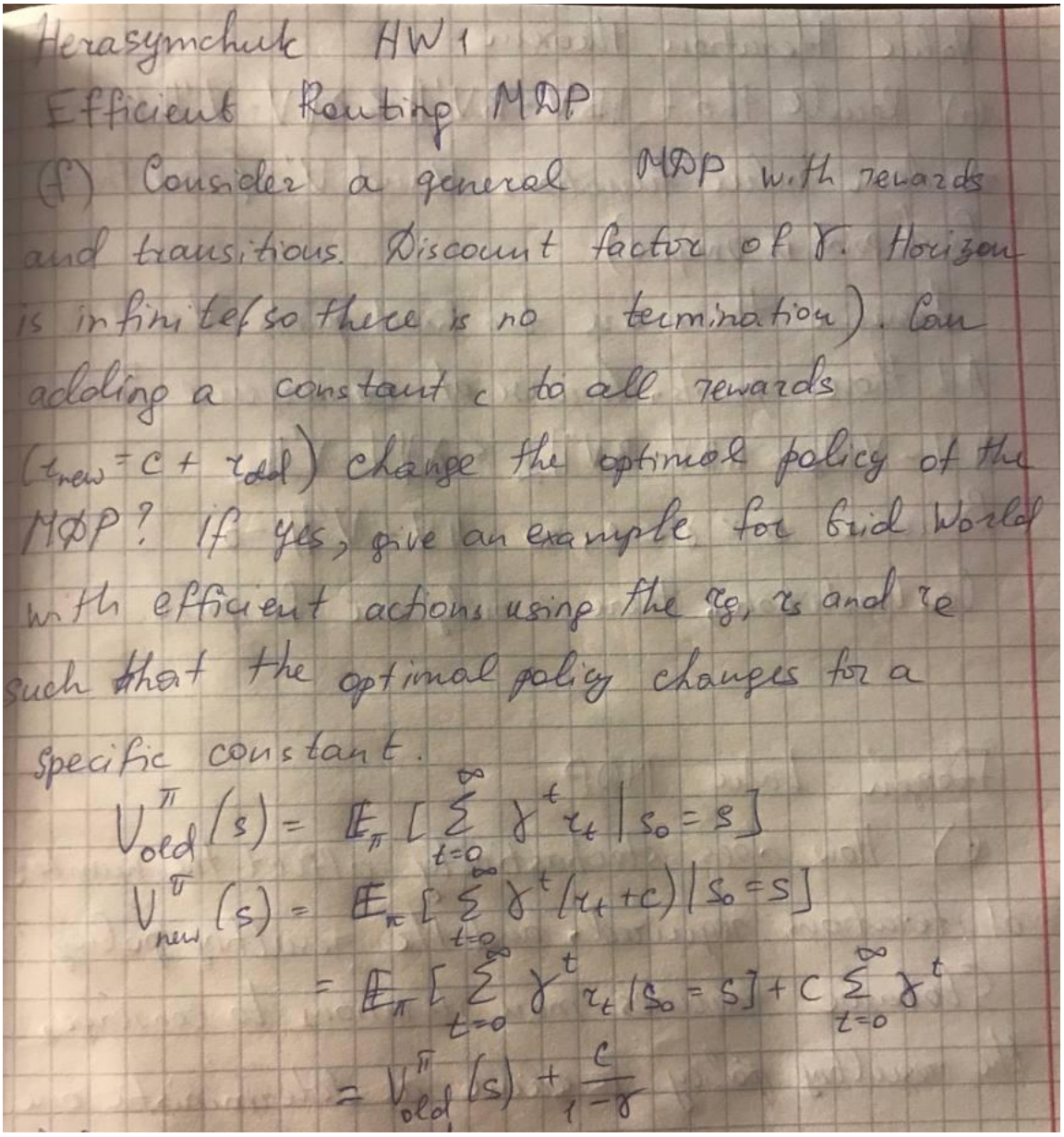
Daria Herasymchuk. Assignment 1.



Herasymchuk HW 1

Efficient Routing MDP

(f) Consider a general MDP with rewards and transitions. Discount factor of γ. Horizon is infinite (so there is no termination). Can adding a constant c to all rewards ($r_{new} = c + r_{old}$) change the optimal policy of the MDP? If yes, give an example for Grid World with efficient actions using the r_g, r_s and r_e such that the optimal policy changes for a specific constant.

$$V^{\pi}_{old}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right]$$

$$V^{\pi}_{new}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t (r_t + c) \mid s_0 = s\right]$$

$$= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right] + c \sum_{t=0}^{\infty} \gamma^t$$

$$= V^{\pi}_{old}(s) + \frac{c}{1-\gamma}$$

Value Iteration theorem

(a) Recall that $\|BV - BV'\| \leq \gamma \|V - V'\|$ for two random value functions V and V' (show that fixed point is unique)

Assume there are two fixed points V and V'

At those points, because of convergence, $BV = V$

$\|V - V'\|_\infty = \|BV - BV'\|_\infty \leq \gamma \|V - V'\|_\infty$

Because $0 < \gamma < 1$, so $\|V - V'\| = 0$

Contradiction, there is only one fix point.

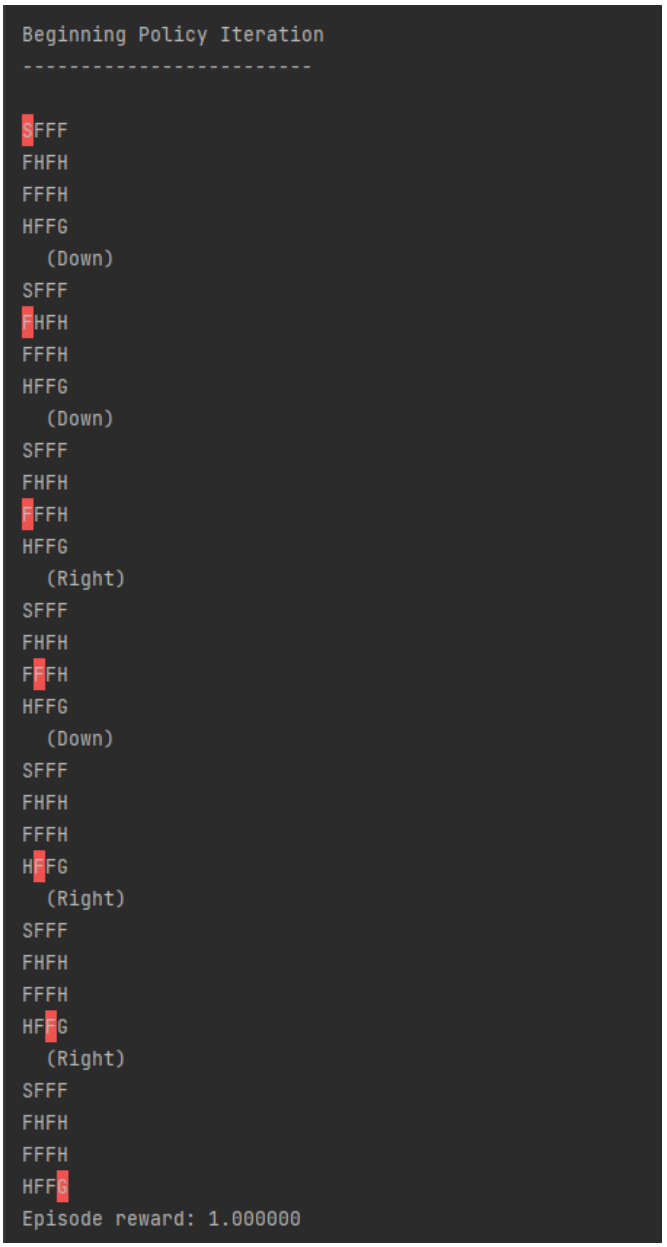
Frozen Lake MDP

(c) How does stochasticity affect the number of iterations required, and the resulting policy?

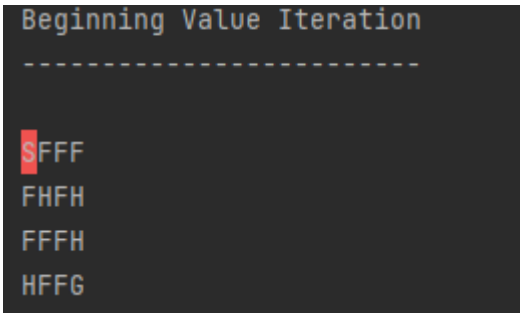
Stochastic requires more iterations to converge, the resulting policy is same as deterministic one

a) Read through vi_and_pi.py and implement policy_evaluation, policy_improvement and policy_iteration. The stopping tolerance (defined as maxs |Vold(s) − Vnew(s)|) is tol = 10−3 . Use γ = 0.9. Return the optimal value function and the optimal policy

policy_evaluation, policy_improvement were implemented in function policy_iteration and the result is



b) Implement value_iteration in vi_and_pi.py. The stopping tolerance is tol = 10^{-3} . Use $\gamma = 0.9$. Return the optimal value function and the optimal policy.



```
SFFF
FHFH
FFFH
HFFG
  (Down)
SFFF
FHFH
FFFH
HFFG
  (Right)
SFFF
FHFH
FFFH
HFFG
  (Down)
SFFF
FHFH
FFFH
HFFG
  (Right)
SFFF
FHFH
FFFH
HFFG
  (Right)
SFFF
FHFH
FFFH
HFFG
Episode reward: 1.000000
```