1. Jelaskan apa yang dimaksud dengan Reinforcement Learning!

   Reinforcement learning merupakan pendekatan komputasi pembelajaran dari aksi. Agen **belajar dari lingkungan dengan cara *trial and error*.** Agen akan menerima penilaian (*reward*) dari aksi yang diambil sebagai umpan balik. Tujuan dari reinforcement learning adalah memaksimalkan nilai tersebut.

2. Jelaskan bagaimana proses dari Q-learning bekerja!

   Q-Learning merupakan algoritma *off-policy* yang bekerja menggunakan metode *value-based* dan menggunakan *temporal difference*. Q-Learning bertujuan untuk mencari/mengoptimalkan Q-Function yang akan memengaruhi jalan yang diambil sebuah agen.

   Di dalam Q-Learning terdapat hyperparameter seperti epsilon dan discount. Epsilon akan mempengaruhi pemilihan state selanjutnya (exploitation vs exploration). Exploitation merupakan teknik pemilihan langkah secara greedy dan exploration adalah pemilihan acak langkah selanjutnya. Hal ini diperlukan agar agen tidak stuck pada local minima. Di sisi lain, discount adalah parameter yang menentukan seberapa besar kita memberi bobot pada reward yang akan datang. Semakin besar discount, semakin kecil pertimbangan reward masa depan.

   Langkah-langkah algoritma Q-Learning adalah sebagai berikut:

   - **Menginisialisasi Q-Function** dengan nol. Q-Function merupakan matrix dengan baris adalah letak state (dalam hal ini 0-9) dan kolom adalah gerak (kiri dan kanan).
   - **Pilih aksi** berdasarkan strategi epsilon-greedy (exploitation vs exploration). Pada implementasi tugas ini, epsilon akan berkurang semakin bertambahnya iterasi. Jadi, agen cenderung melakukan eksplorasi pada awal-awal iterasi saja.
   - Ketika agen mengambil aksi A_t, agen mendapatkan reward R_{t+1} sehingga berada pada state S_{t+1}. Berdasarkan informasi tersebut, kita **melakukan update nilai Q-Function** dengan formula sebagai berikut.

     $$V(S_t) \rightarrow V(S_t) + \alpha[R_{t+1} + \gamma \cdot V^*(S_{t+1}) - V(S_t)]$$

     dengan α adalah learning rate dan γ adalah discount. $V(S_t)$ adalah value pada (S_t, A_t) pada matriks Q-Function, sedangkan $V^*(S_{t+1})$ adalah nilai maksimum pada Q-Function dengan baris S_{t+1}.
   - Update akan **dilakukan terus menerus sampai sebuah episode berakhir** (dalam kasus ini, skor < -200 atau skor > 500). Ketika episode berakhir, skor akan mulai dari nol dan agen akan kembali ke tempat awal.

3. Jelaskan perbedaan algoritma Q-learning dan algoritma SARSA, serta kelemahan dan kekurangan masing-masing algoritma!

   - Q-Learning menggunakan strategi yang **berbeda** saat melakukan **update** Q-Function dan saat **mengambil langkah berikutnya**.
   - SARSA Menggunakan strategi yang **sama** saat melakukan **update** Q-Function dan saat **mengambil langkah berikutnya**.
   - Q-Learning memiliki variansi per sampel yang lebih besar daripada SARSA sehingga bisa menimbulkan masalah saat konvergensi.
   - SARSA **bersifat lebih konservatif** daripada Q-Learning karena algoritma SARSA akan cenderung menghindari jalan yang dekat dengan jebakan (cliff world problem) terutama jika peluang eksplorasinya tinggi.
   - Secara umum, Q-Learning **lebih optimal** daripada SARSA.