Markov decision Process

श्रद्धावान् लभते ज्ञानम्

AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Taxi Problem

A taxi driver is working in a small town with **three main operating zones**:

- **S1: Parking Lot** (where taxis wait for requests)
- **S2: Pickup Zone** (where passengers usually show up)
- **S3: Gas Station** (needed to refuel)

There are also **two possible outcomes** once the taxi makes a trip:

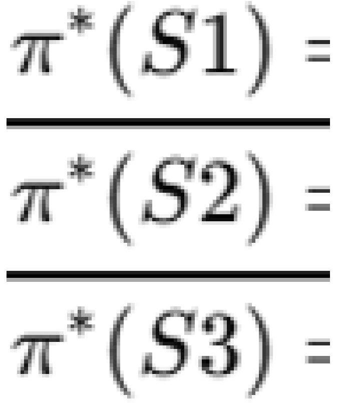
- **S4: City Center** (profitable ride, reward = +15, terminal state)
- **S5: Highway Trap** (wrong route, loss = –10, terminal state)

The driver must decide how to act at each non-terminal zone.

The environment is **stochastic** (things don’t always go as planned).

RL problem

The taxi driver should learn a **policy that maximizes the expected cumulative reward over time** — i.e., earning as much profit as possible from trips to the City Center while avoiding costly mistakes such as unnecessary fuel usage or falling into the Highway Trap.



$\pi^*(S1) =$ ______

$\pi^*(S2) =$ ______

$\pi^*(S3) =$

What can be possible states here??

A taxi driver is working in a small town. The driver faces choices at different locations, and **things don’t always go as planned**. The driver must decide what to do in each place, but the outcomes are **uncertain**.

Scene 1: Parking Lot

The taxi is waiting in the parking lot.

- If the driver decides to **Wait**, sometimes nothing happens (the taxi keeps waiting, wasting fuel and time), but sometimes a request comes and the driver must go pick up a passenger.
- If the driver chooses to **Drive to the Pickup Zone**, usually it works, but occasionally there’s a detour that brings the taxi back to the parking lot.
- The driver could also decide to **Refuel**, driving to the gas station, but it takes extra time and fuel cost.

Scene 2: Pickup Zone

Now the taxi is in the busy pickup zone.

- If the driver **Waits for a Passenger**, often someone arrives and the taxi goes to the city center with a good reward. But sometimes no one shows up, and the taxi just wastes more time.
- If the driver chooses to **Return to the Parking Lot**, most of the time they make it, but sometimes they take a wrong road and end up in a bad situation (the highway trap).
- The driver could also try the **Highway Shortcut** to reach the city center quickly. Sometimes it works and gives a decent reward, but sometimes it leads straight into the highway trap.

Scene 3: Gas Station

At the gas station:

•The driver can try to **Refuel**. Most of the time it works fine, and the taxi returns to the parking lot, ready for the next ride. But rarely, the pump fails and the taxi stays stuck at the station.

Scene 4: City Center

If the taxi reaches here, the driver earns a good fare and the trip ends.

Scene 5: Highway Trap

If the taxi ends up here, it's a costly mistake — the trip ends with a loss.

From this story,

- list the states (places the taxi can be).
- For each state, think of the possible actions the driver can take.
- For each action, discuss what might happen:
  - What are the possible outcomes, and how likely are they? What are the rewards or penalties?

Use this to design the MDP (states, actions, transitions, rewards.

Solution: Designing the Taxi MDP

1. States (places the taxi can be)

- **S1: Parking Lot** (where taxis wait for passengers or decide to move)
- **S2: Pickup Zone** (where passengers may appear)
- **S3: Gas Station** (needed to refuel)
- **S4: City Center** (terminal state, good outcome: profitable ride)
- **S5: Highway Trap** (terminal state, bad outcome: wrong road, penalty)

Actions possible in each state

- **At Parking Lot (S1):**
  - A1: Wait for a request
  - A2: Drive to Pickup Zone
  - A3: Go refuel
- **At Pickup Zone (S2):**
  - B1: Wait for a passenger
  - B2: Return to Parking Lot
  - B3: Take Highway Shortcut
- **At Gas Station (S3):**
  - C1: Refuel
- **At City Center (S4):**
  - No actions (terminal state)
- **At Highway Trap (S5):**
  - No actions (terminal state)

Outcomes (Transitions + Probabilities + Rewards)

At Parking Lot (S1):

•A1 Wait:

- 0.8 → stay in S1 (idle, reward −1)
- 0.2 → move to S2 (request arrives, reward −1)

•A2 Drive to Pickup Zone:

- 0.9 → reach S2 (reward −2 fuel/time)
- 0.1 → detour back to S1 (reward −2)

•A3 Refuel:

- 1.0 → go to S3 (reward −3)

At Pickup Zone (S2):

•B1 Wait for Passenger:

- 0.6 → passenger appears → S4 (reward +15, terminal)
- 0.4 → no passenger → stay in S2 (reward −1)

•B2 Return to Parking Lot:

- 0.9 → go to S1 (reward −2)
- 0.1 → wrong turn → S5 (reward −5, terminal)

•B3 Highway Shortcut:

- 0.5 → succeed → S4 (reward +12, terminal)
- 0.5 → fail → S5 (reward −10, terminal)

Outcomes (Transitions + Probabilities + Rewards)

At Gas Station (S3):

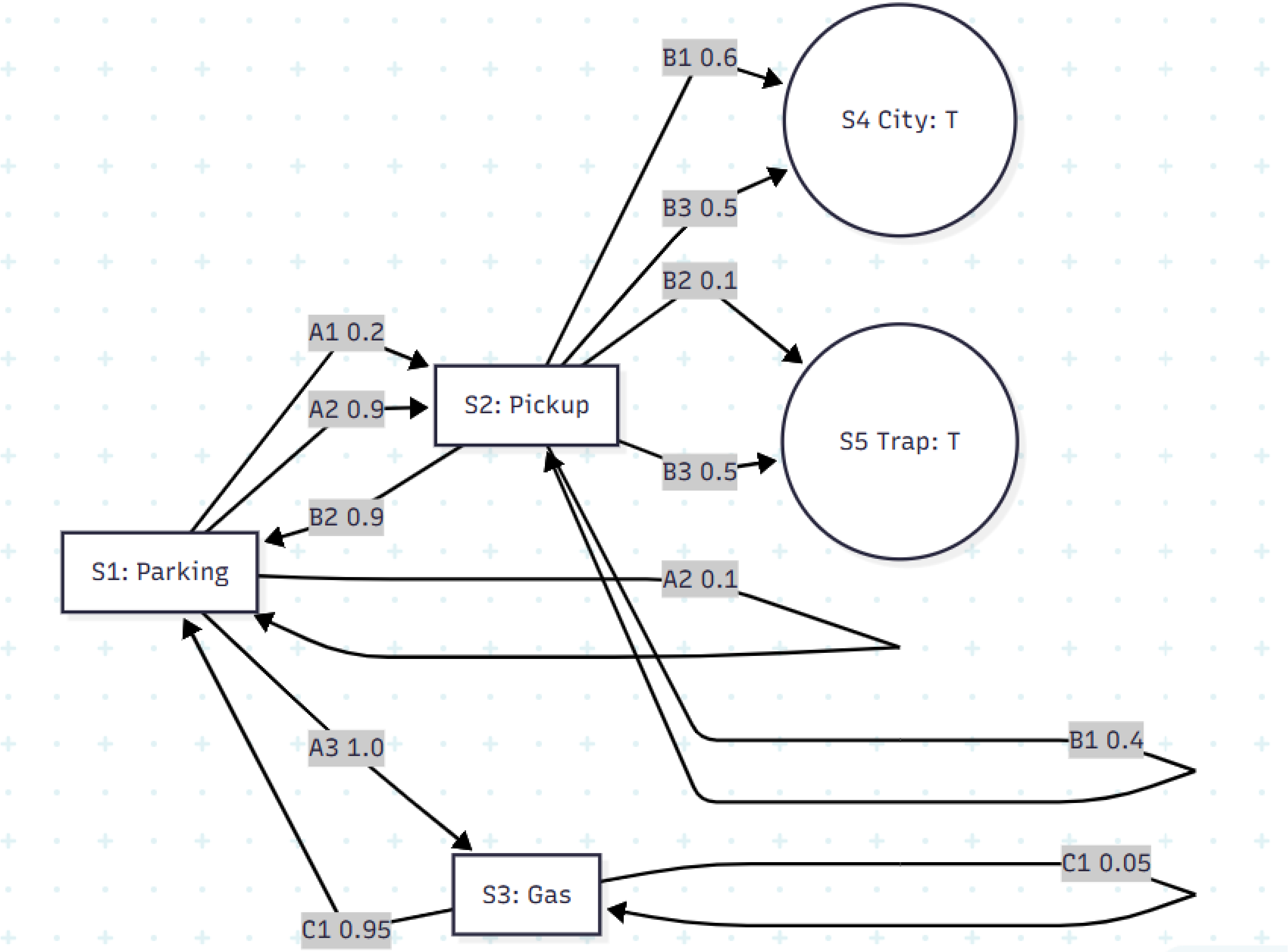
- **C1 Refuel:**
  - 0.95 → refuel, return to S1 (reward −2)
  - 0.05 → pump fails, stay in S3 (reward −2)

At City Center (S4):

- **Terminal**: No further transitions, value fixed after entry.

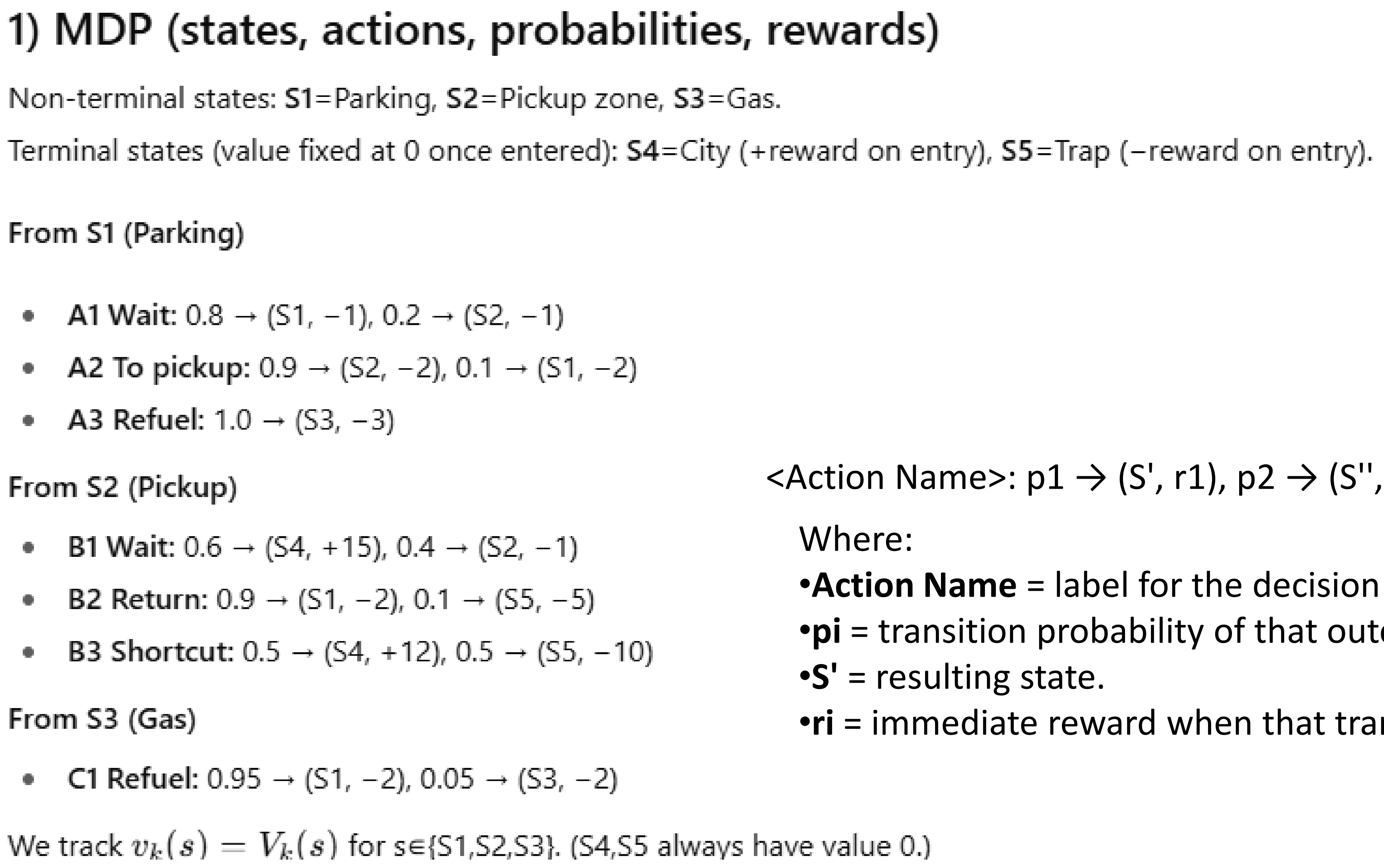
At Highway Trap (S5):

- **Terminal**: No further transitions, value fixed after entry.


B1 0.6
S4 City: T
B3 0.5
B2 0.1
A1 0.2
A2 0.9
S2: Pickup
S5 Trap: T
B3 0.5
B2 0.9
S1: Parking
A2 0.1
B1 0.4
A3 1.0
S3: Gas
C1 0.05
C1 0.95

Final MDP specification

- **States**: {S1, S2, S3, S4, S5}
- **Actions**: {A1, A2, A3, B1, B2, B3, C1}
- **Transition model**: Probabilities given above
- **Rewards**:
  - Negative rewards for time/fuel use: (−1, −2, −3 depending on action)
  - Large positive reward for reaching City Center (S4: +12 or +15 depending on action)
  - Penalties for Highway Trap (S5: −5, −10 depending on action)



1) MDP (states, actions, probabilities, rewards)

Non-terminal states: **S1**=Parking, **S2**=Pickup zone, **S3**=Gas.

Terminal states (value fixed at 0 once entered): **S4**=City (+reward on entry), **S5**=Trap (−reward on entry).

From S1 (Parking)

- **A1 Wait:** 0.8 → (S1, −1), 0.2 → (S2, −1)
- **A2 To pickup:** 0.9 → (S2, −2), 0.1 → (S1, −2)
- **A3 Refuel:** 1.0 → (S3, −3)

From S2 (Pickup)

- **B1 Wait:** 0.6 → (S4, +15), 0.4 → (S2, −1)
- **B2 Return:** 0.9 → (S1, −2), 0.1 → (S5, −5)
- **B3 Shortcut:** 0.5 → (S4, +12), 0.5 → (S5, −10)

From S3 (Gas)

- **C1 Refuel:** 0.95 → (S1, −2), 0.05 → (S3, −2)

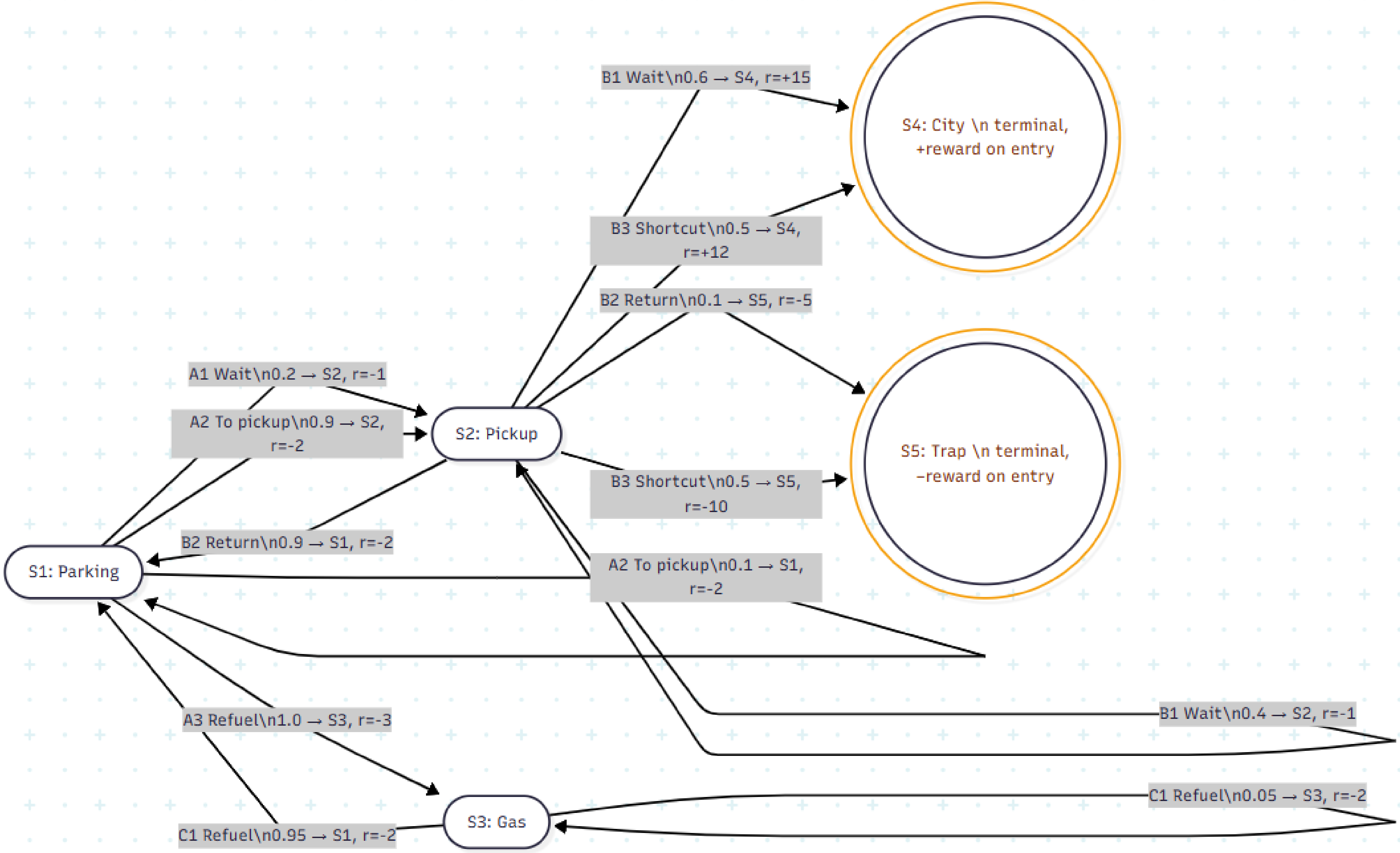
We track $v_k(s) = V_k(s)$ for s∈{S1,S2,S3}. (S4,S5 always have value 0.)

<Action Name>: p1 → (S', r1), p2 → (S'', r2), ...

Where:

- **Action Name** = label for the decision at the current state.
- **pi** = transition probability of that outcome.
- **S'** = resulting state.
- **ri** = immediate reward when that transition happens.

Draw a state Transition Diagram for this


B1 Wait\n0.6 → S4, r=+15
S4: City \n terminal, +reward on entry
B3 Shortcut\n0.5 → S4, r=+12
B2 Return\n0.1 → S5, r=-5
A1 Wait\n0.2 → S2, r=-1
A2 To pickup\n0.9 → S2, r=-2
S2: Pickup
S5: Trap \n terminal, −reward on entry
B3 Shortcut\n0.5 → S5, r=-10
B2 Return\n0.9 → S1, r=-2
S1: Parking
A2 To pickup\n0.1 → S1, r=-2
B1 Wait\n0.4 → S2, r=-1
A3 Refuel\n1.0 → S3, r=-3
C1 Refuel\n0.05 → S3, r=-2
S3: Gas
C1 Refuel\n0.95 → S1, r=-2