float

Parametri nevronskega modela

Nevronski modeli, uporabljeni v tej študiji, temeljijo na tokovno gnanem modelu uhajajočega integrirajočega nevrona (leaky integrate-and-fire). Dinamiko membrane določajo naslednji parametri, ki jih omogoča simulator NEST:

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 odsotnosti od 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*), ki pa je simulator Nest ne podaja kot neodvisen parameter. 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_{reset} — **potencial ponastavitve**
  Ponastavitveni membranski potencial.

- $\tau_{\text{syn,ex}}$ — **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.

- $\tau_{\text{syn,in}}$ — **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**

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 po naslednji enačbi

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

Model z eksponentnim jedrom

V simulatorju NEST je model z eksponentim jedrom (iaf_psc_exp) definiran po sistemu diferencialnih enačb prvega reda, ki jih navaja Tsodyks et. al [**?**]. Postsinaptični tok $y(t)$ se spreminja po sistemu

$$\frac{dx}{dt} = \frac{z}{\tau_{rec}} - ux\delta(t - t_{sp}) \tag{2}$$

$$\frac{dy}{dt} = -\frac{y}{\tau_I} + ux\delta(t - t_{sp}) \tag{3}$$

$$\frac{dz}{dt} = \frac{y}{\tau_I} - \frac{z}{\tau_{rec}} \tag{4}$$

kjer t_{sp} predstavlja čas presinaptičnega impulza, τ_I čas sinaptičnega odtekanja, 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.

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

$$\frac{dy}{dt} = -\frac{y}{\tau_I} \tag{5}$$

rešitev te diferencialne enačbe je tako

$$y(t) = y_0 e^{-t/\tau_I} \tag{6}$$

kjer vidimo, da je jedro res exponentna 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 τ_I, ki pa je v simulatorju NEST predstavljen s $\tau_{\text{syn, X}}$.

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

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

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

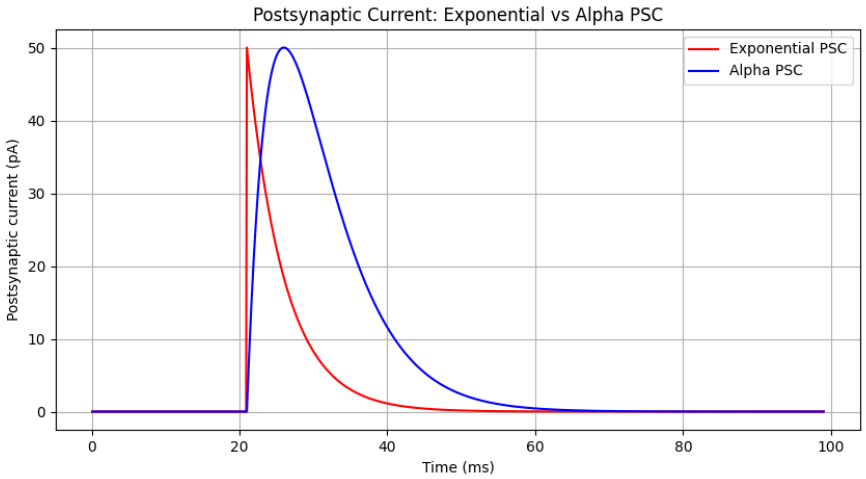

Figure 1: Postsinaptični tok modela z alfa in eksponentim jedrom

Izbira modela nevrona

V sistemih, ki jih bomo implementirali v nadaljevanju skušamo skušamo pri modeliranju mehanizmov v človeških možganih uporabiti čimmanj poenostavitev 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 esponentnim 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_{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 zlahka prilagodimo z nižjimi vrednostmi uteži sinaps. Razlika v varianci frekvenc impulzov je posledica daljšega časovnega intervala, kjer je postsinaptični tok blizu maksimalne vrednosti pri alfa jedru napram eksponentnem, kjer je tok blizu maksimalne vrednosti za zelo kratek čas. Zaradi tega bodo zaporedni postsinaptični impulzi skozi čas precej bolj prekrivni. Pri intergriranju različnih postsinaptičnih tokov sozi čas pride do učinka nizko prepustnega filtra, ki ublaži nenadne spremembe v amplitudi skupnega toka na vhodu v postsinaptični nevron. Posledica so manjše razlike v frekvenci impulzov postsinaptičnega nevrona, če imamo na vhodu sinapse različnih uteži, učinek pa je še bolj opazen pri dodanem šumu. Pri alfa jedru bo namreč šum povzročil manj variance v frekvenci impulzov postsinaptičnega nevrona, kot pri eksponentnem jedru.

Table 1: Parametri simulacije uporabljeni pri primerjavi modelov nevronov.

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
Utež sinapse (Exp PSC)	25.0
Utež sinapse (Alpha PSC)	25.0 / $e \approx 9.20$
Frekvenca Poissonovega šuma	8000 Hz na nevron

Table 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^2)
Exponentno	7.846 ± 0.021	$\mathbf{0.402 \pm 0.028}$
Alfa	7.800 ± 0.023	$\mathbf{0.270 \pm 0.006}$

0.0.1 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 moči 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:

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

kjer so:

- A_+ in A_- multiplikatorja za potenciranje in depresijo,
- τ_+ in τ_- 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:

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

kjer so:

- w — sinaptična utež,
- c — *eligibility 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,
- τ_c, τ_n — časovne konstante odtekanja *eligibility* in dopaminskih sledi.

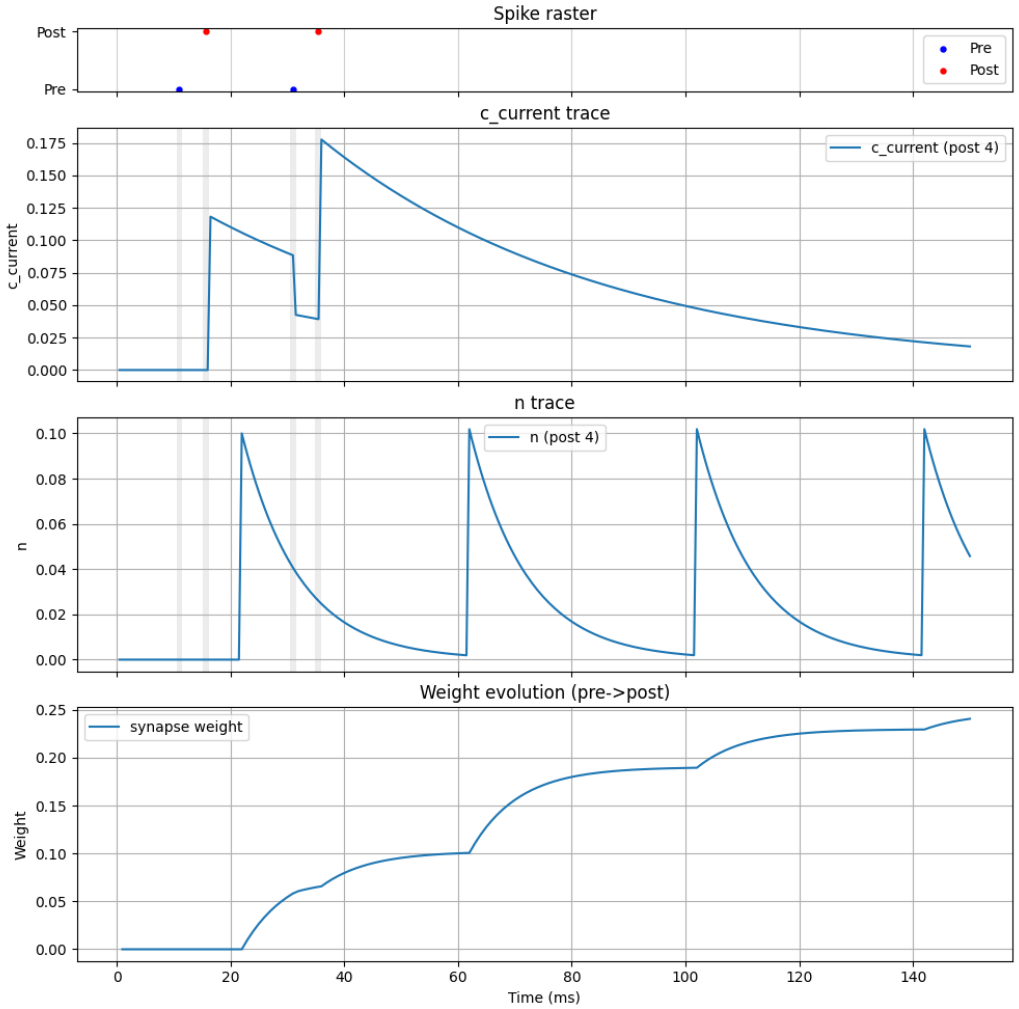

Figure 2: *eligibility* sled c, dopaminska sled n in evolucija sinaptične uteži pri presinaptičnih impulzih pri $[10.0, 30.0]$ ms in postsinaptičnih inpulzih pri $[12.0, 32.0]$ ms, simulirane preko 150 ms pri R-STDP sinapsi z $\tau_c = 50.0$ ms, $\tau_{\text{c,delay}} = 50.0$ ms, $\tau_n = 10.0$ ms, $\tau_{\text{plus}} = 10.0$ ms, $b = 0.0$, $A_{\text{plus}} = 0.2$, $A_{\text{minus}} = 0.2$, in sinaptično zakasnitvijo 0.5 ms.

V poglavju R-STDP 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 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 posodaljanje sinaps zaradi nagrad, ki pridejo prehitro znotraj določenega intervala $\tau_{c,\text{delay}}$. Celotno *eligibility* sled bomo tako premaknili za $\tau_{c,\text{delay}}$

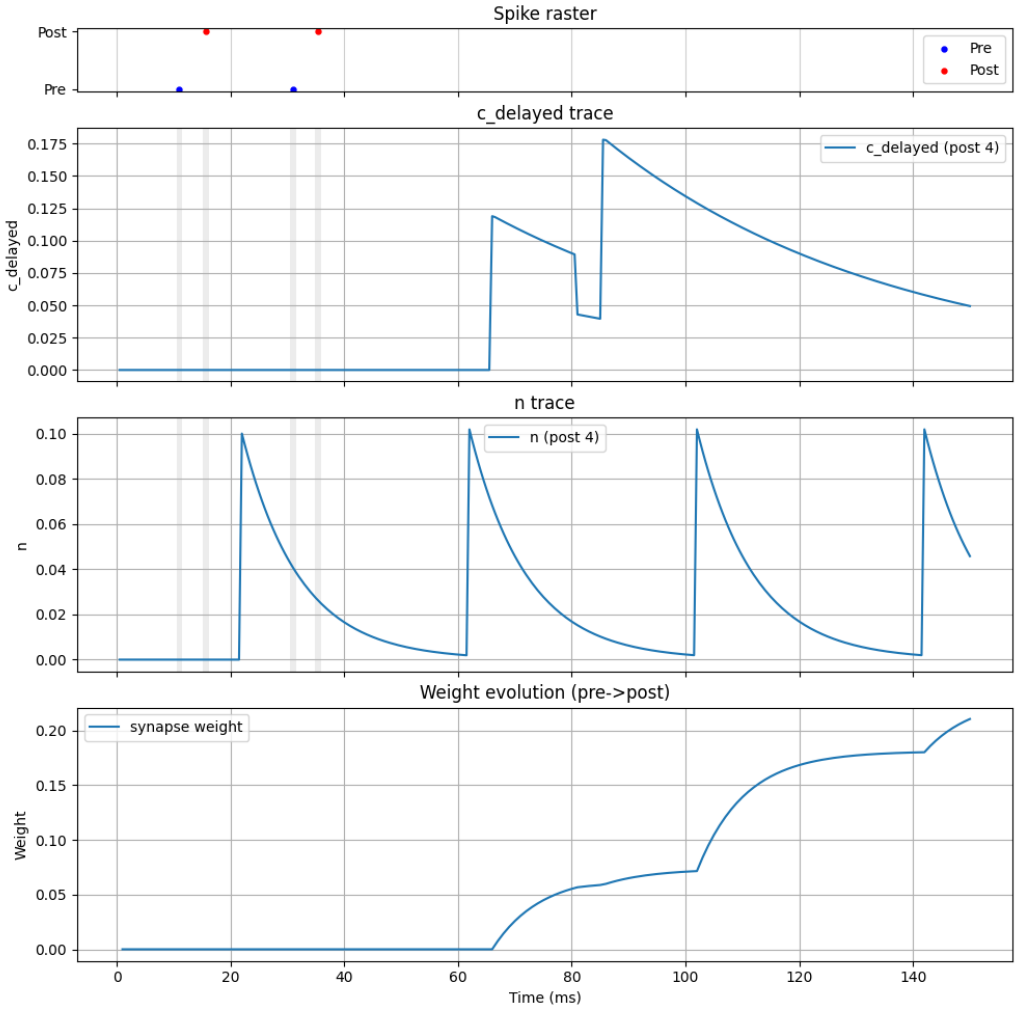

Figure 3: *eligibility* sled c, dopaminska sled n in evolucija sinaptične uteži pri presinaptičnih impulzih pri $[10.0, 30.0]$ ms in postsinaptičnih inpulzih pri $[12.0, 32.0]$ ms, simulirane preko 150 ms pri R-STDP sinapsi z $\tau_c = 50.0$ ms, $\tau_{\text{c,delay}} = 50.0$ ms, $\tau_n = 10.0$ ms, $\tau_{\text{plus}} = 10.0$ ms, $b = 0.0$, $A_{\text{plus}} = 0.2$, $A_{\text{minus}} = 0.2$, in sinaptično zakasnitvijo 0.5 ms.

0.0.2 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_s 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. Vzvratnih povezav tu ne dopuščamo. Za mehanizme ob prisotnosti vzvratnih povezav glej poglavje **Rekurenčne povezave**. Ob prihodu v določeno stanje ustrezen vhodni nevron stimuliramo tako, da oddaja impulze s frekvenco 100 Hz za čas 200 ms. 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 projecirajo dopamin enakomerno med vse povezave med vhodnimi in izhodnimi nevroni.

Nagrada, ki jo neposredno predstavlja aktivnost dopaminskih nevronov bo vedno velčja 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. Varianvo med impulzi pri enakih povezavah dosežemo z zunanjim šumom. Biološko najbolj realističen je poissonski šum, saj prestavlja impulze nevronov, zaradi zunanjih stimulusov nepovezanih s trenutnim stanjem.

$$P(k \text{ impulzov v } \Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \quad k = 0, 1, 2... \tag{7}$$

Naš agent bo uporabljal model nevrona z eksponentim jedrom, saj tako poissonski šum povzroči večjo varianco izhodnih nevronov kot model z alfa jedrom, kot prikazano 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 izhod-

nih nevronov. Tako bo v kasnejših fazah učenja izbira akcije čedalje manj odvisna od šuma.

Figure 4: Primer učenja na preprosti nalogi s tremi stanji. Prehod v vsako stanje je nakjučno, 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 pravilnih 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 $W_{\min} = 500$, $W_{\max} = 2000$, $\tau_c = 5\,\text{ms}$, $\tau_{c,\text{delay}} = 200\,\text{ms}$, $\tau_n = 10\,\text{ms}$, $\tau_+ = \tau_- = 20\,\text{ms}$, $b = 0.1$, $A_+ = 0.7$, $A_- = 0.3$ ter sinaptično zakasnitev $0.5\,\text{ms}$, poissonski šum z $\lambda = 1000$ in utežjo sinaps $w_{\text{poisson}} = 1000$. Sinapse med vhodnimi in izhodnimi nevroni so inicializirane na $w_{\text{motor}} \sim \mathcal{N}(1300, 1)$.

0.1 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 z x,y pozicijo žogice, njeno smerjo in y pozicijo platforme, lahko pa problem poenostavimo v problem sledenja žogici enako kot v delu Wunderlich T, et al. [**?**], 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_{dopa}, ki je sorazmeren razliki med nagrado R_b izračunani glede na oddaljenost željene pozicije j od trenutne y pozicije žogice j in povprečno nagrado $\bar{R}_i$ v iteraciji i. S pomočjo povprečne nagrade omejimo krepitev sinaps v kolikor te ne izboljšajo trenutne politike.

$$R_b = \begin{cases} 1 - |j - k| \cdot 0.3 & \text{if } |j - k| \leq 3, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

$$I_{\text{dopa}} = \max(R_b - \bar{R}_i, 0) \cdot 600 \text{ pA} \tag{9}$$

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.

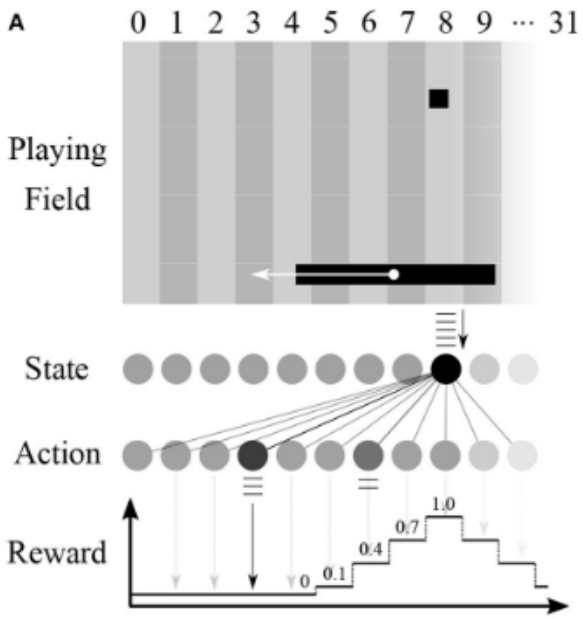

Figure 5: Grafična predstavitev agenta in okolja [**?**]

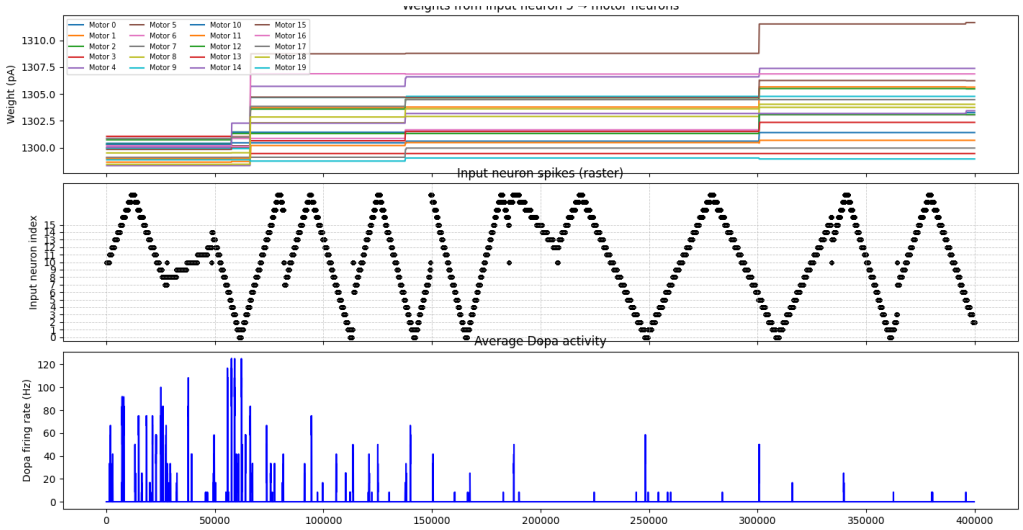

Figure 6: Graf povezav med vhodnim nevronom, ki predstavlja $y = 5$ pozicijo in 20 izhodnimi nevroni, kjer tekom učenja prevladuje motorični nevron 5. Motorična nevrona 4 in 6 pa sta druga po vrsti. Za simulacijo smo uporabili enake parametre kot pri sliki 4

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

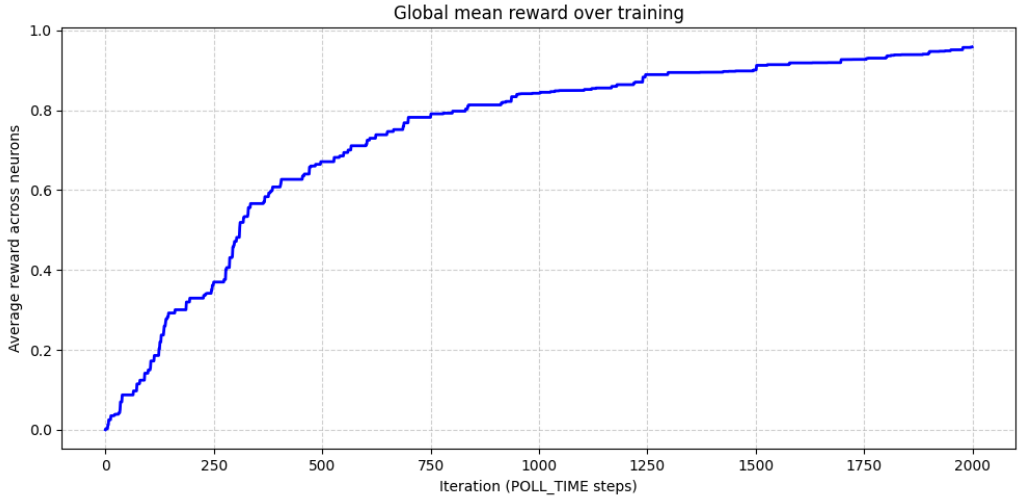

Figure 7: Povprečna nagrada $\bar{R}_i$ tekom 2000 iteracij po 200ms

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 nauči prehoda samo iz stanj neposredno ob cilju. 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.

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

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 rezultat učenja, bomo polja ustrezno obarvali glede na maksimalno razliko med utežmi med vhodnim nevronom i in vsakim od izhodnih nevronov.

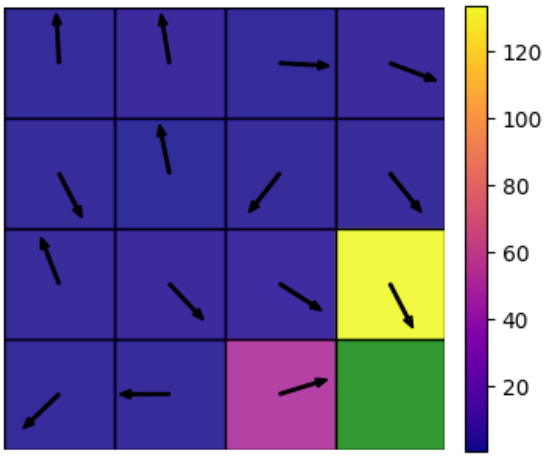

Figure 8: Prikaz politike po 500 iteracijah po 200ms

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

0.2 TD učenje in model actor-critic

ee