DPx,y, z
DPa, b, c

Example Usage of MDPNv1 Notation

Billy Okal

University of Freiburg

1 Introduction

Many Reinforcement learning (RL) research papers contain paragraphs that define Markov decision process (MDP) and related concepts. These paragraphs take up space that could otherwise be used to present more useful/new content. In this paper we specify a notation for MDP that can be used by other papers. Declaring the use this notation using a single sentence can replace several paragraphs of notational specifications in other papers. Importantly, the notation that we define is a common foundation that appears in many RL and related papers, and is not meant to be a complete notation for an entire paper.

2 MDPs using MDPNv1 notation

A MDP is denoted by a tuple, $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, R, d_0, \gamma)$, where;

1. We use $t \in \mathbb{N}_{\geq 0}$ to denotes the time step, where $\mathbb{N}_{\geq 0}$ denotes the natural numbers *including zero*.
2. $\mathcal{S}$ is the set of possible states that the agent can be in, and is called the *state set*. The state of the environment at time t is a random variable that we denote by S_t. We will typically use s to denote an element of the state set.
3. Similarly, $\mathcal{A}$ is the set of possible actions the agent can perform. The action at time t is denoted by A_t, while a denotes an element of the action set.
4. ...
5. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longmapsto [0, 1]$ is called the *transition function*. For all $(s, a, s', t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{N}_{\geq 0}$, let $P(s, a, s') \coloneqq \Pr(S_{t+1} = s' | S_t = s, t = a)$. That is, P characterizes the distribution over states at time $t + 1$ given the state and action at time t.

   We allow three alternate notations for P. First, let $P(s'|s, a) \coloneqq P(s, a, s')$. This form takes approximately the same amount of space, but makes it more clear that P is a conditional distribution over the next state given the current state and action. Second, let $P_s^a(s') \coloneqq P(s, a, s')$. This notation moves terms into subscripts and superscripts in order to save some space. Third, let $P_{s,s'}^a \coloneqq P(s, a, s')$. This final form is particularly useful when space is limited. The author is should select one the four notations for P, and remain consistent within each paper.
6. d_0 is the initial distribution of states defined as $d_0 : \mathcal{S} \longmapsto [0, 1]$.
7. γ is the discount factor defined as $\gamma \in [0, 1)$.
8. π is the policy which is defined as $\pi : \mathcal{S} \times \mathcal{A} \longmapsto [0, 1]$

And so on and so forth

3 Acknowledgments

I want to thank Phillip Thomas for fruitful discussions on the naming and contents of this package.

References

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.