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, 2025

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:

Besedilo teme diplomskega dela študent prepiše iz študijskega informacijskega sistema, kamor ga je vnesel mentor. V nekaj stavkih bo opisal, kaj pričakuje od kandidatovega diplomskega dela. Kaj so cilji, kakšne metode naj uporabi, morda bo zapisal tudi ključno literaturo.

Title: Reinforcement learning on spiking neural networks

Description:

opis diplome v angleščini

Kazalo

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 spodbujevalno učenje na impulznih nevronskih mrežah, tipu nevronskih mrež, ki se obnašajo podobno kot človeški možgani. Ker tu ne moremo uporabiti klasičnih algoritmov spodbujevalnega učenja, postopoma razvijemo kompleksen sistem, navdihnjen z dopaminergičnimi mehanizmi v človeških možganih, ki omogoča učenje na podlagi časovne razlike. Pri tem postopoma rešujemo izzive pri učenju impulznih nevronskih mrež in na koncu uspešno rešimo problem tako s takojšnjo kot z oddaljeno nagrado.

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 explore reinforcement learning in spiking neural networks, a type of neural network that resembles the human brain. As classic reinforcement learning algorithms cannot be directly applied to spiking neural networks, we gradually develop a complex system inspired by dopaminergic mechanisms in the human brain, which implements a form of temporal difference learning. During the development of the final system, we propose solutions to various problems associated with reinforcement learning in spiking neural networks and ultimately solve a problem involving both immediate and delayed rewards.

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

Poglavje 1

Uvod

1.1 Motivacija

Impulzne nevronske mreže so v večini implementacij poskus modeliranja bioloških značilnosti nevronov in sinaps v možganih. Kot izjemno močan računski stroj, so možgani navdih za številne sodobne koncepte v umetni inteligenci. Najbolj očiten primer so nevronske mreže, vendar se po mehanizmih, prisotnih v možganih, lahko zgledujemo tudi pri metodah spodbujevanega učenja.

Delovanje možganov je kljub številnim raziskavam še vedno precej slabo razumljeno, njihovo računalniško modeliranje pa je v času pisanja še razmeroma mlado področje. Odkrivanje mehanizmov in vzorcev, ki se pojavijo med delovanjem in učenjem impulznih nevronskih mrež, ter uporaba teh pri modeliranju mehanizmov, za katere vemo, da so prisotni v možganih, predstavlja velik doprinos tako k področju računske nevroznanosti kot tudi psihoanalizi in sorodnim področjem. V neposredni povezavi s psihoanalizo raziskovanje impulznih nevronskih mrež predstavlja raziskovanje temeljnih vprašanj o človeškem dojemanju in delovanju možganov nasploh.

1.2 Cilji

V tej diplomski nalogi razvijemo kompleksnega agenta, ki je zmožen reševati probleme tako s takojšnjimi kot zakasnjenimi nagradami in temelji izključno na impulznih nevronskih mrežah in spodbujevanem učenju. V poglavju 3 so predstavljeni in ovrednoteni različni modeli bioloških značilnosti nevronov in sinaps. Nato se posvetimo spodbujevanemu učenju, kjer v poglavju 4.1 razvijemo agenta, ki uporablja model sinaptične plastičnosti, odvisne od nagrajevanja in časovne razporeditve impulzov (angl. R-STDP, primer Izhikevich, E. M. 2007), in se je sposoben naučiti igranja igre Pong. Tak agent ni sposoben učenja nalog z zakasnjenimi nagradami, zato v poglavju 4.2 uporabimo TD učenje. Za ta namen modeliramo nevronska vezja in določene mehanizme iz človeškega dopaminskega sistema. S pomočjo TD modela akter-kritik se naučimo poti do zakasnjene nagrade pri nalogi, kjer se premikamo po mreži.

Poglavje 2

Pregled področja in sorodnih del

Na temo impulznih nevronskih mrež v Sloveniji do časa pisanja še ni bila napisana nobena diplomska ali magistrska naloga, doktorska disertacija ali znanstveni članek, kar dodatno motivira pisanje diplomske naloge na to temo. Impulzne nevronske mreže zaradi zahtevnosti učenja (v času pisanja) niso pogosto uporabljene, vse bolj uporabna metoda v umetni inteligenci pa je spodbujevano učenje, ki je tudi prevladujoča in biološko podprta metoda za učenje impulznih nevronskih mrež.

Na temo spobujevanega učenja je na voljo več slovenskih znanstvenih del. Pri spodbujevanem učenju predstavimo svoj sistem kot agenta, ki izvaja aktivnosti nad okoljem, to pa mu kot odziv vrača nagrado in novo stanje okolja. Med drugim je uporabno pri problemih, kot so navigacija in reševanje problemov z roboti, na kar se nanaša članek pod avtorstvom prof. dr. Danijela Skočaja, rednega profesorja na FRI, in dr. Mateja Dobrevskega (Dobrevski M, Skočaj D 2021). Objavljenih je tudi več diplomskih in magistrskih nalog o simuliranih problemih, kot so uporaba spodbujevanega učenja za simulacijo psa ovčarja T 2022, reševanje problemov sorodnih problemu vozička s palico (Svete A 2020), igranje iger (Šutar M 2023) in uporaba TD (angl. Temporal Difference) učenja v Monte Carlo preiskovanju dreves (Deleva A 2015). V

vseh navedenih primerih se zgledujemo po raznolikem procesiranju podatkov iz zunanjega okolja, kjer je v robotiki in podatkih iz resničnega sveta prisoten tudi šum, ki je tako potrebna kot tudi težavna komponenta pri učenju impulznih nevronskih mrež.

Pri spodbujevanem učenju, zlasti v resničnem svetu, imajo impulzne nevronske mreže lahko določene prednosti. Impulzne nevronske mreže namreč naravno upoštevajo časovno komponento in procesirajo sekvenčne podatkovne tokove. Ker so dogodki v teh mrežah v osnovi le propagiranje impulzov sosednjim nevronom v naslednjem časovnem intervalu, je računanje lahko učinkovito in preprosto. Zaradi tega se pojavljajo tudi trdo-ožičene implementacije impulznih nevronskih mrež. V delu Wunderlich T, et al. 2019 je raziskana uporaba TD učenja na trdo-ožičeni impulzni nevronski mreži, kjer je končna naloga igranje igre Pong. Tudi v tej diplomski nalogi bo končna naloga enaka, vendar bodo uporabljeni računalniški in ne trdo-ožičeni modeli ter naprednejši učni algoritmi osnovani na spodbujevanem učenju.

V tej diplomski nalogi je poudarek na simulacijah in snovanju algoritmov za spodbujevano učenje na impulznih nevronskih mrežah ter modeliranju različnih bioloških procesov in možganskih nevronskih vezij. Dober zgled je delo Izhikevich, E. M. 2007, ki poleg modela nevronov in sinaps vpeljuje še način pripisovanja odgovornosti sinapsam za določeno aktivnost nevronske mreže. V postopku nadgradnje algoritmov učenje poteka tudi na osnovi TD učenja in njegove biološko bolj neposredne implementacije akter-kritik (Wiebke P, et al. 2011). V tem postopku implementiramo nevronsko vezje odgovorno za nagrajevanje, kot je bilo to raziskano v človeških možganih.

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.

3.1 Nevronski model

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. 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 impulzu velikost postsinaptičnega toka sledi krivulji, ki jo določa izbrano jedro modela. Primerjali bomo leaky IAF model z eksponentnim in alfa jedrom. Njuna postsinaptična tokova sta prikazana na sliki 3.1. Obe krivulji določajo isti parametri, ki so predstavljeni v nadaljevanju.

Membranski potencial 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 — **mirovalni membranski potencial**
  Električni potencial, na katerega se membranski potencial relaksira skozi 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*). 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} — **refraktorno 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.

Formalno je membranski potencial V_m pri tokovno gnanih modelih uhajajočega integrirajočega nevrona opisan z enačbo

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

.

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. Skupni tok I_{syn}, ki ga postsinaptični nevron prejme prek vseh sinaps, je zato sestavljen iz vzbujajoče in inhibitorne komponente.

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

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.

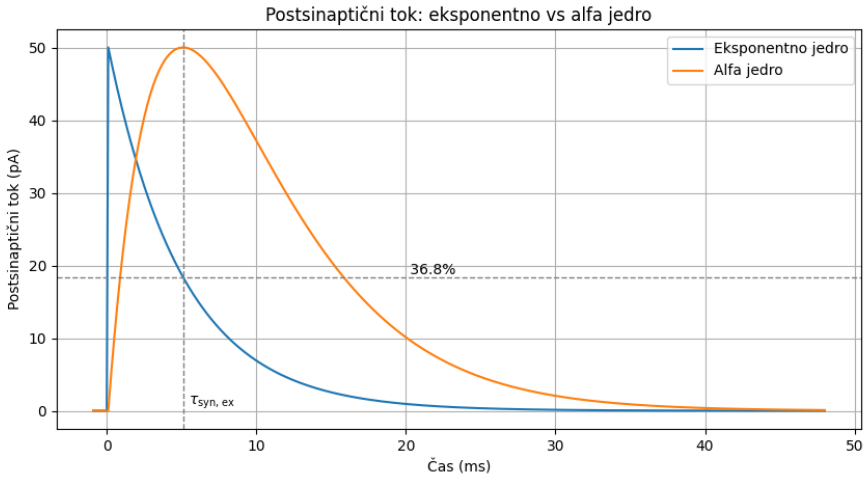
- $\tau_{\text{syn,ex}}$ — **sinaptična časovna konstanta (vzbujajoča)**
  Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju. Pri modelu z alfa-jedrom predstavlja čas dviga alfa-funkcije; pri eksponentnem jedru pa čas padca eksponentne funkcije, pri kateri je čas dviga neskončno majhen.

- $\tau_{\text{syn,in}}$ — **sinaptična časovna konstanta (inhibitorna)**
  Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju, vendar za inhibitorne sinapse.

Na sliki 3.1 je poleg $\tau_{\text{syn, ex}}$ označenih tudi $36,8\%$ maksimalnega postsinaptičnega toka. Če eksponentno jedro predstavlja preprosto eksponentno funkcijo, bo postsinaptični tok to vrednost dosegel natanko pri $\tau_{\text{syn, ex}}$, kar smo izračunali po naslednji enačbi.

$$i_{\text{syn, ex}}(t) = we^{-\frac{t}{\tau_{\text{syn, ex}}}} \tag{3.2}$$

$$i_{\text{syn, ex}}(\tau_{\text{syn, ex}}) = we^{-1} \approx 0.3679w \tag{3.3}$$



Slika 3.1: Postsinaptični tok modela z alfa in eksponentim jedrom, pri sinaptični uteži $w = 50$ in $\tau_{\text{syn, ex}} = 5$ ms.

3.1.1 Model z eksponentnim jedrom

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

$$\frac{dx}{dt} = \frac{z}{\tau_{rec}} - ux\delta(t - t_{sp}) \tag{3.4}$$

$$\frac{di}{dt} = -\frac{i}{\tau_{\text{syn, X}}} + ux\delta(t - t_{sp}) \tag{3.5}$$

$$\frac{dz}{dt} = \frac{i}{\tau_{\text{syn, X}}} - \frac{z}{\tau_{rec}}, \tag{3.6}$$

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, je $\delta(t - t_{sp}) = 0$ in se diferencialna enačba za i poenostavi v

$$\frac{di}{dt} = -\frac{i}{\tau_{\text{syn, X}}}. \tag{3.7}$$

Rešitev te diferencialne enačbe je tako

$$i(t) = i_0 e^{-t/\tau_{\text{syn, X}}}, \tag{3.8}$$

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 3.1.3). Skupni naboj q izračunamo kot

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

3.1.2 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

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

kjer je $\Theta(x)$ enotina stopnica. Postsinaptični tokovi so ob času $\tau_{\text{syn, X}}$ normalizirani na enotski maksimum.

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

Skupni naboj q, ki ga prenese postsinaptični tok pri alfa jedru, izračunamo po naslednji enačbi.

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

3.1.3 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, kot bomo videli v nadaljevanju.

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 preko petih postsinaptičnih nevronov, v katere neodvisno injiciramo šum. Biološko najbolj realističen je Poissonov šum, saj neposredno predstavlja impulze nevronov.

$$P(k \text{ impulzov v } \Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \quad k = 0, 1, 2... \tag{3.9}$$

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 3.1.3. Iz rezultatov simulacije, prikazanih v tabeli 3.1.3, 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)
Exponentno	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 exponentnim jedrom. Povprečje in standardni odklon sta izračunana na vseh postsinaptičnih nevronih.

3.2 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,
- τ_+ in τ_- časovne konstante, ki določajo okno vpliva časovnih razlik.

3.2.1 Dopaminska modulacija

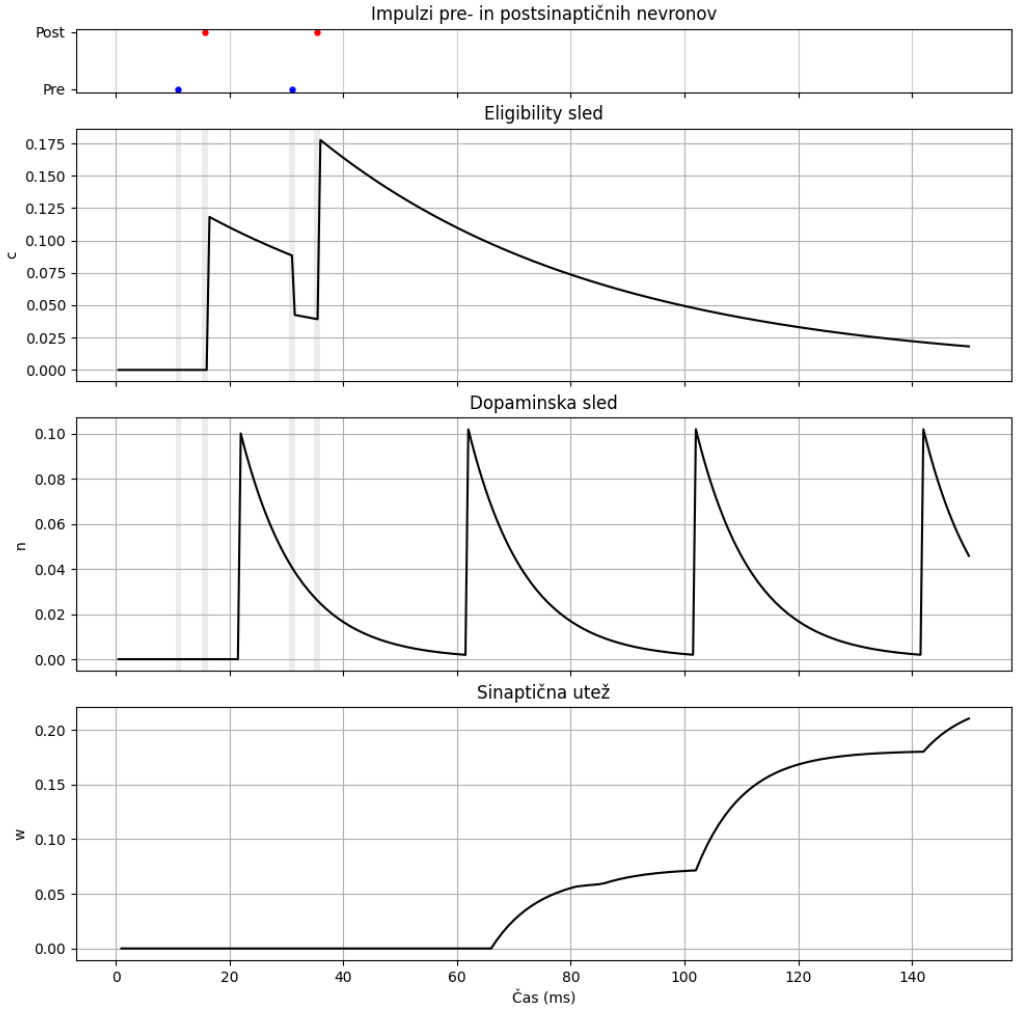
V vseh sistemih, razvitih v tej diplomski nalogi, kot osnovo uporabljamo model R-STDP sinapse, kot ga definira Izhikevich, E. M. 2007. V tem modelu odvisnost od nagrade vpeljemo z nevromodulatorjem dopaminom. Pogosto se koncentracija dopamina obravnava kot neposreden odraz količine nagrade oziroma vrednosti 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:

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

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,
- $s_{\mathrm{pre/post}}$ — čas pre- ali post-sinaptičnega impulza,
- s_n — čas impulzov dopaminskih nevronov,
- C_1, C_2 — konstante,
- τ_c, τ_n — časovne konstante odtekanja *eligibility* in dopaminskih sledi.

Slika 3.2 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 3.2: *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

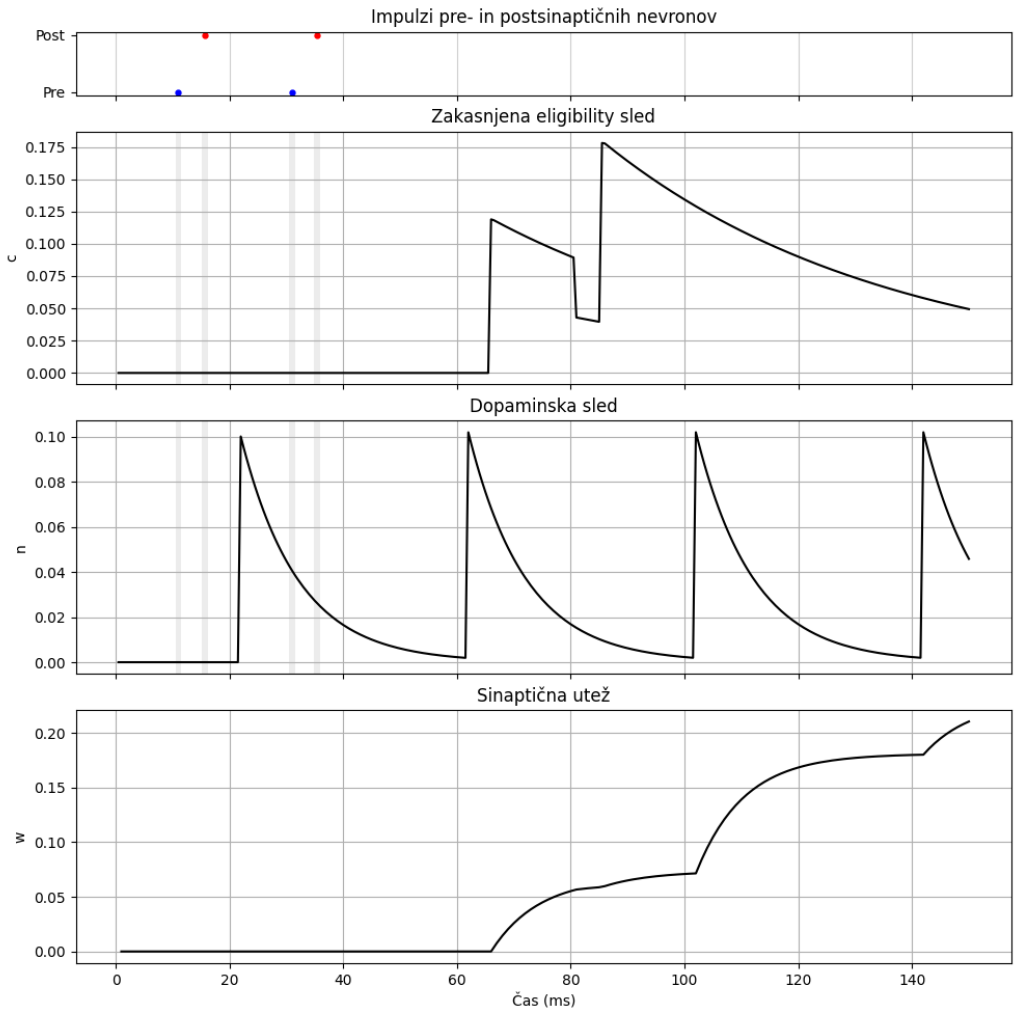
Imamo klasičnega agenta spodbujevanega učenja, ki prejema informacije o zunanjem okolju prek stimulacije vhodnih nevronov, nato pa kot odziv na trenutno stanje izbere akcijo, ki vpliva na okolje. Če se znajde v nagrajenem stanju, agenta nagradimo. S pomočjo nagrajevanja in interakcije z okoljem se agent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



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.

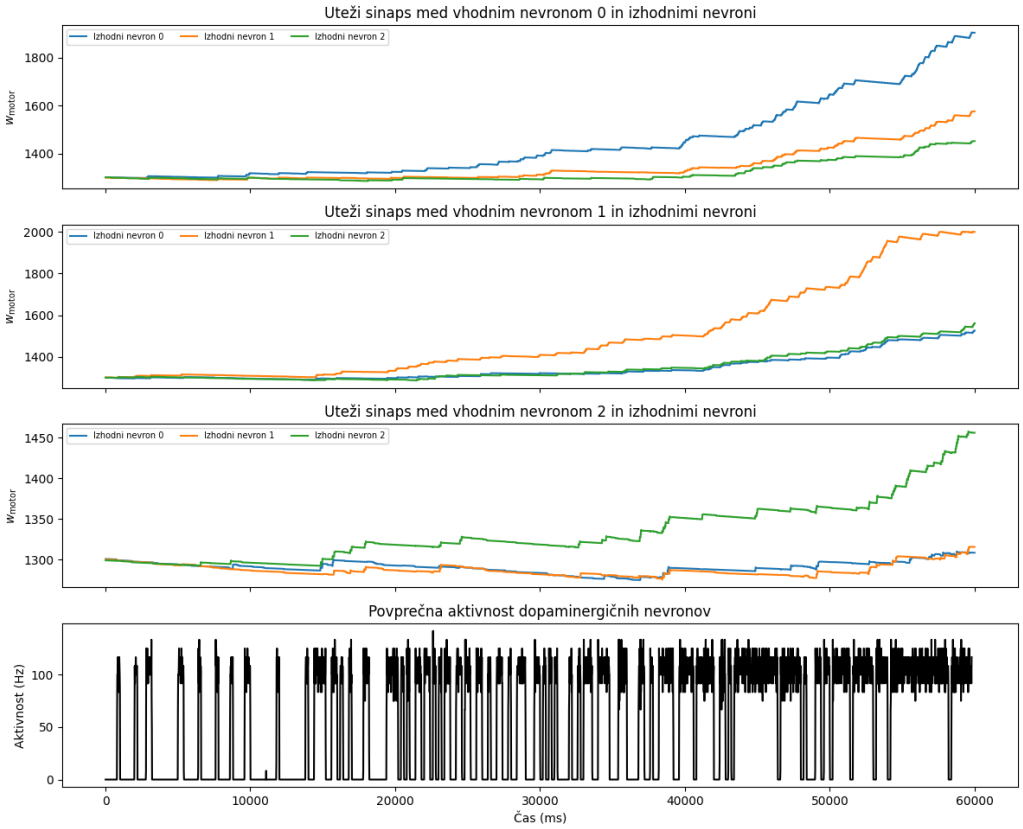
Nagrada, ki jo predstavlja aktivnost dopaminskih nevronov, bo vedno večja ali enaka 0, kar pomeni, da se bodo sinapse skozi čas le krepile. Povezave, odgovorne za izbiro določene akcije v določenem stanju, morajo zato med seboj tekmovati za prevlado. Pri tem moramo omogočiti dovolj veliko

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 3.1.3. 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 nakjučen, 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 pravilnih 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_{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.



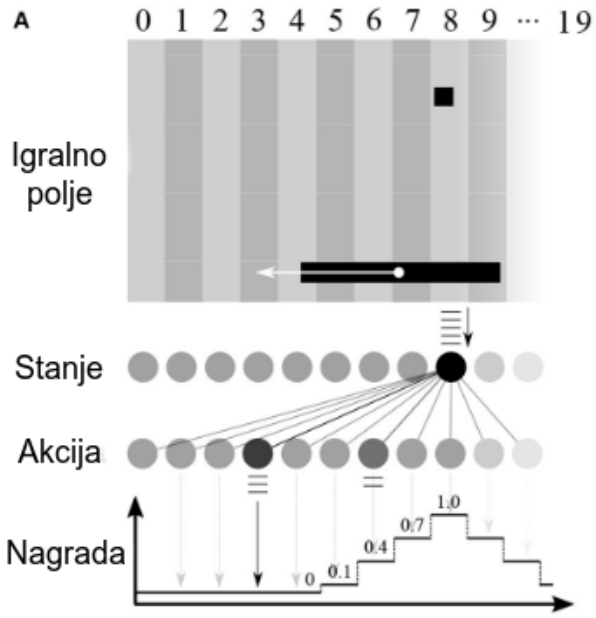
Slika 4.2: Prikaz uteži sinaps med vhodnimi in izhodnimi nevroni ter povprečne aktivnosti dopaminergičnih nevronov med simulacijo R-STDP učenja na preprostem problemu.

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 želeno 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 želene 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.

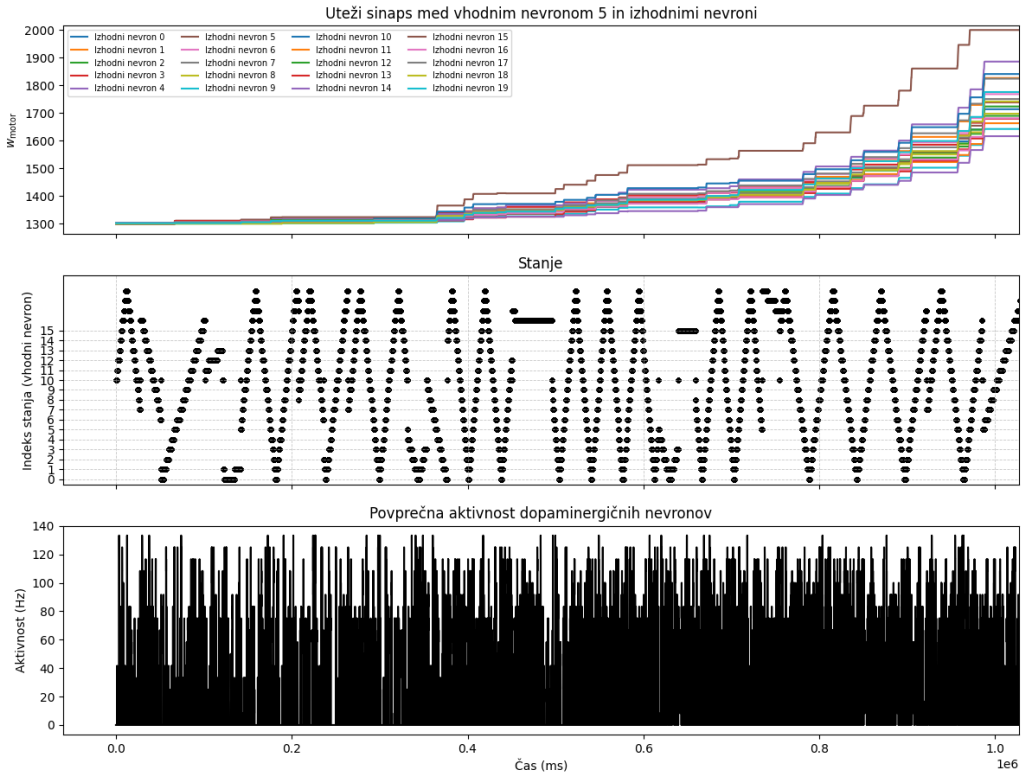
$$R_b = \begin{cases} 1 - |j - k| \cdot 0.3 & \text{if } |j - k| \leq 3, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

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



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

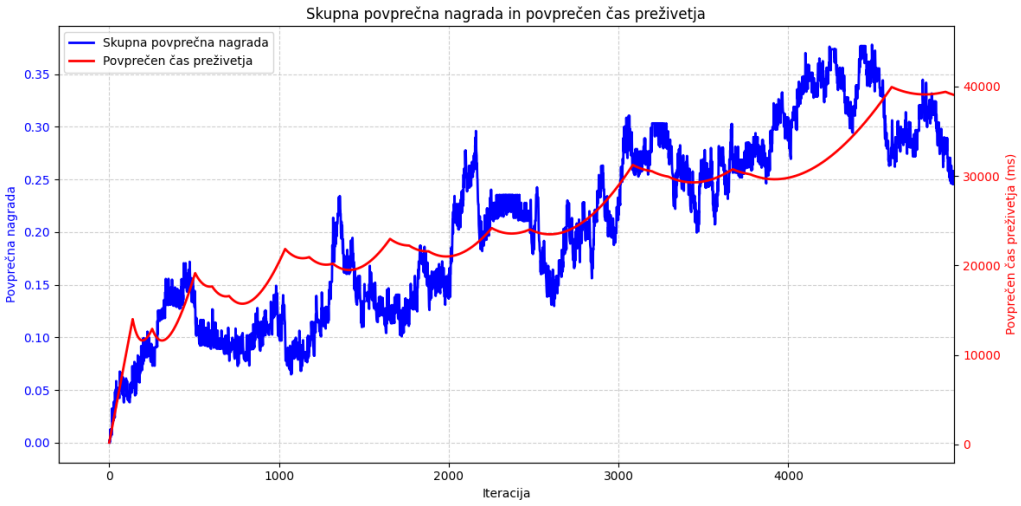
Pričakujemo, da bodo v posameznih stanjih prevladale sinapse, ki iz vhodnega nevrona vodijo do akcij okrog izhodnega nevrona, ki predstavlja isto y pozicijo, kot jo ima takrat žogica. Polje smo po y osi diskretizirali na 20 stanj. Po simulaciji, ki je trajala 6000 ms, na sliki 4.1.1 vidimo graf povezav med vhodnim nevronom, 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 implementaiji 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-

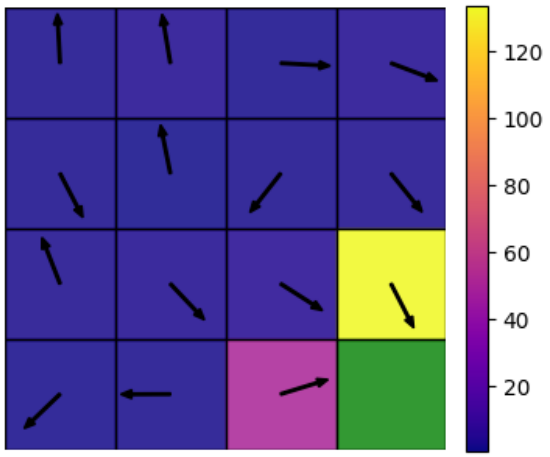
stavili v naključno stanje. V vsakem stanju i bomo trenutno politiko agenta prikazali s puščicami, katerih smer določa normaliziran vektor $\hat{x}_i$, ki predstavlja preferenco akcije glede na razlike v utežeh sinaps.

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

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

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

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 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.

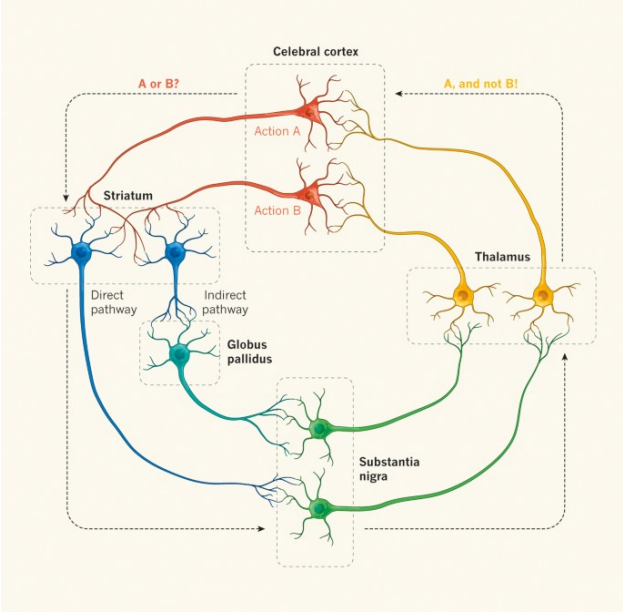
4.2.1 Model akter-kritik

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 ganglijih.

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 ganglijih, 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 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.

Ob prisotnosti osnovne, od 0 različne frekvence nevronov *ventralnega pallidума* 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 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 4.9. 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.

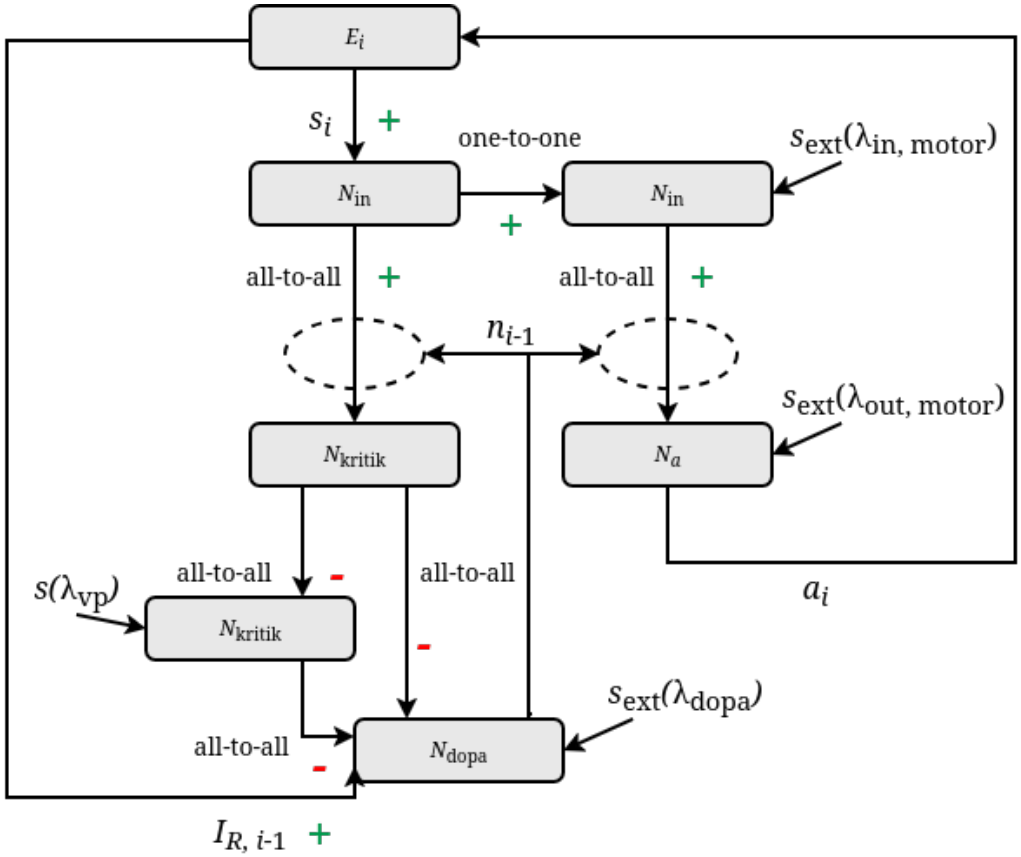
Skupino vhodnih nevronov oziroma korteks bo predstavljalo N_{in} nevronov, ki bodo povezani z N_a motoričnimi oziroma izhodnimi nevroni. V akterju bomo zaradi razlogov, navedenih v poglavju 4.1, uporabljali model nevrona z eksponentnim jedrom, v kritiku pa biološko bolj realistične nevrone z alfa jedrom. 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 $w_{\text{in}\to\text{in, motor}}$ 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



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 preko zakasnjenih R-STDP sinaps z normalno porazdeljenimi utežmi. Prav tako so vhodni nevroni po režimu *all-to-all* povezani z N_{kritik} nevroni striatuma preko zakasnjenih R-STDP sinaps z normalno porazdeljenimi utežmi. Striatum je preko statičnih inhibitornih povezav z utežmi $w_{\text{str}\to\text{vp}}$ povezan z N_{kritik} nevroni ventralnega palliduma in preko 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 preko 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 $\lambda_{\text{out, motor}}$. Prav tako Poissonov šum injiciramo v nevrone ventralnega palliduma s povprečno hitrostjo λ_{vp} in v dopaminergične nevrone s povprečno hitrostjo λ_{dopa}. Generatorji Poissonovega šuma so do posameznih skupin ne-

vronov povezani preko 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 **??**. Opisani model je predstaljven na sliki 4.8.



Slika 4.8: Prikaz implementiranega aktor-kritik sistema

Za razliko od izvornega sistema, ki ga predstavijo **??**, bomo za model sinapse uporabili našo zakasnjeno R-STDP sinapso kot smo jo razvili v poglavjih 3.2 in 4.1, ki poleg presinaptičnih impulzov upošteva tudi postsinaptične, po pravilu STDP. Tako bomo lahko kot akter uporabili R-STDP sistem, kot smo ga razvili v poglavju 4.1. Poleg tega se bo naša implementacija razlikovala tudi v načinu izbire stanja, kjer za izbrano akcijo vzamemo tisto, katere pripadajoč 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. Za ta načim moramo po prvem impulzu v karseda majhnem časovnem intervalu inhibirati vse ostale izhodne nevrone. Tako bomo sicer bolj direktno okrepili povezave odgovorne za izbrano aktivnost, saj smo z inhibicijo izničili *eligibility* sled sinaps do ostalih izhodnih nevronov, poleg tega pa tudi ne bo potrebe po tekmovanju sinaps, kot pri naši metodi, vendar moramo za to 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 100krat poženemo ukaz `nest.Simulate(0.1)` ali enkrat `nest.Simulate(10)`. Naš sistem bo za voljo hitrosti simulacije po številu nevronov manjši od izvornega sistema.

4.2.2 Izbira parametrov

Parametri so bili izbrani eksperimentalno brez oziranja na biološko točnost. Ob spreminjanju velikosti posameznih skupin nevronov moramo pri izbiri parametrov paziti na ohranjanje osnovne frekvence dopaminergičnih nevronov in da sta inhibicija in vzbujanje zaradi direktne in indirektne povezave v ravnovesju. Zmanjšanje frekvence, ki jo opravlja plast nevronov med vhodnimi in izhodnimi, mora biti dovolj velika, da bo pri osnovnih utežeh sinaps med srednjo plastjo in izhodnimi nevroni šum omogočil učenje, kot je to 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 tabeli imajo privzete vrednosti NEST simulatorja.

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 NEST sinapsa
$w_{\text{in}\to\text{in, motor}}$	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
$W_{\text{min},str}$	Minimalna utež	150
$W_{\text{max},str}$	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 NEST sinapsa
$w_{\text{str}\to\text{vp}}$	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

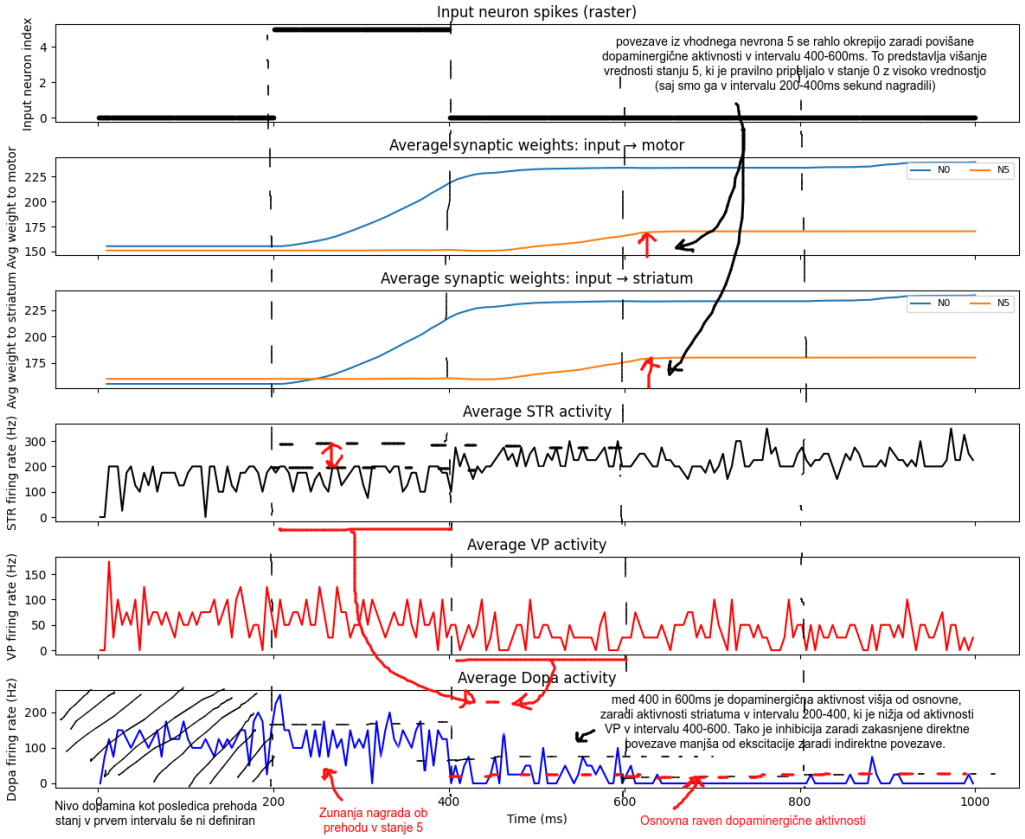
Simbol	Pomen	Vrednost
λ_{vp}	Povprečna hitrost šumnih impulzov nevronov ventral palliduma	5200
λ_{dopa}	Povprečna hitrost šumnih impulzov dopaminergičnih nevronov	4000
$\lambda_{\text{in, motor}}$	Povprečna hitrost šumnih impulzov vhodnih motoričnih nevronov	100 Hz
$\lambda_{\text{out, motor}}$	Povprečna 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

Stanja bomo razdelili v intervale dolžine 200 ms, ob prehodu stranj pa bomo vhodne nevrone stimulirali enako kot smo to počeli v poglavju 4.1. Pri

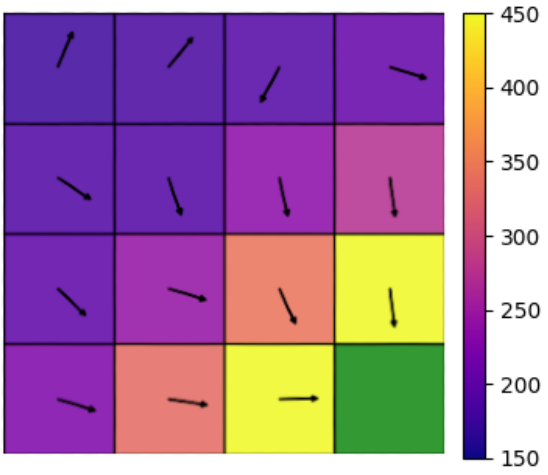
prehodu med stanji se ob dovolj visoki koncentraciji dopamina lahko okrepijo tudi sinapse do izhodnih nevronov, ki so povezane z vhodnimi nevroni prejšnjega stanja, saj *eligibility* sledi lahko ostanejo večje od 0 tudi preko več prehodov stanj. To v osnovi ni napačno in je posledica uporabe RSTDP sinapse, vendar bomo za bolj učinkovito 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 važno v katerem stanju smo se nahajali prej. Mehanizme, ki smo jih opisali v prejšnjem poglavju za začetek preverimo na sekvenci prehodov iz stanja 0, v neko 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 posledica višje pričakovane nagrade. Zunanja nagrada, ki smo jo dovedli ob prehodu v nagrajeno stanje 5 ne pomeni tudi, da v stanju 5 prčakujemo visoko nagrado. Pričakovana nagrada se namreč zviša samo 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 tako preidemo iz stanja z osnovnimi utežmi do striatuma v stanje 0, ki pa ima sedaj okrepljene utežm do striatuma. To predstavlja prehod v stanje z višjo vrednostjo. Posledica tega je, napram osnovni frekvenci 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 prehov med stanji 0 in 5.

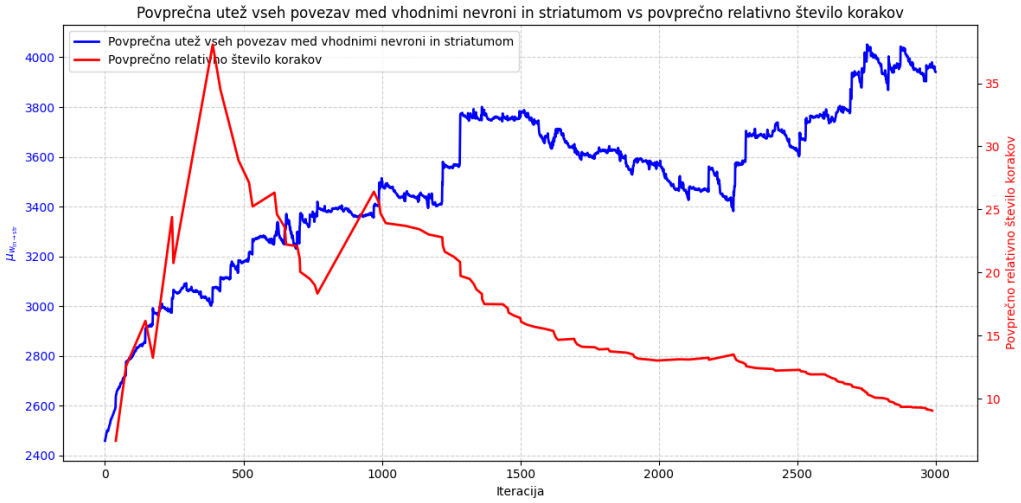
4.2.4 Rezultati

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 lahko vidimo na sliki **??**. Izbira akcije je v posameznem polju 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 pa se oddaljimo od stanja, nižja je pričakovana nagrada stanj. Stanje 0, ki je najdlje od cilja, ima pri trenutno izbiri parametrov minimalno pričakovano nagrado. Propagiranje pričakovane nagrade od končnega stanja lahko pospešimo s tem, da povišamo amplitudo posodobitve povezav do striatuma, vendar bomo s tem zvišali tudi rast motoričnih sinaps. Te bodo zato rasle prehitro in zato tekmovanje sinaps, kot je opisano v poglavju 4.1 ne bo tako učinkovito. Z drugimi besedami, bomo stanja tako nagrajevali prehitro.



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

Podobno kot pri R-STDP učenju bomo učenje spremljali preko povprečne nagrade, vendar tokrat k nagradi ne bo prispevala samo zunanja nagrada temveč tudi pričakovane nagrade. Pričakujemo, da bodo tekom učenja pričakovane nagrade preko vseh stanj rasle. To preverimo proti bolj neposredni evalvaciji učenja na mreži, kjer po vsaki ponastavitvi stanja (ko dosežemo cilj) štejemo korake dokler zopet ne pridemo v ciljno stanje. Ker so različna stanja v katere naključno postavimo agenta različno oddaljena od cilja, bomo število korakov delili z manhattansko razdaljo do cilja. Tako bomo dobili "relativne korake". Povprečno število relativnih korakov proti povprečni uteži vseh povezav med vhodnimi nevroni in nevroni striatuma tekom 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 rasti povprečne pričakovane nagrade.



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

Poglavje 5

Implementacija in uporabljena orodja

Rešitve so implementirane v jeziku Pythonu, kjer za simulacijo uporabljamo simulator NEST, ki ima zaledje implementirano v C++. Za simulator, ki ga uporabljamo preko Pythonovega vmesnika je v sklopu tega diplomskega dela implementiran tudi modul, ki je prav tako implementiran v C++. V sklopu te naloge uporaba impulznih nevronskih mrež zunaj simuliranega okolja, na trdo-ožičenih nevronskih čipih ali na robotih ni pokrito, 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 nevronskih mrežah.

Napiši več o pisanju modula za NEST simulator, s katerim smo v nadaljevanju implementirali zakasnjeno RSTDP sinapso.

Napiši par stavkov o komunikaciji z zaledjem NEST, ki nam je omogočala spremljanje internih parametrov sinaps.

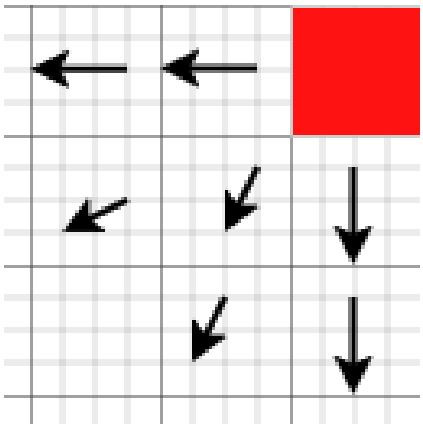
Poglavje 6

Možne razširitve

6.1 Izognitveno obnašanje (*angl. aversive behaviour*)

V večini del, ki se ukvarjajo s spodbujevanim učenjem se izognitev stanj, 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, ki so odgovorne za vstop v neželeno stanje negativno posodobljene. V človeških možganih negativnega dopamina ni. Porodi se ideja, da je izogibanje negativnim stanjem prav tako posledica učenja, kjer je nivo dopamina $n > 0$. Dopamin namreč predstavlja učenje, ne nujno nagrade. Negativo nagrado bi tako lahko predstavili s posebnim vhodom, ki prestavlja nek negativen stimulus ali "bolečino", ki jo tekom učenja želimo zmanjšati. Zopet lahko uporabimo načela R-STDP in TD učenja, kjer zmanjšanje nivoja bolečine prestavlja 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 delujta 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 dovedene negativnega stimulusa pa 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 samo 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 tako oba kritika med seboj tekmovala za nagrajevanje tako akcij, ki vodijo bližje nagradi kot tudi teh, 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 ta, da rekurenčnih povezav ni. Tako so stanja časovno med seboj skoraj popolnoma neodvisna. V kolikor dodamo več vmesnih nivojev in rekurenčne povezave bodo stanja med seboj postala časovno odvisna. Pravzaprav stanja ne mo-

remo več definirati samo z aktivnostjo vhodnih nevronov, saj v vsakem trenutku stanje vsebuje tudi informacijo iz nevronov, ki so se prožili arbitrarno v preteklost 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 200ms stimulacije, ki je do zdaj predstavljala stanje, vseeno v tem intervalu prevladala nevronska aktivnost, ki je neposredno posledica stimulacije vhodnih nevronov.

V eksperimentih izvedenih do sedaj, pravilna akcija določenega stanja ni bila odvisna od akcij, ki so nas privedle v to stanje oziroma zgodovine stanj. V primeru sprehajanja po mreži bomo končno stanje nagradili neglede 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 pripeljalo do nagrade, prehod iz stanja levo pa ne, bi lahko tako končno stanje obravnavali kot dva različna stanja, 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 sistem bi bila na primer potrebna redefinicija trenutnega načina izbire akcij, saj sta lahko dva izhodna nevrona povezana med sabo in se bosta vedno prožila skupaj. Rešitev bi lahko bila dopuščanje izbire večih akcij hkrati, kjer takšno situacijo “kaznujemo”.

Poglavje 7

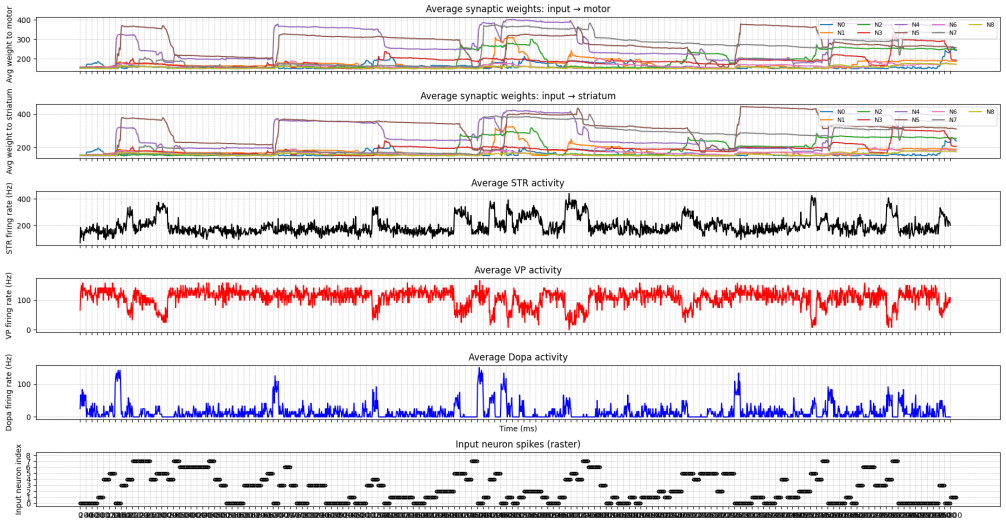
Zaključek

Namen te diplomske naloge je bil predstaviti in rešiti določene izzive pri učenju impulznih nevronskih mrež. V nalogi razvijemo inovativne rešitve, ki upoštevajo tako zahtevnost simulacije, kot tudi smiselnost iz vidika nevrologije in resničnih mehanizmov v možganih. Dodatno smo se v nalogi 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 namreč iskali samo z obzirom na to, da smo dosegli željene mehanizme, ne pa tudi če se te vrednosti skladajo 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 te v bazalnih ganglijih zdijo bolj skupne različnim živalskim vrstam kot parametri nevronov in sinaps. Center, odgovoren za sluh in modulacijo glasilk na primer, sta si po strukturi pri človeku in netopirju podobna, vendar pri netopirju te centri očitno delujejo na precej višji frekvenci. Nadaljnje iskanje hiperparametrov bi najverjetneje lahko privedlo do boljših rezultatov, kot te doseženi v tej diplomski nalogi. Pri sistemih brez rekurenčnih povezav in brez dodatnih popolno povezanih plasti nevro-

nov, kakršni so sistemi implementirani v tej nalogi, višanje števila nevronov v posamezni skupini ne bi nujno privedlo do boljših rezultatov, ni pa bilo to preverjeno v sklopu te diplomske naloge 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/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.