State space S: 10 space_subdivisions, 2D grid $\Rightarrow n_{states} = space\_subdivisions^2 =$ 100

Action space A: $A = \{ left, right, down, up, stay \}$

Hyperparameters α, γ scenarios:

(1) Not improving/learning:

$\alpha = (0; 0.1)$; $\gamma = [0; 1]$

(2) High variance but fast learning:

$\alpha = [0.7; 1]$; $\gamma = (0; 1]$

(3) Low variance and high long-term return:

$\alpha = [0.1; 0.4]$; $\gamma = [0.7; 1]$

(4) High variance and high long-term return:

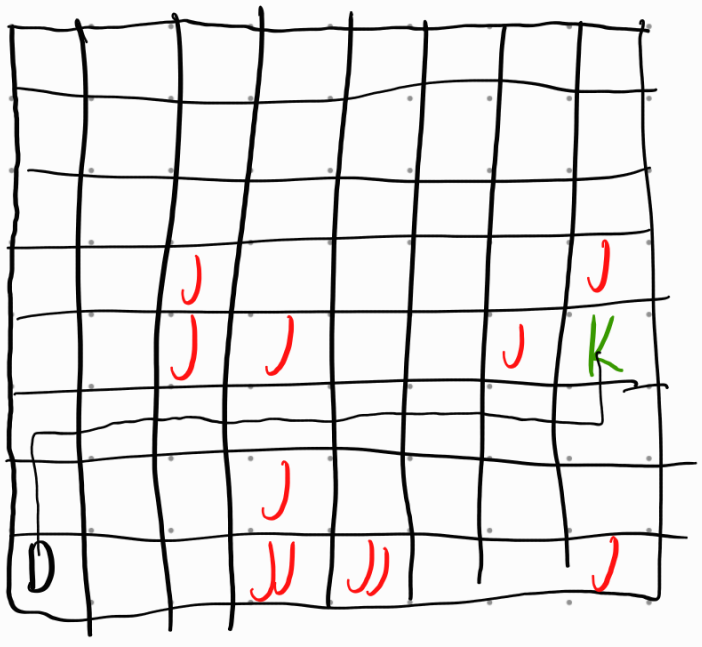
$\alpha = [0.7; 1]$; $\gamma = [0.7; 1]$

Long-term return of the optimal policy in 323.yml:

Reward for golden fish: 50.

Reward for step: -2.

Reward Jelly: -10.



$\Rightarrow R_{opt} = 50 - 2 \cdot \text{shortest path}$

$= 50 - 2 \cdot 10 = 30!$