* **Explain how the cartpole problem can be solved using the REINFORCE algorithm**. Consider using pseudocode, UML, diagrams, or flowcharts to help illustrate your solution.  
  The REINFORCE algorithm solves the cartpole problem by normalizing the rewards, the steps are the number of steps the pole is balanced before it falls. As the agent starts to balance the pole longer the more the negative rewards are removed thus over the course of time leading to a higher probability. (Yoon, Deriving Policy Gradients and Implementing REINFORCE, 2018)

In the case of the cartpole, over the course of time the AI will start to determine the probability of moving left or right given the current state. So, in other words, if the AI determines that moving left is 70% successful and continues to move left with greater and greater success the probability of moving right will get lower and lower. (Yoon, Deriving Policy Gradients and Implementing REINFORCE, 2018)

* **Explain how the cartpole problem can be solved using the A2C algorithm**. Consider using pseudocode, UML, diagrams, or flowcharts to help illustrate your solution.  
  So in the case of the cartpole problem, being A2C is not asynchronous, after each step both the actor and critic are updated. In this case a step would be the movement of the Y-axis. Out actor would present the critic with the action and the critic would then return the value. The policy gradient and advantage would be updated for the actor and the means squared error would be minimized for the critic.  
    
  So, every time the agent moves the pole the critic would return the probability of a success or fail. The parameters would then be updated resulting in the changes to the policy gradient. This would then lead to the decision the agent would make during the next step. If the result would be negative then the agent would reduce the probability resulting in the agent making an action that would lean towards a higher probability.
* **Explain how policy gradient approaches differ from value-based approaches, such as Q-learning**.

Value-based techniques is finding the maximum amount of reward that can be obtained by the state of the environment. This is accomplished by finding or approximating the value function and the extract the policy.

Policy-based throws out the value part and find the policy directly, no middleman. It uses an optimization problem. There is a policy with parameters that outputs a probability distribution over actions. In everyday terms, for every episode which results in a positive reward, the algorithm increases the probability of the actions to be used in the future. The same thing happens with negative rewards and over time are weeded out resulting in positive results being more likely. (Karagiannakos, The idea behind Actor-Critics and how A2C and A3C improve them, 2018)

The difference between the two is how the reward is determined. They both have advantages and disadvantages, scenarios in which they are better suited than the other and how they are developed is very different.

* **Explain how actor-critic approaches differ from value- and policy-based approaches**.

Value- and policy-based approaches differ from actor-critic in fact how Reinforcement Learning methods learn. Value- and Policy-based approaches uses two different functions in which one or the other is used. Value-based approaches maps an action to a value, the higher the value, the better the action. Q learning uses value-based functions to learn. Policy-based algorithms tries to find the optimal policy without having to use a middleman.

(Karagiannakos, The idea behind Actor-Critics and how A2C and A3C improve them, 2018)

The actor-critic approach takes value- and policy-based approach and merges them together. It takes the best characteristics of both algorithm types and removes the worst characteristics. It splits the model into two: one for computing an action and the other to produce values. The theory behind this is the actor takes a state and outputs an action, the critic then takes that action and outputs a value. (Karagiannakos, The idea behind Actor-Critics and how A2C and A3C improve them, 2018). Putting it everyday terms, the actor performs an action and the critic tells it how it did with that move. From these moves the model is trained.

.

# References

Karagiannakos, S. (2018, Nov 17). *The idea behind Actor-Critics and how A2C and A3C improve them*. Retrieved from theaisummer.com: https://theaisummer.com/Actor\_critics/

Karagiannakos, S. (2018, Nov 01). *Unravel Policy Gradients and REINFORCE*. Retrieved from thisaisummer.com: https://theaisummer.com/Policy-Gradients/

Yoon, C. (2018, Dec 29). *Deriving Policy Gradients and Implementing REINFORCE*. Retrieved from medium.com: https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63

Yoon, C. (2019, Feb 05). *Understanding Actor Critic Methods and A2C*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f