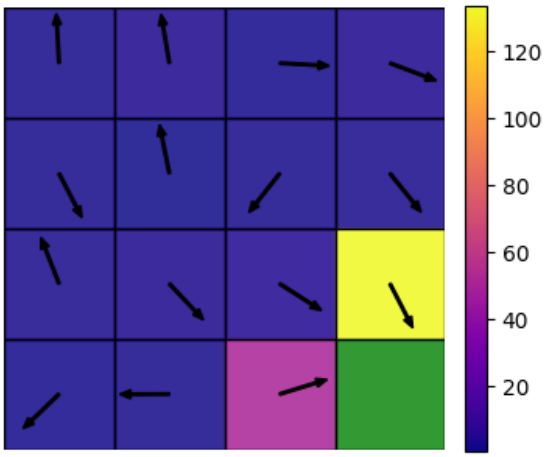
kazali z normaliziranim vektorjem $\hat{x}_i$. Pričakujemo, da bodo vektorji stanj neposredno ob cilju kazali v smer cilja.

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

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

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

Indikator učenja pravilne akcije v stanju i je tudi maksimalna razlika med utežmi med vhodnim nevronom i in vsakim od izhodnih nevronov, ki predstavlja aproksimacijo verjetnosti izbire akcije. Pri dobri politiki želimo, da bo verjetnost izbire pravilne akcije ustrezno večja od ostalih. Pri R-STDP učenju, kot smo ga implementirali do sedaj, pričakujemo opazno diferenciacijo akcij pri stanjih neposredno ob cilju, za ostala stanja pa pričakujemo, da bo makcimalna razlika med utežmi minimalna, saj v teh stanjih ne dovajamo nagrade, ki bi spodbujevala učenje. Vektorji posameznih stanj in maksimalne razlike med utežmi so prek barve polj prikazani na sliki 4.6. Slika prikazuje politiko po učenju 500 iteracij na 4x4 mreži in potrjuje pričakovane rezultate, kjer imajo veliko maksimalno razliko med utežmi in pravilno smer vektorjev le polja neposredno ob cilju. Rezultat 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 modela na 4x4 mreži po 500 iteracijah po 200 ms. Končno stanje je obarvano z zeleno. Barva ostalih stanj prikazuje maksimalno razliko med utežmi med vhodnim nevronom, ki predstavlja to stanje ter posameznimi izhodnimi nevroni, ki predstavljajo akcije. V vsakem polju je s pučico prikazan vektor preferirane akcije (smeri) glede na uteži med vhodnim nevronom in posameznimi izhodnimi nevroni. Vidimo, da so svetlo obarvana (imajo veliko maksimalno razliko med utežmi, ki predstavljajo posamezne akcije) le polja neposredno ob cilju. Podobno imajo tudi le ta polja pravilno obrnjene vektorje.

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 neposredno interakcijo z okoljem. Ključna prednost TD-učenja je njegova sposobnost učenja iz oddaljenih nagrad, saj omogoča postopno razširjanje informacije o nagradi nazaj skozi zaporedje predhodnih stanj. Namesto čakanja na končni izid epizode (sekvence prehajanja stanj, do prihoda v končno stanje) TD-učenje primerja trenutno oceno vrednosti s t. i. TD-napako δ_t, ki temelji na naslednjem stanju in morebitni takojšnji nagradi. Na ta način se tudi dejanja, ki sama po sebi ne prinesejo takojšnje nagrade, a vodijo do kasnejšega uspeha, sčasoma ustrezno ovrednotijo.

Osnovna posodobitvena enačba za vrednostno funkcijo stanja:

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

kjer je α hitrost učenja, TD-napaka δ_t pa je definirana kot

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

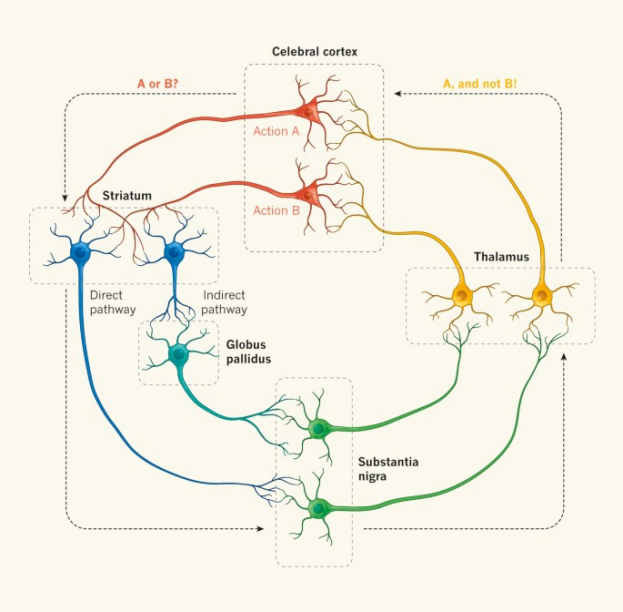
V izrazu je r_{t+1} nagrada ob prehodu iz stanja s_t v stanje s_{t+1}, faktor $\gamma \in [0, 1]$ pa določa relativno težo prihodnjih nagrad. TD-napaka predstavlja razliko med izboljšano napovedjo vrednosti in prejšnjo oceno.

V tem delu bomo implementirali algoritem TD(0) pri katerem se vrednosti posodabljajo izključno na podlagi trenutne nagrade in ocene vrednosti naslednjega stanja. Izbiro akcije v posameznem stanju bomo glede na nagrado posodabljali po enakem R-STDP mehanizmu kot doslej, kjer bomo izbire akcij nagrajevali s pomočjo sledi upravičenosti. Tako bo pravzaprav možno tudi, da nagradimo sinapse, ki so predstavljale izbire akcij v stanjih dlje v preteklost, kar je ideja splošnejše oblike TD-učenja - metodo TD(λ). S tem, da stanja definiramo z 200 ms stimulacijo vhodnega nevrona, pa bomo z ustrezno izbrano konstanto τ_c, ki določa odtekanje sledi upravičenosti lahko

dosegli minimalen vpliv sinaps, ki ne predstavljajo izbire akcije v trenutnem stanju in se tako približali TD(0). To nam bo namreč močno olajšalo spremljanje učenja in iskanje ustreznih hiperparametrov, kar je sicer pri impulznih nevronskih mrežah lahko zahtevno.

4.2.1 Model akter-kritik

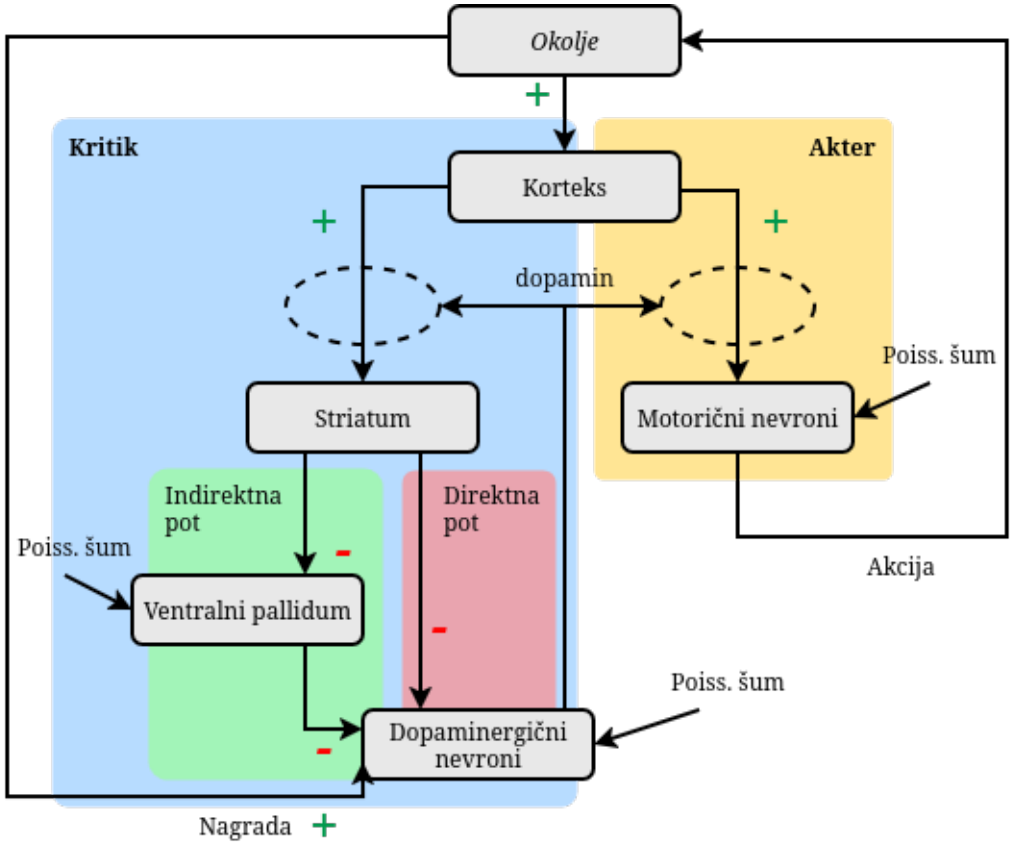
Za TD učenje, potrebujemo poleg že razvitega R-STDP modula za učenje pravilne izbire akcije tudi modul, ki računa TD-napako δ_t. Takemu modelu pravimo akter-kritik (*angl. actor-critic*). V tej nalogi skušamo doseči čimbolj biološko verjetne implementacije, zato želimo, da impulzna nevronska mreža sama računa TD-napako. Za to se bomo zgledovali po bazalnih ganglijih, prikazanih na sliki 4.7, skupini nevronov v človeških možganih, ki med drugim realizira obliko TD-učenja. Za poenostavitev in abstrakcijo kompleksnega sistema in mehanizmov prisotnih v pravih bazalnih ganglijih, se bomo zgledovali po Wiebke P, et al. 2011. Cilj sistema, ki ga bomo razvili v nadaljevanju bo, na nalogi z mrežo, doseči propagiranje zakasnjene nagrade v stanja, ki postopno vodijo do cilja in s tem učenje optimalne politike za navigacijo do nagrade.



Slika 4.7: Diagram skupin nevronov bazalnih ganglijev.

Kot omenjeno, je model akter-kritik kot ga predstavlja Wiebke P, et al. 2011 poenostavitev in abstrakcija resničnih mehanizmov v možganih, ki to omogočajo. V bazalnih ganglijih, kot so prikazani na sliki 4.7, pri tem sodeluje več skupin nevronov in povezav, ki so v tem modelu 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* pravih 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 ganglijih 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. Opisan akter-kritik model je prikazan na sliki 4.8.



Slika 4.8: Model akter-kritik, kot ga predlaga Wiebke P, et al. 2011.

Ob prisotnosti osnovne, stalno prisotne neničelne frekvence nevronov ventralnega palliduma, ki jo povzročimo z injeciranim Poissonovim šumom, bo aktivnost indirektne pot imela na dopaminergične nevrone vzbujajoč učinek. Aktivnost nevronov venralnega palliduma namreč prek inhibitorne povezave znižuje aktivnost dopaminergičnih nevronov. Če prek inhibitorne povezave iz striatuma zmanjšamo aktivnost nevronov ventralnega palliduma, pa bo aktivnost dopaminergičnih nevronov narasla. Indirektna in direktna pove-

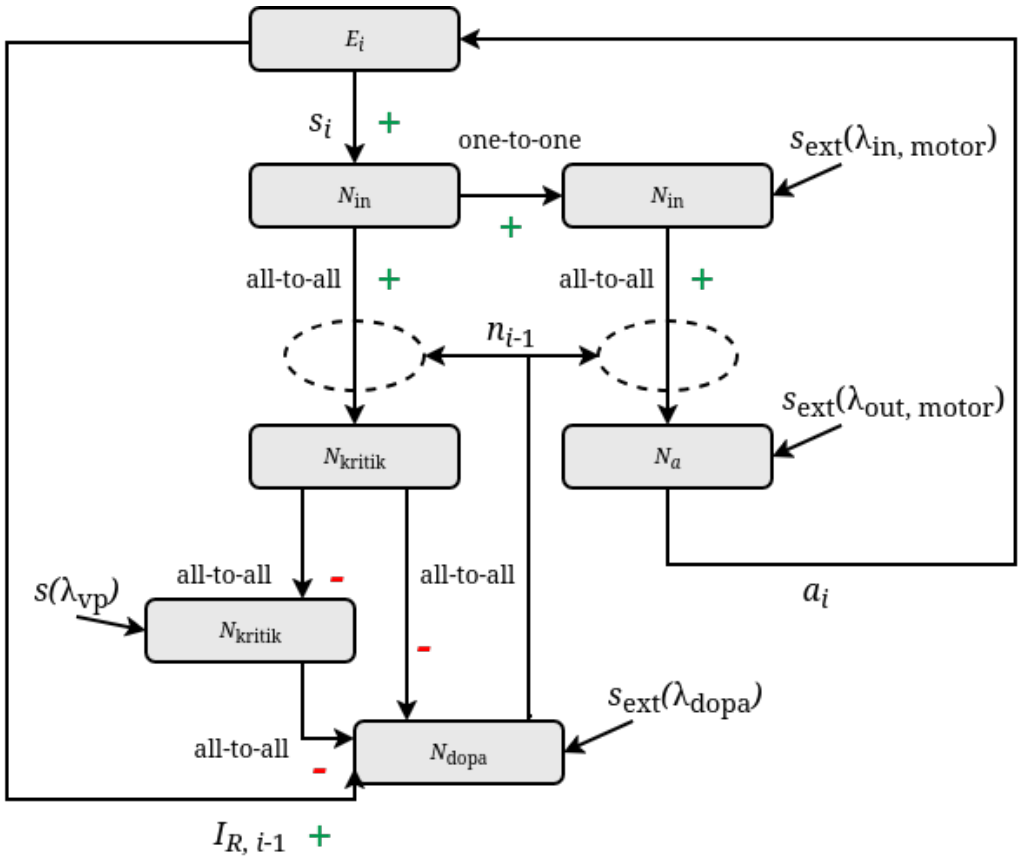
zava na dopaminergične nevrone delujeta konkurenčno. Sinapse na indirektni poti imajo minimalen zamik zaradi prenašanja signala. Indirektna pot tako zvišano aktivnost striatuma v trenutnem stanju z minimalnim zamikom preslika v povišano aktivnost dopaminergičnih nevronov. Hkrati v času nahajanja v trenutnem stanju direktna povezava inhibira dopaminergične nevrone sorazmerno z aktivnostjo striatuma, vendar glede na takšno, kakršna je bila v prejšnjem stanju, saj je direktna pot zakasnjena. Indirektna in direktna povezava skupaj računata TD-napako δ_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 4.10.

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.

Implementacija

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 prek statičnih povezav z utežmi $w_{\text{in}\to\text{in, motor}}$ 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 ustreznih hiperparametrov. Vmesni nevroni so po režimu *all-to-all* (vsak

vhodni nevron je povezan z vsakim izhodnim) povezani z izhodnimi nevroni prek zakasnjenih R-STDP sinaps z normalno porazdeljenimi utežmi. Prav tako so vhodni nevroni po režimu *all-to-all* povezani z N_{kritik} nevroni striatuma prek zakasnjenih R-STDP sinaps z normalno porazdeljenimi utežmi. Striatum je prek statičnih inhibitornih povezav z utežmi $w_{\text{str}\to\text{vp}}$ povezan z N_{kritik} nevroni ventralnega palliduma in prek statičnih inhibitornih povezav z utežmi $w_{\text{str}\to\text{dopa}}$ ter zakasnitvijo d_{dir} z N_{dopa} dopaminergičnimi nevroni. Ventralni pallidum je z dopaminergičnimi nevroni prav tako povezan prek statičnih inhibitornih povezav, ki pa niso zakasnjene in imajo uteži $w_{\text{vp}\to\text{dopa}}$. 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 λ_{vp} in v dopaminergične nevrone s povprečno hitrostjo λ_{dopa}. Generatorji Poissonovega šuma so s posameznimi skupinami nevronov povezani prek statičnih povezav z utežmi $w_{\text{ext, in}}$, $w_{\text{ext, out}}$, $w_{\text{ext, vp}}$ in $w_{\text{ext, dopa}}$. Hiperparametri R-STDP sinaps in modelov nevronov so skupaj z ostalimi hiperparametri navedeni v poglavju 4.2.2. Opisani model je prikazan na sliki 4.9.



Slika 4.9: Prikaz implementiranega sistema akter-kritik.

Za razliko od izvornega sistema (Wiebke P, et al. 2011), bomo za model sinapse uporabili našo zakasnjeno sinapso R-STDP, kot smo jo razvili v poglavjih 3.4 in 4.1, ki poleg presinaptičnih impulzov upošteva tudi postsinaptične po pravilu STDP. Tako bomo lahko kot akter uporabili sistem R-STDP, 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 stanju. V izvornem 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 neposre-

dno okrepili povezave, odgovorne za izbrano aktivnost, saj smo z inhibicijo dosegli ničelne sledi upravičenosti sinaps do ostalih izhodnih nevronov, saj se te ne bodo prožili. V tem primeru 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 poženemo ukaz `nest.Simulate(0.1)` ali enkrat `nest.Simulate(10)`. Naš sistem bo zaradi hitrosti simulacije po številu nevronov manjši od izvornega 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 ohranjanjamo osnovno frekvenco dopaminergičnih nevronov ter ravnovesje med inhibicijo in vzbujanjem zaradi direktne in indirektne 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 **??**-**??** so navedene konstante in parametri implementiranega modela. Parametri, ki niso prikazani v tabelah, imajo privzete vrednosti simulatorja NEST.

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

Tabela 4.2: Parametri simulacije

Simbol	Pomen	Vrednost
Parametri skupin nevronov kritika		
tip	Tip modela nevrona	*iaf_psc_alpha*
$C_{m,\text{in}}$	Membranska kapacitivnost	250.0 pF
$\tau_{m,\text{in}}$	Časovna konstanta membrane	10.0 ms
$V_{\text{reset,in}}$	Potencial ponastavitve	0.0 mV
$V_{\text{th,in}}$	Prag proženja	20.0 mV
$t_{\text{ref,in}}$	Refraktorna doba	0.5 ms
$\tau_{\text{syn,ex,in}} = \tau_{\text{syn,in,in}}$	Ekscitatorna in inhibitorna sinaptična konstanta	2 ms
$\tau_{\text{-,a}}$	Negativna STDP konstanta	20.0 ms
$V_{m,\text{in}}$	Začetni membranski potencial	0.0 mV
$E_{L,\text{in}}$	Mirovalni potencial	0.0 mV
Parametri motoričnih nevronov		
tip	Tip modela nevrona	*iaf_psc_exp*
$C_{m,a}$	Membranska kapacitivnost motornih nevronov	250.0 pF
$\tau_{m,a}$	Časovna konstanta membrane	10.0 ms
$V_{\text{reset,a}}$	Potencial ponastavitve	0.0 mV
$V_{\text{th,a}}$	Prag proženja	20.0 mV
$t_{\text{ref,a}}$	Refraktorna doba	0.1 ms
$\tau_{\text{syn,ex,a}} = \tau_{\text{syn,in,a}}$	Ekscitatorna in inhibitorna sinaptična konstanta	2 ms
$\tau_{\text{-,a}}$	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
Parametri sinaps med vhodnimi in vhodnimi motoričnimi nevroni		
tip	Tip sinapse	Privzeta konstantna sinapsa NEST
$w_{\text{in}\to\text{in, motor}}$	Uteži sinaps med vhodnimi in vhodnimi motoričnimi nevroni	120
Parametri sinaps med vhodnimi in izhodnimi motoričnimi nevroni		
tip	Tip sinapse	Zakasnjena dopaminsko modulirana STDP sinapsa
τ_c	Odtekanje *eligibility* sledi	5 ms
$\tau_{c,\text{ delay}}$	Zakasnitev sledi c	200 ms
τ_n	Odtekanje dopaminske sledi	10 ms
τ_+	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
$w_{\text{in, motor}\to a}$	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
Parametri sinaps med vhodnimi nevroni in striatumom		
tip	Tip sinapse	Zakasnjena dopaminsko modulirana STDP sinapsa
τ_c	Odtekanje *eligibility* sledi	5 ms
$\tau_{c,\text{ delay}}$	Zakasnitev sledi c	200 ms
τ_n	Odtekanje dopaminske sledi	10 ms
τ_+	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
$w_{\text{in}\to\text{str}}$	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 sinapsa NEST
$w_{\text{str}\to\text{vp}}$	Uteži sinps med striatumom in ventral pallidumom	-50
$w_{\text{str}\to\text{dopa}}$	Uteži sinps med striatumom in dopaminergičnimi nevroni	-55
$w_{\text{vp}\to\text{dopa}}$	Uteži sinps med ventral pallidumom in dopaminergičnimi nevroni	-65
d_{dir}	Zakasnitev sinaps direktne povezave	200 ms

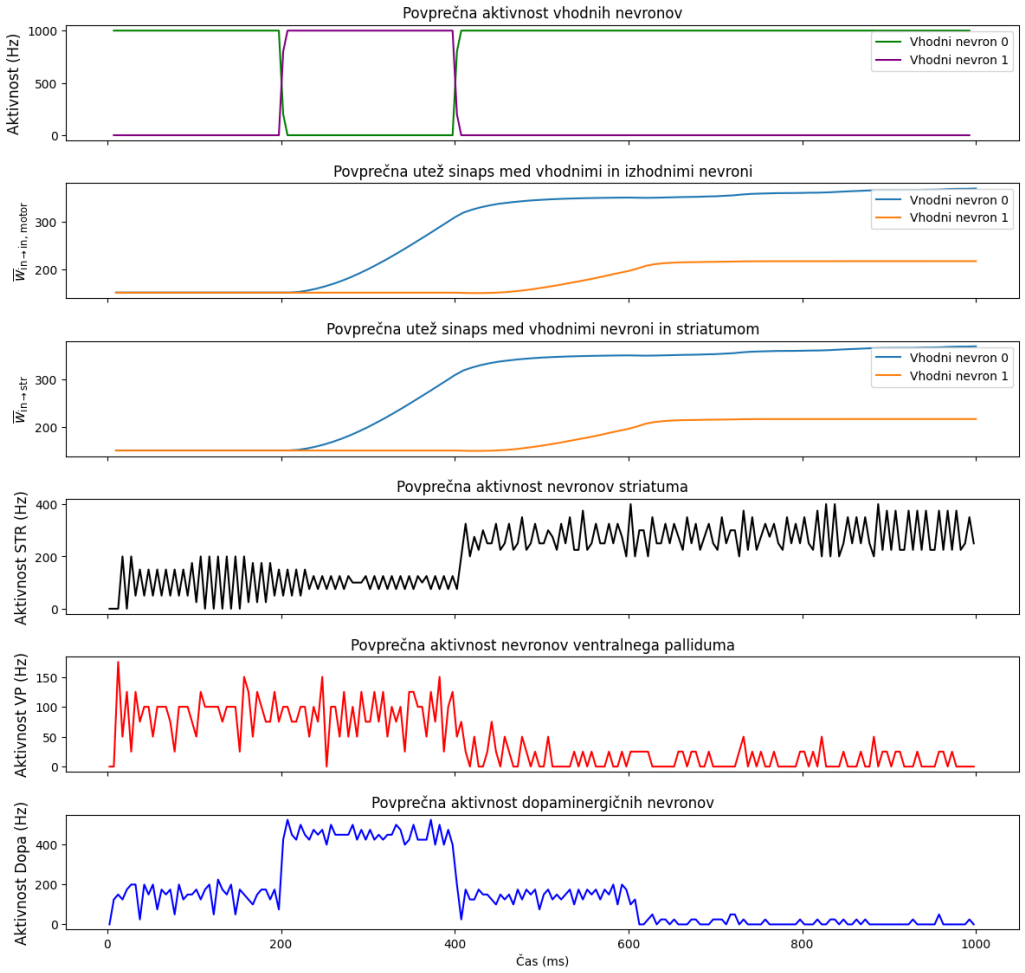
Tabela 4.6: Parametri sinaps kritika

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

Tabela 4.7: Parametri generatorjev šuma

4.2.3 Učenje

Posamezno stanje i bo za akterja predstavljala 200 ms stimulacija vhodnega nevrona i, enako kot v poglavju 4.1. Mehanizme, opisane v prejšnjem poglavju, najprej preverimo na sekvenci prehodov med dvema stanjema 0 in 1. Začeli bomo v stanju 0, se premaknili v stanje 1, kjer agent prejme nagrado, nato pa se vrnemo nazaj v stanje 0, kjer ostanemo. Aktivnosti posameznih skupin nevronov in posodabljanje uteži sinaps akterja in kritika so prikazane na sliki 4.10. Ob prehodu v nagrajeno stanje 1 vidimo, da se uteži povezav med vhodnim nevronom, ki predstavljaja stanje 0 in striatumom okrepijo. To predstavlja zvišanje pričakovane nagrade v stanju 0, nagrada, ki smo jo dovedli ob prehodu v nagrajeno stanje 1, pa ne pomeni, da tudi v stanju 1 pričakujemo visoko nagrado. Pričakovana nagrada se namreč zviša ob prehodu v stanje z višjo vrednostjo ali ob zunanji nagradi in je odvisna od akcije, ki jo izvedemo. Ob prehodu iz stanja 1 nazaj v stanje 0 prehajamo iz stanja z osnovnimo pričakovano nagrado v stanje 0, ki ima tokrat okrepljene uteži do striatuma in povišano pričakovano nagrado. To predstavlja prehod v stanje z višjo vrednostjo, posledica tega pa je, v primerjavi z osnovno frekvenco dopaminergičnih nevronov, zvišana dopaminergična aktivnost. Ta povzroči sorazmerno povišanje uteži sinaps do striatuma v stanju 1. V nadaljevanju ostajamo v stanju 0, kjer se ob prehodu iz stanja 0 v stanje 0 vrnemo k osnovni dopaminergični frekvenci, saj se vpliv direktne in indirektne povezave izničita.

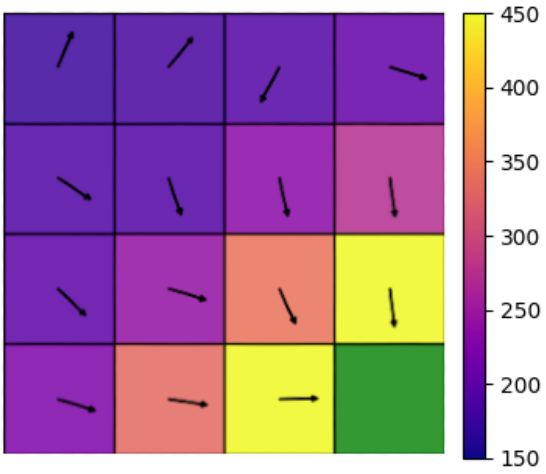


Slika 4.10: Prikaz posodabljanja sinaptičnih uteži ob prehajanju med dvemi stanji. Na grafu povprečne aktivnosti nevronov striatuma spremljamo vrednost trenutnega stanja oziroma pričakovano nagrado. Zvišana aktivnost striatuma brez zamika predstavlja znižano aktivnost ventralnega palliduma. Vsota aktivnosti ventralnega palliduma in aktivnosti striatuma zamaknjene za čas enega stanja (200 ms) v preteklost pa predstavlja aktivnost dopaminergičnih nevronov in dovedeno nagrado v tem stanju. Ker uporabljamo zakasnjeno sinapso R-STDP, bodo ob zvišani dopaminergični aktivnosti okrepljene povezave iz vhodnih nevronov prejšnjega stanja. V intervalu med 200 ms in 400 ms, kjer dovedemo zunanjo nagrado ob prihodu v stanje 1, se bodo okrepile povezave iz vhodnega nevrona 0 do izhodnih nevronov in striatuma. Povezave iz posameznega vhodnega nevrona do striatuma se okrepijo enakomerno, medtem ko pri krepljenju povezav do izhodnih nevronov poteka tekmovanje med povezavami in učenje politike preko R-STDP učenja, kot opisano v poglavju 4.1.

4.2.4 Rezultati

Naučeno politiko bomo prikazali podobno kot v poglavju 4.1, vendar bomo namesto maksimalne razlike med utežmi povezav do različnih akcij uporabili kar povprečno utež povezave od vhodnega nevrona stanja do striatuma. Stanja z največjimi utežmi do striatuma oziroma stanja z najvišjo pričakovano nagrado, so bila tekom učenja deležna največ nagrajevanja, zato pričakujemo, da so v teh stanjih akcije najbolj diferecirane.

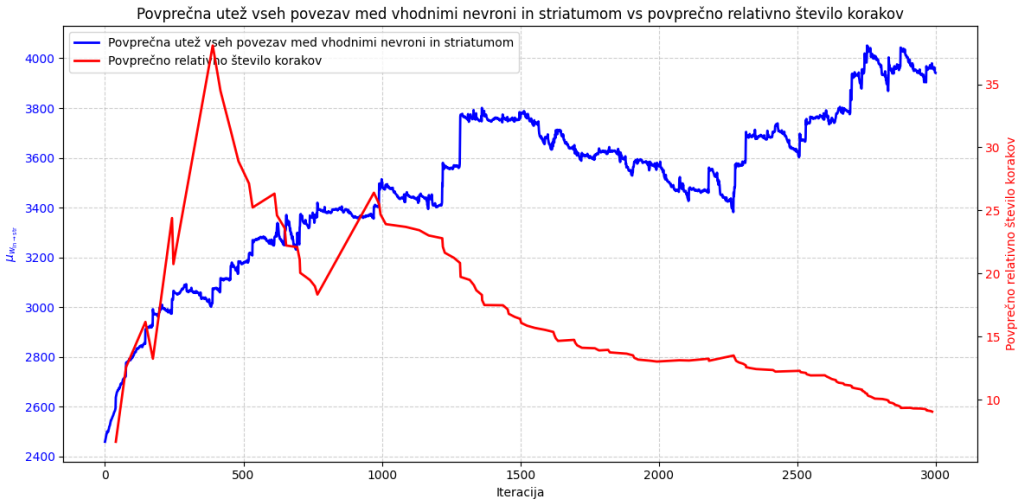
Rezultat učenja na 4x4 mreži po 3000 iteracijah je prikazan na sliki 4.11. Izbira akcije v posameznem polju je razvidna iz smeri puščice, kjer vidimo, da se je agent naučil skoraj optimalne navigacije do cilja iz poljubnega stanja. Pričakovano imajo stanja neposredno ob nagrajenem stanju najvišjo pričakovano vrednost, dlje kot se oddaljimo od tega stanja, nižja je pričakovana nagrada. Stanja, ki so najbolj oddaljena 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 4.11: Prikaz politike modela na 4x4 mreži po 3000 iteracijah po 200 ms. Končno stanje je obarvano z zeleno. Barva ostalih stanj prikazuje povprečno utež povezav med vhodnim nevronom, ki predstavlja to stanje in striatumom, oziroma pričakovano nagrado. V vsakem polju je s pučico prikazan vektor preferirane akcije (smeri) glede na uteži med vhodnim nevronom in posameznimi izhodnimi nevroni. Vidimo, da pričakovana nagrada od končnega stanja, propagira v bolj oddaljena stanja. Smeri vektorjev stanj kažejo v smeri najboljše akcije, kar kaže na uspešno učenje, z izjemo najbolj oddaljenih stanj, kjer se sorazmerno s pričakovano nagrado agent ni naučili optimalne politike.

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 tekom 3000 iteracij je prikazano na grafu 4.12, kjer vidimo, da povprečno število korakov, ki jih agent potrebuje, da pride do cilja, pada sorazmerno z rastjo povprečne pričakovane nagrade.



Slika 4.12: Povprečna utež sinaps med vsemi vhodnimi nevroni in striatumom tekom 3000 iteracij po 200 ms ter povprečno relativno število korakov. Na sliki vidimo, da medtem ko povprečna utež sinaps med vsemi vhodnimi nevroni in striatumom, ki predstavlja globalno pričakovano nagrado glede na naučeno politiko, raste, povprečno relativno število korakov pada, kar je odraz uspešnega učenja politike, ki se približuje optimalni.

Sistem, kot smo ga implementirali izvaja t. i. "on-policy" TD-učenje, saj vrednost posameznih stanj posodablja glede na vrednost naslednjega stanja, kjer naslednjo akcijo izberemo glede na naučeno politiko, za razliko od t. i. "off-policy" algoritmov, kjer vrednosti posodabljamo ob predpostavki, da bomo v naslednjem stanju vedno izbrali najboljšo akcijo.

Poglavje 5

Zaključek

Namen te diplomske naloge je bil predstaviti in rešiti nekatere izzive pri učenju impulznih nevronskih mrež. 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 želenih 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 ganglijih 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.

5.1 Ideje za nadaljnje delo

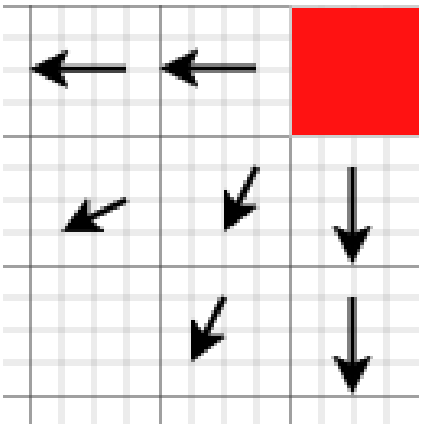
5.1.1 Izognitveno obnašanje (*angl. aversive behaviour*)

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 zunanjosti 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 5.1: Pričakovana politika ob kritiku negativnih stanj (brez kritika nagrajenih stanj)

5.1.2 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”.

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/bhl152.

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". Diplomska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Svete A (2020). "Posplošitev problema vozička s palico na zahtevnejše domene". Diplomska 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.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.