# Introduction to Deep Q-Learning

Alexander Van de Kleut[1]

[1] [1]NeuroCog Lab Cheriton School of Computer Science University of Waterloo

IEEE Control Systems Magazine, 2002

# Outline

# The Reward Hypothesis

All of reinforcement learning is based on the idea that:

### Reward Hypothesis

Every action of a rational agent can be thought of as seeking to maximize some cumulative scalar reward signal

We formalize this idea using a **Markov Decision Process**.

# Markov Process

- A **Markov process** is formally a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$
- $\mathcal{S}$ is a set of states
- $\mathcal{P} : \mathcal{S}^2 \to [0, 1]$ is a transition probability distribution
  - $$\mathcal{P}(s, s') = \mathbb{P}\left[s' | s\right]$$

    the probability of transitioning to state $s'$ given the current state $s$
- Markov processes are used model stochastic sequences of states $s_1, s_2, \ldots, s_T$ satisfying the **Markov property**:
  - $$\mathbb{P}\left[s_{t+1} | s_1, s_2, \ldots, s_t\right] = \mathbb{P}\left[s_{t+1} | s_t\right]$$

    the probability of transitioning from state $s_t$ to state $s_{t+1}$ is independent of previous transitions.
- We can generate **trajectories** of states using $\mathcal{P}$ of the form $\langle s_1, s_2, \ldots, s_T \rangle$

# Markov Reward Process

- A **Markov reward process** is formally a tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R} \rangle$ that allows us to associate with each state transition $\langle s_t, s_{t+1} \rangle$ some reward.
- $$\mathcal{R}(s_t, s_{t+1}) = \mathbb{E}\left[r_t | s_t, s_{t+1}\right]$$
  where $r_t$ is the "instantaneous reward"
- We often simplify this to $\mathcal{R}(s_t)$ the reward of being in a particular state $s_t$
- Given a trajectory beginning at time step $t$ $\langle s_t, s_{t+1}, \ldots, s_T \rangle$ there is an associated sequence of rewards $\langle r_t, r_{t+1}, \ldots, r_T \rangle$
- According to the reward hypothesis, we are interested in trajectories of states that maximize the **return** $R_t$

# Return and Discounted Return

- The **return** $R_t$ is just the cumulative rewards along a trajectory beginning at time stept $t$
- $$R_t = \sum_{k=t}^{T} r_k$$
- For finite $T$, we say the trajectory has a **finite time horizon** and is **episodic**
- For infinite $T$ (trajectories are never-ending) we say the trajectory has an **infinite time horizon**
- In this case, $R_t$ might not converge
- Instead we use the **discounted return** $G_t$
- $$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$$
- where $\gamma$ is a discount factor between 0 and 1

# Value Function

- We can use the expected value of $G_t$ to determine the **value** of being in a state $s_t$
- $$V(s_t) = \mathbb{E}\left[G_t | s_t\right]$$
- We can decompose $V(s_t)$ into two parts: the immediate reward $r_t$ and the discounted value of being in the next state $s_{t+1}$
- $$\begin{aligned} V(s_t) &= \mathbb{E}\left[G_t | s_t\right] \\ &= \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t\right] \\ &= \mathbb{E}\left[r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \ldots) | s_t\right] \\ &= \mathbb{E}\left[r_t + G_{t+1} | s_t\right] \\ V(s_t) &= \mathbb{E}\left[r_t + V(s_{t+1}) | s_t\right] \end{aligned}$$

  which is known as the **Bellman Equation**