Markov? memoriless

$$P(x_{t+1} \mid x_0 \ldots x_t) = P(x_{t+1} \mid x_t)$$

Today

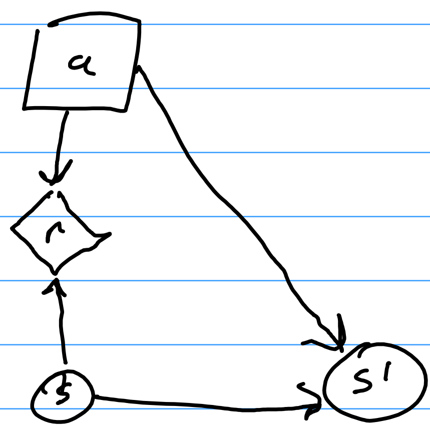
What is an MDP?

What is a policy?

How to evaluate policies

Policy Search

MDPs



maximize $\sum_{t=0}^{\infty} \gamma^t r_t$

"Tuple Definition"

$(S, A, R, T, \gamma) + p_0$

$s \in S$

$s = (x, y) \quad S = \mathbb{R}^4$

S - state space - set of states (e.g. $\{1, 2, 3\}$, $\mathbb{R}^4$, {working, malfunctioning}, $[0,1]^4$)

A - action space - set of actions

$A(s)$

R - reward function $\quad R: S \times A \times S \to \mathbb{R}$

$$R(s,a) \equiv \underset{s'}{E}[R(s,a,s')]$$

T - "transition kernel"

Explicit or Implicit ("Generative Model")

$T(s' | s,a) \qquad s' = G(s,a)$

γ - discount $\qquad \gamma \in [0,1)$

p_0 - initial state distribution

Breakout Rooms

Cooking a pot of Pasta

(S, A, R, T, γ)

Team 5

$$S = \underbrace{\{1 \dots 10\}}_{\text{water temp}} \times \underbrace{\{1 \dots 10\}}_{\text{softness}} \times \underbrace{\{0 \dots 5\}}_{\text{saltiness}}$$

$A = \{$up, down, add salt, eat, taste, add sauce$\}$

$R =$ + tastiness $-(\text{softness}-5)^2$ -1 timestep

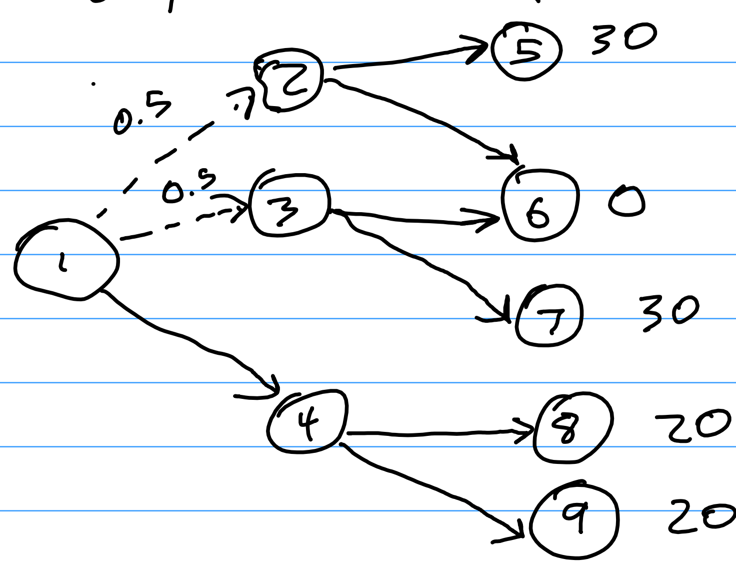
$T =$ if $a =$ high temp increases, if temp $= 10$ softness $+= 1$

$\gamma = 0.99$

Policies determine what actions are taken

Open Loop: sequence

Closed Loop: $\pi : S \to A$



$A = \{\text{up}, \text{down}\}$

open loop policies

$(\uparrow, \uparrow)$	$0.5 \cdot 30 = 15$
$(\uparrow, \downarrow)$	15
$(\downarrow, \downarrow)$	20
$(\downarrow, \uparrow)$	20

Closed Loop $\left(\uparrow, \begin{cases} \uparrow & \text{if } 2 \\ \downarrow & \text{ow} \end{cases}\right)$ $0.5 \cdot 30 + 0.5 \cdot 30 = 30$

Evaluation

estimate $\overset{\text{return}}{u} \equiv \underset{s_0 \sim p_0}{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, a_t = \pi(s_t)\right]$

$$\hat{u} = \sum_{t=0}^{T-1} \gamma^t r_t$$

Simulation

$s \leftarrow \text{sample}(p_0)$

$\hat{u} \leftarrow 0$

for t in $0 \ldots T-1$

  $s', r \leftarrow G(s, a)$ $\leftarrow a \leftarrow \pi(s)$

  $\hat{u} \mathrel{+}= \gamma^t r$

  $s \leftarrow s'$

return $\hat{u}$

$$u \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_i$$

$\leftarrow$ from simulation

$\uparrow$ Monte Carlo Estimate