Applying Reward Shaping and Optimal Reward Search to a Novel Grid-World Setting

Ishank: 16D070012, Rahul: 160070003, Gopinath: 160010053, Anirudh: 160010056

October 9, 2019

1 Introduction

While designing an Intelligent agent, the agent designer has certain expectations from the agent's behaviour. These broader design objectives can be modeled as an extrinsic reward function. The qualifier extrinsic is appropriate since these rewards are obtained from the agents environment. For a majority of tasks, this extrinsic reward function is too sparse for the learning of an optimal policy to be tractable in a reasonable number of iterations. To overcome this problem, multiple solutions have been proposed in the literature. Two of these solutions are **Optimal Reward Search** [S. Singh et al., 2009] and **Policy Invariant Reward Shaping** [A. Ng et al, 1999].
The Optimal Reward Search framework performs a search for an optimal intrinsic reward function over the space of all possible reward functions. Here the optimal reward function, as defined in [S. Singh et al., 2009] is the one that maximizes the expected extrinsic reward thereby complying with the objectives of the designer.
On the other hand Reward Shaping is a method in which a shaping function is added to the sparse extrinsic reward function. Shaping is successful if the modified rewards can guide the learning agent to learn the globally optimal policy in fewer iterations.
In [A. Ng et al, 1999], the authors proved that for the optimal policy to remain invariant under reward shaping, a sufficient and necessary condition is that the shaping functions be given by a conservative potential. That is, the additional (shaping) reward for making a transition is given by a difference in the potentials associated with the start and end states. However, as suggested in the paper other potential-like shaping functions are relevant from an engineering point of view. This is likely due to them facilitating faster convergence to near optimal policies.

2 Scope of Work

Through this project we wish to apply Optimal Reward Search and Reward Shaping to a novel grid-world setting. A Grid-World is a simulated physical space in which an agent lives. Grid-Worlds serve as a popular choice for simulating Reinforcement Learning tasks since they allow for numerous variations in the setup. We call the family of Grid-World tasks we wish to work with as the **Balanced Diet problem**.
The task is to design an agent that maximizes its fitness level (sparse extrinsic reward) over numerous time steps on a 6 x 6 Grid. To remain fit, the agent needs to consume two groups of food - **Protein and Fat**. Specific diagonally opposite locations in the Grid-World are designated sources for these food groups. Whenever the agent lands on one of

these locations, it instantly eats a serving of the food group and moves into a satiated state with respect to the group. Once the agent has entered into a satiated state, it will become unsatiated with respect to the food group with probability 0.2 in any subsequent time step. While satiated with a certain food group (say Fat), if it lands on the source of the other group (Protein) then it will become satiated in both Fat and Protein. Finally, the sparse extrinsic reward framework is such that every time step for which the agent is:

- Unsatiated in both food groups: Its Fitness remains unchanged
- Satiated in a single food group: It gains a +0.01 in its Fitness level
- Satiated in both food groups: It gains a +1 in its fitness level

The inspiration for this problem comes from the Hungry-Thirsty setup discussed in the paper [S. Singh et al., 2009]. In this setup, an agent can increase its fitness only by regularly satiating its hunger. However its ability to satiate its hunger is conditioned on it being not thirsty. Our problem setup also encourages an alternation between two source locations, however both the sources complement each other and consumption from either of them is unconditional.
The problem we wish to tackle is to efficiently train an agent to learn near optimal policies for this family of tasks. Further we wish to compare the two approaches - Optimal Reward Search and Reward Shaping on this family of tasks and answer the following questions:

- Is Reward Shaping effective on this family of tasks?
- If yes, how similar or different are the policies learnt using Optimal Reward Search and Reward Shaping?
- Do subtle variations in the environment eliminate the efficiency of the same reward function in improving the learning rate, or maybe even slow it down in some cases?

We plan to setup the learning rate simulation using a model free learning algorithm such as **Q-learning**. The experiments we shall conduct using this simulation will be similar to the ones in [S. Singh et al., 2009], hence we shall look at the metric of **Mean Cumulative Fitness** achieved by the learned policies under different reward schemes as a function of the Time Step reached in the Agents lifetime.
From the Optimal Reward Search perspective we expect that reward functions which associate a large negative reward with being completely unsatiated and another larger but still negative **relative penalty** with being satiated in a single food group will encourage the agent to learn the optimal policy of alternating between food groups the fastest.
We also expect to able to design potential like Reward Shaping functions which would achieve the same goals as Optimal Reward Search. Further we expect to find a connection between a well designed shaping function and the reward scheme learnt from Optimal Reward Search.

3 References

1. S. Singh, R. L. Lewis, and A. G. Barto. Where do rewards come from? 2009
2. A. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping, 1999
3. Satinder Singh, Richard L. Lewis, Andrew G. Barto, Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective, 2010