DSCI 6650: Reinforcement Learning Master of Data Science (MDSc)

Assignment- 1 Exploring Bandit Algorithms in Stationary and Non-Stationary Environments

Student: Hashna Binte Nahar
Programm: Data Science
Department: Mathematics and Statistics
MUN ID: 202290355
GitHub Repository: https://github.com/Hashna86/Project-1-for-reinforcement-learning.git

Abstract

This study investigates several classical multi-armed bandit (MAB) algorithms under both stationary and non-stationary reward conditions to explore their adaptability and performance. Algorithms including epsilon-greedy, greedy with zero initialization, optimistic greedy, and gradient bandit are implemented and evaluated. Non-stationary environments are simulated via gradual drifts, mean-reverting changes, and abrupt switches to reflect real-world uncertainties. Evaluation is based on average reward and optimal action selection across 1000 simulations of 2000 steps. Results highlight how different strategies perform across environmental dynamics and provide insights into optimal algorithm choices for different scenarios.

1. Stationary Bandit Environment
We first analyzed the algorithms in a 10-armed stationary setting where the true action-values $\mu_k \sim \mathcal{N}(0, 1)$ remain constant throughout each run.

Algorithms Used
- Greedy (Q=0): Exploits the current best estimate without exploration.

- Epsilon-Greedy ($\varepsilon = 0.1$): Chooses a random action with probability ε.

- Optimistic Greedy (Q = 3.89): Initialized with high estimates to encourage exploration.

- Gradient Bandit ($\alpha = 0.2$): Uses preference learning with softmax action selection.

Hyperparameter Tuning
Pilot runs (200 simulations, 400 steps) were conducted to tune the following:
- Epsilon values: {0.01, 0.05, 0.1, 0.2, 0.3, 0.4}
- Alpha values (gradient bandit): {0.1, 0.2, 0.4, 0.6, 0.8, 1.0}
- Optimistic Q-value: Chosen as 99.5th percentile of $\mathcal{N}(\mu\_max, 1)$, estimated as Q = 3.89

Results (Stationary Setting)

Algorithm	Final Avg Reward	% Optimal Action
Greedy (Q=0)	1.028	38.42%
Epsilon-Greedy (ε=0.1)	1.472	86.12%
Optimistic Greedy (Q=3.89)	1.388	65.26%
Gradient Bandit (α=0.2)	1.470	86.89%

Analysis

Greedy performed the worst due to lack of exploration. Epsilon-greedy and gradient bandit achieved higher performance, with gradient bandit being slightly more stable. Optimistic initialization helped early exploration but plateaued quickly.

2. Non-Stationary Bandit Environment

To mimic real-world uncertainties, we introduced changing reward structures. We studied three types of change:

2.1 Gradual Change (Drift)

Action values drifted over time using: $\mu_{k,t} = \mu_{k,t-1} + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, 0.01^2)$

2.2 Mean-Reverting Change

Action values reverted toward zero: $\mu_{k,t} = 0.5 \cdot \mu_{k,t-1} + \varepsilon_t$

2.3 Abrupt Change

At time step t = 501, action means were randomly permuted.

- Without Reset: Agent continues without resetting Q-values or preferences.
- With Reset: Q-values (or preferences) reset to initial state at t = 501.

Results Summary

Change Type	Algorithm	Avg Reward	% Optimal Action
Drift	Greedy	1.178	34.66
Drift	Epsilon-Greedy	1.596	71.27
Drift	Optimistic Greedy	1.466	56.5
Drift	Gradient Bandit	1.559	67.56
Mean-Reverting	Greedy	0.012	9.48
Mean-Reverting	Epsilon-Greedy	-0.003	9.85
Mean-Reverting	Optimistic Greedy	-0.016	8.52
Mean-Reverting	Gradient Bandit	-0.01	9.85
Abrupt (No Reset)	Greedy	0.522	19.05
Abrupt (No Reset)	Epsilon-Greedy	1.361	60.53
Abrupt (No Reset)	Optimistic Greedy	0.81	25.07
Abrupt (No Reset)	Gradient Bandit	0.928	36.51
Abrupt (With Reset)	Greedy	0.994	36.88
Abrupt (With Reset)	Epsilon-Greedy	1.412	81.53
Abrupt (With Reset)	Optimistic Greedy	1.025	36.68
Abrupt (With Reset)	Gradient Bandit	1.462	86.83

Discussion

1. Drift: Gradient bandit and epsilon-greedy adapted well; greedy remained rigid.
2. Mean-Reverting: All algorithms struggled due to high instability in mean updates.
3. Abrupt Changes:
   i) Without reset, algorithms showed degraded performance.
   ii)With hard reset, performance significantly improved, especially for gradient and epsilon-greedy methods.

Conclusion

Exploration-based algorithms such as epsilon-greedy and gradient bandit consistently outperformed pure exploitation strategies across environments. Gradient bandit showed strong robustness in both stationary and dynamic settings. Resetting internal values upon detected change points enhances adaptability. These findings reinforce the need for adaptable and exploratory strategies in uncertain environments.