Reinforcement Learning David Silver - Lecture 5 Notes: Model-Free Control

Name: Eli Andrew

- **MC Learning On Policy**
- **MC Learning Off Policy**
- **TD Learning On Policy**
    - **SARSA**
        - Act ϵ-greedy with respect to current Q on every iteration
        - $Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$
- **TD Learning Off Policy**
    - **Q-learning**
        - No importance sampling required
        - Next action chosen according to behavior policy: $A_{t+1}\ \mu(.|S_t)$
        - But we consider successor action A' according to target policy: $A'\ \pi(.|S_t)$
        - And update $Q(S_t, A_t)$ towards value of alternative action
        - $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A') - Q(S_t, A_t))$
- **N-step Learning On Policy**
- **N-step Learning Off Policy**