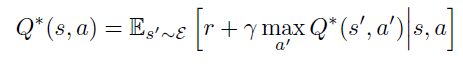
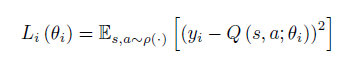
Vision과 Speech 같은 고차원 입력을 이용해 직접적으로 agent 통제를 학습하는 것은 강화학습의 도전 과제였다. 이 논문 이전 대부분의 강화학습 모형은 수작업 특징에 의존해왔다. 여러 지도 및 비지도 학습 모델들이 발달해 왔지만 강화학습에서 기존의 딥러닝 모델들의 기술들을 사용은 어려웠다. 왜냐하면 첫째로 기존의 모델들은 label이 된 학습데이터를 이용한 반면 강화학습은 reward를 이용해 학습해야 하는데 양질의 reward를 얻는 것이 어려웠다. 또한 기존 모델들의 데이터는 독립인 것을 가정하는 반면에 강화학습에 사용되는 데이터는 높은 상관관계를 갖고, 새로운 행동에 따라 데이터의 분포가 달라진다. 이러한 문제들을 해결하기 위해 이 논문에서는 Q-learning 알고리즘을 사용해서 가중치를 업데이트하고, experience reply mechanism을 사용했다.

Agent가 높은 reward를 받는 방향으로 학습을 진행시켜야 한다. 이때 reward는 즉시 보상과 미래의 보상들의 합을 이용해서 정의한다. 이때 미래의 보상들은 time-step에 따라 할인율(γ:gamma)를 등비수열 꼴로 곱해줘서 구한다. Q\*(s,a)(optimal action-value function)를 상태 s와 행동 a 일 때 정책 pi에 대해 보상 기대값의 최대값으로 한다.



Q\*(s’,a’)을 이용해 Q(s,a)를 표현하는 점화식이 되고, 이는 bellman equation을 따른다. 이를 가치 반복 알고리즘을 이용해 Q\_i가 Q\*로 수렴하게 해서 최적의 값을 찾는다. 하지만 이는 실용적이지 못하다. 해당 논문에서는 Q-network로써 신경망 함수를 approximator로 사용해서 loss function을 정의하고, 확률적 경사 하강법을 사용하여 최적화를 수행한다.





θ를 학습해가면서 Q\*(optimal action-value function)를 찾는다.

Environment의 추정치를 명시적으로 구성하지 않고, emulator environment의 샘플들을 사용하여 직접 강화학습을 해결하는 model-free 알고리즘이다. 또한, epsilon-greedy 알고리즘을 활용해 특정 확률에 따라 랜덤 action을 하거나 greedy strategy를 이용한 action(reward를 가장 크게 하는 action)을 하는 off-policy이다. No-policy를 통해 Exploration(탐사)과 Exploitation(개척)을 같이 수행해 나간다.

따라서 정리하자면 model-free하고, no-policy한 알고리즘으로 확률적 경사 하강법을 이용해 최적의 θ를 찾아서 Q\*를 찾아나간다.