TODO describe a decision problem in natural language.

It is assumed that the reader is already familiar with basic machine learning concepts; neural networks & gradient descent.

Reinforcement learning problems are typically viewed as Markov decision processes, these are a special case of decision processes that possess the Markov property. A full specification of the Markov property is outside of the scope of this report, but for the purposes of this report, a process is considered Markov if the future evolution of the process can be predicted using only on the current state.

Before deploying the notation used by the algorithms in this report, some attempt at formally defining an MDP must be made.

There exists a set of states , a set of terminal states 𝕋 (where ), and a set of actions. Time progresses in discrete steps. At each time-step t, the agent receives a state and reward , then takes an action . The process then transitions to a new , chosen stochastically with a probability dependent on and . The process continues until it reaches a point where , then halts.

Note that in the context of reinforcement learning, the manner in which an MDP transitions between states is often called the “transition model”.

Define the term policy. A policy is a mapping , the problem of decision making can be framed as finding a policy that maximises the accumulated reward during an epoch, or over some other period of the process.

Stationary problems VS non-stationary problems (harder).

A non-stationary problem is one in which the transition model changes as the process progresses (significantly harder).

**V-Value.**

This function gives the expected sum of discounted reward after the state , under the policy . Sutton & Barto define the V-value as:

|  |  |
| --- | --- |
|  | [1] |

**Q-Value.**

This function gives the expected sum of discounted reward after taking action in state , then following the policy afterwards. Sutton & Barto defines the Q-value as:

|  |  |
| --- | --- |
|  | [1] |

**Advantage.**

The advantage of an action isolates the action’s contribution to the reward from the inherent value of the state. Sutton & Barto defines the Advantage function as:

There are a variety of ways to classify reinforcement learning methods. [1]

**Policy-Based.** The model produces a probability spread of actions for a given state, this can be sampled from to enact a policy. During training, the model learns to produce the action probabilities that map to the optimal policy.

**State-Value-Based.** The model learns to estimate the V-Values of states. This alone is not useful in constructing a policy. However if the transition model of the process is also known, state-values can be used to construct policies i.e. by calculating the weighted average value of the successor states reached by each action.

**State-Action-Value-Based.** The model learns to approximate the Q-Value of state-action pairs. These values are used to enact a policy i.e. by selecting the action with the highest Q-Value for the current state. Unlike a state-value-based approach, this can be used to construct a policy even if the transition model is not known.

**Actor-Critic.** The model has two components; the actor, and the critic. The actor is policy-based, the critic is value based. Either V-values or Q-values can be used, depending on the algorithm. The critic’s output is used to approximate the advantage function, and the approximation is used to direct the actor during the learning process. Only the actor’s output is necessary to enact the learned policy.

The method by which training data is gathered is also an important factor in the final behaviour of agents. Note that these classifications are orthogonal to the output-structural methods, any combination of the two is possible.

**Monte-Carlo.**

The agent is allowed to interact with the environment until it reaches a terminal state (a full epoch), no learning occurs during the epoch. Once the epoch is over, the model parameters are updated according to the total reward received during the epoch. This means that Monte-Carlo methods can only be applied to problems with epochs of finite length.

**Temporal Difference.**

The agent is allowed to interact with the environment. Model parameters are updated according to the reward received during each transition between states. This allows for learning over the course of a single epoch, which can quicken convergence.

**N-Step.**

This term describes represents a middle-point between Monte-Carlo & Temporal-Difference methods. Updates are performed according to multiple transitions, but not the entire epoch. i.e. training on the first 50 steps of a 100-step epoch.

**~~Dynamic Programming.~~**

~~In contrast to the three prior strategies, in dynamic programming the agent does not interact with the environment. Instead data is gathered by iterating the state space (in an arbitrary order), using the transition model to explore all possible successor states, and updating the model parameters according to the reward obtained. DP requires a finite state and action space, as well as knowledge of the transition model, so it cannot be applied to many real-world problems.~~

There are a number of non-deep approaches; tabular methods, linear regression.

Tabular approaches achieve the desired input-output mapping by constructing a look-up table containing every possible input state, which is held in memory. Tabular methods are backed by theoretical proofs of their ability to converge to optimal policies [3, 4]. Despite this useful property, such methods are not typically practical due to the large state-spaces of real-world machine-learning problems. The associated table would be so large that storing it and updating it during training would be computationally infeasible.

Modern approaches typically use neural networks to approximate TODO. While perceptrons have been proven to converge to the optimal policy over time [5], neural networks in general are not guaranteed to converge, or even to improve. Though in practice they often do.

Methods based on gradient descent can find only the local optimum.

The use of neural networks with hidden layers to solve reinforcement learning problems is known as deep reinforcement learning, and in the last decade it has been applied to achieve human-level performance in complex control problems [6, 7].

Modern RL encompasses a diverse variety of approaches, and there are a variety of overlapping categories by which reinforcement learning algorithms can be classified. The taxonomy of those approaches is of interest if they are to be meaningfully compared.

**Entropy**

The entropy of a policy captures the variance in the actions it prescribes. The entropy of a policy w/r to a particular state is given by:

|  |  |
| --- | --- |
|  | [2] |

Where is the action space, and is the function that gives the probability of the network taking action in state . The policy entropy can be added to the reward at each step of the training process to produce a new target for optimization. Doing so can discourage models from reaching locally optimal policies that prescribe the same action in all, or nearly all states.

*Table 1: Overview of Targeted RL Methods*

|  |  |
| --- | --- |
| Class | Method |
| Q-Value | Deep Q-Learning. |
| SARSA |
| Pure Policy-Based | REINFORCE |
| REINFORCE-MENT |
| Actor-Critic | Basic Actor-Critic |
| Advantage Actor-Critic (A2C) |
| Proximal Policy Optimisation |

**SARSA & Q-Learning**

These are closely related temporal-difference methods.

A transition is a tuple

The network outputs an estimate of If this prediction is accurate, it should be exactly equal to the value of the successor state, plus the observed reward at t. We have access to this observed reward, as well as the network’s prediction of the value of the successor state We can then calculate the squared error and use it as the target of gradient descent.

**Q-Learning Loss Function**

|  |  |
| --- | --- |
|  | [3] |

**SARSA Loss Function**

|  |  |
| --- | --- |
|  |  |

The methods differ in that SARSA is on-policy, whereas Q-learning is off-policy. That is, under SARSA, is the action prescribed by the policy that produced the transition. Whereas under Q-Learning is the action with the highest Q-value (following a greedy policy with respect to ). Even if these policies are initially identical, they will diverge as weight updates are made.

**Policy Gradient Methods.**

The network’s weights are updated according to some rule:

Where is an estimate of , and is the function that gives the probability of the network taking action in state .

If is positive, it is desirable to reinforce this behaviour, so the weights are updated to maximize the probability mass function, making the action more likely to be taken in the future. If it’s negative, they are updated in the opposite direction.

**REINFORCE.**

This is a policy-based Monte-Carlo method, one of the earliest policy-based methods to be discussed. The entire epoch is observed, and the actual sum of discounted rewards after an action is used to calculate . At the end of each epoch, the network parameters are updated according to the rule:

|  |  |
| --- | --- |
|  | [3] |

Where and are hyperparameters of the network, is the length of the epoch, and is the cumulative reward received during the epoch, which serves as an estimate of the action advantage.

**REINFORCE-MENT**

This is a variant of REINFORCE, from [2]. Instead of maximising reward, it targets the sum of reward & policy entropy, discouraging the adoption of low entropy policies.

**Actor-Critic Methods.**

The key feature of actor-critic methods is their use of two networks; the actor & the critic. The actor learns in the same manner as policy-gradient methods, however Â is calculated using the output of the critic.

The critic learns a value function using the same method as the value-based techniques discussed earlier. Note that as the transitions are being produced according to the policy of the actor, critic learning is necessarily off-policy.

There are multiple ways to produce Â, and these methods are described in more detail in the three implementations below. The actor & critic can be two separate networks, or a single network. In the latter, the network has outputs for both action probabilities and the critic outputs.

**Basic Actor-Critic.**

In this algorithm the critic learns to approximate . The critic’s estimate of the Q-value of the actor’s action is normalised and used as . This algorithm includes an entropy term.

**Advantage Actor-Critic (A2C).**

This n-step algorithm is a synchronous version of Asynchronous Advantage Actor-Critic [5]. The critic learns to approximate , and the estimate is used to calculate Â. Every n steps  
- and at the end of the epoch - weight updates are performed on the new transitions, with Â being calculated according to the equation:

|  |  |
| --- | --- |
|  | [5] |

Where is the critic’s approximation of , and is the number of steps since the last weight update. This algorithm includes an entropy term.

**PPO**

This algorithm is similar to A2C, and is calculated in the same manner. The key difference is that in PPO, the magnitude of the weight update is limited within a range, according to:

|  |  |
| --- | --- |
|  | [6] |

Where is a constant such that .

This algorithm includes an entropy term.

REFERENCES

# References

|  |  |
| --- | --- |
| [1] | R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, Second Edition, MIT Press, 2018. |
| [2] | R. J. Williams and J. Peng, “Function Optimization Using Connectionist Reinforcement Learning Algorithms,” *Connection Science,* no. 3, pp. 241-268, 1991. |
| [3] | R. J. Williams, “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning,” *Machine Learning,* vol. 8, p. 229–256, 1992. |
| [4] | V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning,” *arXiv:1312.5602 [cs.LG],* 2013. |
| [5] | M. Volodymyr, A. P. Badia, M. Mirza, A. Graves, T. Harley, L. P. Timothy, S. David and K. Kavukcuoglu, “Asynchronous Methods for Deep Reinforcement Learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, 2016. |
| [6] | J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Proximal Policy Optimization Algorithms,” *arXiv:1707.06347 [cs.LG].* |