Reinforcement Learning

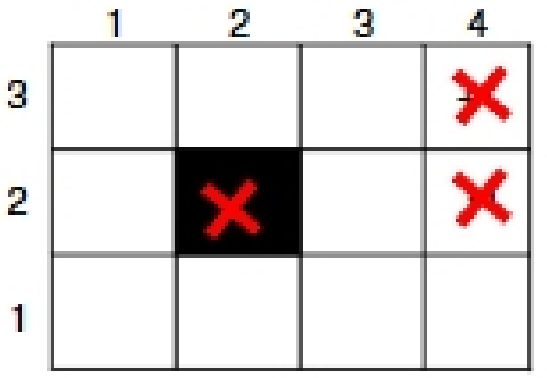
Reinforcement Learning

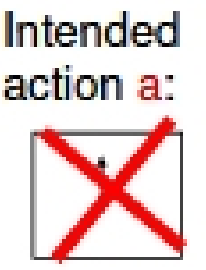
- R&N Chapter 21

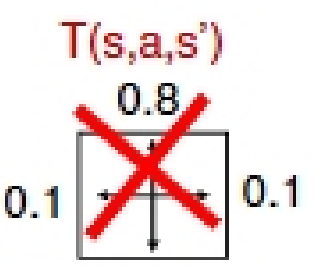
- Demos and Data Contributions from

Vivek Mehta (vivekm@cs.cmu.edu)
Rohit Kelkar (ryk@cs.cmu.edu)

Reinforcement Learning







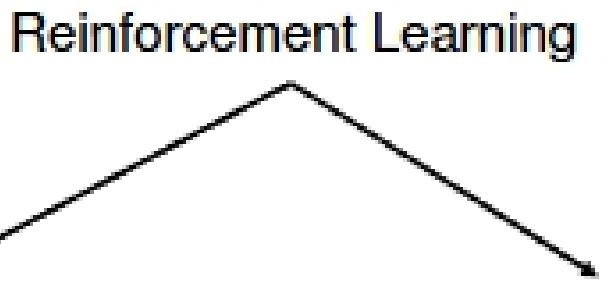
- Same (fully observable) MDP as before except:
  - We don't know the model of the environment
  - We don't know T(.,.,.)
  - We don't know R(.)
- Task is still the same:
  - Find an optimal policy

General Problem

- All we can do is try to execute actions and record the resulting rewards
  - World: You are in state 102, you have a choice of 4 actions
  - Robot: I'll take action 2
  - World: You get a reward of 1 and you are now in state 63, you have a choice of 3 actions
  - Robot: I'll take action 3
  - World: You get a reward of -10 and you are now in state 12, you have a choice of 4 actions
  - .....

Learning from experience ....

Classes of Techniques



Model-Based

- Try to learn an explicit model of $T(.,.,.)$ and $R(.)$

Model-Free

- Recover an optimal policy without ever estimating a model

Model-Based

- If we knew a good estimate $T^{est}(.,.,.)$ of $T(.,.,.)$ and $R(.)$, we could evaluate the optimal policy by solving the fundamental MDP relations:

$$U^{est}(s) = R(s) + \gamma \max_a \left(\sum_{s'} T^{est}(s,a,s')\, U^{est}(s')\right)$$

$$\pi^*(s) = \operatorname{argmax}_a \left(\sum_{s'} T^{est}(s,a,s')\, U^{est}(s')\right)$$