

On the Episode Duration Distribution in Fixed-Policy Markov Decision Processes

Itamar Arel, Andrew Davis

Department of Electrical Engineering and Computer Science
 The University of Tennessee
 Knoxville, TN 37996
 itamar@ece.utk.edu, adavis72@utk.edu

Abstract

This paper presents a formalism for determining the episode duration distribution in fixed-policy Markov decision processes (MDP). To achieve this goal, we borrow the notion of obtaining the n^{th} -step first visit probability from queuing theory, apply it to a Markov chain derived from the MDP, and arrive at the distribution of the episode durations between any two arbitrary states. We illustrate the proposed methodology with an agent navigating a 25-state maze, demonstrating the applicability of the method.

Introduction

A wide range of engineering applications, ranging from automotive systems to robotics, employ Markov decision processes (MDPs) (Sutton & Barto 1998) as an underlying formalism to determine the optimal policy, or control scheme, in a dynamic stochastic environment. An MDP is defined as a quadruple $\langle |S|, A, P_A, R \rangle$, where $|S|$ denotes a finite state space, A contains all possible actions that can be taken at a particular state, P_A represents the probability transition function $|S| \times |S| \times A \rightarrow [0, 1]$, and R represents the mapping of state-action pairs to rewards, $R : |S| \times A \rightarrow \mathbb{R}$. The MDP quadruple serves as a perfect model for an application space, whereas the actual planning, which involves determining an optimal set of actions that must be taken to accumulate maximum reward, is referred to as a dynamic programming (DP) problem.

Occasionally, we may find a particular interest in evaluating the expected duration of a trajectory from state i to state j . For example, in an episodic task a starting state can be arbitrary, but the terminal state may be fixed. In such cases it would be interesting to determine the expected duration from start to terminal/goal state. To date, there has not been a formal method for ascertaining such trajectory duration distributions. This paper presents such a methodology for the general case of MDPs with fixed policies, i.e. policies in which the transition probabilities do not change over time.

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Obtaining the Episode Duration Distribution

The framework proposed for obtaining the episode duration distribution in fixed-policy MDPs is as follows: (1) We extract a Markov chain from the MDP, which expresses the expected transition probabilities from any state to any other state that is directly reachable. (2) Next, we obtain the probability generating function of the expected transition probability matrix. (3) Finally, we utilize the queuing theory result pertaining to the n^{th} -step first visit probability to derive the episode duration distribution. The following sections provide more detail on this methodology.

Extracting the Expected Transition Probability Matrix

An MDP is fully defined by a state space, S , and action set A , along with two constructs: the action-dependent state transition probabilities,

$$P_{ss'}^a = \Pr [s' | s, a], \quad (1)$$

and the expected reward, $R_{ss'}^a$, which expresses the expected reward to be received when transitioning from state s to s' . The policy, $\pi(s, a)$, defines the mapping between states and actions, such that $\pi(s, a) = \Pr [a | s]$. Based on the above, let the expected transition probability matrix be defined as:

$$G_{ss'} = \sum_{a \in A(s)} \pi(s, a) P_{ss'}^a, \quad (2)$$

which reflects the average rate of transition from state s to state s' in S . where $A(s)$ denotes the action set that is permissible at state s . It should be noted that in some cases, not all transitions are possible, so therefore some elements of G may be zero. It is also assumed that the policy is stochastic yet stationary in that its elements do not change over time.

Deriving the n^{th} -step First Visit Distribution Function

Once we obtain G , the next step is to find the probability-generating function (PGF) of the matrix,

$$H(z) = [I - Gz]^{-1}, \quad (3)$$

We note that $H(z)$ requires matrix inversion, which is an $O(N^3)$ operation. It has been shown (Hunter 1998) that we

can obtain the PGF of an n^{th} -step first transition probability distribution through the following expression:

$$F(z) = \frac{H_{ss'}(z) - \delta_{ss'}}{H_{s's'}(z)}, \quad (4)$$

where δ denotes the Kronecker delta. Element (s, s') of the n^{th} -step first transition probability matrix expresses the likelihood that a transition from state s to state s' will occur in precisely n steps. This is the exact interpretation of episodes, if s is the starting state and s' the terminal state. Thus, the inverse PGF transform on $F(z)$ is expected to yield the episode duration probability mass function - the metric which we seek. The relationship between $F(z)$ and the probability mass function for the episode duration is

$$f(k) = \frac{F^{(k)}(0)}{k!}, \quad (5)$$

where $F^{(k)}(0)$ denotes the k^{th} derivative with respect to z evaluated at zero and $k!$ is k factorial.

Example

We illustrate the methodology described using simple 25-state maze, as illustrated in Figure 1. At each state, four actions are permissible: Up, Down, Left, and Right. If the agent chooses an action that results in the collision with a wall, the agent will remain in the same state.

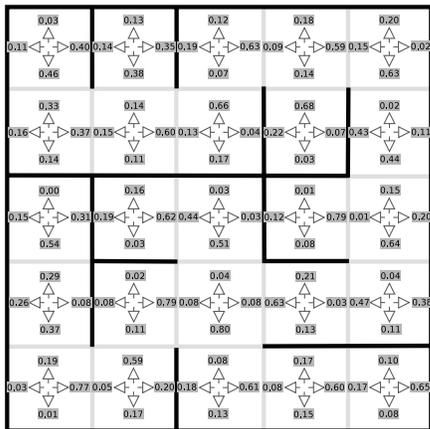


Figure 1: “The 25-state Maze Example, MDP.”

First, we must convert the policy and environment into a Markov chain. By (2), the probability of moving into a state’s neighbor is simply the probability that the agent will take that action. The probability of looping back into the same state is the sum of the probability of all actions that would result in a collision with a wall. By evaluating this simple computation, we now have the MDP of the maze example.

Figure 2 shows the MDP derived from the Markov chain, where the numbers preceding the arrows transitioning into the adjacent cells indicate the probability of taking that action, and the numbers in the top left corner of each cell indicate the probability of colliding with a wall and remaining in that state.

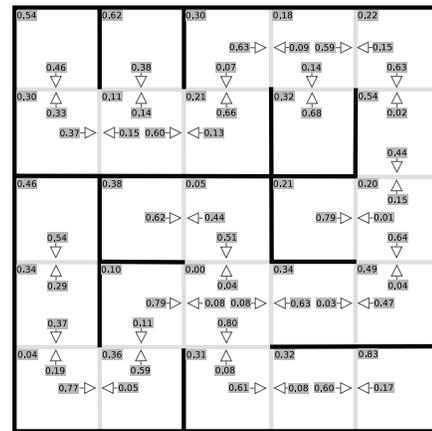


Figure 2: “The 25-state Maze Example, cast into a Markov chain.”

Now that we have the expected transition probability matrix G , we obtain $H(z)$ by applying (3). Next, we must obtain $H_{ss'}(z)$ and $H_{s's'}(z)$ by defining the initial state and the goal state, which will be (1,1) and (5,5), respectively. Applying these two functions to (4) gives us $F(z)$, the PGF of the probability mass function. Finally, we can apply (5) to obtain the probability mass function (see Figure 3).

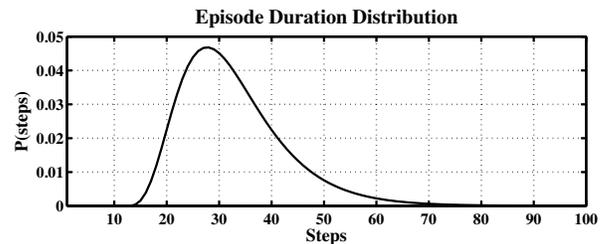


Figure 3: “The Episode Duration Distribution of the Maze Example.”

In conclusion, this paper presented a methodology for analytically determining the episode duration distribution of a Markov decision process. An agent could use the distribution as a simple discriminator between two policies - if episode brevity is of utmost importance, the agent could compare the expected episode duration of two different policies derived from the episode duration distribution. The agent would then choose the policy with the shorter expected value. MDPs that offer a reward at the terminal state are especially applicable to this discriminator, so this methodology could be applied to any problem in this subset of MDPs.

References

- Hunter, J. J. 1998. *Mathematical Techniques of Applied Probability: Discrete Time Models: Techniques and Applications, Vol. 2*. Academic Press.
- Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. The MIT Press.