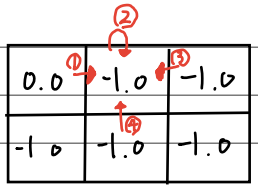
Model-based

Dynamic Programming (동적계획법)

- 모델기반 환경에서 사용하는 MDP 해결방법
- 정책평가와 정책제어를 반복하며 optimal policy를 찾는다.

이러한 과정을 Value Iteration

- 정책평가 (Policy evaluation)
  - (1) 정책을 고정하고 처음 타임스텝과 뒤따르는 스텝들에 대한 가치를 각각 구해서 합산
  - (2) 마지막 타임스텝까지 반복 수행
  - (3) 현재 타임스탭의 가치를 업데이트



$-1.0+(0.0\times0.25)+(-1.0\times0.25)+(-1.0\times0.25)=-1.75$

- 정책제어 (Policy Iteration)
  - (1) 계산한 가치함수를 사용하여 탐욕적으로 정책을 선택해서 현재 정책을 갱신

0	-14	-20
-14	-18	-20
-20	-20	-18

$\pi = \begin{cases} 1 & \text{if } a = \text{argmax} \\ 0 & \text{otherwise} \end{cases}$

※ 무한대로 evaluation (평가)를 진행할 필요 없이 Iteration 을 세 번만 해도 optimal Policy를 구할 수 있다.