

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

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



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

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

**Kandidat:** Matjaž Pogačnik

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

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

**Mentor:** prof. dr. Zoran Bosnić

**Opis:**

Besedilo teme diplomskega dela študent prepíše 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

Povzetek

Abstract

<b>1</b>	<b>Uvod</b>	<b>1</b>
1.1	Motivacija . . . . .	1
1.2	Cilji . . . . .	2
<b>2</b>	<b>Pregled področja in sorodnih del</b>	<b>3</b>
<b>3</b>	<b>Modeliranje nevronov in sinaps</b>	<b>5</b>
3.1	Nevronski model . . . . .	5
3.2	STDP Sinaptični model . . . . .	11
<b>4</b>	<b>Spodbujevano učenje na impulznih nevronskih mrežah</b>	<b>17</b>
4.1	R-STDP učenje . . . . .	18
4.2	TD učenje in model actor-critic . . . . .	25
<b>5</b>	<b>Implementacija in uporabljena orodja</b>	<b>39</b>
<b>6</b>	<b>Diskusija in zaključek</b>	<b>41</b>
	<b>Viri</b>	<b>43</b>



# Seznam uporabljenih kratic

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





# Povzetek

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

**Avtor:** Matjaž Pogačnik

V tem diplomskem delu obravnavamo spodbujevalno učenje na impulznih nevronske mrežah, tipu nevronske mreže, 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 nevronske mreže 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 mehanizmi, 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 nevronske mreže, 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 nevronske mreže 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 zakasnjjenimi 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 zakasnjjenimi 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 nevronske mreže 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 spodbujevanje učenja, ki je tudi prevladujoča in biološko podprta metoda za učenje impulznih nevronske mreže.

Na temo spodbujevanja učenja je na voljo več slovenskih znanstvenih del. Pri spodbujevanju učenja 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 spodbujevanja 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 nevronske mreže.

Pri spodbujanem 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 nevronske mreže. V delu Wunderlich T, et al. 2019 je raziskana uporaba TD učenja na trdo-ožičeni impulzni nevronske 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 spodbujanem učenju.

V tej diplomski nalogi je poudarek na simulacijah in snovanju algoritmov za spodbujano učenje na impulznih nevronske mrežah ter modeliranju različnih bioloških procesov in možganskih nevronske vezij. Pri tem je dober zgled 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 dejansko nevronske vezje odgovorno za nagrajevanje, kot je bilo to raziskano v človeških možganih.



## Poglavje 3

# Modeliranje nevronov in sinaps

Impulzne nevronske mreže so definirane tako z modelom nevrona kot z modelom sinapse, ki nevrone povezujejo. Modelov je sicer veliko, v nadaljevanju pa bosta predstavljena in primerjana dva tipa "integrate and fire" modelov nevronov, glede na njuno uporabnost pri spodbujanem učenju na impulznih nevronskih mrežah. Sinapse, ki so bile uporabljene v nalogi temeljijo v celoti na modelu odvisnem od nagrade in časovne razporeditve impulzov pre in postsinaptičnih nevronov (*NEST: stdp\_dopamine\_synapse*), vendar v sistemih, ki jih razvijemo v nalogi uporabljamo prilagoditev te sinapse, ki ima sled odgovornosti (*angl. eligibility trace*) zamaknjeno v času.

### 3.1 Nevronski model

Nevronski modeli, uporabljeni v tem delu, temeljijo na tokovno gnanih modelih uhajajočega integrirajočega nevrona, pri katerih se membranski potencial spreminja v skladu s pasivnimi električnimi lastnostmi enostavne celične membrane. Dinamika membrane izhaja iz ravnovesja med kapacitivnim nabojem in uhajanjem preko membranske prevodnosti. V simulacijah s simulatorjem NEST ta obnašanja opisujejo naslednji parametri:

- $E_L$  — **mirovalni membranski potencial**

Električni potencial, proti kateremu membrana pasivno relaksira v od-

sotnosti od vhodnih tokov.

- $C_m$  — **membranska kapacitivnost**

Kapacitivnost membrane, ki določa, kako hitro se membranski potencial odziva na vhodne tokove.

- $\tau_m$  — **membranska časovna konstanta**

Čas, v katerem membrana pasivno integrira tok; definiran kot razmerje med kapacitivnostjo  $C_m$  in uhajalsko prevodnostjo  $g_L$  (*leakage conductance*), ki pa je simulator Nest ne podaja kot neodvisen parameter.  $\tau_m$  lahko definiramo tudi kot produkt med kapacitivnostjo in uporom membrane  $\tau_m = C_m R_m = \frac{C_m}{g_L}$

- $t_{ref}$  — **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.

- $V_{reset}$  — **potencial ponastavitve**

Ponastavitveni membranski potencial.

- $\tau_{syn,ex}$  — **sinaptična časovna konstanta (ekscitatorna)**

Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju. Pri modelu z alfa-jedrom (alfa oblikovan postsinaptični tok) predstavlja čas dviga alfa-funkcije; pri eksponentnem jedru pa čas padca eksponentne funkcije, pri kateri je čas dviga sicer neskončno majhen.

- $\tau_{syn,in}$  — **sinaptična časovna konstanta (inhibitorna)**

Čas, ki določa hitrost naraščanja postsinaptičnega toka po proženju, vendar za inhibitorne sinapse.

- $I_e$  — **zunanji konstantni tok**

Dodani tok, ki modelira stalni zunanji šum.

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

Absolutna spodnja meja za membranski potencial.

Membranski potencial  $V_m$  se spreminja v odvisnosti od  $I_{\text{syn}}$  in ostalih parametrov po naslednji enačbi

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

Skupni tok  $I_{\text{syn}}$ , ki ga nevron prejme preko vseh sinaps je sestavljen iz excitatorne 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 ekscitatornih ( $X = \text{ex}$ ) in inhibitornih ( $X = \text{in}$ ) sinapsah z utežmi  $w_j$  do presinaptičnih nevronov,  $k$  teče po časih impulzov nevrone  $j$ ,  $d_j$  pa predstavlja zakasnitev sinapse do nevrone  $j$ . Postsinaptični tokovi  $i_{\text{syn, X}}(t - t_j^k - d_j)$  nevrone  $j$  so odvisni od jedra, ki ga uporablja model.

### 3.1.1 Model z alfa jedrom

V simulatorju NEST je postsinaptični tok modela z alfa jedrom 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 v enotski maksimum.

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

Skupni naboj  $q$ , ki ga prenese postsinaptični tok je tako odvisen od sinaptične časovne konstante in je izračunan po naslednji enačbi

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

### 3.1.2 Model z eksponentnim jedrom

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

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

$$\frac{dy}{dt} = -\frac{y}{\tau_I} + ux\delta(t - t_{sp}) \quad (3.3)$$

$$\frac{dz}{dt} = \frac{y}{\tau_I} - \frac{z}{\tau_{rec}} \quad (3.4)$$

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

Če opazujemo samo speminjanje  $y(t)$  skozi čas brez novih impulzov, bo  $\delta(t - t_{sp}) = 0$  in se diferencialna enačba za  $y$  poenostavi v

$$\frac{dy}{dt} = -\frac{y}{\tau_I} \quad (3.5)$$

rešitev te diferencialne enačbe je tako

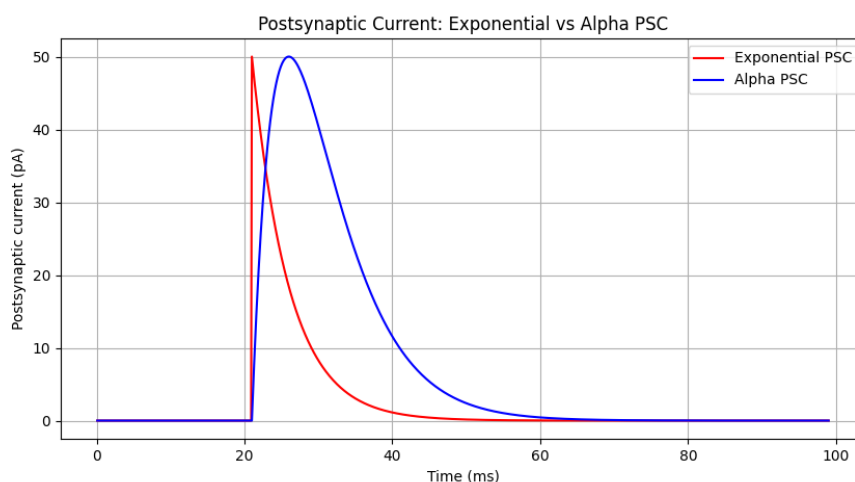
$$y(t) = y_0 e^{-t/\tau_I} \quad (3.6)$$

kjer vidimo, da je jedro res eksponentna funkcija z začetkom v  $y_0$ . Skok potenciala po impulzu je definiran z utežjo sinapse  $w$ , postsinaptični tok pa je sam po sebi definiran samo s hitrostjo padanja funkcije  $\tau_I$ , ki pa je v simulatorju NEST predstavljen s  $\tau_{syn, X}$ .

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

Skupni naboj  $q$ , ki ga prenese postsinaptični tok je tako izračunan po naslednji enačbi

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



Slika 3.1: Postsinaptični tok modela z alfa in eksponentnim jedrom

### 3.1.3 Izbira modela nevrona

V sistemih, ki jih bomo implementirali v nadaljevanju skušamo pri modeliranju mehanizmov v človeških možganih uporabiti čimmanj poenostavitvev ali posplošitev za kar 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 odvisno od nagrade bolje obnese model z eksponentnim jedrom.

Za nas sta najpomembnejši razlika v količini prenesenega naboja  $q$  in, kot je opisano v poglavju spodbujevano učenje z R-STDP, razlika v varianci frekvence impulzov zaradi zunanjega šuma in razlik v utežeh sinaps. Količina prenesenega naboja  $q_{\text{alfa}}$  je pri alfa jedru večja od prenesenega naboja pri eksponentnem jedru  $q_{\text{exp}}$  za faktor  $\frac{q_{\text{alfa}}}{q_{\text{exp}}} = e$ . To razliko zlahka prilagodimo z nižjimi vrednostmi uteži sinaps. Razlika v varianci frekvenc impulzov je posledica daljšega časovnega intervala pri alfa jedru napram eksponentnem, kjer je postsinaptični tok blizu maksimalne vrednosti napram eksponentnem. Zaradi tega bodo zaporedni postsinaptični impulzi skozi čas precej bolj pre-

krivni, pri integriranju različnih postsinaptičnih tokov sozi čas pa pride do učinka nizko prepustnega filtra. Tako so nenadne spremembe v amplitudi skupnega toka na vhodu v postsinaptični nevron ublažene. Posledica so manjše razlike v frekvenci impulzov postsinaptičnega nevrona, če imamo na vhodu sinapse različne uteži, učinek pa je še bolj opazen ob dodanem šumu. Pri alfa jedru bo namreč šum povzročil manj variance v frekvenci impulzov postsinaptičnega nevrona, kot pri eksponentnem jedru.

Primerjamo oba modela preko 5 postsinaptičnih nevronov v katere injeciramo poissonski šum. Da dosežemo karseda enako osnovno frekvenco impulzov postsinaptičnih nevronov pri alfa in eksponentnem jedru je utež sinapse za faktor  $e$  manjša od uteži sinaps do nevronov z eksponentnim jedrom.

Tabela 3.1: Parametri simulacije uporabljeni pri primerjavi modelov nevronov.

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

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

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

## 3.2 STDP Sinaptični model

V sistemih, ki bodo implementirani v tej nalogi bomo uporabljali prilagojeno sinapso s plastičnostjo odvisno od nagrade in časovne razporeditve impulzov (*angl. R-STDP synapse*). 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** post-sinaptičnim ( $\Delta t > 0$ ), se sinapsa **okrepi** (potencira). Če se pre-sinaptični nevron sproži **po** post-sinaptičnem ( $\Delta t \leq 0$ ), se sinapsa **oslabi** (depresira).

Matematično je to opisano s funkcijo okna STDP:

$$\text{STDP}(\Delta t) = \begin{cases} A_+ e^{-|\Delta t|/\tau_+}, & \text{če } \Delta t > 0 \text{ (pre-sinaptični pred post-sinaptičnim)} \\ A_- e^{-|\Delta t|/\tau_-}, & \text{če } \Delta t \leq 0 \text{ (post-sinaptični pred pre-sinaptičnim)} \end{cases}$$

kjer so:

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

### Dopaminska modulacija

Pri neuromodulirani STDP dopaminska koncentracija  $n$  modulira velikost in smer sinaptične plastičnosti tj. velikost in predznak posodobitve uteži

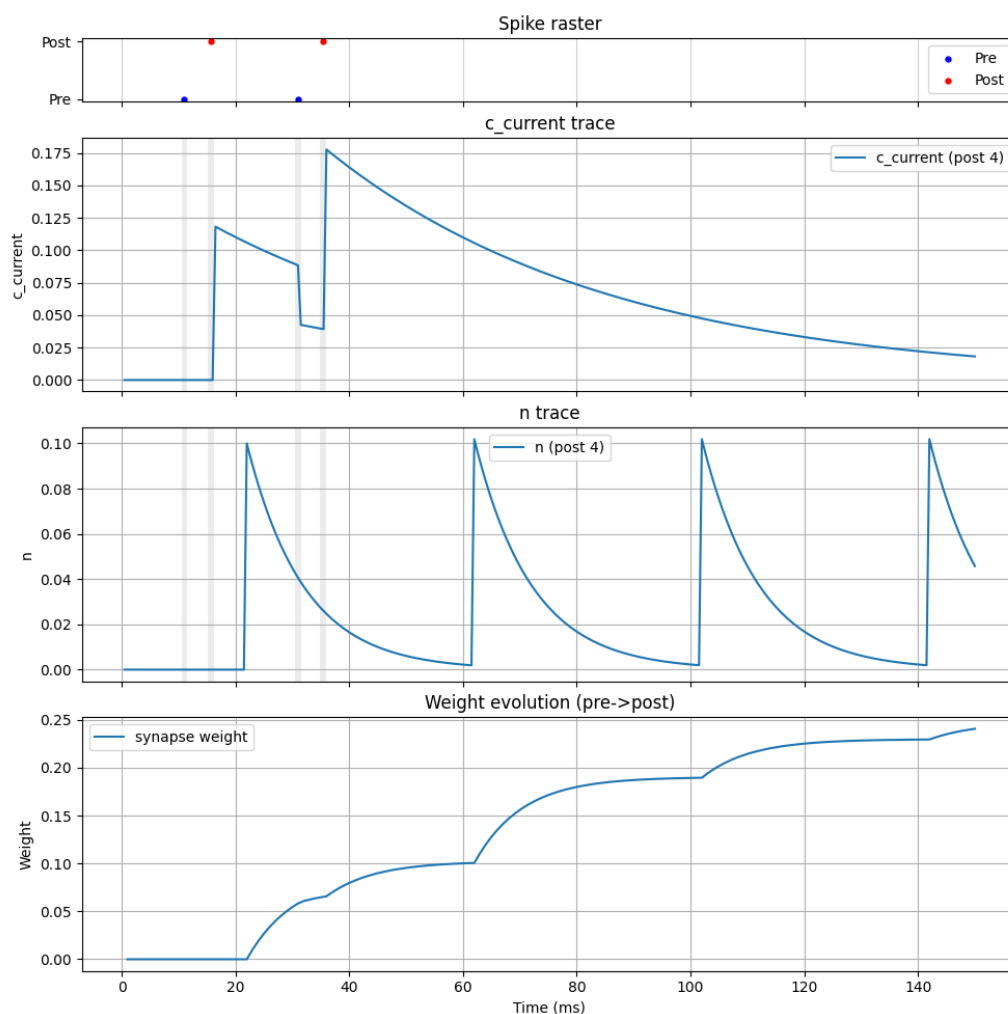
povezave. Sinaptična dinamika je opisana z naslednjimi enačbami:

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

kjer so:

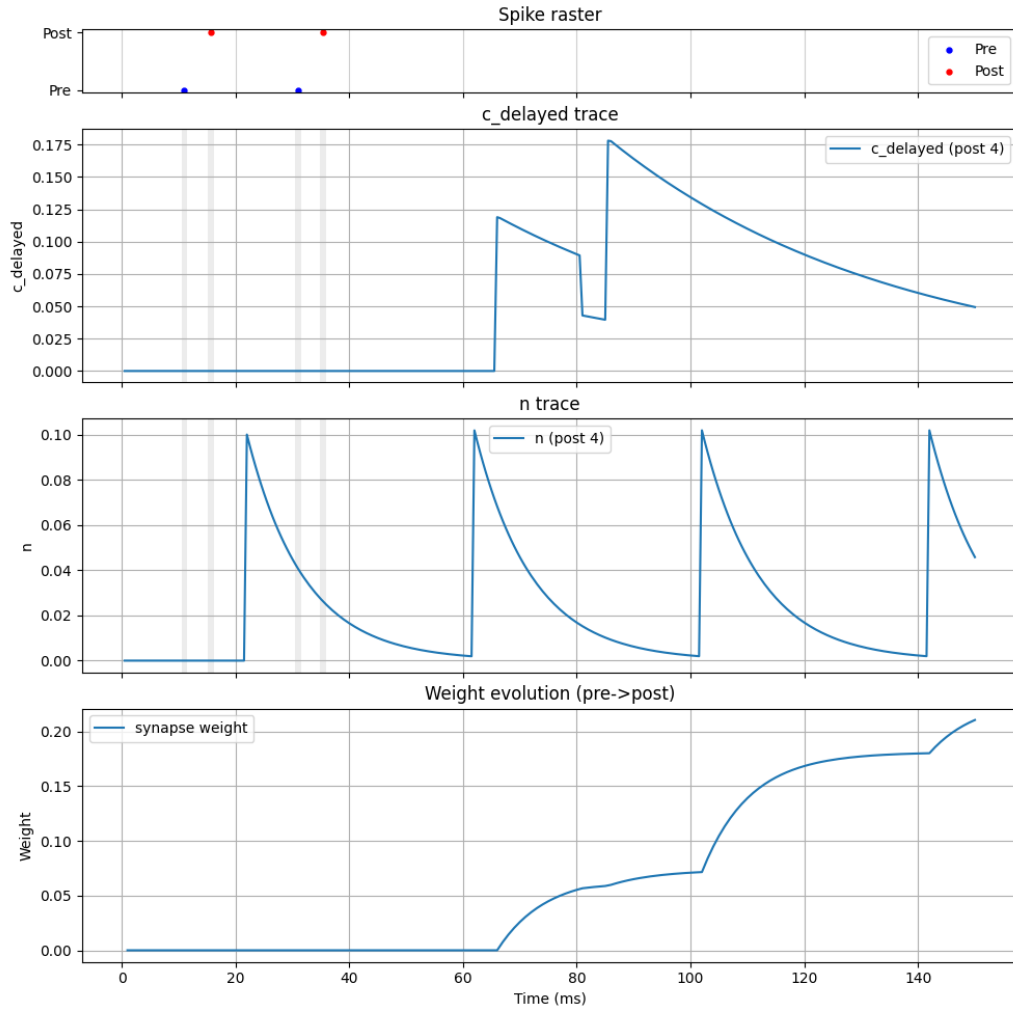
- $w$  — sinaptična utež,
- $c$  — *eligibility trace* (spremlja pare sproženih pre in postsinaptičnih nevronov),
- $n$  — dopaminska koncentracija/sled,
- $b$  — bazalna dopaminska koncentracija,
- $s_{\text{pre/post}}$  — čas pre- ali post-sinaptičnega impulza,
- $s_n$  — čas impulzov dopaminskih nevronov,
- $C_1, C_2$  — konstante,
- $\tau_c, \tau_n$  — časovne konstante odtekanja *eligibility* in dopaminskih sledi.





Slika 3.2: *eligibility* sled  $c$ , dopaminska sled  $n$  in evolucija sinaptične uteži pri presinaptičnih impulzih pri  $[10.0, 30.0]$  ms in postsinaptičnih impulzih pri  $[12.0, 32.0]$  ms, simulirane preko 150 ms pri R-STDP sinapsi z  $\tau_c = 50.0$  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.

V poglavju **R-STDP učenje** bomo R-STDP sinapso uporabili tako, da bomo ob pravilni akciji agenta pri spodbujevanem učenju povezave, ki so bile najbolj odgovorne za izbiro akcije okrepili. To bomo dosegli tako, da za vse povezave hkrati povišamo koncentracijo dopamina, pri tem pa bodo najmočnejše povezave, ki bodo povzročile največ kavzalnih parov pre in postsinaptičnih impulzov imele najvišji eligibility in bodo tako najbolj okrepljene. Agent bo ob prihodu v določeno stanje izbral naslednjo akcijo, kjer bo nagrada na voljo šele ob prihodu v naslednje stanje, v kolikor je to stanje pravilno, zato hočemo posodobiti povezave, ki so bile odgovorne za akcijo, ki nas je do tega stanja pripeljala. Koncentracijo dopamina bomo povišali za čas določenega intervala ob prihodu v nagrajeno stanje, kjer pa bi lahko potemtakem posodabljali že povezave, ki so aktivne v novem stanju. Da se temu izognemo bomo onemogočili posodabljanje sinaps zaradi nagrad, ki pridejo prehitro znotraj določenega intervala  $\tau_{c,\text{delay}}$ . Celotno *eligibility* sled bomo tako premaknili za  $\tau_{c,\text{delay}}$



Slika 3.3: *eligibility* sled  $c$ , dopaminska sled  $n$  in evolucija sinaptične uteži pri presinaptičnih impulzih pri  $[10.0, 30.0]$  ms in postsinaptičnih impulzih pri  $[12.0, 32.0]$  ms, simulirane preko 150 ms pri R-STDP sinapsi z  $\tau_c = 50.0$  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

Verjetno potrebno več uvoda v spodbujevano učenje. Tu morda lahko prikažemo še par osnovnih mehanizmov sinaps, na primer asociacije med nevroni, ki se pogosto prožijo in uporaba tega za primer klasičnega pogojevanja tudi brez nagrade.

## 4.1 R-STDP učenje

Imamo klasičnega agenta spodbujevanega učenja, ki dobi informacijo o zunanjem okolju preko stimulacije vhodnih nevronov, nato pa kot odziv na trenutno stanje izbere akcijo, ki zunanje okolje spremeni. V kolikor smo se znašli v nagrajenem stanju bomo agenta nagradili z nagrado. Preko nagrajevanja in interagiranja z okoljem se bo agent naučil akcij, ki privedejo do nagrade v določenem stanju.

Za začetek bo naš agent sestavljen iz  $N_{in}$  vhodnih nevronov, ki predstavljajo možna stanja in bodo povezani z  $N_a$  nevroni na izhodu. Vhod in izhod sta povezana po režimu *all-to-all*, kjer so vsi nevroni vhoda povezani z vsemi nevroni izhoda. Vzratnih povezav tu ne dopuščamo. Za mehanizme ob prisotnosti vzratnih povezav glej poglavje **Rekurenčne povezave**. Ob prihodu v določeno stanje ustrezen 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 je tekom trenutnega stanja imel najvišje število impulzov. V kolikor vstopimo v nagrajeno stanje, bomo  $N_{dopa}$  dopaminskih nevronov stimulirali s 600 pA tokom. Dopaminski nevroni ob impulzu projicirajo dopamin enakomerno med vse povezave med vhodnimi in izhodnimi nevroni.

Nagrada, ki jo neposredno predstavlja aktivnost dopaminskih nevronov bo vedno veljša ali enaka 0, kar pomeni, da morajo povezave, ki predstavljajo izbiro določene akcije v določenem stanju med seboj tekmovati za prevlado. Pri tem moramo omogočiti dovolj veliko varianco med impulzi izhodnih nevronov predvsem v začetni fazi, ko so vse povezave približno enako velike. V nasprotnem primeru bodo vse povezave posodobljene za približno enako vrednost glede na RSTDP.

Varianco med impulzi pri enakih povezavah dosežemo z zunanjim šumom. Biološko najbolj realističen je poissonski šum, saj predstavlja impulze nevronov, zaradi zunanjih stimulusov nepovezanih s trenutnim stanjem.

$$P(k \text{ impulzov v } \Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \quad k = 0, 1, 2, \dots \quad (4.1)$$

Naš agent bo uporabljal model nevrona z eksponentnim jedrom, saj tako poissonski šum povzroči večjo varianco izhodnih nevronov kot model z alfa jedrom, kot demonstrirano v poglavju **Izbira modela nevrona**. V začetni fazi bodo tako akcije v večini 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 čedalje manj odvisna od šuma.



Slika 4.1: Primer učenja na preprosti nalogi s tremi stanji. Prehod v vsako stanje je naključno, v vsakem stanju pa je samo ena izbira akcije nagrajena. V stanju 0 (input neuron 0) je pravilna akcija 0 (motor neuron 0), v stanju 1 akcija 1, v stanju 2 akcija 2. Razvidna je prevlada pravih sinaps in višanje divergence v sinapsah skozi čas ter višanje povprečne nagrade tekom učenja. V simulaciji uporabljamo privzete NEST parametre za nevrone tipa *iaf\_psc\_exp* ter zakasnjene dopaminsko modulirane sinapse s parametri  $W_{\min} = 500$ ,  $W_{\max} = 2000$ ,  $\tau_c = 5$  ms,  $\tau_{c,\text{delay}} = 200$  ms,  $\tau_n = 10$  ms,  $\tau_+ = \tau_- = 20$  ms,  $b = 0.1$ ,  $A_+ = 0.7$ ,  $A_- = 0.3$  ter sinaptično zakasnitev 0.5 ms, poissonski šum z  $\lambda = 1000$  in utežjo sinaps  $w_{\text{poisson}} = 100$ . Sinapse med vhodnimi in izhodnimi nevroni so inicializirane na  $w_{\text{motor}} \sim \mathcal{N}(1300, 1)$ .



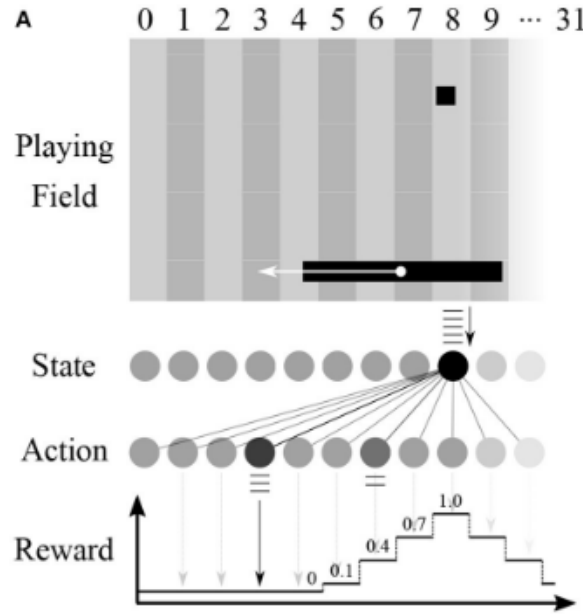
### 4.1.1 Igra Pong

V nadaljevanju bomo R-STDP predstavili na agentu, ki igra *Pong*. R-STDP učenje je kratkovidno, kjer se bomo naučili akcij, samo če nagrada sledi nemudoma, ne pa, če je nagrada zakasnjena. Za zakasnjene nagrade uporabljamo TD (*angl. Temporal Difference*) učenje, ki ga implementiramo v poglavju **TD učenje in model actor-critic**. Igra Pong v osnovi zahteva veliko predvidevanja, vendar lahko igranje igre poenostavimo v obliko, ki se jo lahko naučimo z R-STDP učenjem. Igro bomo v nadaljevanju 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 prožno odbije žogica. V kolikor bi v učenje vključili predvidevanje, bi morali stanja agenta definirati kot kartezični produkt x,y pozicije žogice, njene smeri in y pozicije platforme, lahko pa problem poenostavimo v problem sledenja žogici enako kot v delu Wunderlich T, et al. 2019, kjer agent izira željeno ciljno točko platforme. Tako stanja kot akcije agenta so tako 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 željene pozicije  $j$  od trenutne y pozicije žogice  $k$  in povprečno nagrado  $\bar{R}_i$  v iteraciji  $i$ . S pomočjo povprečne nagrade omejimo krepitev sinaps v kolikor te ne izboljšajo trenutne politike.

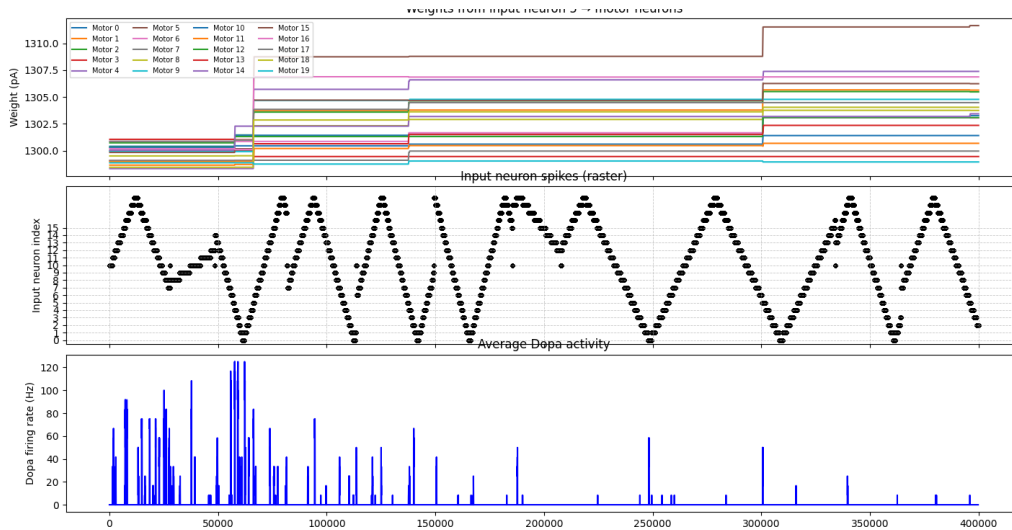
$$R_b = \begin{cases} 1 - |j - k| \cdot 0.3 & \text{if } |j - k| \leq 3, \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

$$I_R = \max(R_b - \bar{R}_i, 0) \cdot 600 \text{ pA} \quad (4.3)$$

Pričakujemo, da bodo sorazmerno oddaljenosti v posameznih stanjih prevladale sinapse, ki iz vhodnega nevrona vodijo do akcij okrog istoležnega izhodnega nevrona. Polje bomo po y osi diskretizirali na 20 stanj.



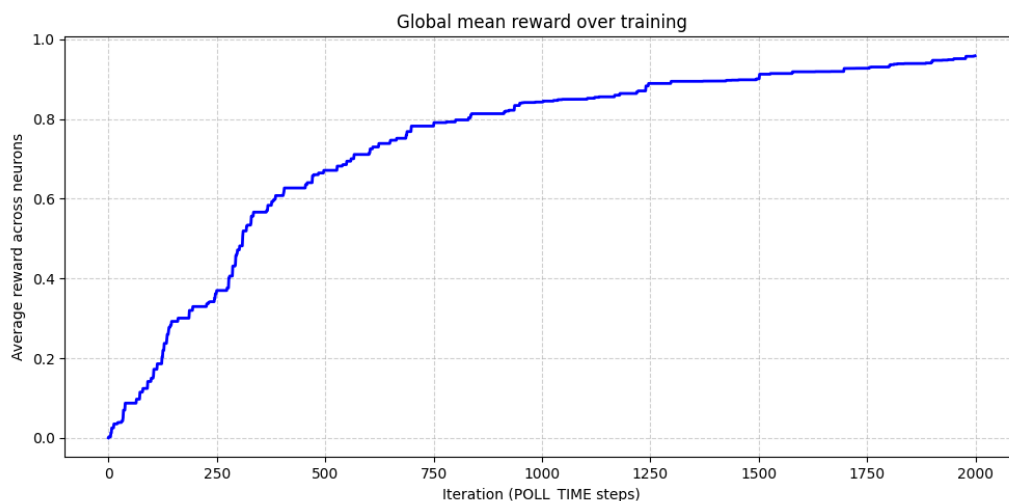
Slika 4.2: Grafična predstavitev agenta in okolja (Wunderlich T, et al. 2019)



Slika 4.3: Graf povezav med vhodnim nevronom, ki predstavlja  $y = 5$  pozicijo in 20 izhodnimi nevroni, kjer tekom učenja prevladuje motorični nevron 5. Motorična nevrona 4 in 6 pa sta druga po vrsti. Za simulacijo smo uporabili enake parametre kot pri sliki 4.1

## Rezultati

Učenje spremljamo preko povprečne nagrade prejete ob prehodih stanj, ki se bliža maksimalni nagradi  $R_{\max} = 1.0$ .



Slika 4.4: Povprečna nagrada  $\bar{R}_i$  tekom 2000 iteracij po 200ms

**Vključi še graf dolžine neprekinjene igre (do zgrešitve) kot posledica učenja.**

Kot že omenjeno je takšno učenje učinkovito samo pri nagradah, ki niso oddaljene, oziroma drugače povedano, se agent ne bo naučil potencialne poti skozi različna nenagrajena stanja, da pride do končne nagrade. To je vidno pri nalogi iskanja oddaljene nagrade v mreži, kjer se agent lahko premika levo, desno, gor in dol. Agent se bo namreč naučil prehoda samo iz stanj neposredno ob cilju.

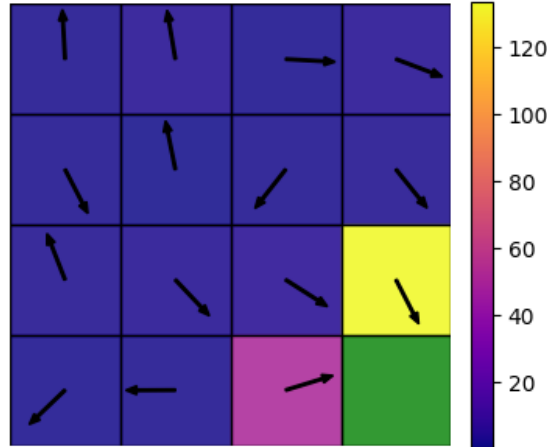
Ob učenju bomo agenta nagradili ko preide v končno stanje in ga po tem postavili v naključno stanje. Trenutno politiko agenta bomo predstavili s puščicami s smerjo, ki jo določa normaliziran vektor  $\hat{x}_i$  v vsakem od stanj  $i$ , ki predstavljajo preferenco akcije glede na medsebojne razlike v utežeh sinaps.

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

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

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

Za prikaz "samozavesti" pri izbiri akcije v stanju  $i$  kot rezultat učenja, bomo polja ustrezno obarvali glede na maksimalno razliko med utežmi med vhodnim nevronom  $i$  in vsakim od izhodnih nevronov.



Slika 4.5: Prikaz politike po 500 iteracijah po 200ms. Končno stanje je obarvano z zeleno.

Rezultat potrjuje, da se agent ni sposoben naučiti poti do nagrade iz poljubnega stanja, vendar samo iz stanj neposredno ob nagradi.

## 4.2 TD učenje in model actor-critic

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 pri TD(0) je

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

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

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

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

Verjetno daljši uvod v TD učenje?

### 4.2.1 Model actor-critic

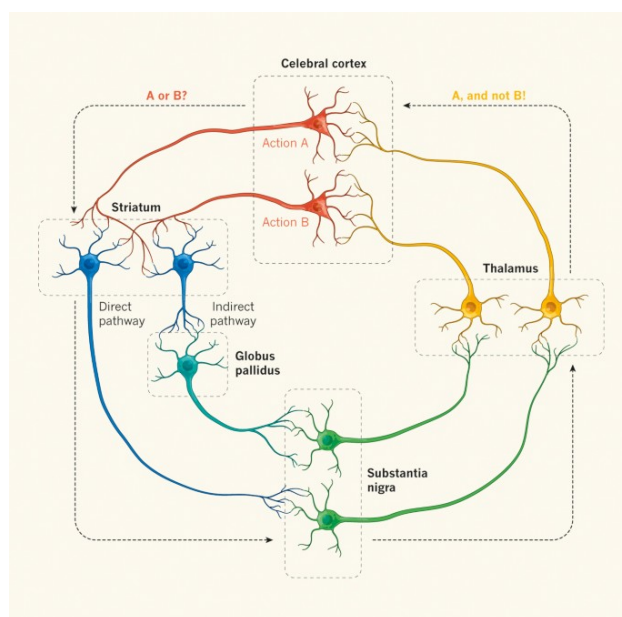
TD učenje bomo implementirali z modelom akter-kritik (*angl. actor-critic*), po zgledu Wiebke P, et al. 2011 na nalogi z mrežo. Za razliko od sistema, ki ga predstavijo Wiebke P, et al. bomo za sinapso uporabili našo zakasnjeno RSTDP, ki omogoča pripisovanje odgovornosti povezavam poljubno v preteklost, kar je pomembno pri učenju, kjer stanja niso definirana v diskretnih intervalih. Poleg tega upoštevamo tako kavzalne kot tudi antikavzalne impulze, po pravilu RSTDP med vsemi pari v zgodovini impulzov (*all-to-all* namesto *next-neighbor* **razloži**). Sistem, ki ga bomo implementirali je tudi

manjši glede na število nevronov, za višjo hitrost simulacije. Stanja bomo razdelili v intervale dolžine 200ms ob prehodu stranj pa bomo vhodne nevrone stimulirali enako kot pri RSTDP. Pri prehodu med stanji lahko določen del stimuliranih nevronov iz prejšnjega stanja postane asociiranih (**pokaži**) z akcijo naslednjega stanja, kar v osnovi ni napačno in je posledica uporabe RSTDP sinapse, vendar bomo za bolj učinkovito učenje med prehodi stanj prekinili stimulacijo 50ms 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.

**tu verjetno potrebna bolj podrobna razlaga...**

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 oziroma bolj konkretno bazalnih ganglijih. Bazalni gangliji so skupina jedrov v možganih, ki igrajo ključno vlogo pri nadzoru gibanja, učenju akcij in odločanju, poleg tega pa realizira obliko TD učenja. Akter-kritik je poenostavitev in abstrakcija resničnih mehanizmov v možganih, vendar uporablja podobne mehanizme. V bazalnih ganglijih in modelu akter-kritik, kot ga predstavlja Wiebke P, et al. razlikujemo dve glavni poti: direktno in indirektno pot, ki vodita iz *striatuma* do dopaminergičnih nevronov. Direktna pot je zakasnjena inhibitorna pot, ki poteka neposredno od striatuma do dopaminergičnih nevronov, indirektna pot pa je inhibitorna do *ventralnega palliduma*, posebne skupine nevronov, ki inhibira aktivnost dopaminergičnih nevronov. Ob prisotnosti neke osnovne od 0 različne frekvence nevronov *ventralnega palliduma* bo tako indirektna pot imela ekscitatoren učinek na dopaminergične nevrone. Indirektna in direktna povezava delujeta konkurenčno. Indirektna pot ima minimalen zamik in aktivnost striatuma v trenutnem stanju neposredno preslika na povišano aktivnost dopaminergičnih nevronov. Hkrati v času nahajanja v trenutnem stanju direktna povezava inhibira dopaminergične nevrone sorazmerno z ak-

tivnostjo striatuma, kot je ta bila v prejšnjem stanju, zaradi zakasnitve. Indirektna in direktna povezava tako skupaj računata TD-napako  $\delta_t$ , ki bo v trenutnem stanju glede na izračunan estimat vrednosti trenutnega stanja okrepila sinapse prejšnjega stanja, ki so izbrale aktivnost, ki nas je pripeljala v to stanje. Povprečna teža sinaps med vhodnim nevromom  $i$  in striatumom tako neposredno predstavlja vrednost 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. V primeru, da se premaknemo v stanje s približno isto povprečno utežjo povezave do striatuma pa se bosta direktna in indirektna povezava izničili, dopaminergični nevroni pa se bodo prožili po frekvenci, ki jo definira externi poissonske šum.



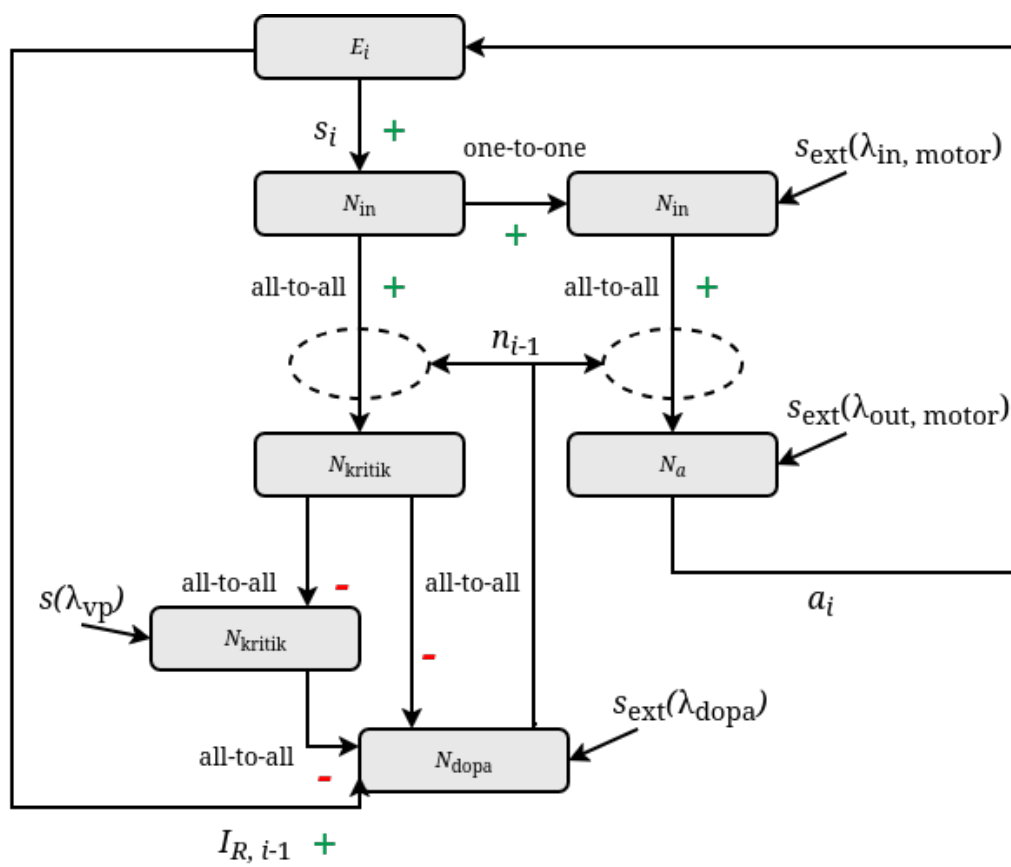
Slika 4.6: Direktna in indirektna pot v bazalnih ganglijah

Model, kot ga predlaga Wiebke P, et al. funkcionalno združuje *substantia nigra* in *thalamus* kot dopaminergični nevroni in signal do povezav med vhom in striatumom in vhom in izhodnimi motoričnimi nevromi. Ventralno *globus pallidus* se nahaja *ventral pallidum*, ki predstavlja del pallidusa povezan z pričakovanjem nagrade in odločanjem. Wiebke P, et al. tako za

nevrone na indirektni poti uporablja ta izraz.

Kot akter bomo v nadaljevanju uporabili RSTDP del kot smo ga implementirali prej. Od implementacije Wiebke P, et al. se poleg sinapse naša implementacija razlikuje tudi v načinu izbire stanja, kjer mi za izbrano stanje vzamemo stanje z maksimalnim številom impulzov za razliko od izbire prvega izhodnega nevrona, ki se je sprožil kot rezultat stimulacije v trenutnem stanju. Ta način bolj direktno okrepi povezave odgovorne za izbrano aktivnost, saj po prvem izhodnem impulzu inhibira vse ostale izhodne nevrone, kar izniči njihovo *eligibility* sled. Tako ni potrebe po tekmovanju sinaps, kot pri naši metodi, vendar moramo preveriti impulze izhodnih nevronov v vsakem koraku simulatorja. V primeru simulatorja NEST je to vsakih 0.1ms, kar pa je problematično, saj simulator teče v C++ zaledju, ki ga zapustimo takoj ko prekinemo simulacijo. Tako je bistvena razlika med tem, da 100krat poženemo ukaz `nest.Simulate(0.1)` ali enkrat `nest.Simulate(10)`. V akterju bomo zaradi razlogov navedenih v poglavju **RSTDP učenje** uporabljali model nevrona z eksponentnim jedrom, v kritiku, pa bomo poskusili uporabiti biološko bolj realistične nevrone z alfa jedrom. Ker bodo vhodni nevroni akterju in kritiku skupni, bomo med vhodnimi nevroni in izhodnimi nevroni dodali dodaten nivo nevronov, ki bo višjo frekvenco potrebno za stimulacijo nevronov kritika z alfa jedrom znižal na frekvenco ustrezno za nevrone akterja z eksponentnim jedrom. S tem smo tudi zmanjšali povezanost dveh delov, kar olajša iskanje ustreznih hiperparametrov, poleg tega pa lahko tudi šum za potrebe RSTDP učenja dovajamo ločeno od kritika.





Slika 4.7: Prikaz implementiranega aktor-kritik sistema

### 4.2.2 Izbira parametrov

Parametri so bili izbrani eksperimentalno in se ne ozirajo 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 ekscitacija zaradi direktne in indirektne povezave v ravnovesju. Redukcija 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 **R-STDP učenje**. Navedene so konstante in parametri implementiranega modela. Parametri, ki niso prikazani v tabeli imajo privzeto vrednost NEST simulatorja.

Tabela 4.1: Parametri simulacije

Simbol	Pomen	Vrednost
POLL_TIME	Čas simulacije na iteracijo	200
$f(s_{in,i})$	frekvenca stimulacije vhodnega nevrona $i$	100 Hz
$n_{critic}$	Število nevronov v skupinah nevronov kritika	8

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

Tabela 4.2: Parametri nevronov

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

Tabela 4.3: Parametri sinaps med vhodnimi in motoričnimi nevroni

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

Tabela 4.4: Parametri sinaps med vhodom in striatumom

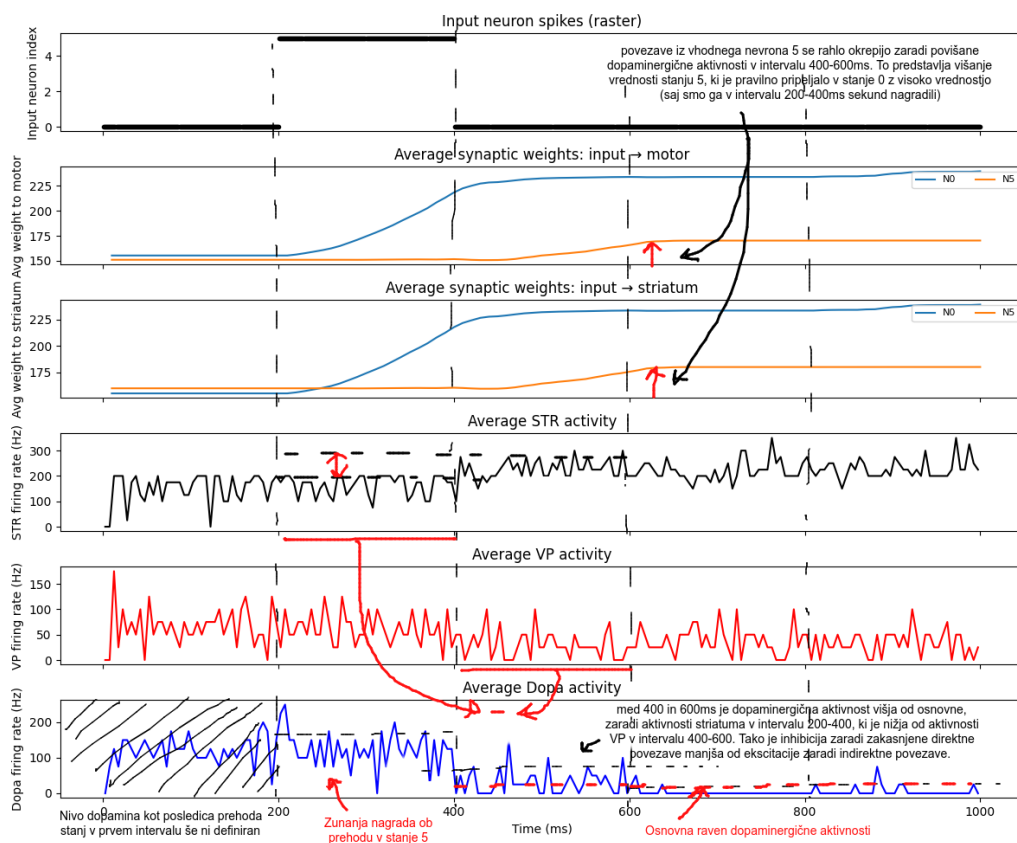
Simbol	Pomen	Vrednost
tip	Tip sinapse	Privzeta konstantna NEST sinapsa
$w_{\text{str} \rightarrow \text{vp}}$	Uteži sinps med striatumom in ventral pallidumom	-50
$w_{\text{str} \rightarrow \text{dopa}}$	Uteži sinps med striatumom in dopaminergičnimi nevroni	-55
$w_{\text{vp} \rightarrow \text{dopa}}$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

Tabela 4.5: Parametri sinaps kritika

Simbol	Pomen	Vrednost
$\lambda_{\text{vp}}$	Povprečna hitrost šumnih impulzov nevronov ventral palliduma	5200
$\lambda_{\text{dopa}}$	Povprečna hitrost šumnih impulzov v 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

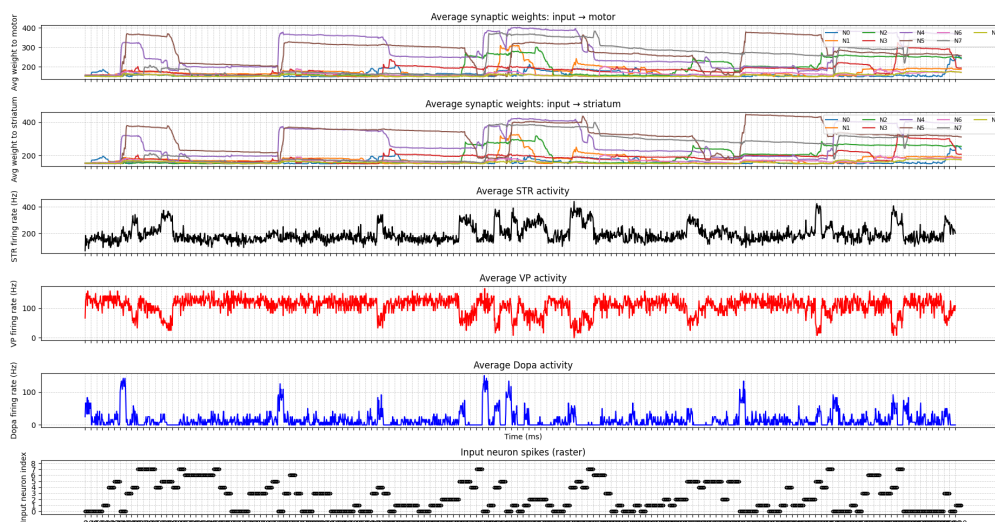
Tabela 4.6: Parametri generatorjev šuma

### 4.2.3 Učenje



Slika 4.8: Prikaz obnašanja sistema ob prehodu iz stanja 0 v nagrajeno stanje 5 in nazaj v stanje 0. Pričakovana nagrada in tako tudi vrednost stanja 0 se ob prehodu v stanje 5 zviša preko povezav do striatuma. Ob prehodu iz nagrajenega stanja (ki pa samega sebe še ne ocenjuje z visoko vrednostjo) nazaj v stanje 0 preidemo iz stanja z osnovnimi utežmi v stanje 0 z okrepljenimi utežmi do striatuma, torej prehod v stanje z višjo vrednostjo. Posledica tega je, napram osnovni frekvenci dopaminergičnih nevronov, povišana dopaminergična aktivnost, ki povzroči aktivnosti sorazmerno povišanje uteži sinaps stanja 5 do striatuma. Ob prehodu iz stanja 0 v stanje 0 se vrnemo k osnovni dopaminergični frekvenci.

### Potrebna podrobnejša razlaga...



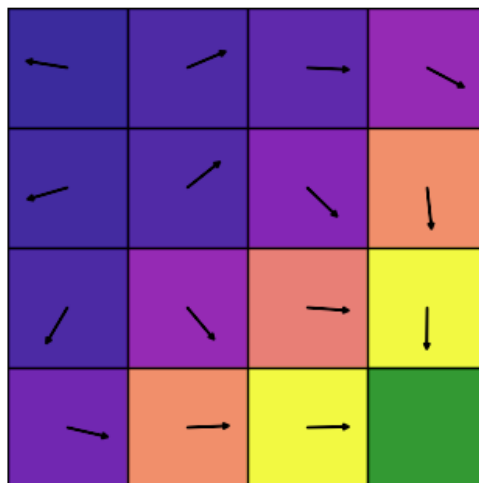
Slika 4.9: Obnašanje sistema tekom učenja na 3x3 mreži. Polja so oštevilčena od leve proti desni od zgoraj navzdol. Cilj se nahaja na polju 8. Povezave vhoda do striatuma stanj 5 in 7 so pričakovano najvišje, sledi pa jim 4, ki neposredno vodi v 5 in 7

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



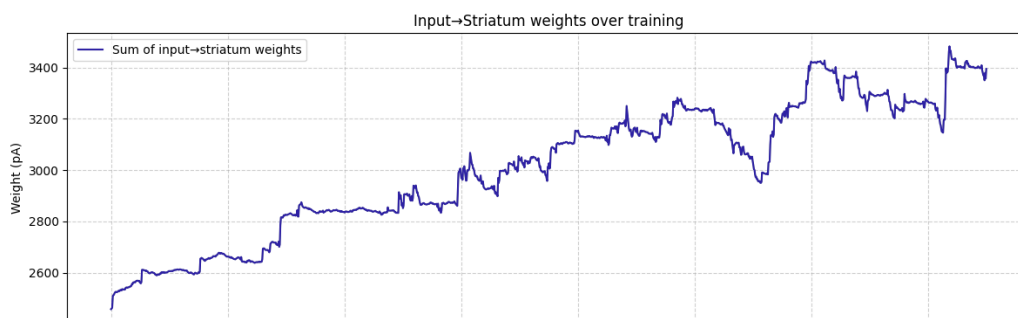
#### 4.2.4 Rezultati

Naučeno politiko bomo prikazali podobno kot v poglavju **R-STDP učenje**, vendar bomo samozavest izbire akcije v določenem stanju prikazali skupaj z povprečno utežjo povezav med vhodnimi nevroni pripadajočega stanja in striatumom.



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

Dodaj skalo in razloži preslikavo povprečne uteži v barve polj.



Slika 4.11: Povprečna utež preko sinaps med vsemi vhodnimi nevroni in striatumom tekom 1500 iteracij po 200 ms.

### Interpretacija rezultatov

Sistem se uspešno uči politike prehodov iz poljubnega stanja do oddaljene nagrade v kolikor ni nagrada preveč oddaljena. Polja 0 in 4 kažeta v napačno smer, vendar vidimo, da je tudi stanje ocenjeno z nizko vrednostjo. Daljše učenje pri trenutni konfiguraciji ne pripelje do boljših rezultatov.

**Potrebna nadaljna analiza.** Glavna omejitev učenja je konstantna rast sinaps tekom učenja. Sinapse v trenutnem modelu skozi čas same po sebi odtekajo v povprečju počasneje kot rastejo. Tako je nastavljeno, saj se pri redkih nagradah in predvsem v zgodnjih fazah učenja razlike v sinapsah zaradi nagrade bolje ohranijo.

## 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 nevronske mreže zunaj simuliranega okolja, na trdo-ožičenih nevronske 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 spodbujevanja učenja na impulznih nevronske 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

### Diskusija in zaključek

To-Do...



# 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.
- Šutar M (2023). “Uporaba predvidevanja akcij nasprotnika pri učenju inteligentnega agenta”. Diplomaska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- Svete A (2020). “Posplošitev problema vozička s palico na zahtevnejše domene”. Diplomaska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- T, Štromajer (2022). “Using machine learning to train a shepherd dog”. Magistrska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.
- Tsodyks, tMisha, Asher Uziel in Henry Markram (2000). “t Synchrony Generation in Recurrent Networks with Frequency-Dependent Synapses”. V: *Journal of Neuroscience* 20.1, RC50–RC50. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.20-01-j0003.2000. eprint: <https://www.jneurosci>.

org/content/20/1/RC50.full.pdf. URL: <https://www.jneurosci.org/content/20/1/RC50>.

Wiebke P, et al. (2011). “An Imperfect Dopaminergic Error Signal Can Drive Temporal-Difference Learning”. V: *PLoS computational biology* 7.5. DOI: 10.1371/journal.pcbi.1001133.

Wunderlich T, et al. (2019). “Demonstrating Advantages of Neuromorphic Computation: A Pilot Study”. V: *Frontiers in neuroscience* 13.260. DOI: 10.3389/fnins.2019.00260.