

---

# Double $Q(\sigma)$ and $Q(\sigma, \lambda)$ Unifying Reinforcement Learning Control Algorithms

---

**Markus Dumke**

Department of Statistics

Ludwig-Maximilians-Universität München

markus.dumke@campus.lmu.de

## Abstract

Temporal-difference (TD) learning is an important field in reinforcement learning. Sarsa and Q-Learning are among the most used TD algorithms. The  $Q(\sigma)$  algorithm (Sutton and Barto (2017)) unifies both. This paper extends the  $Q(\sigma)$  algorithm to an on-line multi-step algorithm  $Q(\sigma, \lambda)$  using eligibility traces and introduces Double  $Q(\sigma)$  as the extension of  $Q(\sigma)$  to double learning. Experiments suggest that the new  $Q(\sigma, \lambda)$  algorithm can outperform the classical TD control methods Sarsa( $\lambda$ ),  $Q(\lambda)$  and  $Q(\sigma)$ .

## 1 Introduction

Reinforcement Learning is a field of machine learning addressing the problem of sequential decision making. It is formulated as an interaction of an agent and an environment over a number of discrete time steps  $t$ . At each time step the agent chooses an action  $A_t$  based on the environment's state  $S_t$ . The environment takes  $A_t$  as an input and returns the next state observation  $S_{t+1}$  and reward  $R_{t+1}$ , a scalar numeric feedback signal.

The agent is thereby following a policy  $\pi$ , which is the behavior function mapping a state to action probabilities

$$\pi(a|s) = P(A_t = a | S_t = s). \quad (1)$$

The agent's goal is to maximize the return  $G_t$  which is the sum of discounted rewards,

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{T-1} \gamma^k R_{t+1+k}, \quad (2)$$

where  $\gamma \in [0, 1]$  is the discount factor and  $T$  is the length of the episode or infinity for a continuing task.

While rewards are short-term signals about the goodness of an action, values represent the long-term value of a state or state-action pair. The action value function  $q_\pi(s, a)$  is defined as the expected return taking action  $a$  from state  $s$  and thereafter following policy  $\pi$ :

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]. \quad (3)$$

Value-based reinforcement learning is concerned with finding the optimal action value function  $q_* = \max_\pi q_\pi$ . Temporal-difference learning is a class of model-free methods which estimates  $q_\pi$  from sample transitions and iteratively updates the estimated values using observed rewards and estimated values of successor actions. At each step an update of the following form is applied:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \delta_t, \quad (4)$$

where  $Q$  is an estimate of  $q_\pi$ ,  $\alpha$  is the step size and  $\delta_t$  is the TD error, the difference between our current estimate and a newly computed target value. The following TD control algorithms can all be characterized by their different TD errors.

When the action values  $Q$  are represented as a table we call this tabular reinforcement learning, else we speak of approximate reinforcement learning, e.g. when using a neural network to compute the action values. For sake of simplicity the following analysis is done for tabular reinforcement learning but can be easily extended to function approximation.

## 2 TD control algorithms: From Sarsa to Q( $\sigma$ )

Sarsa (Rummery and Niranjan (1994)) is a temporal-difference learning algorithm which samples states and actions using an  $\epsilon$ -greedy policy and then updates the  $Q$  values using Equation 4 with the following TD error

$$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t). \quad (5)$$

The term  $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$  is called the TD target and consists of the reward plus the discounted value of the next state and next action.

Sarsa is an on-policy method, i.e. the TD target consists of  $Q(S_{t+1}, A_{t+1})$ , where  $A_{t+1}$  is sampled using the current policy. In general the policy used to sample the state and actions - the so called behaviour-policy  $\mu$  - can be different from the target policy  $\pi$ , which is used to compute the TD target. If behaviour and target policy are different we call this off-policy learning. An example for an off-policy TD control algorithm is the well known Q-Learning algorithm proposed by Watkins (1989). As in Sarsa states and actions are sampled using an exploratory behaviour policy, e.g. an  $\epsilon$ -greedy policy, but the TD target is computed using the greedy policy with respect to the current  $Q$  values. The TD error of Q-Learning is

$$\delta_t = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t). \quad (6)$$

Expected Sarsa generalizes Q-Learning to arbitrary target policies. The TD error is

$$\delta_t = R_{t+1} + \gamma \sum_{a'} \pi(a' | S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t). \quad (7)$$

The current state-action pair is updated using the expectation of all subsequent action values with respect to the action value. Q-Learning is a special case of Expected Sarsa if  $\pi$  is the greedy policy with respect

to  $Q$  (Sutton and Barto (2017)). Of course Expected Sarsa could also be used as an on-policy algorithm if the target policy is chosen to be the same as the behaviour policy (Van Seijen et al. (2009)).

Sutton and Barto (2017) propose a new TD control algorithm called  $Q(\sigma)$  which unifies Sarsa and Expected Sarsa. The TD target of this new algorithm is a weighted mean of the Sarsa and Expected Sarsa TD targets, where the parameter  $\sigma$  controls the weighting.  $Q(1)$  is equal to Sarsa and  $Q(0)$  is equal to Expected Sarsa. For intermediate values of  $\sigma$  new algorithms are obtained, which can achieve better performance (Asis et al. (2017)).

The TD error of  $Q(\sigma)$  is

$$\delta_t = R_{t+1} + \gamma(\sigma Q(S_{t+1}, A_{t+1}) + (1 - \sigma) \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a')) - Q(S_t, A_t). \quad (8)$$

### 3 $Q(\sigma, \lambda)$ : An on-line multi-step algorithm

The TD methods presented so far are one-step methods, which use only rewards and values from the next step  $t + 1$ . These can be extended to use eligibility traces to incorporate data of multiple time steps.

An eligibility trace is a scalar numeric value for each state-action pair. Whenever a state-action pair is visited its eligibility is increased, if not, the eligibility fades away over time. State-action pairs visited often will have a higher eligibility than those visited less frequently and state-action pairs visited recently will have a higher eligibility than those visited long time ago.

The accumulating eligibility trace (Singh and Sutton (1996)) uses an update of the form

$$E_{t+1}(s, a) = \begin{cases} \gamma\lambda E_t(s, a) + 1, & \text{if } A_t = a, S_t = s \\ \gamma\lambda E_t(s, a), & \text{otherwise.} \end{cases} \quad (9)$$

Whenever taking action  $A_t$  in state  $S_t$  the eligibility of this pair is increased by 1 and for all states and actions decreased by a factor  $\gamma\lambda$ , where  $\lambda$  is the trace decay parameter.

Then all state-action pairs are updated according to their eligibility trace

$$Q(s, a) \leftarrow Q(s, a) + \alpha \delta_t E_t(s, a) \quad (10)$$

The corresponding algorithm using the one-step Sarsa TD error and an update using eligibility traces is called Sarsa( $\lambda$ ). Though it looks like a one-step algorithm, it is in fact a multi-step algorithm, because the current TD error is assigned back to all previously visited states and actions weighted by their eligibility.

For off-policy algorithms like Q-Learning different eligibility updates have been proposed. Watkin's  $Q(\lambda)$  uses the same updates as long as the greedy action is chosen by the behaviour policy, but sets the  $Q$  values to 0, whenever a non-greedy action is chosen assigning credit only to state-action pairs we would actually have visited if following the target policy  $\pi$  and not the behaviour policy  $\mu$ . More generally the eligibility is weighted by the target policy's probability of the next action. The update rule is then

$$E_{t+1}(s, a) = \begin{cases} \gamma\lambda E_t(s, a) \pi(A_{t+1}|S_{t+1}) + 1, & \text{if } A_t = a, S_t = s \\ \gamma\lambda E_t(s, a) \pi(A_{t+1}|S_{t+1}), & \text{otherwise.} \end{cases} \quad (11)$$

Whenever an action occurs, which is unlikely in the target policy, the eligibility of all previous states is decreased sharply. If the target policy is the greedy policy, the eligibility will be set to 0 for the complete history.

In this paper we introduce a new kind of eligibility trace update to extend the  $Q(\sigma)$  algorithm to an on-line multi-step algorithm, which we will call  $Q(\sigma, \lambda)$ . Recall that the one-step target of  $Q(\sigma)$  is a weighted average between the on-policy Sarsa and off-policy Expected Sarsa targets weighted by the factor  $\sigma$ :

$$\delta_t = R_{t+1} + \gamma(\sigma Q(S_{t+1}, A_{t+1}) + (1 - \sigma) \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a')) - Q(S_t, A_t) \quad (12)$$

In this paper we propose to weight the eligibility accordingly with the same factor  $\sigma$ . The eligibility is then a weighted average between the on-policy eligibility used in Sarsa( $\lambda$ ) and the off-policy eligibility used in  $Q(\lambda)$ . The eligibility trace is updated at each step by

$$E_{t+1}(s, a) = \begin{cases} \gamma \lambda E_t(s, a)(\sigma + (1 - \sigma)\pi(A_{t+1}|S_{t+1})) + 1, & \text{if } A_t = a, S_t = s \\ \gamma \lambda E_t(s, a)(\sigma + (1 - \sigma)\pi(A_{t+1}|S_{t+1})), & \text{otherwise.} \end{cases} \quad (13)$$

When  $\sigma = 0$  the one-step target of  $Q(\sigma)$  is equal to the Sarsa one-step target and therefore the eligibility update reduces to the standard accumulate eligibility trace update. When  $\sigma = 1$  the one-step target of  $Q(\sigma)$  is equal to the Expected Sarsa target and accordingly the eligibility is weighted by the target policy's probability of the current action. For intermediate values of  $\sigma$  the eligibility is weighted in the same way as the TD target. Asis et al. (2017) showed that n-step  $Q(\sigma)$  with an intermediate or dynamic value of  $\sigma$  can outperform Q-Learning and Sarsa. By extending this algorithm to an on-line multi-step algorithm we can make use of the good initial performance of Sarsa( $\lambda$ ) combined with the good asymptotic performance of  $Q(\lambda)$ . In comparison to the n-step  $Q(\sigma)$  algorithm (Asis et al. (2017)) the new  $Q(\sigma, \lambda)$  algorithm can learn on-line and is therefore likely to learn faster.

Pseudocode for tabular episodic  $Q(\sigma, \lambda)$  is given in Algorithm 1. This can be easily extended to continuing tasks and to function approximation using one eligibility per weight of the function approximator.

---

**Algorithm 1**  $Q(\sigma, \lambda)$

---

```

Initialize  $Q(s, a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ 
Repeat for each episode:
     $E(s, a) \leftarrow 0 \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ 
    Initialize  $S_0 \neq \text{terminal}$ 
    Choose  $A_0$ , e.g.  $\epsilon$ -greedy from  $Q(S_0, \cdot)$ 
    Loop for each step of episode:
        Take action  $A_t$ , observe reward  $R_{t+1}$  and next state  $S_{t+1}$ 
        Choose next action  $A_{t+1}$ , e.g.  $\epsilon$ -greedy from  $Q(S_{t+1}, \cdot)$ 
         $\delta = R_{t+1} + \gamma(\sigma Q(S_{t+1}, A_{t+1}) + (1 - \sigma) \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a')) - Q(S_t, A_t)$ 
         $E(S_t, A_t) \leftarrow E(S_t, A_t) + 1$ 
         $Q(s, a) \leftarrow Q(s, a) + \alpha \delta E(s, a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ 
         $E(s, a) \leftarrow \gamma \lambda E(s, a)(\sigma + (1 - \sigma)\pi(A_{t+1}|S_{t+1})) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ 
         $A_t \leftarrow A_{t+1}, S_t \leftarrow S_{t+1}$ 
        If  $S_t$  is terminal: Break

```

---

## 4 Double $Q(\sigma)$ Algorithm

Double learning is another extension of the basic algorithms. It has been mostly studied with Q-Learning Hasselt (2010) and prevents the overestimation of action values when using Q-Learning in stochastic

environments. The idea is to use decouple action selection (which action is the best one?) and action evaluation (what is the value of this action?). The implementation is simple, instead of using only one value function we will use two value functions  $Q_A$  and  $Q_B$ . Actions are sampled due to an  $\epsilon$ -greedy policy with respect to  $Q_A + Q_B$ . Then at each step either  $Q_A$  or  $Q_B$  is updated, e.g. if  $Q_A$  is selected by

$$Q_A(S_t, A_t) \leftarrow Q_A(S_t, A_t) + \alpha(R_{t+1} + \gamma Q_B(\arg\max_{a \in \mathcal{A}} Q_A(S_{t+1}, a)) - Q_A(S_t, A_t)) \quad (14)$$

$$Q_B(S_t, A_t) \leftarrow Q_B(S_t, A_t) + \alpha(R_{t+1} + \gamma Q_A(\arg\max_{a \in \mathcal{A}} Q_B(S_{t+1}, a)) - Q_B(S_t, A_t)) \quad (15)$$

Double learning can also be used with Sarsa and Expected Sarsa as proposed by Michael Ganger and Hu (2016). Using double learning these algorithms can be more robust and perform better in stochastic environments. The decoupling of action selection and action evaluation is weaker than in Double Q-Learning because the next action  $A_{t+1}$  is selected according to an  $\epsilon$ -greedy behavior policy using  $Q_A + Q_B$  and evaluated either with  $Q_A$  or  $Q_B$ . For Expected Sarsa the policy used for the target in Equation 7 could be the  $\epsilon$ -greedy behavior policy as proposed by Michael Ganger and Hu (2016), but it is probably better to use a policy according to  $Q_A$  (if updating  $Q_A$ ), because then it can also be used off-policy with Double Q-Learning as a special case, if  $\pi$  is the greedy policy with respect to  $Q_A$ .

In this paper we propose the extension of double learning to  $Q(\sigma)$  - Double  $Q(\sigma)$  - to obtain a new algorithm with the good learning properties of double learning, which generalizes (Double) Q-Learning, (Double) Expected Sarsa and (Double) Sarsa. Of course Double  $Q(\sigma)$  can also be used with eligibility traces.

Double  $Q(\sigma)$  has the following TD error when  $Q_A$  is selected,

$$\delta_t = R_{t+1} + \gamma \left( \sigma Q_B(S_{t+1}, A_{t+1}) + (1 - \sigma) \sum_a \pi(a|S_{t+1}) Q_B(S_{t+1}, a) \right) - Q_A(S_t, A_t) \quad (16)$$

and

$$\delta_t = R_{t+1} + \gamma \left( \sigma Q_A(S_{t+1}, A_{t+1}) + (1 - \sigma) \sum_a \pi(a|S_{t+1}) Q_A(S_{t+1}, a) \right) - Q_B(S_t, A_t) \quad (17)$$

if  $Q_B$  is selected. The target policy  $\pi$  is computed with respect to the value function which is updated, i.e. with respect to  $Q_A$  in Equation 16 and with respect to  $Q_B$  in Equation 17.

Pseudocode for Double  $Q(\sigma)$  is given in Algorithm 2.

## 5 Experiments

In this section the performance of the newly proposed  $Q(\sigma, \lambda)$  algorithm will be tested on a gridworld navigation task compared with the performance of classical TD control algorithms like Sarsa and Q-Learning as well as  $Q(\sigma)$ .

The windy gridworld is a simple navigation task described by Sutton and Barto (1998). The goal is to get as fast as possible from a start state to a goal state using the actions left, right, up or down. In each

---

**Algorithm 2** Double  $Q(\sigma)$

---

```

Initialize  $Q_A(s, a)$  and  $Q_B(s, a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ 
Repeat for each episode:
  Initialize  $S_0 \neq \text{terminal}$ 
  Choose  $A_0$ , e.g.  $\epsilon$ -greedy from  $Q_A(S_0, \cdot) + Q_B(S_0, \cdot)$ 
  Loop for each step of episode:
    Take action  $A_t$ , observe reward  $R_{t+1}$  and next state  $S_{t+1}$ 
    Choose next action  $A_{t+1}$ , e.g.  $\epsilon$ -greedy from  $Q_A(S_{t+1}, \cdot) + Q_B(S_{t+1}, \cdot)$ 
    Randomly update either  $Q_A$ :
       $\delta = R_{t+1} + \gamma(\sigma Q_B(S_{t+1}, A_{t+1}) +$ 
         $(1 - \sigma) \sum_a \pi(a|S_{t+1}) Q_B(S_{t+1}, a)) - Q_A(S_t, A_t)$ 
       $Q_A(S_t, A_t) \leftarrow Q_A(S_t, A_t) + \alpha \delta$ 
    or update  $Q_B$ :
       $\delta = R_{t+1} + \gamma(\sigma Q_A(S_{t+1}, A_{t+1}) +$ 
         $(1 - \sigma) \sum_a \pi(a|S_{t+1}) Q_A(S_{t+1}, a)) - Q_B(S_t, A_t)$ 
       $Q_B(S_t, A_t) \leftarrow Q_B(S_t, A_t) + \alpha \delta$ 
     $A_t \leftarrow A_{t+1}, S_t \leftarrow S_{t+1}$ 
  If  $S_t$  is terminal: Break

```

---

column of the grid the agent is pushed upward by a wind. When an action would take the agent outside the grid, the agent is placed in the nearest cell inside the grid. The stochastic windy gridworld (Asis et al. (2017)) is a variant where state transitions are random, with a probability of 0.1 the agent will transition to one of the surrounding eight states independent of the action. The task is treated as an undiscounted episodic task with a reward of -1 for each transition. Figure 1 visualizes the gridworld.

Experiments were conducted using an  $\epsilon$ -greedy behaviour policy with  $\epsilon = 0.1$ . The performance in terms of the average return over the first 100 episodes was measured for different values of  $\sigma$  and  $\lambda$  as a function of the step size  $\alpha$ . For the Expected Sarsa part of the update a greedy target policy was chosen, i.e.  $Q(0)$  is exactly Q-Learning. Results were averaged over 200 independent runs.

Figure 2 shows that an intermediate value of  $\sigma = 0.5$  performed better than Sarsa ( $Q(1)$ ) and Q-Learning ( $Q(0)$ ). The best performance was found by dynamically varying  $\sigma$  over time, i.e. decreasing  $\sigma$  by a factor of 0.99 after each episode. Multi-step bootstrapping with a trace decay parameter  $\lambda = 0.7$  performed better than the one-step algorithms ( $\lambda = 0$ ). Dynamically varying the value of  $\sigma$  allows to combine the good initial performance of Sarsa with the good asymptotic performance of Expected Sarsa. This confirms the results observed by Asis et al. (2017) for n-step algorithms.

## 6 Conclusions

This paper has presented two extensions to the  $Q(\sigma)$  algorithm, which unify Q-Learning, Expected Sarsa and Sarsa.  $Q(\sigma, \lambda)$  extends the algorithm to an on-line multi-step algorithm using eligibility traces and Double  $Q(\sigma)$  extends the algorithm to double learning. Empirical results suggest that  $Q(\sigma, \lambda)$  can outperform classic TD control algorithms like Sarsa( $\lambda$ ),  $Q(\lambda)$  and  $Q(\sigma)$ . Dynamically varying  $\sigma$  obtains the best results.

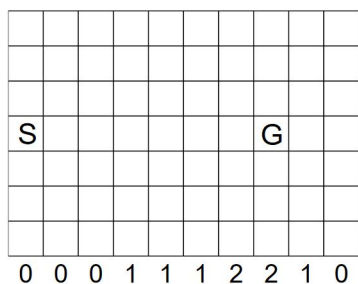


Figure 1: The windy gridworld task. The goal is to move from the start state S to the goal state G while facing an upward wind in the middle of the grid, which is denoted in the numbers below the grid. Described by Sutton and Barto (1998).

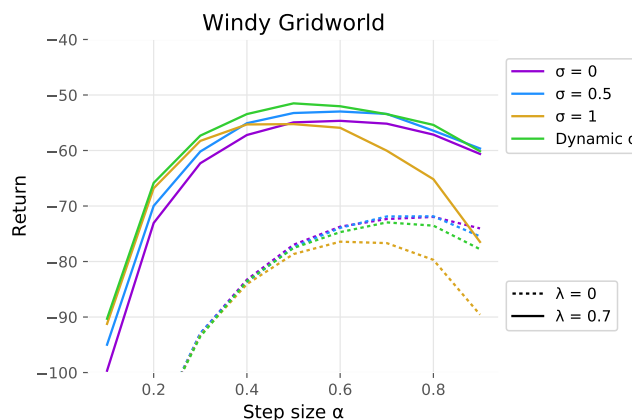


Figure 2: Stochastic windy gridworld results averaged over 100 episodes and 200 independent runs. Performance of  $Q(\sigma, \lambda)$  for different values of  $\sigma$  as a function of the step size  $\alpha$ . For the trace decay parameter  $\lambda = [0, 0.7]$  were used. The best performance was found when using a dynamic value of  $\sigma$  by multiplying  $\sigma$  with a factor 0.99 after each episode.

Future research might focus on performance of  $Q(\sigma, \lambda)$  when used with non-linear function approximation and different schemes to update  $\sigma$  over time.

## References

- Asis, K. D., Hernandez-Garcia, J. F., Holland, G. Z. and Sutton, R. S. (2017). Multi-step reinforcement learning: A unifying algorithm, *CoRR* **abs/1703.01327**.  
 URL: <http://arxiv.org/abs/1703.01327>
- Hasselt, H. V. (2010). Double q-learning, in J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta (eds), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., pp. 2613–2621.  
 URL: <http://papers.nips.cc/paper/3964-double-q-learning.pdf>
- Michael Ganger, E. D. and Hu, W. (2016). Double sarsa and double expected sarsa with shallow and deep learning, *Journal of Data Analysis and Information Processing* **4**: 159–176.
- Rummery, G. A. and Niranjan, M. (1994). On-line q-learning using connectionist systems, *Technical report*.
- Singh, S. P. and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces, *Machine Learning* **22**(1): 123–158.
- Sutton, R. S. and Barto, A. G. (1998). *Introduction to Reinforcement Learning*, 1st edn, MIT Press, Cambridge, MA, USA.
- Sutton, R. S. and Barto, A. G. (2017). Reinforcement learning : An introduction. Accessed: 2017-08-01.

- Van Seijen, H., Van Hasselt, H., Whiteson, S. and Wiering, M. (2009). A theoretical and empirical analysis of expected sarsa, *Adaptive Dynamic Programming and Reinforcement Learning, 2009. AD-PRL'09. IEEE Symposium on*, IEEE, pp. 177–184.
- Watkins, C. J. C. H. (1989). *Learning from delayed rewards*, PhD thesis, King's College, Cambridge.