Intro, MDP

Return

Discounted sum of rewards from time step t to horizon H:

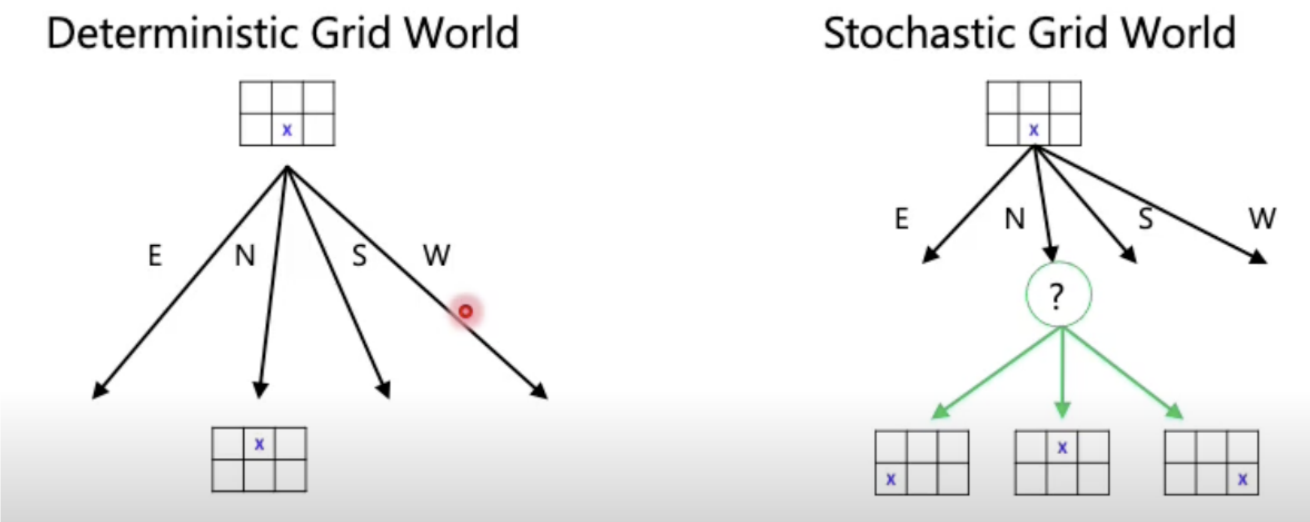
$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{H-1} r_{t+H-1}$$

Value function

Expected return from starting in state s:

$$V(s) = E[G_t | s_t = s] = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{H-1} r_{t+H-1} | s_t = s]$$

If the process is deterministic, the return G_t is equal to value function $V(s)$, the expectation term drops.



MDP

Markov Decision Process is a tuple consisting of S, A, P, R, γ
S: states. 3 different definitions

- Full environment state S_t^e : private to the environment, not visible, maybe irrelevant
- Agent state S_t^a : private to the agent, history of observations, rewards and actions. The agent constructs a state representation using a function of history $S_t^a = f(H_t)$ to decide on the next action
- Information state: useful information from the history. S_t^a with special constraints in $f(H_t)$.
  P: state transition matrix. The rows sum up to 1.0 and P could change over time.
  R: reward function. Can depend on state and action, or only on state.
  γ: discount factor. Mathematically convenient. Humans act as if there's a discount factor < 1.
  If episode lengths are always finite($H < \infty$), can use $\gamma = 1$.

$\gamma = 0$: myopic. Only care about immediate reward
$\gamma = 1$: future reward is as beneficial as immediate reward
A large γ implies we weight delayed/long term rewards more.

Goal

Maximize the expected return $E[G_t]$

Policies

- Deterministic: $a = \pi(s)$
- Stochastic: $\pi(a|s) = P[A_t = a|S_t = s]$