History Dependency

Monkie

6/22/2022

We consider MDP with objective

$$\max_{\pi} \rho[\sum_{t} \gamma^t R_t^{\pi}]$$

where ρ refers to risk measure of interest. It is well known for $\rho = \mathbb{E}$ in finite horizon the optimal policy is time dependent deterministic, and is determistic in infinite horizon. In this document, we will provide a simple example to show that when $\rho = \text{VaR}$ or $\rho = \text{CVaR}$ the optimal policy is history dependent. Let the discount factor $\gamma = 1$ for this example.

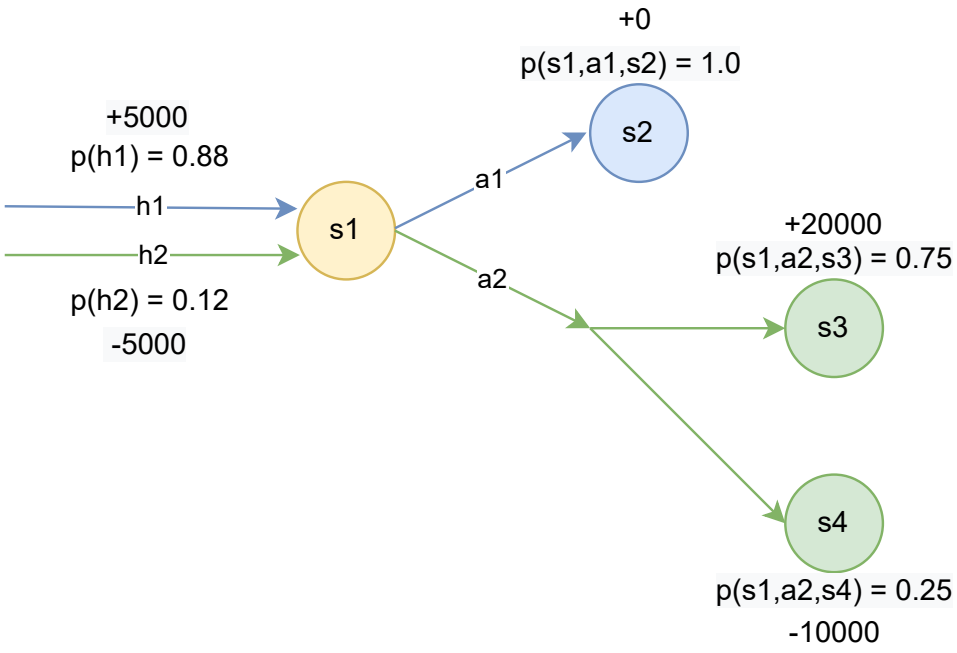

Figure 1: MDP Tree

For expectation, given MDP tree above the decision is independent with respect to the history/ accumulated-total-reward. As a result, regarding history the optimal policy is taking action a2 given state s1.

```
## a1 expected reward-to-go given s1 : 0
```

```
## a2 expected reward-to-go given s1 : 12500
```

However, when $\rho = \text{VaR}$ or $\rho = \text{CVaR}$ we care about the tail distribution of the total discounted reward. As a result,

a1		a2		a1|h1 & a2|h2		a2|h1 & a1|h2	
prob	reward	prob	reward	prob	reward	prob	reward
0.88	5000	0.22	-5000	0.88	5000	0.22	-5000
0.12	-5000	0.66	25000	0.03	-15000	0.66	25000
		0.03	-15000	0.09	15000	0.12	-5000
		0.09	15000				
VaR 10%	-5000	VaR 10%	-5000	VaR 10%	5000	VaR 10%	-5000
CVaR 10%	-5000	CVaR 10%	-8000	CVaR 10%	-1000	CVaR 10%	-5000

In this simple example, the optimal policy for both VaR and CVaR at 10%, is take action a1 if given h1 and take action a2 if given h2 (which is history dependent).