Delay Discounting and Reinforcement Learning

Third year paper proposal

Benjamin Lira

How does hyperbolic discounting relate to reinforcement learning (in particular with delayed rewards)? There is a steep reduction in the value of a reward the further it is in the future. The shape of this discounting has been characterized as hyperbolic. Similarly, there is a steep decline in learning (e.g., lever pressing rate) when a delay is introduced between behavior and reward. Are these two processes the same?

I argue that both reinforcement learning and hyperbolic discounting arise from the same processing of uncertainty inherent in time lags. In the case of delay discounting, there is a lag between decision time and reward claiming time. Hyperbolic discounting arises from the uncertainty of whether or not I will be able to receive the reward (i.e., there is a hazard rate that controls how likely it is that something happens so that I cannot cash the 100 USD?). Sozou@sozou1998 argues that hyperbolic decay arises from integrating over an exponential prior on the hazard rate, given that it is uncertain (the particular shape of the prior is less important, and similar curves will be produced with other priors).

In the case of reinforcement learning, there is a lag between behavior and the reward that needs to be associated with that behavior. Reinforcement delays make learning difficult because there is uncertainty about which of the actions that occurred prior to the reward caused the reward to come about (i.e., the credit assignment problem in AI reinforcement learning). When trying to associate a behavior to a delayed reward, the larger the amount of time between action and reward, the greater the number of potential actions (causes) that might have caused the reward. In much the same way, there is also likely to be uncertainty about the rate with which these extra causes enter per unit time.

I believe the delay discounting literature has underweighted the importance of learning, while the learning literature has underweighted the importance of valuation. Traditional intertemporal choice tasks (e.g., would you take 120 dollars in a month, or 100 dollars today) tend to ignore the learning component. In the real world, when I choose to remain on the couch rather than exercising, not only do I have a valuation problem (I discount the future benefits); I also have a learning problem (I have difficulty learning to associate the activity of exercising to whatever positive delayed consequences it brings me). Conversely, in the reinforcement learning case, not only is behavior muted due to difficulties learning

from delayed rewards (i.e., associating the behavior with its consequence), but the value of the future reward is lower.

In artificial intelligence, these processes relate to the two components of reinforcement learning algorithms in computers: value learning and policy learning. Successful RL systems require both a mechanism to learn the value of states (how much discounted future reward can I expect to receive given that I am in state X), and a policy (what actions should I take when I am in this state). In a sense, we humans also need both components: a valuation process to map external to subjective value, and a policy process to learn what behavior in what situations should be preferred when trying to maximize reward.

I hope to write a perspective style paper to make these three points. I would hope that this would contribute to our multidisciplinary understanding of these processes, which could hopefully lead to new methods and insights on the different fields. From a practical perspective, I hope that incorporating the learning component of real-world policy relevant instances of self-control might lead to strategies to bridge this gap, and hopefully result in positive interventions or messaging.

12 34 56 78 910 sutton1988?sutton1998?

Related Readings Ainslie, G. W. (1974). Impulse control in pigeons. Journal of the Experimental Analysis of Behavior, 21(3), 485–489. https://doi.org/10.1901/jeab.1974.21-485 Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From Covariation to Causation: A Test of the Assumption of Causal Power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(6), 1119–1140. https://doi.org/10.1037/0278-7393.29.6.1119 Farmer, J. D., & Geanakoplos, J. (2009). Hyperbolic Discounting is Rational: Valuing the Far Future with Uncertain Discount Rates. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1448811 Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G., & Larochelle, H. (2019). Hyperbolic Discounting and Learning over Multiple Horizons (arXiv:1902.06865). arXiv. http://arxiv.org/abs/1902.06865 Green, L., & Myerson, J. (1996). Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. American Zoologist, 36(4), 496–505. https://doi.org/10.1093/icb/36.4.496 Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. Journal of Experimental Psychology: General, 139(4), 756–771. https://doi.org/10.1037/a0020976 Greville, W. J., & Buehner, M. J. (2012). Assessing Evidence for a Common Function of Delay in Causal Learning and Reward Discounting. Frontiers in Psychology, 3. https://doi.org/10.3389/fpsyg.2012.00460 Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting*. The Quarterly Journal of Economics, 112(2), 443–478. https://doi.org/10.1162/003355397555253 McNamara, J., & Houston, A. (1980). The application of statistical decision theory to animal behaviour. Journal of Theoretical Biology, 85(4), 673–690. https://doi.org/10.1016/0022-5193(80)90265-9 Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society B: Biological Sciences, 265(1409), 2015–2020. https://doi.org/10.1098/rspb.1998.0534 Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3, 9–44. Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction.

MIT Press.

1. Ainslie, G. W. (1974). Impulse control in pigeons. Journal of the Experimental Analysis of Behavior, 21(3), 485–489. https://doi.org/10.1901/jeab.1974.21-485
2. Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From Covariation to Causation: A Test of the Assumption of Causal Power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(6), 1119–1140. https://doi.org/10.1037/0278-7393.29.6.1119
3. Farmer, J. D., & Geanakoplos, J. (2009). Hyperbolic Discounting is Rational: Valuing the Far Future with Uncertain Discount Rates. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1448811
4. Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G., & Larochelle, H. (2019). Hyperbolic Discounting and Learning over Multiple Horizons (No. arXiv:1902.06865). arXiv. https://arxiv.org/abs/1902.06865
5. Green, L., & Myerson, J. (1996). Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. American Zoologist, 36(4), 496–505. https://doi.org/10.1093/icb/36.4.496
6. Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. Journal of Experimental Psychology: General, 139(4), 756–771. https://doi.org/10.1037/a0020976
7. Greville, W. J., & Buehner, M. J. (2012). Assessing Evidence for a Common Function of Delay in Causal Learning and Reward Discounting. Frontiers in Psychology, 3. https://doi.org/10.3389/fpsyg.2012.00460
8. Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting*. The Quarterly Journal of Economics, 112(2), 443–478. https://doi.org/10.1162/003355397555253
9. McNamara, J., & Houston, A. (1980). The application of statistical decision theory to animal behaviour. Journal of Theoretical Biology, 85(4), 673–690. https://doi.org/10.1016/0022-5193(80)90265-9
10. Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society B: Biological Sciences, 265(1409), 2015–2020. https://doi.org/10.1098/rspb.1998.0534