Sky surveys scheduling using reinforcement learning

Stefan Stefanov

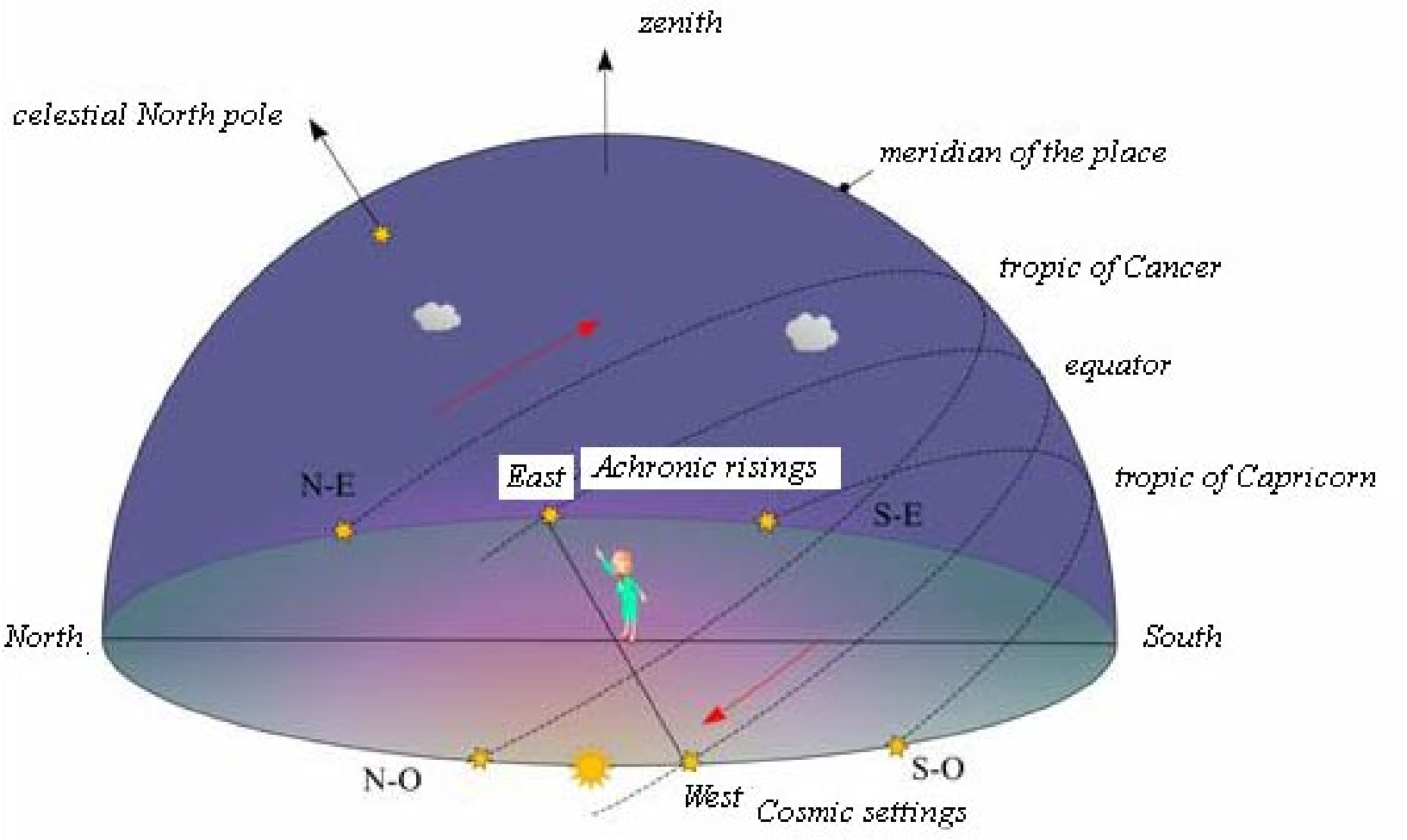
Sofia University

sjonkov@uni-sofa.bg

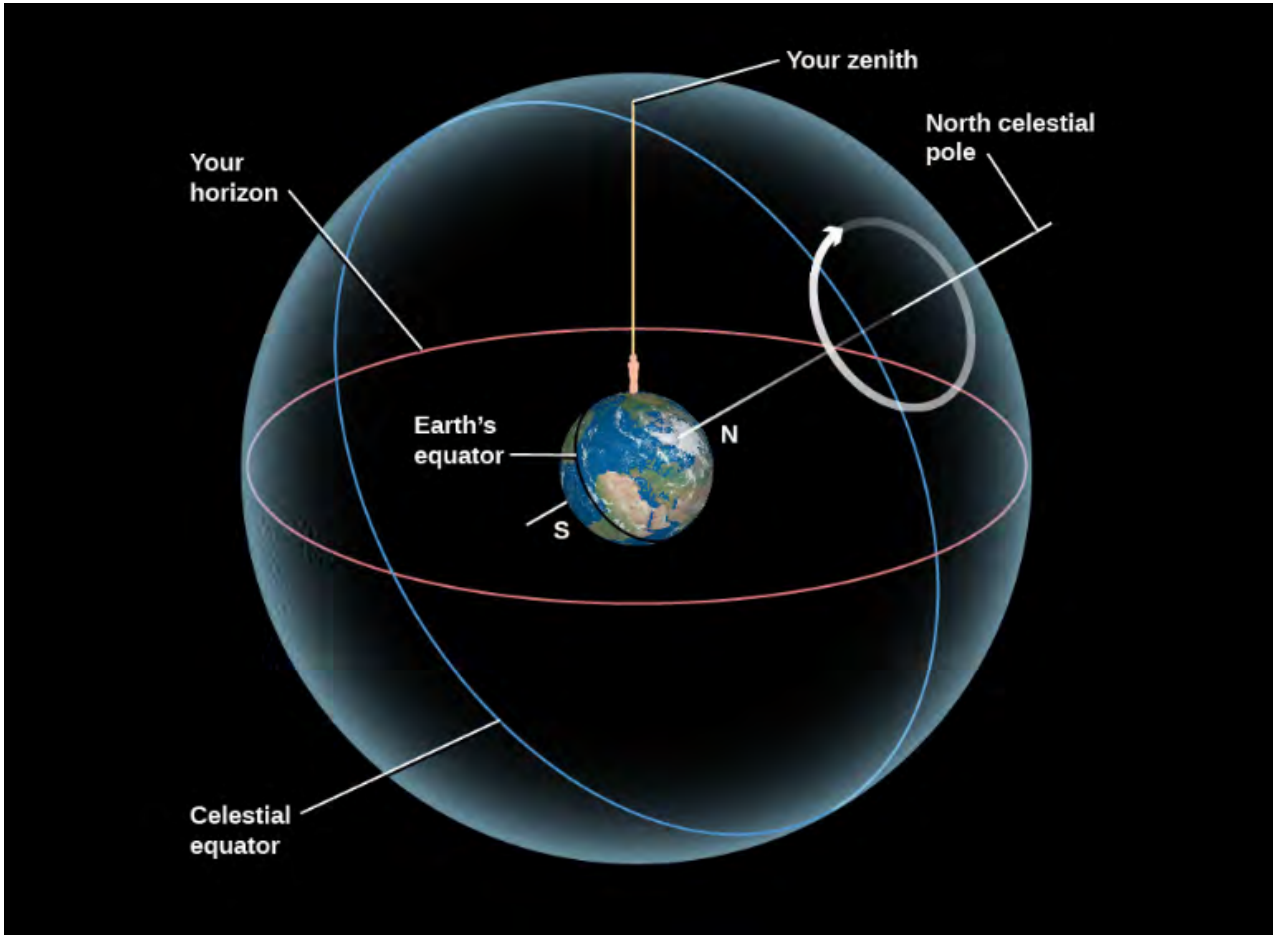
December 12, 2020

Overview

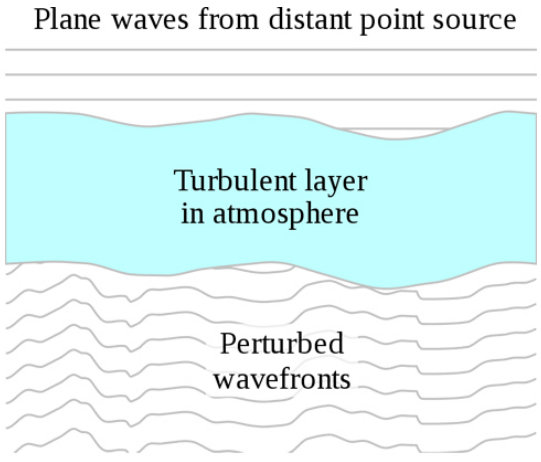
The celestial sphere

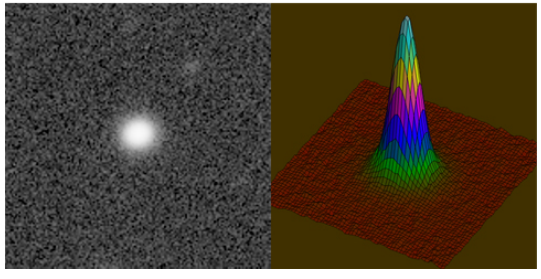


Equatorial coordinates

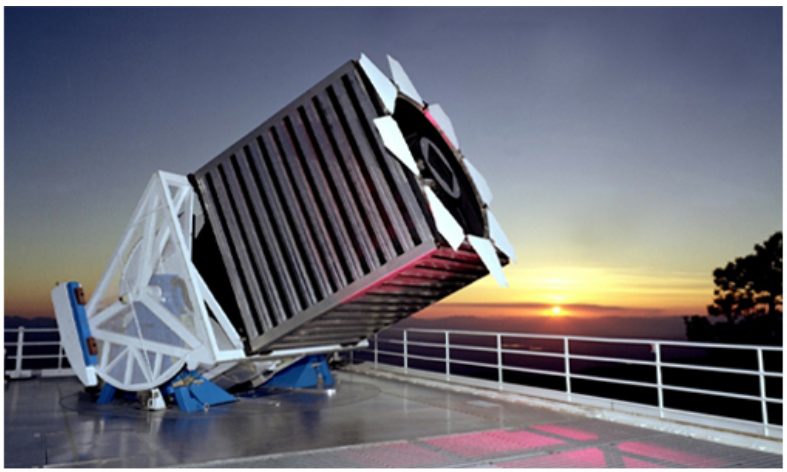


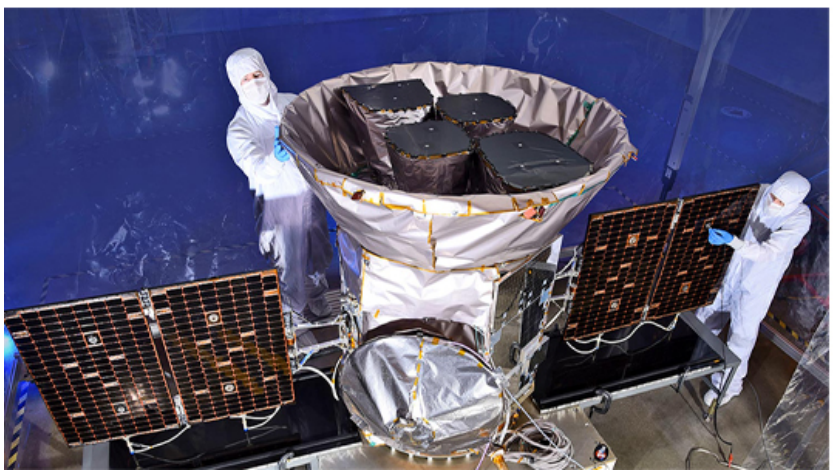
CCD Photometry

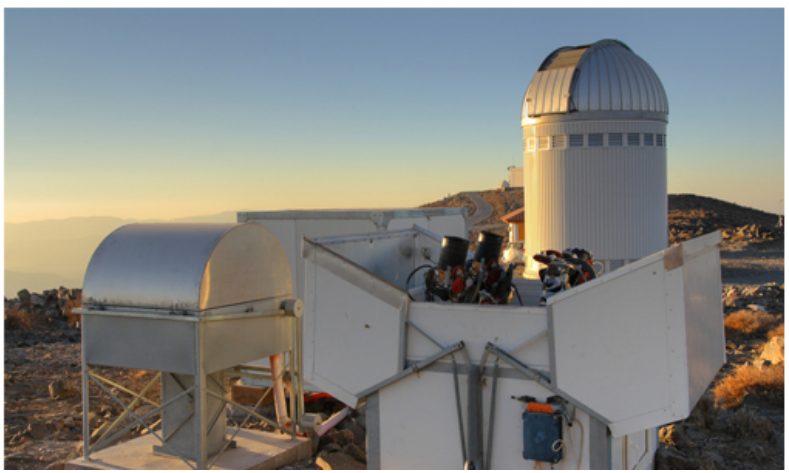




Sky Surveys







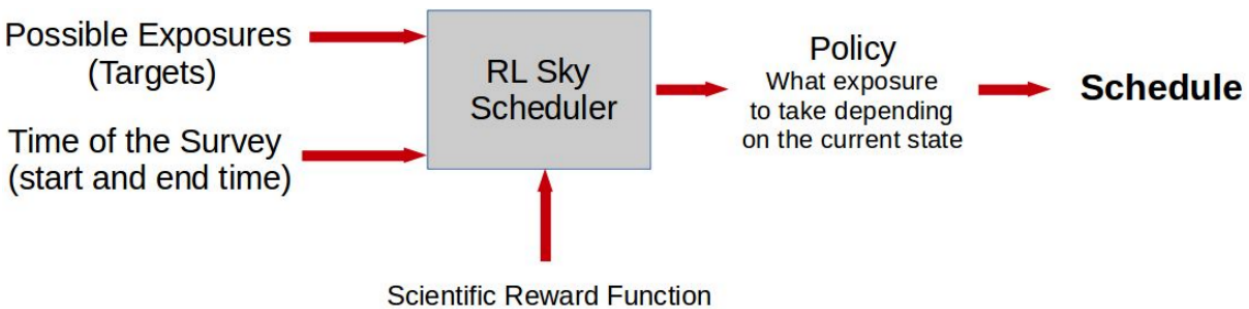
The RL Model

Actions

The agent's actions are the possible fields to take an exposure. The agent takes an action then decides the coordinates of the next field for taking an exposure.

Environment

1. Rewards - The reward is calculated based on the quality of exposures
2. States - The states describe all the possible exposures the agent can take through the course of the survey



Scientific reward function

In order to define the reward the authors introduce two elements:

1. t_{eff} parameter:
   This parameter evaluates the quality of the exposure.
   $t_{eff} = \left(\frac{0.9}{(seeing).(airmass)^{0.6}} \right)^2$
2. V_n - The survey volume:
   This parameter represents the amount of scientific data contained in the exposures taken.
   $V_n \propto (t_{eff})$

The final reward function is the change in the survey volume. Bigger values for t_{eff} produce larger survey volume and positive reward.

$$\boxed{r_n = \Delta V_n(t_{eff}) = V_n - V_{n-1}}$$

State model

The model of the state requires the definition of three elements:

1. Time - Time is defined by using the standard astronomical Julian day (JD). This is a real number, indicating the days passed since 4713 BC. 16:00h today in JD is 2459196.0171
2. State of $\overrightarrow{\mathrm{t_{eff}}}$ array - Every possible exposure in the survey period is going to be treated as a target. The set of all possible targets T conform to the action space of the agent. The accumulated t_{eff} for every target(field) can be represented by a vector:
   $\overrightarrow{\mathrm{t_{eff}}} = [t_{eff_0}, t_{eff_1}, ..., t_{eff_{m-1}}]'_{m\times 1}$
3. Current target - Each position index j in the vector corresponds to a target and the value t_{eff_j} is the accumulated t_{eff} for the j target.

State space

Then the state $s \in S$ is given by $(JD, S_{t_{eff}}, target_j)$, where $S_{t_{eff}}$ is the state of the vector $\overrightarrow{\mathrm{t_{eff}}}$.
The JD and t_{eff} parameters can have continuous values. In order to use tabular methods these parameters have to be quantized. The authors decide to introduce quantization levels of the two parameters.

Parameters	**Values**
Time levels	36
t_{eff} levels	5
Targets	3
$\overrightarrow{\mathrm{t_{eff}}}$ states	125
Number of states	13500

Table: State parameters with a simple quantization

Learning algorithms

In this study the authors use two learning algorithms:

- One-step greedy algorithm
- Q-learning algorithm

Parameter	**Values**
Learning rate	$\alpha = 0.4$
Discount factor	$\gamma = 0.99$
Exploration probability	$\epsilon = 0.005$

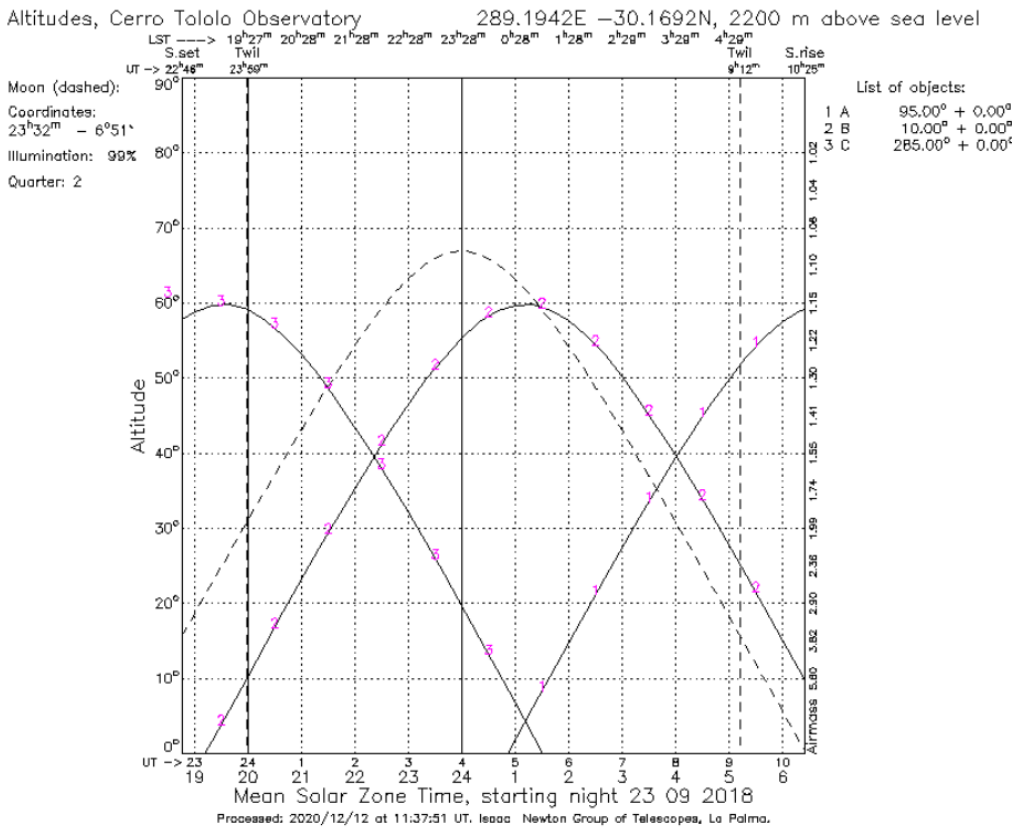
Table: Hyperparameters of the agporithm used for test

The authors made a comparison between the two algorithms and the python package Astroplan.

Three target test

The authors designed a simple test with 3 observational fields for a telescope at Cerro Tololo peak. Each time the agent chooses a field, the telescope does 4 exposures of 120 seconds. The fields A B and C have the following coordinates:

- Field A: $RA = 95^o$ and $DEC = 0^o$
- Field B: $RA = 10^o$ and $DEC = 0^o$
- Field C: $RA = 285^o$ and $DEC = 0^o$



Results

A trajectory made by the authors would be the following: 21 exposures of C, 34 exposures of B and 0 exposures of A, leaving some remaining time.

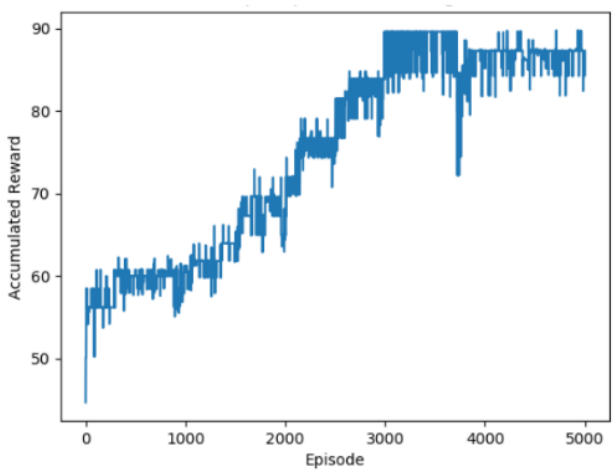
- Astroplan - The schedule obtained is 8 exposures for C, 33 for B, and 2 for A. The total reward accumulated is 73.6.
- One-step greedy - obtained the optimal solution, and in the remaining time did 2 aditional exposures of A, accumulating some negative reward. Total = 96.73
- Q learning after 20000 episodes - 21 exp. of C, 35 for B, and 1 for A. This schedule has a total reward of 97.25.

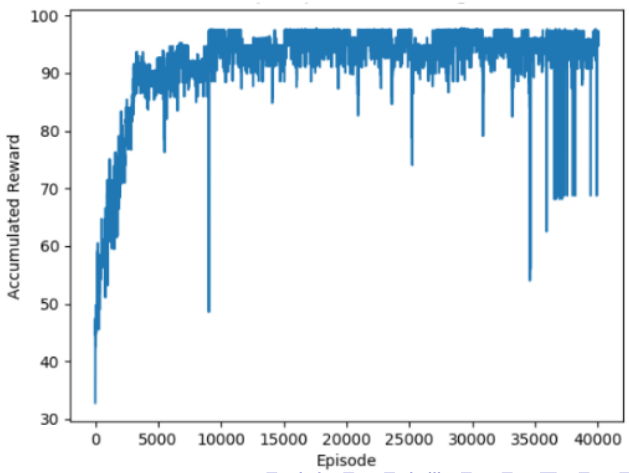
Q Learning performance

The alorithm performance is presented in the table below:

Description	Results 1	Results 2	Results 3	Results 4
Number of episodes	1500	5000	20000	40000
Number of actions	67	59	56	56
Number of valid actions	43	51	55	55
Number of invalid actions	24	8	1	1
Accumulated Reward	68.48	87.28	97.25	97.25
Q learning-reward/Greedy-reward	70.79%	90.79%	103%	103%
Algorithm run time in min	64	253	807	1485

The learing curves for 5000 and 400000 episodes are presented below:





Conclutions

Is this approach for the sky survey scheduling problem practical?

- Advantages
- Disadvantages

The End