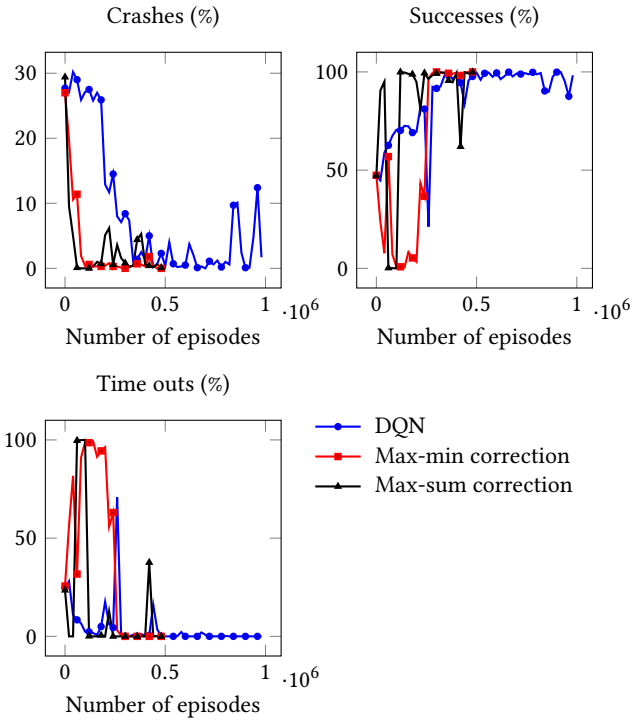

Figure 5: Evaluation of the policy performance throughout the training. The correction function is being trained on twice as less samples than the regular DQN policy but is still converging and outperforming the other for both choices of the decomposition method (max-sum or max-min).

corrective factor can be done with fewer training samples than directly learning the value function of the multi-entity setting.

In the future, more sophisticated multi-fidelity optimization techniques could be used to represent the correction term. Rather than an additive correction, we could try a multiplicative term [6]. Another possibility involves learning the function used for utility fusion itself. A straightforward extension would be to learn a linear weighted combination of the utilities from the single entity problem. Finally, we would like to explore the generality of the correction method. We wish to extend the use of decomposition methods to correcting policies coming from potentially different solving techniques such as an offline POMDP solver. Further experiments for different applications than the one presented in this paper could also highlight the benefit of learning a correction to an existing policy.

REFERENCES

[1] Haoyu Bai, David Hsu, and Wee Sun Lee. 2014. Integrated perception and planning in the continuous space: A POMDP approach. *International Journal of Robotics Research* 33, 9, 1288–1302.

[2] Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. 2012. Intention-Aware Motion Planning. In *Algorithmic Foundations of Robotics X*. 475–491.

[3] Sebastian Brechtel, Tobias Gindele, and Rüdiger Dillmann. 2014. Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*. 392–399.

[4] James P Chryssanthacopoulos and Mykel J Kochenderfer. 2012. Decomposition methods for optimized collision avoidance with multiple threats. *AIAA Journal of Guidance, Control, and Dynamics* 35, 2, 398–405.

[5] Mark Cutler, Thomas J Walsh, and Jonathan P How. 2015. Real-world reinforcement learning via multifidelity simulators. *IEEE Transactions on Robotics* 31, 3, 655–671.

[6] Michael Eldred and Daniel Dunlavy. 2006. Formulations for Surrogate-Based Optimization with Data Fit, Multifidelity, and Reduced-Order Models. *AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*.

[7] Martin Gottwald, Dominik Meyer, Hao Shen, and Klaus Diepold. 2017. Learning to walk with prior knowledge. In *IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. 1369–1374.

[8] Shixiang Gu, Ethan Holly, Timothy P. Lillicrap, and Sergey Levine. 2017. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE International Conference on Robotics and Automation (ICRA)*. 3389–3396.

[9] Mykel J Kochenderfer. 2015. *Decision Making Under Uncertainty: Theory and Application*. MIT Press.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540, 529–533.

[11] Hao Yi Ong and Mykel J. Kochenderfer. 2015. Short-term conflict resolution for unmanned aircraft traffic management. In *Digital Avionics Systems Conference (DASC)*. 5A4–1–5A4–13.

[12] Dev Rajnarayan, Alex Haas, and Ilan Kroo. 2008. A Multifidelity Gradient-Free Optimization Method and Application to Aerodynamic Design. *AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*.

[13] Julio K Rosenblatt. 2000. Optimal selection of uncertain actions by maximizing expected utility. *Autonomous Robots* 9, 1, 17–25.

[14] Stuart J. Russell and Andrew Zimdars. 2003. Q-Decomposition for Reinforcement Learning Agents. In *International Conference on Machine Learning (ICML)*. 656–663.

[15] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. In *International Conference on Learning Representations (ICLR)*.

[16] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*.

[17] William D. Smart and Leslie Pack Kaelbling. 2002. Effective reinforcement learning for mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA)*. 3404–3410 vol.4.

[18] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning - an Introduction*. MIT Press.

[19] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.. In *AAAI Conference on Artificial Intelligence (AAAI)*. 2094–2100.

[20] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*. 1995–2003.