Multi-Armed Bandit Problem:
A New Belief-Resilience Algorithm

Nick Hollman (nhollma@umich.edu)
Weinberg Institute for Cognitive Science, 500 Church St,
Ann Arbor, MI 48109 USA

Qianbo Yin (ygrayson@umich.edu)
Weinberg Institute for Cognitive Science, 500 Church St,
Ann Arbor, MI 48109 USA

Keywords: Multi-Armed Bandit; Belief Resilience; Exploration and Exploitation; Reinforcement Learning

Introduction

The Multi-Arm Bandit (MAB) Problem captures a dilemma in decision-making under uncertainty. Agents are faced with *n* choices that have various unknown rewards, in which they can either exploit choices with greater certainty for rewards or explore the unknown choices hoping for a better result. Ultimately the goal of each agent is to maximize the total rewards as much as possible.

In our current project, we develop a new algorithm based on the resilience of a belief each agent has towards the expected reward. As more information accumulates, the agent's belief becomes more resilient and consequently helps the agent to make better choices.

Existing Algorithms

The Multi-Armed Bandit problem is a well-researched problem in reinforcement learning. To test the performance of our new algorithm, we will compare it with the following previously developed algorithms:

Epsilon Greedy

1 - $\boldsymbol{\varepsilon}$ probability of exploitation

Epsilon First

$\boldsymbol{\varepsilon}$ * N number of random trials (exploration) followed by a phase of exploitation

Epsilon Decreasing

Same as epsilon greedy, but with a decreasing $\boldsymbol{\varepsilon}$:
($\boldsymbol{\varepsilon}$ = 1 / *n* + 1)

Pure Random

Arm is selected at random on each trial.

Upper Confidence Bound

Probability of choosing an arm is proportional to the probability of that arm giving the highest payoff.

Belief Resilience Algorithm

This algorithm is built on the assumption that a belief towards an expected reward falls on a spectrum of resiliency. Resiliency in beliefs relates to the amount of evidence and strength of justification. If a belief is low resilience, it has a high chance to be changed based on future evidence.

According to this algorithm, both the belief resiliency and estimation for reward are used in decision making. The exploration phase aims at increasing belief on all bandits, and the exploitation phase aims at optimizing robust, high rewards. The algorithm is formulated in Figure 1.

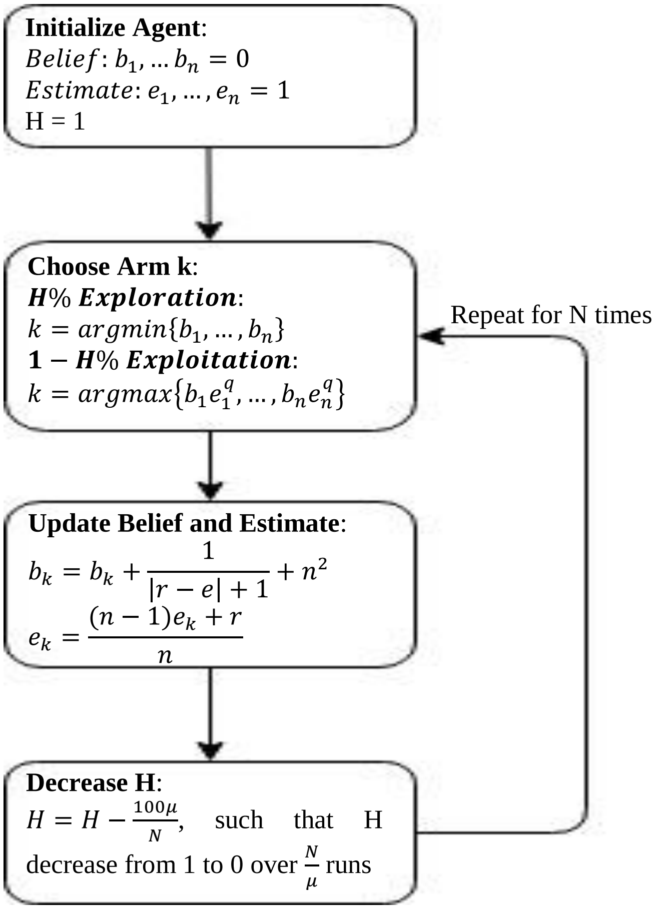
Initialize Agent:
$Belief: b_1, \dots b_n = 0$
$Estimate: e_1, \dots, e_n = 1$
H = 1
Choose Arm k:
$\boldsymbol{H\%\ Exploration}$:
$k = argmin\{b_1, \dots, b_n\}$
$\boldsymbol{1 - H\%\ Exploitation}$:
$k = argmax\{b_1 e_1^q, \dots, b_n e_n^q\}$
Repeat for N times
Update Belief and Estimate:
$b_k = b_k + \frac{1}{|r - e| + 1} + n^2$
$e_k = \frac{(n-1)e_k + r}{n}$
Decrease H:
$H = H - \frac{100\mu}{N}$, such that H decrease from 1 to 0 over $\frac{N}{\mu}$ runs

Figure 1: Belief Resilient Algorithm

Results and Discussion

First, we tested the undetermined parameters q and μ in the Belief-Resiliency Algorithm, generating the reward graph as a function of q and μ, shown in Figure 2. Therefore, we conclude the best parameter for the algorithm.

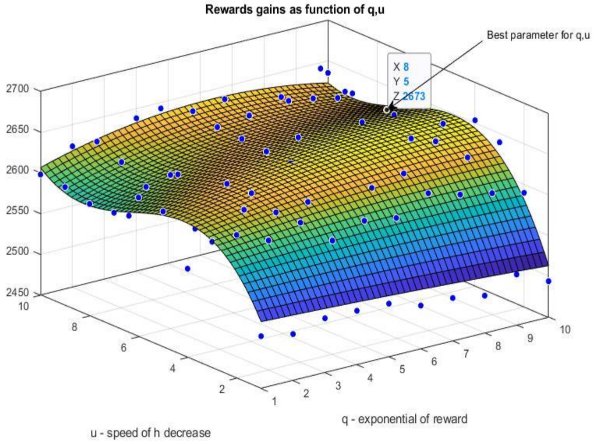

Figure 2: Determine parameters q,u in Belief-Resilience Algorithm. (z-axis: Total Reward)

After determining the parameters in the algorithm, we tested the Belief-Resilient Algorithm against the existing MAB algorithms. Results shown in Figure 3, 4 and 5.

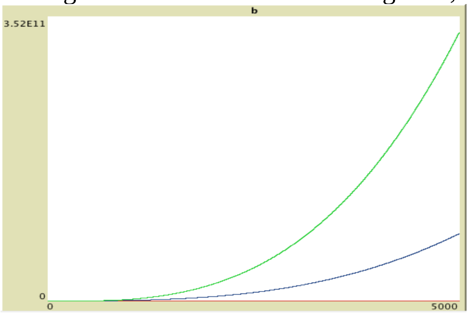

Figure 3: Belief factor (b) increase as a function of trials (plotting 4 arms)

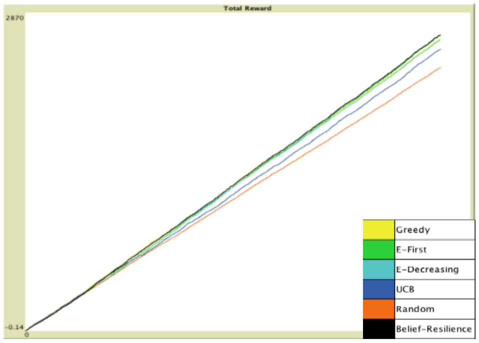

Figure 4: Comparison of 6 different algorithms (Total Rewards)

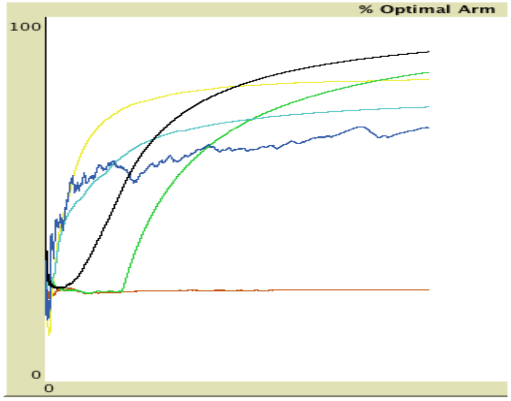

Figure 5: Percentage of Optimal Choices) for 6 different algorithms

Conclusion

After implementing and testing the Belief-Resilience Algorithm, we conclude that this algorithm competes with the standard existing reinforcement learning algorithms, with the optimal parameter q=10 and $\mu = 8$. In some cases, the new algorithm outperforms the leading algorithms in MAB paradigm. Generalizing the idea of belief resiliency in decision making, the robustness of belief can play a crucial role in evaluating a certain decision. Finally, we argue that the Belief-Resilience Algorithm, inspired by human beliefs and decision-making, is potentially an efficient algorithm of human decision making.

Future Directions

In further research, we would like to construct a more sophisticated relationship between *b* and the estimated reward in the Belief Resilience Algorithm. In addition, we would like to further expand the Multi-Armed Bandit problem to more diverse settings to model real-life decision-making situations.

Acknowledgment

We thank the Weinberg Institute for Cognitive Science (WICS) at the University of Michigan, Ann Arbor for its generous support. We highly appreciate the Undergraduate Research Symposium (URS) for providing us the opportunity to present our preliminary results.

References

SUTTON, RICHARD S. BARTO, ANDREW G. (2018). *REINFORCEMENT LEARNING: An introduction*. Cambridge: MIT Press.