VILNIUS TECH
Vilniaus Gedimino technikos universitetas

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Matematinės statistikos katedra

Paulius Janėnas

SKATINAMUOJU MODELIU SUKURTI ŽAIDIMŲ AGENTAI

Game agents created by the incentive model

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 6211AX009
Duomenų mokslo specializacija
Statistikos studijų kryptis

Vilnius, 2022

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Matematinės statistikos katedra

TVIRTINU
Katedros vedėjas

_____
(Parašas)

Doc. Dr. Kazimieras Padvelskis
(Vardas, pavardė)

_____
(Data)

Paulius Janėnas

SKATINAMUOJU MODELIU SUKURTI ŽAIDIMŲ AGENTAI

Game agents created by the incentive model

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 6211AX009
Duomenų mokslo specializacija
Statistikos studijų kryptis

Vadovas	Dr. Doc. Tomas Rekašius (Pedagoginis vardas, vardas, pavardė)	_____ (Parašas)	_____ (Data)
Lietuvių kalbos konsultantė	Dr. Vaida Buivydienė (Pedagoginis vardas, vardas, pavardė)	_____ (Parašas)	_____ (Data)

Vilnius, 2022

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Matematinės statistikos katedra

ISBN ISSN
Egz. sk. .....
Data .....-.....-.....

Antrosios pakopos studijų **Duomenų mokslo ir statistikos** programos magistro baigiamasis darbas

Pavadinimas	**Skatinamuoju modeliu sukurti žaidimų agentai**
Autorius	**Paulius Janėnas**
Vadovas	**Tomas Rekašius**

Kalba: lietuvių

Anotacija

Baigiamajame magistro darbe įgyvendintas mašininio mokymosi modelis TD3, kuris padeda voro agentui judinti fizines galūnes bei judėti link tikslo. Analitinėje dalyje aptariamos skatinamojo mokslo pagrindinės funkcijos ir pirminiai skatinamieji modeliai, neuroniniai tinklai bei kaip jais aproksimuojami šie skatinamieji metodai: gilusis-Q, aktorius-kritikas, DPG, DDPG bei TD3. Praktinėje dalyje pateikiamas sukurtas voro ropojimo algoritmą, kuris buvo įgyvendintas „Unity“ bei „Python“ aplinkoje, naudojantis TD3 modeliu. Apmokytas voras fiziškai judindamas aštuonias galūnes nuropodavo link pažymėtos koordinatės. Išvadose aptariami gauti praktinės dalies sėkmingi rezultatai.

Darbą sudaro 7 dalys: įvadas, teorinė dalis, praktinė dalis, išvados, literatūros sąrašas, priedai, terminų žodynas. Darbo apimtis – 53 p. teksto be priedų, 22 iliustr., 10 bibliografiniai šaltiniai, 6 terminai.

Prasminiai žodžiai: Skatinamasis mokslas, TD3, gilusis-Q, DDPG, neuroniniai tinklai, fizikinis galūnių judinimas.

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Mathematical Statistics

ISBN ISSN
Copies No. .....
Date .....-.....-.....

Master Degree Studies **Data Science and Statistics** study programme Master Graduation Thesis

Title	**Game agents created by the incentive model**
Author	**Paulius Janėnas**
Academic supervisor	**Tomas Rekašius**

Thesis language: Lithuanian

Annotation

The final master thesis implemented a machine learning model, TD3, which helps the spider agent to move its physical limbs towards a goal. The analytical part of the thesis discusses the basic functions of reinforcement learning and the primary reinforcement learning models, neural networks and how they are used to approximate the following reinforcement learning methods: deep-Q, actor-critic, DPG, DDPG and TD3. In the practical part, I present an implementation of the spider crawling algorithm, which was implemented in unity and python environments using the TD3 model. The trained spider physically moved its eight limbs and crawled towards a marked coordinate. The successful results obtained in the practical part are discussed in the conclusions.

The thesis consists of 7 parts: introduction, theoretical part, practical part, conclusions, list of references, appendices, glossary. The thesis consists of 53 pages of text without appendices, 22 illustrations and 10 bibliographical references, 6 terms.

Keywords: Reinforcement learning, TD3, deep-Q, DDPG, neural networks, physical limb movement.

Turinys

Iliustracijų sąrašas **5**

Įvadas **6**

1. Teorinė dalis **8**

1.1. Neuroninio tinklo struktūra 8
1.1.1. Neuroninio tinklo apmokymas naudojantis tiesioginio sklidimo algoritmą 10
1.1.2. Neuroninio tinklo atgalinio sklidimo algoritmas 11
1.2. Neuroninių tinklų klasifikatorių tipai 12
1.2.1. Binarinis klasifikatorius 12
1.2.2. Daugiau nei viena klasė 13
1.2.3. Tiesinės regresijos atsakas 13
1.2.4. Tiesinė regresija su daugybe atsako kintamųjų 14
1.3. Neuroninių tinklų reguliarizacija 15
1.3.1. L1 bei L2 reguliarizacija 15
1.3.2. Atvirkštinis išmetimas 15
1.3.3. Duomenų augmentacija 15
1.3.4. Įvesties normalizacija 16
1.4. Neuroninių tinklų parametrų inicializacija 16
1.4.1. Gradiento sprogimas bei nykimas, aktyvacinių funkcijų parametrų priskyrimas 16
1.5. Mokymosi greičio didinimas 17
1.5.1. Mažos mokymosi imties metodas 17
1.5.2. Stochastinis nuolydis 17
1.6. Gradientinio nuolydžio metodo optimizavimas 17
1.6.1. Momentinis gradientinis nuolydis bei eksponentiškai pasvertų svorių vidurkis 18
1.6.2. RMSProp 19
1.6.3. Adam 19
1.7. Skatinamasis mokymasis 19
1.7.1. Skatinamojo mokslo pagrindiniai kintamieji, stacionariojo pasiskirstymo problema 21
1.7.2. Ne stacionarus apdovanojimo skirstinys 24
1.7.3. Markovo grandinės 25
1.7.4. Apdovanojimai bei tikslai 26
1.7.5. Vertės bei veiksmų taisyklių funkcijos 27
1.7.6. Verčių bei veiksmų taisyklių iteravimas 29

1.7.7. Dinaminis programavimas . . . 30
1.7.8. Monte Carlo metodai . . . 33
1.7.9. TD bei Q-mokymasis . . . 34
1.7.10. Gilusis skatinamasis mokslas, gilusis-Q . . . 35
1.7.11. Veiksmo taisyklių gradientas bei aktorius-kritikas . . . 38
1.7.12. DPG bei DDPG . . . 40
1.7.13. TD3 . . . 41

2. Praktinė dalis **43**
2.1. Agentas, jo tikslai bei aplinka . . . 43
2.2. Mokymosi seka . . . 44
2.3. Aplinka . . . 44
2.4. Duomenys . . . 45
2.4.1. Įvestis - būsenos . . . 46
2.4.2. Apdovanojimas . . . 47
2.4.3. Išvestis - veiksmai . . . 48
2.5. Komunikacijos sluoksnis . . . 48
2.6. Mokymosi procesas . . . 49
2.7. TD3 modelio įgyvendinimas . . . 50
2.8. Rezultatai . . . 51

3. Išvados **53**

Literatūra **54**

A. Priedai **55**

Iliustracijų sąrašas

1. Gilaus neuroninio tinklo bendra struktūra. 9
2. Gradientinio nuolydžio pavyzdys. 10
3. Priekinio sklidimo vieno neuroninio sluoksnio apskaičiavimo grafas. 11
4. Atgalinio sklidimo vieno neuroninio sluoksnio apskaičiavimo grafas 12
5. Sukonstruota stalo teniso simuliacija 14
6. Reguliarizacija - kairėje L1, dešinėje L2 16
7. Gradientinė optimizacija su ir be momentinio greičio. SGD momentinis bei Adam konverguoja į globalų minimumą, kiti konvergavo į lokalų minimumą . 18
8. Bendras skatinamojo mokslo modelis 20
9. Apdovanojimo pavyzdys, pagal voro greičio (raudona rodyklė) bei norimo judėjimo (žalia rodyklė) vektorių atitikimu. 21
10. N lošimo aparatų, su skirtingais vidutiniais laimėjimais 22
11. Dulkio siurblio Markovo grandinė 25
12. Vienas agento epizodas 26
13. Veiksmų taisyklių pasiskirstymas, kur $\epsilon = 5$. 28
14. Viena optimizacijos iteracija 30
15. Dinaminio programavimo pavyzdys. 32
16. Giliojo Q neuroninio tinklo struktūra 36
17. TD3 modelio struktūra 42
18. Voro aplinka 45
19. Voro galūnės. 46
20. Viršutinio sąnario judėjimo trajektorijos. 48
21. Serverio bei agento komunikavimo sluoksnis 50
22. Apdovanojimų grafikas. 51

Įvadas

Šių laikų žaidimų industrijoje sukurti žaidimo agentą, kuris realistiškai reaguoja į aplinką, yra sudėtinga užduotis, reikalaujanti daug laiko bei pastangų. Taip pat neretai prie laiko kaštų prisideda ir tai, kad prieš kuriant žaidimo agentą tam reikia tinkamai paruošti aplinką – sukurti pagalbinius įrankius, leidžiančius padaryti programavimo procesą lengvesnį. Tačiau net įdedant pastangas programuojant agentus, rezultatai yra dažnai prasti, jie nėra pakankamai protingi, žaidėjui nėra didžiulio iššūkio juos įveikti. Tai priveda prie didžiausio žaidėjų nepasitenkinimo – agentų sukčiavimo. Norint agentus padaryti protingesniais, jiems suteikiama žaidimo informacija, kuri yra neprieinama įprastiems žaidėjams.

Per pastarąjį dešimtmetį įvykus neuroninių tinklų proveržiui buvo dedama daug vilčių, kad jie sugebės sukurti protingesnius žaidimo agentus. Tačiau neuroniai tinklai tėra funkcijos aproksimacijos metodas ir sunkiausia šio iššūkio dalis greitai atsiskleidė – kaip sukurti funkciją, kuri padėtų agentams tobulėti, tapti protingesniems? Šis klausimas buvo nagrinėtas per pastaruosius 70 metų, įtraukiant tokias mokslo šakas kaip dinaminis programavimas, Markovo modeliai, neuromokslai. Atradimai minėtose šakose privedė prie skatinamojo mokslo gimimo, kuomet 1987 metais šio mokslo pradininkai Sutton, R. bei Barto, A. moksliniame straipsnyje [1] pirmieji sukūrė optimizavimo funkciją, pavadinta TD, padedančia agentams mokytis iš praeities rezultatų, siekiant pasirinkti tokius ateities veiksmus, kurie pagerintų ateities rezultatą. Funkcijos principas remiasi skatinimo metodais, kaip siunčiant paskatinimo impulsus agentui, jei jis atliko teisinga veiksmą, bei nubaudžiamas jei atliko neteisingą veiksmą. Tokiu principu agentas sukuria pozityvias bei negatyvias asociacijas su galimais veiksmais, kuriuos jis gali atlikti. Šį principą puikiai iliustruoja Pavlovo šunų eksperimentas [2]. Sekančius dešimtmečius sekė šios funkcijos gerinimas, optimizavimas, bandymas išspręsti didelės dimensijos problemą – kuo didesnė būsenos ervės dimensija, tuo ilgiau užtrunka skaičiavimai. Ši problema buvo išspręsta pasitelkiant neuroninius tinklus, kurie aproksimuodavo TD funkciją. Šio metodo kulminacija įvyko 2016 metais, kuomet Google AlphaGo sugebėjo įveikti stalo žaidimą Go, kurio nesugebėjo įveikti nei viena komanda prieš tai dirbusi prie pirmųjų algoritmų, įveikusių šachmatų čempionus. Šiuo metu skatinamieji metodai aprėpia daugybe sričių: automobilių autonominis vairavimas, logistikos grandžių optimizavimas, robotikos judėjimo problemos ir dar daugybe kitų sričių.

Šiame darbe bus išanalizuojamas skatinamasis mokymasis bei įgyvendintas vienas iš šio mokslo metodų - TD3. Teorinėje dalyje bus analizuojama šio metodo veikimo principai. Tai atlikus praktinėje dalyje šis metodas bus pritaikytas išspręsti roboto-voro fizinio judėjimo aplinkoje problemą.

Darbo aktualumas. Dirbtiniui intelektui sugebėjus įveikti šachmatų čempionus buvo manoma, kad žmonių dominavimas tradiciniuose stalo žaidimuose pasibaigė. Tačiau programuotojų ambicijos išblėso, kuomet min-max paremti šachmatų algoritmai nesugebėjo įveikti

kiniečių tradicinio stalo žaidimo – Go. To pagrindinė priežastis – neapskaičiuojamai didelė žaidimo būsenų erdvė. Kompiuteriai užtrukdavo per ilgai planuoti ėjimus ir jam neužtekdavo atminties išanalizuoti visus galimus ėjimus. Todėl iki šiol šis stalo žaidimas buvo neįveikiamas iki kol Google korporacija išleido skatinamojo mokymosi paremto algoritmo, kuris 2016 metais sugebėjo įveikti šių laikų geriausią Go žaidėją. Nuo to laiko skatinamųjų mokymosi metodų aktualumas vis labiau plito įtraukiant sritis kaip autonominiai automobiliai, robotika, finansai bei vis naujai populiarėjanti sritis – žaidimų industrija.

Darbo tikslas. Pritaikyti TD3 skatinamojo mokymosi modelį, kuris apmokytų simuliuojamą vorą vaikščioti 3D aplinkoje.

Uždaviniai.

1. Išanalizuoti gilių neuroninių tinklų struktūrą.
2. Įgyvendinti „Python“ bei „Unity“ duomenų sinchronizaciją.
3. Išanalizuoti skatinamojo mokymosi modelio principus.
4. Išanalizuoti Belmano, Q-mokymosi, TD funkcijas.
5. Sukurti „Unity“ voro aplinką bei agentą, kuris bus apmokomas vaikščioti.
6. Įgyvendinti TD3 mokymosi modelį bei apmokyti vorą vaikščioti sukurtoje aplinkoje.

1. Teorinė dalis

Skatinamojo mokymo metodai remiasi neuroninių tinklų aproksimacijom, kurie bus nagrinėjami tolimesniuose skyriuose. Taigi šioje darbo dalyje nagrinėsiu giliųjų neuroninių tinklų veikimo principus. Tai atlikęs toliau nagrinėsiu skatinamojo mokymosi metodus bei kaip neuroniniai tinklai padeda juos įgyvendinti. Pagrindinis nagrinėjamas skatinamojo mokymosi modelis bus TD3. Tai yra modelis, kuris padeda agentui tolydžioje aplinkoje pasirinkti veiksmus, kurie suteikia didžiausią apdovanojimą.

1.1. Neuroninio tinklo struktūra

Neuroninių tinklų sandarą galimą žiūrėti kaip į logistinės regresijos logistinių funkcijų veikimą. Kitaip tariant vieną neuroną galima traktuoti kaip logistinę funkciją, kuri išsiskaido į dvi dalis: tiesinę bei aktyvacinę. Tiesinėje dalyje suskaičiuojame tiesinę funkciją ir ją įvedame į aktyvacijos funkciją, kuri logistinėje regresijoje yra sigmoido funkcija. Tai yra identiška logistinei regresijai. Aktyvacinės funkcijos gali būti skirtingos, nebūtinai sigmoido. Pagrindinė skirtingų aktyvacinių funkcijų priežastis yra spartesnė svorių konvergacija į optimalią reikšmę. Dažnai naudojama Relu aktyvacinė funkcija 1, kuri neturi sigmoido nykstančio gradiento problemos, aptariamos 1.4.1 skyriuje. Ši problema pasireiškia kuomet sigmoido funkcijos reikšmė yra labai didelė arba maža, tuomet gradientas tampa beveik nulinis ir mokymosi procesas sustoja.

$$f(x) = x = max(0,x) \tag{1}$$

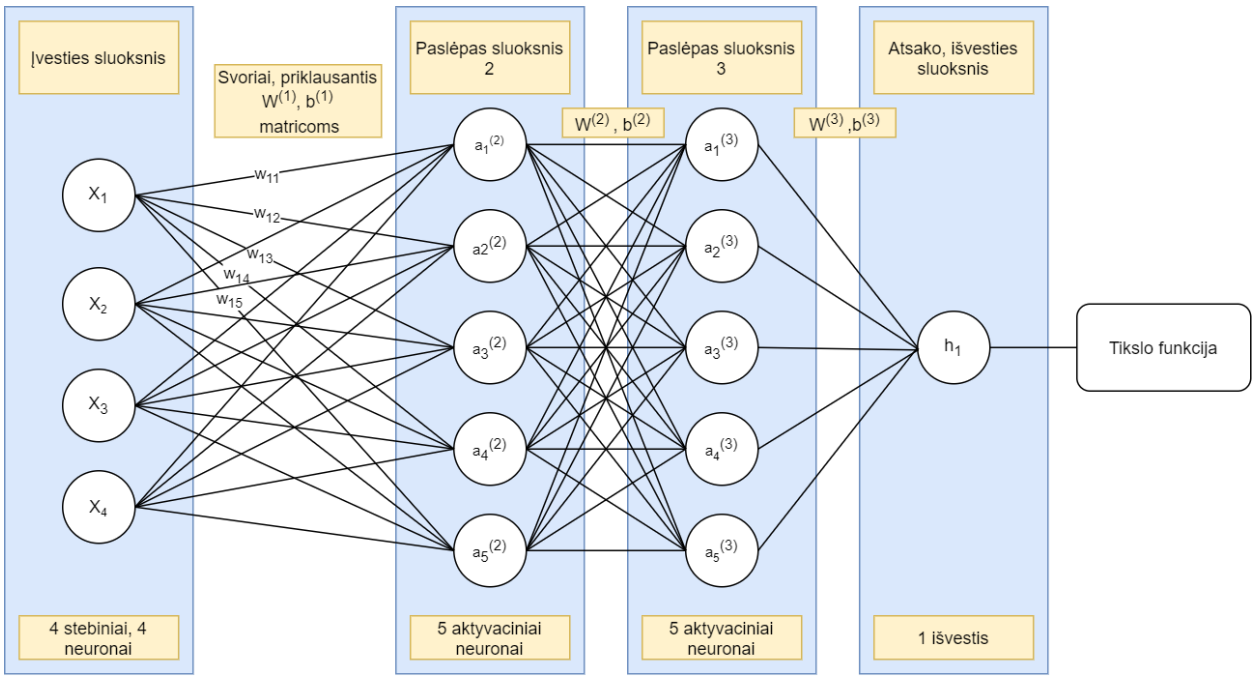
Pats neuroninis tinklas susideda iš logistinių funkcijų (jei naudojame ne sigmoid aktyvacinę funkciją tai jau nebėra logistinė funkcija) sluoksnių, kurie prasideda įvesties sluoksniu, viduryje paslėptuoju sluoksniu ir galiausiai išvesties sluoksniu. Visi klasifikatoriai turi vieną bendrą panašumą – įvesties bei paslėptuosius sluoksnius. Tačiau priklausomai nuo to ar atsakas yra klasifikacinis ar regresinis, kinta išvesties sluoksnis.

Kuriant „Python“ aplinkoje neuroninius tinklus buvo pasitelkta vektorizacija, kuomet neuroninio tinklo įvestis nėra vienas stebinys, bet stebinių matrica. Visas įvestis, svorius, išvestis išsireiškus matriciniu pavidalu gaunamas ryškus neuroninių tinklų paspartėjimas. Neuroninio tinklo bendra naudota struktūra yra pateikta 1 paveikslėlyje.

Vieno sluoksnio įvesties transformacijos vektorizaciją yra taip skaičiuojama:

$$Z^l = W^{l^T} \cdot A^{l-1} + b^l$$

Žymėjimas: A^l yra l sluoksnio neuronų reikšmės (matrica, kuri susideda iš $a^l \in \mathcal{R}^n$ elementų), kurias aktyvacinės funkcijos pateike (jei l yra lygus įvesties sluoksniui, tai A^l tampa duomenų įvesties vektorius X), o W^{l^T} yra l sluoksnio svoriai. Šios transformacijos matricos



1 pav. Gilaus neuroninio tinklo bendra struktūra.

eilutės yra skirtingų neuronų tiesinės transformacijos reikšmės, o stulpeliai žymi skirtingus mokymosi duomenis. Šią transformaciją įkėlus į aktyvacijos funkciją gauname galutinę perceptrono reikšmę. Norint, kad neuroninis tinklas būtų tikslus, mes turime optimizuoti svorius. Tam pasitelkiame gradientinį nuolydį. Gradientinis nuolydis padeda surasti svorius, su kuriais tikslo funkcija įgyja mažiausią paklaidą. Tikslo funkcija sudaro tolygų paviršių, kurio skirtingos ašys yra svorių reikšmės bei viena iš jų yra tikslo funkcijos rezultatas. Pateiktame 2 paveikslėlyje apatinės ašis x bei y galima traktuoti kaip svorius, o viršutinę ašį kaip tikslo funkcijos paklaidą.

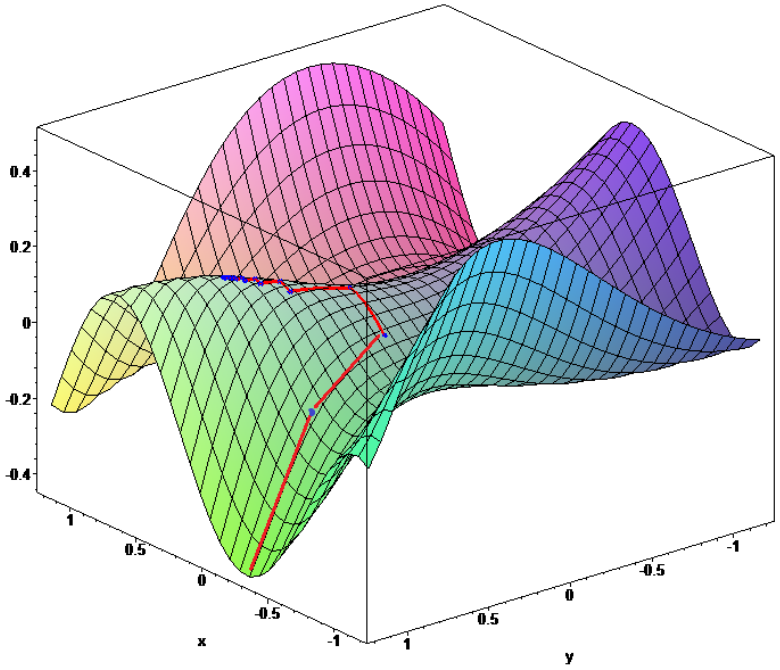
Judindami svorius priešingą gradiento puse, mes judame ties tokiais svoriais, kurie mažina tikslo funkcijos paklaidą. Šią trajektoriją atspindi 2 paveikslėlyje esanti raudona kreivė, kuri kiekvienos iteracijos metu (mėlyni taškai) po truputi slenka žemyn, optimizuoja svorius.

Vėlgi gradientinio nuolydžio vektorizacija yra tokia pati tiek įvesties, tiek paslėptuose sluoksniuose, bet priklausomai nuo analizuojamos problemos - skirtinga išvesties sluoksnyje. Gradiento vektorizacija:

$$dZ^l = W^{l+1^T} dZ^{l+1} \cdot g^l\left(Z^l\right). \tag{2}$$

Žymėjimas: $g^l(Z^l)$ yra dabartinio sluoksnio aktyvacinės funkcijos išvestinė. Ji yra sudauginama su visais matricos nariais. Gavus perceptronų išvestines dZ^l galime toliau gauti svorių gradientus:

$$\begin{aligned} dW^l &= \frac{1}{m} dZ^l A^{(l-1)^T} \\ db^l &= \frac{1}{m}\left(dZ^l \text{ matricos stulpeliu susumavimas }\right) \end{aligned} \tag{3}$$

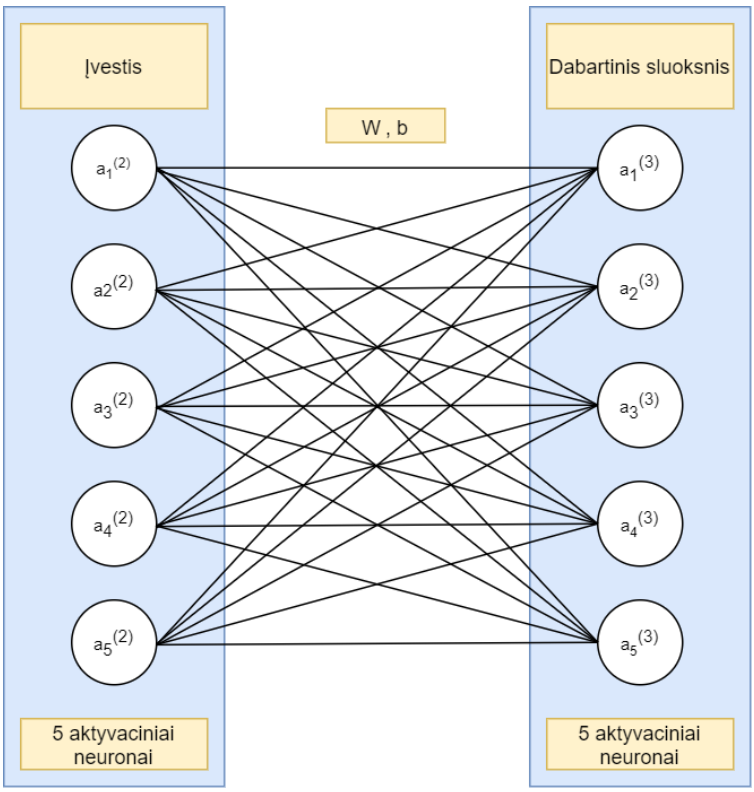


2 pav. Gradientinio nuolydžio pavyzdys.

Žymėjimas: stulpelių skaičius arba mokymosi duomenų kiekis yra žymimas m. Toliau bus aptariamos priekinė, atgalinė sklaida bei tikslo funkcijos, kurios sudaromos atitinkamai pagal turima klasifikacijos problemą.

1.1.1. Neuroninio tinklo apmokymas naudojantis tiesioginio sklidimo algoritmą

Įgyvendinti gilaus neuroninio tinklo struktūra pasirinktoje programavimo kalboje nėra sunku, kuomet pastebimas visiems sluoksniams bendras pasikartojantis skaičiavimo bruožas. Skaičiuojant tiesioginį sklidimą (judame neuroniniu tinklu iš kairės į dešinė) į neuroninį tinklą galima žiūrėti kaip į atskirus blokus susidedančius iš įvesties bei išvesties, kaip tai matosi 3 paveikslėlyje.

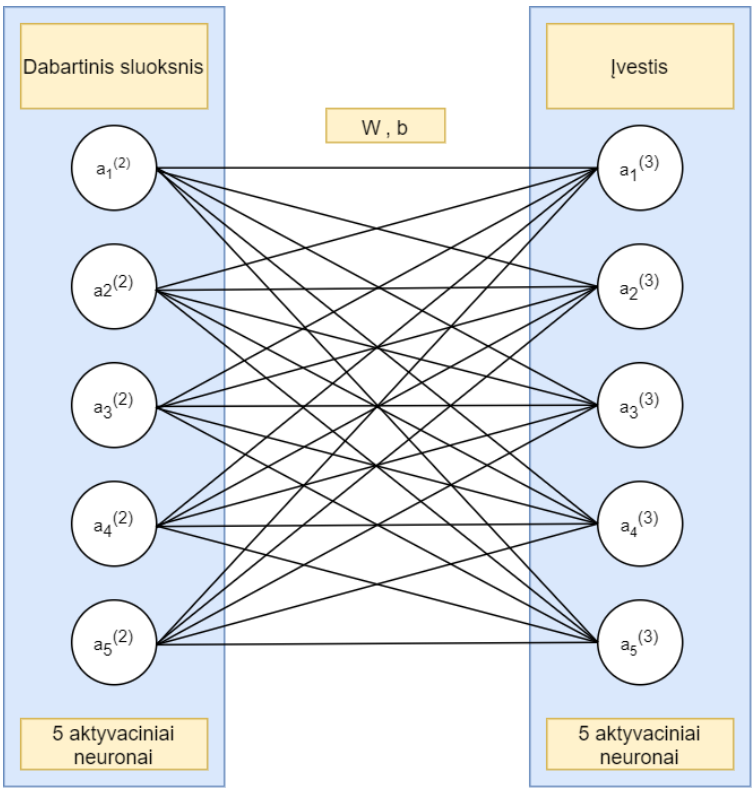


3 pav. Priekinio sklidimo vieno neuroninio sluoksnio apskaičiavimo grafas.

Tokia formą išskaidžius neuroninį tinklą, jį tampa labai paprasta apskaičiuoti, tereikia iteruoti kiekvienu sluoksniu iš kairės į dešinę. Kiekvienos iteracijos metu praeitą sluoksnį traktuojame kaip įvesties sluoksnį ir naudojantis 3 formulėmis apskaičiuojame dabartinio sluoksnio aktyvacinius neuronus. Šios operacijos metu svarbu išsaugoti tarpinius skaičiavimus, kaip tiesinės funkcijos bei aktyvacinę funkcijos apskaičiavimo rezultatą. Jos bus reikalingos apskaičiuoti gradientinį nuolydį, kaip tai matosi 2 bei 3 formulėse. Paskutinės iteracijos metu, atlikus aktyvacinės funkcijos apskaičiavimą, lieka tik apskaičiuoti tikslo funkcijos rezultatą. Šį struktūra pasižymi formulių universalumu, jos tinka neuroniniams tinklams, kurie yra negilieji bei gilieji, kitaip tariant - nepriklauso nuo sluoksnių kiekio. Atlikus tiesioginį sklidimą ir suskaičiavus tikslo funkcijos klaidą, galime pradėti ją mažinti, judėdami priešinga gradiento linkme. Tai atliekama naudojant atgalinio sklidimo metodą.

1.1.2. Neuroninio tinklo atgalinio sklidimo algoritmas

Atgalinės sklaidos struktūra yra vos ne identiška priekiniai sklaidai. Pagrindiniai skirtumai yra sukeistos vietos apskaičiuojamo sluoksnio bei įvesties sluoksnio. Tačiau kaip ir priekinės propogacijos atveju, mes atliekame panašias iteracijas, tik šį kart pradedam nuo tikslo funkcijos ir judame link pirmo sluoksnio, iš kairės į dešinę. Atliekant šias svorių atnaujinimo operacijas, kaip tai matosi 2 bei 3 formulėse, pasitelkiame priekinės sklaidos metu išsaugotais duomenimis.



4 pav. Atgalinio sklidimo vieno neuroninio sluoksnio apskaičiavimo grafas

1.2. Neuroninių tinklų klasifikatorių tipai

1.2.1. Binarinis klasifikatorius

Tai klasifikatorius, kuris vos ne identiškas logistinei regresijai. Klasifikuoja binarinius atsakus. Išvesties sluoksnis turi tik vieną neuroną bei jo reikšmės transformacija sutampa su logistinės regresijos logtikėtinumo funkcija:

$$\log(L) = \ell = \sum_{i=1}^{n} \left[y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right) \right]. \tag{4}$$

Šios funkcijos gradientas susiprastina į šią formą:

$$dZ^l = A^l - Y, \tag{5}$$

kur l yra paskutinis sluoksnis, Y yra tikroji atsako reikšmės, o A^l yra neuroninio tinklo gauti atsakai. Turint dZ^l galima gauti svorių gradientus bei neuronų gradientus, kaip tai buvo parodyta prieš tai buvusiuose skyriuose. Turint svorių gradientus atnaujiname svorius:

$$W^l = W^l - \alpha dW^l, \tag{6}$$

kur α nurodo gradientinio žingsnio dydį.

1.2.2. Daugiau nei viena klasė

Klasifikuojant daugiau nei vieną klasę yra naudojamas entropijos tikslo funkcija. Tikslo funkcijos forma:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i * \log(\hat{y}_i), \tag{7}$$

kur y yra tikroji reikšmė, o $\hat{y}$ yra aproksimuota.

Kuomet paskutinio sluoksnio aktyvacinė funkcija yra „soft-max“:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad kur\ i = 1,2,\ldots,K, \tag{8}$$

šios funkcijos gradientas labai gražiai susiprastina ir paskutinio sluoksnio gradientas atrodo ši taip:

$$dZ^l = (Y - \hat{Y}) \tag{9}$$

kur l yra paskutinio sluoksnio indeksas, o dydžioji Y yra vektorizacija daugybės stebinių į vieną matricą.

Iš šio gradiento galime seniau aptartais metodais gauti W bei b gradiento reikšmes. Galima naudoti ir mažiausių kvadratų metodą, tačiau gradientinis nuolydis konverguotų lėčiau. Taip pat pagrindinis skirtumas tarp binarinio klasifikatoriaus yra aktyvacinė funkcija. Šį kart tai nėra sigmoido funkcija, o „soft-max“ normalizacija, kuomet gautos eksponentinės reikšmės konvertuojamos į tikimybių pasiskirstymą, kaip tai matosi 8 formulėje. Taigi šis metodas yra dažnai naudojamas klasifikuoti reikšmes, kurios turi daugiau nei vieną klasę.

1.2.3. Tiesinės regresijos atsakas

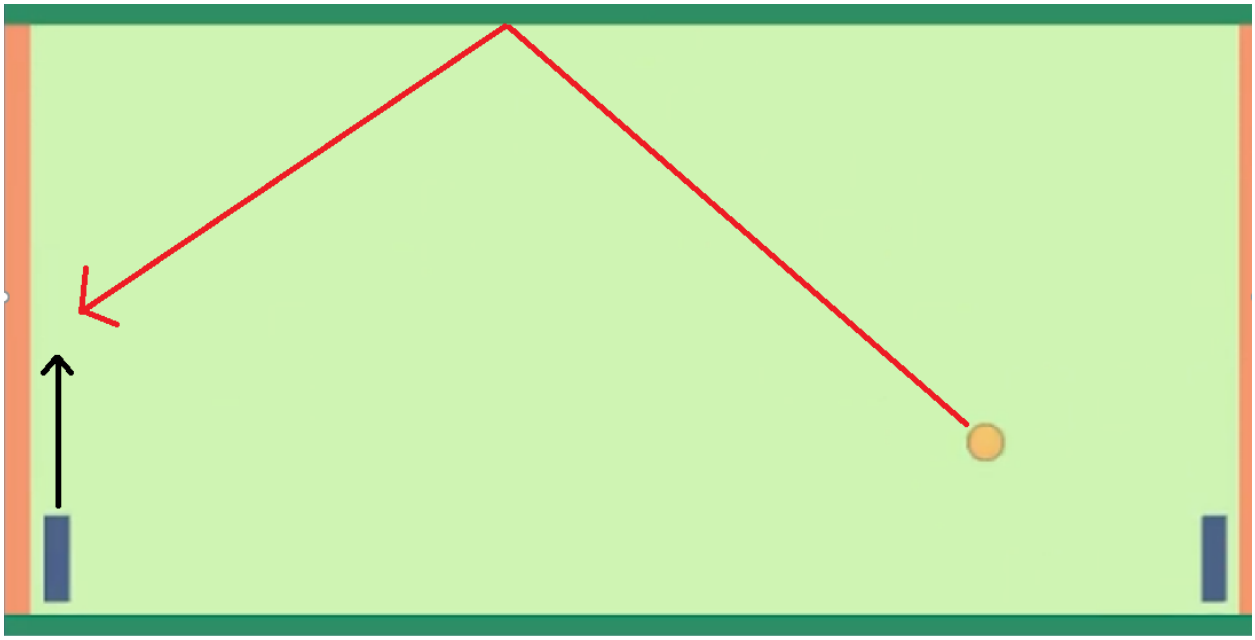
Jei atsakas yra regresinis, tuomet tikslo funkcija tampa didžiausių kvadratų optimizacija ir paskutinis neuronas neturi aktyvacijos funkcijos. Tikslo funkcijos mažiausių kvadratų sumos vidurkio formulė:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{10}$$

kur y yra tikroji reikšmė, o $\hat{y}$ yra aproksimuota.

Neuroniniai tinklai su regresine tikslo funkcija yra puikus agentų apmokymo pavyzdys, įrodantis kad priklausomai nuo užduoties, nebūtina imtis sudėtingų skatinamojo mokymosi metodų, norint sukurti protingą agentą. Su šiuo metodu buvo optimizuojamas stalo teniso žaidimas, kuriame atsakas buvo stalo teniso raketės pozicijos optimali padėtis atmušti atskriejanti kamuoliuką, kaip tai matosi 5 paveikslėlyje. Tai buvo paprastas gilus neuroninis tinklas su dvejais paslėptais sluoksniais po 12 neuronų. Buvo pasitelktas stochastinis gradientinis nuolydis, kuomet kiekvieno žaidimo kadro metu buvo nusiunčiami neuroniniam tinklui

dabartinės raketės padėtis bei kamuoliuko trajektorija. Šiam uždaviniui neuroninio tinklo nereikia, pakaktu paprasto algoritmo ar tiesinės regresijos, tačiau tai gerai iliustruoja jo veikimo principus. Pradžioje raketės juda padrikai, bet bėgant laikui, vykstant gradientiniui apmokymui, rakečių judėjimas vis tikslėja.



5 pav. Sukonstruota stalo teniso simuliacija

5 Paveikslėlyje raudona trajektoriją yra neuroninio tinklo įvestis, apskaičiuojanti kamuoliuko galutinę padėtį. Juoda trajektorija yra neuroninio tinklo išvestis, raketės greičio vektorius, simbolizuojantis pozicija, kurioje raketė turi atsirasti

1.2.4. Tiesinė regresija su daugybe atsako kintamųjų

Identiška tiesinės regresijos funkcijai, tačiau šiuo atveju paskutinis sluoksnis turi daugiau nei vieną neuroną. Tuomet tikslo funkcija atrodo šitaip:

$$L = \frac{1}{m * 2} * \sum \left(Z^l - Y\right)^2 \tag{11}$$

Vektorizacija identiška tiesinės regresijos atveju, tik atsako kintamasis Y turi ne vieną eilutę, o daugiau. Eilutės žymi pasirinkto sluoksnio neuronų reikšmes. Su šiuo metodu buvo optimizuotas automobilio kontroleris, kurio tikslas buvo apvažiuoti trasą. Įvestis buvo normalizuota automobilio sensorių informacija, kurie pateikdavo trasos barjero atstumą. Išvestis buvo automobilio stabdymo ir greičio pedalo stiprumai bei vairo pozicija. Apmokius neuroninį tinklą rezultatai buvo tragiški, automobilis važiuodavo tik tiesiai. Tačiau pritaikius Adamo metodo gradientinį nuolydį ir taip sumažinus modelio klaidą, buvo pasiekti tenkinami rezultatai. Automobilis sugebėdavo apvažiuoti trasą.

1.3. Neuroninių tinklų reguliarizacija

Neuroninių tinklų reguliarizacija reikalinga norint išvengti modelio prisitaikymo prie mokymosi duomenų. Ji padeda pagerinti testavimo duomenų tikslumą. Toliau nagrinėsiu metodus, kurie tai padeda pasiekti.

1.3.1. L1 bei L2 reguliarizacija

Didžiuliai neuroninio tinklo svoriai retai gerai generalizuoja modelį, jie numuša testavimo imties tikslumą. Todėl svarbu neleisti svoriams tapti milžiniškais. L1 bei L2 reguliarizacija tai padeda pasiekti. Prie tikslo funkcijos lygties mes pridedame Lagrandžo svorių apribojimus, L2 atveju formulė išsireiškia tokiu pavidalu:

$$J(W, b, X, y) = \frac{1}{m} \sum L(y, \hat{y}) + \frac{\lambda}{2 * m} \sum \left\| w^l \right\|^2, \tag{12}$$

kur λ yra lagrandžo daugiklis ir taip pat hiperparametras.

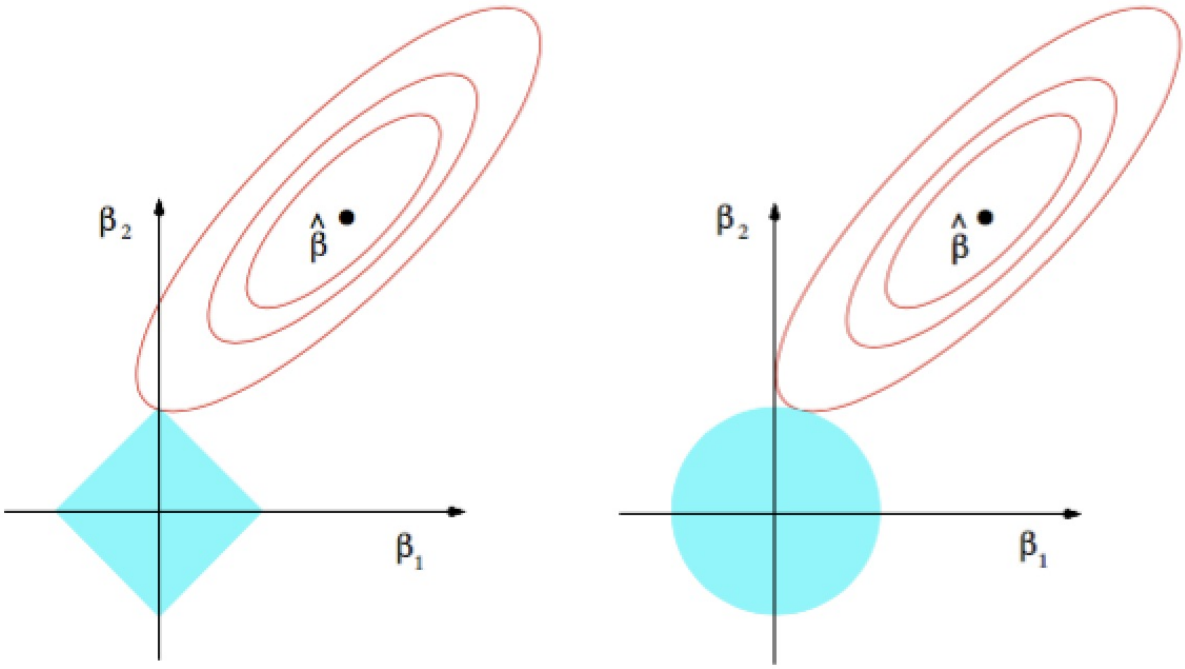
Matome, kad prie pagrindinės tikslo funkcijos prisideda papildoma bauda. Kuo didesni svoriai, tuo didesnė bauda. L1 atveju mes vietoj ilgio kvadrato, naudojame absoliutinę reikšmę. L2 skiriasi nuo L1 tuo, kad L2 padeda tuos svorius paversti nuliais, kurie turi mažai įtakos gradientiniam nuolydžiui, kaip tai matosi 6 paveikslėlyje. Tačiau abejais atvejais mes nepasieksime optimalios reikšmės ir išmainysime dispersiją į prisitaikyma (bias), taip pasiekdami tikslesnį testavimo duomenų klasifikavimą, jei modelis yra per daug prisitaikęs prie mokymo duomenų. Gradientą šiai funkcijai lengva surasti, kadangi matome, kad 12 funkcija išsiskaido į dvi dalis. Telieka surasti išvestines L2 formai, kas yra labai paprasta.

1.3.2. Atvirkštinis išmetimas

Šio metodo esmė, išjunginėti kiekvieno sluoksnio neuronus su tam tikra tikimybe. Juos išjungus, kiti neuronai turės perimti jų darbą. Tai padeda išvengti atveju, kuomet vienas neuronas persimoko. Tačiau išmetimo atveju vidutiniškai neuronų išvestis sumažėja išmetimo tikimybės procentu. Todėl atvirkštiniu išmetimu, mes po išmetimo operacijos kiekvieną neurono Z reikšmę padaliname iš išmetimo tikimybės taip padidindami neurono išvestį ir gražindami to sluoksnio tikėtiną reikšmę į prieš išmetimą buvusią.

1.3.3. Duomenų augmentacija

Šis metodas taikomas, kuomet neturime pakankamai duomenų apmokymui, ko neuroniniams tinklams reikia nepaprastai daug. Jei duomenys susideda iš nuotraukų, tuomet augmentacijos metu mes galime nuotraukas apsukti, priartinti, karpyti ir panašius metodus taikyti, norint padidinti mokymosi duomenų imti.



6 pav. Reguliarizacija - kairėje L1, dešinėje L2

1.3.4. Įvesties normalizacija

Gradientinio nuolydžio konvergacijos greitis priklauso nuo paviršiaus išsidėstymo. Paviršius gali būti labai susispaudęs bei banguotas, kas neleidžia atlikti didžiulių gradientinių žingsnių. Tokiais atvejais dideli žingsniai, gali šokinėti aplink optimalią reikšmę bei tokiu būdu niekados nekonverguoti. Taip pat svoriai nėra proporcingai panašus, vieni turi žymiai didesnę įtaka nuolydžiui, kiti mažesne. Tokiu atveju paimti dideli gradiento žingsniai gali vėlgi nekonverguoti. Tačiau normalizavus įvesties duomenis, paviršius išsilygina bei yra lengviau optimizuoti. Normalizavimas paremtas paprasčiausiu statistiniu Z normalizavimu. Jis atliekamas, kuomet įvesties duomenis pasižymi skirtingais duomenų intervalais – vienos įvesties intervalas [0,1], kitos [0,1000].

1.4. Neuroninių tinklų parametrų inicializacija

1.4.1. Gradiento sprogimas bei nykimas, aktyvacinių funkcijų parametrų priskyrimas

Priekinės sklaidos neuroninio tinklo žingsnyje, kuomet skaičiuojame išvestį, galime pastebėti tokį neuroninio tinklo bruožą

$$\hat{y} = W^l W^{l-1} W^{l-2} \dots X \tag{13}$$

Matome, kad jei W^l svorių parametrai priskiriami su didelėmis reikšmėmis, turint daugybe sluoksnių mūsų tiek neurono tiek svorių reikšmės ekponentiškai išaugs. Tuo tarpu

jei parametro reikšmės W^l priskiriamos su mažesnėms nei vieneto reikšmėm, svoriai greitai tampa nuliais. Norint to išvengti visų neurono sluoksnio aktyvacijų reikšmių vidurkis turi būti lygus nuliui bei variacija visuose sluoksniuose išlikti vienoda. Šių tikslų padeda pasiekti svorių priskyrimo metodai. Dažniausiai svoriai yra paimami iš $\mathcal{N}(0,1) * 0.001$. Šie svoriai nėra blogi jei turime tik pora sluoksnių, tačiau didėjant sluoksniams, kaip matėme prieš tai, neuronų reikšmės taps nuliais. Todėl, priklausomai nuo sluoksnio aktyvacijos, naudojame specifinius metodus priskiriant svorius. Norint kad šie metodai veiktų, įvestis privalo būti normalizuojama Z-normavimo principu.

1. Jei aktyvacijos funkcija yra $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}}$, svoriai imami iš $\mathcal{N}(0, \frac{1}{n^{l-1}})$, kur n yra neuronų skaičius sluoksnyje, šiuo atveju tai būtų praeito sluoksnio neuronų skaičius.
2. Jei aktyvacijos funkcija yra $Relu(z) = max(0, z)$, svoriai imami iš $\mathcal{N}(0,1) * \frac{2}{\sqrt{n}}$.
3. Jei aktyvacijos funkcija yra sigmoidas $\sigma(z) = \frac{1}{1+e^{-z}}$, svoriai imami iš $\frac{\mathcal{N}(n^{l-1}, n^l)}{\sqrt{n}}$.

1.5. Mokymosi greičio didinimas

Jei turime labai daug duomenų, matricų operacijos užima daug laiko. Todėl naudojame tolimesnius metodus, kurie paspartina mokymosi procesą.

1.5.1. Mažos mokymosi imties metodas

Metodo principas labai paprastas – iš visos duomenų imties paimame dalį skirstinio ir su šia gauta maža mokymosi imtimi apmokome modelį. Kiekviena tokia dalis negražins optimalaus gradiento vektoriaus, tačiau kadangi duomenis yra iš to pačio pasiskirstimo mes vis tiek judėsime į optimalias svorių reikšmes ir galiausiai konverguosime į optimalius svorius.

1.5.2. Stochastinis nuolydis

Beveik toks pats kaip praeitas metodas, tik dar ekstremalesnis atvejis – gradientas gaunamas iš vieno imties elemento, stulpelio. Vektorizacija šiuo atveju praranda visą savo greitį, taip pat gradiento vaikščiojimas neatrodo tolygus, juda vos ne į visas puses. Tačiau bendra gradientų krypties tendencija juda link tikslo funkcijos klaidos mažinimo. To priežastis yra išlieka tokia pati, kaip praeito metodo: kadangi duomuo yra iš to pačio pasiskirstimo mes vis tiek judėsime į optimalias svorių reikšmes. Šis metodas yra nuolatos taikomas skatinamuosiuose mokymosi metoduose.

1.6. Gradientinio nuolydžio metodo optimizavimas

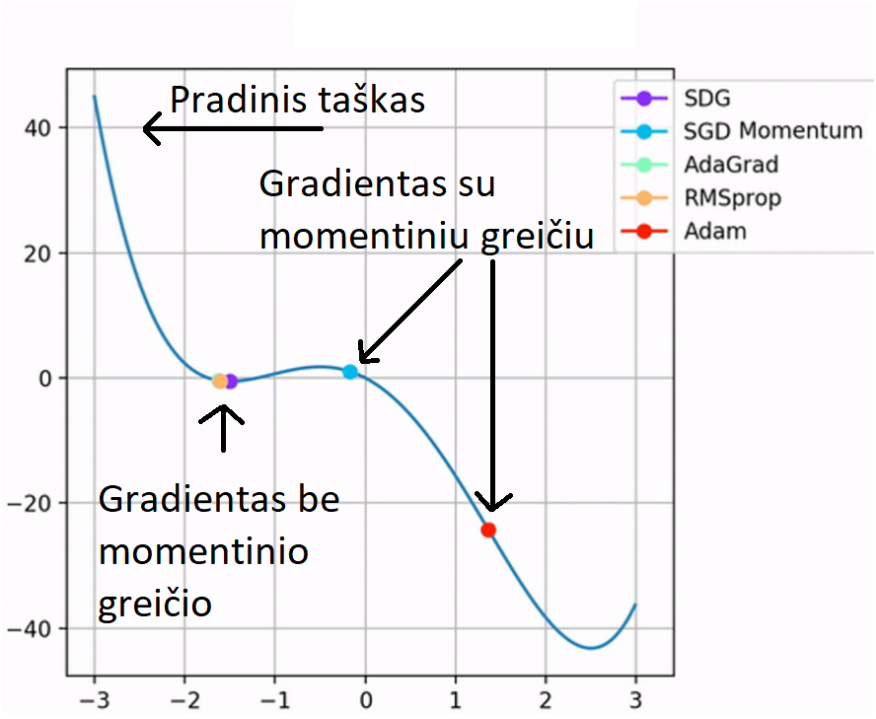
Norint pasiekti greitesni svorių konvergavimą yra sukurta skirtingų gradientinio nuolydžio metodų, kurie netik pagreitina konvergaciją bet ir padeda kovoti su didelėmis gradiento problemomis – keliavimas plokščiu paviršiumi bei lokalus minimumai.

1.6.1. Momentinis gradientinis nuolydis bei eksponentiškai pasvertų svorių vidurkis

Pagrindinis šio optimizavimo principas remiasi momentiniu nuolydžiu, kuomet gradientas įgyja pagreiti ties dažniausiai pasikartojančiu vektoriumi. Primena įsibėgėjusį automobilį. Šis įsibėgėjimas padeda išvengti lokalių minimumu, kaip tai matosi 7 paveikslėlyje. Jame pavaizduoti skirtingi gradientinio mokymosi algoritmai, kurie pradeda tame pačiame taške bei kurių dauguma sustoja mokytis ties lokaliu minimumu. Tačiau momentinio mokymusi paremti gradientinio nuolydžio metodai praskrieja lokalųjį minimumą, kuris tik trumpam sumažina gradiento greitį. Taigi momentinio greičio intuicija labai paprasta: jei gauname naują gradiento reikšmę, kuri nurodo judėti priešinga linkme, ji yra atsveriama gradieno įsibėgėjimo greičiu ir gradientas tik truputi sulėtėja. Šio gradiento formulė primena eksponentiškai pasvertų svorių vidurkio formulę ir išsireiškia tokiu pavidalu:

$$Vdw = \beta_1 \cdot Vdw + (1 - \beta_1) \cdot dw \tag{14}$$

Kur β_1 yra parametras nuo 0 iki 1, dažniausiai būnantis 0.9. Kaip matome naujausi gradientai prie galutinės gradiento reikšmės prisidės tik $(1-\beta_1)$ dydžiu, kas suteikia momentinio greičio pavidalą. Šis pasvertų vidurkių principas yra plačiai naudojamas skatinamuosiuose mokymosi metoduose ir jį galima pastebėti visuose pagrindinėse formulėse. Skatinamuosiuose mokymosi metoduose šios formulės principą galima traktuoti kaip gradientinį nuolydį - po truputi konverguojame ties optimaliomis reikšmėmis.



7 pav. Gradientinė optimizacija su ir be momentinio greičio. SGD momentinis bei Adam konverguoja į globalų minimumą, kiti konvergavo į lokalų minimumą

1.6.2. RMSProp

Tai gradientinio nuolydžio metodas, kurio tikslas pagreitinti gradientą, kuomet judame lygia plokštuma. Tačiau kaip matome iš 7 paveikslėlio RMSProp kenčia nuo momentinio greičio neturėjimo ir sustoja ties lokaliu minimumu. RMSProp forma labai panaši į momentinio gradiento:

$$Sdw = \beta_2 \cdot Sdw + (1 - \beta_2) \cdot (dw)^2. \tag{15}$$

Svorių atnaujinimo formulė:

$$W = W - \alpha \frac{dw}{\sqrt{Sdw}} \tag{16}$$

Empiriškai galime pastebėti, kad didelės gradiento reikšmės sumažės, tačiau mažos padidės. Todėl šis metodas padeda išspręsti lėtą judėjimą lygumomis. Rekomenduojamas β_2 dydis yra 0.99.

1.6.3. Adam

Adam[3] metodas yra vienas populiariausių gradientinio nuolydžio metodų. Jo tikslas sumažinti dvi problemas - lokalius minimumus bei judėjimas plokštuma. Tai pasiekiame sujungiant momentinį greitį bei RMSProp:

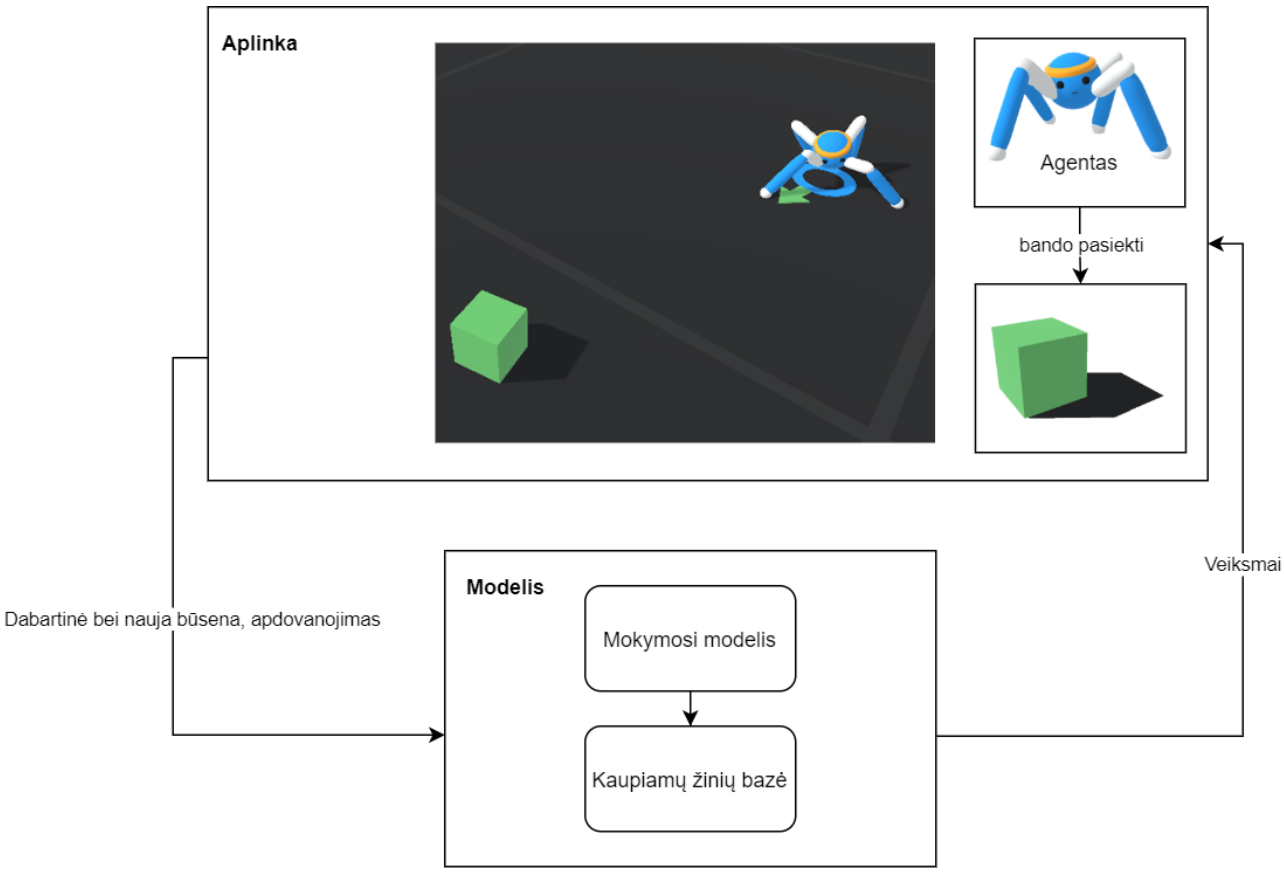
$$W = W - \alpha \frac{Vdw}{\sqrt{Sdw}} \tag{17}$$

Vdw suteikia momentinį greitį kiekvienam svoriui. Bet jei momentinis greitis yra didelis Sdw jį sumažiną. Tačiau jei gradientas užstringa lygumoje, Sdw padeda atstrigti. Šį metodą pritaikiau minėtame automobilio kontrolerio pavyzdyje. Jis padėjo pasiekti trigubai mažesnį klaidos dydį ir taip pagerino automobilio kontrolerį.

1.7. Skatinamasis mokymasis

Skatinamojo mokymosi metodai yra mašininio mokymosi metodai, kurie apdovanoja pageidaujamus veiksmus bei baudžia nepageidautinus veiksmus. Naujuosiuose skatinamuosiuose modeliuose apdovanojimo bei baudos dydis tėra vienintelė informacija, kurią modelis težino. Jei tiesinėje regresijoje mes nuolatos turėdavome tikrąsias reikšmes, ties kuriomis bandėme pritaikyti modelį, šiuo atveju tokia informacija yra nepasiekiama. Todėl mūsų modelis turi nuolatos tyrinėti aplinką, kaupti naujas patirtis. Šiuos duomenis gauname iš aplinkoje esančio roboto ar agento, kuris duotoje aplinkoje stengiasi atlikti optimalius veiksmus ir pagal juos susidaryti geriausių siūlomų veiksmų modelį.

Šią veiksmų eigą apibendrina bendrasis skatinamųjų modelių grafikas, pateiktas 8 paveikslėlyje. Jis iliustruoja praktinės dalies pagrindinės užduoties įgyvendinimą, voro keturių galūnių judinimą, kuriomis bandoma kuo arčiau priartinti vorą ties žaliu kubeliu.



8 pav. Bendras skatinamojo mokslo modelis

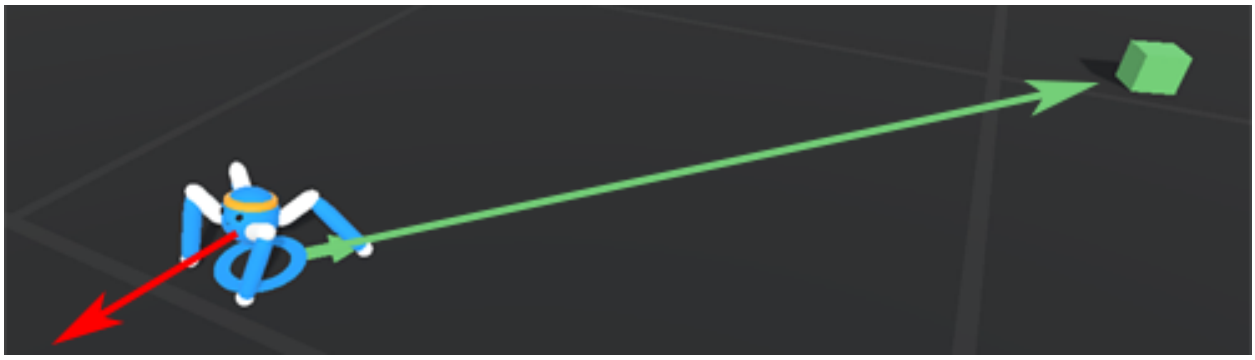
Šio modelio veikimo principai išsiskaido į šias dalis:

1. aplinka.
2. agentas.

Aplinka.

Aplinkoje egzistuoja mūsų agentas, kuris priiminėja modelio sugeneruotus veiksmus. Pateiktame pavyzdyje tai būtų voro galūnių judėjimo trajektorijos. Atlikęs duotus veiksmus, voras įgauną naujas būsenas, pajuda į naują poziciją. Nauja įgyta būsena bei prieš tai buvusi, nusiunčiama modeliui. Taip pat modeliui nusiunčiame apdovanojimo dydį. Jis apskaičiuojamas naudojantis naujai sugeneruotos būsenos naudingumu. Naudojantis voro pavyzdžiu tai atitiktų voro greičio vektorio laipsnio sutapimu su norimos krypties vektoriumi, kaip tai matosi 9 paveikslėlyje. Šiuo atveju apdovanojimas būtų nulinis, nes vektorių trajektorijos visai nesutampa. Atlikus veiksmą bei apskaičiavus jo naudingumą, tokiu būdu gaunamas ryšys tarp veiksmo, kuris sugeneruoja naują būseną, bei įgyto apdovanojimo, naujoje būsenoje. Šis ryšys tarp veiksmo bei gauto rezultato yra neseniai atgimusio skatinamojo mokslo pamatas, kuriuo moderniausi algoritmai ne per seniausiai pradėjo vadovautis, atradus apdovanojimų taisyklių gradientą bei pradėjus taikyti neuroninius tinklus.

Agentas.
Agento sluoksnis susideda iš mokymosi modelio bei žinių bazės, kurioje kaupiami gauti aplinkos rezultatai, susidedantis iš vektorių su šiomis reikšmėmis - praeita būsena, atliktas veiksmas, nauja būsena ir apdovanojimas. Mokymosi modelis naudojasi sukauptų žinių baze gerinti savo modelio rezultatus. Modelio tikslas yra pateikti tokias veiksmų sekas, kurios suteiktų didžiausią įmanoma apdovanojimą. Voro pavyzdžiu tai būtų sėkmingas žalio kubelio pasiekimas. Modelio apmokymui yra sukurta daugybe įvairių metodų ir šiame darbe bus nagrinėjamas vienas iš jų - TD3



9 pav. Apdovanojimo pavyzdys, pagal voro greičio (raudona rodyklė) bei norimo judėjimo (žalia rodyklė) vektorių atitikimu.

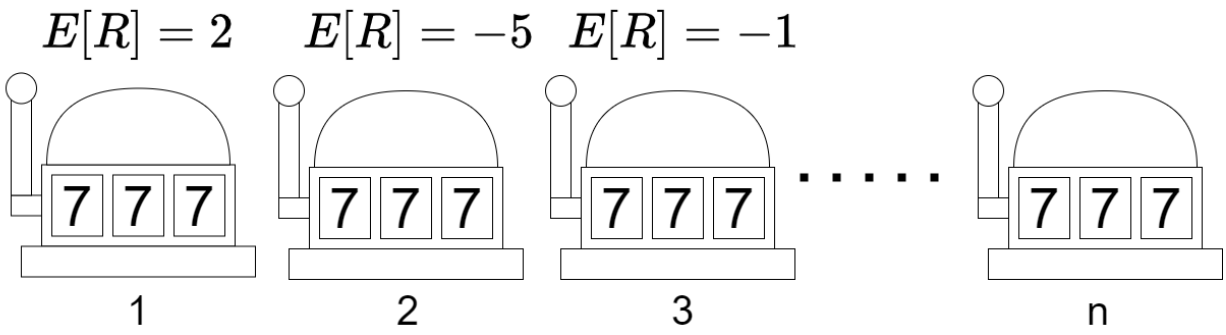
Taigi visas skatinamasis mokslas paremtas tik vienu kintamuoju - apdovanojimo gausumu. Šis apdovanojimas apdovanoja pageidautiną elgesį bei baudžia už neigiamą elgesį. Tokiu būdu agentas skatinamas siekti didžiausio ilgalaikio nuopelno, kas veda į optimalų sprendimą. Tolimesniuose skyriuose bus aptariama kaip iš šio vieno kintamojo yra konstruojami visi skatinamojo mokslo modeliai.

1.7.1. Skatinamojo mokslo pagrindiniai kintamieji, stacionariojo pasiskirstymo problema

Kad ir kokį skatinamojo mokslo uždavinį spręstume, mes visados sutiksime šiuos kintamuosius:

1. t - $\mathbb{N}$ laiko žingsnis. Kiekviena kart atlikus veiksmą bei apskaičiavus jo naudingumą, laiko žingsnis pajuda vienetu į priekį
2. A_t - atliktas veiksmas t laiko žingsnyje.
3. R_t - gautas apdovanojimas atlikus A_t veiksmą.
4. $Q(a)$ - vidurkis, tikėtina apdovanojimų reikšmė pasirinkus veiksmą a. Kitaip tariant $Q(a) = \mathbb{E}[R_t|A_t = a]$. Ši formulė ypatingai svarbi, kadangi iš jos vos ne visi mokymosi metodai yra išvedami.

Dažniausiai pateikiamas pavyzdys, kuris iliustruoja šio mokslo principą bei padeda geriau suprasti šiuos kintamuosius, yra lošimo aparatu, pavaizduotų 10 paveikslėlyje, laimėjimų



10 pav. N lošimo aparatų, su skirtingais vidutiniais laimėjimais

optimizavimas. Šiame uždavinyje iš n aparatų agentas bando surasti lošimo aparatą, kuris suteikia didžiausią įmanomą apdovanojimą. Kiekvienas aparatas duoda atsitiktinį laimėjimą iš $\mathcal{N}(\mu = \mathcal{N}(0{,}2), \sigma = 1)$ pasiskirstymo. Žaidimo sesija, dažniausiai vadiname epizodas, nėra ribojama, agentas gali atlikti begalybę t žingsnių, lošimo aparatų pasirinkimų. Pasirinkimą galima traktuoti kaip veiksmą a ir gautą laimėjimą kaip r. Agentas bando vienos žaidimo sesijos metu įgyti didžiausią sukauptą laimėjimą. Jis pritaiko viena iš populiariausiu skatinamojo mokslo metodu $\epsilon - godusis$. Šis metodas visados renkasi lošimo aparatą, kurio aproksimuotas tikėtinas laimėjimas, žymimas $Q(a)$, yra didžiausias, kaip tai išreikšta 18 formulėje.

$$A_t = argmax_a Q_t(a) \tag{18}$$

Tačiau su maža ϵ tikimybe, agentas pasirenka atsitiktinį lošimo aparatą. Taigi kiekviename t žingsnyje agentas renkasi lošimo aparatą, jį aktyvuoja ir gauną naują apdovanojimą, kuri talpina į to aparato apdovanojimų sąrašą. Šis procesas ypatingai svarbus, jis dominuoja visuose mokymosi modeliuose. Agentas stengiasi ne tik gauti didžiausia įmanomą tikėtiną laimėjimą, tačiau su ϵ tikimybe kartu tyrinėja aplinką, ieško optimalesnių lošimo aparatų, nei surastas dabartinis geriausias. Šis procesas vadinasi apdovanojimų-tyrinėjimo kompromisu. Lošimo aparato tikėtinas laimėjimas tėra to aparato gautų laimėjimų vidurkis, kaip tai matosi 19 formulėje.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i * \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} \tag{19}$$

Kitaip tariant $Q_t(a)$ yra apdovanojimų suma, kuomet pasirinkome a veiksmą (arba lošimo aparatą), atlikus t žingsnių, padalinti iš kiek kartų buvo a veiksmas (arba lošimo aparatas) pasirinktas. Kuomet t artėja link begalybės, vadovaujantis didžiųjų skaičių dėsniu, $Q(a)$ priartėja prie tikrosios reikšmės. Visi naujausi modeliai yra paremti panašiu imties traukimo metodu - simuliuojame skirtingus veiksmus ir stebime gautus rezultatus. Nors pavyzdys atrodo primityvus, tačiau visi naujausi modeliai sprendžia tą pačią užduoti - kokį pasirinkti veiksmą a, kuris suteiktų didžiausią apdovanojimą.

Pagrindiniai skirtumai tarp metodų atsiranda dėl kompiuterio resursų optimizavimo. Viena iš problemų, su kuria iškarto susiduriame, yra duomenų kaupimas. Kiekvienam lošimo aparatui reikia kaupti jo gautus apdovanojimus. Didėjant problemai bei jos duomenų dimensijai, kompiuteriui kaupti duomenis tampa fiziškai nebeįmanoma, todėl yra sukurta daugybe esminių metodų šią problemą mažinti bei išvengti. Vienas iš metodų, kurio formą galima pastebėti TD formulėje, yra rekursyvus vidurkio apskaičiavimas, pateiktas 20 formulėje. Ši formule padeda išvengti apdovanojimų kaupimo, kaip tai matėme lošimo pavyzdyje.

$$\begin{aligned} Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\ &= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\ &= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\ &= \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\ &= \frac{1}{n}\left(R_n + nQ_n - Q_n\right) \\ &= Q_n + \frac{1}{n}\left[R_n - Q_n\right] \end{aligned} \tag{20}$$

Lošimų pseudokodas, kuris palaipsniui apskaičiuoja skirtingų veiksmų imties vidurkius, pateiktas toliau:

Algoritmas 1: Lošimo aparato pseudo algoritmas

$Q(a) \leftarrow 0$;
$N(a) \leftarrow 0$;
for ∞ **do**

$$A \leftarrow \begin{cases} \arg\max_a Q(a) \text{ su tikimybe } 1-\varepsilon \text{ (lygiąsias reikšmes pasirenkame atsitiktinai)} \\ \text{atsitiktinis veiksmas su tikimybe} \varepsilon \end{cases}$$

$R \leftarrow$ LošimoAparatas (A) ; `// Pasirenkame veiksma A ir gauname apdovanojimą R`
$N(A) \leftarrow N(A) + 1$; `// Kiek kartų kiekvienas veiksmas buvo pasirinktas`
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}[R - Q(A)]$;

Rezultatas: Gautos optimalios, tikrosios Q(a) reikšmės. Jas turint žinosime, kuris aparatas duoda didžiausią laimėjimą

Paskutinė 20 formulės išraiška yra ypatingai svarbi, nes ji yra ištaka tiek TD, tiek Q-mokymosi, tiek Bellmano formulėms, kurios yra skatinamojo mokslo pamatas. Apibendrinant 20 formulę ją galima šitaip išreikšti:

$$\text{Naujas vidurkis} = \text{Senas vidurkis} + \frac{1}{n}(\text{Nauja reikšmė - Senas vidurkis}) \tag{21}$$

Jei tęstume formules apibendrinimą išryškėtu Bellmano funkcija, kuri yra viena svarbiau-

siu šiame moksle ir į kuria gilinsimės tolimesniuose skyriuose:

$$\text{Naujas įvertis} = \text{Senas įvertis} + \frac{1}{n} * (\textbf{Tikslas - Senas įvertis}) \tag{22}$$

Paryškinta dalis yra vadinama TD paklaida, ties kuria gilinsimes tolimesniuose skyriuose.

Kol kas nagrinėta problema pasižymi apdovanojimo skirstino stacionarumu, jis nekinta. Tačiau nagrinėjant sudėtingas problemas netik mūsų apdovanojimų pasiskirstymas kinta, tačiau mums įdomios tik paskutinės 10 -100 reikšmės. Taip pat aukštos dimensijos problemose mums neužtenka kompiuterio resursų išsaugoti $N(A)$ reikšmių, ką mes atliekame 1 pseudokode. Todėl įgyvendinamas dar vienas 20 formulės optimizavimas, aptariamas sekančiame skyriuje.

1.7.2. Ne stacionarus apdovanojimo skirstinys

Galima pastebėti daugybę problemų su 20 formule:

1. Toliau plėtojant duotą pavyzdį, galima įsivaizduoti situaciją, kurioje, kas tam tikrą laiko intervalą, kiekvieno aparato apdovanojimo funkcija pakinta ir įgyja nauja $\mathcal{N}(\mu = \mathcal{N}(0{,}2), \sigma = 1)$ pasiskirstymą.
2. Kuomet vieno lošimo aparato apdovanojimų R imtis yra labai didelė, naujai gautos reikšmės turi labai mažai įtakos vidurkio pokyčiui dėl $\frac{1}{n}$ formulėje esančios išraiškos.
3. Kiekviena $Q(a)$ reikšmė turi kaupti $N(A)$. Aukštosiose dimensijos neužtenka kompiuterio resursų išsaugoti šias reikšmes.

Šios problemos sprendimas jau buvo pateiktas Adamo algoritmo įgyvendinime, mes toliau naudojame slenkančio vidurkio principus:

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha\left[R_n - Q_n\right] \\
&= \alpha R_n + (1-\alpha)Q_n \\
&= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right] \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\
&\quad \ldots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i
\end{aligned} \tag{23}$$

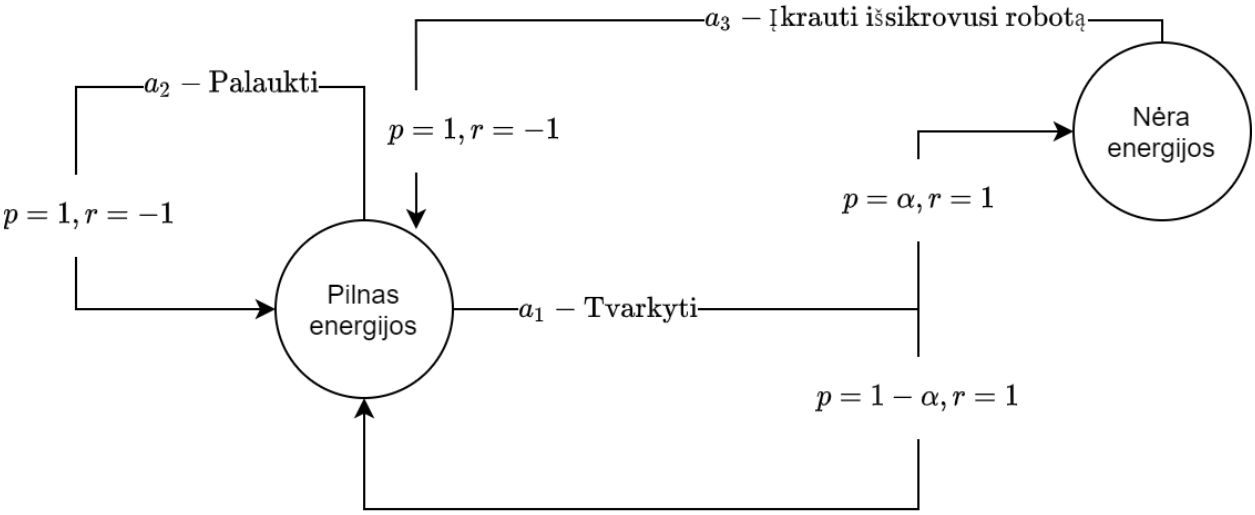
Ši 23 formulė pakeičia $\frac{1}{n}$ į $\alpha \in [0{,}1]$. Dažniausia α reikšmė yra 0.9, su kuria apytiksliai apskaičiuojame paskutinių gautų 10-ties narių vidurkį.

1.7.3. Markovo grandinės

Dauguma mažos dimensijos skatinamojo mokslo problemų galima išreikšti Markovo grandinių pavidalu. Toks išreiškimas padeda geriau suvokti keturias pagrindines skatinamojo mokslo sąvokas:

1. s - esama būsena. Pvz., automobilio momentinis greitis ar automobilio sensorių informacija.
2. a - pasirinktas veiksmas. Pvz., automobilio greičio pedalo paspaudimo stiprumas.
3. r - esamosios būsenos pasirinkto veiksmo apdovanojimas. Pvz., jei simuliuojamas automobilis pateko į avariją, jo apdovanojimas tampa neigiamas, gaunama bauda.
4. s' - nauja būsena, kuri gaunama atlikus a veiksmą.

Nors dažniausiai būsenos būna daugiau nei vienos dimensijos, pavyzdžiui būsenos vektorius susidedantis iš 2 reikšmių - pozicijos bei greičio, lengviausiai jas yra suprasti atvaizduojant vienos dimensijos Markovo grandinių pavidalu, kaip tai matosi 11 paveikslėlyje.



11 pav. Dulkio siurblio Markovo grandinė

Pvz., jei dulkių siurblio robotas yra būsenoje $s =$ (Pilnas energijos), jis gali pasirinkti a_1 arba a_2 veiksmus. Pasirinkęs a_1 veiksmą, robotas su α tikimybe gali patekti į būsena $s' =$ (Nėra energijos) ir su tikimybe $1-\alpha$ į būseną $s' =$ (Pilnas energijos). Už abu šiuos veiksmus veiksmus robotas yra apdovanojamas $r = 1$. Šis procesas galioja ir kitoms būsenoms.

Priklausomai nuo esančios būsenos, mes visą laiką norime pasirinkti tokį veiksmą, kuris duoda didžiausią tikėtiną apdovanojimą:

$$r(s,a) \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p\left(s', r \mid s, a\right) \tag{24}$$

Veiksmų pasirinkimas vadovaujantis didžiausiu tikėtinu apdovanojimu bus nagrinėjamas sekančiame skyriuje.

Markovo grandinių išraiška taip pat yra svarbi ne tik dėl vizualinio aiškumo. Visos skatinamojo mokymo problemų formuluotės privalomai turi galėti išsireikšti Markovo grandinių pavidalu. Šis reikalavimas yra būtinas dėl vienos iš pagrindinių Markovo grandinės savybių - ateities įvykio tikimybė nepriklauso nuo prieš tai įvykusių įvykių:

$$P\left(X_{n+1} = x \mid X_0, X_1, X_2, \ldots, X_n\right) = P\left(X_{n+1} = x \mid X_n\right), \tag{25}$$

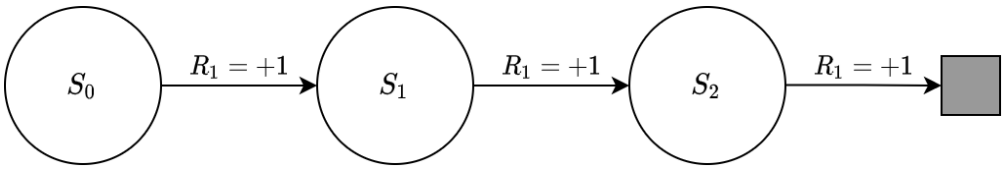
kur X yra prieš tai buvę įvykiai. Ši savybė leidžia apskaičiuoti tikėtiną apdovanojimą pasitelkiant tik turimą būseną, prieš tai buvusios būsenos neturi įtakos tikėtinam apdovanojimui.

Taigi 11 pavyzdys iliustruoja kaip skatinamajame moksle yra lengva apskaičiuoti naudingiausius veiksmus, kuomet turime sudaryta modelį. Tačiau retu atveju mes jį turime ir dažniausiai mums reikia imtis aplinkos tyrinėjimo veiksmų, kurie sudaro modelio aproksimacijas. Taip pat mes ne tik norime apskaičiuoti geriausią sekančio veiksmo apdovanojimą, tačiau geriausią apdovanojimą ilgalaikėje perspektyvoje. Dažnai gyvenime mes pasirenkame mėgautis maloniais dalykais, kuriais vėliau mes gailimės. Būtent šią problemą kitas skyrius ir nagrinės - ilgalaikis apdovanojimas, o ne trumpalaikis.

1.7.4. Apdovanojimai bei tikslai

Apdovanojimų hipotezė: kad visa tai, ką mes turime omenyje kalbėdami apie tikslus ir uždavinius, gali būti gerai suprantama kaip tikėtinos vertės didinimas, kai gaunama ilgalaikė skaliarinio signalo (vadinamo apdovanojimu) suma.

Šią sumą mes vieno epizodo metu stengiamės didinti. Epizodas - būsenų, veiksmų, apdovanojimų seka, kuri užsibaigę agentui pasiekus tikslą arba sustojimo sąlygą, kaip tai matome 12 paveikslėlyje. Agentas renkasi skirtingus veiksmus, įgauna naujas būsenas, kaupia epizodo apdovanojimą, kol galiausiai pasiekę tikslą, pavaizduotų pilku kvadratu.



12 pav. Vienas agento epizodas

Epizodo galutinis sukauptas apdovanojimas užrašomas tokiu pavidalu:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T, \tag{26}$$

kur T yra paskutinis epizodo žingsnis. Tačiau jei agentas optimizuoja savo veiksmus aplinkoje, kurios pabaigą neįmanoma identifikuoti arba epizodas trunka nepaprastai daug žingsnių, mes naudojame apdovanojimų proporcionalius mažinimus:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{27}$$

kur γ yra nuolaidos kintamasis, kurio dažniausia vertė yra 0.9. Ši vertė indikuoja, kad 11 žingsnio apdovanojimas bus proporcionalus 0.9^{11} savo vertės, kitaip tariant nuliui. Tačiau ši formulė kaip ir praeitos pasižymi panašiais trūkumais, kaip nuolatinis duomenų kaupimas. Tai galima išspręsti perrašius formulę į rekursinę formą:

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{28}$$

Taip pat ši formulė padeda pasiekti balansą ties ilgalaikio apdovanojimo bei trumpalaikio apdovanojimo optimizavimo.

Vienas iš svarbiausių praktinių patarimų, padedančių pasiekti didžiausią ilgalaikį apdovanojimą yra priskirti apdovanojimus už pasiektą tikslą, o ne skirti apdovanojimus už tai kaip mes norėtume, kad agentas pasiektų tikslą. Agentas mokymo procese turi pats atrasti optimaliausius veiksmus vedančius ties tiek tikslo pasiekimu, tiek didžiausiu epizodo apdovanojimu. Kitame skyriuje bus apibrėžiama, kaip būtent agentas optimizuoja savo elgesį, jo viena iš tikslo funkcijos konceptų, kurį gradientiniai metodai nuolatos optimizuoja.

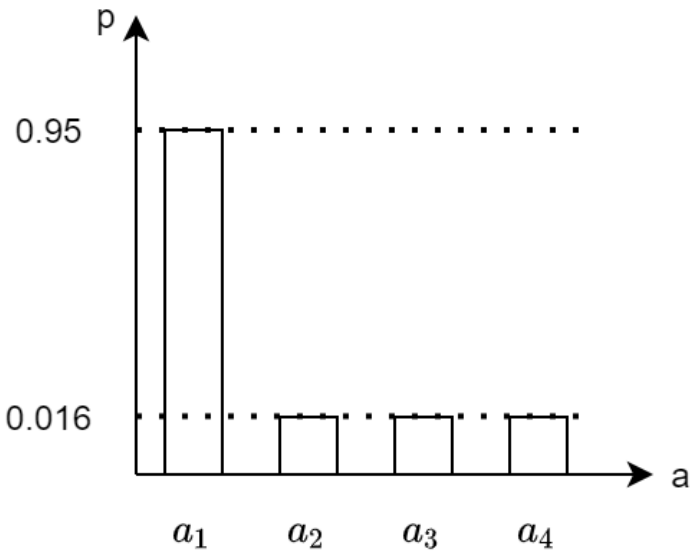
1.7.5. Vertės bei veiksmų taisyklių funkcijos

Norint pasirinkti veiksmus, vedančius į būsenas, kurios suteikia didžiausią įmanoma apdovanojimą, mes turime kiekvienai būsenai priskirti tam tikrą matą. Todėl kiekviena būsena yra nusakoma pagal savo vertę. Ši vertė yra žymima $v_\pi(s)$ ir nusako pasirinktos s būsenos tikėtiną apdovanojimą:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right], \text{ visiems } s \in \mathcal{S} \tag{29}$$

Žinant kiekvienos būsenos tikėtina apdovanojimą, mes visados galime rinktis tik tuos veiksmus, kurie veda į didžiausią apdovanojimą, didžiausią $v_\pi(s')$. Šį pasirinkimą nusako veiksmų taisyklės. Tai yra tikimybių pasiskirstymas žymimas $\pi(a|s)$, nurodantis būsenos s visų veiksmų pasirinkimu pasiskirstymą. Kitaip tariant su kokia tikimybe mes pasirinktume veiksmą a būnant būsenoje s, kaip tai matosi paveikslėlyje 13.

Užrašymas $v_\pi(s)$ nurodo, kad ši būsenos vertė yra apskaičiuota naudojantis π veiksmų taisyklėmis. Veiksmų taisyklės gali būti įvairios. Dažniausiai naudojame $\epsilon-godusis$ metodą, kaip tai matėme ties 18 formule bei kurio pritaikymas Markovo grandinėms yra pateiktas 30 formulėje. Šio metodu mes visą laiką priskiriame $1-\epsilon$ tikimybę veiksmui, kuris gražina didžiausią apdovanojimą apskaičiuojama naudojantis $v_\pi(s)$ formule.



13 pav. Veiksmų taisyklių pasiskirstymas, kur $\epsilon = 5$.

$$\begin{aligned}\pi'(s) &\doteq \arg\max_a q_\pi(s,a) \\ &= \operatorname*{argmax}_a \mathbb{E}\left[R_{t+1} + \gamma v_\pi\left(S_{t+1}\right) \mid S_t = s, A_t = a\right] \\ &= \operatorname*{argmax}_a \sum_{s',r} p\left(s', r \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]\end{aligned} \tag{30}$$

Turint kiekvienos būsenos taisyklių pasiskirstymą, galima suskaičiuoti kiekvienos būsenos tikėtiną reikšmę, kaip tai matosi 31 formulėje.

$$\begin{aligned}v_\pi(s) &\doteq \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \\ &= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p\left(s', r \mid s, a\right)\left[r + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right] \\ &= \sum_a \pi(a \mid s) \sum_{s',r} p\left(s', r \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right], \qquad \text{kiekvienam } s \in \mathcal{S},\end{aligned} \tag{31}$$

Šios formulės paskutinė lygybė yra dar geriau žinoma kaip Bellmano lygybė. Ji padeda rekursiškai apskaičiuoti $v_\pi(s)$ reikšmes bei išvengti apdovanojimų kaupimo kompiuterio atmintyje.

Tačiau praktikoje mes naudojame ne $v_\pi(s)$ - būsenos vertės matą, o $q_\pi(s,a)$ - veiksmo vertės matą esant būsenoj s:

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right] \tag{32}$$

Galime pastebėti, kad $v_\pi(s)$ yra tos būsenos s visų veiksmų $q_\pi(s,a)$ proporcinga $\pi(a|s)$ sudėtis bei taip pat išsireiškia Bellmano lygybe:

$$
\begin{aligned}
&q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a\right] = \\
&= \sum_{s'}\sum_{r} p\left(s', r \mid s, a\right)\left[r + \gamma E_\pi\left[G_{t+1} \mid S_{t+1} = s'\right] = \right. \\
&= \sum_{s'}\sum_{r} p\left(s', r \mid s, a\right)\left[r \cdot \gamma \sum_{a'} \pi\left(a' \mid s'\right) \cdot E_\pi\left[G_{t+1} \mid S_{t+1'} = s', A_{t+1} = a'\right]\right] \\
&= \sum_{s'}\sum_{r} p\left(s'r \mid s, a\right)\left[r + \gamma \sum_{a'} \pi\left(a' \mid s'\right) \cdot q_\pi\left(s', a'\right)\right],
\end{aligned}
\tag{33}
$$

kur $v_\pi(s') = \gamma E_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]$.

Funkcijos $q_\pi(s,a)$ dėka mums nereikia apskaičiuoti s būsenos visų veiksmų apdovanojimus norint gauti $v_\pi(s)$. Užtenka rinktis tik tuos veiksmus, kurie yra jau apskaičiuoti bei optimaliausi.

Taigi turint būsenų verčių bei veiksmų taisyklių apskaičiavimo formules galime pradėti iteracinį procesą, kurio tikslas yra optimizuoti šias reikšmes.

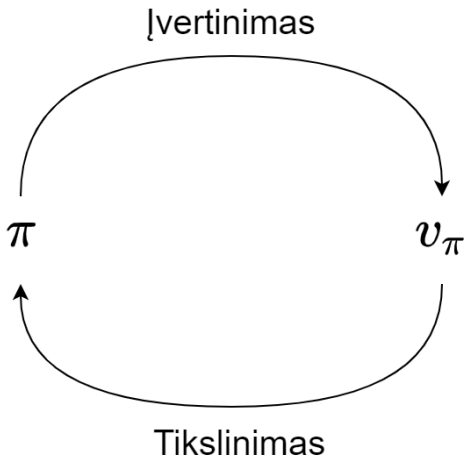
1.7.6. Verčių bei veiksmų taisyklių iteravimas

Optimizavimo metu, mes ieškome tokių būsenų verčių bei optimalių taisyklių, kurios gražintų didžiausią epizodo apdovanojimą. Yra daugybė būdų tai atlikti, kuriuos galima sugrupuoti į tris kategorijas:

1. dinaminio programavimo metodas, kuomet modelis žinomas, mes kiekvienos iteracijos metu perskaičiuojame visas $v_\pi(s)$ reikšmes ir atnaujiname veiksmų taisykles. Šį procesą iteruojame, iki konvergacijos pasiekimo. Nenaudojamas metodas, nes labai retai tenkinama šio metodo privaloma sąlyga - modelio turėjimas. Modelį mes traktuojame kaip Markovo grandinių modelį, su visų veiksmų perėjimo tikimybėmis. Tačiau šio metodo dinaminėmis taisyklėmis remiasi vos ne visi naujausi metodai, todėl svarbu į šį metodą įsigilinti.
2. Monte Carlo metodai - imties metodais apskaičiuojame $q_\pi(s,a)$ reikšmes. Šis metodas taikomas kuomet neturime modelio ir reikia jį aproksimuoti.
3. Neuroninių tinklų aproksimacijos - metodai, kurie aproksimuoja $q_\pi(s,a)$ reikšmes. Jei Monte Carlo metodai apskaičiuodavo kiekvienos būsenos tikslias reikšmes, funkcijų aproksimacijos metodais mes jas apskaičiuojame tik apytiksliai.

Visos šios kategorijos pasižymi bendru optimizavimo principu:

1. įvertinimas - perskaičiuojame $v_\pi(s)$ reikšmes naudojantis atnaujintomis veiksmų taisyklėmis.
2. tikslinimas - turint naujas $v_\pi(s)$ reikšmes, perskaičiuojame optimalias veiksmų taisykles $\pi(a|s)$.



14 pav. Viena optimizacijos iteracija

Šiuos žingsnius kartojame tol, kol pasiekiame $v_\pi(s)$ konvergaciją - naujos iteracijos nepakeičia $v_\pi(s)$ reikšmių. Tolimesniuose skyriuose trumpai apžvelgsiu šio proceso įgyvendinimą dinaminiuose bei Monte Carlo metoduose. Neuroninių tinklų aproksimacijos bus plėtojamos giliausiai, kadangi šiuo metu jie yra labiausiai paplitę bei efektyviausi.

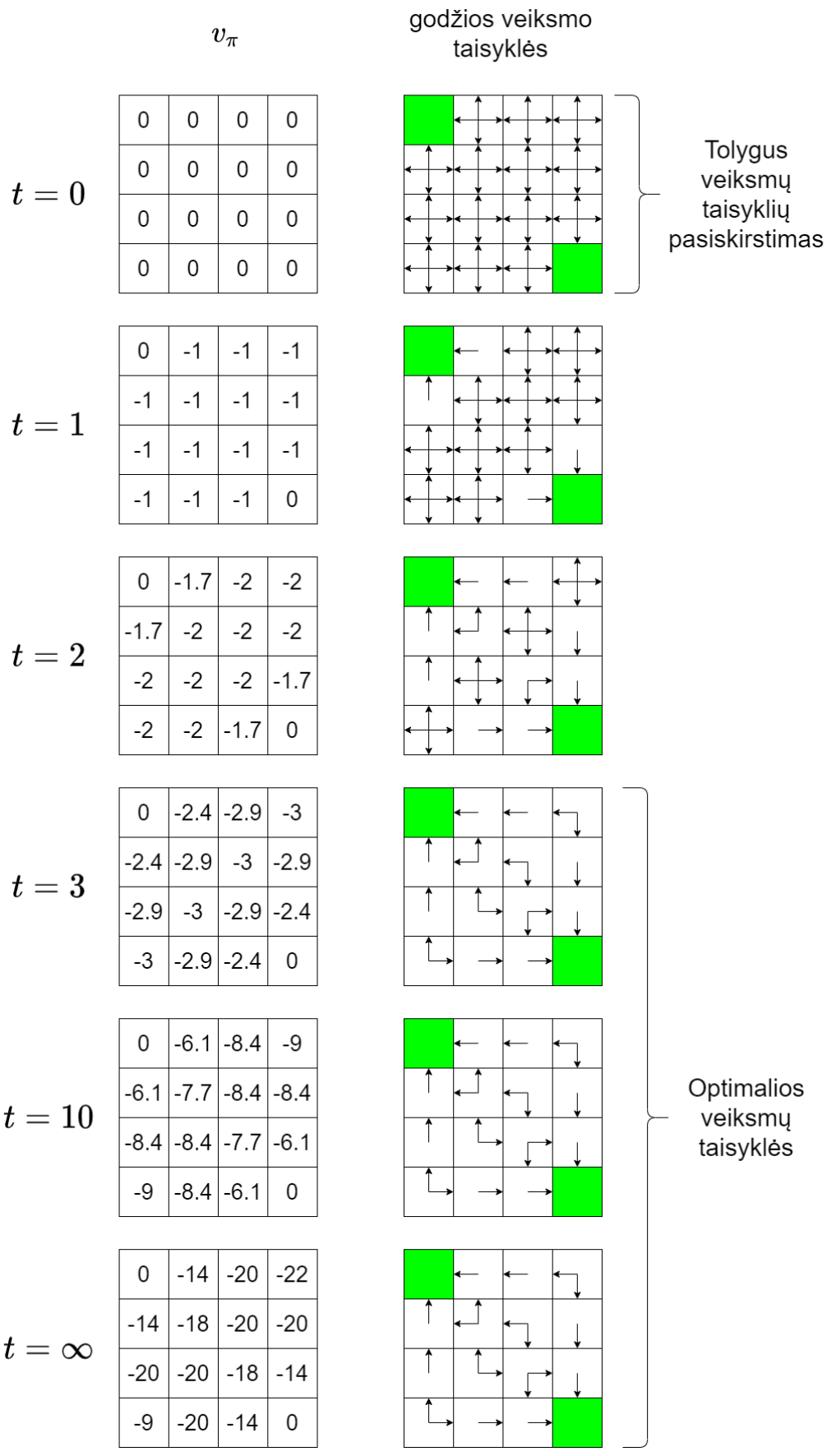
1.7.7. Dinaminis programavimas

Įvertinimo bei tikslinimo procesą, pateikta 14 paveikslėlyje, paprasčiausiai iliustruoja dinaminio programavimo metodas pateiktas 15 pavyzdyje. Šiame uždavinyje yra šešiolika langelių, ties kuriais agentas gali apsistoti. Du iš jų yra žali, žymi pozicijas, ties kuriomis agentas turi atsirasti. Agentas patekęs ant likusių baltų langelių turi nukakti į artimiausią žalią langelį. Agentas gali judėti tik keturiomis trajektorijomis, pateiktomis rodyklėmis. Kairėje paveikslėlio pusėje yra pateikti $v_\pi(s)$ įverčiai naudojantis 31 formule, o dešinėje perskaičiuojame optimalias veiksmų taisykles naudojantis 30 formule. Ypatingai svarbi optimizavimo seka:

1. Įvertinimas - Perskaičiuojame $v_\pi(s)$ reikšmes atnaujinamos naudojantis $t-1$ žingsnio taisyklėmis.
2. Tikslinimas - Perskaičiavus $v_\pi(s)$ reikšmes, perskaičiuojame t žingsnio optimalias veiksmų taisykles $\pi(a|s)$. Pavyzdyje naudojame godų veiksmo taisyklių įvertinima, kuris renkasi tik optimaliausius veiksmus.

Vienas agento žingsnis, žymimas t, susideda iš šios įvertinimo bei tikslinimo sekos. Matome, kad užtenka vos trijų žingsnių veiksmų taisyklėms tapti optimaliomis. Tačiau $v_\pi(s)$ reikšmių konvergacija užtrunka ilgesni laiką. Įvykus konvergacijai, agentui atsiradus ant bet kurio iš galimų baltų langelių, jis, vadovaudamasis veiksmų taisyklėmis, visados optimaliai suras kelią į žalius langelius. Taip pat labai svarbu pastebėti seniau aptarta $v_\pi(s)$ išsiskaidymą į $q_\pi(s,a)$ narių sumą. Šis išsiskaidymas intuityviai matosi pateiktame 15 paveikslėlio

uždavinyje, kuomet kiekvieną rodyklę, galima traktuoti kaip $q_\pi(s, a)$ reikšmę. $q_\pi(s, a)$ Išskaidymas yra svarbus neuroninių tinklų aproksimacijose, kaip tai bus aptarta tolimesniuose mokymosi modeliuose. Tačiau šis algoritmas yra tinkamas tik apskaičiuoti mažų dimensijų problemas bei iliustruoti intuityvius pavyzdžius. Augant tiek veiksmų, tiek būsenų erdvei šie skaičiavimai tampa neįmanomi. Ši problema plėtojama kitame skyriuje.



15 pav. Dinaminio programavimo pavyzdys.

1.7.8. Monte Carlo metodai

Nors specifinių Monte Carlo metodų šiame darbe nenagrinėsiu, tačiau pabrėšiu jų labai svarbų principą, kuriuo visi naujausi metodai remiasi. Monte Carlo metodais mes simuliuojame agento epizodus, gauname veiksmų trajektorijas (kaip tai matosi 12 paveikslėlyje) iš kurių susiskaičiuojame $q_\pi(s, a)$ reikšmes. Kiekvienos iteracijos metu kiekvienai būsenai s bei jos specifiniui veiksmui a mes išsaugome gautą apdovanojimą G_t. Iš šių išsaugotų reikšmių mes susiskaičiuojame vidurkį ir jį priskiriame $q_\pi(s, a)$ nariui. Tai atlikus mes perskaičiuojame veiksmų taisyklių reikšmes, kurios pasirenka naujus optimalius veiksmus. Po daugybes iteracijų, $q_\pi(s, a)$ konverguoja ties specifine, optimalia reikšme. Šio metodo veikimo principai pateikti 2 pseudokode.

Algoritmas 2: Taisyklių bei verčių aproksimavimas Monte Carlo pirmo apsilankymo metodu

Algoritmo parametrai:

$\varepsilon > 0$

Duomenų priskirimas:

$\pi \leftarrow$ atsitiktinė ε-greedy veiksmų taisyklių pasiskirstymas

$Q(s, a) \in \mathbb{R}$ (atsitiktinis priskirimas), visiems $s \in \mathcal{S}, a \in \mathcal{A}(s)$

Vertės $(s, a) \leftarrow$ tuščias sąrašas, visiems $s \in \mathcal{S}, a \in \mathcal{A}(s)$

for ∞ **do**

  Agentas sugeneruoja epizodą sekant $\pi : S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

  $G \leftarrow 0$

  for $t = T-1, T-2, \ldots, 0$ **do**

    $G \leftarrow \gamma G + R_{t+1}$

    if *Jei* S_t, A_t *būsena yra pirmas apsilankymas* $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$ *sekoje* **then**

      Išsaugoti G į Vertės(S_t, A_t) sąraša

      $Q(S_t, A_t) \leftarrow$ vidurkis (Vertės(S_t, A_t))

      $A^* \leftarrow \arg\max_a Q(S_t, a)$

    for *visi* $a \in \mathcal{A}(S_t)$ **do**

      $$\pi(a \mid S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon / |\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon / |\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

Rezultatas: Gautos optimalios $Q(s, a)$ bei $\pi(a|s)$ reikšmės.

2 Algoritmo problema, kaip ir prieš tai buvusių metodų, yra kiekvienos būsenos ir jo veiksmo saugojimas bei ilgas konvergavimas. Augant būsenų bei veiksmų dimensijoms, konvergavimas tampa neįgyvendinama užduotimi. Veiksmams bei būsenoms užtenka būti trečioje dimensijoje ir jau nebeužtenka kompiuterio resursų apskaičiuoti optimalių reikšmių. Taip pat Monte Carlo metodai pasižymi duomenų kaupimo problema, kuomet neužtenka at-

minties juo saugoti. Duomenų kaupimą galima sumažinti pasitelkiant 20 bei 23 formulėmis, kaip tai atlieka pagrindinės šio mokslo formulės - TD ir Q-mokymasis.

1.7.9. TD bei Q-mokymasis

TD formulė sujungia Monte Carlo bei dinaminio programavimo pagrindines idėjas į vieną paprastą 34 formulę.

$$\begin{aligned} V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left[G_t - V\left(S_t\right)\right] = \\ = V\left(S_t\right) + \alpha\left[R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right)\right]. \end{aligned} \tag{34}$$

Šios formulės ištakas galima pastebėti tiek 20 bei 23 formulėse. Šias dvi formules sujungus, galime perrašyti 31 formulę, kuri pasitelkia Markovo grandinių modeliu, į 34 formulę, be Markovo grandinių modelio. TD paklaida yra paklaida tarp $R_{t+1} + \gamma V\left(S_{t+1}\right)$ bei $V\left(S_t\right)$ verčių, kaip tai matosi 34 formulėje. Kuomet $V\left(S_t\right)$ konverguoja, TD paklaida tampa nulinė. Pasiekus nulinę TD paklaidą galima laikyti, kad modelis yra tikslus. Vienas iš pagrindinių šio modelio skirtumu lyginant su 2 algoritmu, yra greitesnė konvergacija. Šis metodas nelaukia epizodo galo, o atlieka $v_\pi(s)$ verčių atnaujinimą kiekvieno agento žingsnio metu. TD metodas yra pagrindas SARSA bei Q-mokymosi metodams, kurie atlieka atnaujinimus $q_\pi(s, a)$ vertėms. Šios dvi metodologijos skiriasi savo optimalumo atradimu, SARSA naudojasi šia TD atnaujinimo formule:

$$Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma Q\left(S', A'\right) - Q(S, A)\right]. \tag{35}$$

o Q-mokymasis šia:

$$Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma \max_a Q\left(S', a\right) - Q(S, A)\right]. \tag{36}$$

Nors šios formulės yra vos ne identiškos, tačiau jos skiriasi skirtingomis optimaliomis veiksmų trajektorijomis. SARSA modelio atveju agentas stengiasi elgtis saugiai, dažniau vengia būsenų, kurių vertė yra neigiama. Šis modelis tinkamas tokiems eksperimentams, kuriuos yra brangu simuliuoti ir norime sumažinti agento kritines klaidas besimokant, vedančias ties agento fiziniu sugedimu, praradimu. Šio atvejo pavyzdys galėtų būti roboto navigacija ties šlaitu. Kiekvienos simuliacijos metu agentui besimokant jis gali nukristi nuo šlaito ir sugadinti savo įranga. Todėl norint šias brangias klaidas sumažinti mes apmokome robotą pasitelkiant SARSA metodu. Bet kadangi didžiąja laiko dalį mokymasis vyksta simuliuotose aplinkose, mūsų agento kritinių klaidų kaina nėra faktorius. Svarbiausias tikslas mums tampa optimalios verčių suradimas, kuo Q-mokymasis ir pasižymi. Q-mokymasis [4] yra vienas iš populiariausių bei seniausių egzistuojančių metodu. Tačiau jis išpopuliarėjo visai neseniai, kuomet 2014 metais Google Deepmind komanda sugebėjo 36 funkcija aproksimuoti pasitelkiant konvoliucinius neuroninius tinklus. Šiomis aproksimacijomis Deepmind

komanda apmokė agentus įveikti Atari žaidimus taip pademonstruojant skatinamojo mokymosi bei neuroninių tinklų aproksimacijų potencialą. Visi dabartiniai modeliai pasitelkia neuroninių tinklų aproksimacijomis. Aproksimacijos padėjo išspręsti viena didžiausių skatinamojo mokslo problemų - aukštų būsenų dimensijų neįmanoma konvergavimą. Dėl šios priežasties toliau bus aptariamas gilusis-Q - vienas iš pirmųjų, jau nebeefektyviu modeliu, sukėlusių skatinamojo mokslo atgimimą.

1.7.10. Gilusis skatinamasis mokslas, gilusis-Q

Visi naujausi, dabartiniai skatinamojo mokslo modeliai yra aproksimuojami neuroniniais tinklais. Neuroniniai tinklai padeda generalizuoti nepaprastai didžiules būsenų erdves, kaip tai pademonstravo Google Deepmind giliojo-Q konvoliucinis neuroninis tinklas[6], kurio įvestis buvo kompiuterio ekrano individualūs pikseliai.

Gilusis-Q metodas parametrizuoja $q_\pi(s,a)$ apdovanojimų vertę ir užrašoma $q_\pi(s,a,\theta)$ funkcijos pavidalu, kur π yra veiksmų taisyklių tikimybių pasiskirstymas, kaip tai matėme 13 paveikslėlyje. Jo θ parametrai tampa neuroninio tinklo θ parametrai, kurie yra gradientiniais metodais optimizuojami, su tikslu pasiekti tikrąsias optimalias $q_\pi(s,a)$ reikšmes. Gilusis-Q neuroninio tinklo išvestis yra aproksimuota $q_\pi(s,a)$ reikšmė, kuri yra pateikta Q-mokymosi 36 formulėje šiuo $[R+\gamma \max_a Q(S',a)]$ pavidalu. Kaip matome, mus domina tik pateiktos funkcijos tikslo dalis, kaip tai išreikšta 22 formulėje. Šis tikslas yra traktuojamas, kaip naujausia $q_\pi(s,a)$ reikšmė. Dažnai norint tai pabrėžti Q-mokymosi formulė yra pertvarkoma į 37 formulėje pateiktą Bellmano pavidalą:

$$Q^{\text{naujas}}(s_t,a_t) \leftarrow (1-\alpha)\cdot \underbrace{Q(s_t,a_t)}_{\text{Sena reikšmė}} + \alpha \cdot \underbrace{(r_t+\gamma\cdot \max_a Q(s_{t+1},a))}_{\text{Tikslas = nauja } q_\pi(s,a) \text{ reikšmė}} \tag{37}$$

Šis pavidalas yra nekas kitas, o daugybę kartu nagrinėta slenkančio vidurkio formulė, kur $[R+\gamma \max_a Q(S',a)]$ yra naujausia gauta $q_\pi(s,a)$ reikšmė.

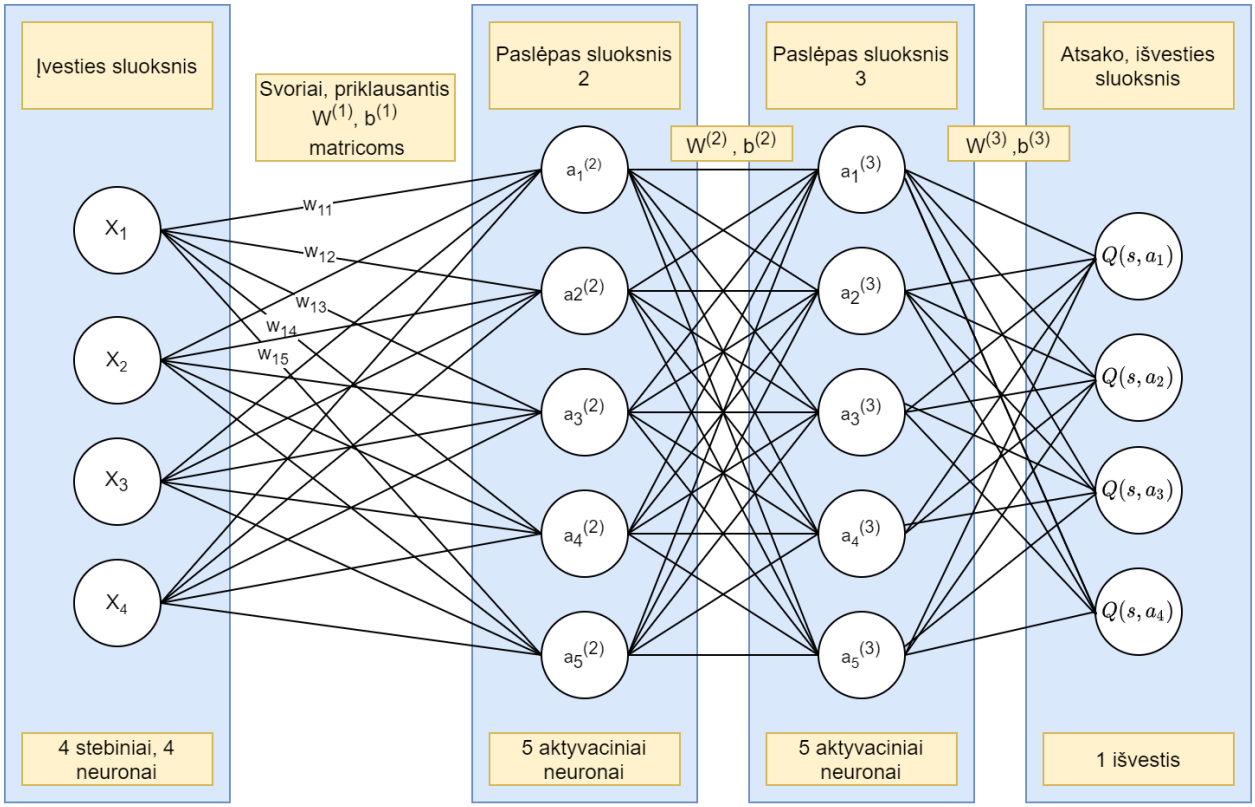
Taigi neuroninis tinklas stengiasi aproksimuoti tikslą, kuris geriau žinomas kaip G_t arba naujausias aproksimuotas $q_\pi(s,a)$ narys. Tuo tarpu neuroninio tinklo įvestis yra būsenos vektorius s. Kadangi naudojamas neuroninis tinklas, vektoriaus s dydis nėra svarbus, konvergacija išlieka sparti.

Optimizavimo procesas susideda iš šių trijų dalių:

1. sukuriama duomenų talpykla bei du neuroniniai tinklai - taikinys bei pagrindinis;
2. pasirenkame veiksmą vadovaujantis godžiuoju ϵ veiksmų taisyklių metodu;
3. svorių atnaujinimas naudojantis Bellmano taisykle, pateikta 37.

Patirčių talpyklos bei dviejų neuroninių tinklų sukūrimas. Agentas tyrinėdamas aplinką, kaupia duomenis patirčių talpykloje, dar kitaip vadinama buferiu. Ši patirčių talpykla yra agento praeities patirtys, kurios padeda optimizuoti neuroninius tinklus. Joje

kaupiamos (s,a,r,s') duomenų eilutės, kurios vos ne identiškos SARSA modelio duomenims. Panašumas ties tuo nesibaigia, modelis taip pat kaupia duomenis tokiu pačiu principu kaip SARSA modelis, Monte Carlo būdu simuliuojant žingsnius. Esant būsenoje s pagrindinis neuroninis tinklas aproksimuoja s būsenos visų a veiksmų $q_\pi(s, a)$ vertes, kaip tai matyti 16 paveikslėlyje ir pasirenka judėti ta trajektorija, kurios $q_\pi(s, a)$ vertė yra didžiausia.



16 pav. Giliojo Q neuroninio tinklo struktūra

Atlikus veiksmą, mes apskaičiuojame gautą apdovanojimą r bei naują būseną s', kurioje atsiradome atlikus veiksmą a. Tokiu būdu turim vieną duomenų eilutę, kurią išsaugome į patirčių talpyklą. Sukaupus pakankamai daug duomenų galime juos pradėti naudoti neuroninio tinklo apmokymui. Apmokant mes pasitelkiame tik mažą dalį duomenų atsitiktinai ištrauktų iš talpyklos. Svarbu neuroninį tinklą apmokyti tik su duomenimis iš patirčių talpyklos, norint išvengti koreliacijos, tarp veiksmų, nutinkančių vienas po kito. Šią koreliacijos tendenciją lengva pastebėti intuityviai išanalizavus patirčių talpyklos veikimo principus.

1. Atsitiktinės duomenų eilutės (s,a,r,s') paimamos iš patirčių talpyklos. Dėl atsitiktinio ėmimo, didžioji dalis patirčių - duomenų eilučių, yra iš skirtingų trajektorijų, epizodų.
2. Su paimtais duomenimis apmokome modelį. Šio proceso metu, dėl atsitiktinio ėmimo, skirtingos būsenos įgyją naujas $q_\pi(s, a)$ aproksimacijas.

Tuo tarpu be patirčių talpyklos, mes apmokytume neuroninį tinklą kiekvieno žingsnio metu. Tačiau kadangi nauji agento žingsniai priklauso nuo senesnių, mūsų duomenis taptų

priklausomi. Tai pažeidžia stochastinio gradientinio nuolydžio vieną iš pagrindinių reikalavimų - imties stebiniai turi būti nepriklausomi. Tai nutikus modelis pradėtų dažniau rinktis pasikartojančias trajektorijas, diverguotų į ne optimalias trajektorijas. Patirties talpykla taip pat turi dar viena papildoma svarbų bruožą - pakartotinas apmokymas su tais pačiais duomenimis. Kadangi duomenis yra išsaugojami talpykloje, mes turime ne mažą tikimybę apmokyti neuroninį tinklą su tuo pačiu dėmeniu ne vieną kartą. Šis bruožas naudingas apmokyti neuroninį tinklą su retai pasitaikančiomis reikšmėmis. Patirčių talpykla dažniausiai turi fiksuotą dydį. Talpykla yra naudojama visuose algoritmuose, kurie naudoja neuroninių tinklų aproksimacijas.

Patirčių talpykla padeda apmokyti pagrindinį neuroninį tinklą, kurio tikslo funkcija yra pateikta žemiau:

$$L(\theta) = \left[\left(\left(r + \gamma \max_{at+1} Q\left(s_{t+1}, a_{t+1}; \theta^{\text{taikinys}}\right)\right) - Q\left(s, a; \theta^{\text{pagrindinis}}\right)\right)^2\right]. \tag{38}$$

Ši funkcija panaši į mažiausių kvadratų optimizavimo tikslo funkciją. Šioje funkcijoje neuroniniai tinklai stengiasi sumažinti paklaidą tarp tikrosios $q_\pi(s, a)$ vertės bei neuroninio tinklo pateiktos aproksimacijos. Pateikti θ parametrai atkeliauja iš taikinio bei pagrindinio neuroninio tinklo. Abu neuroniniai tinklai aproksimuoja $q_\pi(s, a)$. Tačiau taikinio neuroninio tinklo parametrai nėra apmokami, jie kas tam tikra agento žingsnių kieki yra periodiškai paimami iš pagrindinio neuroninio tinklo. Tuo tarpu pagrindinio neuroninio tinklo parametrai yra atnaujinami kas kiekviena žingsnį. Ši metodologija buvo pirma karta pritaikyta Google Deepmind komandos, kuriant Atari agentus [7]. Jos tikslas yra sumažinti mokymosi divergacijas bei nestabilias $q_\pi(s, a)$ verčių svyravimus. Šios nepalankios savybės atsiranda iš neuroninių tinklų generalizacijos. Atnaujinus $q_\pi(s, a)$ reikšmes ir pakeitus θ parametrus, kartu pasikeičia kaimyninės $q_\pi(s', a)$ reikšmės, kitaip tariant mūsų taikinio reikšmės, jei naudotume viena neuroninį tinklą. Nuolat kintančios taikinio reikšmės veda prie nestabilaus konvergavimo, mūsų aproksimuojamos $q_\pi(s, a)$ reikšmės artėja link norimų $q_\pi(s, a)$ reikšmių, kurios dėl kiekvieno atnaujinimo kinta. Todėl yra sukuriamas antras neuroninis tinklas su fiksuotais parametrais, kurie stabilizuoja mokymosi procesą.

Veiksmų pasirinkimas. Gilusis Q naudoja godųjį-ϵ metodą veiksmų pasirinkimui, kur su $1 - \epsilon$ tikimybe pasirinksime veiksmą vedanti į didžiausią $q_\pi(s, a)$. Šias $q_\pi(s, a)$ vertes apskaičiuoja neuroninis tinklas, kaip tai pateikta 16 paveikslėlyje. Tačiau su ϵ tikimybe mes pasirinksime neoptimalią vertę. Tai padeda aplankyti naujas trajektorijas, tikintis surasti optimalesnius veiksmus.

Svorių atnaujinimas. Atlikus veiksmo pasirinkimą bei išsaugojus duomenis susijusius su šiuo veiksmu mes įgyvendiname viena agento žingsnį. Po kiekvieno žingsnio mes apmokome agentą tiksliau generalizuoti $q_\pi(s, a)$ reikšmes. Mokymo procese pasitelkiame patirčių tal-

pyklą, kaip tai jau buvo aptarta 1.7.10 skyriuje. Mokymosi procesas gali trukti porą dienų, priklausomai nuo nagrinėjamos aplinkos sunkumo.

Nors Q-mokymasi procesas pademonstravo skatinamojo mokslo potencialą, šis metodas nebetaikomas. Tačiau svarbios pademonstruotos idėjos kaip $q_\pi(s, a)$ verčių aproksimavimas neuroniniais tinklais, patirčių talpykla, dviejų neuroninių tinklų stabilizavimas yra pamatas aktoriaus-kritiko šablonui, kuriuo vadovaujasi visi moderniausi modeliai. Norint suprasti aktoriaus-kritiko šabloną, aptariama sekančiame skyriuje, pradžioje reik įsigilinti į vieną svarbiausių skatinamojo mokslo teoremų - veiksmo taisyklių gradientas.

1.7.11. Veiksmo taisyklių gradientas bei aktorius-kritikas

Veiksmo taisyklių parametrizavimas bei šios parametrizuotos funkcijos gradientas yra viena iš svarbiausių šio mokslo atradimų. Ilgą laiką buvo manoma, kad veiksmo taisyklių gradientas neegzistuoja, todėl buvo taikomi Giliojo-Q mokymosi algoritmai, kurie optimizuoja $q_\pi(s, a)$ vertes ir neliečia veiksmo taisyklių optimizavimo. Tačiau kaip matome 14 paveikslėlyje, optimizavimo procesas įtraukia nuolatinį verčių bei veiksmo taisyklių iteravimą. Atradus veiksmo taisyklių gradientą, pagaliau tapo įmanoma įgyvendinti 14 ciklą ir tai pasiekus gimė daugybė galingu modelių kaip PPO, DDPG, TD3, SAC.

Veiksmo taisyklių tikslas yra skatinti tuos veiksmus, kurie vedė ties veiksmų trajektorijomis suteikiančiomis aukštus apdovanojimus ir slopinti veiksmus, kurių apdovanojimų trajektorijos yra mažos. Formaliai tai mes užrašome šiuo pavidalu, nurodančiu tikėtiną veiksmų taisyklių apdovanojimą:

$$J\left(\pi_\theta\right) = \underset{\tau\sim\pi_\theta}{\mathrm{E}}[R(\tau)], \tag{39}$$

kur $R(\tau)$ yra gautų epizodo, trukusio T žingsnių, apdovanojimų suma:

$$R(\tau) = \sum_{t=0}^{T} R_t. \tag{40}$$

$J\left(\pi_\theta\right)$ yra tikėtinas apdovanojimas vadovaujantis π veiksmo taisyklėmis, aproksimuotomis neuroninio tinklo θ parametrais.

Tikėtiną veiksmo taisyklių apdovanojimą mes išreiškiame per tikimybę įgyti šiuos apdovanojimus:

$$P(\tau \mid \theta) = \rho_0\left(s_0\right) \prod_{t=0}^{T} P\left(s_{t+1} \mid s_t, a_t\right) \pi_\theta\left(a_t \mid s_t\right) \tag{41}$$

$P(\tau \mid \theta)$ yra trajektorijos tikimybė, kur $\rho_0\left(s_0\right)$ yra tikimybė, kad pradėsime s_0 būsenoje. Šią tikimybę sudauginus su apdovanojimų kiekiu, gauname galutinę $J(\pi)$ išraišką:

$$J(\pi) = \int_\tau P(\tau \mid \pi) R(\tau) = \underset{\tau\sim\pi}{\mathrm{E}}[R(\tau)] \tag{42}$$

Mūsų tikslas yra optimizuoti veiksmų taisyklių $J(\pi)$ parametrus taip, kad jie gražintu kuo didesnį epizodų apdovanojimus:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J\left(\pi_\theta\right)\big|_{\theta_k} \tag{43}$$

Deja mes negalime optimizuoti $J(\pi)$ tiesiogiai, kadangi neturint modelio, mes negalime gauti 41 formulės reikšmes, reikalingas apskaičiuoti gradientą. Todėl imamės analitinių išraiškų, padedančių pakeisti gradiento formulę į naują formulę, kurios rezultatas nors ir nesutampa su $J(\pi)$ bet sutampa jo gradientas. Šios formulės išvedimui buvo pasitelkti šie faktai:

1. Log išvestinė:
$$\nabla_\theta P(\tau \mid \theta) = P(\tau \mid \theta) \nabla_\theta \log P(\tau \mid \theta). \tag{44}$$
2. Log trajektorijos tikimybė:
$$\log P(\tau \mid \theta) = \log \rho_0\left(s_0\right) + \sum_{t=0}^{T} \left(\log P\left(s_{t+1} \mid s_t, a_t\right) + \log \pi_\theta\left(a_t \mid s_t\right)\right). \tag{45}$$
3. Apskaičiuojant gradientinį nuolydį, funkcijos, kurios nepriklauso nuo θ, pranyksta. Kitaip tariant $\rho_0\left(s_0\right), P\left(s_{t+1} \mid s_t, a_t\right)$, ir $R(\tau)$ yra lygus nuliui.
4. Šios pranykusios funkcijos supaprastina log trajektorijos tikimybe:
$$\begin{aligned} \nabla_\theta \log P(\tau \mid \theta) &= \nabla_\theta \log_0\left(s_0\right) + \sum_{t=0}^{T} \left(\nabla_\theta \log P\left(s_{t+1} \mid s_t, a_t\right) + \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right)\right) \\ &= \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right). \end{aligned} \tag{46}$$

Pasitelkus šiuos išvedimus galime išsivesti veiksmo taisyklių gradientą:

$$\begin{aligned} \nabla_\theta J\left(\pi_\theta\right) &= \nabla_{\theta_{\tau \sim \pi_\theta}} \underset{\tau \sim \pi_\theta}{\mathrm{E}}[R(\tau)] \\ &= \nabla_\theta \textstyle\int_\tau P(\tau \mid \theta) R(\tau) \\ &= \textstyle\int_\tau \nabla_\theta P(\tau \mid \theta) R(\tau) && \text{Log išvestinės faktas} \\ &= \textstyle\int_\tau P(\tau \mid \theta) \nabla_\theta \log P(\tau \mid \theta) R(\tau) \\ &= \underset{\tau \sim \pi_\theta}{\mathrm{E}}\left[\nabla_\theta \log P(\tau \mid \theta) R(\tau)\right] \\ \therefore \nabla_\theta J\left(\pi_\theta\right) &= \underset{\tau \sim \pi_\theta}{\mathrm{E}}\left[\textstyle\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right) R(\tau)\right] && \text{4 fakto pritaikymas} \end{aligned} \tag{47}$$

Turint šią formą mes stochastiniu būdu galime pradėti aproksimuoti gradientinį pakilimą. Šio proceso metu mes generuojame naujas imtis bei apskaičiuojame imties vidurkį:

$$\hat{g} = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right) R(\tau), \tag{48}$$

kur $\hat{g}$ yra veiksmo taisyklių gradiento vektorius, o $\mathcal{D}$ yra trajektorijų (imčių) kiekis. Tokiu būdu mes galime labai tiksliai aproksimuoti mūsų veiksmo taisyklių parametrus, vedančius ties didžiausiais apdovanojimais.

Pats metodas savaime nėra efektyvus tačiau sukombinavus su Giliuoju-Q mokymusi mes grįžtame į pagrindinį 14 metodą, kuris buvo taikomas be aproksimacijų. Šį procesą įgyvendina aktoriaus kritiko metodas. Jo pavadinimas kilo iš intuityvios idėjos, kuria galime pastebėti 48 formulėje. Jei $R(\tau)$ yra teigiamas, pateiktas veiksmas paskatinamas, jei $R(\tau)$ neigiamas, pateiktas veiksmas slopinamas. $R(\tau)$ atlieka kritiko rolę, skatina arba slopina veiksmų taisyklių neuroninio tinklo, vadinamo aktoriumi, pateiktas reikšmes. Tačiau aktoriaus-kritiko metoduose mes vietoj tiesioginės $R(\tau)$ reikšmės, pasitelkiame giliojo-Q neuroninį tinklą kurio $q_\pi(s,a)$ reikšmės aproksimuoja $R(\tau)$. Taigi gilusis-Q pateikia $q_\pi(s,a)$ vertes bei jas optimizuoja, kaip tai jau matėme 1.7.10 skyriuje. Taip pat naudojantis $q_\pi(s,a)$ reikšmėmis mes kartu atnaujiname veiksmo taisyklių neuroninį tinklą naudojantis 48 formule. Šis metodas bei jo pseudokodas bus toliau plėtojamas tolimesniuose modeliuose, kurie remiasi jo struktūra.

1.7.12. DPG bei DDPG

DPG[8] metodas yra paremtas aktoriaus-kritiko metodu. Jis sujungia Gilusis-Q metodą su veiksmo taisyklių gradientu. Tačiau šiuo atveju veiksmo taisyklių tikslo formulė įgauna skirtingą pavidalą:

$$J(\theta) = \int_{\mathcal{S}} \rho^{\mu}(s) Q\left(s, \mu_\theta(s)\right) ds, \tag{49}$$

kur $\mu_\theta(s)$ yra deterministinės funkcijos $\pi_\theta(a|s)$ atitikmuo. Kitaip tariant veiksmai parenkami ne iš veiksmų tikimybių pasiskirstymo, o optimaliausias veiksmas yra iškarto funkcijos aproksimuojamas - $a = \mu(s)$. Tuo tarpu $\rho^{\mu}(s)$ žymi tikimybę, kad pradėsime s būsenoje, o $Q\left(s, \mu_\theta(s)\right)$ yra neparametrizuoto $q_\pi(s,a)$ atitikmuo. Veiksmų aproksimacijos padeda praplėsti veiksmų dimensiją iš keleto pasirinkimų, kaip tai matėme 15 pavyzdyje, turint tik keturis galimus veiksmus, į begale. Vienas iš praktinių šio metodo bruožų yra jo gradiento apskaičiavimas:

$$\begin{aligned} \nabla_\theta J(\theta) &= \int_{\mathcal{S}} \rho^{\mu}(s) \nabla_a Q^{\mu}(s,a) \nabla_\theta \mu_\theta(s) \Big|_{a=\mu_\theta(s)} ds \\ &= \mathbb{E}_{s\sim\rho^{\mu}} \left[\nabla_a Q^{\mu}(s,a) \nabla_\theta \mu_\theta(s) \big|_{a=\mu_\theta(s)} \right] \end{aligned} \tag{50}$$

Telieka simuliuoti agento žingsnius ir apskaičiuoti veiksmo taisyklių gradientą, kaip tai matėme 48 formulėje. Praktikoje gradiento apskaičiavimas yra keblesnis, nei 47 metodo, kadangi reikia apskaičiuoti dviejų neuroninių tinklų gradientus bei juos sudauginti. Šis procesas bus giliau aptariamas praktikoje.

Šio modelio didžiausias trūkumas yra neegzistuojantis agento aplinkos tyrinėjimas. Modelis visados renkasi optimalias veiksmų taisykles, netyrinėja aplinkos. DDPG[9] modelis pagerina DPG modelį pridėdamas triukšmą prie veiksmų taisyklių pateiktų reikšmių:

$$\mu'(s) = \mu_\theta(s) + \mathcal{N}, \tag{51}$$

kur $\mathcal{N}$ yra normalusis pasiskirstymas. Jo parametrai yra pasirenkami priklausomai nuo nagrinėjamos problemos. Taigi pridėjus prie aproksimuotos veiksmo reikšmės $\mu_\theta(s) \in \mathcal{R}$ mažą triukšmą, mes skatiname modelį rinktis įvaresnius veiksmus, leidžiančius surasti optimalesnes veiksmų trajektorijas.

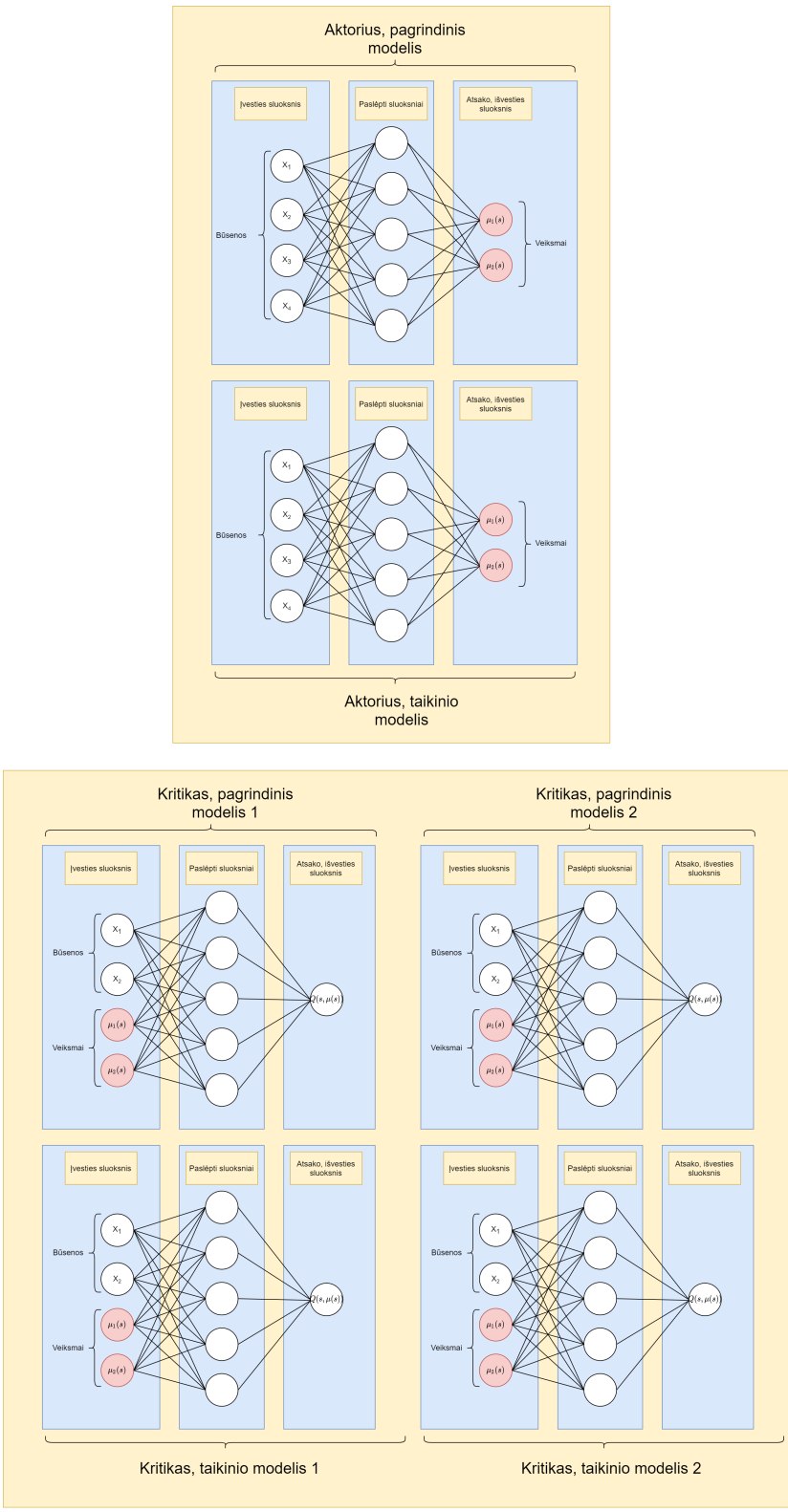
1.7.13. TD3

TD3[10] modelis yra vos ne identiškas DDPG modeliui tačiau turi pora skirtumu:

1. Du gilieji-Q neuroniniai tinklai. Kadangi gilusis-Q neuroninis tinklas dažnai pervertina reikšmes, mes apmokymui naudojame du giliuosius-Q neuroninius tinklus ir apmokome su mažesne gauta reikšme.
2. Vėluojantis veiksmo taisyklių neuroninio parametrų atnaujinimas. Jei DPG atnaujindavo neuroninį tinklą kas kiekvieną agento žingsnį TD3 autorius rekomenduoja atnaujinimą atlikti kas du žingsnius.
3. Reguliarizacija, pateikta 52 formulėje. Ji padeda išvengti persimokymo, kuomet veiksmų taisyklės per greit prisitaiko prie aukštų gilaus-Q reikšmių. Pagrindinė idėja yra pridėti maža triukšmą prie veiksmo taisyklių neuroninio tinklo pateiktos reikšmės. Tokiu būdu neuroninis tinklas bus apmokytas geriau aproksimuoti kaimynines veiksmų reikšmes.

$$a'\left(s'\right) = \text{clip}\left(\mu_{\theta_{\text{taikinio}}}\left(s'\right) + \text{clip}(\epsilon, -c, c), a_{\min}, a_{max}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma). \tag{52}$$

TD3 struktūra susideda iš šešių neuroninių tinklų, kaip tai matosi 17 paveikslėlyje. Jie reikalingi tiktais apmokymo metu. Apmokius neuroninį tinklą užtenka vieno veiksmų taisyklių neuroninio tinklo, kuris pateiktų optimalius veiksmus agentui.



17 pav. TD3 modelio struktūra

2. Praktinė dalis

Praktinėje dalyje bus įgyvendinamas TD3 metodas. Jo įgyvendinimui bus pasitelkta „Unity“ aplinka, kurioje egzistuos apmokamas agentas. Agentas yra keturių galūnių voras, kurio tikslas yra nuropoti į užsibrėžtą poziciją, judinant savo galūnes. Šis agento tikslas bei aplinka bus plėtojama pirmuosiuose praktinės dalies skyriuose. Toliau bus gilinamasi ties komunikacijos sluoksniu, kuriame simuliacijos aplinka bei mokymosi metodas TD3 keičiasi duomenimis. Kadangi TD3 metodas bus įgyvendintas „Python“ aplinkoje, o aplinkos simuliacija „Unity“ aplinkoje, šios dvi aplinkos negali vienas kitai persiųsti bei gauti duomenis. Todėl yra būtinas komunikacijos sluoksnis tarp šių dviejų aplinkų, padedančiu išsiųsti bei gauti žinutes. Galiausiai bus aptartas TD3 įgyvendinamas metodas bei jo rezultatai.

2.1. Agentas, jo tikslai bei aplinka

Per pastarąjį dešimtmeti skatinamieji metodai buvo pradėti taikyti įvairiuose srityse, tačiau viena iš sričių pasižymėjusiu didžiausiu proveržiu - robotų judėjimo kontroleriai. Roboto judėjimo užduotis tradiciniais metodais buvo ne vieną kartą stengiamasi išspręsti, bet rezultatai buvo nepatenkinami. Tačiau skatinamojo mokslo metodai pademonstravo, kad jie yra puikiai pritaikyti kontroliuoti agento galūnes, surasti optimalias judėjimo trajektorijas vedančias į užsibrėžto tikslo pasiekimą. Būtent šią užduoti ir įgyvendinsiu praktinėje dalyje. Bus pasitelktas „Unity“ aplinkoje sukurtas voras, kurio tikslas nuropoti iki specifinės koordinatės, kaip tai matome 8 paveikslėlyje. Pagrindinis užduoties sunkumas kyla iš fizikinio judėjimo. Labai dažnai judėjimo agentai „Unity“ aplinkoje nejuda pagal fizikinius dėsnius, o paprasčiausiai animacijų pagalba jų judesiai yra įrašyti ir atkartojami. Tokios animacijos nėra paremtos fizika, kiekviena agento galūnė kiekvieno kadro metu įgyja įrašytas pozicijų koordinatės. Toks judėjimas neprisitaiko prie kintančios, dinaminės aplinkos ir roboto galūnės dažnai kerta objektus kiaurai. Šias problemas išsprendžia fizika paremtas galūnių judinimas, kurio programiškai iki šiol įgyvendinti buvo neįmanoma. Tik neseniai pasitelkus statistiniais mokymosi metodais kaip TD3 buvo pasiektas fizikinis agentų judėjimas.

Taigi kiekvieno epizodo metu voras yra sukuriamas aukštai ore. Jis turi sugebėti nusileisti ant savo galūnių bei nuropoti ties žaliu kubeliu. Sėkmingai jį pasiekus epizodas nesibaigę, yra sukuriamas naujas žalias kubelis, kurį voras vėl turi pasiekti. Šis būdas skatina vorą aplankyti naujas verčių būsenas susijusias su voro apsisukinėjimu. Epizodas baigiasi kuomet pasiekiamas epizodo maksimalų žingsnį, kurį galime traktuoti kaip laiko intervalą. Taip pat epizodas gali baigtis anksčiau, jei yra pasiekta terminuojanti būsena. Voro aplinkos atveju tai nutinka jei pagrindinė kūno dalis arba viršutiniai sąnariai pasiekę žemę. Kuo ilgiau voras išgyveną bei kuo arčiau jis yra žalio kubelio, tuo agentas gauna didesnį epizodo kaupiamąjį apdovanojimą. Tai ir yra agento pagrindinis tikslas, gauti kuo didesnį apdovanojimą.

2.2. Mokymosi seka

Voro agento mokymosi procesas yra identiškas aptartiems mokymosi metodams ir sudarytas iš šių iteracinių žingsnių:

1. TD3 modelis turi dabartines voro būsenas. Turint šias būsenas TD3 pateikia naujus veiksmus.
2. Voras gauną TD3 modelio pateiktą veiksmą bei jį atlieką.
3. Atlikus veiksmą gaunama nauja būsena bei apdovanojimas. Šios vertės nusiunčiamos TD3 modeliui.
4. TD3 gavęs pateikto veiksmo rezultatą - naują būsena bei apdovanojimą, pradeda modelio apmokymą. Tai atlikus naujos būsenos tampa dabartinėmis ir pradedama naują iteracija. Dažniausiai apmokymas būna atidedamas ir jis įvyksta pasibaigus agento epizodui arba paralelizuotams agentams atlikus tam tikrą kiekį žingsnių.

Svarbu pastebėti, kad apdovanojimų bei būsenų duomenys nusiunčiami TD3 modeliui ne kiekvieno simuliacijos kadro metu, o tik kas penkis „Unity“ fizikos simuliacijos žingsnius. Kitaip tariant agentui atlikus veiksmus jis laukia penkis „Unity“ fizikos atnaujinimus, per kuriuos jo galūnės pakeičią poziciją penkis kartus po labai maža atstumą. Tai atlikus agentas tik tada užfiksuoja naują įgytą būseną. Per šiuos penkis žingsnius jo apdovanojimas yra kaupiamas - kas žingsnį pridedame po apskaičiuota apdovanojimą. Šis penkių kadrų atnaujinimo atidėjimas padeda pasiekti ryškesni naujo veiksmo ir jo pasekmės ryšį. Taip pat tai sumažiną duomenų kiekį bei pagreitiną mokymą.

2.3. Aplinka

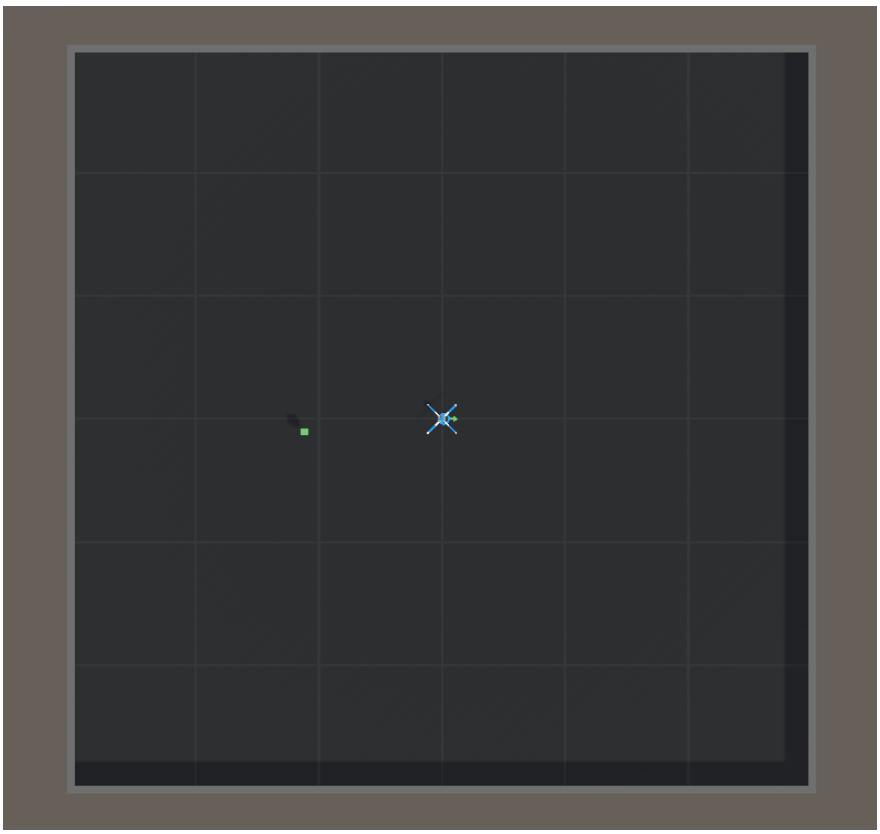
Voro aplinka susideda iš trijų elementų:

1. Voras.
2. Tuščia, lygi arena, su keturiomis, areną apsupančioms, sienomis.
3. Žalias kubelis, iki kurio vorui reikia nuropoti.

Aplinkos atvaizdavimą matome 18 paveikslėlyje. Joje agentas praleidžia visą savo mokymosi laiką. Aplinka bei pats fizikiniais dėsniais paremtas voras yra jau sukurtas „Unity“ programuotojų. Jie pasitelkė skatinamojo mokslo PPO metodus apmokyti vorą ropoti bei pasiekė įspūdingus rezultatus. Mano mokymo metu bus naudojamas TD3 metodas, kuris yra specifiškai pritaikytas deterministiniams veiksmams bei panašiuose užduotyse dažniau pasiekia optimalesnius rezultatus.

Arena įgijo šį specifini išsidėstymą, dėl šių tolimesnių priežasčių:

1. Arena lengva padvigubinti, sukurti daugybe jos kopijų. Tai naudinga paraleliam mokymuisi, kuris bus aptartas tolimesniuose skyriuose.



18 pav. Voro aplinka

2. Lygioje arenoje, be įdubimu bei iškilimu, yra lengviau apmokyti agentą. Nereikia programuoti sudėtingų metodų įdubimams aptikti bei informuoti agentą apie jų egzistavimą. Tai sumažina būsenų erdvę.
3. Žalias kubelis, iki kurio vorui reikia nuropoti. Tai yra paprasčiausias atskaitos taškas, naudojamas įvertinti agento optimalumą.
4. Keturios sienos, neleidžiančio vorui nukristi už simuliacijos ribų.

Taigi ši arena yra paprasčiausia įmanoma aplinka, kuri padeda įsitikinti mokymosi metodo efektyvumu. Jei pasirinktas mokymosi metodas sugeba susidoroti su šia aplinka, ją galima pradėti palaipsniui sunkinti bei pritaikyti savo reikmėms.

2.4. Duomenys

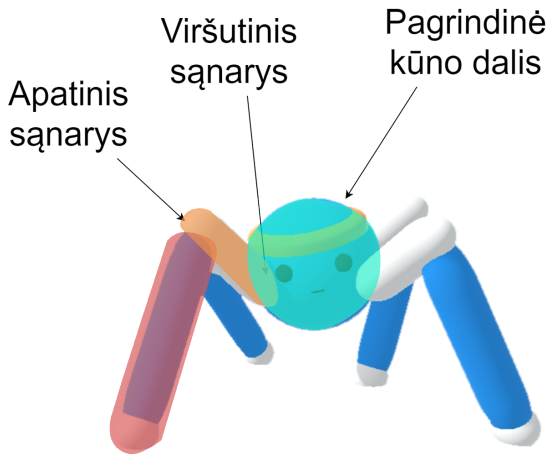
Kaip matome 17 paveikslėlyje, TD3 mokymosi metodų neuroninių tinklų įvestis yra būsenų erdvė bei aproksimuoti veiksmai. Tuo tarpu pagrindinė išvestis yra veiksmų taisyklės - kokius tolimesnius veiksmus turėtų atlikti agentas. Šias veiksmų reikšmes padeda apskaičiuoti apdovanojimai, taip pat viena iš įvesčių reikšmių, apskaičiuojanti taikinio reikšmes, kurias neuroninis tinklas turėtų aproksimuoti. Taigi toliau visi šie duomenys bus aptariami detaliau.

2.4.1. Įvestis - būsenos

Tiek aktoriaus modelis, aproksimuojantis veiksmų taisykles, tiek kritiko modelis, aproksimuojantis būsenų vertes, priima būsenų reikšmes. Voro būsenų reikšmės susideda iš 127 reikšmių. Jas galima suskaldyti į šias kategorijas:

1. Voro galūnės. Kiekviena voro koja susideda iš dviejų sąnarių: viršutinio bei apatinio sąnario.
2. Voro pagrindinis kūnas. Viršutiniai voro sąnariai susijungia su šia dalimi.

Kaip matomo 19 paveikslėlyje apatinis sąnarys kontroliuoja koją, užtušuota rausvai. Viršutinis sąnarys kontroliuoja koją užtušuota oranžine spalva. Ir galiausiai visi viršutiniai sąnariai susijungia į pagrindinę kūno dalį užtušuota žalsvu atspalviu.



19 pav. Voro galūnės.

Visos voro galūnės kaupia ne konkrečią informaciją, bet santykinę. Kitaip tariant nėra kaupiama tiksli voro pozicija, bet atstumo vektorius nuo voro iki taikinio, žalio kubelio. Tokiu būdu voras neprisitaiko prie konkrečios aplinkos, agentas gali lengvai ropoti įvairiose arenose. Tai ypatingai svarbu paralelizuojant mokymosi procesą. Taigi visi sąnariai kaupia šias būsenas:

1. Ar galūnė liečia žemės paviršių. Priklausomai nuo to ar liečia, simuliacija gali būti nutraukiama ir įvardijama kaip nepasisekusi. Ši sąlyga galioja viršutiniam sąnariui, kadangi nenorime, kad jis liestų žemę, jis nuolatos turi būti virš jos.
2. Atstumo vektorius tarp sąnario greičio vektoriaus bei agento-kubelio atstumo vektoriaus. Mes norime, kad atstumas būtų lygus nuliui, kitaip tariant abu vektoriai žiūri į tą pačią trajektoriją.
3. Laipsnių atstumas tarp sąnario pozicijos vektoriaus bei atstumo vektoriaus tarp agento bei kubelio. Turi panašumų su praeitu punktu, tačiau duomenys išreikšti laipsniais.

Pagrindinė kūno dalis kaupia šią informaciją:

1. Ar liečia žemės paviršių. Jei taip, simuliacija yra nutraukiama.
2. Pagrindinės kūno dalies atstumas iki žemės paviršiaus. Kadangi agentas turi sugebėti stovėti su savo kojomis ant žemės paviršiaus, voras turi žinoti kaip arti jis yra iki žemės.
3. Atstumo bei laipsnių vektoriai, gaunami tokiu pat principu kaip sąnariuose.

Taip pat kaupiamas vidutinis agento galūnių greitis. Kuo didesnis greitis, tuo jis gauna didesnį apdovanojimą. Tokiu būdu mes skatiname agentą judėti sparčiai.

Šios visos būsenos supakuojamos į vieną realiųjų skaičių masyvą ir nusiunčiamos TD3 mokymosi modeliui. Taip pat nusiunčiame apdovanojimą, kuris yra apskaičiuojamas „Unity“ aplinkoje.

2.4.2. Apdovanojimas

Duomenų talpykloje kaupiamos ne tik būsenos bet ir apdovanojimai. Jie atkeliauja iš „Unity“ aplinkos. Vorui atlikus TD3 modelio pateiktus veiksmus, jis apskaičiuoja atliktų veiksmų naudą. Vadovaujantis 1.7.4 skyriaus svarbiu patarimu, mes turime suformuoti apdovanojimo funkciją taip, kad jį neformuotų užduoties sprendimo būdo, o tik apdovanotų už pasiektą tikslą. Todėl apdovanojimo funkcija susideda iš šių trijų funkcijų:

1. Skaliarinė sandauga tarp pagrindinės kūno dalies greičio vektoriaus bei normalizuoto krypties vektoriaus nuo pagrindinio kūno į žalią kubelį. Šie vektoriai pavaizduoti 9 paveikslėlyje, tačiau pateiktas žalias vektorius nėra vizualiai normalizuotas. Jei šie vektoriai sutampa, mes gauname vieneto vertę, žyminčią didžiausią įmanomą apdovanojimą. Tokiu būdu mes nurodome agentui judėti link žalio kubelio, bet nepasakome kaip konkrečiai atlikti šią užduoti.
2. Skaliarinė sandauga tarp agento krypties vektoriaus bei normalizuoto krypties vektoriaus nuo pagrindinio kūno į žalią kubelį. Kitaip tariant jei agentas yra atsisukęs į žalią kubelį, jis gauną maksimalų apdovanojimą - vienetą.
3. Apskaičiuojame ar vidutinis visų galūnių greitis atitinka mūsų pateiktą maksimalų greitį. Jei atitinką, priskiriame papildoma vieneto apdovanojimą, jei visai nesutampa, apdovanojimas tampa nulis. Tarpinės reikšmes pateikę sigmoido funkcija. Ši apdovanojimo funkcija buvo pridėta vėliausiai iš visų apdovanojimo funkcijų, kuomet buvo pastebėtas voro lėtas slinkimas į žaliąjį kubelį. Agentas išmokęs stabiliai stovėti ant kojų ir iš lėto ropoti link žaliojo kubelio po truputi pastebi, kad kuo greičiau juda, tuo didesnį pasiekę apdovanojimą.

Šie apdovanojimai, kiekvienos fizikos atnaujinimo metu yra apskaičiuojami bei sudedami į vieną kaupiamąjį apdovanojimą. Jis, kartu su būsenomis, nusiunčiamas TD3 modeliui. Nusiuntus apdovanojimą jis tampa nuliu ir jo skaičiavimo procesas prasideda iš naujo.

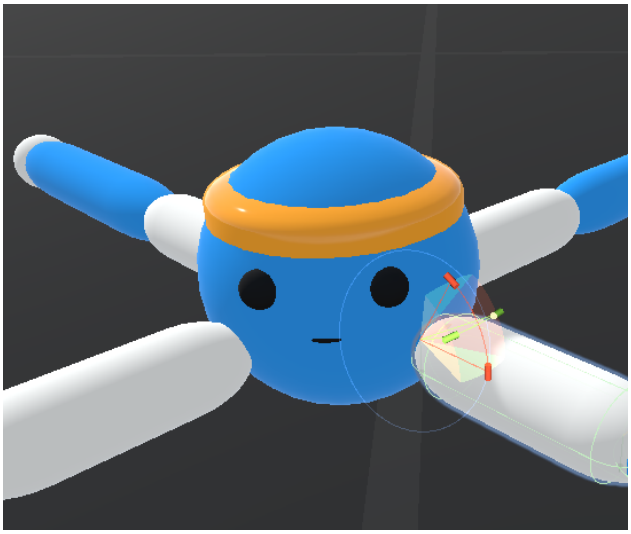
Kiekvieno epizodo metu mes kaupiame epizodo galutinį apdovanojimą. Jis slenkančio vidurkio metodu apskaičiuoja paskutinių penkiasdešimties epizodų apdovanojimo vidurkį. Šis kiekvieno epizodo metu kintantis matas yra pagrindinis kriterijus pagal kurį vizualiai vertiname modelio efektyvumą. Mokymuisi vykstant jis turi kilti. Tai parodo vis didėjanti epizodų apdovanojimą.

2.4.3. Išvestis - veiksmai

TD3 aktoriaus neuroninis tinklas priėmęs būsenas išveda veiksmus. Iš viso yra apskaičiuojami dvidešimt veiksmo verčių, kuriuos voras gauną. Šios veiksmų vertės kontroliuoja voro viršutines bei apatines galūnes. Toliau bus vardijama galūnių veiksmų vertės:

1. Viršutinis voro sąnarys gali sukinėtis aplink x bei y koordinačių ašį, tačiau ne z, kaip tai matome 20 paveikslėlyje. Veiksmų vertės nurodo galutinę šio sąnario poziciją laipsniais bei kiek jėgos bus suteikta šiai pozicijai pasiekti.
2. Apatinio sąnario veiksmų vertės yra identiškos, tačiau šis sąnarys gali sukiotis tik aplink x ašį.

Kadangi TD3 aktoriaus neuroninio tinklo išvestis yra tanh funkcija, mes nuolatos turime konvertuoti $[-1,1]$ intervalo reikšmes į konkrečias laipsnių bei jėgų reikšmes. Taigi agentas gauna naujas galūnių judėjimo reikšmes ir tokiu būdu jis jas po truputi krutina.



20 pav. Viršutinio sąnario judėjimo trajektorijos.

2.5. Komunikacijos sluoksnis

TD3 modelis bei voro aplinka yra įgyvendinta skirtinguose programose. Voro aplinka yra įgyvendinta „Unity“ programoje, o TD3 modelis „Python“ aplinkoje. Kadangi „Unity“ variklis negali kompiliuoti „Python“ kodo, teko įgyvendinti šias dvi aplinkas atskirose

programose. Tai sukėlė pastovių keblumų, kadangi duomenys tarp šių dviejų programų turi nuolatos bei užtikrintai keistis, kas dažnai neįvykdavo. Taigi buvo sukurtas komunikacinis sluoksnis, kuris užtikrindavo pastovų žinučių siuntimą iš TD3 modelio į „Unity“ agento aplinką. TD3 atliko serverio rolę, nuolatos pirmutinis siusdavo duomenis agento aplinkai bei iš jo laukdavo atsakymo. Žinutėms siųsti buvo pasirinktas json formatas dėl dviejų priežasčių: abi skirtingos aplinkos turi json nuskaitymo bibliotekas bei į json lengva įrašyti norimą informaciją.

Buvo įgyvendintos dvi esminės žinutės: naujo epizodo bei duomenų rinkimo. Šios dvi žinutės pasižymi panašia duomenų forma - išsaugome tekstinę komandą bei pridedame norimus duomenis. Toliau pateiktos dvi skirtingos siunčiamos komandos:

1. Naujo epizodo komanda. Serveris praneša agentui reikalavimą pradėti naują epizodą. Jį gavęs agentas inicializuoja vorą į pradinę epizodo poziciją ir išsiunčią būsenos duomenis į serverį. Serveris gavęs šiuos duomenis gali pradėti mokymosi procesą.
2. TD3 gavęs voro būsenas aproksimuoją naują veiksmą bei nurodo serveriui išsiųsti jį agentui. Voras gavęs veiksmus juos atlieka. Juos atlikęs atgal išsiunčią serveriui naujai įgytas būsenas, apdovanojimą bei ar ši nauja būsena terminuoją epizodą.

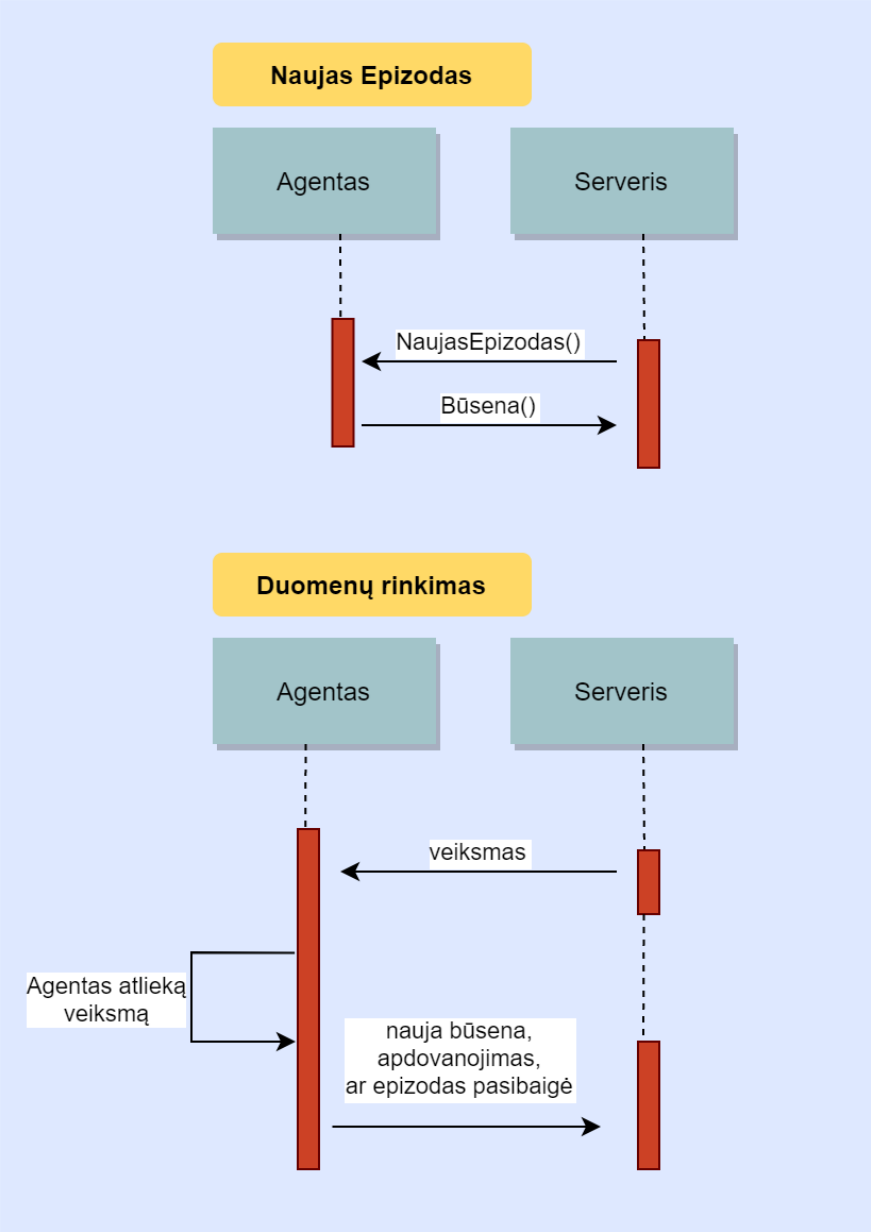
Šis duomenų keitinėjimasis pateiktas 21 paveikslėlyje.

2.6. Mokymosi procesas

Vienas didžiausių skatinamojo mokslo trūkumų yra labai lėtas mokymosi procesas, reikalaujantis daug resursų. Su šia problema buvo susidurta vos ne iš anksto. Pradžioje įgyvendinant TD3 modelį duomenys į patirčių talpyklą buvo renkami tik iš vieno agento. Nors modelio vidutinis apdovanojimų kiekis po truputi didėjo, tačiau tai buvo lėtas procesas. Net vienai mokymo dienai praėjus, voras sugebėdavo išmokti tik stovėti ant savo keturių galūnių bei nedemonstruodavo jokio progreso ties ropojimu į žalio kubelio poziciją. Todėl buvo pasitelktas paralelinis mokymasis, kuris sparčiai pagreitino mokymosi procesą. Vietoj vienos aplinkos, buvo sukurtos dešimt identiškų aplinkų. Sukurtų aplinkų vorų mokymosi seka išliko tokia pati, buvo apmokamas vienas modelis. Tačiau šį karta patirčių talpykla buvo pildoma ne vieno agento pagalba bet dešimties. Ši technika paspartina mokymosi procesą iki dešimties kartų.

Taip pat buvo pritaikytas simuliacijos spartinimas. „Unity“ variklyje kiekvienas fizikinis kadras užtrunka 0.2 ms. Norint pasiekti greitesnį agentų simuliavimo greitį, ši vertė buvo pakeista į 0.01333 ms. Kitaip tariant įvykdavo daugiau fizikinių atnaujinimų, dažniau pakisdavo vorų galūnių pozicijos. Tokiu būdu paspartinus simuliaciją, vizualiai agentai vos ne dvigubai greičiau atlikinėdavo savo veiksmus.

Tačiau net su įvardintais mokymosi pagreitinimais, mokymosi procesas ilgai užtrukdavo. Galutinis voro modelis buvo apmokamas vos ne dvi dienas. Šis ilgo mokymosi didžiausias



21 pav. Serverio bei agento komunikavimo sluoksnis

trūkumas yra lėtas modelio iteravimas. Pakeitus pora modelio parametrų, rezultatų reikia vėl iš naujo laukti pora dienų.

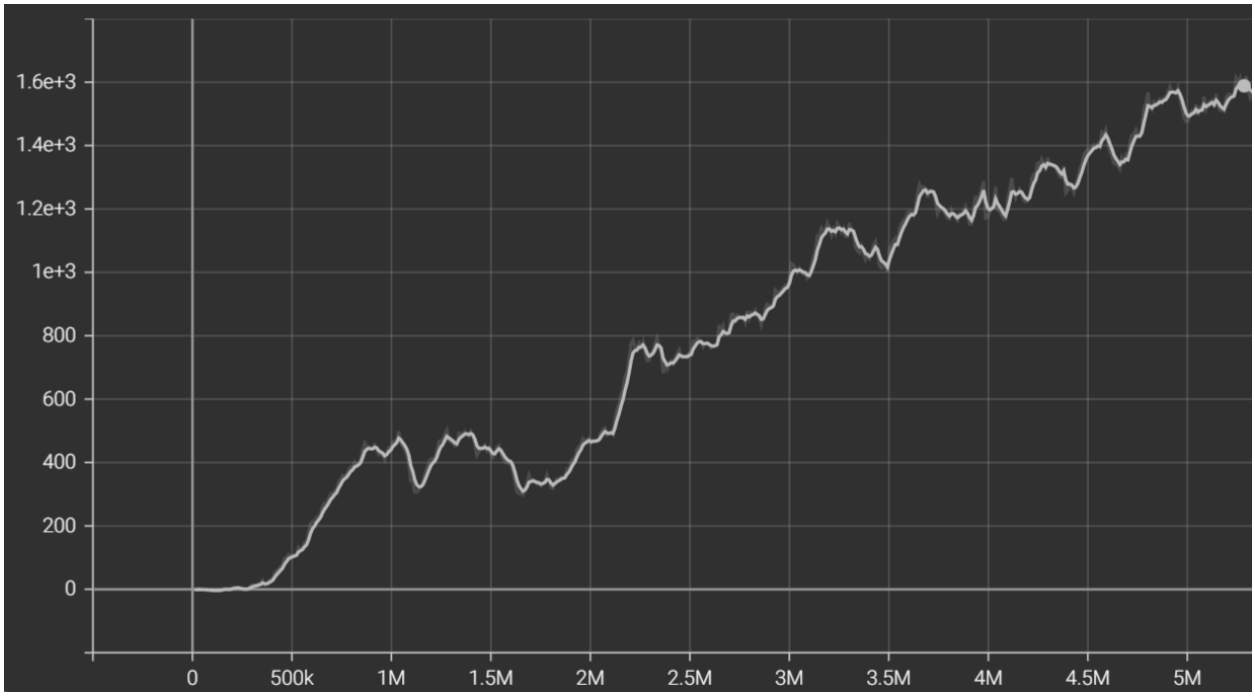
2.7. TD3 modelio įgyvendinimas

Nors buvo planuojama sudaryti TD3 modelį naudojantis savo įgyvendintais neuroniniais tinklais, tačiau pastebėjus lėta mokymosi procesą ši idėja buvo atšaukta. Buvo pasitelkti Tensorflow bibliotekos sukurti neuroniai tinklai, kurių pagrindinis privalumas yra gebėjimas atlikti matricines operacijas paralelizuotai bei vaizdo plokštėje, paspartinant skaičiavimus, o ne procesoriuje. Norint įgyvendinti šiuos du punktus, jie reikalauja nepaprastai daug žinių ir laiko, todėl TD3 modelio verčių bei veiksmų taisyklių aproksimacijos yra įgyvendintos Tensorflow neuroninių tinklų pagalba. Taip pat vienas didžiausių privalumu naudojantis

Tensorflow neuroniais tinklais, yra tai kad TD3 yra paplitęs metodas ir jo įgyvendinimo procesą galima lengvai surasti bei perprasti. Vienas iš susidurtų didžiausių sunkumų buvo veiksmo taisyklių neuroninio tinklo parametrų gradiento radimas, kaip matome 50 formulėje. Šis gradientas apima dviejų neuroninių tinklų gradientą, kurio apskaičiavimas pasirodė keblus. Iki šiol gradientinis nuolydis buvo atliekamas tik vieno neuroninio tinklo parametrams. Tačiau pateiktoje formulėje gradientas keliauja per du neuroninius tinklus. Įsigilinus į 49 formulę, galime pastebėti paprastą gradiento formą: kaip gautos veiksmo taisyklių reikšmės, keičia 49 formulę. Norint atsakyti į šį klausymą pradžioje reikia surasti vieno iš pasirinkto pagrindinio kritiko modelio įvesties neuronų veiksmų išvestines, pavaizduoto raudonai 17 paveikslėlyje, bei jas sudauginti su aktoriaus paprastu neuroniniu gradientiniu nuolydžiu. Tokiu būdu optimizuojame aktoriaus pagrindinio neuroninio tinklo parametrus. Tuo tarpu kritiko modelio optimizavimas yra identiškas giliojo-Q metodui, aptartam 1.7.10 skyriuje.

2.8. Rezultatai

Voro mokymasis truko porą dienų ir mokymuisi pasibaigus agentas gebėdavo dideliu greičiu ropoti link žaliojo kubelio. Mokymosi proceso gerumo kriterijus buvo vidutinis judantis apdovanojimas, kuris pateiktas 22 paveikslėlyje. Viršutinė ašis žymi vieno epizodo vidutinį judanti apdovanojimą, o apatinė ašis žingsnių kiekį.



22 pav. Apdovanojimų grafikas.

Agentui besimokant buvo pastebėti šie mokymosi etapai:

1. Nuo 0 iki 500 tūkst. žingsnio agentas išmoko stovėti ant keturių kojų.
2. Nuo 500 tūkst. iki 1.5 mln. žingsnio agentas išmoko judėti link žaliojo kubelio. Ties 1.2 mln. žingsniu matome smarku pagrindinio rodiklio kritimą, susijusi su katastrofinio

užmiršimo problema. Ši problema atsiranda, kuomet senos patirčių talpyklos reikšmės yra pakeičiamos naujomis, agentas užmiršta optimalias strategijas. Kadangi mano talpyklos dydis yra 1 mln. reikšmių, netrukus šis katastrofinis užmiršimas pasireiškė. Tačiau tai buvo trumpalaikis efektas bei agentas sugebėjo toliau tobulėti.

3. Ties 2 mln. žingsniu agentas išmoko efektyviau sukiotis ir užtikrinčiau stovėti ant savo keturių galūnių. Voras rečiau gaudavo baudą už nukritimą ant pagrindinės kūno dalies.
4. Nuo 2.5 mln. žingsnio agento greičio apdovanojimo funkcija pradėjo įgyti didžiulę įtaką. Agentas pastebėjo, kad kuo greičiau jis juda, tuo didesnį nuopelną gauną. Tai privedė ties greitu galūnių klibinimu.
5. Ties 5 mln. žingsniu voras įgijo optimalų judėjimą. Pasiekus žalią kubelį voras staigiai apsisukdavo bei greitai nuropodavo link naujo kubelio. Taip pat pastebėtas įdomus stabdymo bruožas. Agentui priartėjus prie žalio kubelio, jis smarkiai pradeda stabdyti, mažindamas susidariusi momentinį greitį. Pasiekus žalią kubelį bei esant vos ne stovinčioje pozicijoje, jis gali žymiai greičiau atsisukti į kubelio trajektoriją ir vėl pradėti ropoti link jo.

Taigi TD3 modelis sugebėjo sėkmingai apmokyti vorą judinti fizikines galūnes bei užtikrintai pasiekti žaliąjį kubelį.

3. Išvados

Atlikus skatinamųjų modelių analizę, gautos išvados:

1. Neuroninis tinklas yra efektyvus įrankis aproksimuoti skatinamuosius modelius. Tačiau neuroniniai tinklai yra labai nestabilūs, prasta jų inicializacija priveda prie prastos konvergacijos. Svarbu teisingai priskirti neuroninio tinklo parametrus, pritaikyti reguliarizacijas, neužmiršti gradientinio nuolydžio momentinius algoritmus kaip Adam. Pagal šiuos principus buvau sukūręs savo neuroninius tinklus, tačiau dėl lėtos konvergacijos buvo pasitelkta Tensorflow neuroninių tinklų biblioteka.
2. Ne generalizuoti skatinamieji modeliai yra prasti metodai. Šie metodai gali būti pritaikomi tik mažos dimensijos problemose. Kylant būsenų dimensijai, konvergavimo laikas greitai pakyla į begalines reikšmes. Todėl skatinamieji metodai dešimtmečius po jų atradimo buvo retai pritaikomi. Dažniausiai šie metodai buvo pritaikomi plėtoti skatinamąjį mokslą demonstruojant jų efektyvumą primityviuose pavyzdžiuose. Buvo įgyvendinti SARSA, Q-mokymosi metodai, kurie gali optimizuoti tik mažos būsenų erdvės uždavinius.
3. Aproksimuojami bei generalizuojami skatinamieji metodai transformavo mirusį mokslą į didžiulį potencialą turinti. Atradus veiksmo taisyklių gradientą, neuronų tinklais aproksimuojami skatinamieji metodai išsprendė didžiausią šio mokslo problemą - aukštų dimensijų aproksimavimas.
4. Pritaikytas giliojo-Q mokymosi metodas pasižymi tik būsenos verčių aproksimavimu, neturi veiksmo taisyklių aproksimizacijos. Jis yra silpniausias iš visų aproksimuojamų skatinamųjų modelių, tačiau vienas iš labiausiai įtakingiausių. Iš jo kilo daugybe metodų, kaip PPO, DPG, DDPG, TD3.
5. Buvo įgyvendintas TD3 metodas, kuris sugeba pateikti deterministines veiksmo taisykles. Metodas buvo sėkmingai pritaikytas spręsti voro fizikinę judėjimo problemą. Apmokytas metodas sugebėjo tiksliai pateikti optimalius veiksmus voro galūnėms, kuriomis voras ropojo iki nusibrėžtos pozicijos.

Literatūra

[1] Sutton, R. & Barto, A. 1987, 'A Temporal-Difference Model of Classical Conditioning', in Proceedings of the Ninth Annual Conference of the Cognitive Science Society, Seattle, WA, pp. 355–78.PDF.

[2] Ivan P. Pavlov(1927). CONDITIONED REFLEXES: AN INVESTIGATION OF THE PHYSIOLOGICAL ACTIVITY OF THE CEREBRAL CORTEX html.

[3] Diederik P. Kingma, Jimmy Lei Ba (2015). ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION PDF

[4] Watkins, C.J.C.H. (1989). Learning from Delayed Rewards. PDF

[5] Methods and Apparatus for Reinforcement Learning, US Patent #20150100530A1. US Patent Office. 9 April 2015. Retrieved 28 July 2018. PDF

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller (2013). Playing Atari with Deep Reinforcement Learning. PDF

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver (2015). Human-level control through deep reinforcement learning PDF

[8] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra (2014). Deterministic Policy Gradient Algorithms. ICML, Jun 2014, Beijing, China. ffhal-00938992f PDF

[9] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess,Tom Erez, Yuval Tassa, David Silver & Daan Wierstra (2016). CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING. Google Deepmind London, UK PDF

[10] Scott Fujimoto, Herke van Hoof, David Meger (2018). Addressing Function Approximation Error in Actor-Critic Methods PDF

A. Priedai

Visą programos kodą bei įgyvendintą projektą galima surasti nurodytoje github paskyroje.

Terminai

$\boldsymbol{\pi(a|s)}$ – Nurodo su kokia tikimybe būsenoje s, pasirinksime veiksmą a. 27

$\boldsymbol{q_\pi(s,a)}$ – Veiksmo a apdovanojimo vertės matas esant būsenoj s. 28

$\boldsymbol{v_\pi(s)}$ – Pasirinktos būsenos $\boldsymbol{s}$ tikėtinas apdovanojimas. 27

a – Agento pasirinktas veiksmas. 25

r – Agento esamosios būsenos pasirinkto veiksmo apdovanojimas. 25

s – Esama agento būsena. 25

Žaidimo agentas – Žaidimo agentas yra suprogramuotas žaidėjas kuris stebi žaidėjo ėjimus, protingai mąsto ir atlieka savo ėjimus. Tai galėtų būti suprogramuotas šachmatų varžovas. 6

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Paulius Janėnas, 20154453

(Studento vardas ir pavardė, studento pažymėjimo Nr.)

Fundamentinių mokslų fakultetas

(Fakultetas)

Duomenų mokslas ir statistika, DMfm-20

(Studijų programa, akademinė grupė)

BAIGIAMOJO DARBO (PROJEKTO)

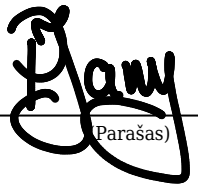
SĄŽININGUMO DEKLARACIJA

2022 m. gegužės 30 d.

Patvirtinu, kad mano baigiamasis darbas tema „Skatinamuoju modeliu sukurti žaidimų agentai“ yra savarankiškai parašytas. Šiame darbe pateikta medžiaga nėra plagijuota. Tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos pažymėtos literatūros nuorodose.

Parenkant ir įvertinant medžiagą bei rengiant baigiamąjį darbą, mane konsultavo mokslininkai ir specialistai: docentas daktaras Tomas Rekašius. Mano darbo vadovas docentas daktaras Tomas Rekašius.

Kitų asmenų indėlio į parengtą baigiamąjį darbą nėra. Jokių įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs (-usi).



(Parašas)

Paulius Janėnas

(Vardas ir pavardė)