How to miss data? Reinforcement learning for environments with high observation cost

Mehmet Koseoglu, Ayca Ozcelikkale

Motivation and the Setting:

- There is a cost associated with making accurate observations. As the accuracy increases, the cost of the observation increases.
  - Examples: Medical applications with expensive tests, usage of wireless communication channels with high power
  - A limiting scenario: Measure with a given accuracy (with a given cost) or Miss (with no cost)
- We would like to perform a given task using a small number of observations as possible or with the smallest cost as possible (i.e. with low accuracy measurements.)
- **Can a RL agent automatically determine these uninformative observations and decide to miss them or sample them inaccurately? YES!**

Proposed Approach: In addition to original reward, also reward the agent for inaccurate or missing samples.

$$\bar{r} = f(r; \beta)$$

- r: the old reward
- $\bar{r}$: the new reward
- β: represents the accuracy of the observations
- f(r; β): monotonically increasing function of r and β

Example: Inverted Pendulum – I

POMDP with the following noisy state measurements:

$$\tilde{\theta} = \theta + C_\theta \times \mathcal{U}(-\beta, \beta)$$

$$\tilde{\dot{\theta}} = \dot{\theta} + C_{\dot{\theta}} \times \mathcal{U}(-\beta, \beta)$$

Original state

Scaling

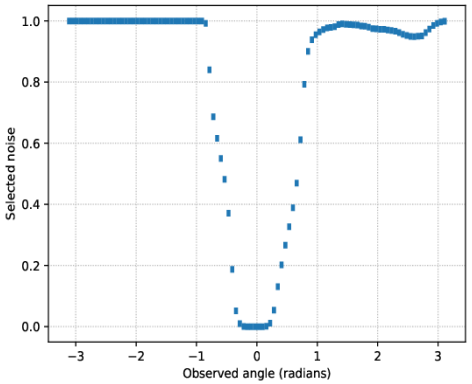
Agent decides on β

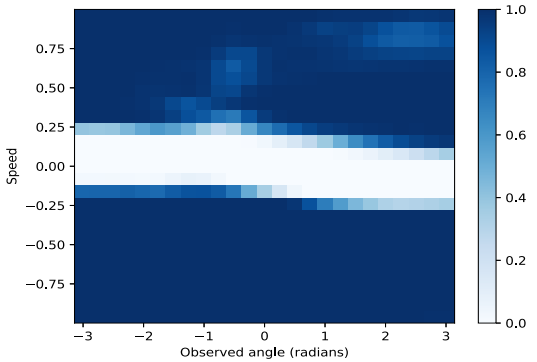
Additive reward shaping is used:

$$\bar{r} = r + g(\beta)$$

Original reward

Measurement accuracy versus angle and speed





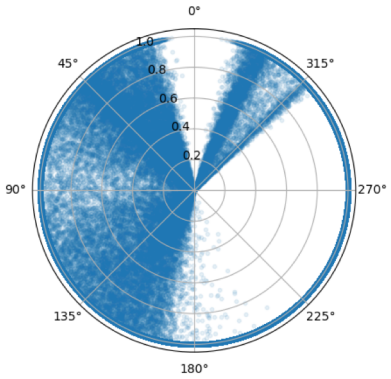
The agent takes accurate samples when the observed position of the pendulum is close to the upright position and hence the speed is low:

- The upright position is a vulnerable position, a badly chosen action based on a noisy measurement may cause the pendulum to drop

Example: Inverted Pendulum – II

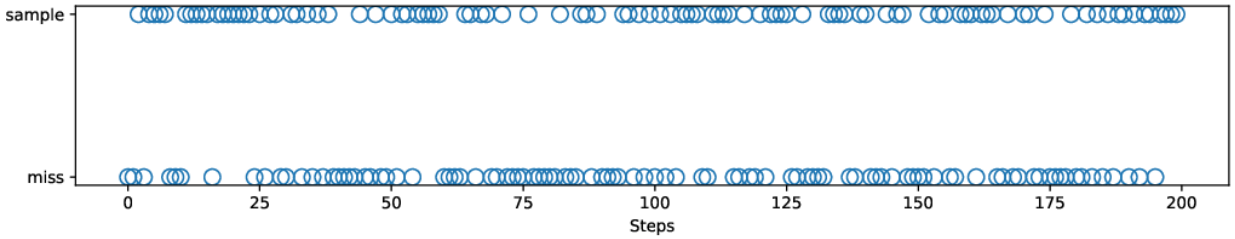
Measurement accuracy versus position

The noise levels of the samples for different positions of the pendulum, where the distance from the origin represents the noise level.



- Agent typically climbs up from LHS taking accurate samples and lets the pendulum fall from RHS.

When the agent is forced to decide between either "sample" or "miss":



- The agent only takes half of the samples (101 out of 200) but is still able to exert correct actions and hold the pendulum in the upright position.
- Comparison with the standard case: As a result of sparse sampling, it takes on average 47 steps to get to the upright position with this RL agent whereas it takes 43 steps with a standard RL agent that has access to all noiseless samples.

Discussion and Conclusions

- **This type of problems with high observation cost arise in various real-life applications:**
  - Expensive but high-accuracy (versus inexpensive but inaccurate) medical tests
  - Physics/chemistry experiments with high labor/material cost
  - Power control in wireless communications
    - Increasing power results in lower noise but consumes more energy.

- **We provided a self-tuning RL agent that learns to successfully adjust the accuracy of the samples:**
  - Proposed approach: Reward shaping which penalizes accurate measurements as well as rewards success in the original task

- **Future work:**
  - Generalizations to other observation accuracy models.
  - Autonomous reward shaping for different environments and accuracy models.
  - Optimized handling of missing and inaccurate samples.