Interpreting the Evolution of an Isolated Physical System in a Deterministic World as a Markov Reward Process

Huichang Chae*

(Dated: April 17, 2024)

Abstract

This paper argues that explaining the evolution of an isolated physical system in a deterministic world using Markov reward processes implies that Nature does not prioritize immediate rewards over future rewards. The resulting equations of motion appear as differential forms of conservation laws, where the conserved quantities are the modified value functions which incorporate the reward of being in the initial state.

* hcc136741@gmail.com

In a Markov reward process [1], the system evolves through a sequence of states, with transitions between states governed by the Markov property, meaning that future states depend only on the current state and not on the sequence of events that preceded it. Additionally, each state transition in an MRP is associated with a reward (or penalty) that quantifies the desirability or cost of moving from one state to another. Key components of an MRP include:

- State Space: The set of all possible states of the system.
- Transition Probabilities: The probabilities governing the transition from one state to another.
- Rewards: A reward is assigned to each state transition.
- Discount Factor: The tendency to prioritize immediate rewards over future rewards.
- Value Function: The expected cumulative reward starting from a given state.

In this paper, we explore the application of Markov reward processes to interpret the evolution of an isolated physical system within a deterministic framework. For a finite number of states, the Bellman equation for MRPs is

$$v(s) = \sum_{s'} p(s'|s) \left[r(s', s) + \gamma v(s') \right] \tag{1}$$

where $v(s)$ is the value function for state s, $p(s'|s)$ is the transition probability from state s to s', $r(s', s)$ is the reward for transitioning from state s to s', and γ is the discount factor. There can be many independent value functions for a system and the associated Bellman equations. In this paper, we will assume there is only one value function, but the generalization to many independent value functions is trivial. For continuous states, we will make three assumptions:

1. A state can be effectively represented by a real number or numbers.
2. The reward $r(s', s)$ depends only on the difference between states s' and s such that one can rewrite the reward as $r(s' - s)$.
3. The reward for transitioning from a state to the same state $r(0)$, is 0.

Then, the Taylor expansions of $v(s')$ and $r(s'-s)$ up to the first order in infinitesimal time dt are as follows:

$$v(s') = v(s,t) + \frac{\partial v(s,t)}{\partial t}dt + \frac{\partial v(s,t)}{\partial s}\dot{s}(t)dt + o(dt) \tag{2}$$

$$r(s'-s) = \frac{\partial r(s,t)}{\partial t}dt + \frac{\partial r(s,t)}{\partial s}\dot{s}(t)dt + o(dt) \tag{3}$$

Plugging equations (2) and (3) into equation (1) gives the equation (after changing the probability mass function to a probability density function and the discrete sum to an integral),

$$(\gamma - 1)v(s,t) + \int p(s'|s)\left[\frac{\partial V(s,t)}{\partial t} + \frac{\partial V(s,t)}{\partial s}\dot{s}(t)\right]dt\,ds' = 0 \tag{4}$$

where $V(s) := \gamma v(s) + r(s)$ is the "proper" value function for state s, in the sense that it incorporates the reward for transitioning from state 0 to state s, whereas the value function $v(s)$ does not. If equation (4) were to be satisfied *without knowing the value function and transition probabilities*, the following two equations must be true:

$$\gamma = 1 \tag{5}$$

$$\frac{\partial V(s,t)}{\partial t} + \frac{\partial V(s,t)}{\partial s}\dot{s}(t) = 0 \tag{6}$$

Equation (5) implies that Nature does not prioritize immediate rewards over future rewards. Equation (6) is the equation of motion where the proper value V is a conserved physical quantity.

---

[1] Silver, D. UCL course on reinforcement learning lecture 2: Markov decision processes (2020). URL `https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf`.