Literature review: Reinforcement Learning

Daniel Hernandez

Contents

1 Introduction

subfiles

Reinforcement learning (RL) is an optimization framework. A problem can be considered a reinforcement learning problem if it can be framed in the following way: Given an enviroment in which an agent can take actions, receiving a reward for each action, find a policy that maximizes the expected cumulative reward that the agent will obtain by acting in the environment.

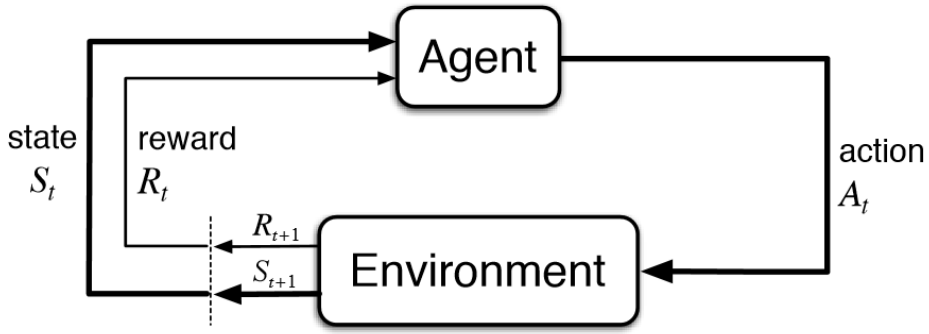

Figure 1: Reinforcement learning loop

1.1 Markov Decision Processes

subfiles

The most famous mathematical structure used to represent reinforcement learning environments are Markov Decision Processes (MDP) (Bellman, 1957). Bellman introduced the concept of a Markov Decision Process as an extension of the famous mathematical construct of Markov chains. Markov Decision Processes are a standard model for sequential decision making and control problems. An MDP is fully defined by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}(\cdot|\cdot,\cdot), \mathcal{R}(\cdot,\cdot), \gamma)$. Whereby:

- $\mathcal{S}$ is the set of states $s \in \mathcal{S}$ of the underlying Markov chain, where $s_t \in \mathcal{S}$ represents the state of the environment at time t.
- $\mathcal{A}$ is the set of actions $a \in \mathcal{A}$ which are the transition labels between states of the underlying Markov chain. $A_t \subset \mathcal{A}$ is the subset of available actions in state s_t at time t. If an state s_t has no available actions, it is said to be a *terminal* state.
- $\mathcal{P}(s_{t+1}|s_t, a_t) \in [0,1]$, where $s_t, s_{t+1} \in \mathcal{S}$, $a_t \in \mathcal{A}$. $\mathcal{P}$ is the transition probability function[1]. It defines the probability of transitioning to state s_{t+1} from state s_t after performig action a_t. Thus, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to [0,1]$. Given a state s_t and an action a_t at time t, we can find the next state s_{t+1} by sampling from the distribution $s_{t+1} \sim P(s_t, a_t)$.
- $\mathcal{R}(s_t, a_t, s_{t+1}) \in \mathbb{R}$, where $s_t, s_{t+1} \in \mathcal{S}$, $a_t \in \mathcal{A}$. $\mathcal{R}$ is the reward function, which returns the immediate reward of performing action a_t in state s_t and ending in state s_{t+1}. The real-valued reward[2] r_t is typically in the range $[-1,-1]$. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. If the environment is deterministic, the reward function can be rewritten as $\mathcal{R}(s_t, a_t)$ because the state transition defined by $\mathcal{P}(s_t, a_t)$ is deterministic.
- $\gamma \in [0,1]$ is the discount factor, which represent the rate of importance between immediate and future rewards. If $\gamma = 0$ the agent cares only about the immediate reward, if $\gamma = 1$ all rewards r_t are taken into account. γ is often used as a variance reduction method, and aids proofs in infinitely running environments (Sutton et al., 1999).

[1] The function $\mathcal{P}$ is also known in the literature as the transition probability kernel, the transition kernel (Tamar et al., 2017) or the dynamics of the environment in model-based contexts. The word kernel is a heavily overloaded mathematical term that refers to a function that maps a series of inputs to value in $\mathbb{R}$.

[2] The reward r_t can be equivalently written as $r(s_t, a_t)$.

The environment is sometimes represented by the Greek letter ξ. The tuple of elementes introduced above are the core components of any environment ξ, but a lot of work in RL literature also presents a distribution over initial states ρ_0 of the MDP, So that the initial state can be sampled from it: $s_0 \sim \rho_0$.

An environment can be episodic if it presents terminal states, or if there are a fixed number of steps after which the environment will not accept any more actions. However environments can also run infinitely (TODO add examples).

Acting inside of the environment, there is the agent, and through its actions the transitions between the MDP states are triggered, advancing the environment state and obtaining rewards. The agent's behaviour is fully defined by its policy π. A policy $\pi(a_t|s_t) \in [0, 1]$, where $s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$ is a mapping from states to a distribution over actions. Given a state s_t it is possible to sample an action a_t from the policy distribution $a_t \sim \pi(s_t)$. Thus, $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

The reinforcement learning loop presented in figure 1 can be represented in algorithmic form as follows:

Algorithm 1: Reinforcement Learning loop.

Sample initial state from the initial state distribution $s_0 \sim \rho_0$;
$\mathtt{t} \leftarrow 0$;
repeat
  Sample action $a_t \sim \pi(s_t)$;
  Sample successor state from the transition probability function $s_{t+1} \sim P(s_t, a_t)$;
  Sample reward from reward function $r_t \sim R(s_t, a_t, s_{t+1})$;
  $\mathtt{t} \leftarrow \mathtt{t} + 1$;
until *Termination*;

For an episode of length T, The objective for the agent is to find an *optimal* policy π^*, which maximizes the cumulative sum of (possibly discounted) rewards.

$$\pi^* = \max_{\pi} \mathbb{E}_{s_0 \sim \rho_0, s \sim \xi, a \sim \pi}[\sum_{t=0}^{T} \gamma r_t] \tag{1}$$

The equation 1 above, algorithm 1 and figure 1 represent the same concept. All of the reinforcement learning research focuses on solving this problem.

There are two functions of special relevance in reinforcement learning, the *state value* function $V^\pi(s)$ and the *action value* function $Q^\pi(s, a)$:

- The state value function $V^\pi(s)$ under a policy π, where $s \in \mathcal{S}$, represents the expected sum of rewards obtained by starting in state s and following the policy π until termination. Formally defined as $V^\pi(s) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} r(s_t, a_t)|s_0 = s]$
- The state-action value function $Q^\pi(s, a)$ under a policy π, where $s \in \mathcal{S}$, $a \in \mathcal{A}$, represents the expected sum of rewards obtained by performing action a in state s and then following policy π. Formally defined as: $Q^\pi(s, a) = \mathbb{E}^\pi[r(s_0, a_0) + \sum_{t=1}^{\infty} r(s_t, a_t)|s_0 = s, a_0 = a]$

The Bellman equations are the most straight forward, dynamic programming approach at solving MDPs (Bertsekas, 2007; Bellman, 1957).

1.2 Bellman equations and optimality principle

The optimality principle, found in Bellman (1957), states the following: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. The optimality principle, coupled with the proof of the existance of a deterministic optimal policy for any MDP as outlined in (Borkar, 1988) give rise to the optimal state value function $V^*(s) = \text{argmax}_\pi V^\pi(s) = V^{\pi^*}(s)$ and the optimal action value function $Q^*(s, a) = \text{argmax}_\pi Q^\pi(s, a) = Q^{\pi^*}(s, a)$. The optimal value functions determine the best possible performance in a MDP. An MDP is considered *solved* once the optimal value functions are found.

Most of the field of reinforcement learning research focuses on approximating these two equations (Tamar et al., 2017) (Watkins and Dayan, 1992) (Mnih et al., 2013). (cite many more)

Bellman (1957) outlined two analytical equations for the state value and action value function:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) * (r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) * V^{\pi}(s')) \tag{2}$$

$$Q^{\pi}(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) * (\sum_{a' \in \mathcal{A}} \pi(a'|s') Q^{\pi}(s',a')) \tag{3}$$

2 Environment models

subfile

Even though Markov Decision Processes are the most famous mathematical structure used to model an environment in reinforcement learning, there are other types of possible models for RL environment which act as extensions to vanilla MDPs. This section concerns itself to defining these extensions, and making links between them.

This is not an exhaustive list of all possible mathematical models used to represent environments in the RL literature. However, these are some of the most used or fundamental models in the field, on which the majority of the research is conducted, and on top of which most niche extensions are built.

2.1 Semi Markov Decision Process (SMDP)

As stated in (Barto, 2003), in an MDP, only the sequential nature of the decision process is relevant, not the amount of time that passes between decision points. Semi Markov Decision Processes are a generalization of this idea, where the time elapsed in between decision points, also known as decision stages, is taken into account. Every action taken in an SMDP has an assigned delay τ, known as *holding time*. When an action with holding time τ is decided at time t, the agent waits for τ timesteps before the action is executed and the next decision point is reached. At this decision point, the agent recieves state s‘ and a cumulative reward which includes the individual reward of all elapsed timesteps, $r = \sum_{l=t}^{t+\tau} r_{t+l}$. The time until the next decision point τ can only depend on the action a and state t and thus τ is independent of the history of the environment. SMDPs can also be used for real-valued time systems instead of discretely timed environments. This holding time allows for a gap in time between sensorial input reaching the agent and the agent's action being executed on the environment.

This type of process is considered Semi Markovian because as the holding time is elapsing, the agent cannot know how the system is evolving. Thus, in order to determine when the next state (decision point) will be reached, it is necessary to know how much time has elapsed, introducing temporal dependency, breaking the Markov property. To iterate on this point, the probability of reaching state state $s_{t+\tau}$ depends only on s_t and action a_t with associated holding time τ. Once the action a_t has been decided, estimating when the agent will recieve state $s_{t+\tau}$ depends on how much time has elapsed since the action a_t was decided.

A Semi Markov Decision Process is defined by a 5-element tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}(\cdot,\cdot|\cdot,\cdot), \mathcal{R}(\cdot,\cdot,\cdot,\cdot), \gamma)$:

- $\mathcal{S}, \mathcal{A}$ and γ express the same concepts as in classical MDPs.
- $\mathcal{P}(s', \tau|s, a)$, where $s', s \in \mathcal{S}, \tau \in \mathbb{N}, a \in \mathcal{A}$, is the transition probability function. Which states the probability of transitioning to state s‘ after a holding time of τ.
- $\mathcal{R}(s, a, s', \tau)$, where $s', s \in \mathcal{S}, a \in \mathcal{A}$, is the reward function. It represents the expected reward of deciding on action a on state s with an assigned holding time of τ timesteps and landing on state s'.

A useful properties of SMDPs is that they can be reduced to regular MDPs through the *data-transformation method* (Piunovskiy and Zhang, 2012). This introduces the possibility of using MDP solving methods to solve SMDPs. SMDPs have recieved a lot of attention in the field hierarchical learning, especially with regards to options (Sutton and Barto, 1998).

2.2 Partially Observable Markov Decision Process (POMDP)

In an MDP, the internal representation of the environment is the same representation that the agent receives at every timestep. POMDPs introduce the idea that what the agent observes at every timestep t is only a partial representation o_t of the real environment state s_t. This partial observation o_t alone is not

enough to reconstruct the real environment state s_t, which entails that $o_t \subset s_t$. A Partially Observable Markov Decision process is defined by a 6-element tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}(\cdot|\cdot,\cdot), \mathcal{R}(\cdot,\cdot), \Omega, \mathcal{O}(\cdot|\cdot,\cdot), \gamma)$:

- $\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}$ and γ express the same concepts as in classical MDPs.
- Ω is the set of all possible agent observations. Notably $\Omega \subset \mathcal{S}$, meaning that some of the state properties may never be available to the agent.
- $\mathcal{O}(o|s,a)$, where $o \in \mathcal{O}, s \in \mathcal{S}, a \in \mathcal{A}$, represents the probability of the agent recieving observation o after executing action a in state s.

In an POMDP the goal of the environment is not to find an optimal policy π^* which will maximize the expected cumulative reward. This is because the observations that the agent receives as it acts inside of the environment may not be informative enough to allow the agent to learn an optimal policy. Because of this, the goal of the agent becomes to find an optimal policy that maximizes the cumulative reward *conditioned* on the history of observations that the agent has recieved from the environment. The agent samples actions from its policy, which is no longer conditioned on the state of the environment, as the agent does not have access to it, but rather it is conditioned on the sequence of the observations that the agent has obtained so far, $a_t \sim \pi(o_{\leq t})$. This goal is formalized as:

$$\pi = \max_{\pi} \mathbb{E}_{s_0 \sim \rho_0, s \sim \xi, a \sim \pi(\cdot|o_{\leq t})}[\sum_{t=0}^{T} \gamma r_t] \tag{4}$$

A POMDP can be reduced to an MDP iff, for all timesteps t the agent's observation o_t and the environment state s_t are equal $o_t = s_t$.

2.3 Decentralized Markov Decision Process (dec-MDP)

A major shortcoming of MDPs is that they assume stationary environments, which by definition entails that the environment does not change over time. This assumption makes MDPs unsuitable for modelling multi-agent environments. Agents must be considered as non-stationary parts of the environment, because their behaviour changes over time through the course of learning.

Dec-POMDPs form a framework for multiagent planning under uncertainty (Oliehoek and Amato, 2014). This uncertainy comes from two sources. The first one being the partial observability of the environment, the second one stemming from the uncertainy that each agent has over the other agent's policies. It is the natural multi-agent generalization of POMDPs, introducing multi-agent concepts from game theory. They are considereded *decentralized* because there is no explicit communication between agents. Agents do not have the explicit ability of sharing their observations and action choices with each other. Every agent bases its decision purely on its individual observations. On every timestep each agent chooses an action simultaneously and they are all collectively submitted to the environment. It is important to note that all agents share the same reward function, making the nature of dec-POMDPs collaborative. A decentralized Partially Observable Markov Decision Process is defined by an 8-element tuple $(I, \mathcal{S}, \mathcal{A}_{1..k}, \mathcal{P}(\cdot|\cdot,\cdot), \mathcal{R}(\cdot,\cdot), \Omega_{1..k}, \mathcal{O}(\cdot|\cdot,\cdot), H)$:

- $\mathcal{S}$ and express the same concepts as in classical MDPs.
- I is the set of all agent policies present in the dec-POMDP, indexed $1..k$. As stated in Section 1.1, agent's are fully defined by their policies.
- $\mathcal{A}_{1..k}$ is a collection of action sets, on for each agent in the environment, with $\mathcal{A}_i$ corresponding to the action set of the ith agent.
- $\mathcal{P}(s'|s,\mathbf{a})$, where $s \in \mathcal{S}$, $\mathbf{a} = \{a_1, \ldots, a_k\}$ and $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$, represents the probability transition function. It states the probability of transitioning from state s to state s' after executing the *joint* action $\mathbf{a}$. The joint action is a vector containing the action performed by every agent at a certain timestep.
- $\mathcal{R}(s,\mathbf{a})$, where $s \in \mathcal{S}$, $\mathbf{a} = \{a_1, \ldots, a_k\}$ and $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$, represents the reward function associated to the ith agent, indicating the reward that the ith agent will obtain after the joint action vector $\mathbf{a}$ is executed in state s.

- $\Omega_{1..k}$ represents the joint set of all agent observations, with Ω_i representing the set of all possible observations for the ith agent.
- $\mathcal{O}(\mathbf{o}|s, \mathbf{a})$, where $\mathbf{o} = \{o_1, \ldots, o_k\}$, $o_1 \in \Omega_1, \ldots, o_k \in \Omega_k$, $s \in \mathcal{S}$, $\mathbf{a} = \{a_1, \ldots, a_k\}$ and $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$, represents the probability of observing the joint observation vector $\mathbf{o}$, containing an observation for each agent, after executing the joint action $\mathbf{a}$ in state s.
- $H \in \mathbb{N}$ represents the finite time horizon, the number of steps over which the agents will try to maximize their cummulative reward.

A dec-POMDP featuring a single agent, $|I| = 1$, can be treated as a POMDP. Furthermore, when agents are permitted to have different reward functions, this model becomes a Partially Observable Stochastic Game (POSG) (Hansen et al., 2004).

2.4 Markov Game

Owen and Owen (1982) first introduced the notion of a Markov Game. Markov Games also serve to model multi-agent environments. They came to be as a crossbreed between game theoretic structures such as extended-form games and Markov Decision Processes. A Markov Game with k different agents is denoted by a 5-element tuple $(\mathcal{S}, \mathcal{A}_{1..k}, \mathcal{P}(\cdot|\cdot,\cdot), \mathcal{R}_{1..k}(\cdot,\cdot), \gamma)$

- $\mathcal{S}$ and γ express the same concepts as classical MDPs.
- $\mathcal{A}_{1..k}$ is a collection of action sets, on for each agent in the environment, with $\mathcal{A}_i$ corresponding to the action set of the ith agent.
- $\mathcal{P}(s'|s, \mathbf{a})$, where $s \in \mathcal{S}$, $\mathbf{a} = \{a_1, \ldots, a_k\}$ and $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$, represents probability transition function. It states the probability of transitioning from state s to state s' after executing the *joint* action $\mathbf{a}$. The joint action represents all of the actions that were executed at timestep t.
- $\mathcal{R}_i(s, \mathbf{a})$, where $s \in \mathcal{S}$, $\mathbf{a} = \{a_1, \ldots, a_k\}$ and $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$, represents the reward function associated to the ith agent, indicating the reward that the ith agent will obtain after the joint action $\mathbf{a}$ is executed in state s.

Each agent independently tries to maximize its expected discounted cumulative reward, $\mathbb{E}[\sum_{j=0}^{\infty} \gamma^j r_{i,t+j}]$, where $r_{i,t+j}$ is the reward obtained by agent i at time $t + j$.

Markov Games have several important properties Owen and Owen (1982); Littman (1994). Like MDP's, Every Markov game features an optimal policy for each agent. Unlike MDPs, these policies may be *stochastic*. The need for stochastic action choice stems from the agent's uncertainty about the opponent's pending moves. On top of this, sthocastic policies make it difficult for opponents to "second guess" the agent's action, which makes the policy less exploitable.

When the number of agents in a Markov Game is exactly 1, the Markov Game can be considered an MDP. When $|\mathcal{S}| = 1$, the environment can be considered a normal-form game from game theory literature. TALK ABOUT GAMA BEING THE PROBABILITY OF ENDING THE EPISODE AND THE KEY DIFFERENCE AGAINST DEC-pomdpS THAT WE HAVE PER AGENT REWARD FUNCTIONS.

3 Categorization of RL algorithms

subfiles

Explain control vs prediction.

Most RL algorithms can be devided into the following categories:

3.0.1 Policy based

These algorithms tend to represent a policy through a parameter vector $\theta \in \mathbb{R}^D$. The goal then becomes to improve the choice of parameters $\theta_i \in \theta$ to improve the expected sum of rewards $\mathbb{E}[\sum_T r(t)]$. Policy based algorithms are the main focus on Section 5.

Famous policy based algorithms: vanilla policy gradient, REINFORCE(Williams, 1992) and the REINFORCE family of algorithms.

3.0.2 Value based

Value based, or critic only methods, rely exclusively on value function approximation and aim at learning an approximate solution to the Bellman equation. The underlying assumption is that from the value function, a near optimal policy can be computed.

Famous value based algorithms: Value iteration, Policy iteration and Sarsa Sutton and Barto (1998), Q-learning (Watkins, 1989)

3.0.3 Actor critic

Actor critic methods combine the strong points of both policy based and value based algorithms, and overcome some of their individual weaknesses. The critic assumes the role of learning a value function, which is then used as part of the update for the actor's policy. The individual critic is analogous to value based algorithms, and the actor to policy based methods.

Famous actor critic algorithms: A3C (Mnih et al., 2016), A2C a variant of A3C where a single worker is used, PPO (Schulman et al., 2017), TRPO (Schulman 2015, find paper), ACKTR (Wu et al., 2017).

3.0.4 Model based and model free approaches

In the reinforcement learning literature the *model* or the dynamics of an environment is considered to be the transition function $\mathcal{P}(s_{t+1}|s_t, a_t)$ and reward function $\mathcal{R}(s_t, a_t, s_{t+1})$. Model free algorithms aim to approximate an optimal policy π^* without explicitly using either $\mathcal{P}$ or $\mathcal{R}$ in their calculations.

A further categorization of algorithms is the notion of *model free* and *model based* algorithms. Consider a *model* of an environment to be the transition function $\mathcal{P}$ and reward function $\mathcal{R}$. Model free algorithms aim to approximate an optimal policy without them. Model based algorithms are either given a prior model that they can use for planning (Browne et al., 2012; Soemers, 2014), or they learn a representation via their own interaction with the environment (Sutton, 1991; Guzdial et al., 2017). Note that an advantage of learning your own model is that you can choose a representation of the environment that is relevant to the agent's actions, which can have the advantage of modelling uninteresting (but perhaps complicated) environment behaviour (Pathak et al., 2017).

Model based algorithms are either given a prior model that they can use for planning (Browne et al., 2012; Soemers, 2014), or they learn a representation via their own interaction with the environment (Sutton, 1991; Guzdial et al., 2017; Deisenroth and Rasmussen, 2011). Note that an advantage of learning your own model is that you can choose a representation of the environment that is relevant to the agent's actions, which can have the advantage of modelling uninteresting (but perhaps overly complicated) environment behaviour (Pathak et al., 2017). Another advantage of having a model is that it allows for forward planning, which is the main method of learning for search-based artificial intelligence (throw mcts papers here).

3.0.5 on-policy, off-policy

On policy methods use a policy π to sample experiences on an environment in order to improve that same policy π. Off policy methods use a behavioural policy μ to carry out actions in an environment, and use this information to improve a target policy π. The learning that takes place in off policy methods can be regarded as learning from somebody else's experience, whilst on policy methods focus on learning from an agent's own experience.

Historically On-policy and off-policy methods have been used separatelly. Sutton and Barto (1998) describes SARSA, Monte Carlo for value function estimation, Monte Carlo for prediction and Monte Carlo Tree Search (MCTS) methods as on policy methods. In the field of deep reinforcement learning, Williams (1992) introduces de REINFORCE algorithm family of policy gradient methods as on-policy methods.

A limitation of on-policy learning is that it forces the agent to commit its resources to learning *only* a policy. Off-policy learning can be used to learn multiple tasks in parallel. Sutton et al. (2011) use sensorimotor interaction with an environment to learn a multitude of pseudoreward funcitons (READx paper). Jaderberg et al. (2016) takes this idea further by using an off-policy methods to learn auxiliary extra tasks. Most notably, these tasks include predicting immediate rewards[3], pixel control[4].

[3]This is different from value function estimation because the value that the off-policy algorithm is trying to predict is expected immediate reward, instead of expected future cummulative reward.

[4]Given a matrix of pixels as input, the authors define pixel control as a separate policy that tries to maximally change the pixels in the following state. The reasoning behind this approach is that big changes in pixel values may correspond to

A method to alllow algorithms to perform off policy updates to their policies is to introduce the notion of an *experience replay* (Lin, 1993), which was made famous after the success of Mnih et al. (2013). An experience replay is a list of experiences, where each experience is a 5 element tuple $< s_t, a_t, r_t, s_{t+1}, a_{t+1} >$. As an agent acts in an environment, in the same fashion as in the reinfocement learning loop presented in (reference), the experience replay is filled. At the time of updating the policy, the agent does not choose to update its policy using the last action function that was taken, as it is the case with Q-learning. Instead, the agent samples an experience (or batch of experiences) from the experience replay. Because these sampled experiences may have been generated using a previous policy, experience replay allows for policy updates to happen in an off-policy fashion. The experience replay buffer has been the focus on future research such as (Schaul et al., 2015; Hessel et al., 2017) where the authors use a *prioritized* experience replay. The difference is that experiences are not sampled uniformly from the replay buffer. Instead, experiences are weighted according to (READ paper). Andrychowicz et al. (2017) expands on the idea by introducing the Hindsight Experience Replay (HER) where an experience replay is used to learn a task from failures by treating certain state transitions as goals.

Some notable on policy algorithms: Sarsa, Monte Carlo estimation for prediction, Monte Carlo Tree search methods, Many vanilla policy gradient methods such as REINFORCE(Williams, 1992) family of algorithms, asynchronous advantage actor critic method (A3C and A2C). $Q(\sigma)$ (De Asis et al., 2017).

Some notable off policy algorithms are. Q-learning, deep Q-learning, deterministic policy gradient (Silver et al., 2014). Deep deterministic policy gradient (Lillicrap et al., 2015)

3.1 Common problems in Reinforcement Learning

3.1.1 Credit assignment problem

As we have previously discussed, one of the difficulties of reinforcement learning is that feedback for actions is often delayed. The credit assignment problem focuses on the question: if a sequence of actions led to a reward, how much credit should each action take for obtaining that reward? It is against common sense to assume that every action should be equally rewarded or punished. This is commonly known as the credit assignment problem.

3.1.2 Exploration vs exploitation

The exploration vs exploitation dilemma corresponds to finding a good middle point between exploting the best action available at any given time (exploitation), or performing a sub optimal action in the hopes that the agent will learn more about the environment that could lead to higher expected rewards in the long run. A common approach in stochastic environments, the trade off between exploration and exploitation is tackled by the following idea. For each state $s_t \in \mathcal{S}$ we want to maximize the expected sum of rewards $\mathbb{E}_\pi[V_\pi(s_t)]$ and reduce the variance $\sigma^2_{s_t}$. The Upper Bound Confidence Monte Carlo Tree Search algorithm UCT-MCTS has a remarkable example of this tradeoff as part of its selection policy (explain selection policy or just chuck equation?):

4 Q-learning

subfiles

The Q-learning algorithm was first introduced by Watkins (1989), and is arguably one of the most famous and widely implemented methods in the entire field. Given an MDP, Q-learning aims to calculate the corresponding optimal action value function Q^*, following the principle of optimality and proof of the existence of an optimal deterministic policy in an MDP as described in Section 1.1. It is model free, learning via interaction with the environment, and it is an off-policy algorithm. The latter is because, even though we are learning the optimal action value function Q^*, we can choose any *behavioural* policy to gather experience from the environment. Researchers like Tijsma et al. (2017) benchmarked the efficiency of using various exploratory policies in grid world stochastic maze environments.

Q-learning has been proven to converge to the optimal solution for an MDP under the following assumptions:

1. The Q^{π^*} function is represented in tabular form, with each state-action pare represented discretely Watkins and Dayan (1992).

important events inside of the environment.

2. Each state-action pair is visited an infinite number of times. (Watkins, 1989)

3. The sequence of updates of Q-values has to be monotonically increasing $Q(s_i, a_i) \leq Q(s_{i+1}, a_{i+1})$. (Thrun and Schwartz, 1993).

4. The learning rate α must decay over time, and such decay must be slow enough so that the agent can learn the optimal Q values. Expressed formally: $\sum_t \alpha_t = \infty$ and $\sum_t (\alpha_t)^2 < \infty$. (Watkins, 1989)

Algorithm 2: Q-learning

1 Initialize Q table $\forall s \in \mathcal{S} \wedge \forall a \in \mathcal{A}$, $Q(s, a) = 0$;
2 Sample $s_0 \sim \rho_0$;
3 $s = s_0$;
4 **repeat**
5 | Select action `a` $= \pi(s)$, where $\pi(s) = \epsilon - greedy(Q(s, \cdot))$;
6 | Observe successor state `s'` and reward `r` after taking action `a` ;
7 | Update $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \text{argmax}_{a'} Q(s', a') - Q(s, a)]$;
8 **until** *termination*;

Q-learning features its own share of imperfections. If there is a function approximator[5] in place, Thrun and Schwartz (1993) shows that if the approximation error is greater than a threshold which depends on the discount factor γ and episode length, then a systematic overestimation effect occurs, negating convergance. This is mainly due to the joint effort of function approximation methods and the $argmax$ operator used in step 7 of the algorithm. On top of this, Kaisers and Tuyls (2010) introduces the concept of *Policy bias*, which states that state-action pairs that are favoured by the policy are chosen more often, biasing the updates. Ideally all state-action pairs are updated on every step. However, because agent's actions modify the environment, this is generally not possible in absence of an environment model. Frequency Adjusted Q-learning (FAQL) proposes scaling the update rule of Q-learning inversely proportional to the likelihood of choosing the action taken at that step (Kaisers and Tuyls, 2010). Abdallah et al. (2016) introduces Repeated Update Q-learning (RUQL), a more promising Q-learning spin off that proposes running the update equation *multiple times*, where the number of times is inversely proportional to the probability of the action selected given the policy being followed.

5 Policy gradient methods

subfiles

5.1 Policy gradient theorem

In order to comply with notation used in the field of direct optimization, we shall use u for actions.

Consider a stochastic control policy $\pi_\theta(s)$ parameterized by a parameter vector θ, that is, a distribution over the action set $\mathcal{A}$ given a state $s \in \mathcal{S}$. θ is a D-dimensional real valued vector, $\theta \in \mathbb{R}^D$, where D is the number of parameters / dimensions and $D << |\mathcal{S}|$. This prameterized policy function will be denoted by π_θ.[6].

There are strong motivations for using policy gradient approaches versus the already discussed RL methods:

1. A more direct way of approaching the problem. Instead of computing the value functions V or Q and from those deriving a policy function, we are calculating the policy function directly.

2. Using stochastic policies smoothes the optimization problem. With a deterministic policy, changing which action to do in a given state can have a dramatic effect on potential future rewards[7]. If we assume a stochastic policy, shifting a distribution over actions slightly will only slightly modify the potential future rewards. Furthermore, Many problems, such as partially observable environments or adversarial settings have stochastic optimal policies (Degris et al., 2012).

[5] With neural networks being the most famous function approximators in reinforcement learning at the time of writing.
[6] Some researchers prefer the notation $\pi(\cdot, \theta)$, $\pi(\cdot \mid \theta)$ or $\pi(\cdot; \theta)$. These notations are equivalent.
[7] An example of this concept are *greedy* or *ϵ-greedy* policies derived thus: $\pi(s) = \text{argmax}_{a \in \mathcal{A}} Q(s, a)$.

3. Often π can be simpler than V or Q.

4. If we learn Q in a large or continuous actions space, it can be tricky to compute $\underset{u}{argmax}\ Q(s,u)$.

(they are not, though. REINFORCE is, but dqn, ddpg arent!) Policy gradient methods are on-policy. In them, the agent acts with using a policy which is improved gradually over time. This contrasts with off-policy algorithms, such as the Q-learning algorithm introduced in Section 4, which allows the agent to interact with the environment with a policy while it is simultaneously learning another policy. There is ongoing research looking at off-policy variants of policy gradient methods (Mnih et al., 2013, 2016).

Let's assume an stochastic environment E from which to sample states and rewards, and an stochastic policy π_θ parameterized by a vector θ from which to sample actions. The agent acting under policy π_θ is to maximize the (possibly discounted)[8] sum of rewards on environment E, over a time horizon H (possibly infinitely long). We reach the following optimization problem:

$$\max_\theta = \mathbb{E}_{s_t\sim E, u_t\sim\pi_\theta}[\sum_{t=0}^{H} r(s_t,u_t)|\pi_\theta] \tag{5}$$

For an episode of length H let τ be the trajectory followed by an agent in an episode. This trajectory τ is a sequence of state-action tuples $\tau = (s_0, a_0, \ldots, s_H, a_H)$. We overload the notation of the reward function $\mathcal{R}$ thus: $\mathcal{R}(\tau) = \sum_{t=0}^{H} r(s_t,u_t)$, indicating the total accumulated reward in the trajectory τ. We will also use $r(s_t) \in \mathbb{R}$ to refer to the scalar reward obtained at timestep t in the trajectory. From here, the utility of a policy parameterized by θ is defined as:

$$U(\theta) = \mathbb{E}_{s_t\sim E, u_t\sim\pi_\theta}[\sum_{t=0}^{H} r(s_t,u_t)|\pi_\theta] = \sum_\tau P(\tau;\theta)\mathcal{R}(\tau) \tag{6}$$

Where $P(\tau;\theta)$ denotes the probability of trajectory τ happening when taking actions sampled from a parameterized policy π_θ. More informally, how likely is this sequence of state-action pairs to happen as a result of an agent following a policy π_θ. Linking equations 5 and 6, our optimization problem becomes:

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)\mathcal{R}(\tau) \tag{7}$$

Policy gradient methods attempt to solve this maximization problem by iteratively updating the policy parameter vector θ in a direction of improvement w.r.t to the policy utility $U(\theta)$. This direction of improvement is dictated by the gradient of the utility $\nabla_\theta U(\theta)$. The update is usually done via the well known gradient descent algorithm. This idea of iteratively improving on a parameterized policy is was introduced by Williams (1992) under the name of *policy gradient theorem*. In essence, the gradient of the utility function aims to increase the probability of sampling trajectories with higher reward, and reduce the probability of sampling trajectories with lower rewards.

Equation 8 presents the gradient of the policy utility function. The Appendix section shows the derivation from equation 6 to equation 8.

$$\nabla_\theta U(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\tau)R(\tau)] \tag{8}$$

A key advantage of the policy gradient theorem, as inspected by Sutton et al. (1999) (and formalized in the Appendix Section), is that equation 8 does not contain any term of the form $\nabla_\theta \mathcal{P}(\tau;\theta)$. This means that we don't need to model the effect of policy changes on the distribution of states. Policy gradient methods therefore classify as model-free methods.

We can use Monte Carlo methods to generate an empirical estimation of the expectation in equation 8. This is done by sampling m trajectories under the policy π_θ. This works even if the reward function R is unkown and/or discontinuous, and on both discrete and continuous state spaces. The equation for the empirical approximation of the utility gradient is the following:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=0}^{m} \nabla_\theta \log \pi_\theta(\tau^{(i)})R(\tau^{(i)}) \tag{9}$$

[8] Williams (1992); Sutton et al. (1999) present proofs of this same derivation using a discount factor, which makes policy gradient methods work for environments with infinite time horizons.

The estimate $\hat{g}$ is unbiased estimate and it works in theory. However it requries an impractical amount of samples, otherwise the approximation is very noisy. In order to overcome this limitation we can do the following tricks:

- Add a baseline
- Add temporal structure (advantage function)
- Use trust region and natural gradient.

5.2 Baselines

Intuitively, we want to reduce the probability of trajectories that are worse than average, and increase the probability of trajectories that are better than average. Williams (1992), in the same paper that introduces the policy gradient theorem, explores the idea of introducing a baseline b as a method of variance reduction, where $b \in \mathbb{R}$. These authors also prove that introducin a baseline keeps the estimate unbiased (have proof in appendix?). It is imporant to note that this estimate is not biased as long as the baseline at time t does not depend on action u_t. Introducing a baseline in equation 8 yields the equation:

$$\nabla_\theta U(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(\tau)(R(\tau) - b)] \tag{10}$$

The most basic type of baseline is the global average reward, which keeps track of the average reward across all episodes. Greensmith et al. (2004) derives the optimal constant value baseline. We can also add time dependency to the baseline. It is not optimal to scale the probability of taking an action by the whole sum of rewards. A better idea is, for a given episode, to weigh an action u_t by the reward obatined from time t onwards, otherwise we would be ignoring the Markov property underlying the environment's Markov Decission Process. This changes equation 10 to:

$$\nabla_\theta U(\theta) = \mathbb{E}_{s_t \sim E, u_t \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(u_t \mid s_t)(\sum_{k=t}^{H-1} R(s_k, u_k) - b)] \tag{11}$$

A powerful idea is to make the baseline state-dependent $b(s_t)$ (Baxter and Bartlett, 2001). For each state s_t, This baseline should indicate what is the expected reward we will obtain by following policy π_θ. By comparing the empirically obtained reward with the estimated reward given by the baseline $b(s_t)$, we will know if we have obtained more or less reward than expected. Note how this baseline is the exact definition of the state value function V_{π_θ}, as shown in equation 12. This type of baseline allows us to increase the log probability of taking an action proportionally to how much its returns are better than the expected return under the current policy.

$$b(s_t) = \mathbb{E}[r_t + r_{t+1} + r_{t+2} + \cdots + r_{H-1}] = V_{\pi_\theta}(s_t) \tag{12}$$

Consider a further improvement: the term $\sum_{k=t}^{H-1} R(s_k, u_k)$ can be regarded as an estimate of $Q_{\pi_\theta}(s_t, u_t)$ for a single roll out. This term has high variance because it is sample based, where the amount of variance depends on the stochasticity of the environment. A way to reduce variance is to include a discount factor γ, rendering the equation: $\sum_{k=t}^{H-1} \gamma^k R(s_k, u_k)$. However, this still keeps the estimation sample based, which means that it is not generalizable to unseen state-action pairs. This issue can be solved by using function approximators to approximate the function Q_{π_θ}. We can define another real valued parameter vector $\phi \in R^F$, where F is the dimensionality of the parameter vector. From here, we can use ϕ to parameterize the function approximator $Q^\phi_{\pi_\theta}$. This function will be able to generalize for unseen state-action pairs.

$$\begin{aligned} Q^\phi_{\pi_\theta}(s, u) &= \mathbb{E}[r_0 + r_1 + r_2 + \cdots + r_{H-1} \mid s_0 = s, u_0 = u] && (\infty\text{-step look ahead}) \\ &= \mathbb{E}[r_0 + V^\phi_{\pi_\theta}(s_1) \mid s_0 = s, u_0 = u] && (1\text{-step look ahead}) \\ &= \mathbb{E}[r_0 + +r_1 + V^\phi_{\pi_\theta}(s_2) \mid s_0 = s, u_0 = u] && (2\text{-step look ahead}) \end{aligned} \tag{13}$$

Notice how we use parameter vector ϕ to approximate the state value function V_{π_θ}. This approach can be viewed as an actor-critic architecture where the policy π_θ is the actor and the baseline b_t is the critic (Sutton and Barto, 1998; Degris et al., 2012) (read these 2 papes). Konda and Tsitsiklis (2000) make the key observation that in actor critic methods, the actor parameterization θ and the critic parameterization ϕ should *not* be independent. The choice of critic parameters should be directly prescribed by the choice of the actor parameters.

5.3 Advantage functions

Let the advantage function $A_\pi(s_t, a_t) \in \mathbb{R}$ be the numerical advantage of taking action a_t in state s_t under policy π. The advantage function is often depicted as:

$$A_\pi(t) = A_\pi(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t) \tag{14}$$

5.4 Trust region optimization, and natural gradient?

5.5 Off-policy policy gradient methods

Off-PAC, The first off-policy policy gradient method introduced by Degris et al. (2012) used importance sampling techniques to weigh the actor gradient update against the behavioural policy being used (look this up on paper). They also used eligibility traces for a critic with linear function approximator, similar to a TD(λ). Reply buffer, introduced in Lin (1993), has seen a lot of use recently (Mnih et al., 2013, 2016).

6 Multi-agent reinforcement learning

subfiles

Reread multi agent deep deterministic policy gradient MADDPG Lowe et al. (2017). Unified game theoretic approach to MA systems (Lanctot et al., 2017)

- Stochastic games.
- Enumerate elements.
- Link to MDP and matrix games.
- Explain that actions are taken simultaneously, as opposed to sequentially
- Notion of strategy in stochastic games and joint strategies. The $\pi = \langle \pi_i, \pi_{-1} \rangle$. And itroduce the π_{-i} suffix notation.
- Best Responose definition
- Nash equilibria
- Playing against stationary policies simplifies SG to MDP.
  - Show V and Q equations.
  - Briefly explain how this notation allows for any algorithm in sections (find sections) can be used.

useful link (where I got all of this info from): http://users.isr.ist.utl.pt/ mtjspaan/readingGroup/learningNeto05.pdf Opponent modeling (Uther and Veloso, 1997). The agent treats all opponents as a single agent that is able to take joint actions.

Stochastic games are an extension of matrix games (von Neumann and Morgenstern, 1947; Owen, 1995) Matrix game iff $\mid S \mid = 1$. So a SG is a succession of matrix games. They also superset Markov Decision Processes. Stochastic games can be thought of as MDPs where the number of agents $n = 1$. Definition of Nash Equilibrium: A Nash equilibrium is a collection of strategies, one for each player, that are best response strategies, which means that none of the players can do better by changing strategy, if all others continue to follow the equilibrium Multiagent goodness

7 Learning Environments

- OpenAI Gym (Brockman et al., 2016)

- Real time strategy games have always been a famous testbed for many reinforcement learning algorithms. The most popular game used in RTS AI research is Starcraft: Broodwar. There are several frameworks that either simulate the game (Heinermann, 2009; Wender and Watson, 2012) or inject code into the game to allow external processes to communicate with the real game (Synnaeve et al., 2016) such as TorchCraft. Vinyals et al. (2017) uses the same concept for Starcraft II. Followed by microRTS (Ontañón, 2013), finishing by saying that (Andersen, 2017) Creates a unifying and highly customizable RTS environment that allows for POMDP with variyng degrees of partial observability. Compare simulation speed.
  - Real time strategy games are really appealing for reinforcement scenarios because a good agent should be able to perform effective search and meaningful planning, over environments with enormous branching factors (Soemers, 2014)(Wender and Watson, 2012).
- Arcade Learning Environment (ALE) (Bellemare et al., 2015)
- VizDoom (Kempka et al., 2017)
- Unity ML-agents (Unity, 2017). Currently only performance boost of 100x.

8 Appendix

subfiles

Take equation 6 from Section **??**, representing the utility of a policy π_θ parameterized by a D-dimensional real valued parameter vector $\theta \in \mathbb{R}^D$

$$U(\theta) = \sum_\tau P(\tau;\theta)R(\tau) \tag{15}$$

The goal is to find the expression $\nabla_\theta U(\theta)$ that will allow us to update our policy parameter vector θ in a direction that improves the estimated value of the utility of the policy π_θ. Taking the gradient w.r.t θ gives:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) && \text{(Move gradient operator inside sum)} \\
&= \sum_\tau \nabla_\theta \frac{P(\tau;\theta)}{P(\tau;\theta)} P(\tau;\theta)R(\tau) && \text{(Multiply by} \frac{P(\tau;\theta)}{P(\tau;\theta)}\text{)} \\
&= \sum_\tau P(\tau;\theta)\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}R(\tau) && \text{(Rearrange)} \\
&= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau) && \text{(Note:} \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} = \nabla_\theta \log P(\tau;\theta)\text{)}
\end{aligned} \tag{16}$$

$$\nabla_\theta U(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log P(\tau;\theta)R(\tau)] \qquad (\mathbb{E}[f(x)] = \sum_x xf(x)) \tag{17}$$

This leaves us with an expectation for the term $\nabla_\theta \log P(\tau;\theta)R(\tau)$. Note that as of now we have not discussed how to calculate $P(\tau;\theta)$. Let's define the probability of a trajectory under a policy π_θ as:

$$P(\tau;\theta) = \Pi_{t=0}^{H} \underbrace{P(s_{t+1}|s_t,u_t)}_{\text{dynamics models}} \underbrace{\pi_\theta(u_t|s_t)}_{\text{policy}} \tag{18}$$

From here we can calculate the term $\nabla_\theta \log P(\tau;\theta)$ present in equation 16

$$\begin{aligned}\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log[\Pi_{t=0}^{H} P(s_{t+1}|s_t,u_t)\pi_\theta(u_t|s_t)] \\ &= \nabla_\theta[(\sum_{t=0}^{H} \log P(s_{t+1}|s_t,u_t)) + (\sum_{t=0}^{H} \log \pi_\theta(u_t|s_t))] \\ &= \sum_{t=0}^{H} \underbrace{\nabla_\theta \log \pi_\theta(u_t|s_t)}_{\text{no dynamics required!}}\end{aligned} \tag{19}$$

Plugging the result of equation 19 into equation 16 we obtain the following equation for the gradient of the utility function w.r.t to parameter vector θ:

$$\nabla_\theta U(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\tau) R(\tau)] \tag{20}$$

We can compute an empirical approximation of that expresion by taking m sample trajectories (or paths) under the policy π_θ. This works even if the reward function R is unkown and/or discontinuous. This works in discrete state spaces. The likelihood ratio changes the probability of experienced paths. That is, the probability of sampling trajectories. Thus we use a Monte Carlo approach to approximate the gradient of the utility of π_θ: (REPHRASE) Which, if we plug into the original equation, we get:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=0}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)}) \tag{21}$$

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=0}^{m}\sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})(\sum_{k=0}^{H} R(s_t^{(i)},u_t^{(i)})) \tag{22}$$

Sutton et al. (1999) offers a different approach to this derivation by calculating the gradient for the state value function on an initial state s_0, calculating $\nabla_\theta V_{\pi_\theta}(s_0)$.

References

Abdallah, S., Org, S., Kaisers, M., and Nl, K. (2016). Addressing Environment Non-Stationarity by Repeating Q-learning Updates. *Journal of Machine Learning Research*, 17:1–31.

Andersen, P.-a. (2017). Deep RTS : A Game Environment for Deep Reinforcement Learning in Real-Time Strategy Games.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. (2017). Hindsight Experience Replay. (Nips).

Barto, A. G. (2003). Recent Advances in Hierarchical Reinforcement Learning Markov and Semi-Markov Decision Processes. *Most*, 13(5):1–28.

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2015). The arcade learning environment: An evaluation platform for general agents. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2015-Janua, pages 4148–4152.

Bellman, R. (1957). *Dynamic Programming*, volume 70.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control, Vol. II*, volume 2.

Borkar, V. S. (1988). A convex analytic approach to Markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. pages 1–4.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of {Monte Carlo Tree Search} methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

De Asis, K., Ca, K., Hernandez-Garcia, J. F., Ca, J., Holland, G. Z., and Sutton, R. S. (2017). Multi-step Reinforcement Learning: A Unifying Algorithm.

Degris, T., White, M., and Sutton, R. S. (2012). Off-Policy Actor-Critic.

Deisenroth, M. P. and Rasmussen, C. E. (2011). PILCO: A Model-Based and Data-Efficient Approach to Policy Search.

Greensmith, E., Bartlett, P., and Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *The Journal of Machine Learning ...*, 5:1471–1530.

Guzdial, M., Li, B., and Riedl, M. O. (2017). Game Engine Learning from Video. *International Conference on Artificial Intelligence (IJCAI)*.

Hansen, E., Bernstein, D., and Zilberstein, S. (2004). Dynamic Programming for Partially Observable Stochastic Games. *Nineteenth Conference on Artificial Conference (AAAI)*, pages 709–715.

Heinermann, A. (2009). BWAPI: Brood war api, an api for interacting with starcraft: Broodwar.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. (2017). Rainbow: Combining Improvements in Deep Reinforcement Learning.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement Learning with Unsupervised Auxiliary Tasks. pages 1–14.

Kaisers, M. and Tuyls, K. (2010). Frequency Adjusted Multi-agent Q-learning. *Learning*, pages 309–316.

Kempka, M., Wydmuch, M., Runc, G., Toczek, J., and Jaskowski, W. (2017). ViZDoom: A Doom-based AI research platform for visual reinforcement learning. In *IEEE Conference on Computatonal Intelligence and Games, CIG*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-Critic Algorithms. *Nips*, 42(4):1143–1166.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. (2017). A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. (Nips).

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning.

Lin, L.-j. (1993). Reinforcement Learning for Robots Using Neural Networks. *Report, CMU*, pages 1–155.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *Machine Learning Proceedings 1994*, pages 157–163.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. (2017). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous Methods for Deep Reinforcement Learning. 48.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. pages 1–9.

Oliehoek, F. A. and Amato, C. (2014). Best Response Bayesian Reinforcement Learning for Multiagent Systems with State Uncertainty. *AAMAS Workshop on Multiagent Sequential Decision Making Under Uncertainty, MSDM 2014*, (May).

Ontañón, S. (2013). The combinatorial multi-armed bandit problem and its application to real-time strategy games. *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 58–64.

Owen, G. and Owen, G. (1982). Game Theory. *Collection*.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-Driven Exploration by Self-Supervised Prediction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:488–489.

Piunovskiy, A. and Zhang, Y. (2012). The Transformation Method for Continuous-Time Markov Decision Processes. *Journal of Optimization Theory and Applications*, 154(2):691–712.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized Experience Replay. pages 1–21.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. pages 1–12.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic Policy Gradient Algorithms. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 387–395.

Soemers, D. (2014). Tactical Planning Using MCTS in the Game of StarCraft. page 12.

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*.

Sutton, R. S., Mcallester, D., Singh, S., and Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. *In Advances in Neural Information Processing Systems 12*, pages 1057–1063.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., and White, A. (2011). Horde : A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction Categories and Subject Descriptors. *In Practice*, 2(1972):761–768.

Synnaeve, G., Nardelli, N., Auvolat, A., Chintala, S., Lacroix, T., Lin, Z., Richoux, F., and Usunier, N. (2016). TorchCraft : a Library for Machine Learning Research on Real-Time Strategy Games arXiv : 1611 . 00625v2 [cs . LG] 3 Nov 2016. pages 1–6.

Tamar, A., Wu, Y., Thomas, G., Levine, S., and Abbeel, P. (2017). Value iteration networks. *IJCAI International Joint Conference on Artificial Intelligence*, (Nips):4949–4953.

Thrun, S. and Schwartz, A. (1993). Issues in Using Function Approximation for Reinforcement Learning. *Proceedings of the 4th Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*, pages 1–9.

Tijsma, A. D., Drugan, M. M., and Wiering, M. A. (2017). Comparing exploration strategies for Q-learning in random stochastic mazes. In *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*.

Unity (2017). Unity Machine Learning Agents.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., Quan, J., Gaffney, S., Petersen, S., Simonyan, K., Schaul, T., van Hasselt, H., Silver, D., Lillicrap, T., Calderone, K., Keet, P., Brunasso, A., Lawrence, D., Ekermo, A., Repp, J., and Tsing, R. (2017). StarCraft II: A New Challenge for Reinforcement Learning. *arXiv preprint arXiv:1708.04782*.

Watkins, C. J. (1989). *Learning from delayed rewards*. PhD thesis, King's College.

Watkins, C. J. and Dayan, P. (1992). Technical Note: Q-Learning. *Machine Learning*, 8(3):279–292.

Wender, S. and Watson, I. (2012). Applying Reinforcement Learning to Small Scale Combat in the Real-Time Strategy Game StarCraft : Broodwar. pages 402–408.

Williams, R. J. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3):229–256.

Wu, Y., Mansimov, E., Liao, S., Grosse, R., and Ba, J. (2017). Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. pages 1–14.