**CS6362: Advanced Machine Learning**
**Assignment 3: Reinforcement Learning** Rana Banik

The table below provides the sequence of experience from each episode:

| #Episodes | States | Actions | Rewards | Return |
|---|---|---|---|---|
| 1 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, S] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 2 | [RU8p, RD10p, RD8a, RD10a] | [S, P, R, P] | [-1.0, 2.0, 0.0, 4.0] | 5 |
| 3 | [RU8p, RU10p, RU10a] | [R, P, R] | [0.0, 2.0, 0.0] | 2 |
| 4 | [RU8p, RD10p, RD8a, RD10a] | [S, P, R, P] | [-1.0, 2.0, 0.0, 4.0] | 5 |
| 5 | [RU8p, RD10p, RD8a, TD10a] | [S, R, P, R] | [-1.0, 0.0, 2.0, 3.0] | 4 |
| 6 | [RU8p, RD10p, RD10a] | [S, P, P] | [-1.0, 2.0, 4.0] | 5 |
| 7 | [RU8p, TU10p, RU8a, RD10a] | [P, R, S, R] | [2.0, 0.0, -1.0, 4.0] | 5 |
| 8 | [RU8p, RU10p, RU8a, RD10a] | [R, P, S, R] | [0.0, 2.0, -1.0, 4.0] | 5 |
| 9 | [RU8p, TU10p, RU10a] | [P, P, R] | [2.0, 2.0, 0.0] | 4 |
| 10 | [RU8p, RU10p, RU10a] | [R, P, P] | [0.0, 2.0, 0.0] | 2 |
| 11 | [RU8p, RU10p, RU8a, RD10a] | [R, P, S, P] | [0.0, 2.0, -1.0, 4.0] | 5 |
| 12 | [RU8p, RD10p, RD8a, RD10a] | [S, P, R, P] | [-1.0, 2.0, 0.0, 4.0] | 5 |
| 13 | [RU8p, TU10p, RU10a] | [P, P, P] | [2.0, 2.0, 0.0] | 4 |
| 14 | [RU8p, TU10p, RU8a, RU10a] | [P, R, R, R] | [2.0, 0.0, 0.0, 0.0] | 2 |
| 15 | [RU8p, RD10p, RD8a, TD10a] | [S, R, P, S] | [-1.0, 0.0, 2.0, 3.0] | 4 |
| 16 | [RU8p, TU10p, RU10a] | [P, P, R] | [2.0, 2.0, 0.0] | 4 |
| 17 | [RU8p, TU10p, RU10a] | [P, P, S] | [2.0, 2.0, 0.0] | 4 |
| 18 | [RU8p, RD10p, RD10a] | [S, P, P] | [-1.0, 2.0, 4.0] | 5 |
| 19 | [RU8p, RU10p, RU8a, TU10a] | [R, R, P, S] | [0.0, 0.0, 2.0, -1.0] | 1 |
| 20 | [RU8p, RD10p, RD8a, TD10a] | [S, R, P, R] | [-1.0, 0.0, 2.0, 3.0] | 4 |
| 21 | [RU8p, TU10p, RU8a, RU10a] | [P, R, R, R] | [2.0, 0.0, 0.0, 0.0] | 2 |
| 22 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, S] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 23 | [RU8p, TU10p, RU10a] | [P, P, P] | [2.0, 2.0, 0.0] | 4 |
| 24 | [RU8p, TU10p, RU10a] | [P, P, P] | [2.0, 2.0, 0.0] | 4 |
| 25 | [RU8p, TU10p, RU8a, RD10a] | [P, R, S, R] | [2.0, 0.0, -1.0, 4.0] | 5 |
| 26 | [RU8p, RU10p, RU8a, RU10a] | [R, P, R, S] | [0.0, 2.0, 0.0, 0.0] | 2 |
| 27 | [RU8p, RU10p, RU8a, TU10a] | [R, R, P, R] | [0.0, 0.0, 2.0, -1.0] | 1 |
| 28 | [RU8p, RU10p, RU8a, RD10a] | [R, P, S, R] | [0.0, 2.0, -1.0, 4.0] | 5 |
| 29 | [RU8p, RU10p, RU8a, TU10a] | [R, R, P, P] | [0.0, 0.0, 2.0, -1.0] | 1 |
| 30 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, P] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 31 | [RU8p, RU10p, RD8a, TD10a] | [R, S, P, P] | [0.0, -1.0, 2.0, 3.0] | 4 |
| 32 | [RU8p, RU10p, RD8a, RD10a] | [R, S, R, R] | [0.0, -1.0, 0.0, 4.0] | 3 |
| 33 | [RU8p, RD10p, RD8a, TD10a] | [S, P, P, P] | [-1.0, 2.0, 2.0, 3.0] | 6 |
| 34 | [RU8p, RU10p, RU10a] | [R, P, R] | [0.0, 2.0, 0.0] | 2 |
| 35 | [RU8p, RD10p, RD8a, RD10a] | [S, R, R, P] | [-1.0, 0.0, 0.0, 4.0] | 3 |
| 36 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, P] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 37 | [RU8p, TU10p, RU8a, RD10a] | [P, R, S, S] | [2.0, 0.0, -1.0, 4.0] | 5 |

| | | | | |
|---|---|---|---|---|
| 38 | [RU8p, RU10p, RU10a] | [R, P, S] | [0.0, 2.0, 0.0] | 2 |
| 39 | [RU8p, RD10p, RD8a, RD10a] | [S, P, R, S] | [-1.0, 2.0, 0.0, 4.0] | 5 |
| 40 | [RU8p, RU10p, RD8a, RD10a] | [R, S, R, R] | [0.0, -1.0, 0.0, 4.0] | 3 |
| 41 | [RU8p, TU10p, RU10a] | [P, P, R] | [2.0, 2.0, 0.0] | 4 |
| 42 | [RU8p, RU10p, RU8a, TU10a] | [R, R, P, R] | [0.0, 0.0, 2.0, -1.0] | 1 |
| 43 | [RU8p, TU10p, RU8a, RD10a] | [P, R, S, S] | [2.0, 0.0, -1.0, 4.0] | 5 |
| 44 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, P] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 45 | [RU8p, RD10p, RD8a, TD10a] | [S, P, P, P] | [-1.0, 2.0, 2.0, 3.0] | 6 |
| 46 | [RU8p, RU10p, RU8a, RU10a] | [R, P, R, R] | [0.0, 2.0, 0.0, 0.0] | 2 |
| 47 | [RU8p, TU10p, RU8a, TU10a] | [P, R, P, P] | [2.0, 0.0, 2.0, -1.0] | 3 |
| 48 | [RU8p, RU10p, RU10a] | [R, P, R] | [0.0, 2.0, 0.0] | 2 |
| 49 | [RU8p, RU10p, RD8a, RD10a] | [R, S, R, R] | [0.0, -1.0, 0.0, 4.0] | 3 |
| 50 | [RU8p, RU10p, RU8a, RU10a] | [R, P, R, R] | [0.0, 2.0, 0.0, 0.0] | 2 |
| avg: | | | | 3.5 |

Also, the average return of 50 episodes is: 3.5

The values of each state:

| No. | States | Values |
|---|---|---|
| 1 | RU8p | 3.513889 |
| 2 | TU10p | 1.666667 |
| 3 | RU10p | 2.5 |
| 4 | RD10p | 5.375 |
| 5 | RU8a | 1.333333 |
| 6 | RD8a | 4.5 |
| 7 | TU10a | -1 |
| 8 | RU10a | 0 |
| 9 | RD10a | 4 |
| 10 | TD10a | 3 |
| 11 | 11a | 0 |

3b.

Policy Evaluation:

| Iterations -> | 1 | | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|---|---|
| RU8p | P | 2 | P | 4 | P | 4 | P | 4 |
| TU10p | P | 2 | P | 2 | P | 2 | P | 2 |
| RU10p | P | 2 | P | 3 | P | 2.5 | P | 2.5 |
| RD10p | P | 2 | P | 5 | P | 6.5 | P | 6.5 |
| RU8a | P | 2 | P | 1 | P | 1 | P | 1 |
| RD8a | P | 2 | P | 5 | P | 5 | P | 5 |
| TU10a | P | -1 | P | -1 | P | -1 | P | -1 |
| RU10a | P | 0 | P | 0 | P | 0 | P | 0 |
| RD10a | P | 4 | P | 4 | P | 4 | P | 4 |
| TD10a | P | 3 | P | 3 | P | 3 | P | 3 |
| 11a | P | 0 | P | 0 | P | 0 | P | 0 |

Policy Improvement:

| | RU8p | TU10p | RU10p | RD10p | RU8a | RD8a | TU10a | RU10a | RD10a | TD10a | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Iterations: 1 | P | P | P | P | P | P | P | P | P | P | P |
| | 4.0 | 2.0 | 2.5 | 6.5 | 1.0 | 5.0 | -1.0 | 0.0 | 4.0 | 3.0 | 0.0 |
| Iterations: 2 | S | P | R | P | S | P | P | P | P | P | P |
| | 5.5 | 2.0 | 3.0 | 6.5 | 3.0 | 5.0 | -1.0 | 0.0 | 4.0 | 3.0 | 0.0 |
| Iterations: 3 | S | R | R | P | S | P | P | P | P | P | P |
| | 5.5 | 3.0 | 3.0 | 6.5 | 3.0 | 5.0 | -1.0 | 0.0 | 4.0 | 3.0 | 0.0 |
| Iterations: 4 | S | R | R | P | S | P | P | P | P | P | P |
| | 5.5 | 3.0 | 3.0 | 6.5 | 3.0 | 5.0 | -1.0 | 0.0 | 4.0 | 3.0 | 0.0 |

The bottom row actions and values show that after 3 iterations policy and value functions reach optimal point.