Actor-Critic Based Solutions for Scheduling Astronomical Observations

Hasti Abbasi Kenarsari, Department of Computer Science

Dr. John Korah (Faculty Mentor)

Abstract

- Efficiently scheduling the observation of celestial bodies is a critical optimization problem
- Variant of the Traveling Salesman Problem in combinatorial optimization (Astro-TSP)
  - Specific observation windows
  - Variable observation lengths
  - Prioritization of distinct observations
  - Intricacies of telescope movement
- Reinforcement learning can be utilized for combinatorial optimization problems with a large solution space
- Actor-Only Methods
  - Utilize parametrized policies to estimate the gradient of performance according to actor parameters
  - New gradients are estimated as the policy changes to adhere to performance improvement
- Critic-Only Methods
  - Work towards an approximate solution to the Bellman Equation by solely utilizing value function optimization
- The Actor-Critic framework combines these two categories of reinforcement learning, ultimately resulting in variance reduction, faster convergence, and versatility across diverse action spaces [1]

Introduction

- Astronomers at the Palomar Observatory create celestial observation schedules by hand
  - Inefficient allocation of labor and resources
  - Exigency for a solution that considers the countless variables that comprise each observation

- Observation Factors
  - Characteristics of the target celestial object
  - Enforced methodologies of the observatory
  - Weather conditions at the time of the observation

- Constraints
  - Time-dependent observation lengths are a product of differing signal-to-noise ratios, which consider background noise and atmospheric interference [2]
  - Each celestial body must be observed for a variable observation time depending on the time of night
  - Each body has a different priority, in recognition of specific targets that are more valuable than others

- Reinforcement Learning is an emerging methodology in response to this problem given the dynamic characteristic of the environment
  - Generalization of agent strategies
  - Long-term reward maximization
  - Analyzation of initial state sequences

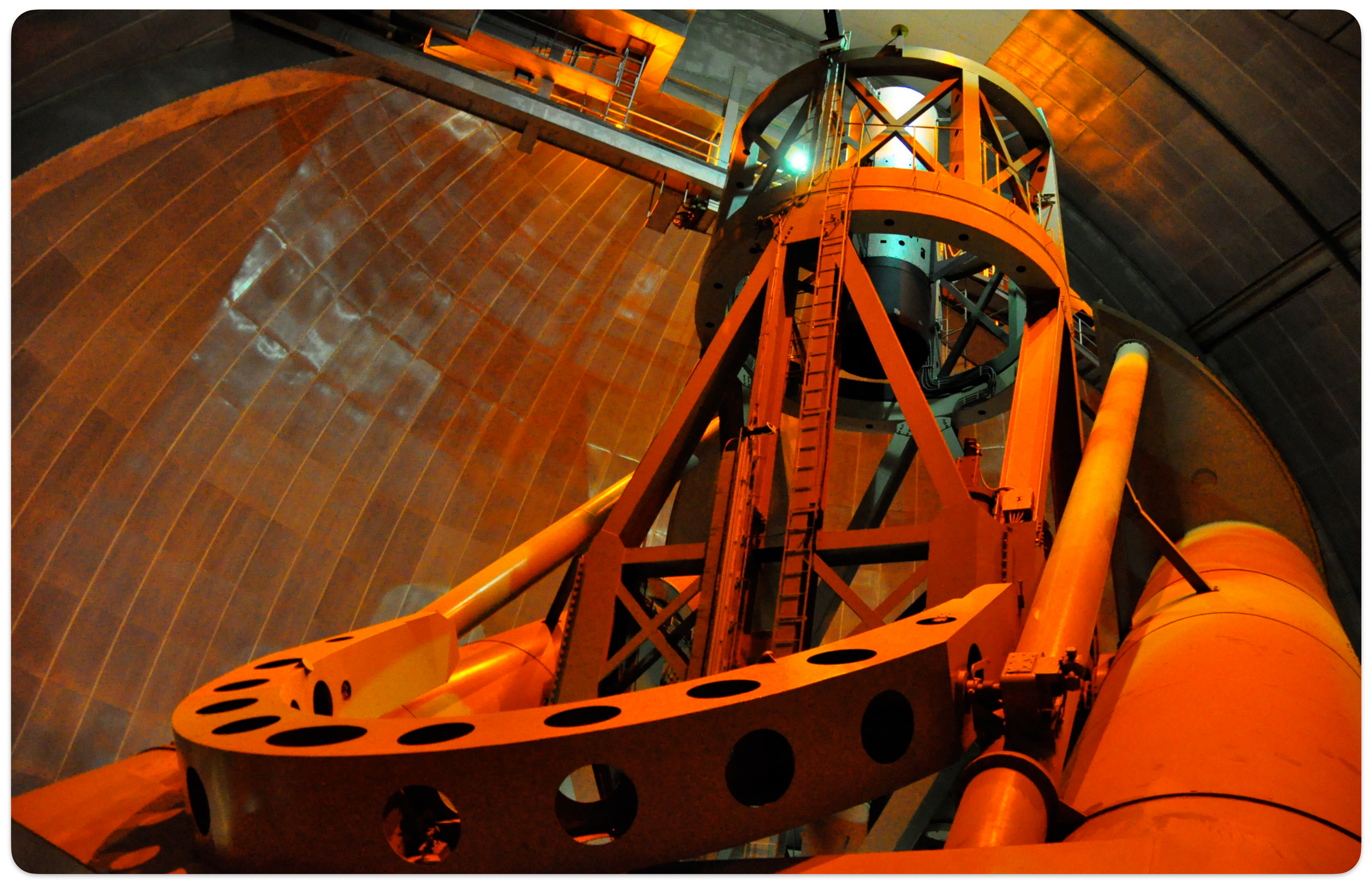

Fig 1: An internal depiction of the 200-inch (5.1-meter) Hale Telescope located within the Palomar Observatory at the NASA Jet Propulsion Laboratory. [5]

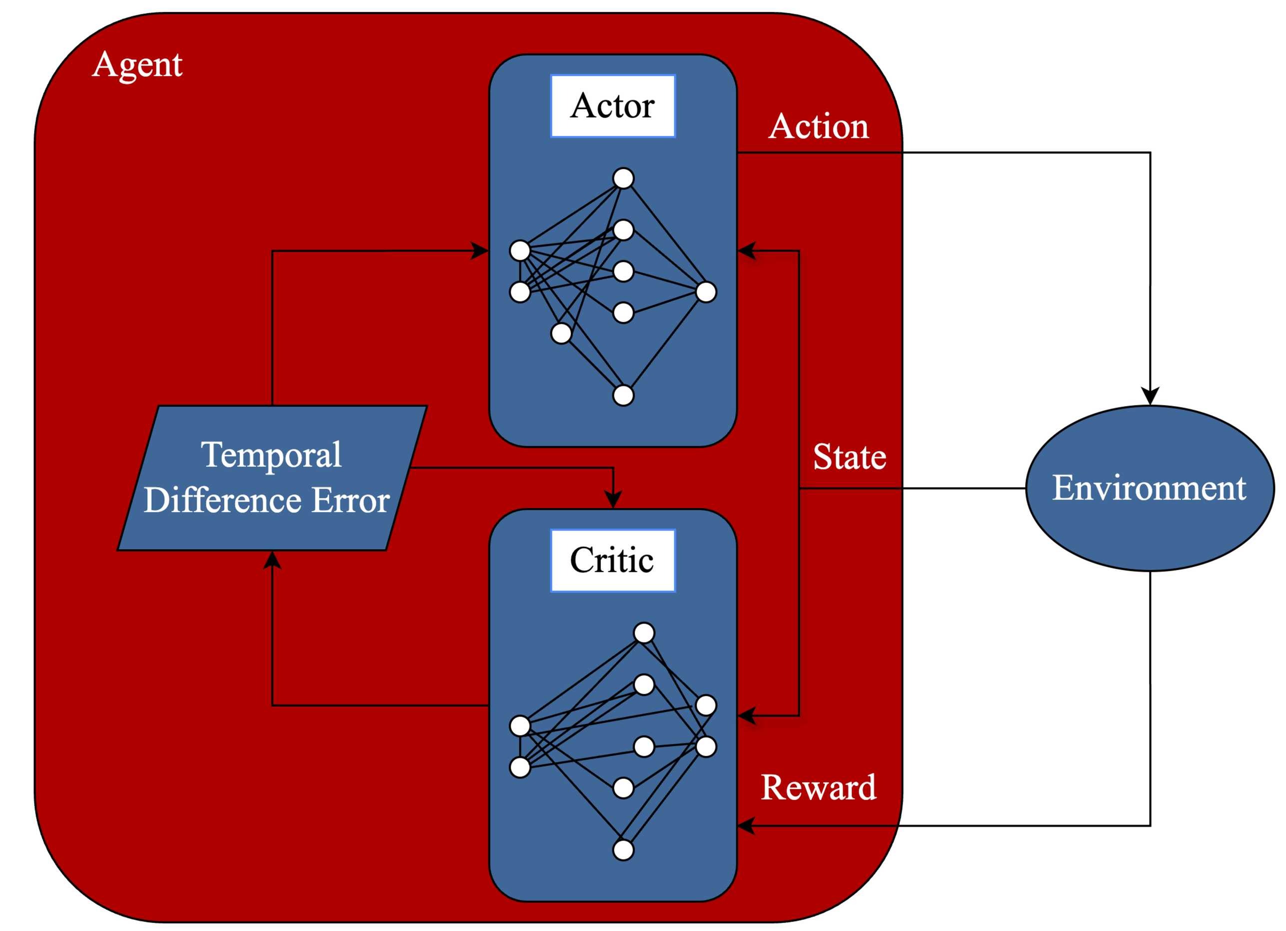

Fig 2: A diagram demonstrating the different components of the Actor-Critic Algorithm. [3]

Methodology

- Agent: the scheduling algorithm within the decision-making model that selects the next observation for each designated time step

- Environment: the telescope in the Palomar Observatory, withholding the available list of observations and serving as the external system that the agent interacts with

- State (s): the depiction of the current scheduling scenario, represented as a feature vector containing the telescope's current position, current time, list of remaining observations, and remaining time in the observation window

- Action (a): selecting the next observation from the available celestial objects

- Reward (r_t): feedback accumulated after the series of actions performed by the agent, guiding the agent in the process of scheduling efficiently in the long-term
  - Encourage observations that are closer (minimize movement of the telescope) & have low airmass (better visibility)
  - Penalize scheduling observation that are far part & have passed the available time window

- Policy ($\pi_\theta(a_i|s_i)$): decision-making guide that serves as the set of rules for the agent
  - Outputs a probability distribution given the available observations (Actor Network)
  - Predicts the expected reward of the schedule (Critic Network)

- In each time step, the actor selects an observation & the critic provides a value prediction (feedback)
- The advantage (difference between the value prediction and actual reward) is used to update the networks
  - Actor-network adjusts its policy to improve the selection of observations
  - Critic-network fluctuates its value estimates

Implementation

- The policy gradient equation adjusts to increase the probability of actions that yield high rewards & decrease the probability of actions that yield low rewards

$$\underbrace{\nabla_\theta J(\theta)}_{\text{Expected Return}} \approx \frac{1}{N}\Sigma_{i=0}^{N} \nabla_\theta \underbrace{log\pi_\theta(a_i|s_i)}_{\text{Policy}} \bullet \underbrace{A(s_i, a_i)}_{\text{Advantage}}$$

- The value function update equation evaluates the actions taken by the actor, allowing for a balance between exploration and exploitation

$$\underbrace{\nabla_w J(w)}_{\text{Loss Gradient}} \approx \frac{1}{N}\Sigma_{i=1}^{N} \nabla_w \underbrace{(V_w(s_i) - Q_w(s_i, a_i))^2}_{\text{Value Estimates}}$$

- After the initialization of the policy & function parameters, the agent takes actions according to the pre-defined policy. The advantage is computed using the equation

$$\underbrace{A^\pi(s,a)}_{\text{Advantage}} = \underbrace{Q^\pi(s,a)}_{\text{Action-Value}} - \underbrace{V^\pi(s)}_{\text{State-Value}}$$

- The policy and the value are updated simultaneously
  - The policy gradient is utilized to update the actor's parameters
  - The policy gradient increases the probability of actions that have higher rewards
  - The value-based method is utilized to update the critic's parameters by minimizing the temporal difference error, defined using the equation

$$\underbrace{\delta_t}_{\text{Temporal Difference Error}} = \underbrace{r_t}_{\text{Reward}} + \overbrace{\gamma}^{\text{Discount Factor}} \underbrace{V(s_{t+1}) - V(s_t)}_{\text{State Values}}$$

Conclusion

- The schedule is optimized by
  - Maximizing the quantity of unique observations
  - Minimizing the time spent observing the objects
  - Reducing the time spent waiting for available observations
  - Maximizing the total priority of the objects

- The generalization of the Astro-TSP can be applied to similar astrophysics problems, concurrently advancing the original Traveling Salesman Problem

- The scope of the proposed solution can be applied to optimization problems that require strong heuristic solutions and have similar constraints

- The Actor-Critic algorithm is expected to outperform previous optimization techniques such as Look-Ahead Greedy, Q-Learning, and simple ordering heuristics

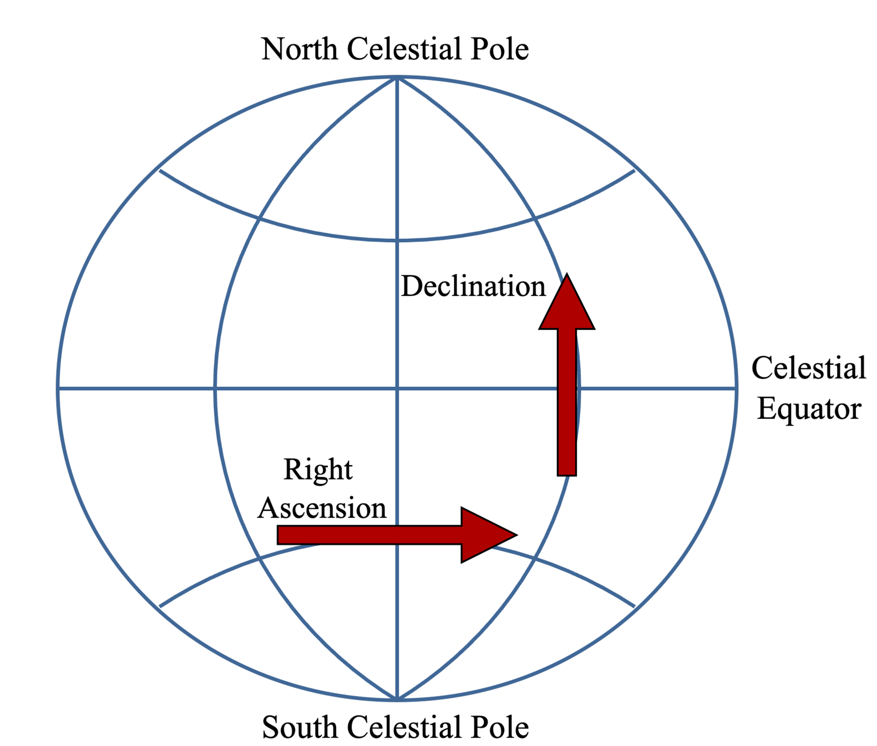

Fig 3: A visual overview of the celestial coordinate system utilized to pinpoint observations. [4]

References & Acknowledgements

[1] V. R. Konda and J. N. Tsitsiklis, "On Actor-Critic Algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, 2003. doi: 10.1137/S0363012901385691.

[2] M. Humphries, "Astro-TSP: Traveling Salesman Problem Based Solutions for Scheduling Astronomical Based Observations," California State Polytechnic University, Pomona, 2024. http://hdl.handle.net/20.500.12680/f7623m070

[3] H. Giang, T. Hoan, P. Thanh, and I. Koo, "Hybrid NOMA/OMA-Based Dynamic Power Allocation Scheme Using Deep Reinforcement Learning in 5G Networks," *Appl. Sci.*, vol. 10, no. 12, p. 4236, 2020, doi: 10.3390/app10124236.

[4] B. Sun, J. Wang, H. Zhou, H. Liu, E. Wei, and X. Zhou, "Measurement sensitivity analysis and on-orbit calibration of systematic errors for a narrow field-of-view camera," *Opt. Express*, vol. 31, 2023, doi: 10.1364/OE.479984.

[5] Gracey, Bill. "Hale Telescope". October 11, 2009. Flickr. https://flic.kr/p/76JJhW