Jelaskan cara kerja dari algoritma Q-Learning dan SARSA!

A. Q-Learning

Q-Learning adalah algoritma pembelajaran penguatan (reinforcement learning) yang digunakan untuk menemukan kebijakan optimal dalam lingkungan yang diwakili sebagai Markov Decision Process (MDP). Algoritma ini bersifat *off-policy*, artinya pembaruan nilai Q dilakukan berdasarkan tindakan yang optimal, bukan berdasarkan tindakan yang sebenarnya dilakukan oleh agent.

Algoritma Q-Learning memiliki beberapa tahap:

1. Inisialisasi

Inisialisasi nilai Q untuk semua pasangan keadaan-tindakan (s, a) dengan nilai sebarang, biasanya 0.

2. Iterasi

Pada setiap langkah, agent berada dalam keadaan s dan memilih tindakan a berdasarkan kebijakan tertentu, seperti ε-greedy, yang memungkinkan agent untuk mengeksplorasi tindakan baru sesekali.

Agent kemudian melakukan tindakan a, bergerak ke keadaan berikutnya s' dan menerima reward r.

Nilai Q diperbarui menggunakan persamaan:

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

Dengan:

- α adalah laju pembelajaran (learning rate).
- γ adalah faktor diskon (discount factor).
- maxa'Q(s',a') adalah nilai Q terbaik untuk keadaan s' berikutnya.

Kedua langkah di atas akan diulang terus hingga mencapai batas iterasi tertentu.

SARSA (State-Action-Reward-State-Action) adalah algoritma pembelajaran penguatan yang juga digunakan untuk menemukan kebijakan optimal dalam MDP. Berbeda dengan Q-Learning, SARSA bersifat *on-policy*, artinya pembaruan nilai Q dilakukan berdasarkan tindakan yang sebenarnya dilakukan oleh agent.

B. SARSA (State-Action-Reward-State-Action)

Algoritma SARSA memiliki beberapa tahap:

1. Inisialisasi

Sama seperti Q-Learning, nilai Q untuk semua pasangan keadaan-tindakan (s, a) diinisialisasi dengan nilai sebarang, biasanya 0.

2. Iterasi

Pada setiap langkah, agent berada dalam keadaan s dan memilih tindakan a berdasarkan kebijakan tertentu, seperti ε-greedy. Agent kemudian melakukan tindakan a, bergerak ke keadaan berikutnya s', menerima reward r, dan memilih tindakan baru a' di s'.

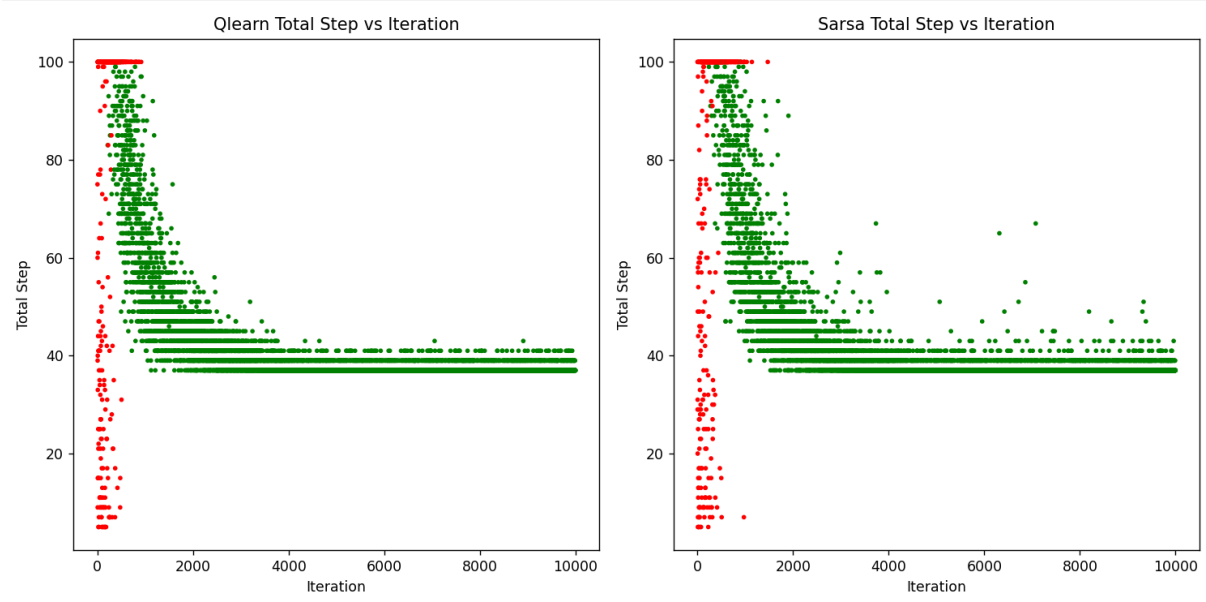
Nilai Q diperbarui menggunakan persamaan

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

Keterangan: Q(s',a') adalah nilai Q dari tindakan a' yang dipilih di keadaan s'.

Kedua langkah di atas akan diulang terus hingga mencapai batas iterasi tertentu.

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



Kedua algoritme awalnya menunjukkan beberapa episode dengan jumlah langkah yang tinggi, yang menunjukkan bahwa agen kesulitan untuk mencapai goal secara efisien. Seiring bertambahnya iterasi, kedua algoritme menunjukkan penurunan signifikan dalam jumlah langkah yang diperlukan untuk mencapai sasaran, yang menunjukkan proses belajar menuju perilaku optimal.

Qlearning menemukan langkah optimal lebih cepat dan lebih stabil. Sedangkan SARSA membutuhkan iterasi yang lebih lama untuk mencapai langkah optimal. SARSA juga memiliki jumlah langkah yang lebih tidak stabil. Hal ini terjadi mungkin karena agen pada SARSA lebih berhati-hati dalam eksplorasinya sehingga mengarah ke hasil yang lebih bervariasi mengingat SARSA merupakan algoritma on-policy yang mengubah policy sesuai langkah yang benar-benar dijalankan.