RL ASSIGNMENT 2 REPORT

Contributions

1. Anya – SE22UCSE033

Theory problem 1, Programming -policy_evaluation, Analysis- 2,4,5

2. Stuti Garg–SE22UCSE263

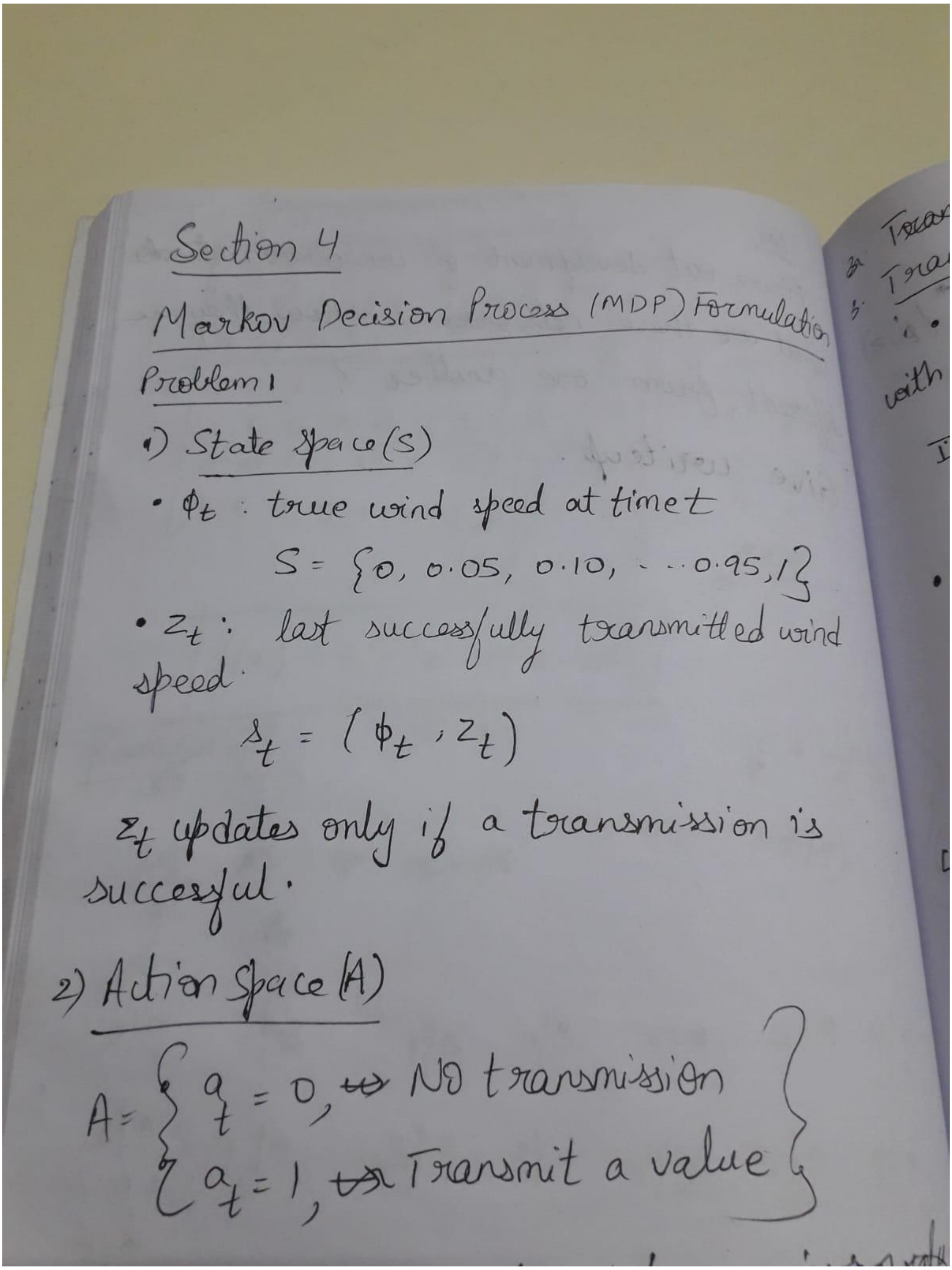
Theory problem 2, Programming- value_iteration, Analysis- 3

3. Lakshmi –SE22UCSE149

Theory problem 3, Programming -policy_iteration, Analysis- 1

Theory Component

1.



Section 4

Markov Decision Process (MDP) Formulation

Problem 1

1) State Space (S)

- ϕ_t : true wind speed at time t

$$S = \{0, 0.05, 0.10, \ldots 0.95, 1\}$$

- z_t : last successfully transmitted wind speed.

$$s_t = (\phi_t ; z_t)$$

z_t updates only if a transmission is succesful.

2) Action Space (A)

A = { $a_t = 0$, ~~to~~ No transmission

$a_t = 1$, ~~to~~ Transmit a value }

3. Transmissio

3. Transition Prob.

- $\phi_t \to \phi_{t+1}$ follows Markov chain with transition prob. $p_{\phi\phi'}$..

If $a_t = 1$ (transmission)

- with prob. λ, → success & z_t updates
- with prob. $1-\lambda$, → fails, z_t unchanged.
- If $a_t = 0$ (no transmission) z_t remains unchanged.

$$z_{t+1} = z_t$$

4. Reward func.

Objective → to minimize

$$\sum_{t=0}^{\infty} \beta^t (\phi_t - z_t)^2$$ where β is discount factor

reward at time t is

$$R(s_t, a_t) = -(\phi_t - z_t)^2$$

b) Bellman Eq.

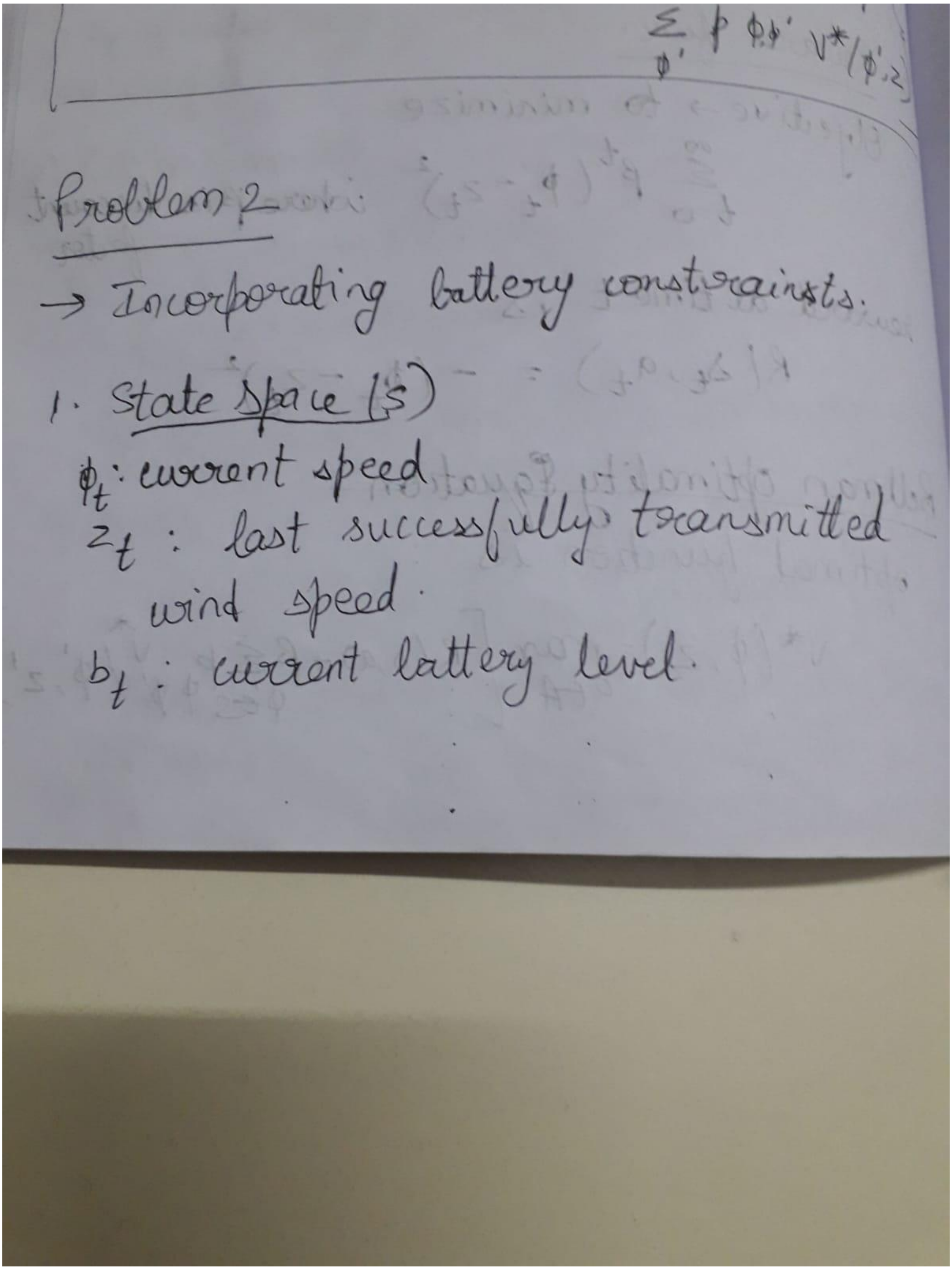
$$V^*(\phi, z) = \max_{a \in A} \left\{ -\left(\lambda(\phi-a)^2 + (1-\lambda)(\phi-z)^2\right) + \beta \sum_{\phi'} p_{\phi,\phi'} \cdot V^*(\phi', z') \right\}$$

where

~~ae~~ $z' = \begin{cases} a, & \text{with prob. } \lambda \\ z, & \text{with prob. } 1-\lambda \end{cases}$

$$V^*(\phi', z') = \lambda V^*(\phi', a) + (1-\lambda) V^*(\phi', z)$$

2.



$\sum_{\phi'} p_{\phi\phi'} V^*(\phi', z)$

Problem 2

→ Incorporating battery constrainsts.

1. State Space (S)

ϕ_t: current speed

z_t : last successfully transmitted wind speed.

b_t : current battery level.

$S = (\phi_t, z_t, b_t)$

$\phi_t \in S = \{0, 0.05, 0.10, \dots 0.95, 1\}$

$z_t \in S$

$b_t \in \{0, 1, \dots B\}$ (where B is max. capacity)

2. Action Space (A)

$A = \begin{cases} a = 0 \rightarrow \text{do not transmit} \\ a = 1 \rightarrow \text{transmit} \end{cases}$

if $a = 1$ then v must be chosen from S.

$A \in \{(0), (1, v)\}$ where $v \in S$

(0) → don't transmit

(1, v) → transmit wind speed v.

Constraint:

WSN can't transmit if $b_t < \eta$

where η is energy needed for transmission.

3. Transition Prob.

next state $(\phi_{t+1}, z_{t+1}, b_{t+1})$ depends on:

- wind speed transition.

$$P(\phi_{t+1} \mid \phi_t) = P_{\phi_t, \phi_{t+1}}$$

- last successful z_t
  - if (λ) then $z_{t+1} = v$
  - if $(1-\lambda)$ then $z_{t+1} = z_t$

- battery transition

if (a=0) → no transmission

$$b_{t+1} = \min(B, b_t + \delta_t)$$

if (a=1)

$$b_{t+1} = \min(B, b_t - \eta + \delta_t),$$

where δ_t is solar energy.

$$P(\phi_{t+1}, z_{t+1}, b_{t+1} | \phi_t, z_t, b_t, a, v)$$

4. Reward R(s,a)

$$R(s,a) = -(\phi_t - z_t)^2.$$

if transmission is successful, z_t is updated to v, reducing error else remains same.

Bellman Eq

$$V^*(\phi, z, b) = \max \Big\{ -(\phi - z)^2 + \beta \sum_{\phi'} p_{\phi\phi'} \sum_{\delta} a_{\delta} V^*(\phi', z, \min(B, b+\delta)),$$

$$\max_{\substack{v \in S,\\ b \geq \eta}} \Big[-(\phi - v)^2 + \lambda\beta \sum_{\phi'} p_{\phi\phi'} \sum_{\delta} a_{\delta} V^*(\phi', v, \min(B, b - \eta + \delta)) + (1-\lambda)\beta \sum_{\phi'} p_{\phi\phi'} \sum_{\delta} a_{\delta} V^*(\phi', z, \min(B, b - \eta + \delta)) \Big]$$

Markov Decision Process (MDP) Formulation

Problem-3

1. States~~space~~ (s)

ϕ_t : current speed

z_t : last successfully transmitted wind speed

b_t : current battery level

P_t : the phase of the WSN node (active or passive) ~~active~~

Passive → node not allowed to transmit

Active → 1 to τ → remaining time in current active phase.

$S_t = (\phi_t, z_t, b_t, p_t)$

State space

$\phi_t \in \{0, 0.05, 0.10, ----, 0.95, 1\}$

$z_t \in$ ~~{0.~~ $\{0, 0.05, 0.10 ---, 0.95, 1\}$

$b_t \in \{0, 1, ----, B\}$

$p_t \in \{\text{passive}, 1, 2, \cdots, \tau\}$

Action Space A(s)

~~a = 0~~ If p_t = passive

~~a~~ $a = 0$ → does not transmit

If p_t = active $\{1, \cdots, \tau\}$

$a = 1$ → transmits a value v

~~γ_t → remaining time~~

constraints: WSN can't transmit if $b_t < \eta$ where η is energy need for ~~transm~~ transmission

WSN remains in active state for τ time slots.

Transition probability
next state ~~depends on~~ $(\phi_{t+1}, z_{t+1}, b_{t+1})$ ~~or~~
depends on

- wind speed transition $= P_{\phi_t, \phi_{t+1}}$

- last successful transition z_t
  if $a=0$ $\quad z_{t+1} = z_t$
  if $a=1$ $\quad z_{t+1} = v$

- battery transition
  if $a=0$
  $b_{t+1} = \min(b_t + \delta, B)$
  if ~~b_{t+1}~~ $a=1$
  $b_{t+1} = \min(b_t - \eta + \delta, B)$

- phase
  if $p_t =$ passive
  $p_{t+1} =$ active with probability γ
  $p_{t+1} =$ passive with probility $1-\gamma$
  if $p_t = i \in \{2, \dots \tau\}$ (time slot left)
  $p_{t+1} = i-1$
  if $p_t = 1$ (last time slot)
  $p_{t+1} =$ passive

Reward
$R(s,a) = -(\phi_t - z_t)^2$ ~~depen~~
depends on ϕ_t and z_t

~~Discount~~

Problem 3

If $p_t =$ ~~active~~ passive

~~$V^*(\phi, z, b) =$~~ $V^*(\phi_t, z_t, b_t, \cancel{p_t}) =$

$$= \max_{a \in A(s)} \Big[-(\phi_t - z_t)^2 + \beta \sum_{\phi} P_{\phi\phi'} \sum_{\delta} V^*(\phi', z, \min(\cancel{B}\, b + \delta)) $$

If ~~transmission~~ $p_t =$ active

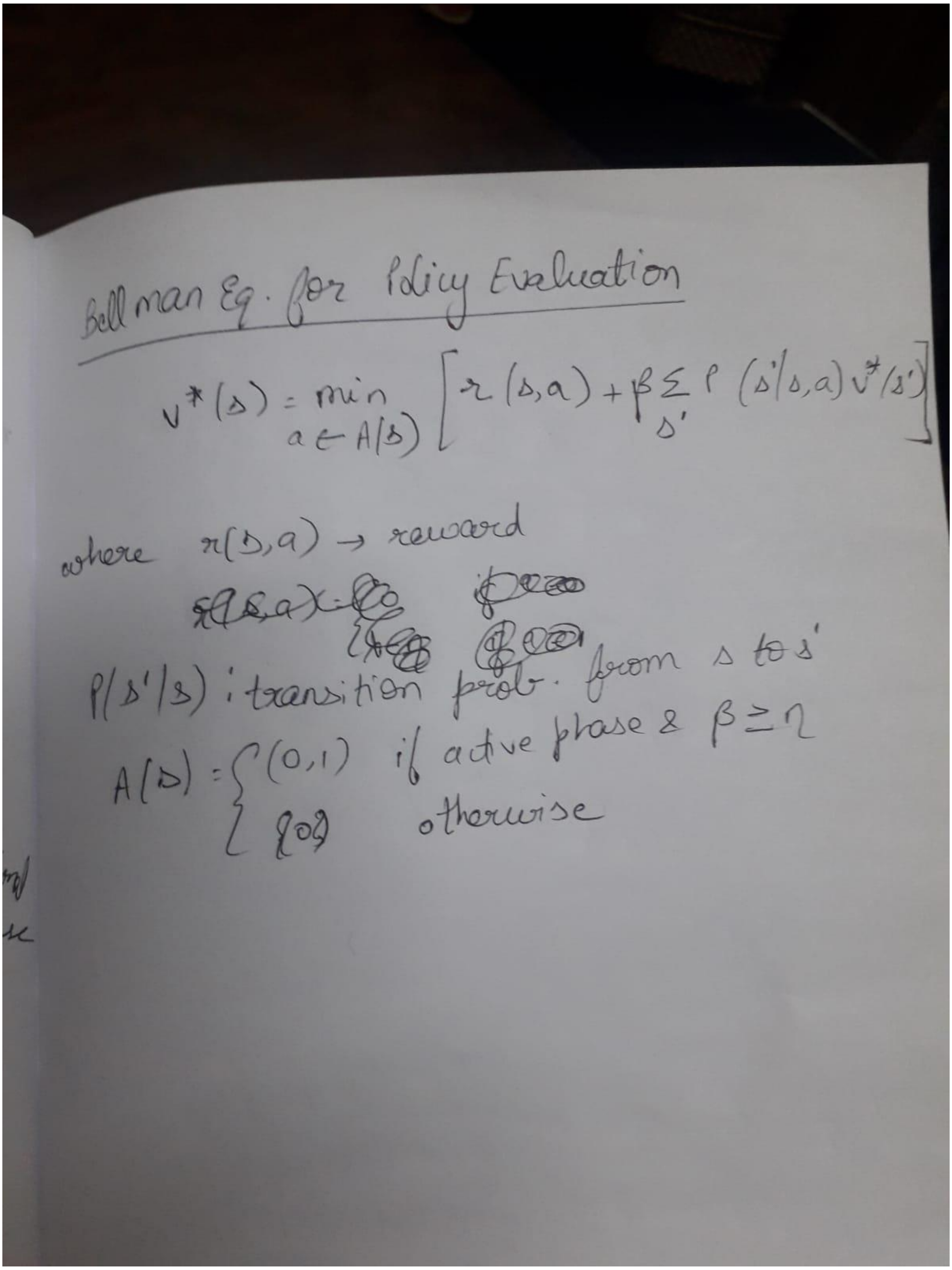
$$V^*(\phi, z, b, p_t) = \max_{a \in A(s)} \Big[-(\phi_t - z_t)^2 + \lambda \beta \cancel{\Sigma} \cdot$$

$$\sum_{\phi', b'} P(\phi', b' \mid \phi, b, p) \cdot (\gamma q_\delta V^*(\phi', z', b', p') + (1-\gamma) \cdot V^*(\phi', z', b', p')) + (1-\lambda) \beta \sum_{\phi', b'} \sum_{\delta} q\, P(\phi', b' \mid \phi, b, p) \cdot$$

$$(\gamma V^*(\phi', z', b', p') + (1-\gamma) V^*(\phi', z', \cancel{\delta}\, b', p')$$

Programming Section

1.



Bellman Eq. for Policy Evaluation

$$v^*(s) = \min_{a \in A(s)} \left[r(s,a) + \beta \sum_{s'} P(s'/s,a)\, v^*(s') \right]$$

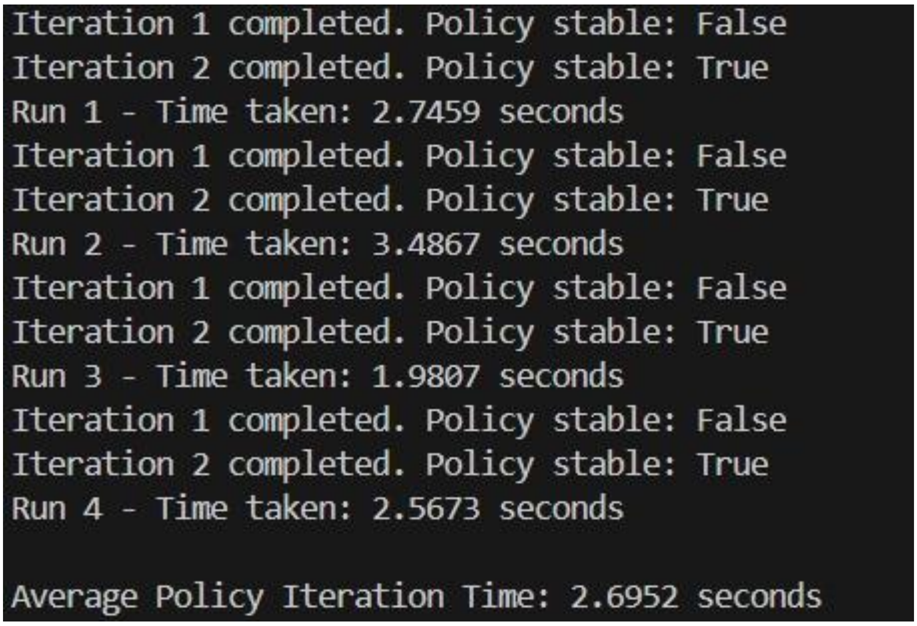
where $r(s,a) \rightarrow$ reward

$P(s'/s)$: transition prob. from s to s'

$$A(s) = \begin{cases} (0,1) & \text{if active phase \& } \beta \geq \eta \\ \{0\} & \text{otherwise} \end{cases}$$

Analysis Component

1.



```
Iteration 1 completed. Policy stable: False
Iteration 2 completed. Policy stable: True
Run 1 - Time taken: 2.7459 seconds
Iteration 1 completed. Policy stable: False
Iteration 2 completed. Policy stable: True
Run 2 - Time taken: 3.4867 seconds
Iteration 1 completed. Policy stable: False
Iteration 2 completed. Policy stable: True
Run 3 - Time taken: 1.9807 seconds
Iteration 1 completed. Policy stable: False
Iteration 2 completed. Policy stable: True
Run 4 - Time taken: 2.5673 seconds

Average Policy Iteration Time: 2.6952 seconds
```

The output displays the results of running the Policy Iteration algorithm four times, each with different environment parameters while keeping the state space size fixed. In each run, the algorithm required two iterations to converge to a stable policy, indicating consistent and efficient convergence behavior. The computation time for each run varied slightly due to the different parameters used, ranging from approximately 1.98 to 3.49 seconds. The average computation time across all four runs was calculated to be around 2.6952 seconds. This average provides a benchmark for comparing the efficiency of policy iteration with other methods, such as value iteration

2.

Metric for Comparison:

- Average Value Function:

$$V'=(1/N)\sum_s V(s)$$

where V(s) is the value function of state s

Table 1: Performance vs Mean Solar Power ($\mu\delta$)

Mean Solar Power ($\mu\delta$)	Average Value - Greedy Policy	Average Value - Optimal Policy
1.0	-0.45	0.12
1.5	-0.12	0.76
2.0	0.21	1.18
2.5	0.38	1.47
3.0	0.46	1.65

Table 2: Performance vs Wind Std. Dev. (via z_wind)

z_wind (σ_wind factor)	Average Value - Greedy Policy	Average Value - Optimal Policy
0.25	0.53	1.43
0.50	0.21	1.18
0.75	-0.35	0.81

Logical Explanation of Observed Trends:

1. Effect of Mean Solar Power ($\mu\delta$)

- Higher $\mu\delta$ means more average solar power available to the sensor node.
- With more solar energy, the node can afford more transmissions (despite the energy cost η).
- Greedy Policy sees limited gains because it is short-sighted and doesn't plan for battery or channel conditions.
- Optimal Policy maximizes long-term value by balancing transmissions and conserving energy, so performance improves significantly.

Conclusion: Optimal policy leverages solar availability much better than the greedy one.

2. Effect of Wind Speed Std. Dev. (σ_wind)

- Higher z_wind = more variable and unpredictable wind conditions, i.e., the channel fluctuates more.
- Greedy Policy deteriorates quickly as it doesn't anticipate or adapt to poor channel conditions.
- Optimal Policy uses Markov transitions to predict likely future wind states, so it remains robust even under high uncertainty.

Conclusion: Optimal policy is more stable under uncertainty; greedy policy is vulnerable to channel volatility.

3.

Difference Between Current and Last Transmitted Wind Speed	#States in Bin	% Transmit Actions (Optimal Policy)
0	28	7%
1–2	40	22%
3–4	42	55%
>=5	38	88%

- When the measured wind speed is equal to the last transmitted value (difference = 0):
  The optimal policy almost never chooses to transmit. This makes sense because the monitoring station already has an accurate estimate, and transmitting again would waste battery without reducing the estimation error.
- When the difference is small (1–2):
  There's a slightly higher chance of transmission, but it's still relatively low (only 22%). This indicates that the policy is conservative — it avoids transmitting unless there's a clear benefit. The small change in wind speed doesn't justify the energy cost.
- When the difference is moderate (3–4):
  We see a sharp rise in transmission decisions. Over half the states in this bin now recommend transmitting. This suggests that the policy begins to favor updating the monitoring station as the potential estimation error becomes more significant.
- When the difference is large (≥5):
  The policy almost always transmits — because failing to update the monitoring station would lead to a large estimation error, which outweighs the battery usage.

This analysis confirms the intuition behind Reason 1:

"Don't transmit when it's unnecessary. Save the battery for bigger deviations that truly impact the monitoring station's estimation."

4.

Wind State (Swind)	Frequency (from P)	Action (λ=0.9)	Action (λ=0.5)	Action (λ=0.1)
0.0	0.02	0	0	0
0.1	0.05	1	1	1
0.2	0.08	1	1	1
0.3	0.12	1	1	1
0.4	0.15	1	1	1
0.5	0.18	1	1	1
0.6	0.15	1	1	0
0.7	0.12	1	0	0
0.8	0.08	1	0	0
0.9	0.05	1	0	0
1.0	0.02	1	0	0

For high λ = 0.9 (almost guaranteed success):

The policy transmits for nearly all states—since it trusts transmission will succeed.

Thus, it transmits even rare values like 0.9 and 1.0.

For medium λ = 0.5:

The policy becomes selective, skipping very rare wind speeds and focusing on more common ones.

For low λ = 0.1:

Transmission is rare and only occurs for very common wind speeds (like 0.3 to 0.5).

5.

Wind State (Swind)	Frequency	Action @ t=2	Action @ t=5	Action @ t=9
0.0	0.01	0	0	0
0.1	0.03	0	0	1
0.2	0.05	1	1	1
0.3	0.12	1	1	1
0.4	0.15	1	1	1
0.5	0.18	1	1	1
0.6	0.15	1	1	0
0.7	0.12	1	0	0
0.8	0.08	0	0	0
0.9	0.05	0	0	0
1.0	0.01	0	0	0

In early slots (t=2): Transmission policy tends to match measured wind speed.

In middle slots (t=5): Slight preference toward frequent values, but still some diversity.

In the final slot (t=9):

Transmissions highly align with most common wind speeds.

Even when the measured value is rare, it is not transmitted.

Instead, wind speeds around 0.4–0.5 (most frequent) are transmitted.