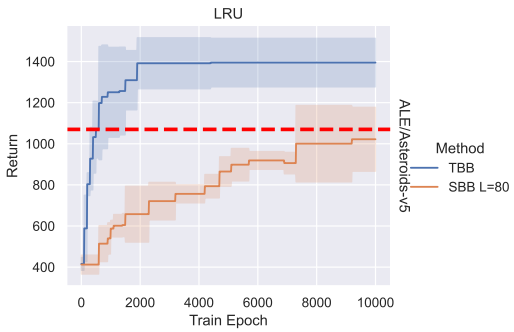
We have added two new experiments comparing TBB and SBB. In the first, we focus on the Asteroids Atari environment. In the second, we train our policies on CountRecall for longer. In both experiments, TBB outperforms SBB.



Hyperparameter	Value
Hidden size	512
Recurrent size	512
Episodes per epoch	1
Updates per epoch	5
Evaluation interval	100
Batch size	16,000
CNN filter sizes	8, 4, 3
CNN filter channels	32, 64, 64
Post-LRU MLP layers	2

Figure 1: Comparison of TBB and SBB on the Atari Asteroids task for the LRU model. Unknown velocity and flickering make this task partially obseravable [Hausknecht and Stone, 2015]. We plot the mean and 95% bootstrapped confidence interval of the best evaluation epoch over three random seeds. We use $L = 80$ following Kapturowski et al. [2019]. The dotted red line denotes the results from Hausknecht and Stone [2015], indicating consistentcy with prior work. We use the same recurrent and hidden size as Hausknecht and Stone [2015].

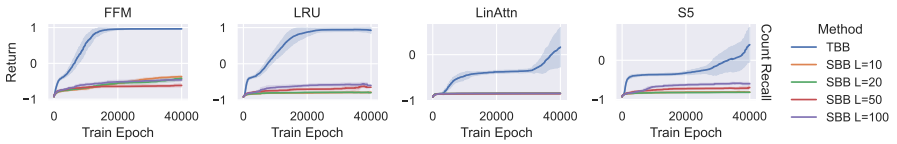


Figure 2: We reran the CountRecall experiments from Figure 6 for longer. We show that the FFM and LRU models converge to the optimal return with TBB, while the SBB models do not.

References

M. Hausknecht and P. Stone. Deep Recurrent Q-Learning for Partially Observable MDPs. In *2015 AAAI Fall Symposium Series*, Sept. 2015. URL `https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11673`.

S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney. Recurrent Experience Replay in Distributed Reinforcement Learning. In *International conference on learning representations*, 2019.