Generalised Discount Functions applied to a Monte-Carlo AIμ Implementation

Sean Lamont[1], John Aslanides[1], Jan Leike[2] and Marcus Hutter[1]

[1]Research School of Computer Science, Australian National University. [2]Google Deepmind, London; Future of Humanity Institute, Oxford

Objectives

- Experimentally reproduce behaviour characteristic of agents using generalised discount functions.
- Modify the GRL demonstration platform AIXIjs to facilitate this, and demonstrate these results in a simple MDP.

Motivation

- General Reinforcement Learning (GRL) : Domain independent Reinforcement Learning agents
- Many theoretical results proven for GRL, but few examples demonstraring these in a concrete setting
- **Our Goal**: Use the platform AIXIjs to experimentally verify theoretical results regarding general discount functions

Background

Samuelson's [3] standard model of discounted utility:

$$V_k := \sum_{t=k}^{\infty} \gamma_{t-k} r_t$$

Where γ is the discount function, and r_t is the reward at time t.
Hutter and Lattimore [2] extend this to include discount functions which can change over time.
This generalised model allows for policies which:

- Are **time inconsistent**, where actions may not always align with previous plans.
- Cause future rewards to become relatively more desirable as time progresses (from a **growing effective horizon**)

The discount functions we investigate are:
Hyperbolic Discounting: $\gamma_t^k = \frac{1}{(1+\kappa(t-k))^\beta}$
Thought to model human discounting,and explains many irrational (time inconsistent) behaviours.
Power Discounting: $\gamma_t^k = t^{-\beta}$
Time consistent, causes a growing effective horizon.

AIXIjs

- Online, JavaScript based platform showcasing theoretical results from GRL in Gridworld environments. [1]
- **Open Source**, allows researchers to add and modify demos as necessary
- Adapted to include arbitrary discount functions, and to include a simple MDP assessing agent far-sightedness.
- We derive the number of time inconsistent actions by recording the MCTS plan and comparing this with future actions.

AIXIjs Source Code/ Web Page

- Source: https://github.com/aslanides/aixijs
- Web Page: http://aslanides.io/aixijs/
- Also: http://www.hutter1.net/aixijs/

Environment

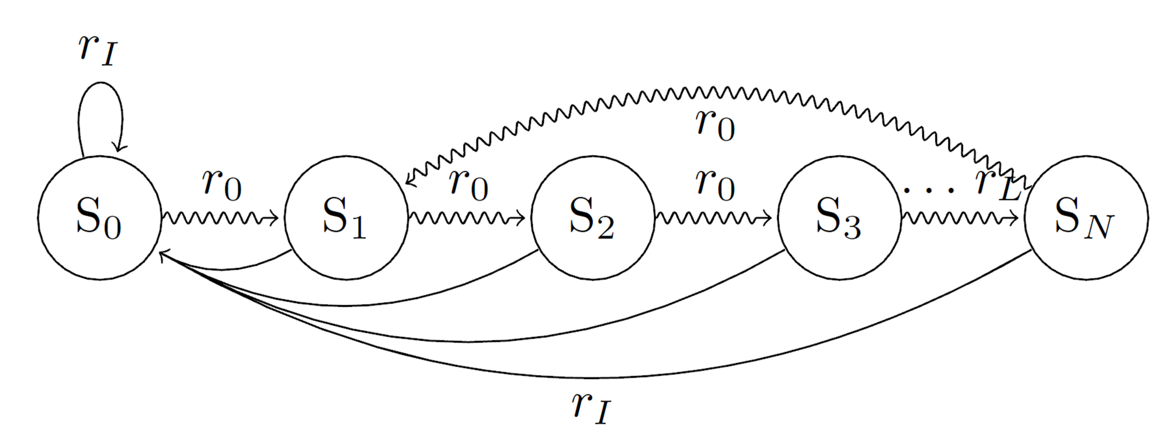

Figure 1: MDP Used for Discounting Experiments

This environment gives the agent 2 actions:

- a_1 (Straight Line): Return a small reward r_I every time a_1 is taken
- a_2 (Squiggly Line): Return very large reward $r_L > N r_I$ only if the agent follows a_2 for N consecutive steps. Return 0 reward r_0 otherwise.

If the agent is far-sighted enough, it will ignore the low instant reward and plan ahead to reach the very large reward every N time steps.

Key Results

- We observed the impact of a growing effective horizon when using power discounting, which resulted in a change in policy over time
- We observed time inconsistent agent behaviour (procrastination) under hyperbolic discounting

AIμ with ρUCT Monte-Carlo

- We use the informed agent AIμ for our experiments.
- Knows the true environment dynamics a priori, eliminating any uncertainty in the agents model
- In combination with a deterministic MDP, ensures any observed change in behaviour is from the discount function, as opposed to any stochasticity in the environment/model.
- ρUCT [4] Monte-Carlo Tree Search used to approximate expectimax
- UCT would suffice, though AIXIjs has ρUCT as default

Results: Reward Plot

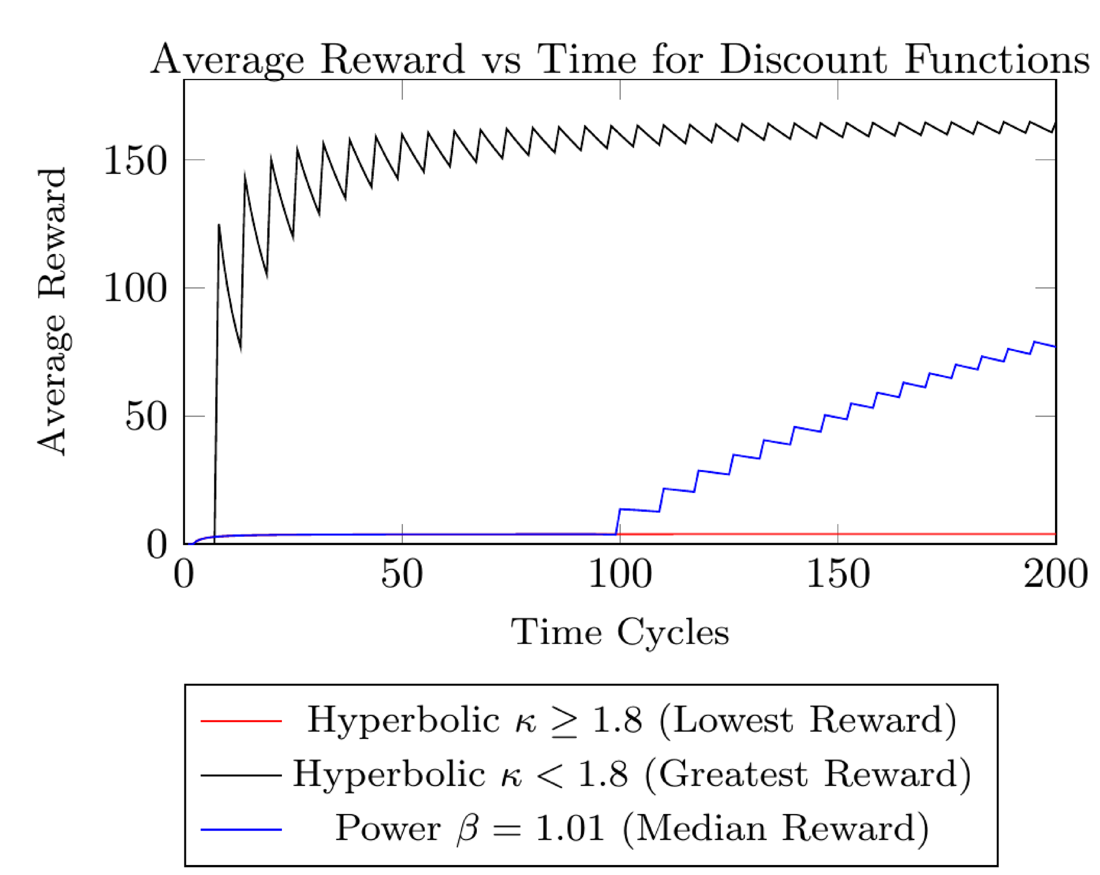

Figure 2: Reward Plot for Discounting Experiments

Summary of Results

- The agent plan enumeration gave time consistent results for all trials of power discounting, and for $\kappa \neq 1.8$ with hyperbolic discounting
- For $\kappa = 1.8$, the agent plan enumeration showed the agent was consistently planning to take the far sighted policy 1 step in the future, yet remained on the instant reward state.
- By observing the graph, we see that the agent using power discounting changed to a far-sighted policy around step 100. This is directly caused by the growing effective horizon.

Contact Details

Please feel free to contact me about any questions regarding this project.

- Email: sean.a.lamont@outlook.com

References

[1] J Aslanides. AIXIjs: A software demo for general reinforcement learning, Australian National University, 2016.
[2] T. Lattimore and M. Hutter. General time consistent discounting. *Theoretical Computer Science, 519:140-154*, 2014.
[3] P. Samuelson. A note on measurement of utility. *The Review of Economic Studies, 4(2) : 155-161*, 1937.
[4] J. Veness, M. Hutter, W. Uther, D. Silver, and K. S. Ng. A Monte-Carlo AIXI Approximation. *Journal of Artificial Intelligence Research 40: 95-142*, 2011

Australian National University

NATURAM PRIMUM COGNOSCERE RERUM