## Response to Reviewer Comments on M.S. Thesis

*April 18, 2018*

Name: Subhojyoti Mukherjee

Roll Number: CS15S300

Registered for: Master of Science (by Research)

Guide : Dr. B. Ravindran, Co-Guide : Dr. N. Sudarsanam

# 1 Response to Reviewer 1

We thank the reviewer for his cogent comments and have made a few necessary changes as suggested by the reviewer.

1. **Abstract:**

   *Correction Suggested:* Start with a transparent sentence.

   *Correction Done:* Wrote a more specific sentence outlying the main theme of the thesis.

   This thesis studies the following topics in the area of Reinforcement Learning: classic Multi-armed bandits in stationary distribution with the goal of cumulative regret minimization using variance estimates and Thresholding bandits in pure exploration fixed-budget setting with the goal of instantaneous regret minimization also using variance estimates.

2. **Chapter 1, subsection 1.1 (Introduction to Bandits):**

   *Correction Suggested:* "In pursuit of these, the agent can be thought of as making a series of *sequential decisions*" to "series of tautological decisions".

   *Correction Done:* Kept the sentence in original form. It should be a "series of sequential decisions" in the setting considered.

   In pursuit of these, the agent can be thought of as making a series of *sequential decisions*...

3. **Chapter 2, subsection 2.5.2(The Upper Confidence Bound Approach)**

   *Correction Suggested:* Break the paragraph, starting from "Empirically...".

   *Correction Done:* Wrote it into two paragraphs now.

   UCB-Improved incurs a gap-dependent regret bound of $O\left(\frac{K\log(T\Delta^2)}{\Delta}\right)$, which is better than that of UCB1. On the other hand, the worst case gap-independent regret bound of UCB-Improved is $O\left(\sqrt{KT\log K}\right)$.

   Empirically, UCB-Improved is out-performed by UCB1 in almost all environments. This stems from the fact that UCB-Improved is pulling all arms equal number of times in each round and hence spends a significant number of pulls in initial exploration as opposed to UCB1 thereby incurring higher regret.

4. **Chapter 4, subsection 4.1 (Introduction to Thresholding Bandits):**

   *Correction Suggested:* Can go to section 1.9.

   *Correction Done:* Removed the line that "An interested reader can read through the previous chapters or can continue from here."

5. **Chapter 4, subsection 4.2 (Notations and Assumptions):**

   *Correction Suggested:* In the sentence "Whenever there is no ambiguity about the underlaying...." correct "underlaying"

   *Correction Done:* Corrected the spelling to, "Whenever there is no ambiguity about the underlying...."

6. **Chapter 4, subsection 4.6 (Challenges in the TBP seeting):**

   *Correction Suggested:* Write more succinctly.

   *Correction Done:* Rewrote the paragraph.

   Further, if we look closely into the TBP setting we will see that there are several similarities between the challenges in SMAB setting and the TBP setting. These challenges are as follows:-

   (a) The more closer an arm's expected reward mean ($r_i$) is to $\tau$, the more harder is the problem. This stems from the fact that it becomes increasingly difficult to discriminate between the arms lying above and below $\tau$.

   (b) Lesser the budget $T$, the more harder is the problem. This is because there are lesser number of pulls available and so the number of samples collected tends to be low.

   (c) The higher the variance of an arm's reward distribution $D_i$, the more harder is the problem. This is similar to the first case as it becomes harder to discriminate between the arms lying close to the threshold.

7. **Chapter 5, subsection 5.1 (Introduction):**

   *Correction Suggested:* Capitalize proper nouns such as "section 1" to "Section 1", "chapter 1" to "Chapter 1", "table 1" to "Table 1" and "figure 1" to "Figure 1".

   *Correction Done:* Rewrote all such usages in the thesis.

   The rest of the chapter is organized as follows. We elaborate our contributions in Section 5.2 and we present the AugUCB algorithm in Section 5.3. Our main theoretical result on expected loss is stated in Section 5.4. Section 5.5 contains numerical simulations on various testbeds to show the performance of AugUCB against state-of-the-art algorithms and finally, we summarize in Section 5.6.

8. **Chapter 5, subsection 5.4 (Proofs):**

   *Correction Suggested:* Proofs can be non-italics format style.

   *Correction Done:* I did not tamper with the format specified for MS Thesis, so kept the proof format style unchanged.

9. **Chapter 6, subsection 6.1 (Conclusions):**

   *Correction Suggested:* Some word changes.

   *Correction Done:* Rewrote the paragraph with the changes..

   In this thesis, we studied two different bandit problems, the stochastic multi-armed bandit (SMAB) with the goal of cumulative regret minimization and pure exploration stochastic thresholding bandit problem (TBP) with the goal of expected loss minimization. For the first problem, we devised a novel algorithm called Efficient UCB Variance (EUCBV) which enjoys an order optimal regret bound and performs very well in diverse stochastic environments. In the second part, the thresholding bandit problem, we came up with the novel algorithm called Augmented UCB (AugUCB) which is the first algorithm to use variance estimation for the considered TBP setting and also empirically outperforms most of the other algorithms.