



VICTORIA UNIVERSITY OF  
**WELLINGTON**  
TE HERENGA WAKA

**School of Engineering and Computer Science**  
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600  
Wellington  
New Zealand

Tel: +64 4 463 5341  
Internet: [office@ecs.vuw.ac.nz](mailto:office@ecs.vuw.ac.nz)

**Reinforcement Learning**  
**Part 1: Introduction and Modern Algorithms**

James Thompson

Supervisor: Dr Aaron Chen

April 3, 2025

Submitted in partial fulfilment of the requirements for  
Master of Artificial Intelligence.

**Abstract**

This is submitted for the completion of AIML440 Directed individual study as part of the Master of Artificial Intelligence programme at Victoria University of Wellington. This report will have three sections. Firstly an Introduction of reinforcement learning and basic algorithms. Secondly I will evaluate modern algorithms. Lastly I will describe a new algorithm and compare it to the state of the art models. This version contains just the first part which is an introduction and modern algorithms explanations.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.0.1	The reinforcement learning problem . . . . .	1
1.0.2	Formalism . . . . .	2
<b>2</b>	<b>Algorithms</b>	<b>5</b>
2.1	Traditional algorithms . . . . .	5
2.1.1	Value Based Methods . . . . .	5
2.1.2	Policy Based Methods . . . . .	6
2.1.3	Challenges of traditional methods . . . . .	7
2.2	Modern Deep learning algorithms . . . . .	7
2.2.1	Deep Q Networks . . . . .	7
2.2.2	Soft Actor-Critic . . . . .	9
2.2.3	Proximal Polixy Optimizatin . . . . .	10
2.2.4	Twin Delayed Deep Deterministic Policy Gradient . . . . .	12
2.3	State of the Art Algorithms . . . . .	13
2.3.1	Truncated Quantile Critics . . . . .	14
2.3.2	Randomized Ensemble Double Q-Learning . . . . .	15
2.3.3	Dropout Q-functions . . . . .	16
2.3.4	Cross Q-learning . . . . .	17
<b>3</b>	<b>Conclusions</b>	<b>19</b>



# Chapter 1

## Introduction

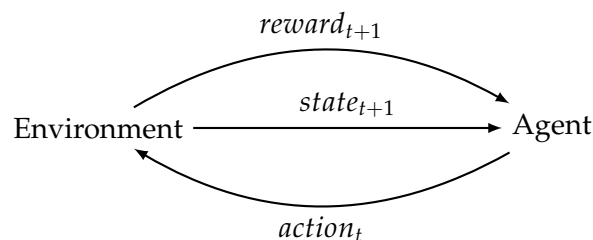
### 1.0.1 The reinforcement learning problem

Reinforcement learning is a framework that describes how an agent learns by interacting with the environment. The framework involves 2 entities the agent and environment. The agent is the only entity that we as designers have direct control of. We decide how its learns and decides on its actions. The environment is the context and situation that the agent is in. There are three important flows of information; the state of the environment to the agent, the action decision to the environment and lastly the reward to the agent. This can be best be understood in the figure below.

---

**Figure 1.1** The flow of information between th environment and agent

---



---

The state is a representation of the information within the environment that the agent observes to make its decision on which action to take. The reward signal is treated as the final word on how good the previous state action was, the more reward the better, always. Lastly the action taken by the agent is sent to the environment and the environment will in turn return the next state and reward. This brings us to the reinforcement learning problem which can be framed as such "How does the agent decide which actions to take given the state such that it maximises the future cumulative reward". It is important that the agent maximises all *future cumulative* reward otherwise short term gains could be made to the sacrifice of larger long term gains. This idea has been formalised by Richard Sutton as the reward hypothesis

"That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)." [25]

### 1.0.2 Formalism

The reinforcement problem can be formalised as a Markov Decision Process (MDP). The MDP is a collection of states, actions and rewards along with a transition function which states the probability of the next reward and state given a state and reward. This makes the MDP a 4-tuple  $(S, \mathcal{A}, \mathcal{P}, R)$  where  $S$  is the set of states,  $\mathcal{A}(s)$  is the set of actions that can be taken from state  $s$ ,  $\mathcal{P}$  is the transition function and  $R \subset \mathbb{R}$  is the set of rewards. The transition function is defined as:

$$\mathcal{P}(s', r, s, a) = \Pr \{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\} \quad (1.1)$$

Where  $s, s' \in S, a \in \mathcal{A}(s)$  and  $r \in R$ .

This transition function completely characterises the dynamics of the environment. The abstraction of the environment to a MDP is widely applicable and serves as the basis for much of reinforcement learning. We can see here that transition function only looks at the previous state. It can do this because we assume that the state representation has the *Markov property* [25]. The Markov property states that including previous states in the conditional won't change the probability of next state reward tuple. In other words  $\mathcal{P}(s_{t+1}, r_{t+1}, s_t, a_t) = \mathcal{P}(s_{t+1}, r_{t+1}, s_t, a_t, s_{t-1}, s_{t-2}, \dots, s_0)$ . Therefore a state with the Markov property is a sufficient representation of the history of the agent-environment interaction. In some problems the agent only partially observes the state meaning the Markov property is not satisfied. This is called a partially observable MDP (POMDP) and is a more complex problem to solve which won't be addressed in this report.

The agent interacts with the MDP to produce a trajectory of states, actions and rewards. Because the agent will not know the transition function 1.1 it will have to learn about the MDP from the information in a trajectory 1.2.

$$S_1, A_1, R_2, \dots, S_n, A_n, R_{n+1}, \dots \quad (1.2)$$

In this trajectory 1.2 we have an infinite sequence as this is an example from a continuing MDP that has no end. Alternatively you could have episodic MDPs that have a start and end with a terminal state. Episodic MDPs require practical considerations in implementing algorithms. However for most of the theory it makes no difference as you can think of an episode MDP as a continuing MDP with a state that transitions to itself and gives reward 0.

"The cumulative sum of a received scalar signal" part of the reward hypothesis 1.0.1 can be formalised to be the return  $G$ .

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots = \sum_{k=t}^{\infty} \gamma^{k-t} R_{k+1} \quad (1.3)$$

$\gamma$  is called the discounting factor and weights future rewards to have less effect on the return. It is usually added for two reasons. Firstly because it makes the return finite which simplifies the mathematics. Secondly is due to the very natural intuition that the future is less predictable than the present, thus more distant rewards should have less weight as they are less certain.

A agent will use policy  $\pi$  which provides the probability of taking action  $a$  given state  $s$ . This policy could be deterministic or stochastic. To evaluate how good a policy is or how much reward we can expect at a state we need a function to tell us. This is called the value function and it is an expectation of the future rewards.

$$v_{\pi}(s) = \mathbb{E} [G_t | S_t = s] \quad (1.4)$$

The state value function depends on the state, policy and discounting factor  $\gamma$  which is used by the return. In a similar vain to the state value function we have the action-value function which is the expected future reward given a state and action taken.

$$q_{\pi}(s, a) = \mathbb{E} [G_t | S_t = s, A_t = a] \quad (1.5)$$

We can understand how to compute the expectation of the value function by looking at the Bellman equation [25]. The Bellman equation is based off the notion that the value of a situation should be the immediate reward you get plus the value of the situation you end up in. This can be written for the state value function as follows:

$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r + \gamma v_{\pi}(s')] \quad (1.6)$$

$\pi(a|s)$  is the probability of taking action  $a$  given state  $s$  and  $p(s', r|s, a)$  is the transition function. This equation can be solved for  $v_{\pi}$  by iterating over all states and actions until the value function converges. This is called the iterative policy evaluation and is a dynamic programming technique [3].

The maximising part of the reinforcement problem can be solved by the finding optimal actions. We can define both  $v_{\star} = \max_{\pi} v_{\pi}(s)$  and  $q_{\star} = \max_{\pi} q_{\pi}(s, a)$  as the optimal value given optimal actions afterwards. If we can find  $q_{\star}$  then the problem is solved as we could make a policy  $\pi_{\star}$  that is greedy with respect to  $q_{\star}$ . Two things stop this from being so simple. Firstly is that in many real problems we don't know the true transition function, therefore we either have to learn/approximate or learn the value/policies in a way that doesn't involve the transition function. Secondly is the computational complexity of iterating over all states and actions as well as all possible policies. This is called the curse of dimensionality and is a major problem in reinforcement learning. If we try to learn and transition function and come up with a model of the environment then we are using model based methods, alternatively we ignore the model and learn the value or policy directly and these are called model free methods. In the rest of the report we will be focusing on model free methods.

The form of explicitly learning a value function then implicitly getting actions from it is called value based methods. Alternatively you can learn the policy directly with policy based methods. The effectiveness of these methods depend heavily on how easy the value function or policy is to learn.





## Chapter 2

# Algorithms

In the real world we don't get access to the transition function of the MDP. The agents only method to learn is through interacting with the environment. From this interaction we get a sample of the environment and must learn from it. Therefore stochastic learning methods must be used. Two learning ideas are core for many reinforcement learning algorithms. These are Monte Carlo [25] and Temporal Difference learning [23] [24].

### 2.1 Traditional algorithms

#### 2.1.1 Value Based Methods

As mentioned earlier value based method focus on learning a value function and from that function we can derive a policy. The action value function is more powerful as simply knowing the value of a state is not enough to make a decision. The action value function is defined as the expected return given a state and action. Thus all you need to do is take the action that gives you the maximum  $Q$  value.

Monte Carlo learning works by estimating the value function as being the average return from that state. A basic algorithm would look like this.

---

**Algorithm 1** Monte Carlo Control with Exploring Starts

---

```
1: Arbitrarily initialize  $\pi(s) \in \mathcal{A}(s)$  and  $Q(s, a) \in \mathbb{R}$ 
2: Returns( $S, A$ )  $\leftarrow$  empty list
3: for each episode do
4:   Arbitrarily choose  $S_0 \in \mathcal{S}, A_0 \in \mathcal{A}$ 
5:   Generate an episode by following  $\pi$ 
6:    $G \leftarrow 0$ 
7:   for each step of the episode  $t = T - 1, T - 2, \dots, 0$  do
8:      $G \leftarrow \gamma G + R_{t+1}$ 
9:     if  $(S_t, A_t)$  does not appear earlier in the episode then
10:      Append  $G$  to Returns( $S_t, A_t$ )
11:       $Q(S_t, A_t) \leftarrow \text{average}(\text{Returns}(S_t, A_t))$ 
12:       $\pi(S_t) \leftarrow \arg \max_a Q(S_t, a)$ 
13:     end if
14:   end for
15: end for
```

---

Because we need a well defined return Monte Carlo methods only work on episodic tasks. It will only update the states that it actually visits. This gives us the option of just

exploring the state space area we are interested in learning about.

Temporal difference learning works different incrementally updating the value of a state with the received reward and the value of the next state 2.1. As the update is using its own estimate it is *bootstrapping*. Because of bootstrapping it can work with both episodic and continuing environments.

$$V(S_t) = R_{t+1} + \gamma V(S_{t+1}) \quad (2.1)$$

Below is a basic implementation of TD learning for a continuing environment, however it is easy to change it to episodic by changing the loop to be over episodes.

---

**Algorithm 2** TD(0)

---

- 1: Initialize  $Q(s, a)$  arbitrarily and set  $Q(\text{terminal-state}) = 0$
  - 2: Initialize  $S$
  - 3: **while** not converged **do**
  - 4:   Take action  $A$ , observe  $R, S'$
  - 5:   Choose  $A' \in \mathcal{A}(S')$  using policy derived from  $Q$
  - 6:    $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$
  - 7:    $S \leftarrow S'; A \leftarrow A'$
  - 8: **end while**
- 

This particular update is known as TD(0) because it only looks one step into the future. However it can be generalised to be TD(n). One can see that if we had TD( $\infty$ ) we would be back at Monte Carlo if working in episodic tasks.

The two methods of TD(n) and Monte carlo can be to get the best of both worlds using TD( $\lambda$ ) methods [24].

### 2.1.2 Policy Based Methods

A different way of the approaching the problem is by learning the policy directly without a learning a concept of how good a state is. The most common policy based method is called policy gradient methods [26]. To use policy gradient methods you need to parametrize the policy  $\pi$  in any way as long as  $\pi(a|s, \theta)$  is differentiable with respect to  $\theta$ .

The first of these policy gradient methods was the Reinforce algorithm [30]. It is a Monte Carlo policy gradient method. The updates are made by moving the policy in the direction of the gradient of the log probability of the action taken. This works because it is proportional to the return which is given to use by the policy gradient theorem [26].

---

**Algorithm 3** REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for  $\pi_*$ 


---

**Require:** A differentiable policy parameterization  $\pi(a|s, \theta)$

**Require:** Step size  $\alpha > 0$

- 1: Initialize policy parameter  $\theta \in \mathbb{R}^d$  (e.g., to 0)
  - 2: **loop**
  - 3:   Generate an episode  $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$ , following  $\pi(\cdot|\cdot, \theta)$
  - 4:   **for**  $t = 0, 1, \dots, T-1$  **do**
  - 5:      $G \leftarrow \sum_{k=t+1}^T \gamma^{k-t-1} R_k$
  - 6:      $\theta \leftarrow \theta + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \theta)$
  - 7:   **end for**
  - 8: **end loop**
-

### 2.1.3 Challenges of traditional methods

There are main two issues with the simple methods of TD and Monte Carlo mentioned above. This is scalability and exploration.

Firstly with scalability we are storing the action value of a particular state in a large lookup table. This will immediately become a problem when dealing with large state or action spaces. Most real world application have tremendously large state/action spaces. Furthermore it is easy to imagine how most of the states have overlapping information and are actually really quite similar. Therefore instead of learning  $Q$  directly we can try and learn an approximator  $\hat{Q}$ .

Second problem is exploration. As we learn we are finding better and better actions. These better action are taken which will form our trajectory. What can happen here is that we miss large chunks of the state space. To solve this we can instead use a sub optimal policy that deliberately takes exploratory actions.  $\epsilon$ -greedy is a good example of this as it will take a random action with probability  $\epsilon$  and an optimal action the rest of the time. Alternatively you can use a different exploration policy to interact with the environment while you are learning the optimal. This method of learning from experience using a different policy is called off-policy learning. The previous methods we looked at are all on-policy.

## 2.2 Modern Deep learning algorithms

One of the problems with the above tabular methods is that the policy and value functions are stored explicitly in effectively a large table lookup. This means that for every state and or action there needs to be a new entry. This is problematic as the state and action space of the problems get very large. To resolve this problem we can use function approximators and the best ones we have are neural networks [10]. This allows for the combination of the deep learning ideas within the reinforcement learning framework to give rise to deep reinforcement learning.

### 2.2.1 Deep Q Networks

One can take the TD(0) learning algorithm and apply the notion of off-policy learning we get SARSA MAX otherwise known as Q-Learning [28] [29]. It works almost the same except the action value function is updated using the best possible action taken at the next state.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left( R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right) \quad (2.2)$$

As with TD(0) this fails when faced with a sufficiently large state and/or action space. Therefore we apply a neural network as an approximator  $\hat{Q}$  for  $Q$ . The algorithm we get is called Deep-Q-Learning [17][18]. It was applied very successfully to match and exceed human level performance on a large variety of Atari 2600 games (Space invaders, pong etc).

Here is the algorithm that Mnih et al used in their papers [17] [18], with some notational differences to make it more explicit.

---

**Algorithm 4** Deep Q-Learning (DQN)

---

**Require:** Number of episodes  $M$ , replay memory capacity  $N$ , minibatch size  $K$ , discount factor  $\gamma$ , learning rate  $\alpha$ , update frequency  $C$ , exploration probability  $\epsilon$ . A neural network function approximator  $Q$

- 1: Initialize replay memory  $\mathcal{D} \leftarrow \emptyset$  with capacity  $N$
- 2: Initialize action-value function  $Q_\theta$  with random weights  $\theta$
- 3: Initialize target action-value function  $\hat{Q}_{\theta^-}$  with weights  $\theta^- \leftarrow 0$
- 4: **for** episode in range(1,  $M$ ) **do**
- 5:   Initialize sequence  $s_1 \leftarrow \{x_1\}$  and preprocessed sequence  $\phi_1 \leftarrow \phi(s_1)$
- 6:   **for** each step  $t$  in episode **do**
- 7:     With probability  $\epsilon$ , select a random action  $a_t$
- 8:     Otherwise, select  $a_t \leftarrow \arg \max_a Q(\phi(s_t), a; \theta)$
- 9:     Observe reward  $r_t$  and next state  $x_{t+1}$
- 10:    Set  $s_{t+1} \leftarrow (s_t, a_t, x_{t+1})$  and preprocess  $\phi_{t+1} \leftarrow \phi(s_{t+1})$
- 11:    Store transition  $(\phi_t, a_t, r_t, \phi_{t+1})$  in  $\mathcal{D}$
- 12:    Sample random minibatch  $B$  of size  $K$  of transitions from  $\mathcal{D}$
- 13:    **for** each transition  $(\phi_j, a_j, r_j, \phi_{j+1})$  in  $B$  **do**

$$y_j \leftarrow \begin{cases} r_j, & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-), & \text{otherwise} \end{cases}$$

- 14:     Compute loss  $L_j \leftarrow (y_j - Q(\phi_j, a_j; \theta))^2$
  - 15:   **end for**
  - 16:   Perform gradient descent step on the average loss:  $\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{|B|} \sum_j L_j$  ▷
  - Alternative gradient steps could be used here (e.g Adam was used in the paper)
  - 17:   **if**  $t \bmod C == 0$  **then**
  - 18:     Update target network:  $\theta^- \leftarrow \theta$
  - 19:   **end if**
  - 20: **end for**
  - 21: **end for**
- 

There are more features of this algorithm that are different than traditional Q-learning. These are important to solve problems that arise when you add a neural network approximator into the learning process of Q-learning.

The first of these ideas is called experience replay [15], which is collecting your experience in a replay memory  $D$ . Then to learn you can loop through your experience and extract as much as you can from your previous experiences. The idea in Deep Q-learning is that at each time step rather than updating  $\hat{Q}$  with your current experience you update it using a sampled transition from  $D$ . This helps with an important problem which is that stochastic gradient descent assumes independent and identically distributed data (\*\*i.i.d\*\*). As experience is collected each transition will be dependent on the previous transition and will affect the next transition thus the independent assumption is not held. Thus randomly sampling from  $D$  will give you an independent data set. This also makes DQN an off policy learning method

The second assumption is that the training data is identically distributed. Gradient descent is working towards the target  $\gamma$ . The problem is that the target involves our current prediction, so as we learn a better prediction our target will move. This results in a chase which decreases learning efficiency. The solution to this as seen in the algorithm is fixing  $\hat{Q}$  for a fixed number of steps  $C$ . It uses  $\hat{Q}$  in the target  $\gamma_j$  but then updates  $Q$  every  $C$  updates.

This gives the network a reasonable opportunity to tend towards the target  $\gamma_j$  while still using a recent estimate of the action value.

Lastly is the concept of pre-processing the state to speed up the learning process. One can decrease the computational complexity of the neural network by simplifying the state using a transformation that in theory doesn't lose the meaningful information. Minh et al used this in their 2013 paper to make the video input; smaller, conform to the ratio their model needed and making it gray scale. We can see that this pre-processing step is the only example of domain specific knowledge that would be needed for an implementation of the algorithm to a task.

## 2.2.2 Soft Actor-Critic

The DQN method discussed above, along with TD and MC learning, are all value-based methods. That is, they learn to evaluate how good a current situation is (whether that be a state or state-action), and then a policy can be generated by maximizing over the value function. Another approach is to directly learn the policy function. A combination of these ideas leads to the actor-critic framework.

In an actor-critic framework, there is a policy that controls how the agent acts,  $\pi$ , as well as a value function that evaluates how good the action taken was. A good way to update the policy is based on whether the result was better or worse than what the critic expected. Using the TD error of the critic for this is called A2C [16].

However, there are still challenges in expanding A2C to high-dimensional continuous control tasks. Many current methods are difficult to stabilize and require carefully tuned hyperparameters. To address this, several improvements can be introduced to the actor-critic framework, resulting in Soft Actor-Critic (SAC) [7].

First, similar to DQN [17, 18], making it off-policy allows for much better sampling efficiency as more information is extracted from experience. More importantly, SAC employs entropy maximization. Traditional RL agents aim to maximize the expected sum of rewards. However, the maximum entropy RL framework modifies this objective to maximize both the expected reward and the entropy of the policy, leading to the following objective function:

$$J(\pi) = \sum_{t=0}^T \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | s_t))] \quad (2.3)$$

This entropy term encourages exploration, with the temperature parameter  $\alpha$  determining the strength of this encouragement. Initially a hyperparameter in [7], the SAC algorithm was later modified to learn and adjust  $\alpha$  throughout training [8]. This not only improves efficiency—since choosing an optimal temperature is task-dependent and non-trivial [8]—but also allows the policy to explore uncertain regions while being more exploitative in familiar areas.

The final SAC algorithm [8] modified for clarity is shown below:

---

**Algorithm 5** Soft Actor-Critic Algorithm

---

**Require:** Replay buffer size  $N$ , Minibatch size  $K$ , discount factor  $\gamma$ , learning rates  $\lambda$ ,  $\lambda_\pi$  and  $\lambda_Q$ , update frequency  $C$ , target smoothing coefficient  $\tau$ , Updates to Data ratio UTD, environemt steps  $S$ , neural network function approximator  $Q$ , neural network function approximator  $\pi$  and initial entropy coefficient  $\alpha$ .

- 1: Initialize replay memory  $\mathcal{D} \leftarrow \emptyset$  with capacity  $N$
  - 2: Initialize policy  $\pi_\phi$  with random weights  $\phi$
  - 3: Initialize both action-value function  $Q_{\theta_i}$  with random weights  $\theta_i$
  - 4: Initialize  $\bar{\theta}_i \leftarrow \theta_i$  for  $i \in \{1, 2\}$
  - 5: **repeat**
  - 6:   **for**  $S$  steps **do**
  - 7:     select action:  $a_t \sim \pi_\phi(a_t|s_t)$
  - 8:     Observe reward  $r_t$  and next state  $s_{t+1}$
  - 9:     Store transition  $(s_t, a_t, r_t, s_{t+1})$  in  $\mathcal{D}$
  - 10:   **end for**
  - 11:   **for** UTD times **do**
  - 12:     Update Q-functions:  $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$  for  $i \in \{1, 2\}$
  - 13:     Update policy:  $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
  - 14:     Update entropy coefficient:  $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$
  - 15:     Update target Q-networks:  $\bar{\theta}_i \leftarrow \tau \theta_i + (1 - \tau) \bar{\theta}_i$  for  $i \in \{1, 2\}$
  - 16:   **end for**
  - 17: **until** stopping criteria is met
- 

One noteworthy difference from conventional actor-critic methods is that SAC uses two Q-functions. While not strictly necessary, this significantly speeds up training [8].

Between 2018 and 2019, when SAC was first introduced, it outperformed other popular methods such as TD3 [6], DDPG [14], and PPO [21] in continuous control benchmarks.

### 2.2.3 Proximal Polixy Optimizatin

One of the instabilities in RL algorithms is that a change in the actions taken by a policy can have large consequences, both good and bad. A natural idea is to try and limit the changes that can happen to a policy at any step. You can do this by using a surrogate objective function that is limited to be at least somewhat similar to the current policy. This was first implemented using a trust region (i.e an area that we can safely move the policy and is guaranteed to improve) and resulted in the TRPO algorithm [19]. Yet the most popular method is that of PPO [21], as it is computationally simpler. The method used in PPO is to use a clipped objective which results in a loss function like so:

$$L^{\text{CLIP}}(s, a, \theta_k, \theta) = \min \left( \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), \text{clip} \left( \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(s, a) \right) \quad (2.4)$$

Where the advantage function ( $A^{\pi_{\theta_k}}(s, a)$ ) is a way of measuring how much better the situation is than expected. The ratio  $\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}$  is used to measure the difference between the current policy and the old policy. This ratio is what is clipped and used as a scalar for the advantage function. As we see in the situation that the action was better ( $A > 0$ ) the clip function will apply and make sure that  $L$  wont be too large, however it can be as small as 0. In the other case when things were worse than expected ( $A < 0$ ) it capped to make  $L$  not close to zero (note that  $L$  will be negative), yet  $L$  could be as large (in the negative sense) as

it would like. The reasoning is that allowing for these large and negative  $L$  means that we can do a better job of making sure that this action is not taken again.

The advantage function can be estimated in many ways . A simple advantage function could be actual return - estimated return ( $G_t - V(s_t)$ ) for episodic tasks, or TD error ( $r_t + \gamma V(s_{t+1}) - V(s_t)$ ) for continuing tasks. The implementation in the paper use generalised advantage estimation which is an exponentially weighted sum of TD errors [20], as it better balances bias and variance. GAE is defined as:

$$\hat{A}_t^{GAE(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}^V \quad (2.5)$$

Where  $\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$ .  $\gamma$  is the usual discounting factor but this is expanded on with  $\lambda$  which is from 0-1, at 0 it is TD(0) error and at 1 would be the full complete return.

With this advantage estimator we need to learn a value function. This puts PPO in the same style of actor-critic methods. It is however regarded as a Policy gradient method starting with the author Schulman designating it so [21].

The CLIP loss function can be augmented with other objectives to increase its effectiveness. For example using entropy maximisation to encourage exploring like what is done above with 2.2.2. This leads to this objective:

$$L^{\text{CLIP}+S}(s, a, \theta_k, \theta) = \hat{\mathbb{E}}_t \left[ L^{\text{CLIP}}(s, a, \theta_k, \theta) + cS[\pi_\theta](s) \right] \quad (2.6)$$

Where  $S$  is the entropy of the policy. There is a large space of exploring different surrogate objective functions which incorporate the CLIP function.

Below is the algorithm from the paper [21] with some modifications to make it clearer.:

---

**Algorithm 6** PPO Algorithm, Actor Critic style

---

**Require:** Clipping amount  $\epsilon$ , Minibatch size  $K$ , epoch  $M$ , discount factor  $\gamma$ , learning rates  $\alpha_\theta$  and  $\alpha_\phi$ , number of actors  $N$ , environment steps  $T$ , neural network function approximator  $V$ , neural network function approximator  $\pi$ , value function coefficient  $c_1$  and entropy coefficient  $c_2$

- 1: Initialize policy  $\pi_\theta$  and value function  $V_\phi$  with random weights  $\theta$  and  $\phi$
- 2:  $\theta_{old} \leftarrow \theta$
- 3: **repeat**
- 4:   **for** actor = 1, 2, ..., N **in parallel do**
- 5:     Run policy  $\pi_{\theta_{old}}$  in environment for  $T$  timesteps
- 6:     Collect trajectory  $(s_t, a_t, r_t, s_{t+1})_{t=0}^{T-1}$
- 7:     Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$  using GAE
- 8:     Compute returns  $\hat{R}_t \leftarrow \sum_{i=t}^{T-1} \gamma^{i-t} r_i$
- 9:   **end for**
- 10:   Combine data from all actors:  $\mathcal{D} = \{(s_t^n, a_t^n, \hat{A}_t^n, \hat{R}_t^n)\}$  for  $n = 1, \dots, N$  and  $t = 1, \dots, T$
- 11:   **for** epoch = 1, 2, ..., M **do**
- 12:     Shuffle data  $\mathcal{D}$  and split into minibatches of size  $K \leq NT$
- 13:     **for** each minibatch  $\mathcal{B} \subset \mathcal{D}$  **do**
- 14:       For each  $(s, a, \hat{A}, \hat{R}) \in \mathcal{B}$ :
- 15:        Compute ratio  $r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}$
- 16:        Compute clipped surrogate objective:
- 17:         $L_{clip}(\theta) = \min(r(\theta)\hat{A}, \text{clip}(r(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A})$
- 18:        Compute value function loss:
- 19:         $L_{VF}(\phi) = (V_\phi(s) - \hat{R})^2$
- 20:        Compute entropy bonus:
- 21:         $S[\pi_\theta](s) = -\sum_a \pi_\theta(a|s) \log \pi_\theta(a|s)$
- 22:        Total objective:  $L(\theta, \phi) = \mathbb{E}[L_{clip}(\theta) - c_1 L_{VF}(\phi) + c_2 S[\pi_\theta](s)]$
- 23:        Update  $\theta$  using SGD or Adam optimizer:  $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta L(\theta, \phi)$
- 24:        Update  $\phi$  using SGD or Adam optimizer:  $\phi \leftarrow \phi + \alpha_\phi \nabla_\phi L(\theta, \phi)$
- 25:     **end for**
- 26:   **end for**
- 27:    $\theta_{old} \leftarrow \theta$
- 28: **until** stopping criteria is met

---

## 2.2.4 Twin Delayed Deep Deterministic Policy Gradient

A problem with using value function based methods with approximation is that they can lead to overestimation bias (an agent thinking a situation is better than it really is). This has been shown to be a problem in the continuous control setting [6]. A way to resolve this overestimation error is to use two critics to estimate the value of the action taken and then use the minimum of the two to update the policy. This is called Twin Delayed Deep Deterministic Policy Gradient [6] and is a variant of DDPG [13] that builds on Double Q-Learning [27]. It is an actor critic method that also uses experience replay, target networks and policy noise to stabilise learning.

The policy noise method is not seen in the other algorithms mentioned above. The clipped policy noise in the critic target is used to make sure that the critics aren't overfitting to peaks. This works on the intuition that similar actions should have similar values.

The algorithm from the original paper [6] with slight modifications for clarity:



---

**Algorithm 7** Twin Delayed Deep Deterministic Policy Gradient (TD3)

---

**Require:** Replay buffer size  $N$ , Minibatch size  $K$ , discount factor  $\gamma$ , learning rates  $\alpha_\pi$  and  $\alpha_Q$ , target smoothing coefficient  $\tau$ , policy noise  $\sigma$ , target policy noise clip  $c$ , target policy noise  $\tilde{\sigma}$ , update frequency  $d$  and neural network function approximator  $Q$  and  $\pi$

- 1: Initialize critic networks  $Q_{\theta_1}, Q_{\theta_2}$ , and actor network  $\pi_\phi$  with random parameters  $\theta_1, \theta_2, \phi$
- 2: Initialize target networks  $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2, \phi' \leftarrow \phi$
- 3: Initialize replay buffer  $\mathcal{D}$
- 4: **repeat**
- 5:   increment timestep  $t$
- 6:   Select action with exploration noise  $a \sim \pi_\phi(s) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$  and observe reward  $r$  and new state  $s'$
- 7:   Store transition tuple  $(s, a, r, s')$  in  $\mathcal{D}$
- 8:   Sample mini-batch of  $K$  transitions  $(s, a, r, s')$  from  $\mathcal{D}$
- 9:   Initialize critic loss  $L_{\theta_1} \leftarrow 0, L_{\theta_2} \leftarrow 0$
- 10:   **for**  $i = 1$  to  $K$  **do**
- 11:     Add noise to target action:  $\tilde{a}_i \leftarrow \pi_{\phi'}(s'_i) + \text{clip}(\epsilon_i, -c, c)$  where  $\epsilon_i \sim \mathcal{N}(0, \tilde{\sigma})$
- 12:     Compute target Q-value:  $y_i \leftarrow r_i + \gamma \min(Q_{\theta'_1}(s'_i, \tilde{a}_i), Q_{\theta'_2}(s'_i, \tilde{a}_i))$
- 13:     Accumulate critic losses:  $L_{\theta_j} \leftarrow L_{\theta_j} + (y_i - Q_{\theta_j}(s_i, a_i))^2$  for  $j = 1, 2$
- 14:   **end for**
- 15:   Update critics:  $\theta_i \leftarrow \theta_i - \alpha_Q \nabla_{\theta_i}(L_{\theta_i}/K)$  for  $i = 1, 2$
- 16:   **if**  $t \bmod d$  **then**
- 17:     Update  $\phi$  by the deterministic policy gradient:
- 18:      $\nabla_\phi J(\phi) = K^{-1} \sum_i^K \nabla_a Q_{\theta_1}(s_i, a)|_{a=\pi_\phi(s_i)} \nabla_\phi \pi_\phi(s_i)$
- 19:     Update target networks:
- 20:      $\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i$
- 21:      $\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$
- 22:   **end if**
- 23: **until** stopping criteria is met

---

## 2.3 State of the Art Algorithms

Since the introduction of the modern deep reinforcement learning algorithms, there have been many new algorithms that have been developed. These generally build on the ideas from the first generation of deep reinforcement learning algorithms. I will go into detail about 4 algorithms which are actor critic methods and build on the ideas of SAC 2.2.2. These are:

Truncated Quantile Critics (TQC) [12] which uses an ensemble of critics and a truncated distributional critic with the goal of reducing the overestimation bias.

Randomized Ensemble Double Q-Learning (REDQ) [5] which uses an ensemble of critics which are trained towards a target made up of the minimum of a random subset of the critics. This allows for a more stable learning process that can handle Updates to Data ratio (UTD) which are greater than 1.

Dropout Q-functions (DROQ) [9] which is a variant of REDQ that uses smaller number of critics and instead adds layer dropout and layer normalization to handle a UTD  $\gg 1$ .

Cross Q-learning (CrossQ) [4] which builds upon SAC by removing the target networks by having the critic network take both the current and next time step as input and applying batch normalization.

### 2.3.1 Truncated Quantile Critics

One of the challenges in learning an accurate critic function is the overestimation bias. This comes from the fact that the critic is updated towards the maximum of the Q-values. TQC mitigates this by using multiple distributional critics and truncating the top quantile of the distributions.

The distributional reinforcement learning framework [2] is different from traditional Q-learning because rather than learning the expected value for the state action pair, it learns the return.

$$Z_{\psi_n}(s, a) := \frac{1}{M} \sum_{m=1}^M \delta(\theta_{\psi_n}^m(s, a))$$

It approximates the distribution with  $M$  number of quantiles. each predicting the value of the return at a given quantile  $\theta_{\psi}^m$ .

Rather than just having a single critic, TQC uses  $N$  critics. It takes all of the quantiles from the critics and mixes them together to form a single distribution. From there it truncates the top  $k$  values. Kuznetsov et al [12] posits that the order of mixing then truncating is important. These quantiles are then used to create a target which the critics are trained towards. In practice target networks are used as well to stabilize the learning process.

Note that the policy is optimized using the untruncated distributions from the critics.

The algorithm from the paper [12] with slight modifications is shown below:

---

#### Algorithm 8 Truncated Quantile Critics (TQC)

---

**Require:** Replay buffer size  $N$ , Minibatch size  $K$ , discount factor  $\gamma$ , learning rates;  $\lambda_\alpha, \lambda_\pi, \lambda_Q$ , number of critics  $C$ , truncation amount  $k$ , target critic update rate  $\beta$ , update frequency  $d$ , Updates to Data ratio UTD, and neural network function approximator  $Z$  and  $\pi$

- 1: Initialize replay buffer  $\mathcal{D}$
  - 2: Initialize policy  $\pi_\phi$  and  $C$  critics  $Z_{\psi_i}, \bar{Z}_{\bar{\psi}_i}$  with random parameters  $\psi_i$
  - 3: Set  $\alpha = 1$  and  $\mathcal{H}_T = -\dim \mathcal{A}$
  - 4: **repeat**
  - 5:   Run policy  $\pi_\phi$  in environment until termination or truncation
  - 6:   Append trajectories  $(s_t, a_t, r_t, s_{t+1})_{t=0}^{T-1}$  to  $\mathcal{D}$
  - 7:   **for** UTD times **do**
  - 8:     sample a batch of size  $K$  from the replay  $\mathcal{D}$
  - 9:      $\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha J(\alpha)$
  - 10:     $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
  - 11:     $\psi_n \leftarrow \psi_n - \lambda_Z \hat{\nabla}_{\psi_n} J_Z(\psi_n)$  for  $n \in [1..C]$
  - 12:     $\bar{\psi}_n \leftarrow \beta \psi_n + (1 - \beta) \bar{\psi}_n$  for  $n \in [1..C]$
  - 13:   **end for**
  - 14: **until** stopping criteria is met
-

### 2.3.2 Randomized Ensemble Double Q-Learning

Chen et al [5] made an algorithm to increase its updates to data ratio (UTD) to be greater than 1. Naively increasing the UTD to be greater than 1 in SAC 2.2.2 has poor results, which comes from the instability in the learning. To stabilize the learning REDQ using an ensemble of critics with a target made up of the minimum of a random subset of the critics.

The ensemble of  $C$  critics is trained towards a target which is the minimum of a random subset  $M$  of the critics. This target is in turn used to update all of the critics. The policy is optimized using the average of the critics. With an  $C = 10$  and  $M = 2$  the algorithm was able to handle a UTD of 20 [5]. The authors reference implementation of the algorithm is built of SAC 2.2.2 and interesting turns back into SAC if  $C = M = 2$  and  $UTD = 1$ .

At the time it was the first model-free algorithm to be able to handle a UTD greater than 1. This higher ratio allows for a increase in sample efficiency as it is extracting more information from the replay buffer per time step. This does however increase the computational cost as each time step requires  $C \times UTDC$  gradient descent steps and  $M$  forward passes.

Pseudo code for the algorithm taken from the paper [5] with some modifications for clarity is shown below:

---

#### Algorithm 9 Randomized Ensembled Double Q-learning (REDQ)

---

**Require:** Replay buffer size  $N$ , Minibatch size  $K$ , discount factor  $\gamma$ , learning rates  $z\lambda_\pi, \lambda_Q$ , number of critics  $C$ , target critic update rate  $\beta$ , entropy coefficient  $\alpha$  critic subsets  $M$ , Updates to Data ratio UTD, and neural network function approximator  $Q$  and  $\pi$

- 1: Initialize policy parameters  $\theta$ ,  $C$  Q-function parameters  $\phi_i, i = 1, \dots, C$ , empty replay buffer  $\mathcal{D}$ . Set target parameters  $\phi_{\text{targ},i} \leftarrow \phi_i$ , for  $i = 1, 2, \dots, C$
- 2: **repeat**
- 3:   Select action:  $a_t \sim \pi_\theta(a_t | s_t)$  and observe reward  $r_t$  and next state  $s_{t+1}$
- 4:   Store transition  $(s_t, a_t, r_t, s_{t+1})$  in  $\mathcal{D}$
- 5:   **for** UTD times **do**
- 6:     Sample a mini-batch  $B = \{(s, a, r, s')\}$  from  $\mathcal{D}$  of size  $K$
- 7:     Sample a set  $\mathcal{M}$  of  $M$  distinct indices from  $\{1, 2, \dots, C\}$
- 8:     Compute the Q target  $y$  (same for all of the  $C$  Q-functions):

$$y = r + \gamma \left( \min_{i \in \mathcal{M}} Q_{\phi_{\text{targ},i}}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}' | s') \right), \quad \tilde{a}' \sim \pi_\theta(\cdot | s')$$

- 9:     **for**  $i = 1, \dots, C$  **do**
- 10:       Update  $\phi_i$  with gradient descent using

$$\nabla_{\phi} \frac{1}{K} \sum_{(s,a,r,s') \in B} (Q_{\phi_i}(s, a) - y)^2$$

- 11:       Update target networks with  $\phi_{\text{targ},i} \leftarrow \beta \phi_{\text{targ},i} + (1 - \beta) \phi_i$
- 12:     **end for**
- 13: **end for**
- 14:   Update policy parameters  $\theta$  with gradient ascent using

$$\nabla_{\theta} \frac{1}{K} \sum_{s \in B} \left( \frac{1}{C} \sum_{i=1}^C Q_{\phi_i}(s, \tilde{a}_\theta(s)) - \alpha \log \pi_\theta(\tilde{a}_\theta(s) | s) \right), \quad \tilde{a}_\theta(s) \sim \pi_\theta(\cdot | s)$$

- 15: **until** stopping criteria is met
-

### 2.3.3 Dropout Q-functions

The big draw back of REDQ 2.3.2 is that it requires a large number of critics to be able to handle a UTD greater than 1. Hiraoka et al [9] proposed a method called Dropout Q-functions (DROQ) which uses a smaller number of critics and instead uses dropout and layer normalization to handle the UTD. This allows for a similar sample efficiency at about half the computational cost (DroQ is about 2x faster than REDQ) [9].

Dropout layers [22] are used within the critic networks to randomly drop out neurons during training. It is used within the deep learning community to help prevent overfitting. The intuition being that by randomly removing neurons the networks has to learn features that are not reliant on single neurons. However in this setting it is being used to model the uncertainty in the Q-value. Layer normalization [1] is used after the dropout layer which increases the effectiveness of the dropout layer.

Because it is using a smaller ensemble of critics it uses all of the critics when creating the target. This is different from REDQ 2.3.2 which uses a random subset of the critics. Other than this the algorithm is almost identical to REDQ as seen below:

---

#### Algorithm 10 Randomized Ensembled Double Q-learning (REDQ)

---

**Require:** Replay buffer size  $N$ , Minibatch size  $K$ , discount factor  $\gamma$ , learning rates  $z\lambda_\pi, \lambda_Q$ , number of critics  $C$ , target critic update rate  $\beta$ , entropy coefficient  $\alpha$ , Updates to Data ratio UTD, and neural network function approximator  $Q$  and  $\pi$

1: Initialize policy parameters  $\theta$ ,  $C$  Q-function parameters  $\phi_i, i = 1, \dots, N$ , empty replay buffer  $\mathcal{D}$ . Set target parameters  $\phi_{\text{targ},i} \leftarrow \phi_i$ , for  $i = 1, 2, \dots, N$

2: **repeat**

3:   Select action:  $a_t \sim \pi_\theta(a_t | s_t)$  and observe reward  $r_t$  and next state  $s_{t+1}$

4:   Store transition  $(s_t, a_t, r_t, s_{t+1})$  in  $\mathcal{D}$

5:   **for** UTD times **do**

6:     Sample a mini-batch  $B = \{(s, a, r, s')\}$  from  $\mathcal{D}$  of size  $K$

7:     Compute the Q target  $y$  (same for all of the  $C$  Q-functions):

$$y = r + \gamma \left( \min_{i \in C} Q_{Dr, \phi_{\text{targ},i}}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}' | s') \right), \quad \tilde{a}' \sim \pi_\theta(\cdot | s')$$

8:     **for**  $i = 1, \dots, C$  **do**

9:       Update  $\phi_i$  with gradient descent using

$$\nabla_{\phi} \frac{1}{K} \sum_{(s,a,r,s') \in B} (Q_{Dr, \phi_i}(s, a) - y)^2$$

10:       Update target networks with  $\phi_{\text{targ},i} \leftarrow \beta \phi_{\text{targ},i} + (1 - \beta) \phi_i$

11:     **end for**

12:   **end for**

13:   Update policy parameters  $\theta$  with gradient ascent using

$$\nabla_{\theta} \frac{1}{K} \sum_{s \in B} \left( \frac{1}{C} \sum_{i=1}^C Q_{Dr, \phi_i}(s, \tilde{a}_\theta(s)) - \alpha \log \pi_\theta(\tilde{a}_\theta(s) | s) \right), \quad \tilde{a}_\theta(s) \sim \pi_\theta(\cdot | s)$$

14: **until** stopping criteria is met

---

### 2.3.4 Cross Q-learning

The goal of both REDQ 2.3.2 and DROQ 2.3.3 is to increase the sample efficiency and they achieve it by using a UTD  $\geq 1$  (typically around 20). However they both come at a computational cost which increases with the number of updates to the data ratio. CrossQ [4] is simple improvement on SAC 2.2.2 which removes the target networks which allows it to use batch normalization [11]. This combination allows it to achieve higher sample efficient than REDQ 2.3.2 and DROQ 2.3.3 while being about 4 times faster. Furthermore much wider critic networks (2000) are used to as it further increases the sample efficiency however not as much as the combination of the previous two.

To allow the batch normalization to work without the target networks we combine the calculation current Q-value and the next Q-value. This is done by concatenating the current state and action with the next state and action. This means that the loss function is changed from what we see in SAC 2.2.2:

$$J_Q(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[ (Q_\theta(s, a) - r - \gamma Q_{\bar{\theta}}(s', \pi_\phi(s')))^2 \right] \quad (2.7)$$

To no longer use the target network and instead the output from a single Q-function:

$$J_Q(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[ (q_t - r - \gamma q_{t+1})^2 \right] \quad (2.8)$$

Where  $Q_\theta \left( \begin{bmatrix} S_t \\ S_{t+1} \end{bmatrix}, \begin{bmatrix} A_t \\ A_{t+1} \end{bmatrix} \right) = \begin{bmatrix} q_t \\ q_{t+1} \end{bmatrix}$ .

These changes are applied to SAC 2.2.2 in the paper however the changes could be applied to other off policy TD learning based algorithms. The algorithm is identical to the SAC algorithm 2.2.2 and so wont be shown again here. The differences are in the loss function and the batch normalization in the critic networks. Furthermore CrossQ uses a UTD of 1 with a policy delay of 3.



## Chapter 3

# Conclusions

Reinforcement Learning is a powerful framework for learning how to act. The recent combination of addition of deep learning has made reinforcement learning more powerful than ever. In this report I have covered the basics of reinforcement learning and the algorithms that are used to solve the problem. I have also evaluated some foundational algorithms and shown that they are capable of solving complex problems.

There are many ways of splitting the field of reinforcement learning. The binary splits we have looked at are model based and model free, value based and policy based and on policy and off policy. However the different solutions can merge the boundary between these splits. For example actor critic methods are explicitly both value based and policy based. Below is a table which helps summarise the differences between the modern algorithms.

Feature	DQN	SAC	PPO	TD3
Algorithm type	Value based	Actor-critic	Policy based (with actor critic style)	Actor-critic
Policy learning	Off policy	Off policy	On policy	Off policy
Learnt policy	Deterministic	Stochastic	Stochastic	Deterministic
Exploration strategy	$\epsilon$ -greedy	entropy maximisation	entropy maximisation	noise
Action space	Discrete	Continuous	Both	Continuous
Extra features	Experience replay, ensemble Q-functions	Experience replay, ensemble Q-functions	-	Experience replay

Table 3.1: Differences between modern deep reinforcement learning algorithms.

These models at their time of release were all considered state of the art but due to the fast moving field are now transitioning to more foundational models which the latest algorithms are building on. There is unknown potential in the applicability of current algorithms as well as unknown power in new algorithms. Some of the current limitations of the algorithms are the sample efficiency and the stability of the algorithms. Furthermore there is the larger problem of generalisation, which is getting an agent to act intelligently on varied environments that it has never seen before with minimal to no new training.

The field of reinforcement learning is still in its infancy and there is much to be discovered. The combination of deep learning and reinforcement learning has opened up many possibilities and the future is bright and exciting.





# Bibliography

- [1] BA, J. L., KIROS, J. R., AND HINTON, G. E. Layer Normalization, July 2016.
- [2] BELLEMARE, M. G., DABNEY, W., AND MUNOS, R. A Distributional Perspective on Reinforcement Learning, July 2017.
- [3] BELLMAN, R. *Dynamic Programming*. Princeton University Press, 1957.
- [4] BHATT, A., PALENICEK, D., BELOUSOV, B., ARGUS, M., AMIRANASHVILI, A., BROX, T., AND PETERS, J. CrossQ: Batch Normalization in Deep Reinforcement Learning for Greater Sample Efficiency and Simplicity, Mar. 2024.
- [5] CHEN, X., WANG, C., ZHOU, Z., AND ROSS, K. Randomized Ensembled Double Q-Learning: Learning Fast Without a Model, Mar. 2021.
- [6] FUJIMOTO, S., VAN HOOFF, H., AND MEGER, D. Addressing Function Approximation Error in Actor-Critic Methods, Oct. 2018.
- [7] HAARNOJA, T., ZHOU, A., ABBEEL, P., AND LEVINE, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, Aug. 2018.
- [8] HAARNOJA, T., ZHOU, A., HARTIKAINEN, K., TUCKER, G., HA, S., TAN, J., KUMAR, V., ZHU, H., GUPTA, A., ABBEEL, P., AND LEVINE, S. Soft Actor-Critic Algorithms and Applications, Jan. 2019.
- [9] HIRAOKA, T., IMAGAWA, T., HASHIMOTO, T., ONISHI, T., AND TSURUOKA, Y. Dropout Q-Functions for Doubly Efficient Reinforcement Learning, Mar. 2022.
- [10] HORNIK, K., STINCHCOMBE, M., AND WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 5 (Jan. 1989), 359–366.
- [11] IOFFE, S. Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. In *Advances in Neural Information Processing Systems* (2017), vol. 30, Curran Associates, Inc.
- [12] KUZNETSOV, A., SHVECHIKOV, P., GRISHIN, A., AND VETROV, D. Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics, May 2020.
- [13] LILLICRAP, T. P., HUNT, J. J., PRITZEL, A., HEESS, N., EREZ, T., TASSA, Y., SILVER, D., AND WIERSTRA, D. Continuous control with deep reinforcement learning, Feb. 2016.
- [14] LILLICRAP, T. P., HUNT, J. J., PRITZEL, A., HEESS, N., EREZ, T., TASSA, Y., SILVER, D., AND WIERSTRA, D. Continuous control with deep reinforcement learning, July 2019.

- [15] LIN, L.-J. *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, Carnegie Mellon University, USA, 1992.
- [16] MNIH, V., BADIA, A. P., MIRZA, M., GRAVES, A., LILICRAP, T. P., HARLEY, T., SILVER, D., AND KAVUKCUOGLU, K. Asynchronous Methods for Deep Reinforcement Learning, June 2016.
- [17] MNIH, V., KAVUKCUOGLU, K., SILVER, D., GRAVES, A., ANTONOGLOU, I., WIERSTRA, D., AND RIEDMILLER, M. Playing Atari with Deep Reinforcement Learning, Dec. 2013.
- [18] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K., OSTROVSKI, G., PETERSEN, S., BEATTIE, C., SADIK, A., ANTONOGLOU, I., KING, H., KUMARAN, D., WIERSTRA, D., LEGG, S., AND HASSABIS, D. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533.
- [19] SCHULMAN, J., LEVINE, S., MORITZ, P., JORDAN, M. I., AND ABBEEL, P. Trust Region Policy Optimization, Apr. 2017.
- [20] SCHULMAN, J., MORITZ, P., LEVINE, S., JORDAN, M., AND ABBEEL, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation, June 2015.
- [21] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A., AND KLIMOV, O. Proximal Policy Optimization Algorithms, Aug. 2017.
- [22] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958.
- [23] SUTTON, R. S. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 1984.
- [24] SUTTON, R. S. Learning to predict by the methods of temporal differences. *Machine Learning* 3, 1 (Aug. 1988), 9–44.
- [25] SUTTON, R. S., AND BARTO, A. G. *Reinforcement Learning, Second Edition: An Introduction*, 2nd edition ed. Bradford Books, Nov. 2018.
- [26] SUTTON, R. S., MCALLESTER, D., SINGH, S., AND MANSOUR, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems* (1999), vol. 12, MIT Press.
- [27] VAN HASSELT, H., GUEZ, A., AND SILVER, D. Deep Reinforcement Learning with Double Q-learning, Dec. 2015.
- [28] WATKINS, C. *Learning from Delayed Reward*. PhD thesis, King’s College London, England, May 1989.
- [29] WATKINS, C. J. C. H., AND DAYAN, P. Q-learning. *Machine Learning* 8, 3 (May 1992), 279–292.
- [30] WILLIAMS, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8, 3 (May 1992), 229–256.