C1M3: Q&A and Worksheets

CMPUT 365
Fall 2021

Reminders: Sept 24, 2021

- Will post some extra material on derivation of Bellman Equations and other theory stuff in the schedule

- Practice quiz is due Monday at noon

Plan for today

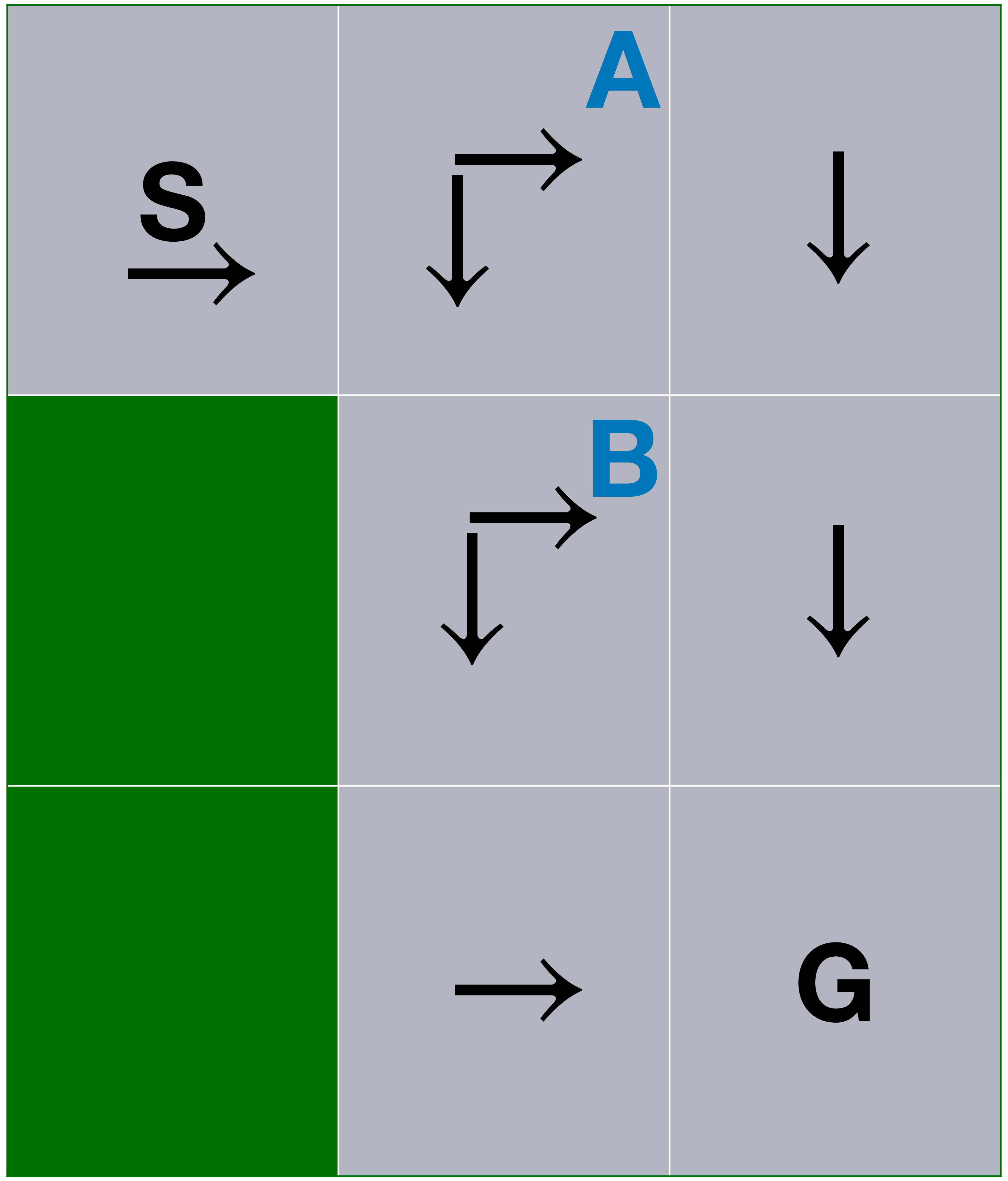
- Class questions
- A fun worksheet question
- Review graded quiz

Stochastic vs Deterministic Policies

- Q: “Is a stochastic policy ever optimal? Or for a policy to be optimal, must it be deterministic?”
    - Yes we can have a stochastic optimal policy. It makes proofs simpler to reason about that there is always at least one deterministic optimal policy
- Q: “If it's deterministic, does it mean the policy is optimal?”
    - No. Example: imagine a maze. Always go up is deterministic but not optimal

Multiple Optimal Policies

- Q: “If there are multiple optimal action at a state, is the probability of picking them always evenly distributed or can there be cases where it is not?”
  - How many unique paths are from **S to G,** deterministically taking optimal actions? **How many deterministic optimal policies are there?**
  - We could define a **stochastic optimal policy** here too:
    - take **down** wp 0.2 and **right** wp 0.8, in state **A**
    - take **down** wp 0.8 and **right** wp 0.2, in state **B**
    - deterministically follows arrows in all other states



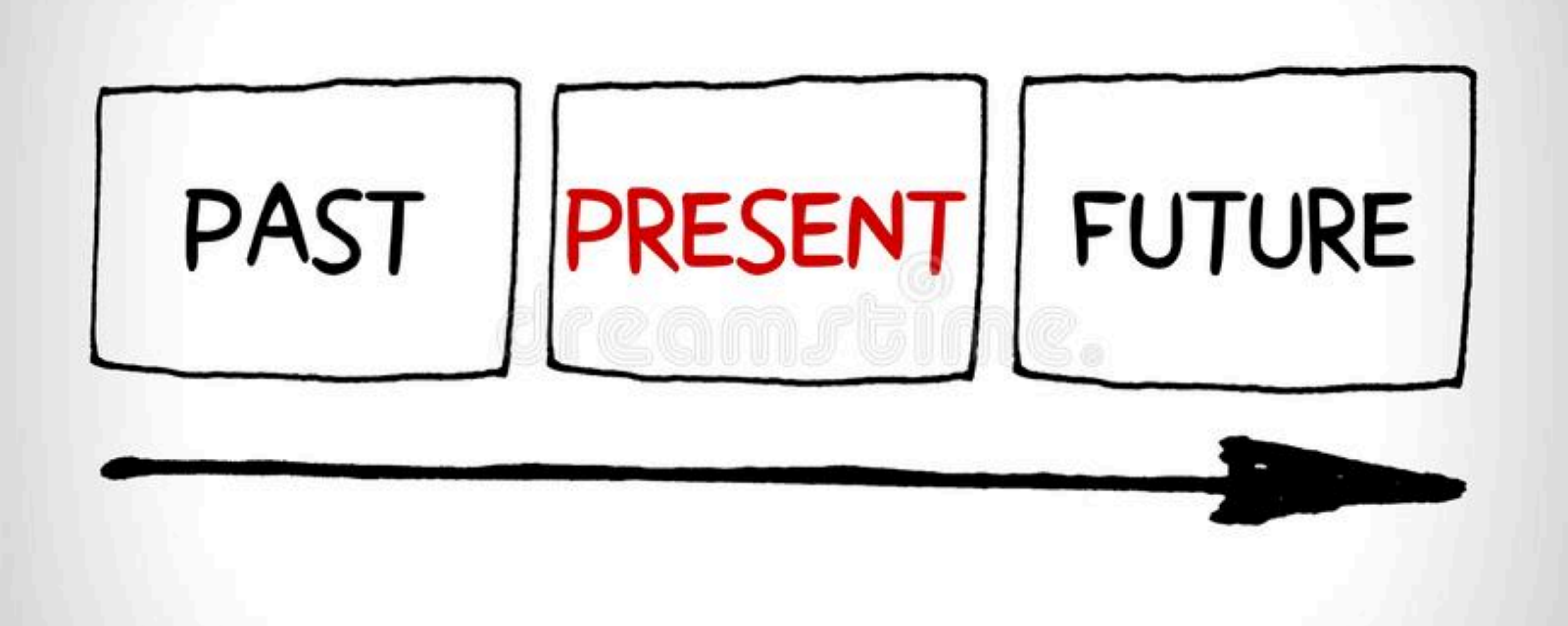
The Role of Gamma

- Q: “In week 1 lectures, it made sense to diminish past rewards when calculating cumulative rewards, because we don't care about the past. Using the same logic, in week 2, shouldn't we discount rewards received at present rather than discounting future rewards (because future rewards should matter more)?”

  - The **objective of learning** is about the **future**. We want to achieve goals, that is about the future: $max_\pi \mathbb{E}[R_{t+1} + \gamma R_{t+1} + \ldots | A \sim \pi]$

  - Learning algorithms cannot use future data; they must compute statistics of their experience.
    Do stuff, see what happens. Learn from it!

Goal: take actions now to lead to a better (reward) future



Updating: learn from the past

The Role of Gamma

- **The objective of learning is about the future.** We want to achieve goals, that is about the future:

  $$max_{\pi}\mathbb{E}[R_{t+1} + \gamma R_{t+1} + \ldots | A \sim \pi]$$

- Learning algorithms cannot use future data; they must compute statistics of their experience. **Do stuff, see what happens. Learn from it!**

$$Q_{n+1} \quad = \quad (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i .$$

$$\alpha(1-\alpha)^{(n-n)} = \alpha$$

In both cases the next reward is the most important !!!!!!

Reward specification

- Q: “I’m wondering about the rewarding intermediate steps, in the textbook they say it shouldn't be done as the agent could find a way to optimize this without achieving the goal. In the video, it mentioned providing an incentive for long stretch goals. What is best?”
    - **Rewards that frequently give the agent feedback can be good!**
    - Like me telling you every 30 mins how well you are doing—you could adapt as soon as things go wrong; or focus on things your are doing well
    - **DANGER**: the more detailed the reward specification, the bigger chance we make the optimal policy something we didn’t expect. **See the chess example in book**

Implementation

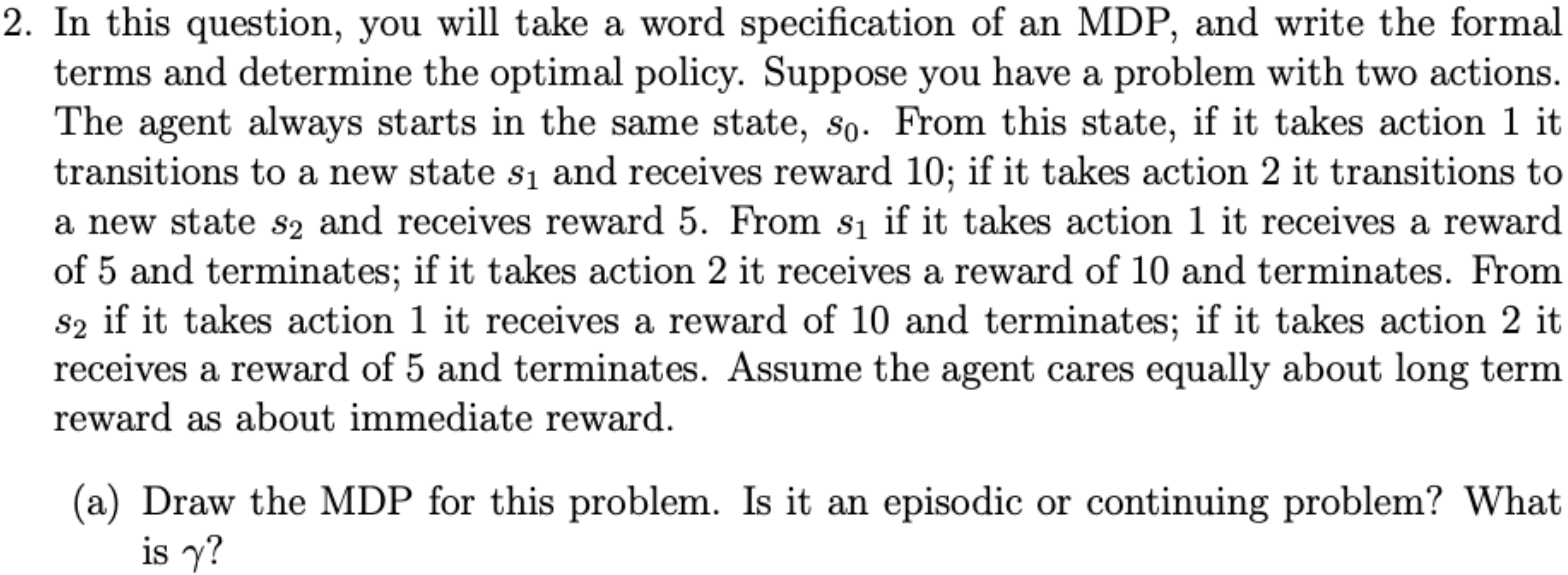
- Q: “How are policies implemented in code? Can they be like a python list that gets updated and changed after each episode so that it gets closer to the optimal policy?”
  - Think of implementing \epsilon-greedy for bandits. Did you write down all the probabilities? No! You used an if statement and a random number …

```
if np.random.rand() < \epsilon:
        return np.random.randint(num_actions)
else:
        return argmax(self.Q, s)
```

➡

$$\pi(a \mid s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathscr{A}|} & a = \textbf{argmax}_b Q(s, b) \\ \frac{\epsilon}{|\mathscr{A}|} & a \neq \textbf{argmax}_b Q(s, b) \end{cases}$$

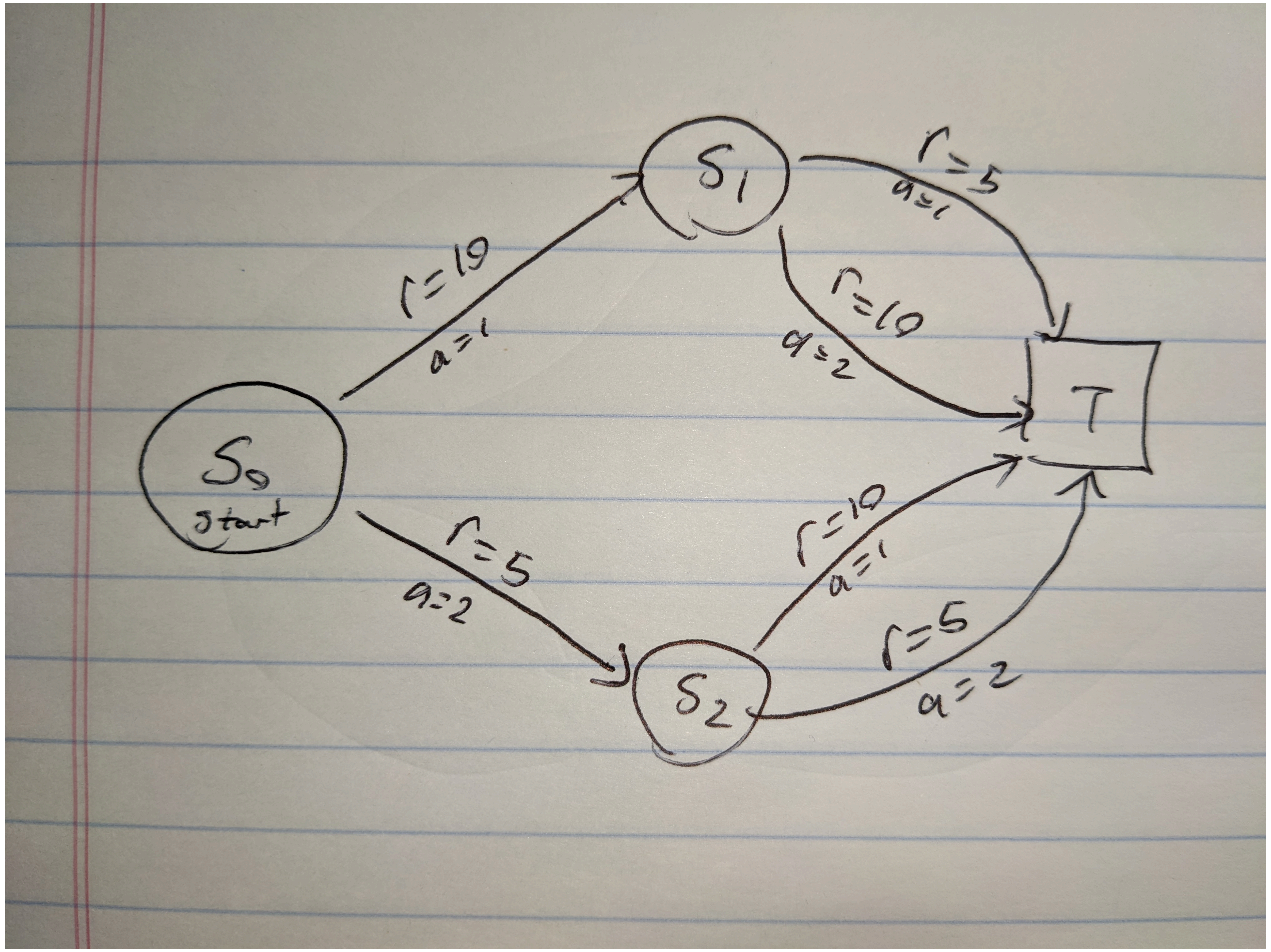
Worksheet Question 1



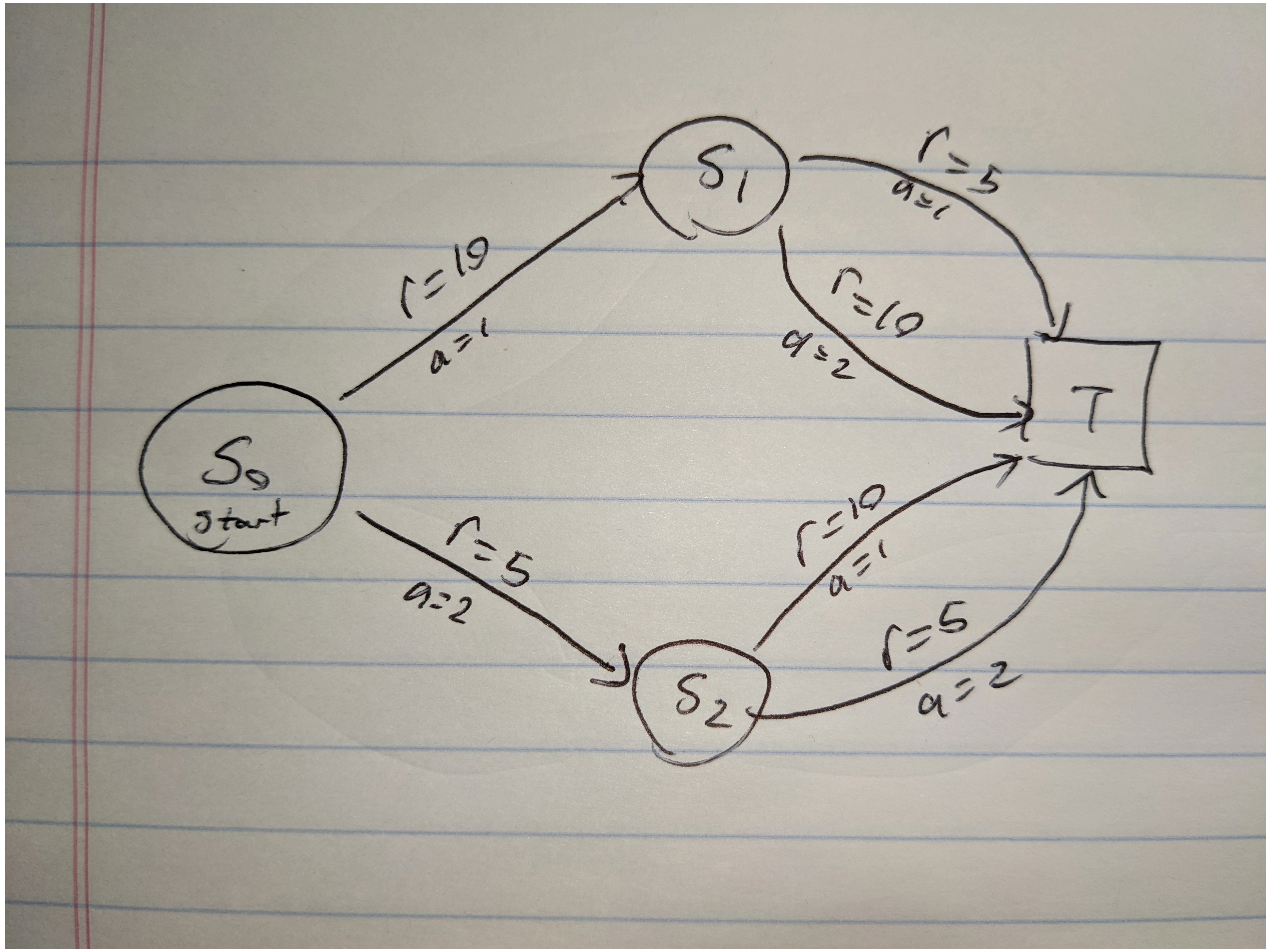
2. In this question, you will take a word specification of an MDP, and write the formal terms and determine the optimal policy. Suppose you have a problem with two actions. The agent always starts in the same state, s_0. From this state, if it takes action 1 it transitions to a new state s_1 and receives reward 10; if it takes action 2 it transitions to a new state s_2 and receives reward 5. From s_1 if it takes action 1 it receives a reward of 5 and terminates; if it takes action 2 it receives a reward of 10 and terminates. From s_2 if it takes action 1 it receives a reward of 10 and terminates; if it takes action 2 it receives a reward of 5 and terminates. Assume the agent cares equally about long term reward as about immediate reward.

   (a) Draw the MDP for this problem. Is it an episodic or continuing problem? What is γ?

Worksheet Question 1



Worksheet Question 1

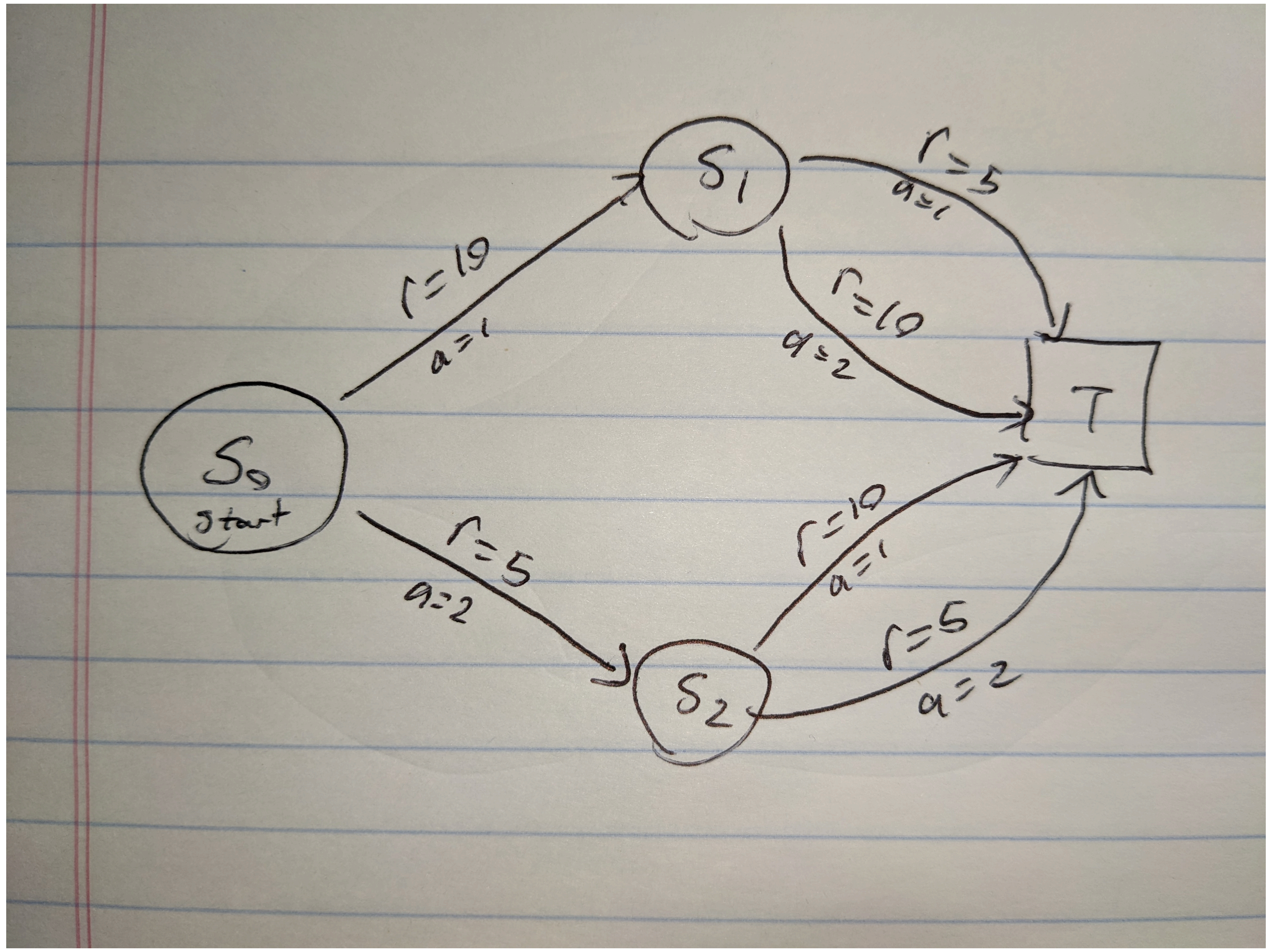


- $\gamma = ?$
  - One
- Episodic or continuing?
  - Episodic

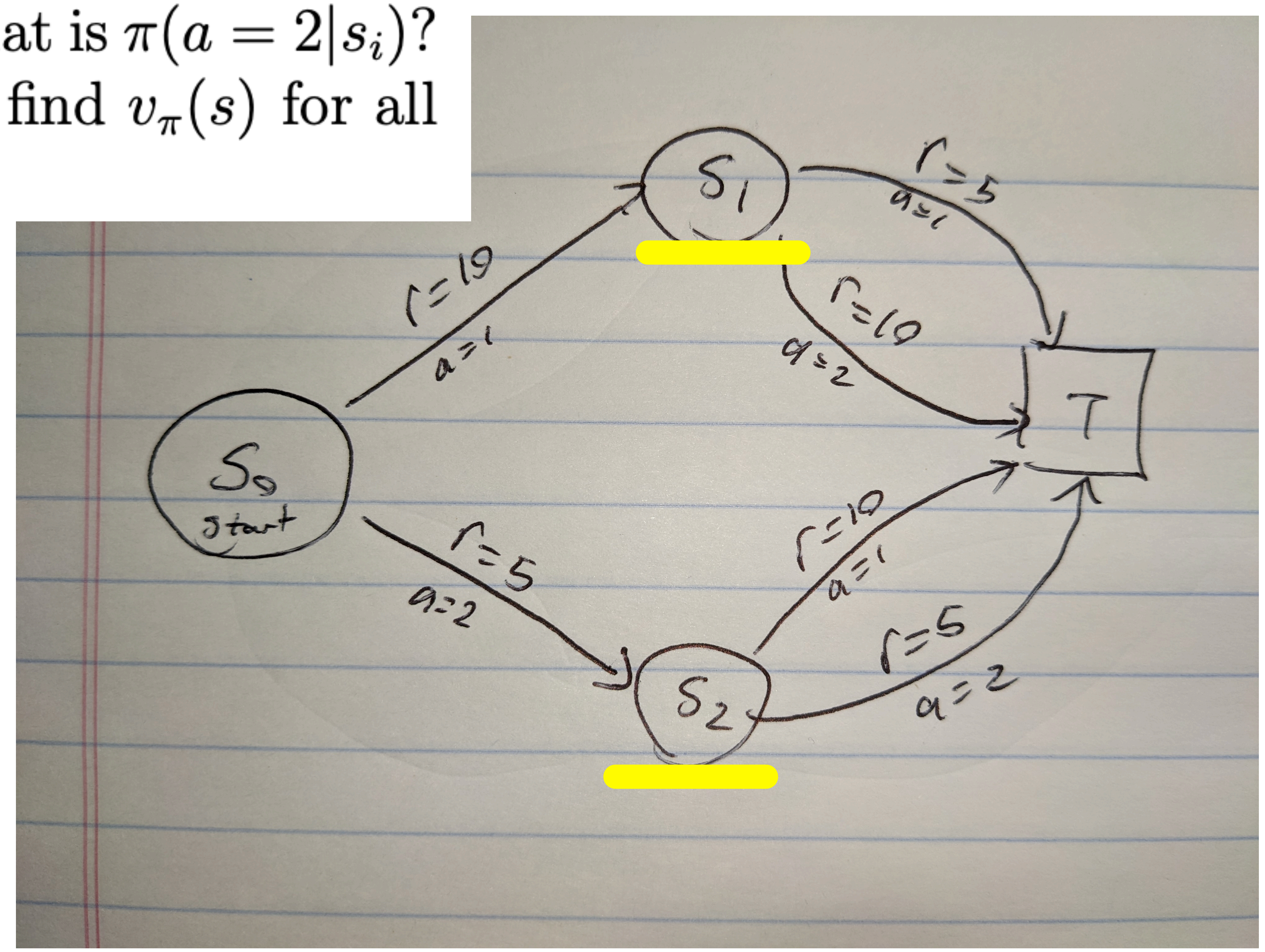
(a) Draw the MDP for this problem. Is it an episodic or continuing problem? What is γ?

(b) Assume the policy is $\pi(a = 1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a = 2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

- \pi(a=1 | s_i) = 0.3
- \pi(a=2 | s_i) = 1 - 0.3 = 0.7



(b) Assume the policy is $\pi(a = 1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a = 2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.



v(s1) = E[R] + \gamma v(sT)

= 0.3*5 + 0.7*10 + 1.0*0

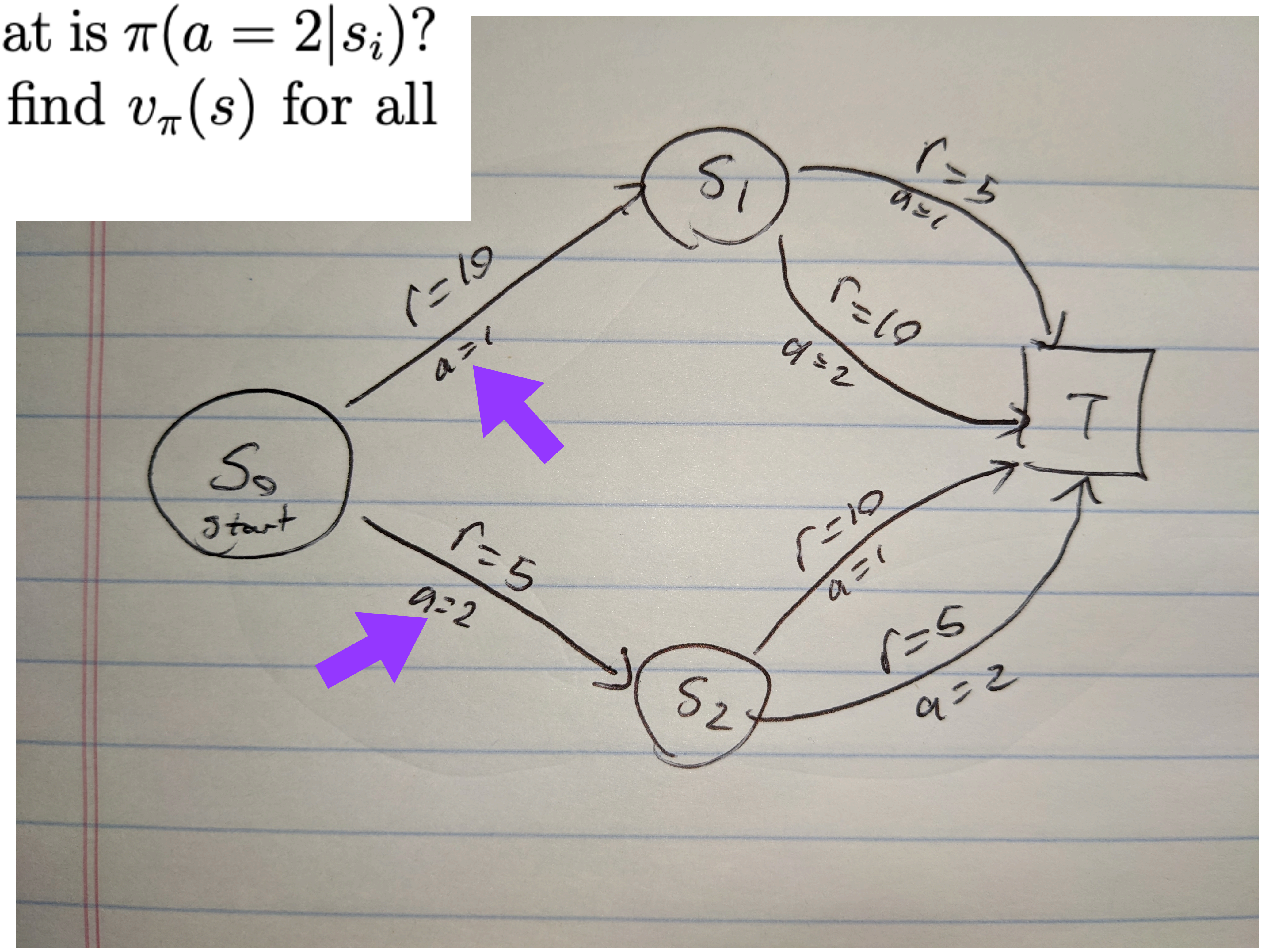
= 8.5

v(s2) = E[R] + \gamma v(sT)

= 0.3*10 + 0.7*5 + 1.0*0

= 6.5

(b) Assume the policy is $\pi(a=1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.



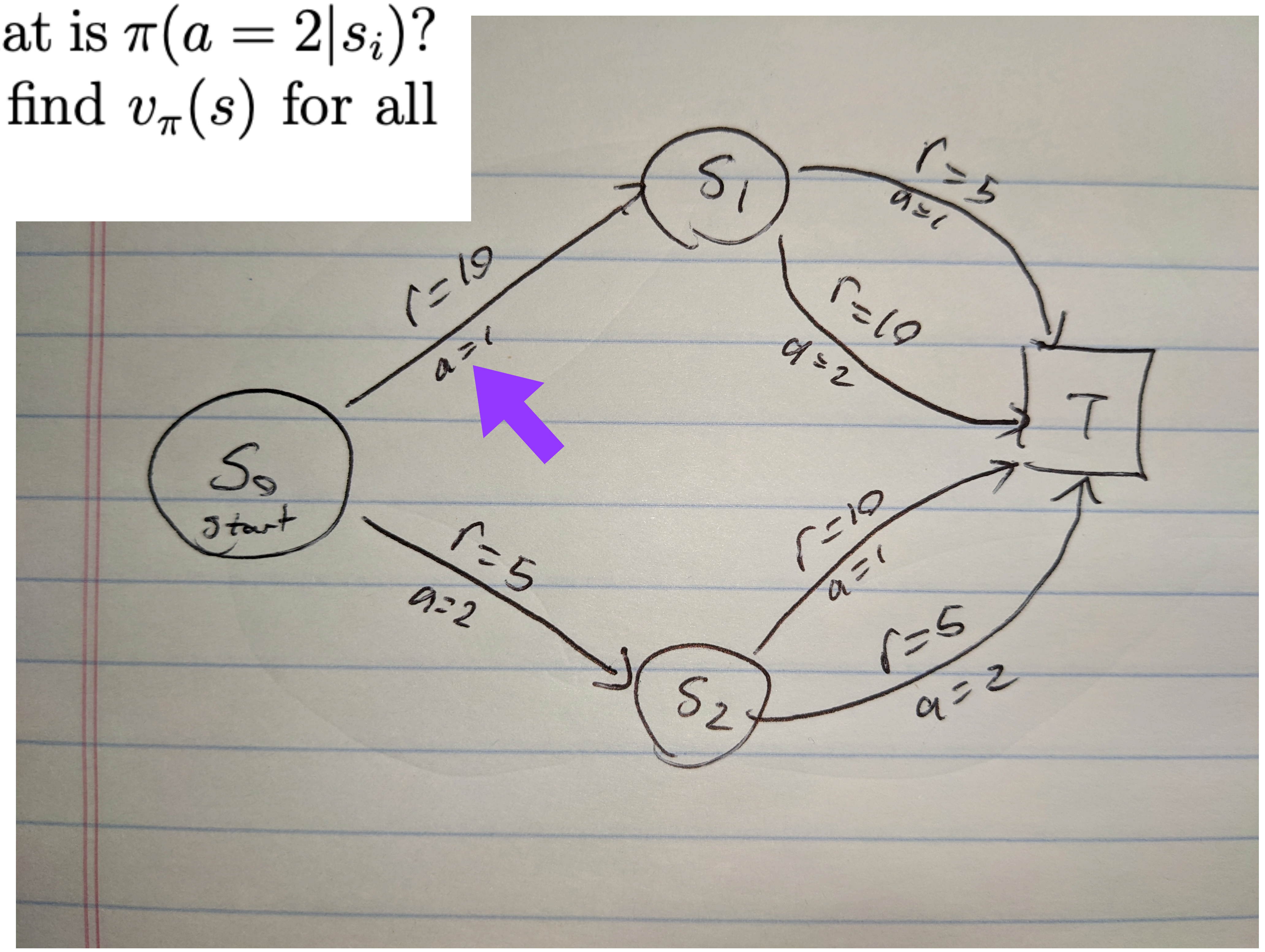
$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma v_\pi(s')\big]$$

There are two actions; so we reason about those two cases:

- a=1
- a=2

(b) Assume the policy is $\pi(a = 1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a = 2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

$$v_\pi(s) = \sum_a \boxed{\pi(a|s)} \sum_{s',r} p(s', r\,|\,s, \underline{a})\big[r + \gamma v_\pi(s')\big]$$



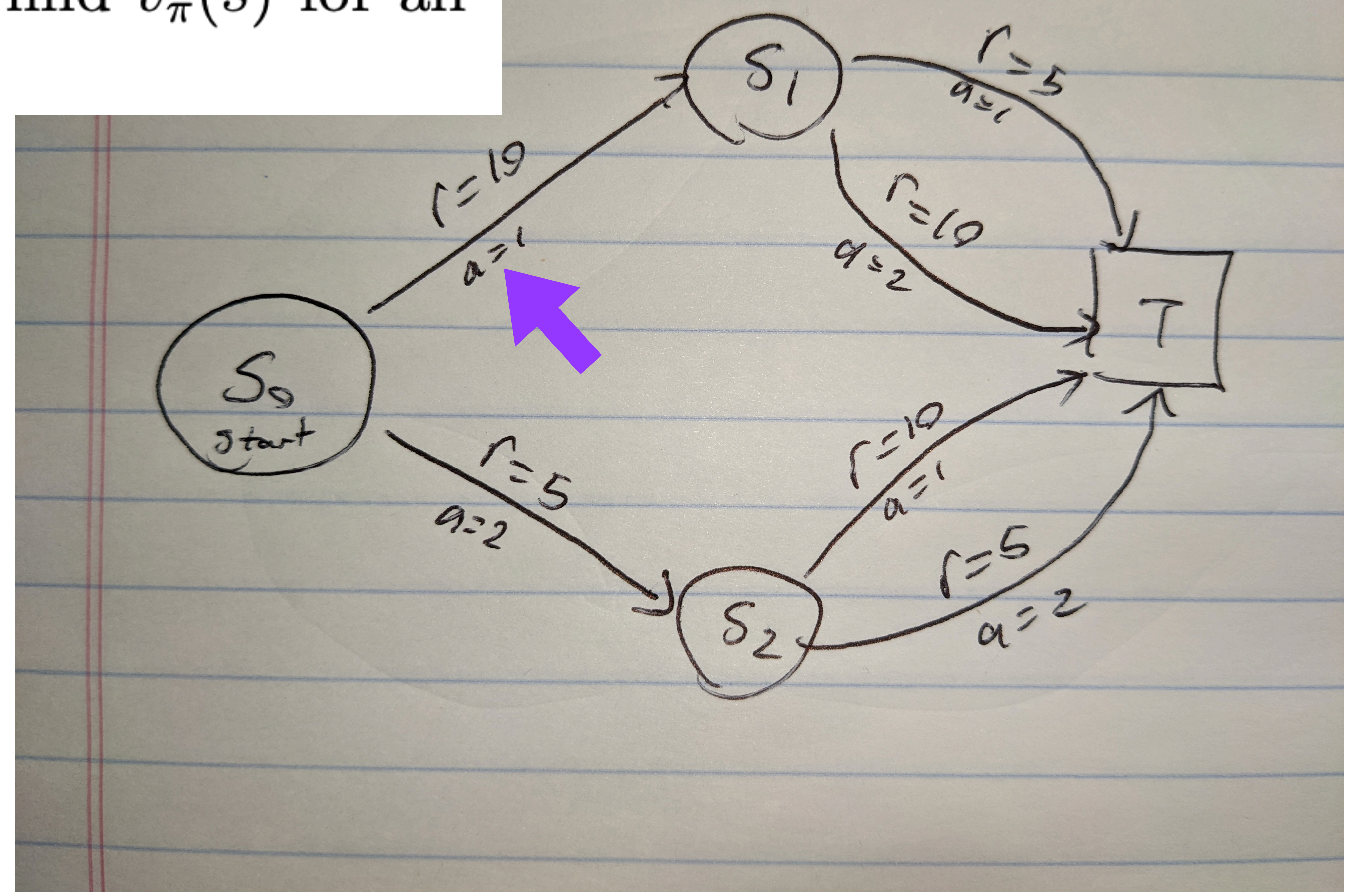
There are two actions; so we reason about those two cases:

- a=1: \pi(1|s0) = 0.3
- a=2: \pi(2|s0) = 0.7

recall: \pi(a=1 | s_i) = 0.3

(b) Assume the policy is $\pi(a=1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

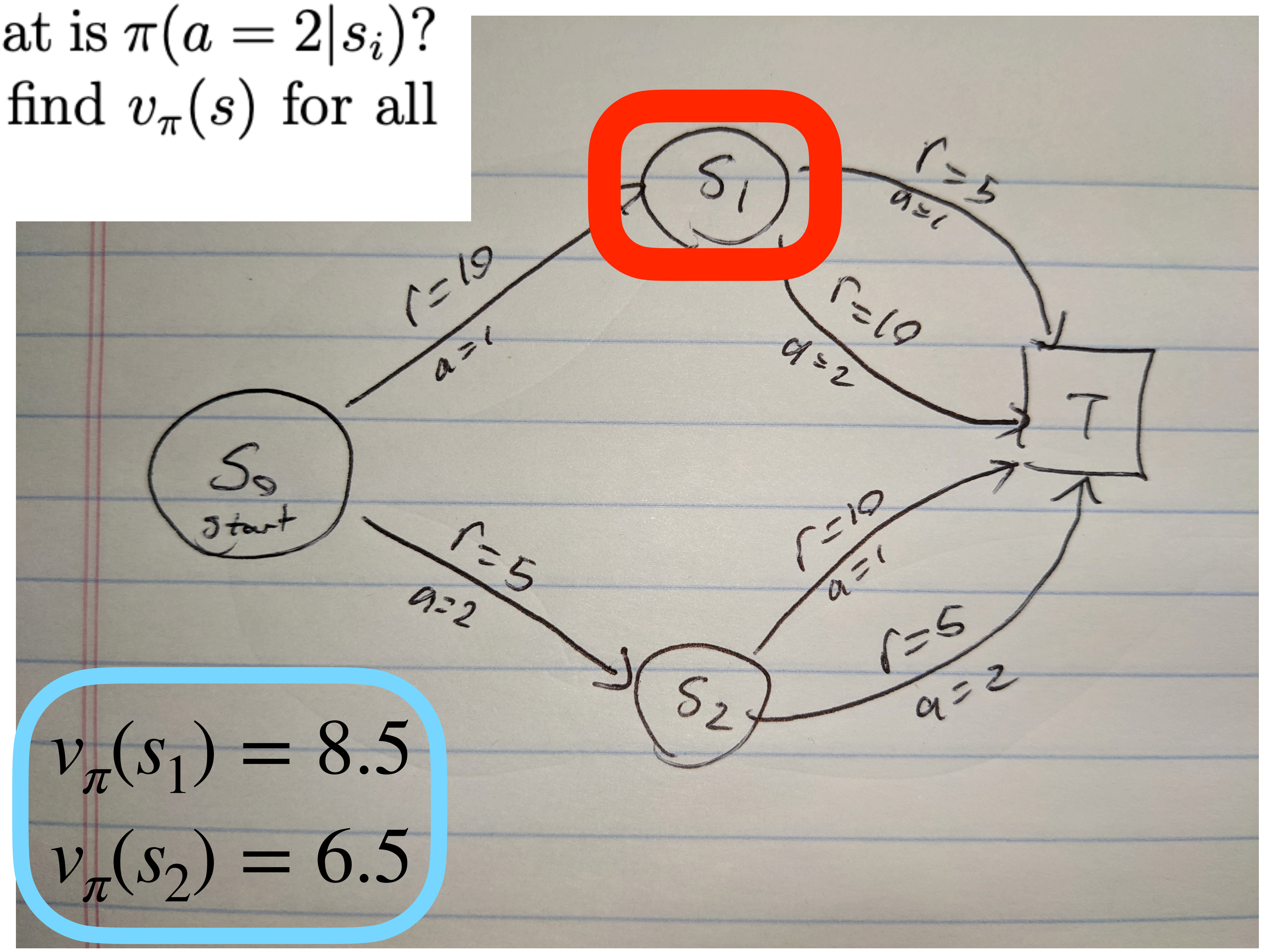
$$v_\pi(s) = \sum_a \boxed{\pi(a|s)} \sum_{s',r} p(s',r\,|\,s,\underline{a})\big[r + \gamma v_\pi(s')\big]$$



Lets work out everything for action 1:

- a=1: \pi(1|s0) = 0.3
- What is the next state and reward?
- s’ = s1 and r = 10
- What is p(s1,10 | s=s0, a=1)? 1.0

(b) Assume the policy is $\pi(a=1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.



$v_\pi(s_1) = 8.5$

$v_\pi(s_2) = 6.5$

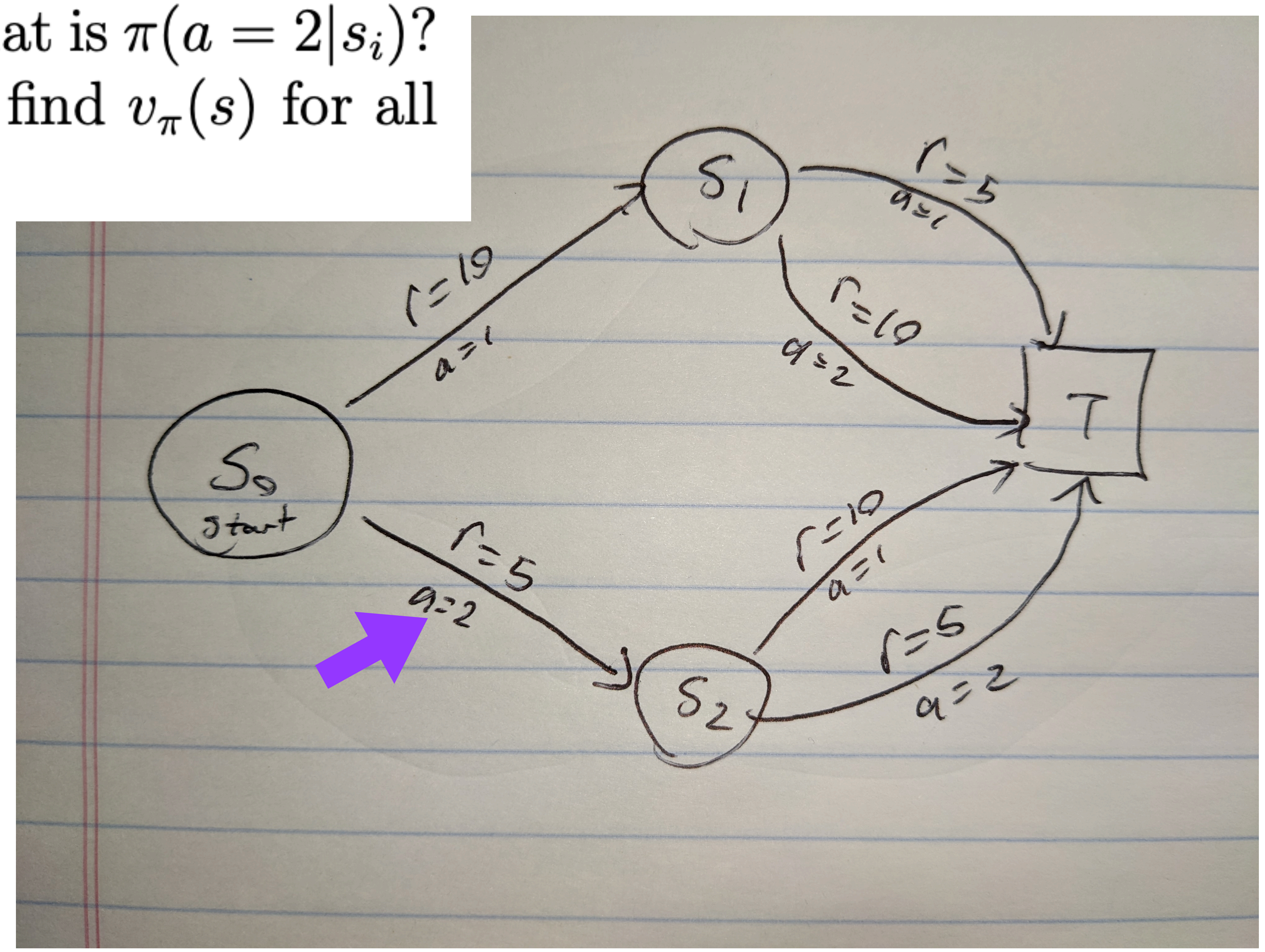
Lets work out everything for action 1:

- \pi(1|s0) = 0.3
- s' = s1 and r = 10 & p(s1,10 | s=s0, a=1) =1.0
- Therefore:
- $$\pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v(s')]$$
- $$= \pi(1 \mid s_0) p(s_1, 10 \mid s_0, 1)[10 + 1 * \boxed{v(s_1)}]$$
- $$= 0.3 * 1.0 * [10 + \boxed{8.5}] = 5.55$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma v_\pi(s')\big]$$

(b) Assume the policy is $\pi(a = 1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a = 2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

$$v_\pi(s) = \sum_a \boxed{\pi(a|s)} \sum_{s',r} p(s', r | s, \underline{a})\big[r + \gamma v_\pi(s')\big]$$

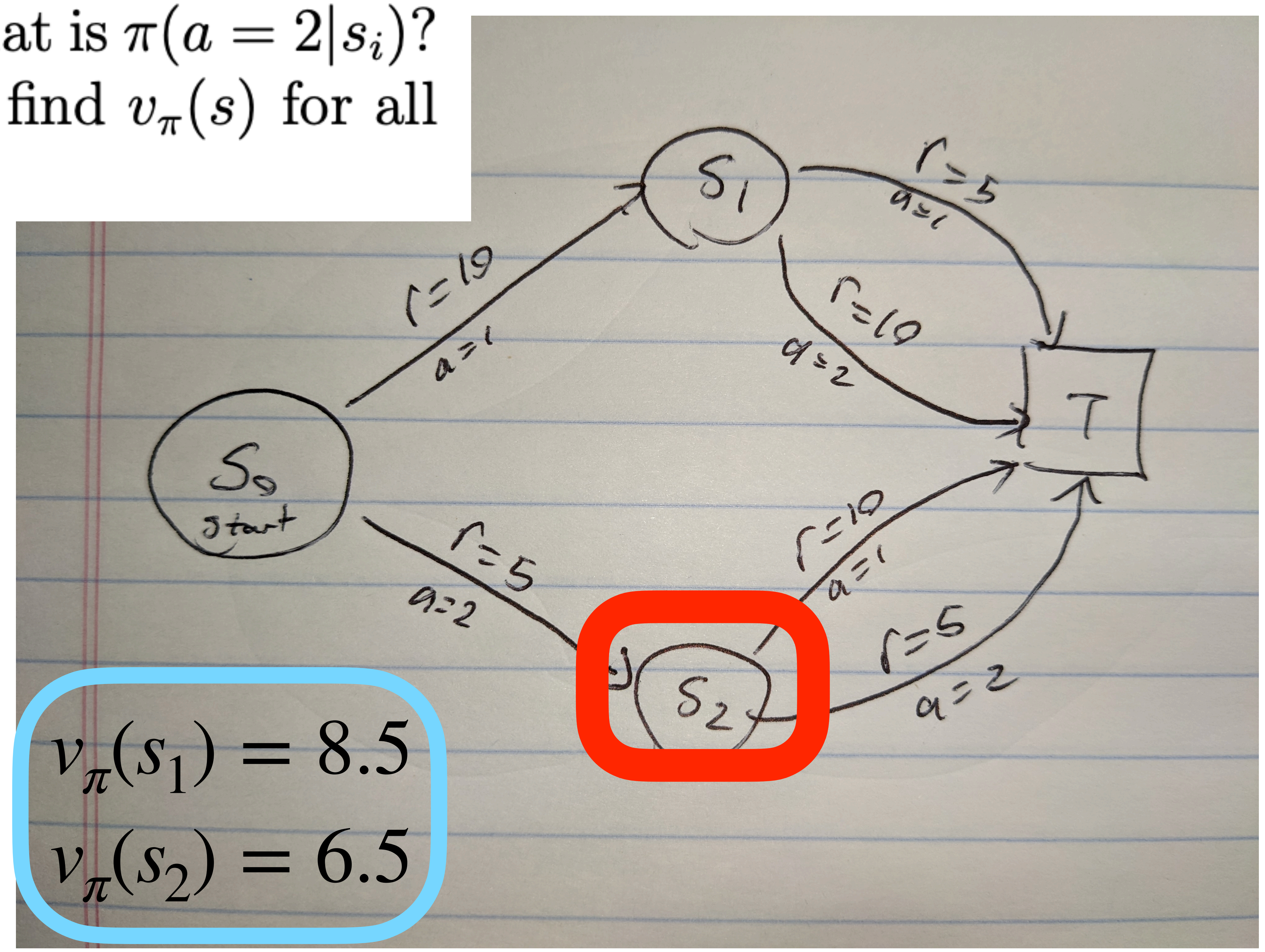


Lets work out everything for action 2:

- a=2: \pi(2|s0) = 0.7
    - What is the next state and reward?
    - s’ = s2 and r = 5
    - What’s the p(s2,5 | s=s0, a=2)? 1.0

recall: \pi(a=2 | s_i) = 0.7

(b) Assume the policy is $\pi(a=1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.



$v_\pi(s_1) = 8.5$

$v_\pi(s_2) = 6.5$

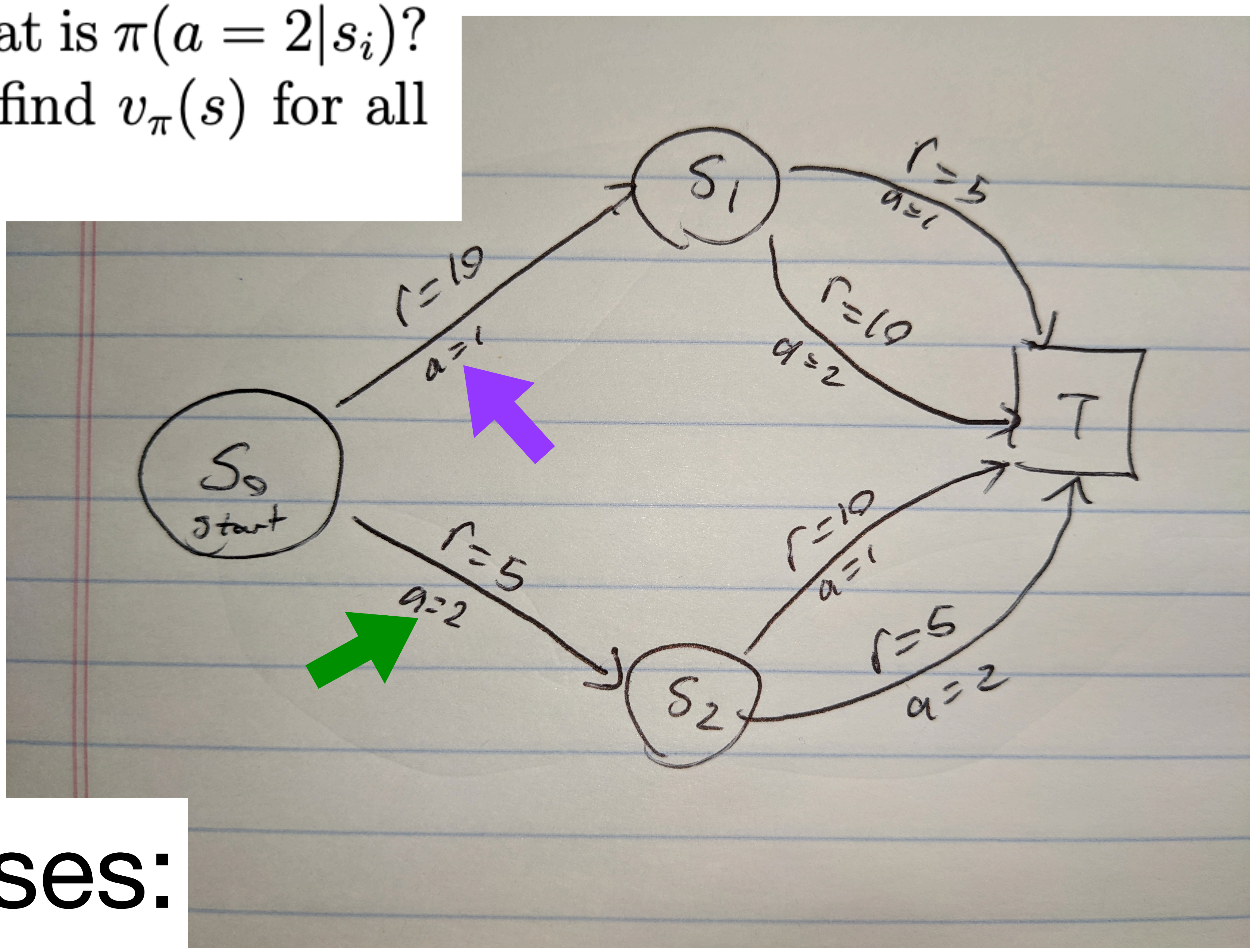
Lets work out everything for action 2:

- \pi(2|s0) = 0.7
- s' = s2 and r = 5 & p(s2,5 | s=s0, a=2) =1.0
- Therefore:
- $$\pi(a\,|\,s)\sum_{s',r} p(s',r\,|\,s,a)[r+\gamma v(s')]$$
- $$= \pi(2\,|\,s_0)p(s_2,5\,|\,s_0,2)[5 + 1 * \boxed{v(s_2)}]$$
- $$= 0.7 * 1.0 * [5 + \boxed{6.5}] = 8.05$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r\,|\,s,a)\big[r + \gamma v_\pi(s')\big]$$

(b) Assume the policy is $\pi(a=1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

$$v_\pi(s) = \sum_a \boxed{\pi(a|s)\sum_{s',r} p(s',r|s,a)\left[r+\gamma v_\pi(s')\right]}$$



Putting it all together. We have two cases:

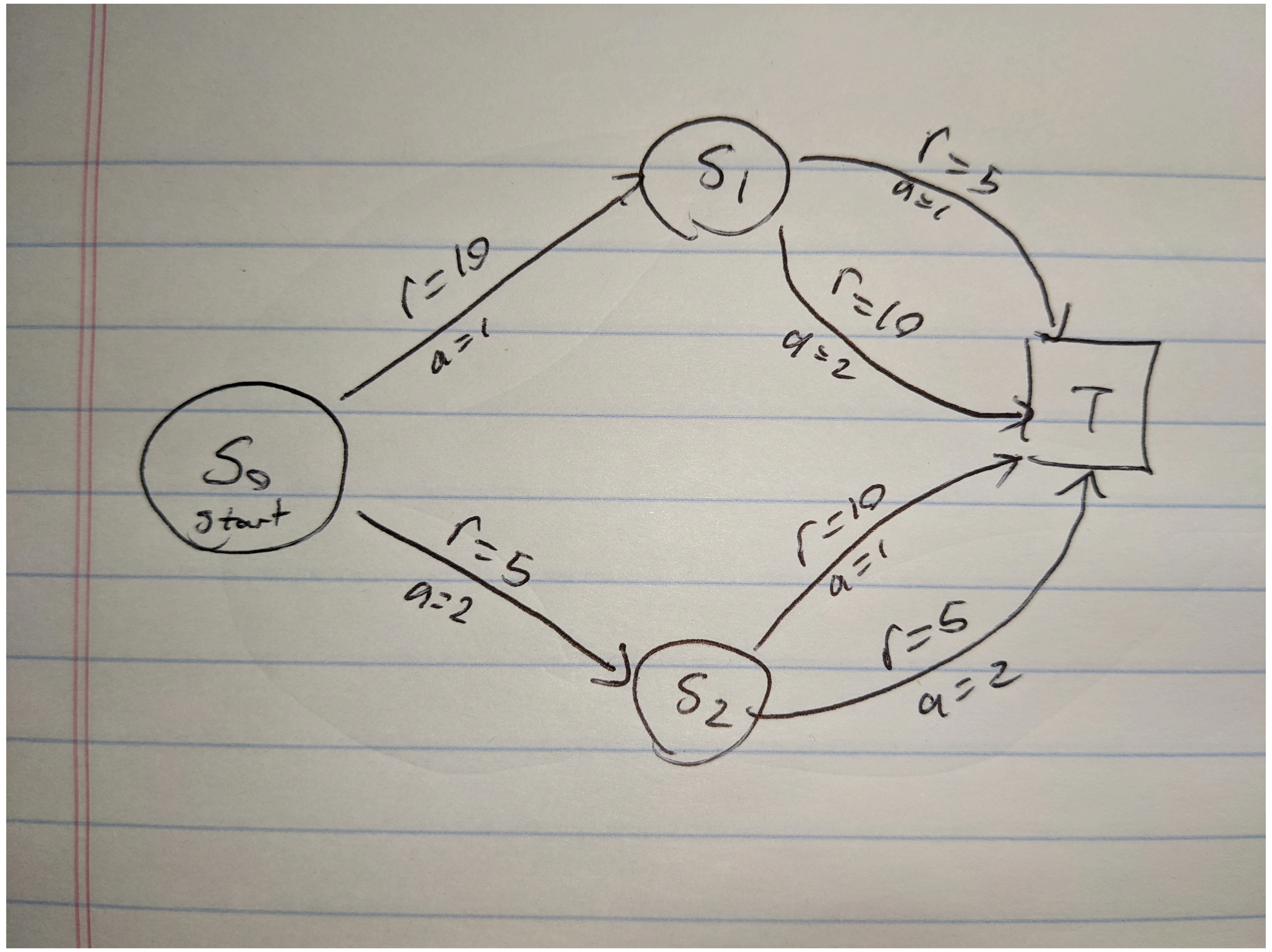
a=1: $0.3 * 1.0 * [10 + 8.5] = 5.55$

a=2: $0.7 * 1.0 * [5 + 6.5] = 8.05$

Add the two: 5.55 + 8.05 = 13.6

$$v_\pi(s_1) = 8.5$$
$$v_\pi(s_2) = 6.5$$
$$v_\pi(s_0) = 13.6$$

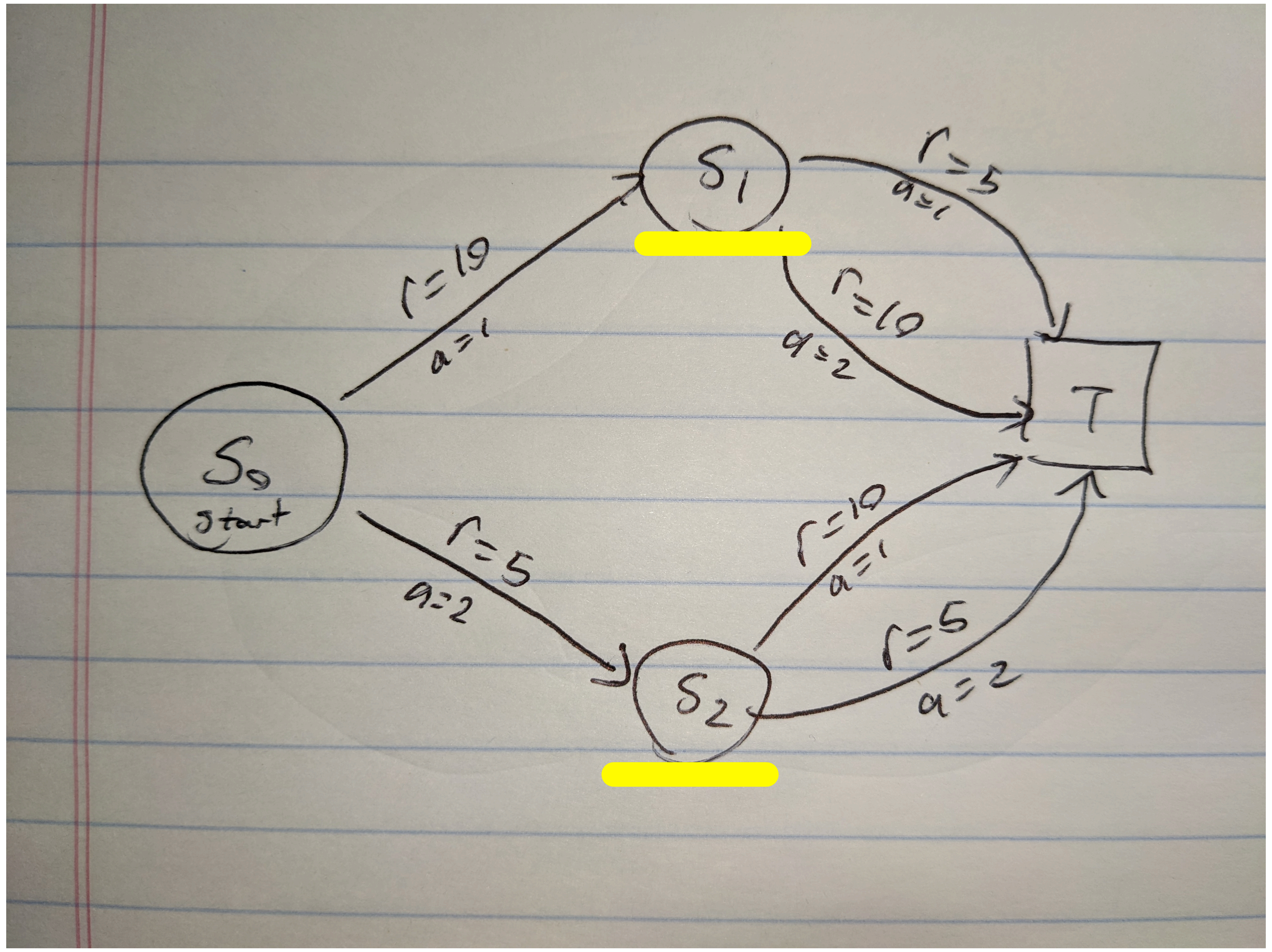
What is the optimal policy?



Again, work backwards!

- Best action in s1? => a=2
- Best action in s2? => a=1
- Best action in s0? Which is better s1 or s2? Does't matter pick based on immediate reward => a=1
  - pi(a=1|s0) = 1, pi(a=2|s1) =1, pi(a=1|s2)=1

Exercise: can do this computation for v(s1) and v(s2) using the bellman equation of v_\pi?



$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

v(s1) = E[R] + \gamma v(sT)

= 0.3*5 + 0.7*10 + 1.0*0

= 8.5

v(s2) = E[R] + \gamma v(sT)

= 0.3*10 + 0.7*5 + 1.0*0

= 6.5

Graded Quiz review

- https://www.coursera.org/teach/fundamentals-of-reinforcement-learning/authoringBranch~NvD-BMC4EeqciA6D1ug0Gw/content/edit/quiz/5AizQ