RS-HL-9: Time Variant MDP

UT Austin Oden Institute

Hongseok Kim

07/02/2024

Scope

- In this report, the time variant MDP process of satellite network structure is presented.
- We have dataset of satellite inter-connection status of each time vector, and we want to optimize the data packet transmission pathway from ground station n to ground station m.
- We are using Markov Decision Process for this framework. Initally, for the last iteration, we will use conventional MDP policy iteration process which is presented in RS-HL-5 and RS-HL-7, then we will propagate the state value function according to inverse time vector (T = 99, 98, 97, ... 1). Then, we will generate the policy according to conveged state value parameter.
- **Question: Is there any difference between state value iteration and policy iteration on T = 100?**

I. Key Theory

I.1 Structure from Time-Invariant MDP

Structure Reminder

Level 1 : MDP

Level 2 : State

Level 3 : Action, **State Value**

Level 4 : Next State, **Action Value**, **Policy Function**

Level 5 : State Transition Probability, Reward

I.2 Steps for Time Variant MDP using Dynamic Programming

Steps for Time Variant MDP − DP

Step 1 : Policy Iteration for $T = T_0 + \Delta T$

Output info for the input : $V(s, T_0 + \Delta T)$

Step 2 : State Value Propagation from $T = T_0 + \Delta T$ to $T = T_0$

$q(s, a, t) = \sum_{s',r} p(s', r|s, a)(r(s') + \gamma V(s', t + 1)) \rightarrow$ Level 4

$V(s, t) = \max_a \sum_{s',r} p(s', r|s, a)(r(s') + \gamma V(s', t + 1)) \rightarrow$ Level 3

Input info from $T = t$: $p(s', r|s, a) \Rightarrow$ From Satellite to Satellite Contact Matrix at $T = t$

− We have to define action from $s(t)$ to $s'(t + 1)$ and corresponding rewards

Input info from $T = t + 1$: $V(s', t + 1) \Rightarrow$ From State value (level 3) at $T = t + 1$

Output info for $T = t$: $V(s, t), q(s, a, t) \Rightarrow$ Reuse this information to calculate $V(s, t - 1), q(s, a, t - 1)$

Step 3 : Policy Determination Based of State Value

For each $T = t$,

$\pi(s, t) = \underset{a}{\text{argmax}} \sum_{s',r} p(s', r|s, a)(r(s') + \gamma V(s', t + 1))$

If there is n multiple actions which have same action value, $\pi = \frac{1}{n}$ for each action

Step 4 : Test

Input : S_n at $T = T_0$

Output : S_m at $T = T_0 + \Delta T$

Or any time, any input can be possible, check wheter the output is intended destination

I.3 New Structure Based on Time Variant MDP

Structure Reminder

Level 1 : MDP

Level 2 : Time $(T, T + t_1, T + t_2, \cdots T + \Delta T)$

Level 2 : State

Level 3 : Action, **State Value**

Level 4 : Next State, **Action Value**, **Policy Function**

Level 5 : State Transition Probability, Rewards

Key Equations to calculate

Action Value : $q(s,a,t) = \sum_{s',r} p(s',r|s,a)(r(s') + \gamma V(s',t+1)) \rightarrow$ Level 4

State Value : $V(s,t) = \max_a \sum_{s',r} p(s',r|s,a)(r(s') + \gamma V(s',t+1))$

II. Code Demonstration

II.1 Variable Initialization

```
% clear;clc;

% Define destination satellite number
destination_state = 38;

% 1.1 Initialize the MDP Structure (Level 1)
MDP = struct();

% 1.2 Load the Satellite Contat Dataset
load('/workspace/RS_Dataset/RS_HL_3_dataset.mat')

% 1.2.1 Select Elapsed time slot for the simulation:
%   15 seconds timestep,so n time_indices indicates start time = 15*n seconds

time_index = 1000;
sat_to_sat_contact_matrix = sat_to_sat_contact_3d_matrix(:,:,time_index);
```