# AGENTNI O‘RGATISHDA Q-LEARNING VA SARSA ALGORITMLARINI TAQQOSLASH

***Umarov Bekzod Azizovich***
*Farg‘ona davlat unversiteti amaliy matematika va informatika kafedrasi o‘qituvchisi*
ubaumarov@mail.ru
***Tursunova Azimaxon Alijon qizi***
*Farg‘ona davlat unversiteti talabasi*
azimaxontursunova676@gmail.com

**Annotatsiya**

Ushbu maqolada Q-learning va SARSA (State-Action-Reward-State-Action) algoritmlarining taqqoslanishi, ularning o‘rganish jarayonlari tahlil qilinadi. Q-learning va SARSA algoritmlari kuchaytirilgan o‘qitish (Reinforcement Learning) metodlaridan bo‘lib, agentning maqsadga erishish uchun optimal strategiyalarni o‘rganishiga yordam beradi. Bu algoritmlar, o‘rganish jarayonida qanday turdagi ma'lumotlar va tajribalarga asoslanib qarorlar qabul qilishni o‘rganishni ko‘rsatadi. Maqolada ikkala algoritmning asosiy farqlari va qo‘llanilish sohalari tahlil qilinadi, shuningdek, Python dasturlash tilida kod namunasi ko‘rsatiladi.

**Kalit so‘zlar:** Q-learning, SARSA, kuchaytirilgan o‘qitish, agent, optimal strategiya, o‘rganish, Python, algoritmlar, taqqoslash.

Kuchaytirilgan o‘qitish (Reinforcement Learning) - bu agentning o‘z tajribalari orqali optimal qarorlar qabul qilishni o‘rganish jarayonidir. RLda agent ma'lum bir muhitda faoliyat yuritadi, unga harakatlar bajarish va ular natijasida mukofot yoki jazo olish imkoniyati beriladi. Maqsad agentning eng yuqori mukofotni olishga intilishidir. Bu jarayonda agentning harakatlari va qarorlari orqali o‘rganish amalga oshiriladi.

## Kuchaytirilgan o‘qitishda Q-Learning

Q-learning - bu mashinani o‘rganish va sun’iy intellekt ilovalarida qo‘llaniladigan mashhur modelsiz mustahkamlovchi o‘rganish algoritmidir. Bu vaqtinchalik farqni o‘rganish usullari toifasiga kiradi, bunda agent natijalarni kuzatish, atrof-muhit bilan o‘zaro ta’sir qilish va mukofotlar shaklida fikr-mulohazalarni olish orqali yangi ma’lumotlarni oladi. Q-learningning asosiy xususiyati shundaki, u **off-policy** algoritmi hisoblanadi, ya'ni agent o‘zining oldingi tajribalarini ishlatib, muhitni qanday o‘rganishi bilan bog‘liq emas.

Q-learningning umumiy formulasi:

**Q(st,at)=Q(st,at)+α(rt+1+γmax⁡a′Q(st+1,a′)−Q(st,at))Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left( r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right)Q(st,at)=Q(st,at)+α(rt+1+γa′maxQ(st+1,a′)−Q(st,at))**

Bu yerda:

**Q(st,at)Q(s_t, a_t)Q(st,at)** — joriy holat va harakatning qiymati,

**α\alphaα** — o‘rganish tezligi,

**γ\gammaγ** — chegara omili (discount factor),

**rt+1r_{t+1}rt+1** — mukofot,

**max⁡a′Q(st+1,a′)\max_{a'} Q(s_{t+1}, a')maxa′Q(st+1,a′)** — keyingi holatda eng yaxshi harakatning qiymati.

Q-learningning asosiy komponentlari

1. Q-qiymatlar yoki harakat-qiymatlar

Q-qiymatlari holatlar va harakatlar uchun aniqlanadi. Q(S, A)Q(S,A) - S holatida A harakatini bajarish qanchalik yaxshi ekanligini baholash. Q(S, A)Q(S,A) ning bu taxmini keyingi bo‘limlarda ko‘rib chiqiladigan TD- Update qoidasi yordamida iterativ tarzda hisoblanadi.

2. Mukofotlar va epizodlar

Agent o‘z hayoti davomida boshlang‘ich holatdan boshlanadi va o‘z tanlovi, shuningdek agent o‘zaro ta'sir qiladigan muhit asosida joriy holatidan keyingi holatga bir necha marta o‘tadi. O‘tishning har bir bosqichida agent holatdan harakat qiladi, atrof-muhitdan mukofotni kuzatadi va keyin boshqa holatga o‘tadi. Agar istalgan vaqtda agent tugatuvchi holatlardan biriga tushib qolsa, bu boshqa o‘tishlar mumkin emasligini anglatadi. Bu epizodning tugashi deyiladi.

3. Vaqtinchalik farq yoki TD-yangilash

Vaqtinchalik farq yoki TD-Update qoidasi quyidagicha ifodalanishi mumkin:

Q(S,A)\chap Q(S,A) + \alfa (R + \gamma Q({S}',{A}') – Q(S,A))Q(S,A)←Q(S,A)+a(R+gQ(S',A')–Q(S,A))

4. s - ochko‘zlik siyosati bilan harakat yo‘lini tanlash

Q-qiymatining joriy baholari asosida amalga oshiriladigan harakatni tanlashning oddiy usuli bu s-ochko‘zlik siyosatidir. U shunday ishlaydi:

1.Yuqori Q-qiymati harakati (ekspluatatsiya):

a) Ko‘pchilik holatlarni ifodalovchi 1−s ehtimolligi bilan,

b) Ayni paytda eng yuqori Q-qiymatli amalni tanlang.

c) Ushbu ekspluatatsiya holatida agent o‘zining hozirgi tushunchasini hisobga olgan holda, u eng maqbul deb hisoblaydigan harakat yo‘nalishini tanlaydi.

2. Tasodifiy harakat orqali kashfiyot:

a) s ehtimollik bilan, vaqti-vaqti bilan,

b) Q-qiymati eng yuqori bo‘lgan harakat yo‘nalishini tanlash o‘rniga,

c) Q-qiymatlaridan qat'i nazar, tasodifiy har qanday harakatni tanlang.

d) Yangi harakatlarning mumkin bo‘lgan afzalliklarini bilish uchun agent tadqiqot turi bilan shug‘ullanadi.

Kuchaytirilgan o‘qitishda SARSA algoritimi

SARSA (State-Action-Reward-State-Action) - bu ham kuchaytirilgan o‘qitish algoritmi bo‘lib, u Q-learningga qaraganda **on-policy** algoritmidir. SARSA algoritmi o‘zining o‘rganish jarayonida agentning joriy holati va harakatini hisobga oladi va keyingi holatdagi qarorni (harakatni) baholashga asoslanadi. SARSAda harakatlar va holatlar bir-biriga mos keladi.

SARSA algoritmi mashhur Q-Learning algoritmining biroz o‘zgarishi. Har qanday mustahkamlashni o‘rganish algoritmidagi o‘quv agenti uchun uning siyosati ikki xil bo‘lishi mumkin:

a) **On Policy:** Bunda o‘quv agenti hozirda foydalanilayotgan siyosatdan olingan joriy harakatga muvofiq qiymat funksiyasini o‘rganadi.

b) **Off Policy:** Bunda o‘quv agenti boshqa siyosatdan olingan harakatga muvofiq qiymat funksiyasini o‘rganadi.

SARSAning umumiy formulasi:

**Q(st,at)=Q(st,at)+α(rt+1+γQ(st+1,at+1)−Q(st,at))Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left( r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right)Q(st,at)=Q(st,at)+α(rt+1+γQ(st+1,at+1)−Q(st,at))**

Bu yerda:

**Q(st,at)Q(s_t, a_t)Q(st,at)** — joriy holat va harakatning qiymati,

**rt+1r_{t+1}rt+1** — mukofot,

**γ\gammaγ** — chegara omili,

**Q(st+1,at+1)Q(s_{t+1}, a_{t+1})Q(st+1,at+1)** — keyingi holatdagi harakatning qiymati.

SARSA uchun yangilanish tenglamasi joriy holatga, joriy harakatga, olingan mukofotga, keyingi holatga va keyingi harakatga bog‘liq. Ushbu kuzatish o‘rganish texnikasini nomlanishiga olib keladi, chunki SARSA (s, a, r, s', a') timsoli bo‘lgan Davlat Harakati Mukofoti Davlat Aktsiyasini anglatadi.

**Q-learning va SARSA taqqoslash**

**Q-learning** off-policy algoritm bo‘lib, agent o‘zining strategiyasidan mustaqil ravishda qaror qabul qiladi. Bu shuni anglatadiki, agar agent o‘zining tanlangan harakatlari orqali optimal strategiyani o‘rganmasa ham, Q-learningda maqsadli strategiya topiladi.

**SARSA** esa on-policy algoritmidir, ya'ni agent o‘zining strategiyasiga amal qiladi va harakatni shunga muvofiq baholaydi. Bu holatda agentning amaliyoti o‘rganish jarayonida bevosita ta'sir qiladi.

**Xulosa**

Q-learning va SARSA algoritmlari kuchaytirilgan o‘qitish sohasida muhim o‘rin tutadi. Har ikkala algoritmning o‘ziga xos afzalliklari va kamchiliklari mavjud. Q-learning agentning joriy harakati va keyingi qaroridan mustaqil bo‘lsa, SARSA esa agentning strategiyasini hisobga olib qaror qabul qiladi. Q-learning ko‘proq o‘rganish jarayonida optimal strategiya olish uchun foydalidir, SARSA esa, odatda, tezroq va amaliyroq o‘rganish imkonini beradi.

## FOYDALANILGAN ADABIYOTLAR VA SAYTLAR

1.Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

2.Watkins, C. J. C. H., & Day

3. **https://www.geeksforgeeks.org**