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

Andrew Davis, Itamar Arel, *Senior Member IEEE*
Department of Electrical Engineering and Computer Science
The University of Tennessee
Knoxville, TN 37996-2100

***Abstract*—A method for obtaining a computationally efficient closed-form solution for the episode duration distribution in finite-horizon, fixed-policy Markov decision processes (MDP) is presented. The approach is based on finding the n^{th}-step first visit probability mass function between any two states of a Markov chain derived from the MDP. Simulation results of the technique applied to a 25-state maze clearly support the analytical derivations.**

I. Introduction

A wide range of engineering applications, ranging from automotive systems to robotics, employ Markov decision processes (MDPs) [1] 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.

It is often desired to evaluate the expected duration of a trajectory originating from state i to state j. For example, in an episodic task, a starting state can be arbitrary while the terminal state may be fixed. In such cases it would be valuable to obtain a closed-form expression for the expected duration of the trajectories from start to terminal/goal states. To date, there has not been a formal method for ascertaining such trajectory duration distribution. This letter presents such a methodology for the general case of MDPs with fixed policies, i.e. policies in which the action-dependent transition probabilities do not change over time.

II. Extracting the Expected Transition Probability Matrix

The framework proposed for obtaining the episode duration distribution in fixed-policy MDPs comprises of the following three steps: (1) derive a Markov chain from the MDP, which expresses the expected transition probabilities from any state to any other state that is directly reachable; (2) obtain the probability generating function of the expected transition probability matrix; (3) utilize the n^{th}-step first visit probability mass function to attain the episode duration distribution.

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

$$P^a_{ss'} = \Pr\left[s'|s,a\right], \tag{1}$$

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

$$G^\pi = \sum_{a \in A(s)} \pi(s,a) P^a_{ss'}, \tag{2}$$

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

The n^{th}-step First Visit Distribution Function - Once G^π is obtained, the next step is to find the probability-generating function (PGF) of the matrix,

$$G(z) = \left[I - zG^\pi\right]^{-1}. \tag{3}$$

We note that $G(z)$ requires matrix inversion, which is an $O(N^3)$ operation. It has been shown [2] that we can obtain the PGF of an n^{th}-step first transition probability distribution through the following expression:

$$F_{ss'}(z) = \frac{G_{ss'}(z) - \delta_{ss'}}{G_{s's'}(z)} \tag{4}$$

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

$$f(k) = \frac{F^{k}_{ss'}(0)}{k!}, \tag{5}$$

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

Given the great cost of evaluating k derivaties and factorials, we must seek a more computationally efficient method for determining the probability mass function. Because $F_{ss'}(z)$ is a quotient of two functions, we can apply $F_{ss'}(z)$ to a recursive method of determining the k^{th} derivative of a function [3].

$$\frac{F^{(k)}_{ss'}(z)}{k!} = \frac{1}{G_{s's'}(z)} \left(\frac{G^{(k)}_{ss'}(z)}{k!} - \sum_{j=1}^{k} \frac{G^{(k+1-j)}_{s's'}(z)}{(k+1-j)!} \frac{F^{(j-1)}_{ss'}(z)}{(j-1)!} \right) \tag{6}$$

Now we must determine a closed-form solution for the k^{th}derivative of $G^{(k)}(z) \mid_{z=0}$ so we may directly substitute these closed-form solutions into (6).

$$G^{(k)}(z) \mid_{z=0} = G^{\pi^{k}}(k!) \tag{7}$$

Now that we have a closed-form solution for $G^{(k)}(z) \mid_{z=0}$, we can insert the solution into (6) to obtain

$$\frac{F^{(k)}_{ss'}(z)}{k!} = G^{\pi^{k}}_{ss'} - \sum_{j=1}^{k} (G^{\pi^{k+1-j}}_{s's'}) F^{(j-1)}_{ss'}(z) \tag{8}$$

Now, we can equate (5) to (8) to obtain the final recursive solution:

$$f(k) = G^{\pi^{k}}_{ss'} - \sum_{j=1}^{k} (G^{\pi^{k+1-j}}_{s's'}) f(k-1) \tag{9}$$

III. Simulation

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.

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 Markov chain of the maze example. Figure 2 shows this 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.

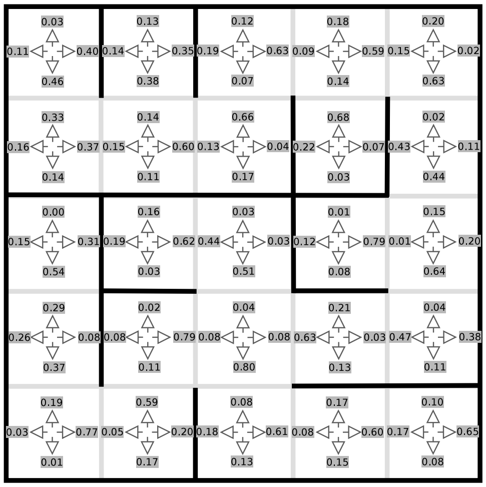

Figure 1. The 25-state Maze Example, MDP.

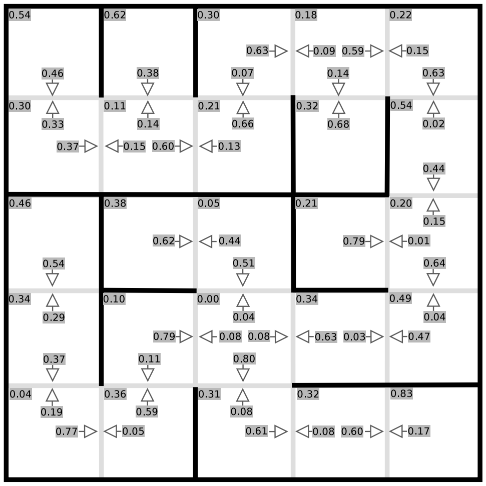

Figure 2. The 25-state Maze Example, cast into a Markov chain.

By simply substituting our example Markov chain matrix G^{π} into (9), we can efficiently compute the expected duration distribution, which is shown in Figure 3.

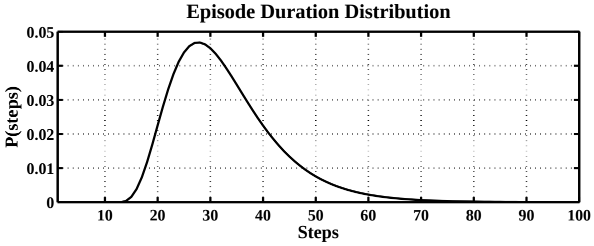

Figure 3. The Episode Duration Distribution of the Maze Example.

IV. Conclusions

This paper presented a methodology for analytically determining the episode duration distribution of a Markov decision process. There are many scenarios where such derivation could be useful – in situations where episode brevity is of essential importance, an agent could favor a policy with a shorter mean episode length over another policy with a longer mean episode length, but with a higher mean reward. To that end, the applicability of the proposed scheme is broad.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning : An Introduction*. The MIT Press, 1998.

[2] J. J. Hunter, *Mathematical Techniques of Applied Probability: Discrete Time Models: Techniques and Applications, Vol. 2*. Academic Press, 1983.

[3] C. Xenophontos, “A formula for the nth derivative of the quotient of two functions.”.