R05/Sarthak Mittal/200050129

February 5, 2023

This paper provides a set of **polynomial-time** algorithms for determining a **Markov reward** that can allow an agent to **optimize** tasks (finding a set of behaviors, a partial ordering over behaviors, or over trajectories) or conclude that such a reward function **does not exist**. Reward being the main **incentive** that drives any reinforcement learning (RL) agent to learn, this paper studies the reward hypothesis by examining how **expressive** the reward is.

They study interactions between a **designer** (thinks of a task) and a **learner** (incentivized to learn a task). They focus on finding out whether there are tasks that cannot be characterized by a Markov reward. They chose Markov functions because many applications rely on **immediate worth**, and history-based rewards have to deal with an additional parameter for the length of the history. They work with **environment-task** pairs and determine whether a Markov reward exists to capture the task in the given environment.

They model the environment as a **Controlled Markov Process (CMP)**, which is a Markov Decision Process (MDP) without a reward function. They assume that reward functions are **deterministic**, and depend only on state, state-action pairs or state-action-state triplets. They assume that the agent will maximize value for a particular discount factor γ. They also highlight the other perspectives of reward in related works. They claim that a suitably rich description of task could help in distinguishing non-optimal behaviors by obtaining an ordering over behaviors.

They define a **Set Of Acceptable Policies (SOAP)** as a non-empty subset of deterministic policies Π from $\mathcal{S}$ to $\mathcal{A}$ for a given E. They claim that a reward function captures a task $\mathscr{T}$ in E when the **start-state value** exactly adheres to the constraints of $\mathscr{T}$ (**optimal** for all good policies and strictly **greater** than other values). They define range-SOAP with the condition that there exists an $\epsilon \geq 0$ such that every $\pi \in SOAP$ is **ϵ-optimal** in start state value.

They define **Partial Ordering on Policies (PO)** as a partial order of the deterministic policies Π. They claim that a reward function captures a PO in E if and only if it produces a start-state value that **orders** Π according to PO.

They define **Partial Ordering on Trajectories (TO)** of length $N \in \mathbb{N}$ as a PO with each trajectory consisting of N state-action pairs. They claim that a reward function captures a TO if the ordering as per the cumulative discounted N-step return from start-state matches the PO.

They claim that there exist $(E, \mathscr{T})$ for which no Markov reward exists. This is due to tasks that have policies or trajectories that are **correlated** in value, due to which the reward is unable to find a PO that distinguishes them. They show this via simple 2-state MDPs where the SOAP is a subset of actions chosen such that PO that distinguishes optimal policies is **not possible**. They also show that even if the transition function or γ are taken as part of the reward specification, it is not sufficient.

They formulate the reward design problem as a **linear program** that matches the constraints with the requirement that reward function has **infinitely many** outputs. They make use of the **"fringe"** (set of policies that differ from a policy in SOAP by exactly one action) to ensure maximality of start-state value of good policies. They also give a **generalization** where they can find a reward function that realizes a task in all environments in a finite set with shared state-action space. However, this is **not closed** under sets of CMPs.

They conclude with an empirical analysis followed by possible relaxation of main **assumptions** - environment may not be a finite CMP, designer may not know environment precisely, reward may depend on hsitory, designer may not know how learner manages states - along with the consideration about how reward impacts the learner's **dynamics**, and to assess whether it is capable of developing **cognition** attributes.