CS3243 **AY20/21 Sem1**

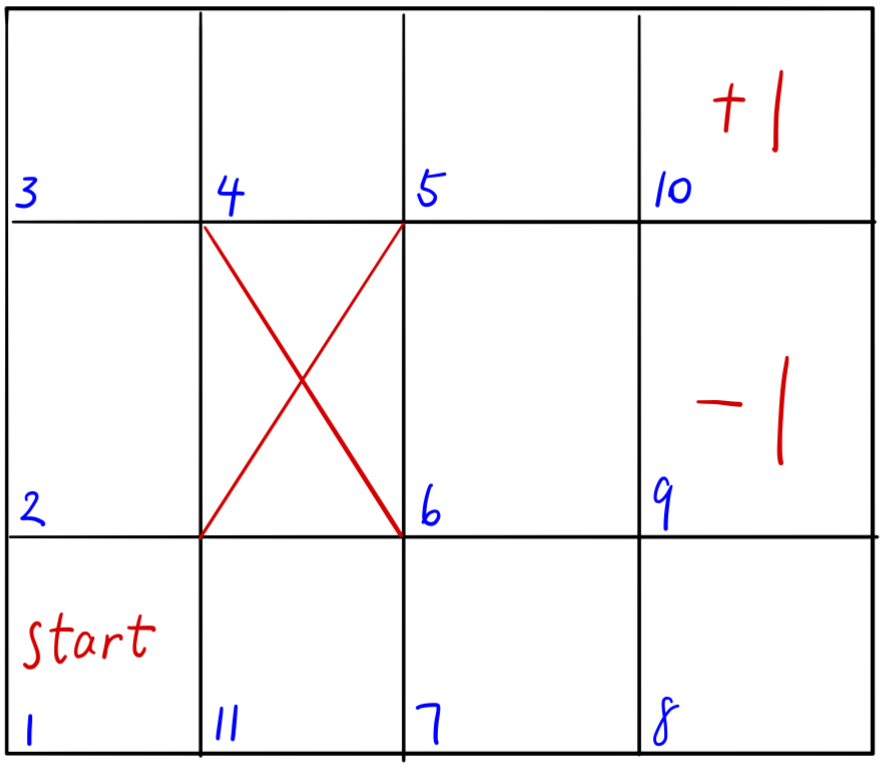
Lecture 8: October 14

Lecturer: MEEL, Kuldeep S. *Scribes: Song Qifeng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

8.1 From stochastic environment to Markov Decision Process

8.1.1 Define a stochastic environment



We will use the grid environment as shown above to represent a stochastic environment.

We will first define the stochastic environment as follows:

States: Every grid in the environment except the one with red cross represents a state. For a grid with number i, we will denote the state as s_i. Note that the 's' here is in lower case.

Actions: At each state, the agent can take an action $a \in \{up, down, left, right\}$, however, due to the stochasticity of the environment, the agent may not behave as expected. For example, suppose the agent is at s_1 and it takes the action 'up', it may end up in s_{11}.

Transition model: The transition model takes in the current state and an action as input, and returns a probability distribution over all the possible states the agent will with.

Reward function: The reward function is similar to the cost function we have seen before, with two differences: 1. What the rewards function returns is the magnitude of reward instead of cost, hence the agent should try to maximize it. 2. The reward function is designed as $state \rightarrow reward \in \mathbb{R}$.

Initial state: The initial state of this environment is s_1 as labeled in the image.

Goal: There is no specific or permanent goal of this game. The agent is utility-based, meaning that it's only purpose is to maximize its reward when it terminates at the terminal states.

Terminal states: For this environment, there are two terminal states, s_9 and s_{10}. The game ends when the agent reaches any of the terminal states.

8.1.2 Define a plan for the agent

8.1.2.1 Define as a sequence of actions

In deterministic environments, a plan can be defined as a sequence of actions. Once a plan is given, the path of the agent is determined. However, in a stochastic world, an action may not result in expected outcome, therefore it is risky to continue defining a plan as a sequence of actions.

For example, suppose for any actions, there is a chance of 0.7 that the agent will behave exactly as told. Then the chance of the agent behaving correctly in a consecutive 5 times is only 0.16807. Hence if we still stick to the traditional plans, we will face high risks.

8.1.2.2 Define as a policy

Since the outcome of the agent's movement may differ from the expectation, hence we conclude that the situations in a stochastic world is very changeable, and we need to design a more flexible plan such that it can overcome the changeability of the environment.

Here we define a plan as a mapping: $state \rightarrow action$, and we call this kind of plan as policy.

8.1.3 Define the utility

As we have stated before, the agent is utility-based, it compares the utility of different policies and choose the one of highest utility. Thus it is important to define the utility of a given plan.

8.1.3.1 Utility of a sequence of states

We first define the utility of a given sequence of states $\tau = [s_0, s_1, s_2, \ldots, s_n]$ or $\tau = [s_0, s_1, s_2, \ldots]$.

Let $U_h(\tau)$ be the utility of the sequence of states τ,

$$\begin{aligned} U_h(\tau) &= U_h([s_0, s_1, s_2, \ldots]) \\ &= \sum_{t=0}^{\infty} R(s_t) \\ &= \sum_{t=0}^{\infty} \gamma^t R(s_t) \end{aligned} \tag{8.1}$$

As shown in math, there are two ways of calculating the utility of a given sequence of states. The first one is additive of all the rewards provided at each state. The second one discounts the rewards of later states by a discounting factor $\gamma \in (0, 1)$. The second way of calculating utility corresponds to an economic concept "impatience" which serves as the basis for interest. To make it simple, an 100 dollar note today is more valuable than an 100 dollar note tomorrow, even without considering the inflation. In the context of a stochastic environment, an early reward is more desirable than a late reward, hence we discount the reward by γ for every step.

Additionally, if there is a maximum reward R_{max}, then the utility is upper bounded by $\frac{R_{max}}{1-\gamma}$.

8.1.3.2 Utility of a policy

Since a policy does not guarantee a certain sequence of states, the state at each stage can only be represented as a probability distribution over all possible states. Hence we define the utility of a policy as a mathematical expectation. We denote each state as a variable S_i, note that the 'S' here is in upper case.

$$\begin{aligned} U^{\pi}(S) &= E(U_h(\tau)) \\ &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t)] \end{aligned} \tag{8.2}$$

Now we can calculate the utility of any policy $\pi(s)$ at s. The optimal policy at s is $\pi^*(s) = argmax_{\pi} U^{\pi}(s)$. We can also define the optimal policy as $\pi^*(s) = argmax_{a \in actions} \sum_{s'} P(s'|s, a) U^{\pi^*}(s')$.

To simplify the notation, let $U(s) = U^{\pi^*}(s)$. Then,

$$\pi^*(s) = argmax_{a \in actions} \sum_{s'} P(s'|s, a) U(s') \tag{8.3}$$

8.1.3.3 Bellman equation for utilities

$$
\begin{aligned}
U(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t)] \\
&= E[R(s_0) + \sum_{t=1}^{\infty} \gamma^t R(s_t)] \\
&= R(s) + E[\sum_{t=1}^{\infty} \gamma^t R(s_t \mid s_0 = s)] \\
&= R(s) + \sum P(s' \mid s, \pi^*(s))(\gamma R(s') + E[\sum_{t=2}^{\infty} \gamma^t (R(s_t \mid s_1 = s'))]) \\
&= R(s) + \sum P(s' \mid s, \pi^*(s))\gamma(R(s') + E[\sum_{t=2}^{\infty} \gamma^{t-1} (R(s_t \mid s_1 = s'))]) \\
&= R(s) + \sum P(s' \mid s, \pi^*(s))\gamma(R(s') + E[\sum_{t=1}^{\infty} \gamma^{t'} (R(s_{t'} \mid s_0 = s'))]) \\
&= R(s) + \gamma \sum P(s' \mid s, \pi^*(s))U(s') \\
&= R(s) + \gamma max_{a \in A(s)} (\sum_{s'} P(s' \mid s, a)U(s'))
\end{aligned}
$$