E1:277 Reinforcement Learning

Assignment 1

Adaptive Gradient Bandit in a Non-Stationary Environment

Name: Varun
SR No.: 05-05-00-10-12-22-1-21129
Date: February 1, 2026

---

1 Experiment Summary

We evaluate a gradient bandit algorithm with an adaptive baseline in a non-stationary 10-armed bandit problem. The true action values drift every 500 steps using a Gaussian random walk.

Setup:

- 10 arms
- 200 runs, 2000 steps each
- Learning rate $\alpha = 0.1$
- Window size = 50
- $\beta \in \{0, 0.3, 0.6\}$

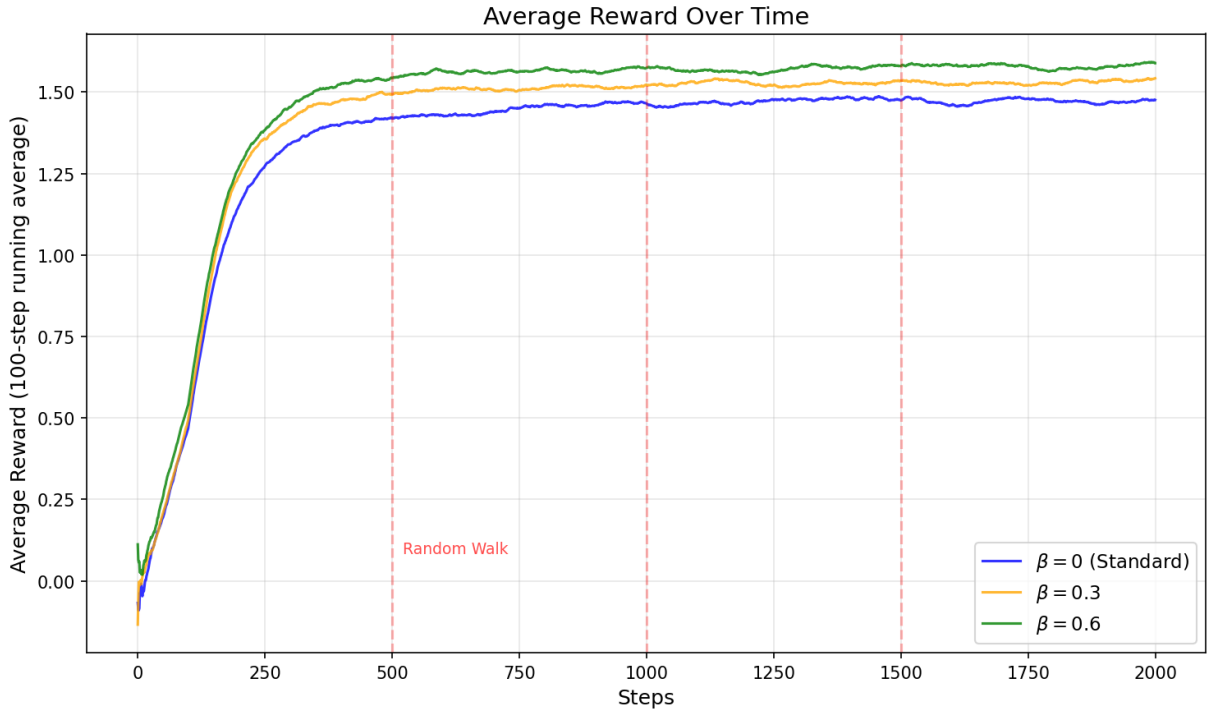
The adaptive baseline is

$$b_t = (1 - \beta)\bar{R}_t + \beta\tilde{R}_t$$

where $\bar{R}_t$ is the running mean and $\tilde{R}_t$ is a recent variance-adjusted mean.

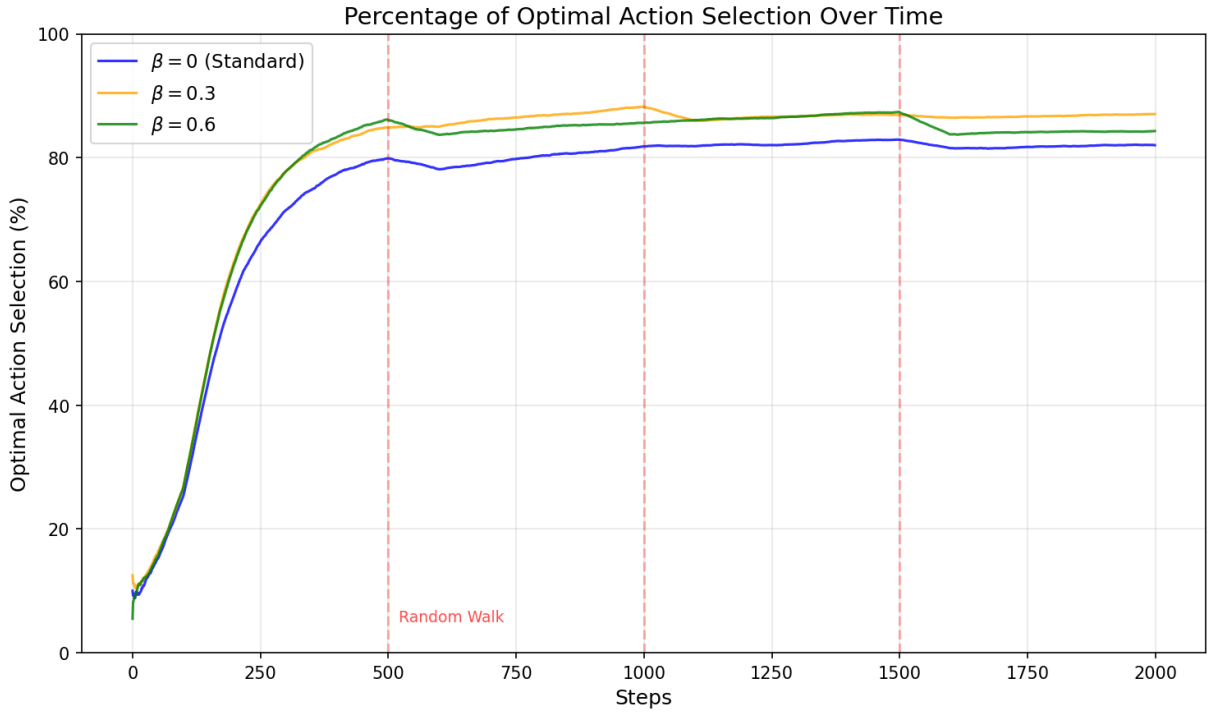
2 Results

2.1 Average Reward



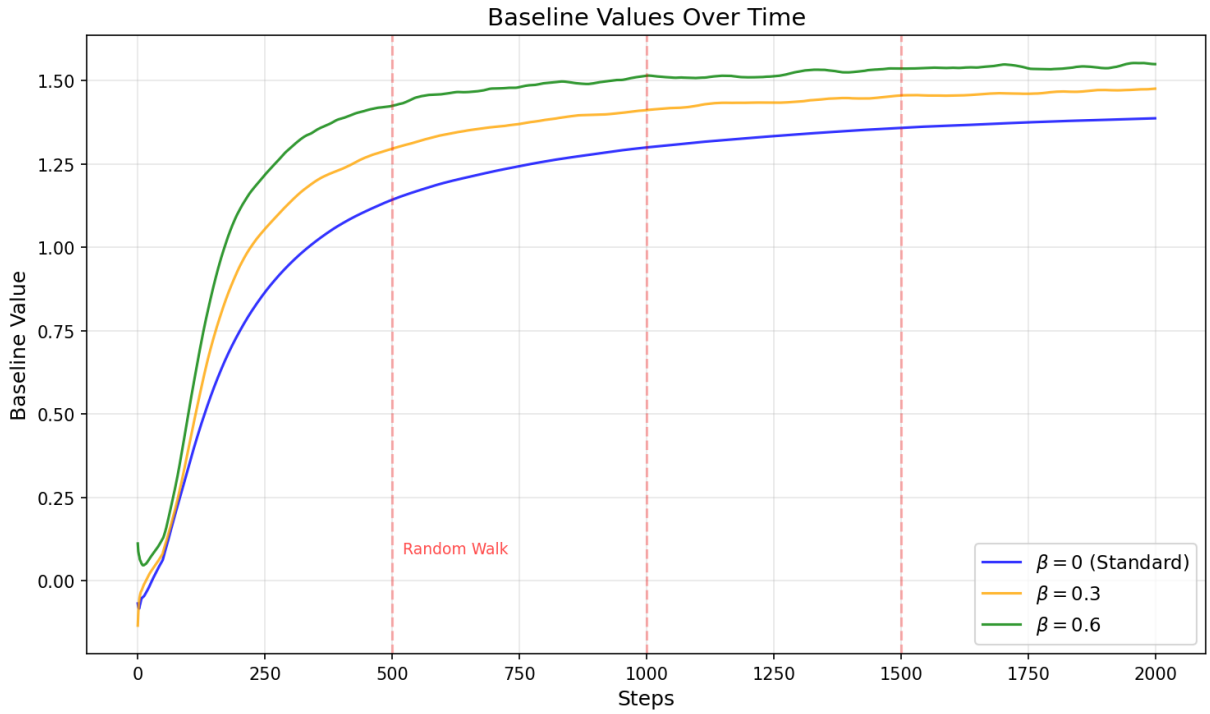
This plot shows the running average reward. After each environment drift (every 500 steps), adaptive baselines recover faster and maintain higher reward levels. The standard baseline reacts slowly because it relies heavily on older rewards.

2.2 Optimal Action Percentage



Optimal action percentage measures how often the agent selects the true best arm (the arm with the highest current mean). This metric directly reflects decision quality. Adaptive baselines achieve higher optimal-action rates, meaning they detect changes in the best arm more quickly. $\beta = 0.3$ shows the most consistent improvement.

2.3 Baseline Tracking



This figure shows how the baseline value evolves. Higher β values adjust the baseline faster after reward changes. $\beta = 0.6$ reacts the fastest but also fluctuates more, indicating higher variance.

3 Conclusions

Adaptive baselines clearly improve performance in non-stationary environments.

- Both $\beta = 0.3$ and $\beta = 0.6$ outperform the standard baseline.
- $\beta = 0.3$ provides the best balance between stability and responsiveness.
- $\beta = 0.6$ adapts quickly but shows slightly higher variability.

Thus, moderate adaptivity ($\beta = 0.3$) is the most reliable choice.