Rewarding Air Combat Behavior in Training Simulations[1]

Armon Toubman[a] Jan Joris Roessingh[a] Pieter Spronck[b]
Aske Plaat[c] Jaap van den Herik[c]

[a] *National Aerospace Laboratory NLR, P.O. Box 90502, 1006 BM Amsterdam*
[b] *Tilburg University, P.O. Box 90153, 5000 LE Tilburg*
[c] *Leiden University, P.O. Box 9500, 2300 RA Leiden*

Abstract

Air combat training simulations require high quality virtual agents for optimal training. Reinforcement learning techniques offer the ability to let these virtual agents learn good behavior by rewarding them for their performances. In air combat, the (human or virtual) pilot's missiles should hit the opponent, while the pilots should avoid being hit by their opponent's missiles. However, missile hits depend on chance factors called the *probability-of-kill* (P_k). This makes learning good air combat behavior quite difficult, as low P_k missiles may be rewarded and high P_k missiles may be punished. We propose a method of incorporating the P_k into the learning process. Simulations show that rewarding virtual agents via reinforcement learning based on the P_k of their missiles leads to a 10%-19% increase in performance under various conditions.

1 Introduction

Air combat training simulations are inhabited by virtual agents that perform a variety of roles. These agents should display realistic and adaptive behavior for effective training of fighter pilots. Traditionally, the behavior of these agents is scripted. However, good scripts are hard to develop and maintain. Therefore, we turn to reinforcement learning (RL) to let the agents discover improved behavior from their performances [1, 2], before using these agents as enemies in human-in-the-loop simulations.

Using RL methods, the virtual agents perform actions in many distinct environments and receive feedback on their actions in the form of reward signals [3]. Typically, these signals are 0 (failure) or 1 (success), but complex learning problems may require more complex reward signals.

In air combat, 'success' can be defined as hitting your enemy with a missile. However, missile hits are subject to the *probability-of-kill* (P_k), which is the product of the many factors that influence whether a missile will hit its target. These factors include, e.g., the distance flown by the missile, and the use of countermeasures by the target. Essentially, the P_k is the expected value of a missile hit. Due to a given P_k it is possible for an agent to miss out on a reward because its missile misses the target, despite having a near-optimal policy. Similarly, an agent can have a sub-optimal policy, yet still manage to hit its target and collect a reward. These chance hits and misses obstruct learning, as good behavior is not always rewarded and therefore not reinforced, while sub-optimal behavior may be reinforced. We believe that learning air combat behavior can be improved by taking the P_k into account in the rewards.

---

[1] The full paper has been published in *Systems, Man and Cybernetics (SMC), 2015 IEEE International Conference on*.

2 Method

Two blue RL agents learned to fight a scripted red agent in an air combat simulator. The blue agents learned using the dynamic scripting technique [4], which recombines behavior rules into scripts.

The blues were rewarded using one of three reward functions. The first reward function (binary) rewarded the blues with 0 if they lost and 1 if they won an encounter. The second reward function (domain knowledge) rewarded the blues based on a combination of factors, such as how many missiles they fired during an encounter, and how long they took to end the encounter. The third reward function (P_k) rewarded the blues proportionally to the P_k of the missiles that they fired, rather than the actual hits.

The red agent used one of three statically scripted tactics. Furthermore, red also used a mixed tactic by which it would select one of the three scripted tactics. Red used this tactic until it lost, at which point it would select a new tactic at random.

3 Results

Figure 1 shows the learning curves of the blues, using each of the reward functions, fighting against each of red's tactics. The benefit of using the P_k rewards over the other rewards ranged from a 10%-19% increase in final performance (in terms of trials won).

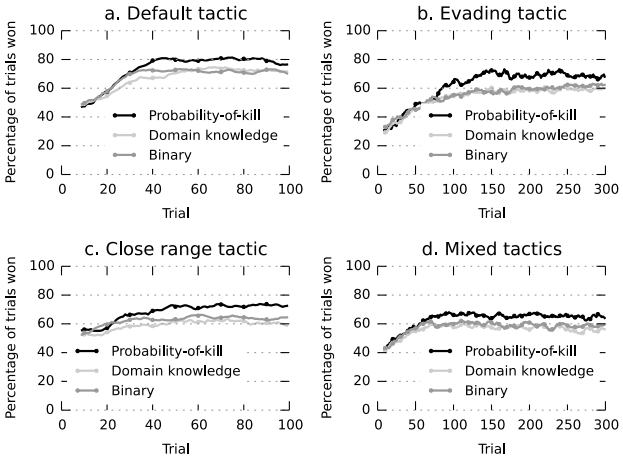

Figure 1. Percentages of trials won by the blue team using each of the three reward functions (probability-of-kill, domain knowledge-based and binary), against each of red's tactics. Rolling mean, window size 10.

4 Conclusion

Rewarding the virtual air combat agents via RL using the P_k of their missiles leads to (1) less rewards for missiles that hit by chance, and to (2) less punishment for missiles that miss by chance. These effects result in a 10%-19% increase in performance against enemies using various tactics. In essence, the use of P_k rewards allows us to automatically generate better virtual agents. The application of P_k rewards as presented in this paper is not limited to the air combat domain, as rewards on the basis of expected values can be used in the RL of any stochastic process model.

References

[1] A. Toubman, J. J. Roessingh, P. Spronck, A. Plaat and H. J. van den Herik, "Dynamic Scripting with Team Coordination in Air Combat Simulation," in *Modern Advances in Applied Intelligence*, Kaohsiung, Taiwan, 2014.

[2] A. Toubman, J. J. Roessingh, P. Spronck, A. Plaat and H. J. van den Herik, "Centralized Versus Decentralized Team Coordination using Dynamic Scripting," in *Proceedings of the 28th European Simulation and Modelling Conference*, Porto, Portugal, 2014.

[3] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, Cambridge, MA, USA: MIT Press, 1998.

[4] P. Spronck, M. Ponsen, I. Sprinkhuizen-Kuyper and E. Postma, "Adaptive game AI with dynamic scripting," *Machine Learning 63.3,* pp. 217-248, 2006.