Reinforcement Learning

1. Jelaskan cara kerja dari algoritma Q-Learning dan SARSA!

   Q-Learning

   Q-Learning adalah metode pembelajaran tanpa model yang mengoptimalkan kebijakan dengan cara memperbarui nilai Q (fungsi nilai aksi) menggunakan estimasi reward maksimal dari aksi yang akan datang. Q-Learning menggunakan pendekatan off-policy, dalam hal ini, agen mempelajari fungsi nilai berdasarkan tindakan saat ini yang berasal dari kebijakan yang sedang digunakan. Prosesnya adalah sebagai berikut:

   - Inisialisasi Q-Table:

     Q-Table diinisialisasi dengan nilai awal, biasanya nol, dengan ukuran [state, action].
   - Pengambilan Aksi:

     Agent memilih aksi berdasarkan eksplorasi (random) atau eksploitasi (memilih aksi terbaik berdasarkan Q-Table).
   - Transisi dan Pemberian Reward:

     Setelah melakukan aksi, agent bergerak ke state berikutnya, dan menerima reward berdasarkan aksi tersebut.
   - Update Q-Table:

     Q-Table diupdate menggunakan rumus:

     $$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

     di mana α adalah learning rate, γ adalah discount factor, r adalah reward yang diterima, dan max_α′ Q(s′, α′) adalah nilai Q maksimal dari state berikutnya.
   - Pengulangan:

     Proses ini diulang sampai agent mencapai kondisi berhenti (win atau lose).

   SARSA (State-Action-Reward-State-Action)

   SARSA adalah algoritma reinforcement learning yang menggunakan pendekatan on-policy, dalam hal ini, agen pembelajaran mempelajari fungsi nilai sesuai dengan tindakan yang berasal dari kebijakan lain. Prosesnya adalah sebagai berikut:

   - **Inisialisasi Q-Table**:

     Sama dengan Q-Learning, Q-Table diinisialisasi dengan nilai awal.
   - **Pengambilan Aksi**:

     Agent memilih aksi berdasarkan strategi epsilon-greedy yang sejalan dengan kebijakan yang sedang digunakan.

- **Transisi dan Pemberian Reward**:
  Setelah melakukan aksi, agent bergerak ke state berikutnya dan menerima reward.
- **Pengambilan Aksi Berikutnya**:
  Aksi berikutnya diambil berdasarkan Q-Table pada state yang baru.
- **Update Q-Table**:
  Q-Table diupdate menggunakan rumus:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$$

  di mana α′ adalah aksi berikutnya yang diambil oleh agent di state s′.
- **Pengulangan**:
  Proses ini diulang sampai agent mencapai kondisi berhenti.

2. Bandingkan hasil dari kedua algoritma tersebut, bagaimana hasil perbandingannya? Jika ada perbedaan, jelaskan alasannya!

```
Q-LEARNING
Path: [2, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3]
Total Points: 569
Q-Table:
[[   0.           0.        ]
 [2170.84610695 1648.88952154]
 [1968.45651124 2223.65203605]
 [2207.53841984 2428.90359271]
 [2300.04398558 2446.02546709]
 [2358.1549668  2462.53055152]
 [2333.39872647 2470.03231601]
 [2379.79358883 2494.19504113]
 [2386.79662253 2516.64154306]
 [   0.           0.        ]]



SARSA
Path: [2, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3, 4, 5, 6, 7, 8, 3]
Total Points: 569
Q-Table:
[[   0.           0.        ]
 [ 781.29245635 2231.93507136]
 [2071.7525889  2380.56694504]
 [2348.64695417 2400.12292568]
 [2387.60016645 2418.95773081]
 [2382.58694796 2445.52325624]
 [2417.23320655 2486.06959892]
 [2431.59381195 2529.35232382]
 [2439.49170201 2542.66196285]
 [   0.           0.        ]]
```

- **Total Points**
  Keduanya memiliki total poin yang sama yaitu 569. Ini menunjukkan bahwa baik Q-Learning maupun SARSA mencapai tujuan yang sama dalam hal perolehan poin dalam permainan ini.
- **Path**
  Path yang dilalui oleh kedua algoritma sangat mirip, dengan pola pergerakan yang repetitif dari posisi 2 hingga 8 dan kembali ke posisi

3 secara berulang. Ini menunjukkan bahwa kedua algoritma mengikuti strategi yang serupa dalam permainan.

- **Q-Table**
  - Q-Learning:

    Q-Table menunjukkan nilai yang lebih tinggi pada state tertentu dibandingkan SARSA, terutama pada state yang dekat dengan tujuan (posisi 9). Ini menunjukkan bahwa Q-Learning cenderung memberikan lebih banyak nilai pada aksi yang dianggap optimal secara teori, berfokus pada eksploitasi aksi terbaik berdasarkan informasi saat ini.
  - SARSA:

    Q-Table menunjukkan nilai yang lebih rendah pada state tertentu dibandingkan Q-Learning. Ini karena SARSA lebih berhati-hati, karena nilai Q diperbarui berdasarkan aksi yang benar-benar diambil, bukan hanya berdasarkan aksi terbaik secara teoretis. Ini membuat SARSA lebih stabil tetapi mungkin tidak memberikan nilai yang sama tinggi untuk state tertentu.

Perbedaan tersebut dapat terjadi karena Q-Learning memperbarui nilai Q berdasarkan aksi terbaik yang mungkin diambil di masa depan (off-policy), sehingga dapat lebih agresif dalam mengejar reward maksimum. Sementara SARSA memperbarui nilai Q berdasarkan aksi yang diambil oleh agent saat ini (on-policy), yang membuatnya lebih konservatif dan lebih terfokus pada kebijakan aktual yang diikuti.

Selain itu, Q-Learning mungkin memiliki nilai yang lebih tinggi karena fokus pada aksi optimal teoritis. Namun, hal ini dapat menyebabkan overestimation pada beberapa nilai Q. Sementara SARSA cenderung lebih stabil dan lebih berhati-hati, tetapi mungkin lebih lambat dalam mencapai nilai Q yang sangat tinggi karena selalu memperbarui berdasarkan aksi yang diambil saat ini.

Secara keseluruhan, perbedaan ini menunjukkan bahwa Q-Learning cenderung memberikan hasil yang lebih optimis dalam hal nilai Q, sementara SARSA lebih stabil dan lebih sesuai untuk kebijakan yang lebih hati-hati. Hasil yang serupa dalam poin total menunjukkan bahwa dalam kasus ini, kedua algoritma berhasil memecahkan masalah dengan cara yang berbeda, namun mendekati hasil yang serupa.