Phani Ram Sayapaneni
University at Buffalo

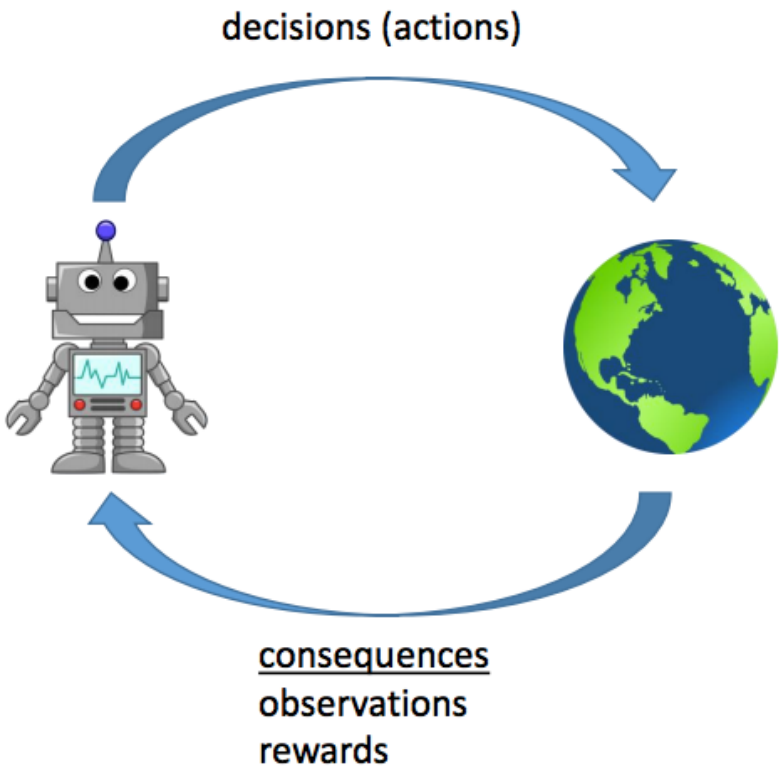
REINFORCEMENT LEARNING

Motivation:
Reinforcement Learning is one of the core elements for many intelligent systems. A simple analogy in human beings would be the dopamine system. We are rewarded when we take food, explore new concepts/places in the world. Most of it is inherited through evolution and few have been acquired from the environment we interact with. Developing intelligent systems would never be complete with out this reward center.

Also there is a huge economic value($7 trillion) around reinforcement learning due to their applications in self driving cars.

Introduction:

What is reinforcement Learning: ?
It is the learning mechanism/response to the outputs we receive when we interact any environment(world)



For every interaction, we have observations, rewards as a consequence.

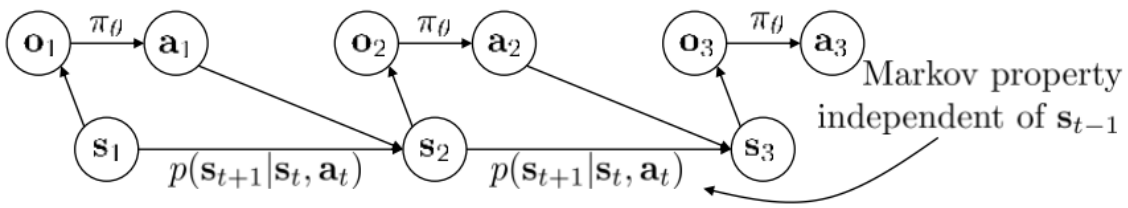
Terminology:
State (s_t): A state is configuration of the environment that the agent is interacting with.

Observation (o_t): Observation is the configuration of the environment that the agent observes/perceives. Usually state fully encapsulates everything about the world, but observation need not.

Action (a_t): Action is a typical interaction with the environment, example: running, holding objects etc.

Policy (p_t): Policy is the function that hold distribution of actions conditioned on observations.
$p_t = \Pi(a_t \mid o_t)$

Mathematical relation between the terms:
All the terms could nicely be represented using a probabilistic graph, where action, state influences future state as shown below.



Here the states hold markov property, where as observations need not. Policy decides what actions to pursue when conditioned on observations.

Probabilistic graph to a typical machine learning formulation:

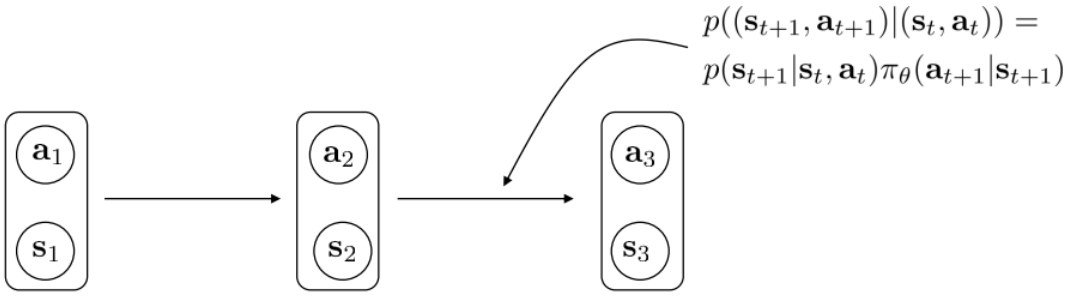
Our goal is to find the the best parameters which maximize the expectation of reward.
In policy gradient based algorithms, the expected reward gradient is computed directly w.r.t these parameters.

$$\underbrace{p_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)}_{\pi_\theta(\tau)} = p(\mathbf{s}_1) \prod_{t=1}^{T} \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\theta^\star = \arg\max_\theta E_{\tau \sim p_\theta(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

Value based reinforcement learning:
In value based learning, we have values associated with each:

1. State and action(q learning): Q (s, a)
   We club state and action into a single node and we have node transition probability across each nodes.



Training process:

1. We fit the the model with (s,a) as X and it's expected marginal value as Y
2. We update the policy to most likely pick valuable actions.
   $\Pi(a_t \mid o_t) = \text{argmax}_a\, Q(s,a)$

   $Q(s,a) = \Sigma E\,[\text{reward}(s_{t+!},a_{t+!}) \mid s_t\ a_t]$

2. State: V(s)
   Everything is same as Q learning except that we use only values of the state.

   $V(s) = \Sigma E\,[\text{reward}(s_{t+!},a_{t+!}) \mid s_t]$

Experiments performed using gym libraries:

Using gym, we can now have an agent interacting with the environment and the environment provides the observation and reward.

In reality we might have to compute/sense observation using CNNs +LSTMs

1. Game configurations:
   Mountain Car
   The car needs to build the momentum to climb the hill.
   Input: move left or right [-1, +1]
   Observations: car's position, car's velocity
   Reward: given by the API
   Done: if the game is completed

   RL algorithms used:
   Policy gradient algorithm

   Modelling part:
   The policy is modeled as a normal distribution. Whose parameters (mean, variance) were learned to successfully take right actions to finish the game.

   Interaction:
   Make random perturbations to actions and record the rewards and pick the best policy model.

   Number of episodes played: 300

References:

1. http://rail.eecs.berkeley.edu/deeprlcourse/
2. https://gym.openai.com/envs/MountainCar-v0/
3. http://www.deeplearningbook.org/
4. https://sites.ualberta.ca/~szepesva/RLBook.html