B9120-001 Dynamic Programming Winter 2020
Prof. Daniel Russo

Homework Assignment 5: Due Monday March 2

Reading

Read through `this paper` to review the proof of the Gittins Index Theorem. Ask yourself the following questions to enforce your comprehension:

- Where did we leverage the independence structure of bandit processes?
- Lemma 2.1 might make you think the Gittins Index is somehow myopic. Why is the process of "reducing" the bandit i^* capturing the problem dynamics, discount factor, etc?
- Why did this proof require a semi-Markov formulation?
- How would this proof break if the problem had a finite horizon?

For those of you who are interested in reinforcement learning, you may want to read this landmark paper which introduced the "Options framework" for Hierarchical RL. There are natural connections to the ideas we've just studied (Semi-Markov processes, 'reducing' a bandit arm etc.). http://www-anw.cs.umass.edu/ barto/courses/cs687/Sutton-Precup-Singh-AIJ99.pdf

Computation of The Gittins Index

Solve problem 1.8. of Bertsekas Volume II. You may assume that the state space of each bandit process is finite.

Coin Tossing

Suppose we are faced with 10 biased coins. Each coins bias (probability of heads) is either 2/3 or 1/3. Our prior probabilities over the possible biases of each coin are independent and uniform. At each time, we choose one coin to flip and receive \$1 if the coin lands heads and nothing if it lands tails. Each coin can be flipped at most 100 times. Our objective is to maximize expected discounted revenue, with a discount factor of 0.99. Describe an optimal strategy that selects the next coin to flip given observations to date.

Remark: The restriction that each coin can be flipped at most 100 times ensures the state space is finite.