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

Na tem mestu zapišite, komu se zahvaljujete za pomoč pri izdelavi diplomske naloge oziroma pri vašem študiju nasploh. Pazite, da ne boste koga pozabili. Utegnil vam bo zameriti. Temu se da izogniti tako, da celotno zahvalo izpustite.

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	Temporalna razlika

Povzetek

Naslov: Spodbujevano učenje na impulznih nevronskih mrežah

Avtor: Matjaž Pogačnik

V vzorcu je predstavljen postopek priprave diplomskega dela z uporabo okolja LaTeX. Vaš povzetek mora sicer vsebovati približno 100 besed, ta tukaj je odločno prekratek. Dober povzetek vključuje: (1) kratek opis obravnavanega problema, (2) kratek opis vašega pristopa za reševanje tega problema in (3) (najbolj uspešen) rezultat ali prispevek diplomske naloge.

Ključne besede: računalnik, računalnik, računalnik.

Abstract

Title: Reinforcement learning on spiking neural networks

Author: Matjaž Pogačnik

This sample document presents an approach to typesetting your BSc thesis using LaTeX. A proper abstract should contain around 100 words which makes this one way too short.

Keywords: computer, computer, computer.

Poglavje 1

Uvod

1.1 Motivacija

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

Delovanje možganov je kljub mnogim raziskavam še vedno precej slabo razumljeno, njihovo računalniško modeliranje pa je v času pisanja še precej mlado področje. Odkrivanje kakršnihkoli 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 drugim 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 reševanja tako instantnih kot zakasnjenih nagrad in temelji izključno na impulznih nevronskih mrežah in spodbujevanem učenju. V prvi fazi so predstavljeni in ovrednoteni različni modeli bioloških značilnosti nevronov in sinaps **??**. V nadaljevanju se posvetimo spodbujevanem učenju. V prvi fazi razvijemo agenta, ki uporablja model sinaptične plastičnost odvisne od nagrajevanja in časovne razporeditvre impulzov (angl. R-STDP, primer Izhikevich EM [3]) in se je sposoben naučiti igranja igre Pong.Tak agent ni sposoben učenja nalog, ki imajo zakasnjene nagrade zato v nadaljevanju 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

Iz diplomskega seminarja... Na temo impulznih nevronskih mrež v Sloveniji v času pisanja še ni bila napisana nobena diplomska ali magistrska naloga, doktorska dizertacija ali znanstveni članek, kar je dodatna motivacija za pisanje diplomske naloge na to temo. Impulzne nevronske mreže zaradi zahtevnosti učenja (v času pisanja) niso kaj dosti uporabljene, čedalje bolj uprabna 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, ki mu kot odziv vrača nagrado in novo stanje okolja. Med drugim je uporabno v problemih kot so navigacija in reševanje problemov z roboti, na temo česar je bil objavljen članek pod avtorstvom prof. dr. Danijela Skočaja, rednega profesorja na FRI in dr. Mateja Dobrevskega [2] . Objavljenih je tudi več diplomskih in magistrskih nalog bolj simuliranih problemov, kot so uporaba spodbujevanega učenja za simulacijo psa ovčarja Štromajer T [7], reševanje problemov sorodnih problemu vozička s palico Svete A [6], igranje iger Šutar M [5] in uporaba TD [8] (angl. Tem-

poral Difference) učenja v Monte Carlo preiskovanju dreves Deleva A [1]. 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, sploh v resničnem svetu, imajo impulzne nevronske mreže lahko določene prednosti. Impulzne nevronske mreže namreč naravno upoštevajo časovno komopnento in procesirajo sekvenčne podatkovne tokove. Ker so dogodki v teh mrežah v osnovi samo 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. [10] 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 za to 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. Pri tem je dober zgled delo Izikhevich EM [3], ki poleg modela nevronov in sinaps vpeljuje še način pripisovanja odgovornosti sinapsam za določeno aktivnst nevronske mreže. V postopku nadrgadnje algoritmov učenje poteka tudi na osnovi TD učenja in njegovi biološko bolj neposredni implementaciji Actor-Critic, Wiebke P, et al. [9]. V tem postopku implementiramo dejansko nevronsko vezje odgovorno za nagrajevanje, kot je bilo to raziskano v človeških možganih.

Poglavje 3

Metodologija 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 4

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 nadaljeanju pa bodo predstavljeni "integrate and fire"modeli nevronov z alpha (*NEST: iaf_psc_alpha*) ali exponentno (*NEST: iaf_psc_exp*) oblikovanimi postsinaptičnimi tokovi. 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.

4.1 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.

- 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**
  Absolutna spodnja meja za membranski potencial.

Membranski potencial V_m se spreminja v odvisnoti od I_{syn} in ostalih parametrov po naslednji enačbi

$$\frac{dV_m}{dt} = -\frac{V_m - E_L}{\tau_m} + \frac{I_{\text{syn}} + I_e}{C_m} \tag{4.1}$$

.

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

4.1.1 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 po naslednji enačbi

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

4.1.2 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 [**expModel**]. Postsinaptični tok $y(t)$ se spreminja po sistemu

$$\frac{dx}{dt} = \frac{z}{\tau_{rec}} - ux\delta(t - t_{sp}) \tag{4.2}$$

$$\frac{dy}{dt} = -\frac{y}{\tau_I} + ux\delta(t - t_{sp}) \tag{4.3}$$

$$\frac{dz}{dt} = \frac{y}{\tau_I} - \frac{z}{\tau_{rec}} \tag{4.4}$$

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

$$\frac{dy}{dt} = -\frac{y}{\tau_I} \tag{4.5}$$

rešitev te diferencialne enačbe je tako

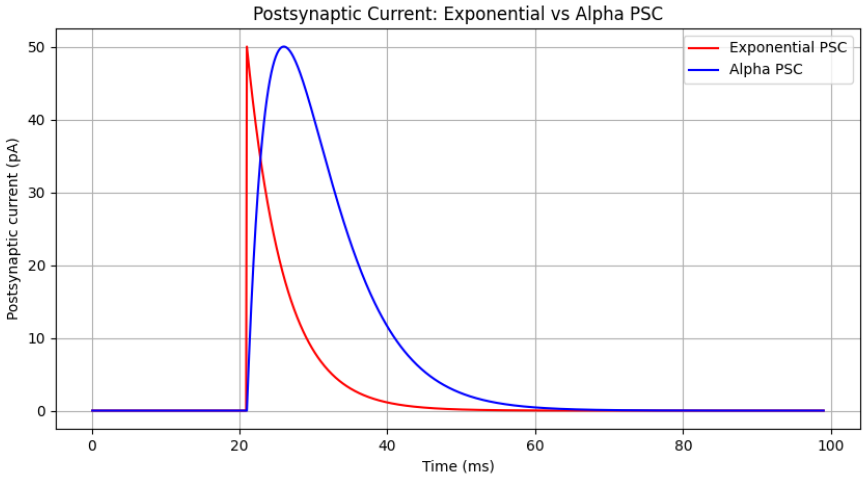
$$y(t) = y_0 e^{-t/\tau_I} \tag{4.6}$$

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

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



Slika 4.1: Postsinaptični tok modela z alfa in eksponentim jedrom

4.1.3 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 zapore-

dni 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.

Tabela 4.1: 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

Tabela 4.2: Povzetek statistike medimpulznih intervalov nevronov z alfa in exponentnim 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	$\mathbf{0.270 \pm 0.006}$

4.2 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 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:

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

kjer so:

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

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

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,
- τ_c, τ_n — časovne konstante odtekanja *eligibility* in dopaminskih sledi.

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 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

$\tau_{c,\text{delay}}$

Poglavje 5

Spodbujevano učenje na impulznih nevronskih mrežah

Verjetno potrebno več uvoda. 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.

5.1 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_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.

Varianco 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.

$$P(k \text{ impulzov v } \Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \quad k = 0, 1, 2... \tag{5.1}$$

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

Slika 5.1: 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 $w_{\text{poisson}} = 100$. Sinapse med vhodnimi in izhodnimi nevroni so inicializirane na $w_{\text{motor}} \sim \mathcal{N}(1300, 1)$.

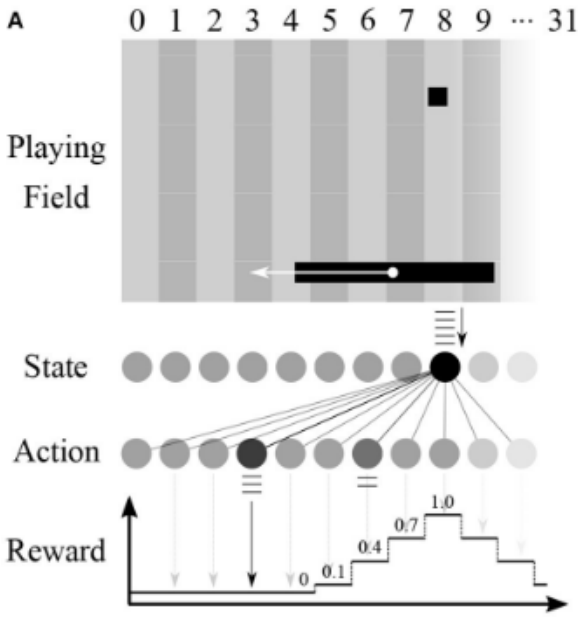
5.1.1 Igra Pong

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 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. [10], 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 j in povprečno nagrado $\bar{R}_i$ v iteraciji i. S pomočjo povprečne nagrade omejimo krepitev sinaps v kolikor te ne izboljšajo trenutne politike.

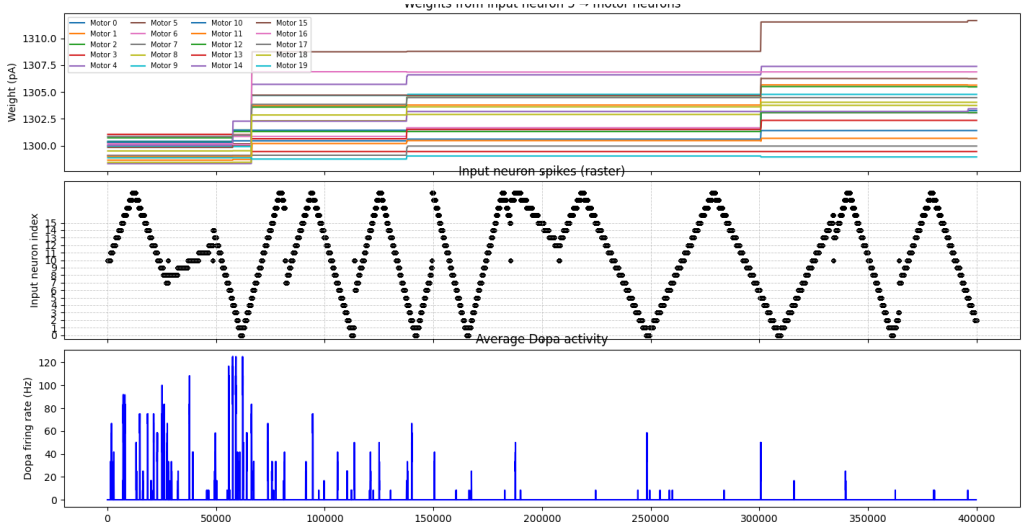
$$R_b = \begin{cases} 1 - |j-k| \cdot 0.3 & \text{if } |j-k| \leq 3, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

$$I_R = \max(R_b - \bar{R}_i, 0) \cdot 600 \text{ pA} \tag{5.3}$$

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.



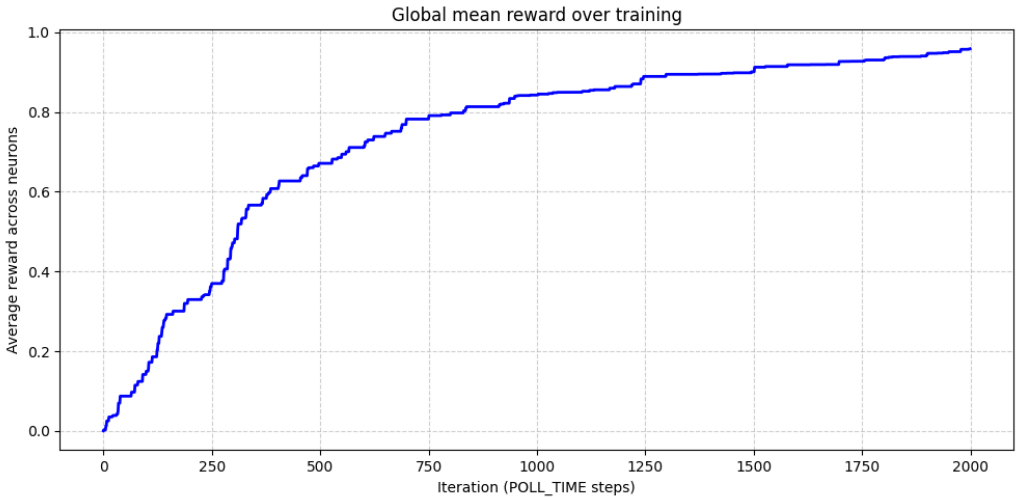
Slika 5.2: Grafična predstavitev agenta in okolja [10]



Slika 5.3: 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

Rezultati

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



Slika 5.4: 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 bo namreč naučil prehoda samo iz stanj neposredno ob cilju.

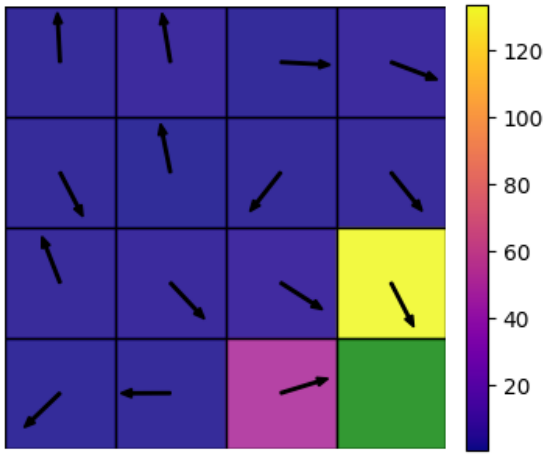
Ob učenju bomo agenta nagradili ko preide v končno stanje in ga po tem postavili v naključno stanje. Trenutno politiko agenta bomo predstavili s puščicami s smerjo, ki jo določa normaliziran vektor $\hat{x}_i$ v vsakem od stanj i, ki predstavljajo preferenco akcije glede na medsebojne razlike v utežeh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

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

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 šamozavesti"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 5.5: 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.

5.2 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 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.

Verjetno daljši uvod v TD učenje?

5.2.1 Model actor-critic

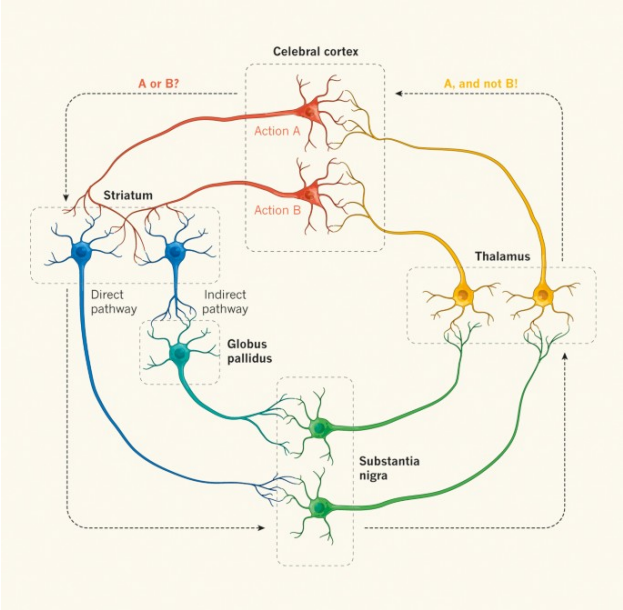
TD učenje bomo implementirali z modelom akter-kritik (*angl. actor-critic*), po zgledu Wiebke P, et al. [9] na nalogi z mrežo. Za razliko od sistema, ki ga predstavijo Wiebke P, et al. bomo za sinapso uporabili našo zakasnjeno RSTDP, ki omogoča pripisovanje odgovornosti povezavam poljubno v preteklost, 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 (*allto-all* namesto *next-neighbor* **razloži**)Sistem, ki ga bomo implementirali je tudi

manjši glede na število nevronov, za višjo hitrost simulacije, zato bomo stanja razdelili v intervale dolžine 200ms. Pri prehodu med stanji lahko določen del stimuliranih nevronov iz prejšnjega stanja postane asociiranih 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 aktivnostjo striatuma, kot je ta bila v prejšnjem stanju, zaradi zakasnitve.

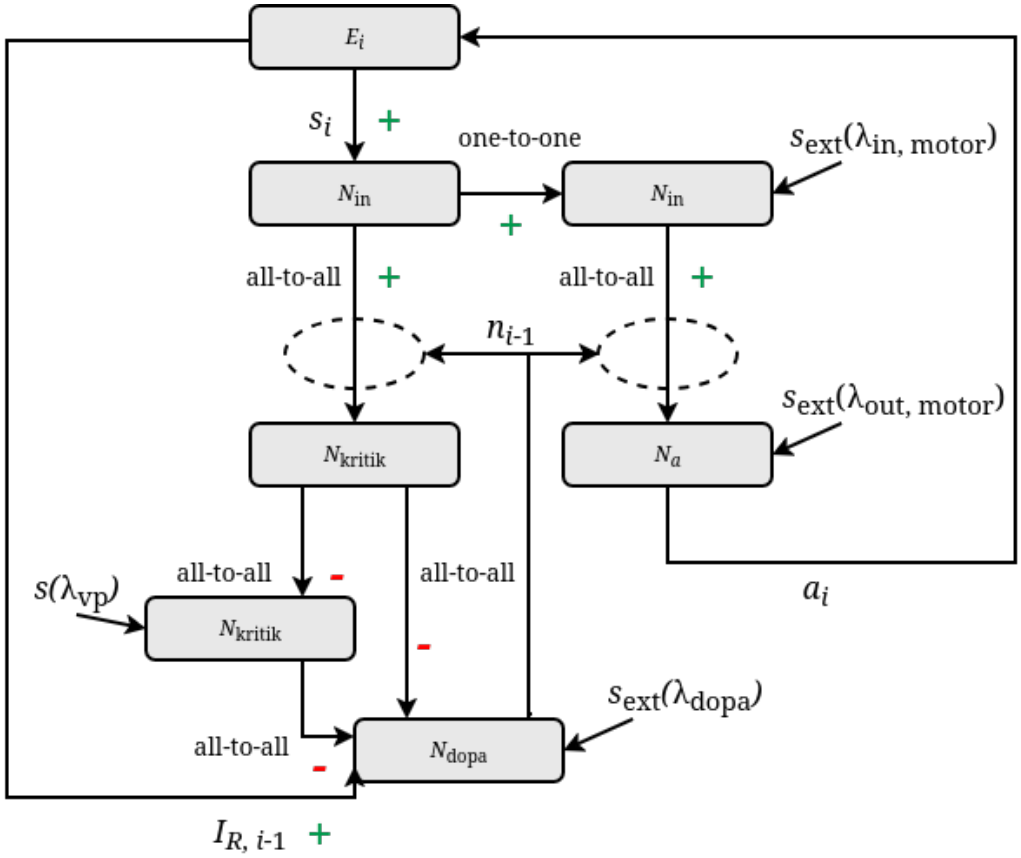
Indirektna in direktna povezava tako skupaj računata TD-napako δ_t, ki bo v trenutnem stanju glede na izračunan estimat vrednosti trenutnega stanja okrepila sinapse prejšnjega stanja, ki so izbrale aktivnost, ki nas je pripeljala v to stanje. Povprečna teža sinaps med vhodnim nevronom i in striatumom tako neposredno prestavlja 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 poissonski šum.



Slika 5.6: Direktna in indirektna pot v basalnih ganglijih

Model, kot ga predlaga Wiebke P, et al. funkcionalno združuje *substantio nigro* in *talamus* kot dopaminergični nevroni in signal do povezav med vhodom in striatumom in vhodom in izhodnimi motoričnimi nevroni. Ventralno *globusu pallidusu* se nahaja *ventral pallidum*, ki prestavlja 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 5.7: Prikaz implementiranega aktor-kritik sistema

5.2.2 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 osnovih 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 5.1: Parametri simulacije

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

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 5.2: 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 5.3: 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 5.4: Parametri sinaps med vhodom in striatumom

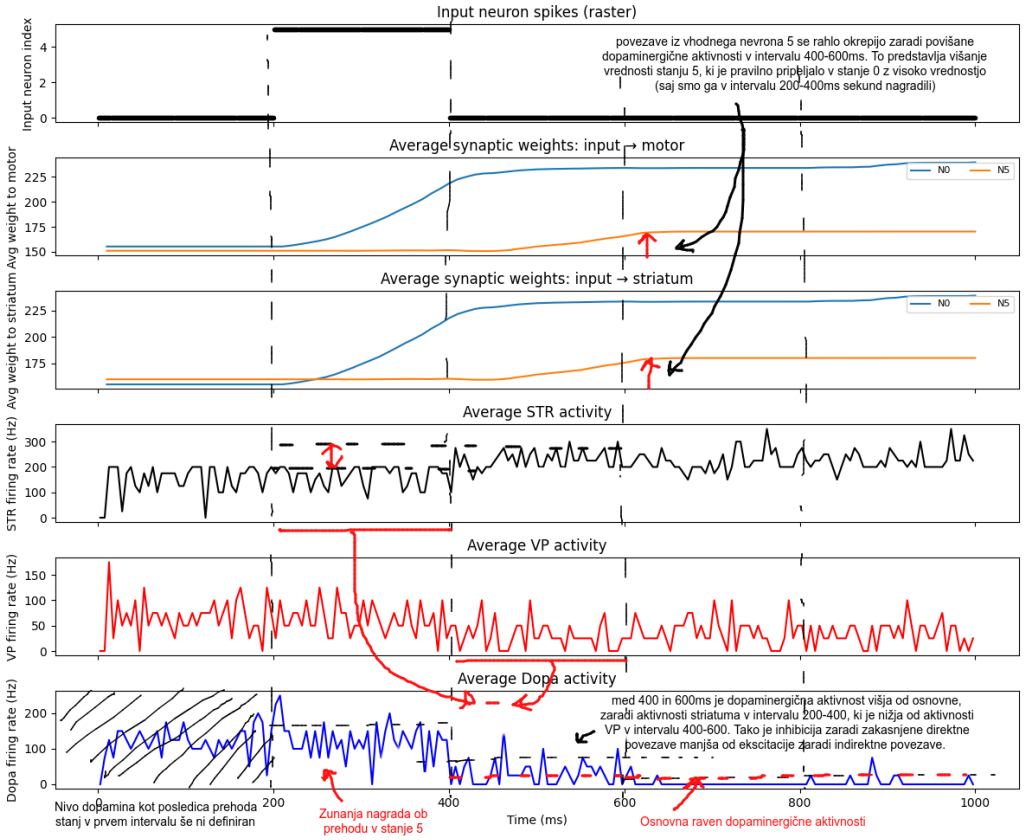
Simbol	Pomen	Vrednost
tip	Tip sinapse	Privzeta konstantna NEST sinapsa
$w_{\text{str}\to\text{vp}}$	Uteži sinps med striatumom in ventral pallidumom	-50
$w_{\text{str}\to\text{dopa}}$	Uteži sinps med striatumom in dopaminergičnimi nevroni	-55
$w_{\text{vp}\to\text{dopa}}$	Uteži sinps med ventral pallidumom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

Tabela 5.5: Parametri sinaps kritika

Simbol	Pomen	Vrednost
λ_{vp}	Povprečna hitrost šumnih impulzov nevronov ventral palliduma	5200
λ_{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

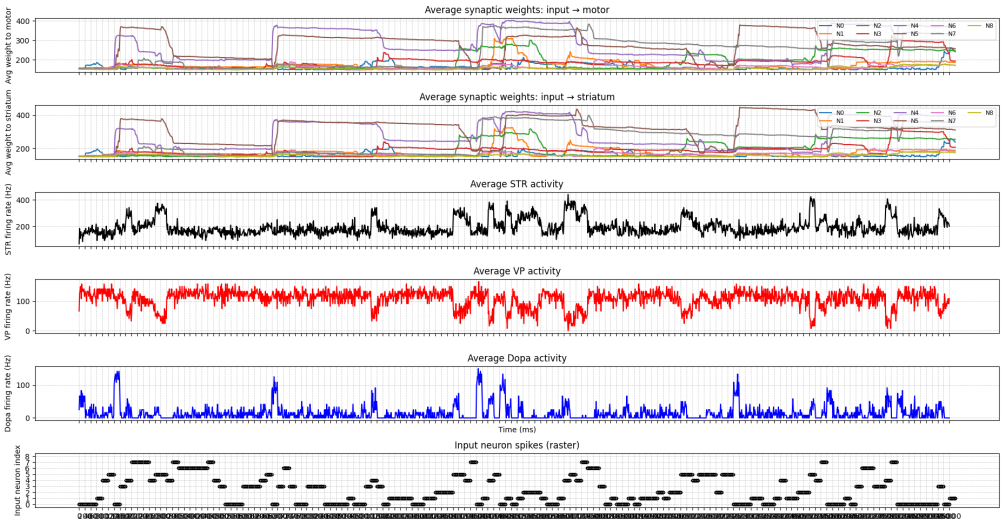
Tabela 5.6: Parametri generatorjev šuma

5.2.3 Učenje



Slika 5.8: 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...

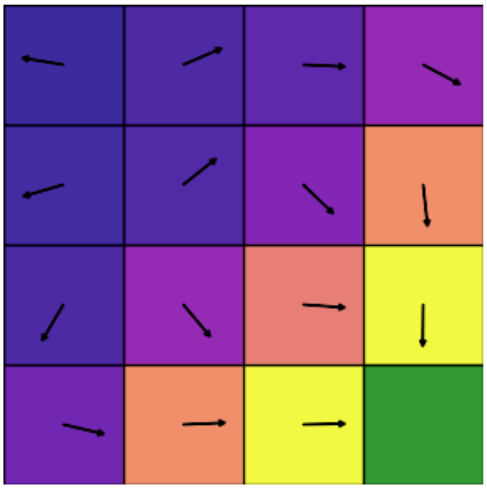


Slika 5.9: Obnašanje sistema tekom učenja na 3x3 mreži. Polja so oštevilčena od leve proti desni od zgoraj navzdol. Cilj se nahaja na polju 8. Povezave vhoda do striatuma stanj 5 in 7 so pričakovano najvišje, sledi pa jim 4, ki neposredno vodi v 5 in 7

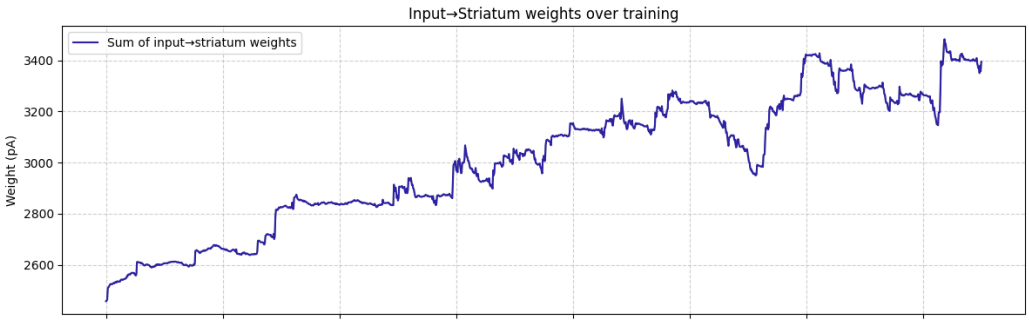
Iz zgornje kolekcije grafov izberi izseke, ki predstavljajo ključne situacije med učenjem opisane mehanizme ocenjevanja nagrade in učenja.

5.2.4 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 5.10: Rezultati učenja modela tekom 3000 iteracij po 200 ms.



Slika 5.11: 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.

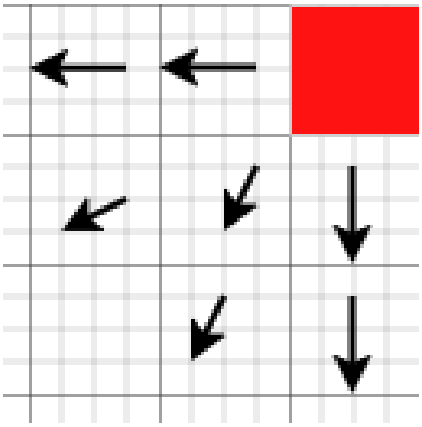
Potrebna nadaljna analiza. Glavna omejitev učenja je kosntantna rast sinaps tekom učenja. Sinapse v trenutnem modelu skozi čas same po sebi odtekajo v povprečju počasneje kot rastejo. Tako je nastavljeno, saj se pri redkih nagradah in predvsem v zgodnjih fazah učenja razlike v sinapsah zaradi nagrade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

5.3 Razširitve modela

Potencialne nadaljne raziskave in razširitve modela:

5.3.1 Izognitveno obnašanje

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 tako v enačbi sinapse RSTDP obrne predznak posodobitve in so tako sinapse, ki so odgovorne za vstop v neželeno stanje najbolj negativno posodobljene. V človeških možganih negativnega dopamina ni. Porodi se ideja, da je izogibanje negativnim stanjem prav tako posledica učenja, kjer je nivo dopamina > 0. Dopamin namreč predstavlja učenje, ne nujno nagrade. Negativna nagrada je predstavljena s posebnim vhodom, ki prestavlja abstraktno "bolečino", ki jo tekom učenja želimo zmanjšati. Tako lahko prav tako uporabimo načela STDP in TD učenja, kjer zmanjšanje nivoja bolečine prestavlja nagrado. Trenutnemu akter-kritik sistemu bi dodali še eno kopijo kritika, ki računa temporalno razliko nivoja bolečine in deluje na dopaminergične nevrone, ki so skupni obema kritikoma. Ob prehodu iz stanja z visokim nivojem bolečine v stanje z nizkim dopaminergične nevrone ekscitiramo, v obratnem primeru inhibiramo, v primeru enakega nivoja dovedene bolečine pa kritik negativne nagrade ne vpliva na dopaminergične nevrone. Sistem se v tem primeru obnaša kratkovidno, kljub računanju temporalne razlike. Nagrajene bodo samo povezave, ki so nas vodile stran od bolečine, ker pa je vhod bolčine eksteren, za razliko od striatuma, ki nagrado napoveduje sam, bodo nagrajene povezave samo v stanja neposredno ob negativnem stanju. Če želimo okrog negativnega stanja negativno označiti tudi stanja, ki nas potencialno vodijo vanj, pa moramo v sistem dodati skupino nevronov, ki stanja asociirajo z bolečinskim vhodom in jo tako napovedujejo. Pričakujemo, da tako oba kritika med seboj tekmujeta za nagrajevanje akcij, ki vodijo bližje nagradi in sinaps, ki vodijo stran od negativnega stanja.



Slika 5.12: Pričakovana politika ob kritiku negativnih stanj (brez kritika nagrajenih stanj)

5.3.2 Rekurenčne povezave

Velika predpostavka trenutnega sistema je ta, da rekurenčnih povezav ni. Tako so stanja časovno med seboj skoraj popolnoma neodvisna (med prehodi stanj se sinapse še vedno lahko križno asociirajo, zaradi česar smo v našem modelu uvedli 50ms pavzo stimulacije pred prehodom v naslednje stanje. To je opisano v poglavju **TD učenje in model actor-critic**). V kolikor dodamo več vmesnih nivojev in rekurenčne povezave pa bodo stanja med seboj časovno odvisna. Pravzaprav stanja ne moremo več definirati samo z vhodno stimulacijo, saj v vsakem trenutku stanje vsebuje tudi vplive rekurenčnih povezav, ki nosijo informacijo iz stanj arbitrarno v preteklost. V primeru našega akter-kritik sistema bi tako v vsakem trenutku t kritik računal temoralno 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 je to samo miselna prilagoditev. 200ms stimulacija, ki je do zdaj predstavljala stanje, bi vseeno v tem intervalu vodila do nevronske aktivnosti, ki je v glavnem pogojena s to vhodno stimulacijo in bi tako trenutki (oziroma neskončno kratka stanja) vseeno bili v naboru trenutkov značilnih za trenutno vhodno stimulacijo.

V eksperimentih izvedenih do zdaj, 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. Rekurenčne povezave bi tako predvideno 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. Zanimivo bi bilo realna stanja neke naloge tako razviti v drevo abstraktnih stanj in označiti pričakovane nagrade in preferirane akcije, ki jih sistem napoveduje. Sistem je še vedno stimuliran glede na realna stanja naloge, vendar bi s pomočjo rekurenčnih povezav interno predstavljal nabor abstraktnih stanj.
Rekurenčne povezave privedejo tudi do določenih situacij, kjer bi bila potrebna redefinicija trenutnega načina izbire stanj. Dva izhodna nevrona sta namreč lahko povezana med sabo in se bosta vedno prožila skupaj. V tem primeru moramo v sistemu dopuščati izbiro večih akcij hkrati in takšno situacijo na primer "kaznovati".

Poglavje 6

Diskusija in zaključek

To-Do...

Članki v revijah

[2] Dobrevski M, Skočaj D. "Deep reinforcement learning for map-less goal-driven robot navigation". V: *International Journal of Advanced Robotic Systems. 2021* 18.1 (2021). DOI: 10.1177/1729881421992621.

[3] Izhikevich, E. M. "Solving the distal reward problem through linkage of STDP and dopamine signaling". V: *Cerebral cortex (New York, N.Y. : 1991)* 17.10 (2007). DOI: 10.1093/cercor/bhl152.

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

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

Celotna literatura

[1] Deleva A. "TD learning in Monte Carlo tree search : masters thesis". Magistrska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2015.

[2] Dobrevski M, Skočaj D. "Deep reinforcement learning for map-less goal-driven robot navigation". V: *International Journal of Advanced Robotic Systems. 2021* 18.1 (2021). DOI: 10.1177/1729881421992621.

[3] Izhikevich, E. M. "Solving the distal reward problem through linkage of STDP and dopamine signaling". V: *Cerebral cortex (New York, N.Y. : 1991)* 17.10 (2007). DOI: 10.1093/cercor/bhl152.

[4] *Klasično pogojevanje*. URL: https://sl.wikipedia.org/wiki/Klasi%C4%8Dno_pogojevanje (pridobljeno 20. 3. 2025).

[5] Šutar M. "Uporaba predvidevanja akcij nasprotnika pri učenju inteligentnega agenta". Diplomska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2023.

[6] Svete A. "Posplošitev problema vozička s palico na zahtevnejše domene". Diplomska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2020.

[7] Štromajer T. "Using machine learning to train a shepherd dog". Magistrska naloga. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2022.

[8] *Temporal difference learning*. URL: https://en.wikipedia.org/wiki/Temporal_difference_learning (pridobljeno 20. 3. 2025).

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

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