

Variable discount factor learning in Markov Decision Process

Davide Tateo, Alessandro Nuara, Carlo D'Eramo
Dipartimento di Elettronica, Informazione e Biongegneria Politecnico Di Milano
Milano, Italy
Email: {davide.tateo, alessandro.nuara, carlo.deramo}@polimi.it

Abstract—TODO

I. INTRODUCTION

Motivations.

State of the art (Q-Learning [1], SARSA, Double, Weighted, R-Learning)

A. Subsection Heading Here

1) Subsubsection Heading Here:

II. PRELIMINARIES

A. Decomposition of the TD error

Decompose Q function:

$$\begin{aligned} Q(x, u) &= \mathbb{E} [R(x, u, x') + \gamma Q(x', \pi(x'))] \\ &= \mathbb{E} [R(x, u, x')] + \gamma \mathbb{E} [Q(x', \pi(x'))] \\ &= \tilde{R}(x, u) + \gamma \tilde{Q}(x, u) \end{aligned} \quad (1)$$

Decomposed TD update:

$$\tilde{R}(x, u) \leftarrow \tilde{R}(x, u) + \alpha(R(x, u, x') - \tilde{R}(x, u)) \quad (2)$$

$$\tilde{Q}(x, u) \leftarrow \tilde{Q}(x, u) + \beta(Q(x', \pi(x')) - \tilde{Q}(x, u)) \quad (3)$$

Update of the Q function:

$$\begin{aligned} Q(x, u) &\leftarrow \tilde{R}(x, u) + \alpha(R(x, u, x') - \tilde{R}(x, u)) \\ &\quad + \gamma \left(\tilde{Q}(x, u) + \beta(Q(x', \pi(x')) - \tilde{Q}(x, u)) \right) \\ &= Q(x, u) + \alpha(R(x, u, x') - \tilde{R}(x, u)) \\ &\quad + \gamma\beta(Q(x', \pi(x')) - \tilde{Q}(x, u)) \end{aligned} \quad (4)$$

B. Analysis of the decomposed update

If $\alpha = \beta$

$$\begin{aligned} Q(x, u) &\leftarrow Q(x, u) + \alpha(R(x, u, x') + \gamma Q(x', \pi(x')) \\ &\quad - Q(x, u)) \end{aligned} \quad (5)$$

That is the classical Q-Learning update

If $\beta = \delta\alpha$

$$\begin{aligned} Q(x, u) &\leftarrow Q(x, u) + \alpha(R(x, u, x') + \gamma\delta Q(x', \pi(x')) \\ &\quad - (\tilde{R}(x, u) + \gamma\delta\tilde{Q}(x, u))) \\ &= Q(x, u) + \alpha(R(x, u, x') + \gamma'Q(x', \pi(x')) \\ &\quad - (\tilde{R}(x, u) + \gamma'\tilde{Q}(x, u))) \\ &= Q(x, u) + \alpha((R(x, u, x') + \gamma'Q(x', \pi(x')) \\ &\quad - Q'(x, u)) \end{aligned} \quad (6)$$

With $\gamma' = \gamma\delta$. Notice that $Q'(x, u)$ is the current Q function with a different learning rate.

C. Variance dependent learning rate

$$\alpha = \frac{\sigma^2}{\sigma^2 + 1} \quad (7)$$

III. EXPERIMENTAL RESULTS

IV. CONCLUSION

REFERENCES

- [1] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.