Solving Complex Sparse Reinforcement Learning Environments

Javier Rando Ramírez   
javier.rando@tum.deShreyash Agarwal  
shreyash.agarwal@tum.de

Keywords – reinforcement learning, sparse rewards, Scheduled Auxiliary Control (SAC-X)

# Objective

In this project, we will try to extend the current state-of-the-art systems to solve complex sparse reinforcement learning tasks. This sparse setting is very useful to avoid reward shaping; i.e. the agent overfitting to the rewards found in the environment. The idea behind these scenarios is that the reward is rare and usually found after a complex series of events. Therefore, discovering this reward by random exploration is unfeasible, especially for complex environments.

We will build on top of SAC-X [1], a technique proposed to solve these tasks by creating auxiliary policies that allow the agent to efficiently explore the environment. We will analyze potential shortcomings of this implementation and explore how it can be used for object manipulation, e.g., with a hand. Furthermore, we will try to implement temporal coherency using recurrent neural networks such as LSTM.

# Related Work

## Reward shaping

This was the first and most intuitive idea introduced to solve the problem [2]. The idea is including additional reward features to reward or punish interactions with the environment before the final state. However, this may need experts to come up with the correct additional rewards that will help the agent solve the task. This is not usually easy.

## Curiosity driven methods

In this approach, the agent must explore the environment in order to discover the final reward state. Therefore, curiosity must be modelled as a reward signal. There have been several contributions following this idea. For instance, curiosity can be seen as an intrinsic reward for the agent [3] or we may create self-supervised exploration [4].

## Curriculum learning

The idea is presenting the agent increasingly complex tasks over time until it is able to solve the initial sparse task. One technique used for curriculum learning is GoalGAN which uses a Generative Advesarial Network to create new goals that are harder but still solvable for the agent [5].

## Auxiliary tasks

In this case, we will show the agent auxiliary policies that lead to an efficient exploration of the environment. Riedmiller et al [1] put forward Scheduled Auxiliary Control (SAC-X). Based on the auxiliary tasks presented to the agent, it will explore the environment until the external reward is found.

# Technical Outline

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

1. M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Van de Wiele, V. Mnih, N. Heess, T. Springenberg, “Learning by Playing – Solving Sparse Reward Tasks from Scratch”. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:4344-4353
2. M. J. Mataric, “Reward functions for accelerated learning.”
3. D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction. ”
4. R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, “Planning to explore via self-supervised world models.”
5. C. Florensa, D. Held, X. Geng, and P. Abbeel, “Automatic goal generation for reinforcement learning agents.”