Restless Bandits

Deep Reinforcement Learning

CS698R - Prof. Ashutosh Modi

Group 4 - Brute Force

Members:

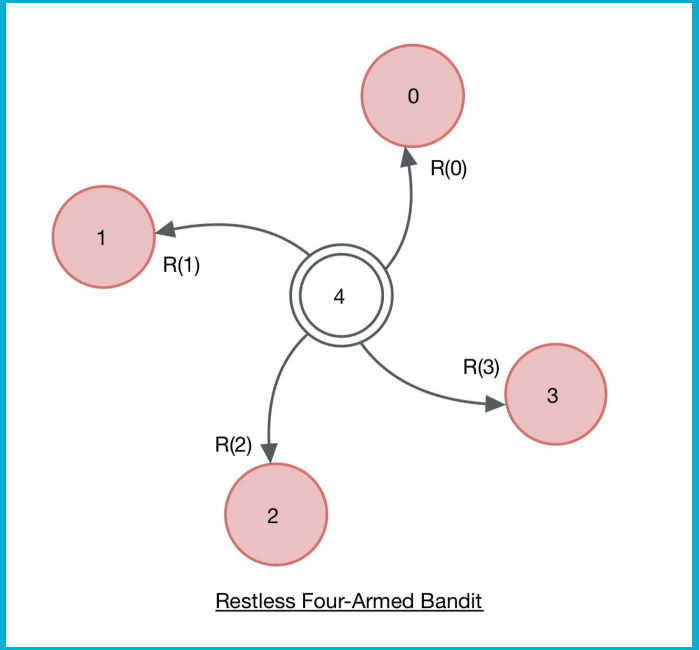
Arpit Agarwal	180139
Abhinav Kumar	16907018
Vartika Gupta	180849
Suman Singha	180793

Mentors: Harsh and Samik

Restless Bandit Problem & Motivation

In a nutshell

- A game between a player and an environment
- K arms to pull
- Each arm's reward evolves with an underlying distribution at each time step
- At each time step, player pulls an arm and receives a reward
- Goal: Maximize reward collected at the end of total T time steps

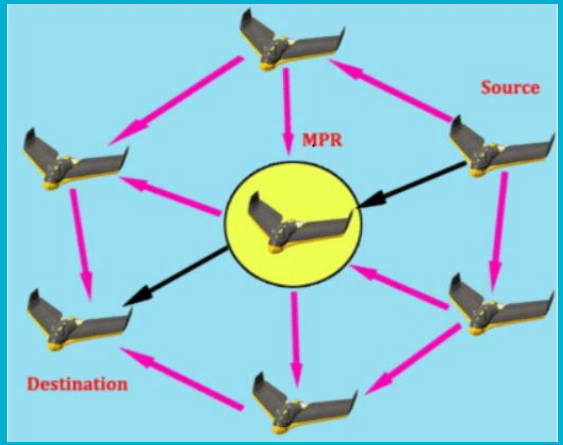


Restless Four-Armed Bandit

Restless Bandit Problem & Motivation

Has myriad applications in the real world

- Resource and Job Allocation in a Server
- Preventive Health Care Systems
- Dynamic Posted Pricing
- Dynamic UAV Routing
- Channel Detection in a Wireless Network



Dynamic UAV Routing [src] using Weighted RMAB

Related Work

01

Whittle Index & Q-Learning

1. On an index policy for restless bandits [1]
2. Towards Q-learning the Whittle Index for Restless Bandits [2]
3. A Reinforcement Learning Algorithm for Restless Bandits [3]
4. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare [4]

02

Bayesian Methods

1. Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits [6]

03

Convex Optimization

1. Weighted Restless Bandit and Its Applications [5]

04

Deep Reinforcement Learning

1. Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks [7]
2. Actor-Critic Deep Reinforcement Learning for Dynamic Multichannel Access [8]
3. Robust Restless Bandits: Tackling Interval Uncertainty with Deep Reinforcement Learning [9]

Related Work

- Extensions used in Related Work
  - Choosing K arms from total N arms
  - Minimizing cost associated with choosing K arms
  - Multi-agent setting for K different arms

Environment Details: Theory

- A single Non-Terminal starting state and four Terminal states
- Each episode consist of a single time step (1-step horizon)
- No stochasticity during transition from one state to another
- Reward corresponding to each action is sampled from a gaussian distribution with
  - A time-varying mean and
  - A fixed variance

➔ $R_j(t) = \mu_j(t) + \epsilon_j(t)$ $\quad \epsilon_j(t) \sim N(0,\sigma_\epsilon) \quad \forall\ j \in \{0,1,2,3\}$

➔ $\mu_j(t) = \lambda \mu_j(t-1) + k_j + \xi_j(t)$ $\quad \xi_j(t) \sim N(0,\sigma_\xi(t)) \quad \forall\ j \in \{0,1,2,3\}$

- k_j - trend parameter, λ - decay parameter
- Four versions of the environment:
  a. No trend and low volatility
  b. No trend and high volatility
  c. Trend and low volatility
  d. Trend and high volatility

Environment Details: Implementation

P = {

4 : {

0: [(1, 0, $N(\mu_0(t), \sigma_e)$, true] # state is same as action

1: [(1, 1, $N(\mu_1(t), \sigma_e)$, true] # transition probability is 1

2: [(1, 2, $N(\mu_2(t), \sigma_e)$, true] # μ_j's are function of time t

3: [(1, 3, $N(\mu_3(t), \sigma_e)$, true]

},

0 : {

0: [(1, 0, 0, true)] # all the terminal state list will be similar

1: [(1, 0, 0, true)]

2: [(1, 0, 0, true)]

3: [(1, 0, 0, true)]

},

}

Environment Details: Testing & Demo

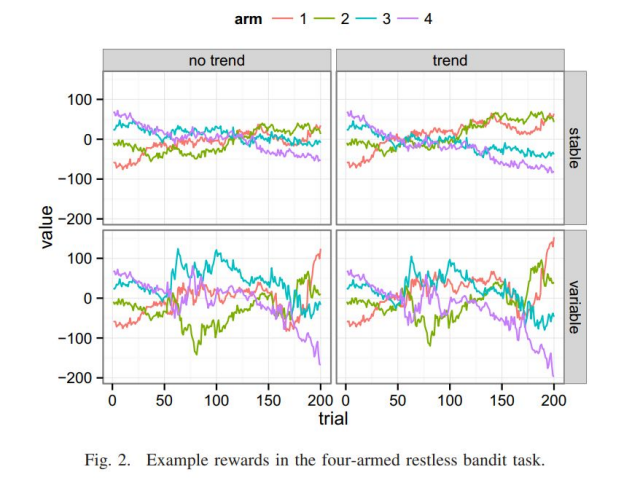
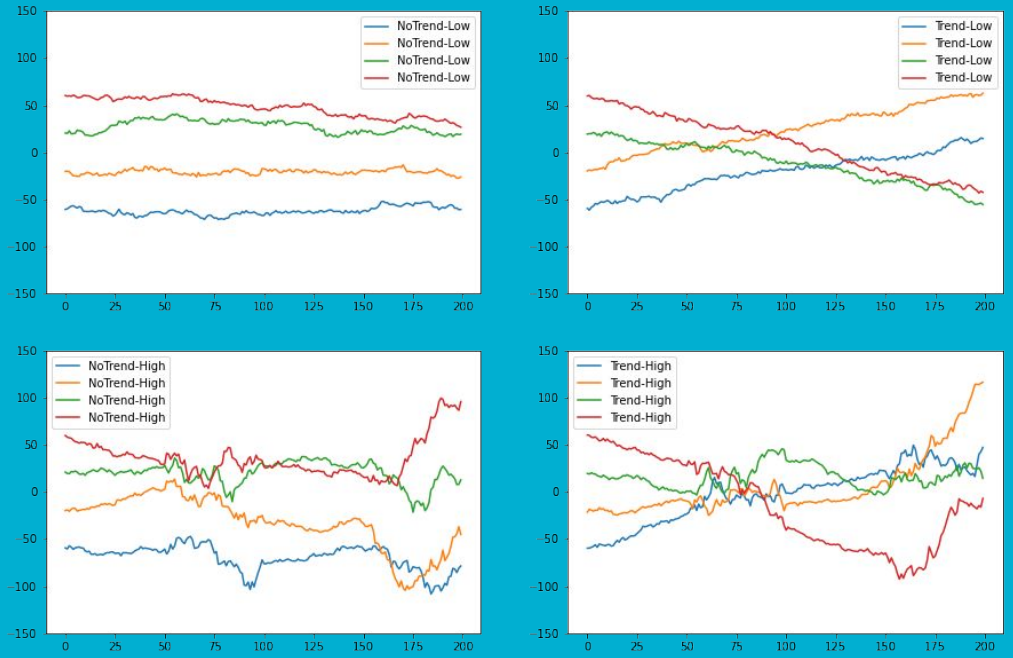


Fig. 2. Example rewards in the four-armed restless bandit task.

Image taken from
Uncertainty and Exploration in a Restless Bandit Problem [10]



Environment Test Average Results over 10 runs

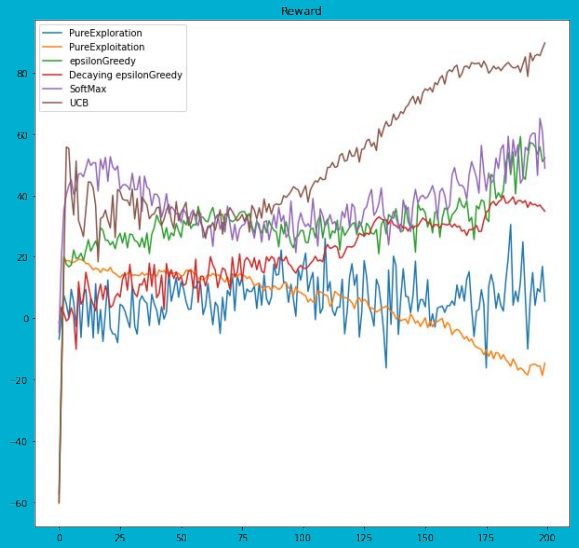
Demo

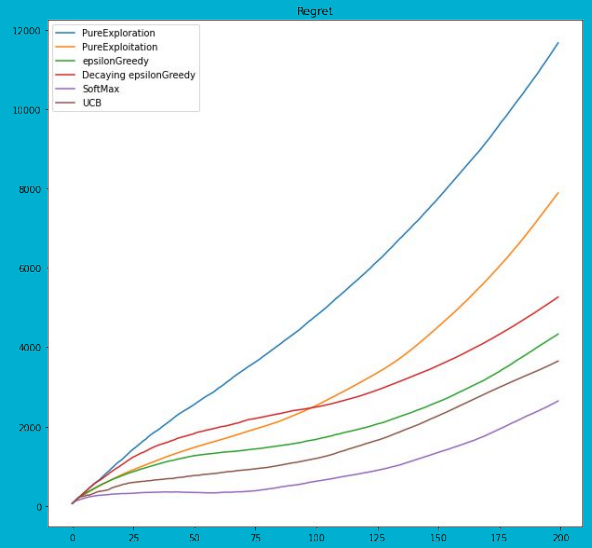
Baseline Implementation

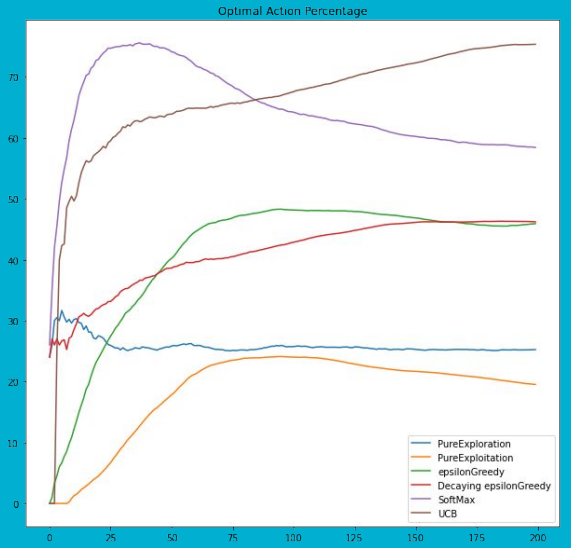
- Testing the environment with classical RL strategies
  - Pure Exploitation
  - Pure Exploration
  - Epsilon Greedy
  - Decaying Epsilon Greedy
  - Softmax
  - UCB

Trend & Low Volatility

Trend-Low

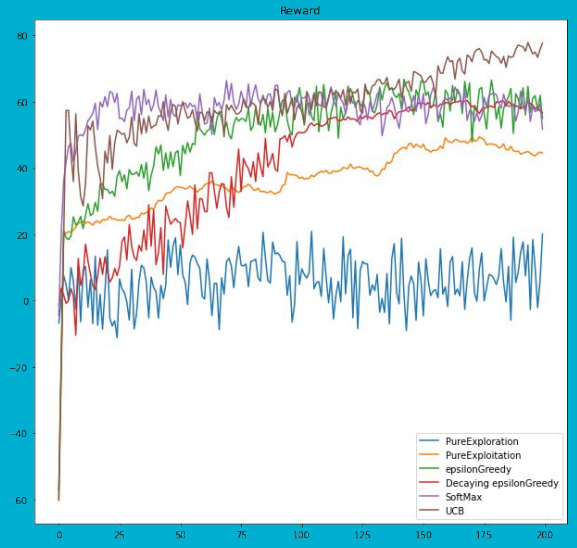




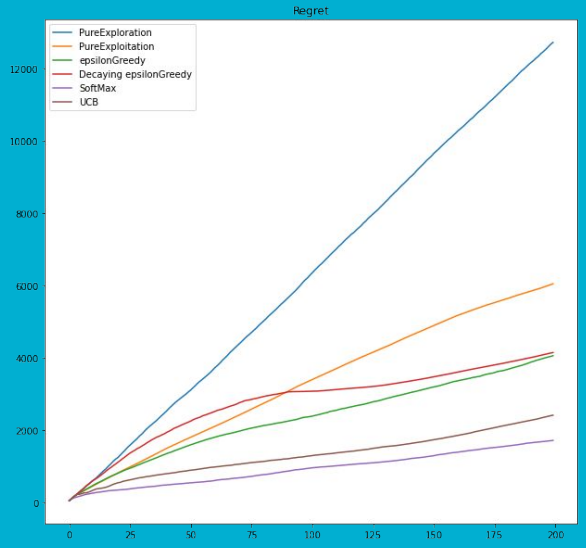


No Trend & Low Volatility

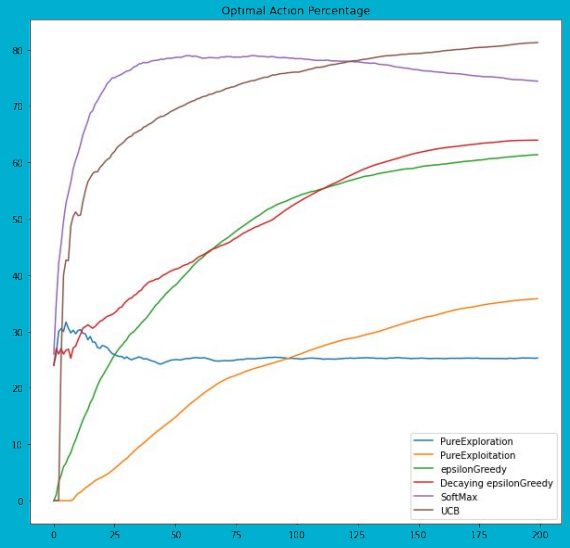
No Trend-Low



Reward



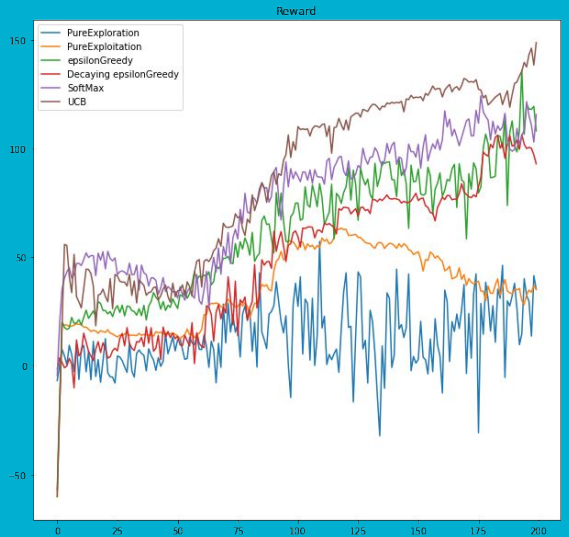
Regret



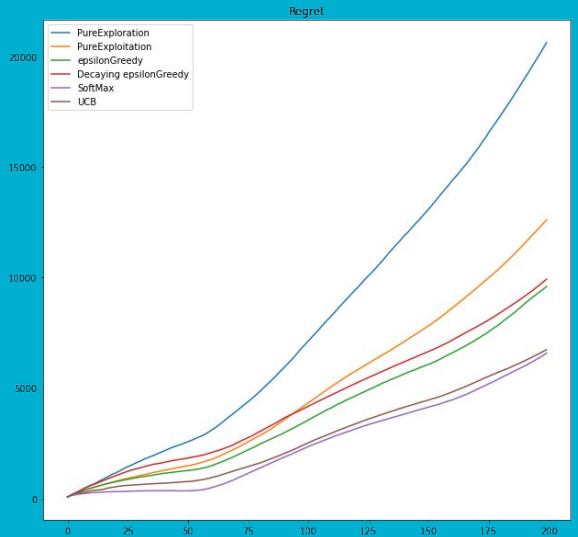
Optimal Action Percentage

Trend and High Volatility

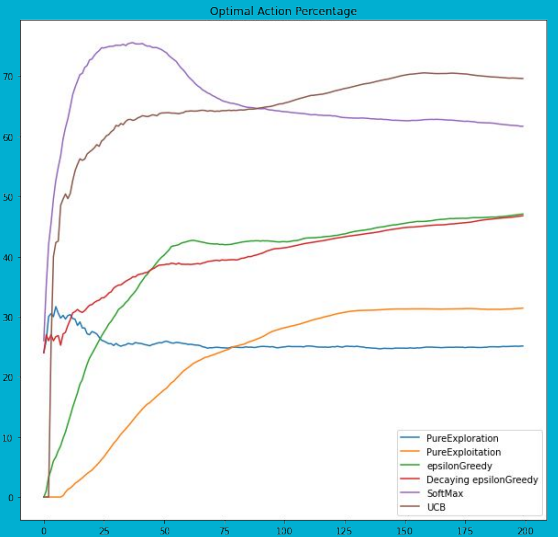
Trend-High



Reward



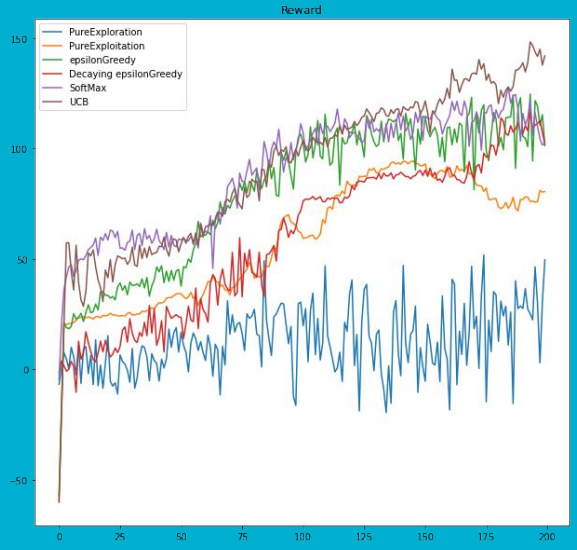
Regret

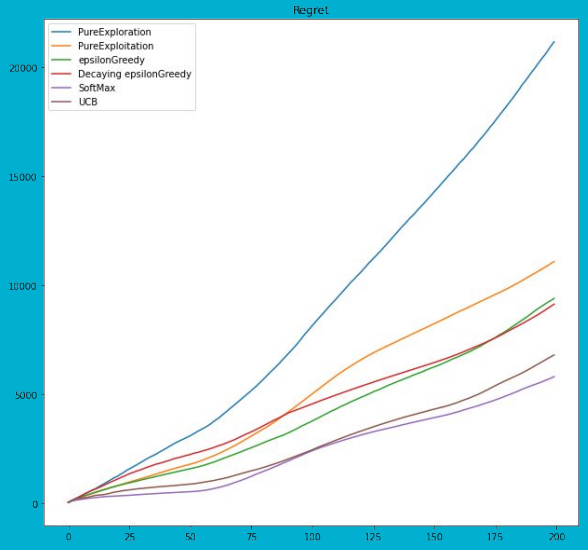


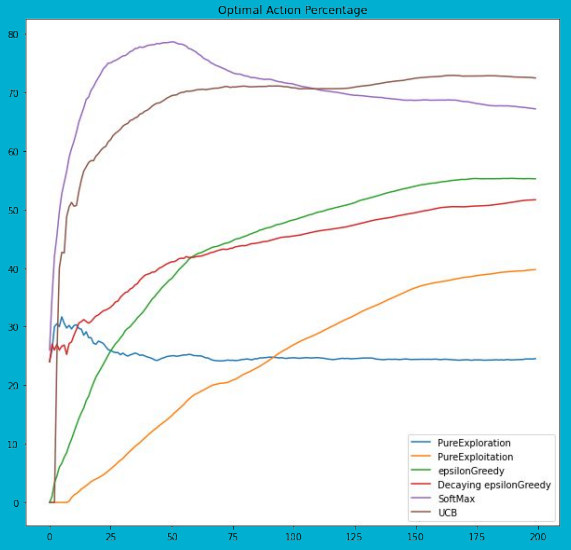
Optimal Action Percentage

No Trend & High Volatility

No Trend-High







Future Work

- Approaches
  - Deep RL ✔
  - Bayesian Methods
  - Classical RL

- Challenges
  - Scaling: in case of K actions (${}^{N}C_{K}$, eg: ${}^{50}C_{20} \cong 10^{12}$)
  - Online vs Offline: known and unknown MDP
  - Robustness: dealing with uncertainty

Future Work

- Robust Restless Bandits with Proximal Policy Optimization [9]
  - Minimax regret objective (policy that minimizes the maximum regret possible by transitions)
  - Real world scenarios with no knowledge of MDP and multi-action RMABs
  - Interval uncertainty in probability of transition instead of a fixed value - Robustness
  - Decoupling each arm's learning
  - Multi-agent setting where agents learn optimal policies and generates maximum possible regret

- Deep Q-Networks [7]
  - Learns underlying distributions of different arms over time
  - DQN comparison with
    - Myopic Policy (focuses on immediate reward)
    - Whittle Index Policy (assigns index values to states and selects the maximum)

- Actor-Critic [8]
  - Actor-Critic comparison with above mentioned DQN
  - Actor scores all possible actions for a state and selects the best one
  - Critic calculates TD error using the actor's action and provides feedback to actor

Timeline

Understanding and reviewing the mentioned approaches

Implementing and testing selected approaches

Improving upon the basic implementation and including extensions if possible

Collecting results and drawing observations

Contributions

Members	Work Done		
Arpit Agarwal	Literature Review	Future Work	PPT and Report
Abhinav Kumar	Literature Review	Future Work	PPT and Report
Vartika Gupta	Literature Review	Future Work	PPT and Report
Suman Singha	Environment Implementation	Future Work	PPT and Report

Thank You