Algorithm	Overall				Last 25k steps			
	Standard ER		DER		Standerd ER		DER	
	Mean reward	vs. baseline	Mean reward	vs. baseline	Mean reward	vs. baseline	Mean reward	vs. baseline
CN*	–	–	8.883	–	–	–	7.085	–
DCRAC	8.923	+0.450%	9.433	+6.192%	7.642	+7.862%	8.525	+20.325%
DCRAC-M	8.437	-5.021%	9.131	+2.792%	6.653	-6.097%	8.843	+24.813%
	Mean regret	vs. baseline	Mean regret	vs. baseline	Mean regret	vs. baseline	Mean regret	vs. baseline
CN*	–	–	4.689	–	–	–	2.937	–
DCRAC	4.856	+3.562%	4.053	-13.564%	3.360	+14.402%	2.591	**-11.781%**
DCRAC-M	4.385	-6.483%	4.321	-7.848%	3.164	+7.729%	2.676	-8.887%

*CN+DER is the baseline used for comparison.

Table 2: Cumulative Average Episodic Reward and Regret for 3×3 partial view in DST

5 CONCLUSION

We propose DCRAC, a Deep Recurrent Actor-Critic approach, for decision making in partially-observable multi-objective environments. The actor-critic network is conditioned on the weights, i.e., the preferences of different objectives. Hence, DCRAC can generalize to different weights and can handle scenarios where the weights change dynamically over time. We also propose DCRAC-M, which uses memory networks for remembering long-term dependencies in the action-observation histories. We use diverse experience replay to sample transitions while training the network to prevent overfitting on recently trained weights. Experiments on the partially-observable version of DST shows that the DCRAC outperforms Conditioned Networks. As future-work, we plan to conduct experiments on more complex scenarios such as minecart.

REFERENCES

[1] Axel Abels, Diederik M Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. 2018. Dynamic Weights in Multi-Objective Deep Reinforcement Learning. *arXiv preprint arXiv:1809.07803* (2018).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[4] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*.

[5] Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. 2015. Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455* (2015).

[6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[7] Seungchan Kim, Kavosh Asadi, Michael Littman, and George Konidaris. 2019. DeepMellow: Removing the Need for a Target Network in Deep Q-Learning. (2019).

[8] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[10] Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. 2016. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707* (2016).

[11] Junhyuk Oh, Valliappa Chockalingam, Satinder Singh, and Honglak Lee. 2016. Control of memory, active perception, and action in minecraft. *arXiv preprint arXiv:1605.09128* (2016).

[12] Christos H Papadimitriou and John N Tsitsiklis. 1987. The complexity of Markov decision processes. *Mathematics of operations research* 12, 3 (1987), 441–450.

[13] Julien Perez and Tomi Silander. 2017. Non-markovian control with gated end-to-end memory policy networks. *arXiv preprint arXiv:1705.10993* (2017).

[14] Diederik Marijn Roijers, Shimon Whiteson, and Frans A Oliehoek. 2015. Point-based planning for multi-objective POMDPs. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[15] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.

[16] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* (2015).

[17] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.

[18] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

[19] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. 2015. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581* (2015).

[20] Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. 2007. Solving deep memory POMDPs with recurrent policy gradients. In *International Conference on Artificial Neural Networks*. Springer, 697–706.

[21] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. 2017. On improving deep reinforcement learning for pomdps. *arXiv preprint arXiv:1704.07978* (2017).