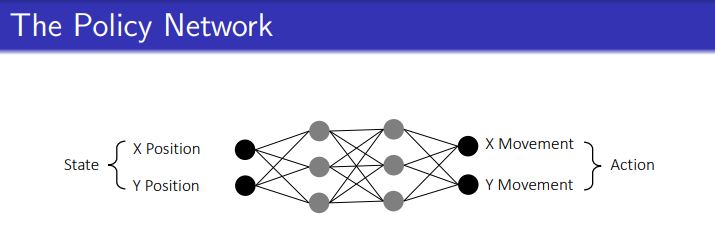
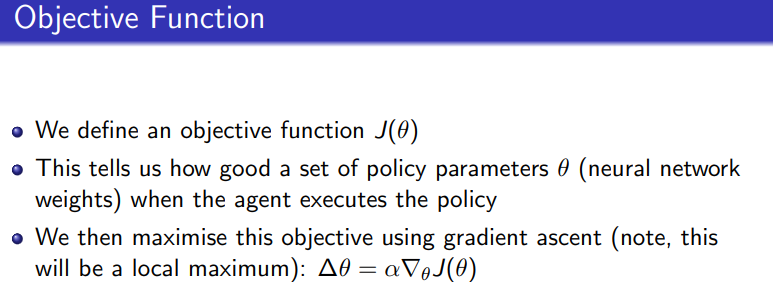
Answers are given but for discussion if wanted

1. C The state, from the early slides 
2. B standard gradient (ascent) Change of objection function w.r.t. network params. Not policy output action because the network is trained on the expected return for each trajectory
3. D product of state probabilities times the probability of choosing that action
4. B Node Y. We have value of node = mean + 2 x sqrt(ln(t)/n). The lecture gives constant 2
   1. Node X: mean 1/3 + 2 sqrt(2.3/3) = 0.33 + 0.88 x 2
   2. Node y: mean 0 + 2 sqrt(2.3/2) = 1.07 x 2
   3. Node z: mean 3/5 + 2 sqrt(2.3/5) = 0.6 + 0.68 x 2

So as long as the constant C isn’t too small (2 in slides), node Y has the highest value

1. C. MCTS requires more nodes eventually, CEM is a distribution so doesn’t need more memory
2. E. Pretty straightforward (straight hehe) relation between x+1 and x