Tugas Seleksi Asisten Laboratorium AI 2023
Reinforcement Learning

September 2, 2025

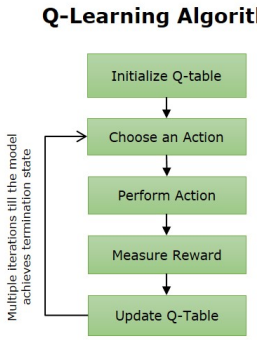
Adinda Putri
13523071

---

1. Jelaskan cara kerja dari algoritma Q-Learning dan SARSA, terutama perbedaan fundamental antara keduanya (*on-policy vs off-policy*).

Ans: Q-Learning dan SARSA adalah algoritma *model-free* dalam *reinforcement learning* yang memungkinkan *agent* mempelajari *policy* optimal melalui interaksi dengan lingkungan tanpa model yang telah ditentukan.
Agent menerima *reward* (imbalan) untuk tindakan yang menguntungkan dan *punishment* (hukuman) untuk tindakan merugikan. Umpan balik tersebut digunakan untuk memperbarui Q-value, yaitu perkiraan efektivitas suatu tindakan pada keadaan tertentu. Seluruh Q-value disimpan dalam sebuah Q-table, yang berfungsi sebagai panduan bagi *agent* dalam memilih tindakan terbaik pada setiap keadaan. Namun, terdapat perbedaan di antara kedua algoritma tersebut, yaitu:

(a) Q-Learning merupakan algoritma *off-policy*, artinya meskipun *agent* memilih tindakan secara eksploratif, pembaruan Q-value dilakukan berdasarkan tindakan terbaik yang mungkin diambil (*max* Q). Hal ini memungkinkan Q-Learning mempelajari *policy* secara optimal meskipun tindakan aktual selama pelatihan tidak selalu optimal.



Gambar 1. ***Cara Kerja Algoritma Q-Learning***
Sumber: *tutorialspoint*

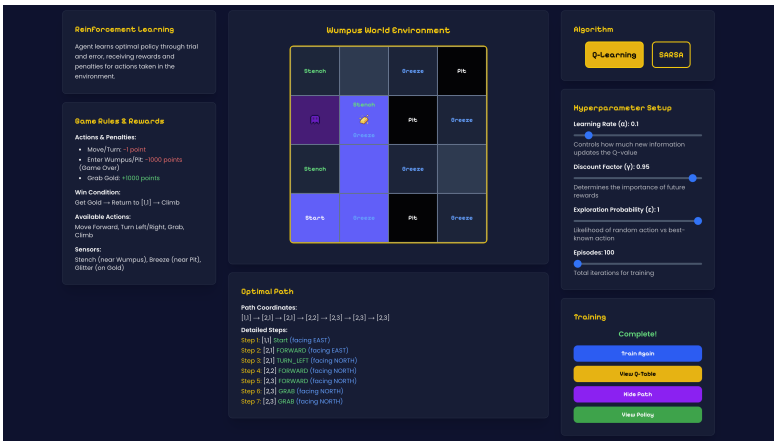
Proses Q-Learning berjalan secara iteratif. Pada setiap langkah, *agent* mengamati keadaan, memilih tindakan, menerima umpan balik dari lingkungan, lalu memperbarui Q-Table menggunakan persamaan *Temporal Difference*. Proses ini diulang hingga Q-value konvergen.

(b) SARSA merupakan algoritma *on-policy*, yang berarti pembaruan Q-value dilakukan berdasarkan tindakan yang dipilih oleh *agent* sesuai dengan *policy* yang sedang dijalankan. Dengan demikian, *policy* yang terbentuk melalui SARSA merepresentasikan strategi yang benar-benar diterapkan oleh *agent* selama proses pelatihan.
Proses SARSA juga berjalan secara iteratif. Pada setiap langkah, *agent* mengamati keadaan dan memilih tindakan, mengeksekusi tindakan tersebut, menerima umpan balik dari lingkungan, lalu memilih tindakan berikutnya. Pembaruan Q-Table dilakukan menggunakan persamaan *Temporal Difference* berdasarkan keadaan dan aksi yang benar-benar diambil. Proses ini diulang hingga Q-value konvergen.

2. Bandingkan hasil dari kedua algoritma tersebut dalam konteks Wumpus World ini. Analisis perbandingan bisa mencakup:
   (a) Kecepatan konvergensi (jumlah episode yang dibutuhkan untuk belajar).
   (b) *policy* (*policy*) final yang dihasilkan. Apakah ada perbedaan?
   (c) Jalur (*path*) yang ditempuh. Apakah Q-Learning cenderung mengambil rute yang lebih berisiko dibandingkan SARSA? Jelaskan mengapa!

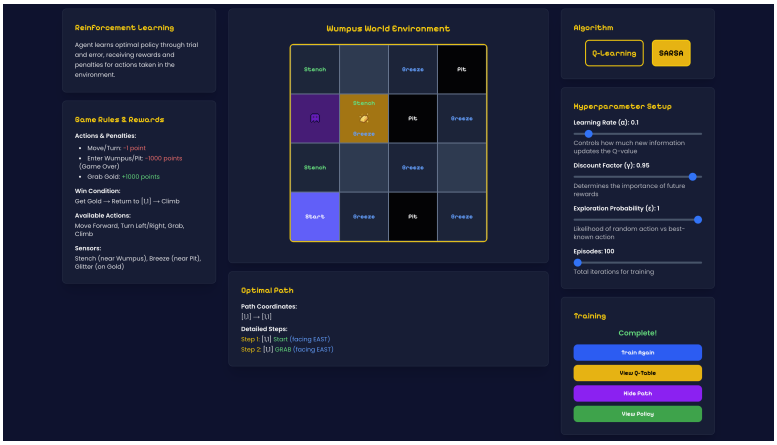
Ans:

(a) **Kecepatan Konvergensi**
Berdasarkan hasil pengujian pada *web-application* yang telah dibuat, Q-Learning menunjukkan kecepatan konvergensi yang lebih tinggi dibandingkan SARSA. Dengan parameter yang sama, ketika jumlah episode diset sebanyak 100, *agent* dengan algoritma Q-Learning sudah mampu mencapai *gold*, sementara *agent* dengan algoritma SARSA hanya berhenti pada state [1,1] dan langsung memilih aksi CLIMB.



Gambar 2. ***Algoritma Q-Learning dengan 100 Episode***
Sumber: Penulis

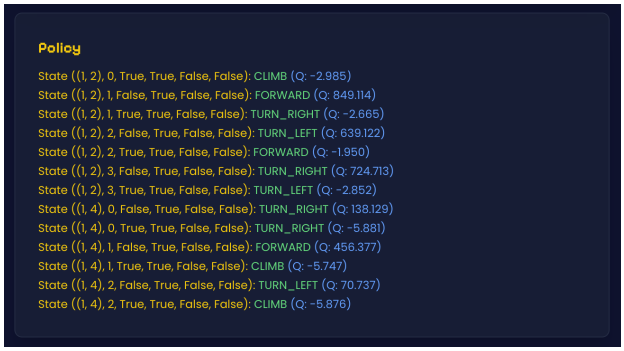
Hal ini terjadi karena Q-Learning bersifat *off-policy*, di mana pembaruan Q-value selalu berdasarkan tindakan terbaik yang mungkin diambil pada state berikutnya terlepas dari tindakan nyata yang dipilih. Ketika *agent* melakukan kesalahan atau menempuh jalur yang kurang optimal, proses pembaruan tetap mengarahkan *agent* untuk belajar menuju *policy* optimal sehingga Q-Learning membutuhkan lebih sedikit episode untuk menemukan jalur menuju gold secara konsisten. Sebaliknya, SARSA bersifat *on-policy* dan memperbarui Q-value berdasarkan tindakan nyata saat eksplorasi. Kondisi ini membuat *agent* lebih berhati-hati dalam mempertimbangkan risiko sehingga membuat laju pembelajaran menjadi lebih lambat.



Gambar 3. ***Algoritma SARSA dengan 100 Episode***
Sumber: Penulis

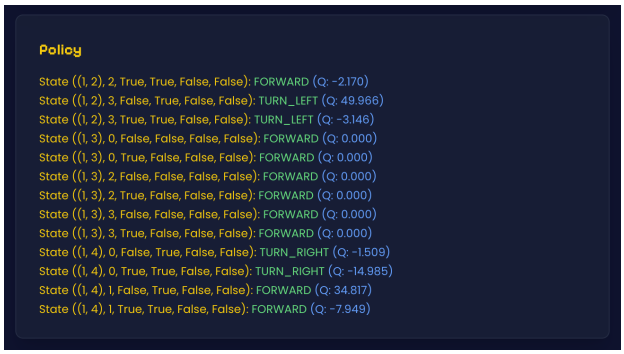
(b) **Kebijakan (*Policy*) Final**
Dari hasil pengujian pada *web-application* yang telah dibuat, Q-value yang

dihasilkan oleh algoritma SARSA cenderung lebih kecil dibandingkan dengan algoritma Q-Learning. Selain itu, *policy* dari Q-Learning tidak mengeksplor state berbahaya seperti state [1,3] tempat Wumpus berada.



Gambar 4. *Q-Learning Final Policy*
Sumber: Penulis

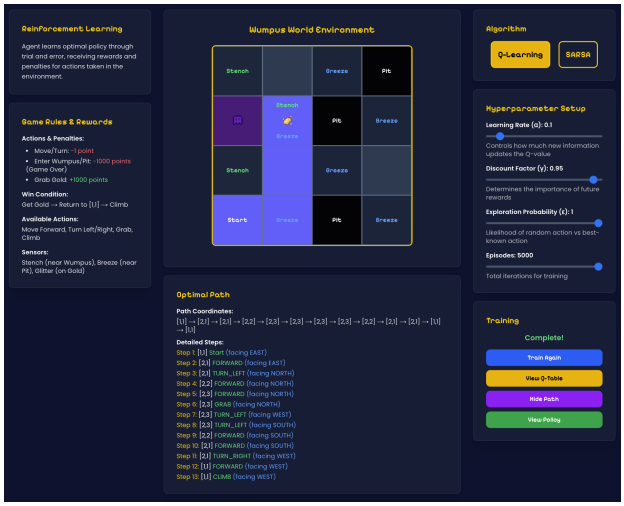
Perbedaan ini terjadi karena sifat dasar kedua algoritma. Q-Learning yang bersifat *off-policy* sehingga tidak wajib mengeksplor semua state karena pembaruan Q-value selalu menggunakan nilai maksimum pada state berikutnya. Sebaliknya, SARSA yang bersifat *on-policy* memperbarui Q-value berdasarkan tindakan nyata yang dilakukan *agent*, termasuk ketika melewati state berbahaya. Hal ini membuat SARSA lebih sering mencatat Q-value pada state berisiko sehingga *policy*-nya lebih lengkap, walaupun Q-value yang dihasilkan lebih kecil dibandingkan Q-Learning.



Gambar 5. *SARSA Final Policy*
Sumber: Penulis

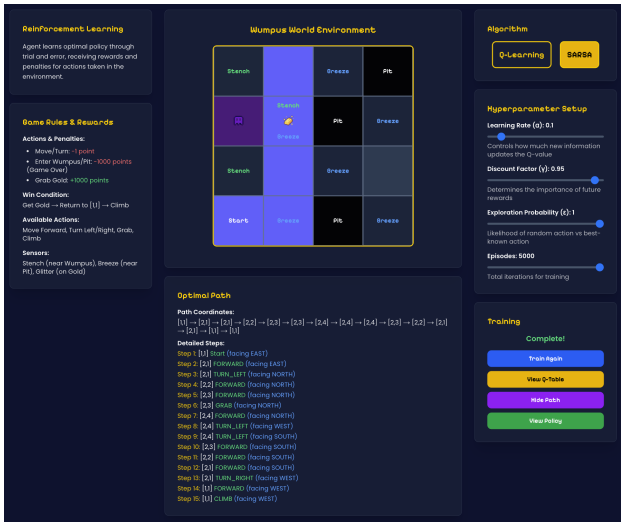
(c) **Jalur (*Path*) Tempuh**
Pada hasil pengujian dengan 5000 episode, jalur yang ditempuh oleh *agent* dengan algoritma Q-Learning lebih optimal dibandingkan SARSA. Ketika berada pada state [2,3], *agent* Q-Learning langsung memilih aksi TURN LEFT karena pembaruan Q-value-nya selalu mengacu pada tindakan terbaik yang tersedia. Dengan sifat *off-policy*, Q-Learning melewati tindakan yang

kurang efisien dan tetap mengarahkan *agent* pada jalur terpendek untuk kembali ke awal setelah mengambil gold.



Gambar 6. ***Jalur Tempuh dengan Algoritma Q-Learning***
Sumber: Penulis

Sebaliknya, SARSA yang bersifat *on-policy* memperbarui Q-value berdasarkan tindakan nyata yang dilakukan selama eksplorasi. Jika pada suatu episode *agent* pernah maju ke [2,4] dan tidak langsung gagal, maka tindakan tersebut akan tercatat sebagai pengalaman yang aman dan memengaruhi pembaruan Q-value. Akibatnya, *policy* akhir SARSA memasukkan aksi FORWARD ke [2,4], sehingga jalur yang ditempuh menjadi lebih panjang.



Gambar 7. ***Jalur Tempuh dengan Algoritma SARSA***
Sumber: Penulis

References

[1] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition, Prentice Hall, 2010.

[2] TutorialsPoint. *Machine Learning - SARSA Reinforcement Learning*. Diakses dari: `https://www.tutorialspoint.com/machine_learning/machine_learning_sarsa_reinforcement_learning.htm`, 2025.

[3] TutorialsPoint. *Machine Learning - Q Learning*. Diakses dari: `https://www.tutorialspoint.com/machine_learning/machine_learning_q_learning.htm`, 2025.