

Credit Assignment with Bayesian Reward Estimation

Constantin A. Rothkopf¹, Dana H. Ballard²

¹ University of Rochester, ²University of Texas at Austin

Evidence suggests that neurotransmitter dopamine is involved in the process of learning visuo-motor behaviors¹. Moreover, reinforcement learning (RL) algorithms have been formulated that characterize well the neuronal signals of dopaminergic neurons in response to the occurrences of stimuli associated with rewards and the delivery of the rewards across learning². Such algorithms have been successful in modeling those responses in cases where only a single variable describes the current state of the world. But most of the time, multiple goals have to be pursued simultaneously. This problem has proven to be intractable for even small numbers of goals because of the exponential growth of the state space. One solution to this problem is to realize a single RL algorithm that uses a high dimensional state space with compositions of RL algorithms that utilize lower dimensional state spaces. Such an approach allows composite visuo-motor behaviors to be synthesized from simpler such behaviors. However taking such an approach introduces problems of its own. First, an action selection mechanism has to decide which action to choose, given that the different algorithms may suggest different actions at the same time. Secondly, given the total observed reward, the organism has to learn what fraction of this belongs to each algorithm. Sprague³ showed in simulation that individual behaviors can be learned in combination by reinforcement learning. However that simulation assumed that the rewards associated with the individual behaviors were known. In practice this is an unreasonable assumption for biological systems where only the total reward for the composite behavior is likely to be available. One would like to learn the contributions from the different behaviors from the obtained total reward. This is a long-standing credit assignment problem in learning. Chang et al.⁴ showed that an estimate for individual rewards could be obtained if the total reward was assigned to each behavior and the variations in that reward were assumed to be noise. This model made sense in their setting, which had the individual behaviors embedded in different agents, but had problems in that the resultant reward estimates could have a constant bias and be suboptimal.

We showed that the credit assignment problem has a solution when all the behaviors are embedded in the same agent⁵. This allows us to model reward not as a globally broadcasted number⁴, but as a consumable entity. The difference is that in the latter case, the individual reward estimates must add up to the total reward estimate whereas in the former case any reward not assigned to an individual behavior is assumed to be entirely noise. In our algorithm, each behavior only needs to know which subset of other behaviors is simultaneously active and their reward estimates. It can then keep a running estimate of its share as its current estimate adjusted by the total instantaneous reward minus the estimates of the concurrent behaviors. Simulations using a standard multiple Predator-Prey problem showed that when the order that the behaviors update is chosen randomly, the estimated reward for each behavior converges to its true value. We have extended this result to show that: 1) It can be formulated as a temporal difference (TD) algorithm 2) Even when the algorithm chooses a compromise (suboptimal) action, individual rewards are learned correctly and 3) Standard Bayesian cue combination can be used in this setting to weight estimates in the TD formulation according to their inverse variance. Simulations using this new weighting exhibit very fast convergence.

References

- [1] W. Schultz, *J Neurophysiol* 80: 1-27, 1998
- [2] W. Schultz, P. Dayan and P. Read Montague, *Science*, 275 1998
- [3] N. Sprague and D. Ballard, *NIPS*, 2003
- [4] Y. H. Chang, T. Ho and L. P. Kaelbling, *NIPS*, 2003
- [5] C. Rothkopf and D. Ballard, *COSYNE*, 2006