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*—**

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:

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

Decomposed TD update:

$$\tilde{R}(x,u) \leftarrow \tilde{R}(x,u) + \alpha(R(x,u,x') - \tilde{R}(x,u)) \tag{2}$$
$$\tilde{Q}(x,u) \leftarrow \tilde{Q}(x,u) + \beta(Q(x',\pi(x')) - \tilde{Q}(x,u)) \tag{3}$$

Update of the Q function:

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

B. Analysis of the decomposed update

If $\alpha = \beta$

$$Q(x,u) \leftarrow Q(x,u) + \alpha(R(x,u,x') + \gamma Q(x',\pi(x'))) \tag{5}$$
$$\tag{6}$$

That is the classical Q-Learning update

If $\beta = \delta\alpha$

$$\begin{aligned} Q(x,u) \leftarrow\ & Q(x,u) + \alpha(R(x,u,x') + \gamma\delta Q(x',\pi(x')) \\ & - (\tilde{R}(x,u) + \gamma\delta\tilde{Q}(x,u))) \\ & = Q(x,u) + \alpha(R(x,u,x') + \gamma' Q(x',\pi(x')) \\ & - (\tilde{R}(x,u) + \gamma'\tilde{Q}(x,u))) \\ & = Q(x,u) + \alpha((R(x,u,x') + \gamma' Q(x',\pi(x'))) \\ & - Q'(x,u)) \end{aligned} \tag{7}$$

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

C. Variance dependent learning rate

$$\alpha = \frac{\sigma^2}{\sigma^2 + 1} \tag{8}$$

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.