남이 써놓은 자료를 계속 보자니, 왜 이소리가 여기서 나오는건지 이해가 잘 안 되는 부분이 있는 것 같다. 영어 논문을 복사해서 처음부터 끝까지 해석해보자.

제목. **Policy Gradient Methods** for Reinforcement Learning with **Function Approximation**

function approximation은 실제 함수를 여러가지 샘플을 뽑아서 근사값으로 가정한다는 말일 것이다. 그렇다면 어떤 함수를 approximate한다는 것일까? state-action value function? state-value function?

Abstract

Function approximation is essential to reinforcement learning, but the standard approach of approximating a value function and determining a policy from it has so far proven theoretically intractable.

Fuctional approximation은 강화학습에 있어서 필수적이지만, 일반적인 접근법으로 value-function을 근사하고 이를 바탕으로 policy를 결정하는 것은 이론적으로 난해한 것으로 지금까지 증명되었다.

In this paper we explore an alternative approach in which the policy is explicitly represented by its own function approximator, independent of the value function, and is updated according to the gradient of expected reward with respect to the policy parameters.

이 논문에서 우리는 새로운 어프로치를 발견했다 – policy가 명시적으로 그 자신의 function approximator로 표현되어 value function과 무관하고, 그리고 정책 파라미터에 관한 기대 reward의 그래디언트에 의해 업데이트 되는.

Williams's REINFORCE method and actor-critic methods are examples of this approach. Our main new result is to show that the gradient can be written in a form suitable for estimation from experience aided by an approximate action-value or advantage function.

윌리엄의 REINFORCE 방식과 actor-critic 방식이 이 접근법의 예가 될 수 있다. 우리의 주된 새로운 결론은 그래디언트가 approximate action-value 또는 advantage function에 의해 만들어진 experienc로부터 추측되는데 알맞은 형태로 쓰여질 수 있음을 보여주기 위함이다.

윌리엄의 Reinforce 방식?

앞서 유도해내었던 이 식, monte carlo gradient estimation을 나타내는 이 식을 이용하는 방식이다.

Using this result, we prove for the first time that a version of policy iteration with arbitrary differentiable function approximation is convergent to a locally optimal policy.