“All have their worth,” said Yavanna,

“and each contributes to the worth of the others”.

J.R.R. Tolkien, *The Silmarillion*

CMPUT 365
Introduction to RL

Marlos C. Machado

Class 13/35

Plan

- Dynamic programming
    - Finally, a solution method (albeit limited)!

Reminder I

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks.

You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

The deadlines in the public session **do not align** with the deadlines in Coursera.

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

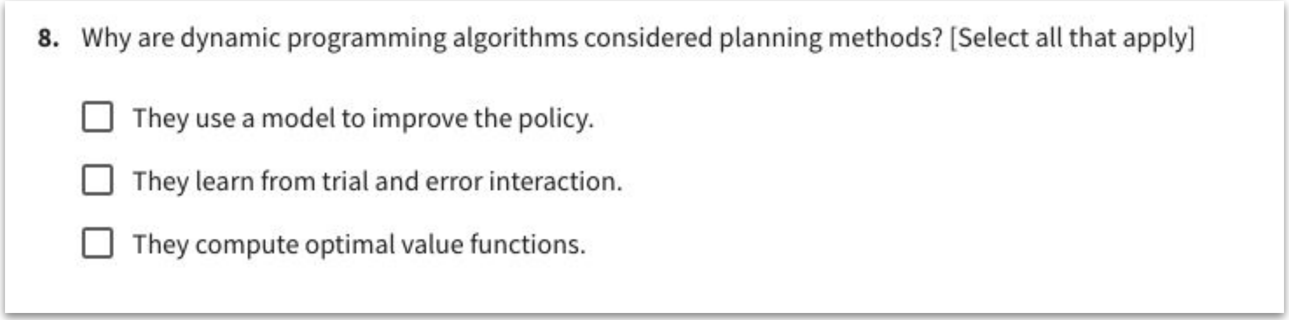
Reminder II

- Progr. assign. for Coursera's Dynamic Programming module is due Wednesday.
  Fundamentals of RL: Dynamic Programming – Week 4.
- Midterm is next Wednesday, after Thanksgiving.
  - There will be an attendance sheet.
  - You need to have your OneCard with you and show it to us.

Please, interrupt me at any time!

https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Quiz Question – Week 4



8. Why are dynamic programming algorithms considered planning methods? [Select all that apply]

- □ They use a model to improve the policy.
- □ They learn from trial and error interaction.
- □ They compute optimal value functions.

Policy Improvement

Given a value function for a policy π, how can we get a better policy π'?

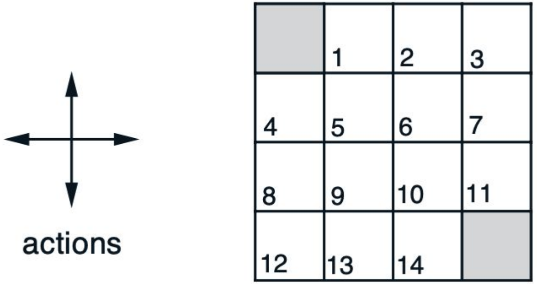
We already know how good policy π is, what if we acted differently now? What if instead of selecting action π(s) we selected action a ≠ π(s), but then we followed π?

We know the value of doing that!

$$\begin{aligned} q_\pi(s,a) &\doteq \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \sum_{s',r} p(s',r \mid s,a)\Big[r + \gamma v_\pi(s')\Big]. \end{aligned}$$

If this new action is better, in general this new policy is better overall

Policy Improvement – Intuition

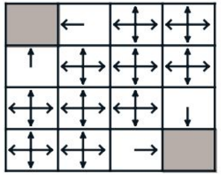


$R_t = -1$
on all transitions

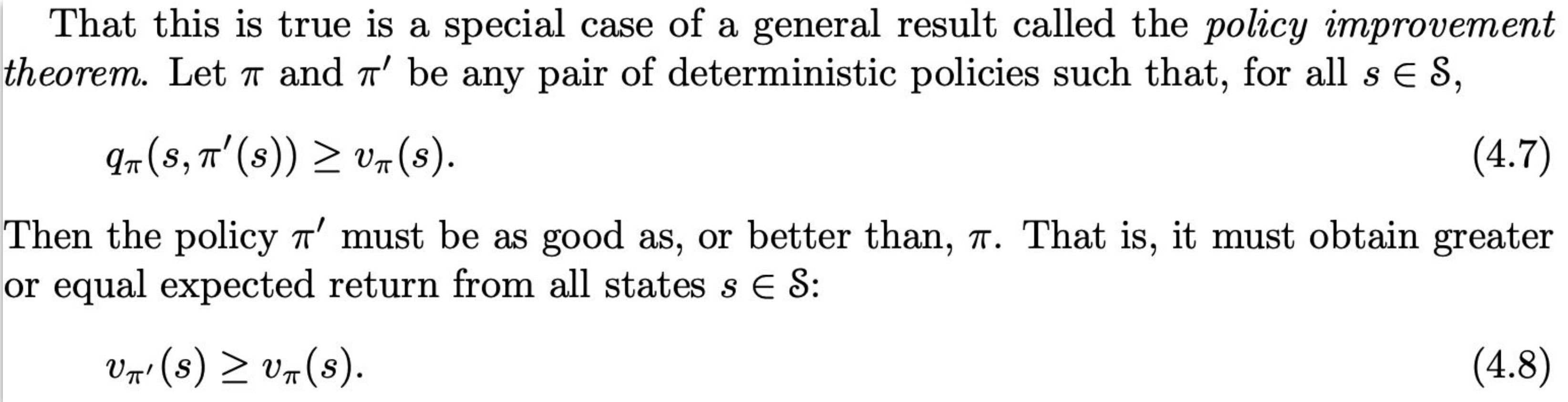
v_k for the random policy

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



Policy Improvement Theorem



That this is true is a special case of a general result called the *policy improvement theorem*. Let π and π' be any pair of deterministic policies such that, for all $s \in \mathcal{S}$,

$$q_\pi(s, \pi'(s)) \geq v_\pi(s). \tag{4.7}$$

Then the policy π' must be as good as, or better than, π. That is, it must obtain greater or equal expected return from all states $s \in \mathcal{S}$:

$$v_{\pi'}(s) \geq v_\pi(s). \tag{4.8}$$

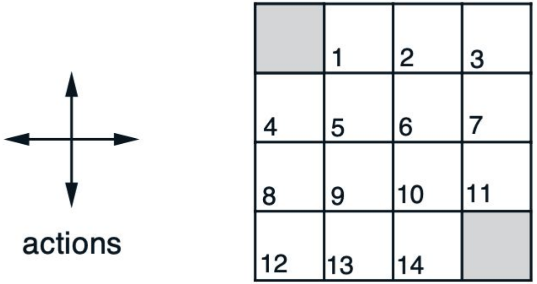
A more aggressive update

Instead of doing it for a particular action in a single state, we can consider changes at *all* states and to *all* possible actions.

$$
\begin{aligned}
\pi'(s) &\doteq \operatorname*{arg\,max}_a q_\pi(s,a) \\
&= \operatorname*{arg\,max}_a \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \operatorname*{arg\,max}_a \sum_{s',r} p(s',r \mid s,a)\Big[r + \gamma v_\pi(s')\Big],
\end{aligned}
$$

This is called *policy improvement*. And eventually it converges to the optimal policy.

Policy Improvement – Intuition

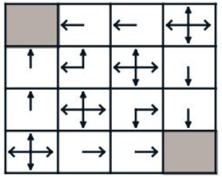


$R_t = -1$
on all transitions

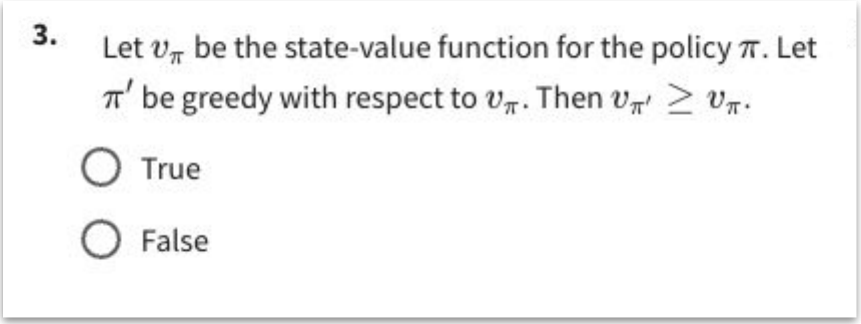
v_k for the random policy

$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0



Quiz Question – Week 4

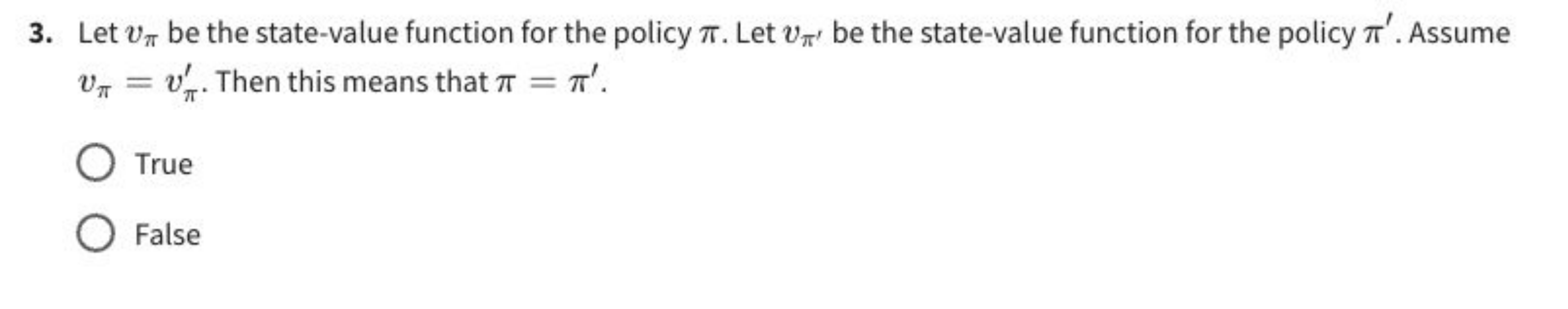


3. Let v_π be the state-value function for the policy π. Let π' be greedy with respect to v_π. Then $v_{\pi'} \geq v_\pi$.

○ True

○ False

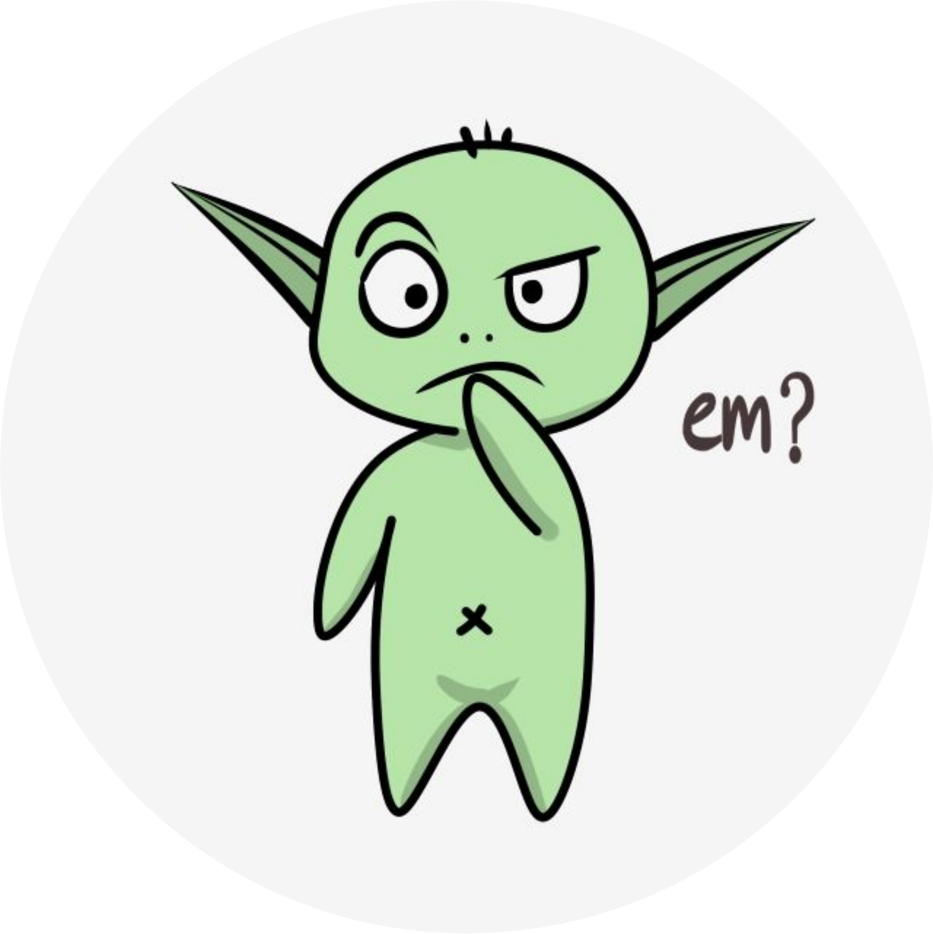
Quiz Question – Week 4



3. Let v_π be the state-value function for the policy π. Let $v_{\pi'}$ be the state-value function for the policy π'. Assume $v_\pi = v'_\pi$. Then this means that $\pi = \pi'$.

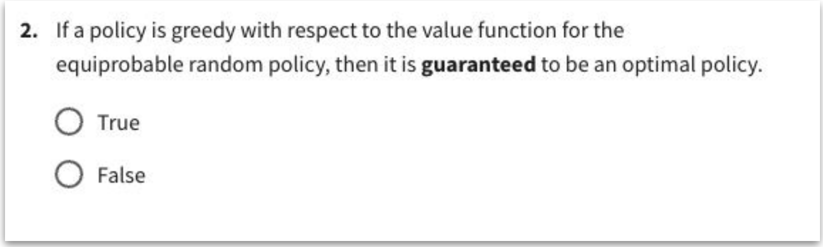
○ True

○ False



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Quiz Question – Week 4

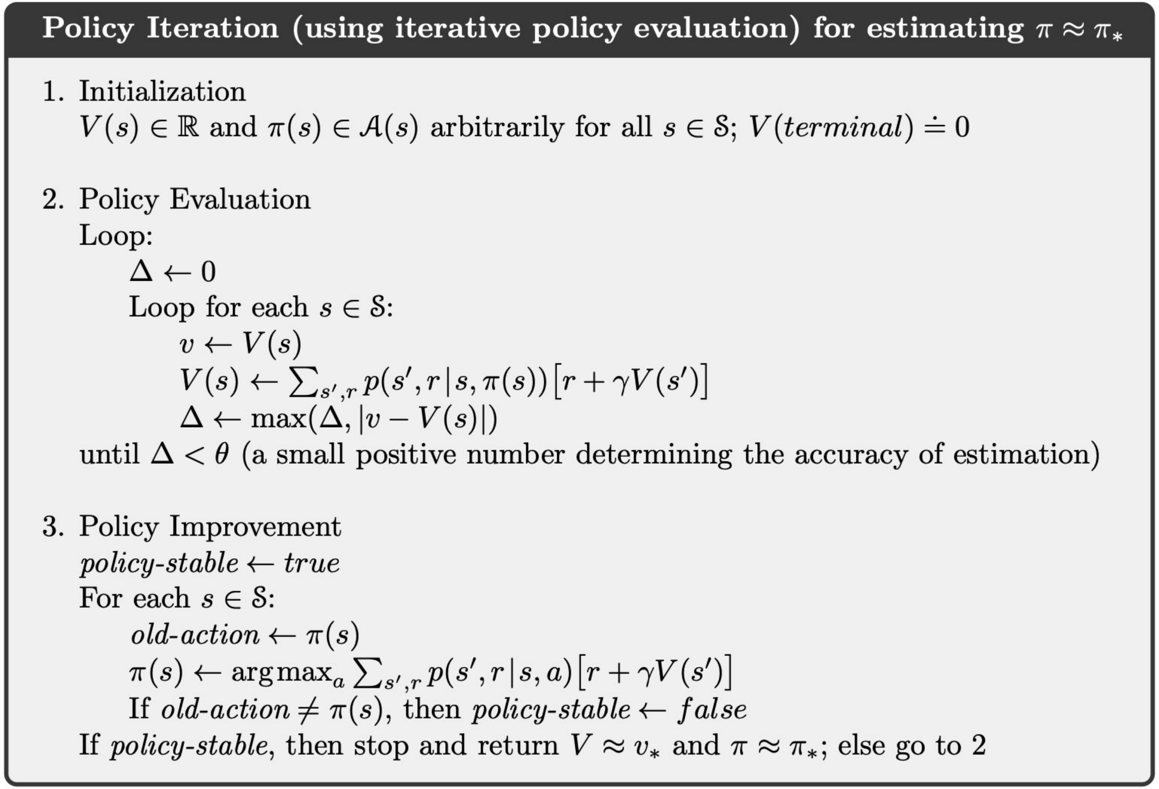


2. If a policy is greedy with respect to the value function for the equiprobable random policy, then it is **guaranteed** to be an optimal policy.

○ True

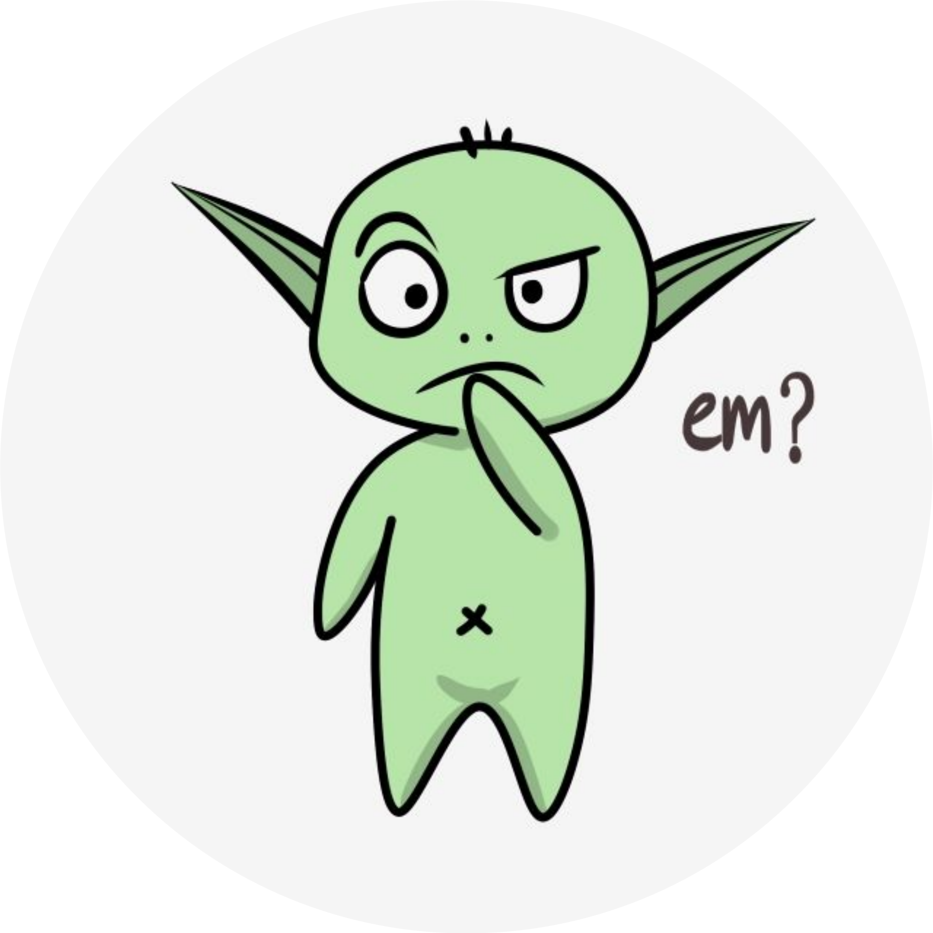
○ False

Policy Iteration: Interleaving Policy Eval. and Improvement



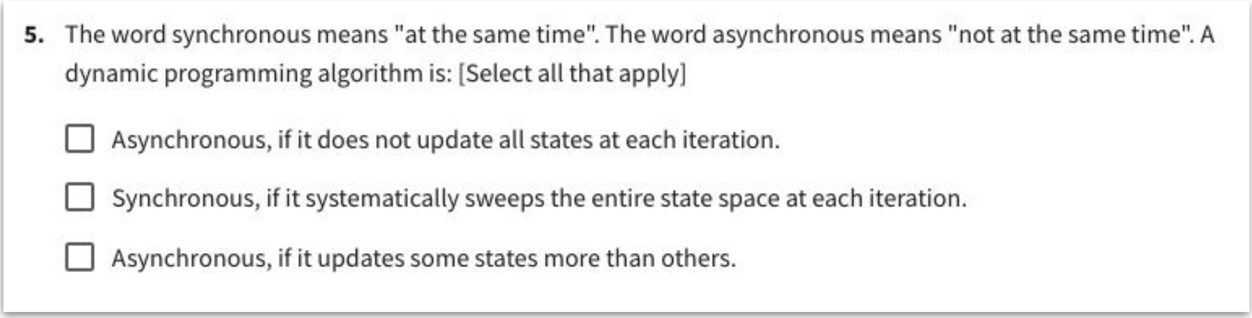
Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

```
1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$; $V(terminal) \doteq 0$

2. Policy Evaluation
   Loop:
       $\Delta \leftarrow 0$
       Loop for each $s \in \mathcal{S}$:
           $v \leftarrow V(s)$
           $V(s) \leftarrow \sum_{s',r} p(s', r \,|\, s, \pi(s))\big[r + \gamma V(s')\big]$
           $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   $policy\text{-}stable \leftarrow true$
   For each $s \in \mathcal{S}$:
       $old\text{-}action \leftarrow \pi(s)$
       $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$
       If $old\text{-}action \neq \pi(s)$, then $policy\text{-}stable \leftarrow false$
   If $policy\text{-}stable$, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2
```



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Quiz Question – Week 4



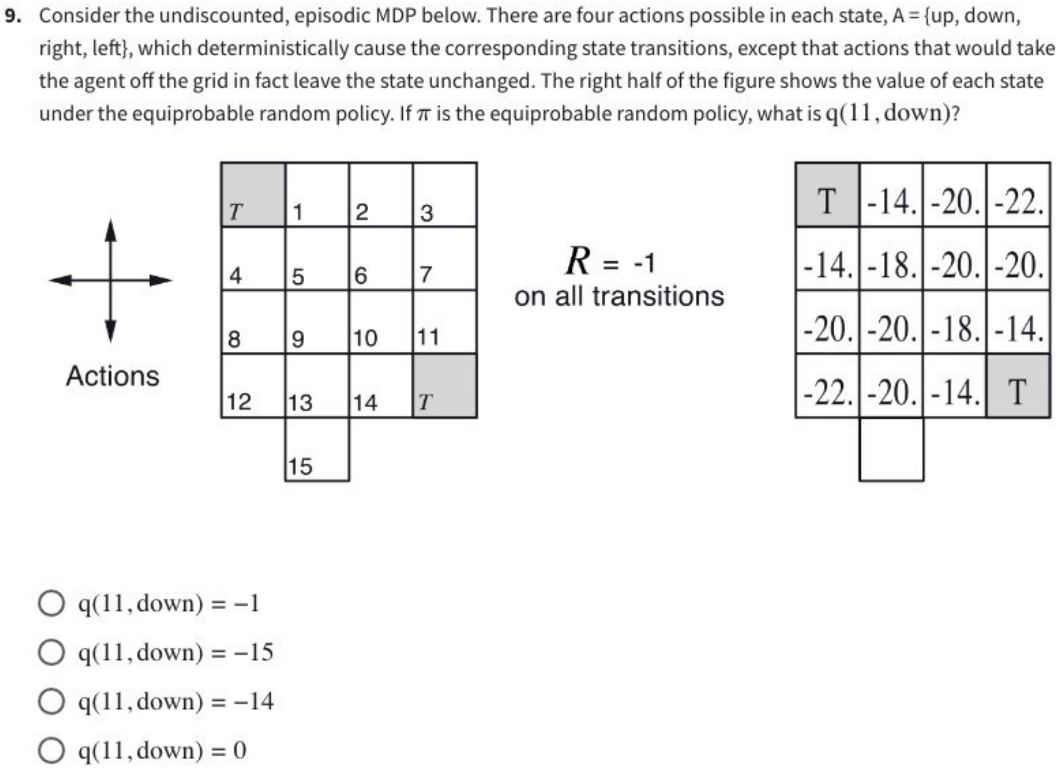
5. The word synchronous means "at the same time". The word asynchronous means "not at the same time". A dynamic programming algorithm is: [Select all that apply]

□ Asynchronous, if it does not update all states at each iteration.

□ Synchronous, if it systematically sweeps the entire state space at each iteration.

□ Asynchronous, if it updates some states more than others.

Quiz Question – Week 4



9. Consider the undiscounted, episodic MDP below. There are four actions possible in each state, A = {up, down, right, left}, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. The right half of the figure shows the value of each state under the equiprobable random policy. If π is the equiprobable random policy, what is $q(11, \text{down})$?

Actions

T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	*T*
	15		

R = -1
on all transitions

T	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	T

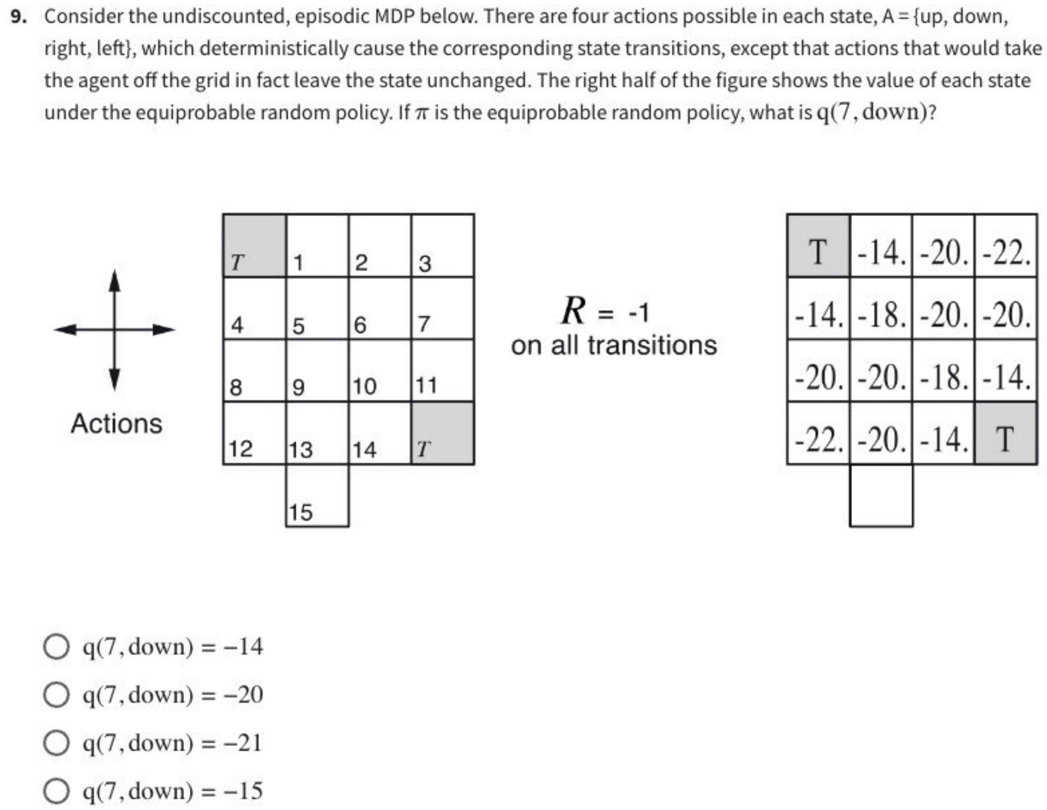
○ $q(11, \text{down}) = -1$

○ $q(11, \text{down}) = -15$

○ $q(11, \text{down}) = -14$

○ $q(11, \text{down}) = 0$

Quiz Question – Week 4



9. Consider the undiscounted, episodic MDP below. There are four actions possible in each state, A = {up, down, right, left}, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. The right half of the figure shows the value of each state under the equiprobable random policy. If π is the equiprobable random policy, what is $q(7, \text{down})$?

T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	*T*
	15		

R = -1
on all transitions

T	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	T

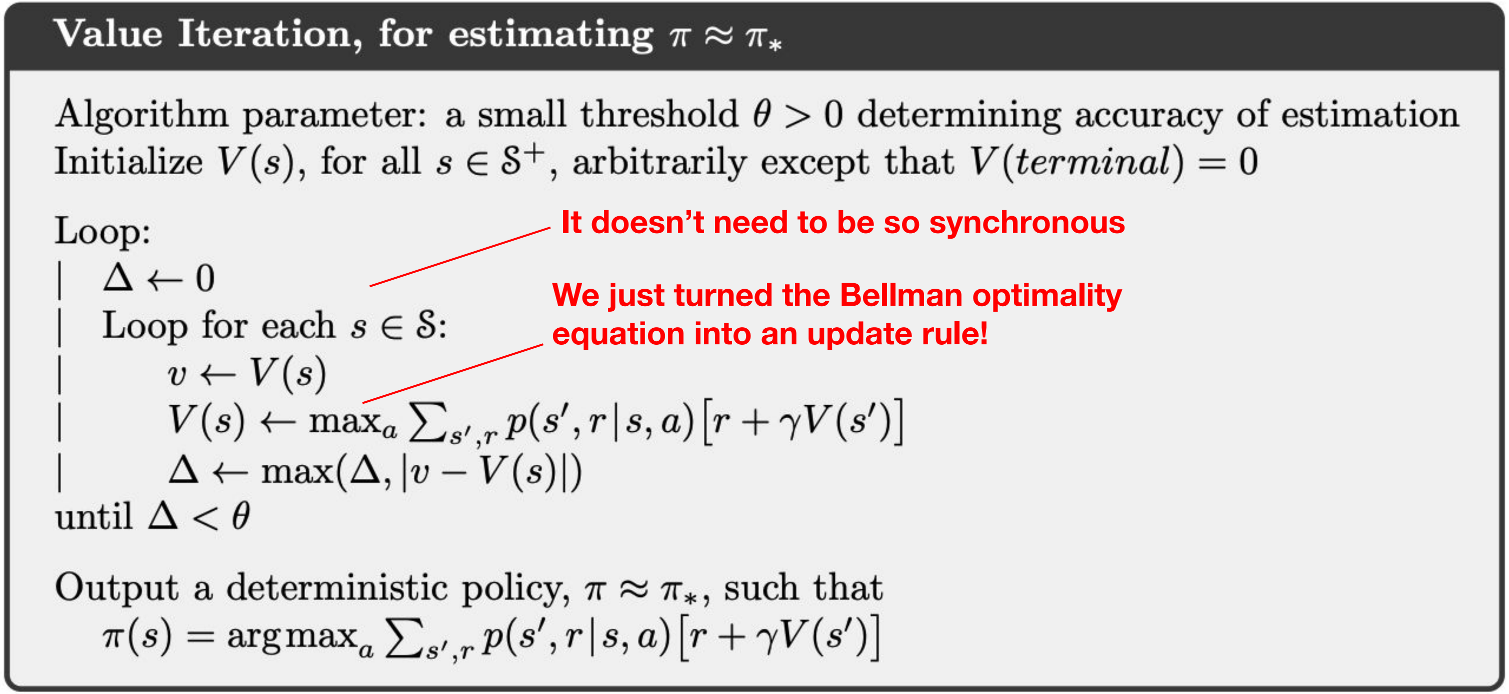
○ $q(7, \text{down}) = -14$

○ $q(7, \text{down}) = -20$

○ $q(7, \text{down}) = -21$

○ $q(7, \text{down}) = -15$

Value Iteration



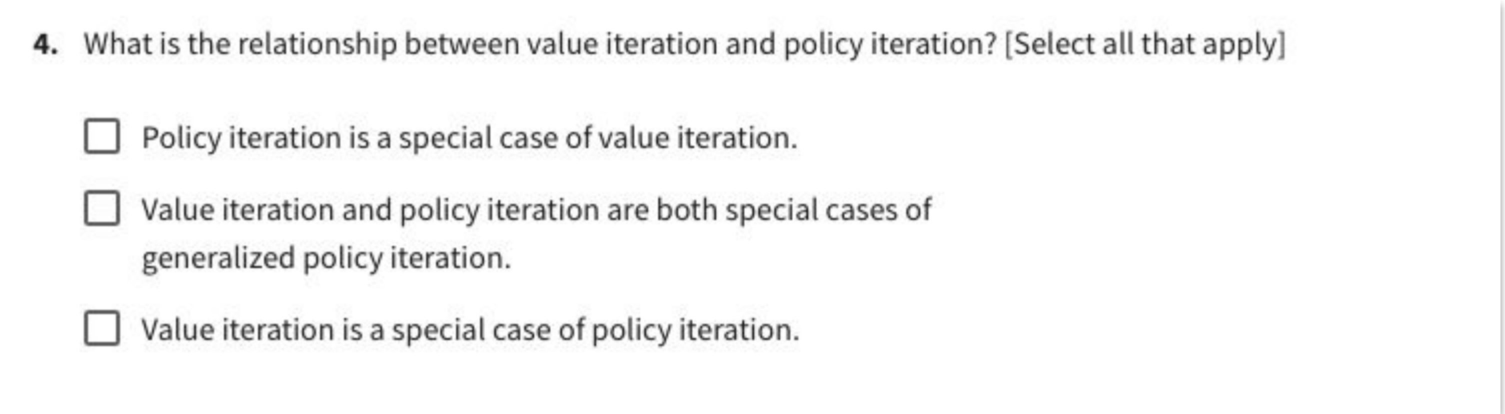
Value Iteration, for estimating $\pi \approx \pi_*$

```
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  | $\Delta \leftarrow 0$
  | Loop for each $s \in \mathcal{S}$:
  |     $v \leftarrow V(s)$
  |     $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
  |     $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
  $\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
```

It doesn’t need to be so synchronous

We just turned the Bellman optimality equation into an update rule!

Quiz Question – Week 4

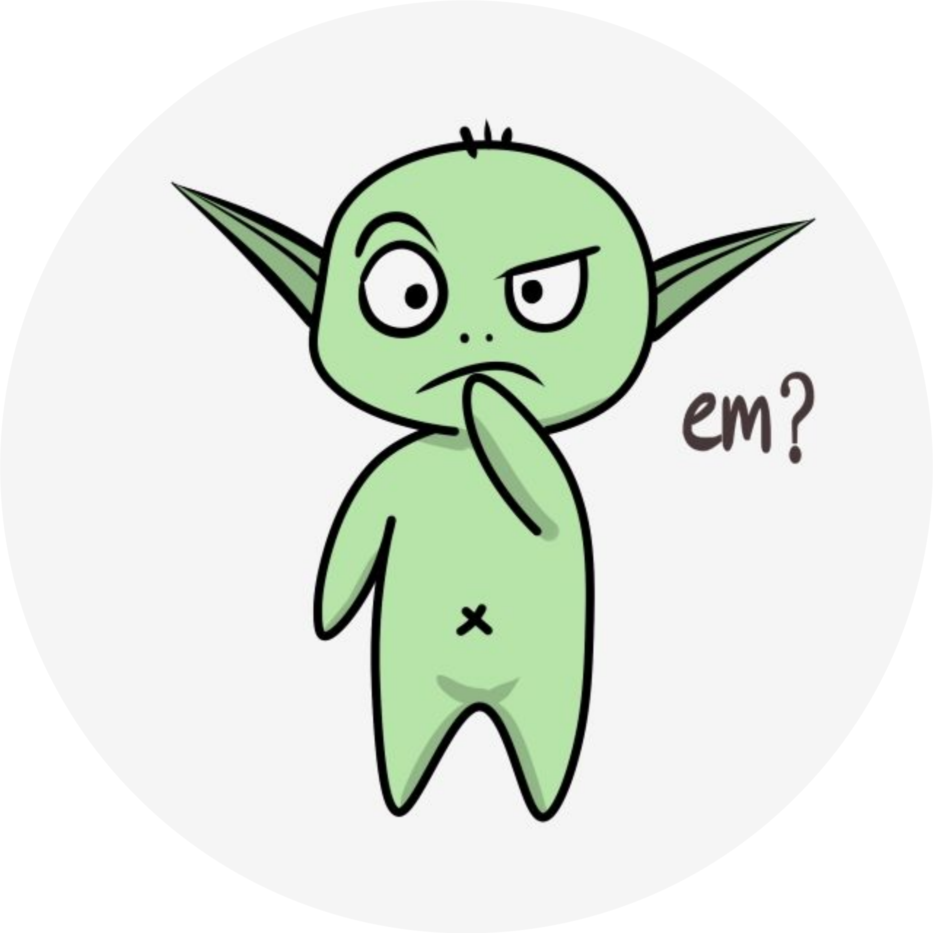


4. What is the relationship between value iteration and policy iteration? [Select all that apply]

□ Policy iteration is a special case of value iteration.

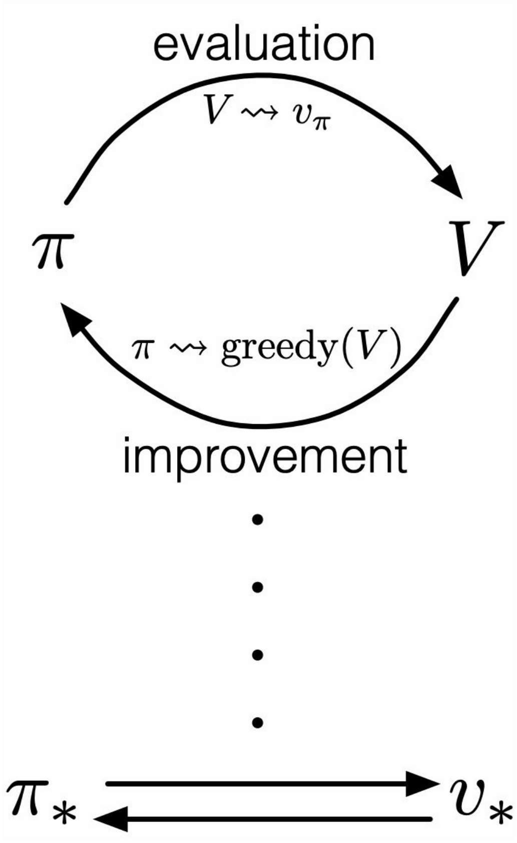
□ Value iteration and policy iteration are both special cases of generalized policy iteration.

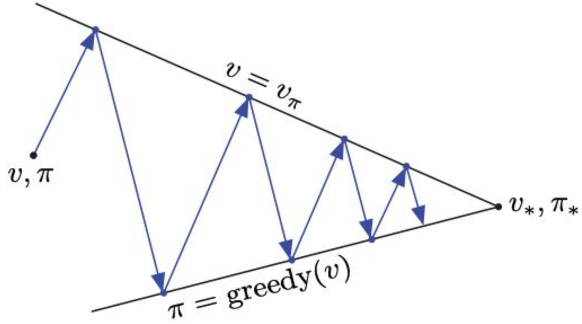
□ Value iteration is a special case of policy iteration.

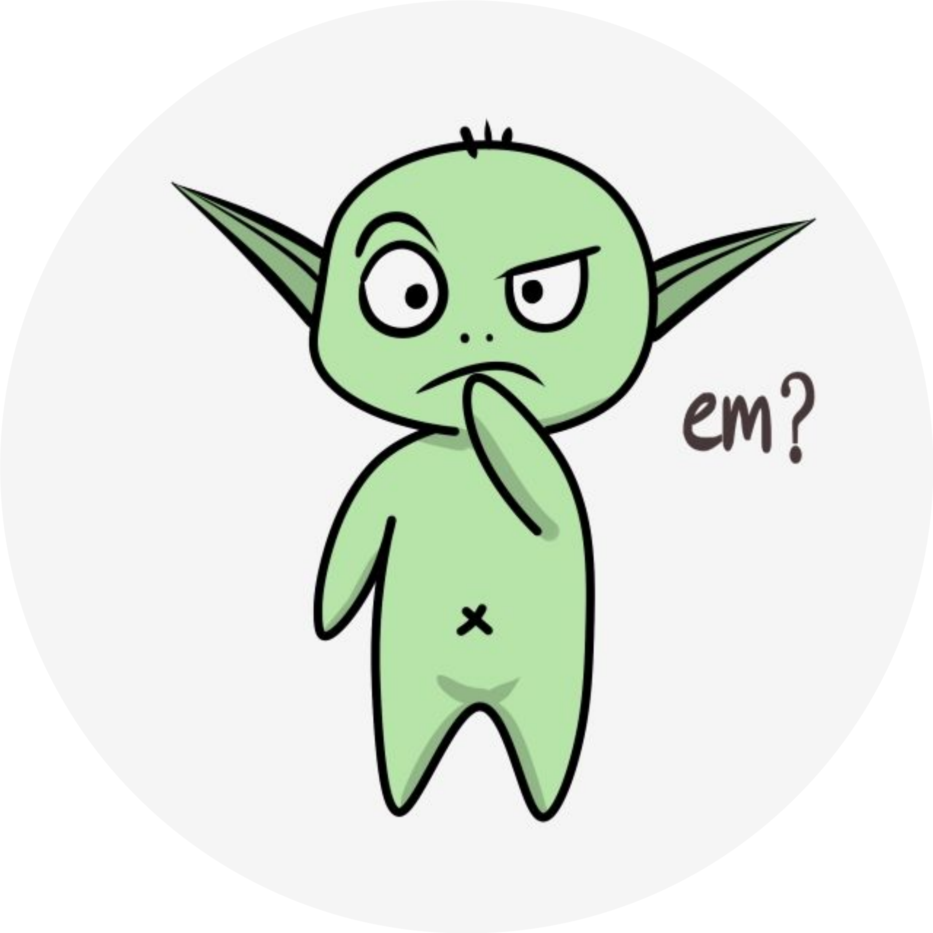


https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Generalized Policy Iteration

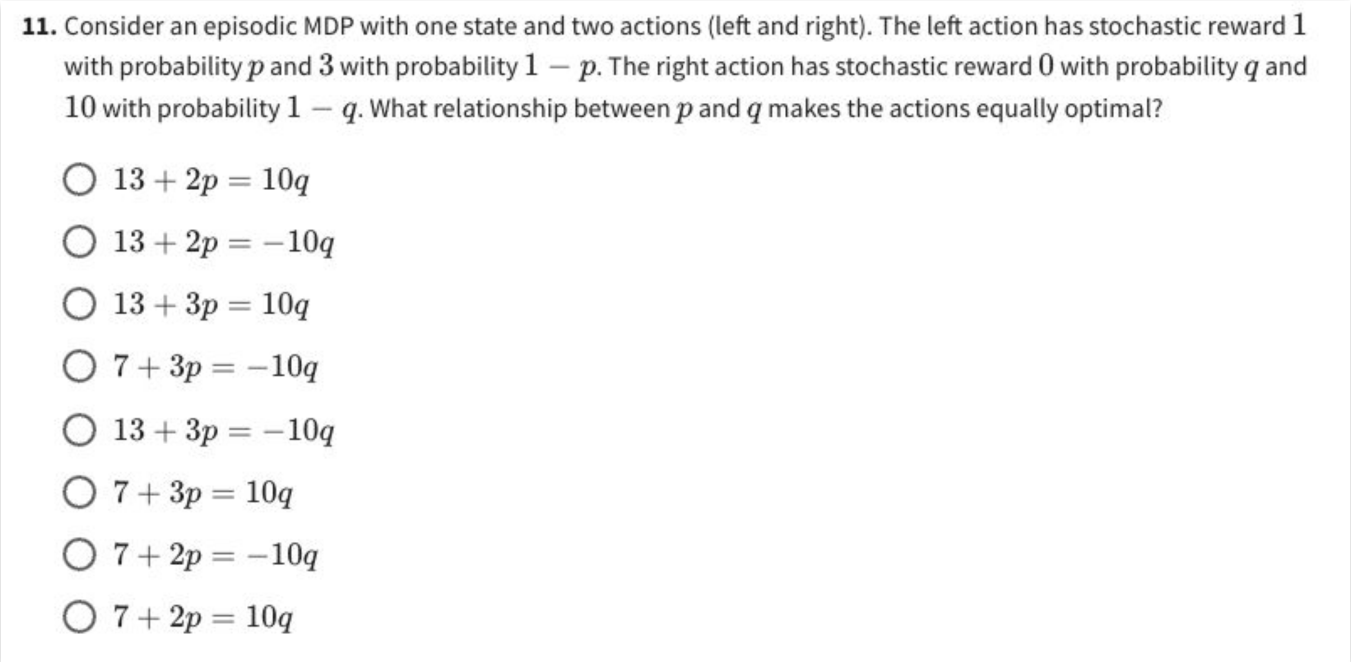

evaluation
$V \rightsquigarrow v_\pi$
π
V
$\pi \rightsquigarrow \mathrm{greedy}(V)$
improvement
π_*
v_*





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Quiz Question – Week 3



11. Consider an episodic MDP with one state and two actions (left and right). The left action has stochastic reward 1 with probability p and 3 with probability $1-p$. The right action has stochastic reward 0 with probability q and 10 with probability $1-q$. What relationship between p and q makes the actions equally optimal?

○ $13+2p=10q$

○ $13+2p=-10q$

○ $13+3p=10q$

○ $7+3p=-10q$

○ $13+3p=-10q$

○ $7+3p=10q$

○ $7+2p=-10q$

○ $7+2p=10q$