Learning Reliable Policies in the Bandit Setting with Application to Adaptive Clinical Trials

Hossein Aboutalebi*[1], Doina Precup[1], and Tibor Schuster[2]

[1]Department of Computer Science, McGill University. Mila Quebec AI Institute
[2]Department of Family Medicine, McGill University

Abstract

The stochastic multi-armed bandit problem is a well-known model for studying the exploration-exploitation trade-off. It has significant possible applications in adaptive clinical trials, which allow for a dynamic change of patient allocation ratios. However, most bandit learning algorithms are designed with the goal of minimizing the expected regret. While this approach is useful in many areas, in clinical trials, it can be sensitive to outlier data especially when the sample size is small. In this article, we propose a modification of the BESA algorithm [Baransi, Maillard, and Mannor, 2014] which takes into account the variance in the action outcomes in addition to the mean. We present a regret bound for our approach and evaluate it empirically both on synthetic problems as well as on a dataset form the clinical trial literature. Our approach compares favorably to a suite of standard bandit algorithms.

Introduction

The multi-armed bandit is a standard model for researchers to investigate the exploration-exploitation trade-off, see e.g [Baransi, Maillard, and Mannor, 2014; Auer, Cesa-Bianchi, and Fischer, 2002; Sani, Lazaric, and Munos, 2012a; Chapelle and Li, 2011; Sutton and Barto, 1998] . Unlike fully sequential decision-making problems, multi-armed bandit problems are simple enough to allow for theoretical studies.

The multi-armed bandit problem consists of a set of arms, each of which generates a stochastic reward from a fixed but unknown distribution associated to it. The standard goal in this setting is to find the arm $\star$ which has the maximum expected reward $\mu_\star$ (or equivalently, minimum expected regret). The expected regret R_T is defined as the sum of the expected difference between the mean reward of the chosen arm a_t and the optimal arm until $t = T$:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{a_t})\right]$$

While this objective is very popular, there are practical applications, for example in medical research and AI safety [Garcıa and Fernández, 2015] where maximizing expected value is not sufficient, and it would be better to have an algorithm sensitive also to the variability of the outcomes of a given arm. For example, consider multi-arm clinical trials where the objective is to find the most promising treatment among a pool of available treatments. Due to heterogeneity in patients' treatment responses, considering only the expected mean may not be of interest [Austin, 2011]. Specifically, as the mean is usually sensitive to outliers and does not provide information about the dispersion of individual responses, the expected reward has only limited value in achieving a clinical trial's objective. This is especially true if some outcomes are very bad for the patients. Due to this issue, analysis of variance approaches for studying the effectiveness of the treatments were recently propose [Corbin-Berrigan et al., 2018]. In other studies like [Carandini, 2004], the variance itself is the source of interest. Also, the consistency of treatment effects among patients is essential, with the ideal treatment usually defined as the one which has a high positive response rate while showing low variability in response among patients.

In this paper, we tackle the problem of designing bandit algorithms that reflect both the mean and variability of the arms, by extending one of the recent algorithms in the bandit literature called BESA (Best Empirical Sampled Average) [Baransi, Maillard, and Mannor, 2014]. One of the main advantage of BESA compared to other existing bandit algorithms is that it is a non-parametric learning algorithm. This is especially useful when one does not have any prior knowledge or has insufficient prior knowledge about the different arms in the beginning. We establish regret bounds for the proposed algorithm, and we show that this new algorithm is superior to some of the past risk-averse learning algorithms like MV-LCB and ExpExp [Sani, Lazaric, and Munos, 2012a] in both simulated tasks as well as in some clinical trial tasks.

Background and Notation

We consider the standard bandit setting with action (arm) set $\mathcal{A}$, where each action $a \in \mathcal{A}$ is characterized by a reward distribution r_a bounded in the interval $[0, 1]$. The distribution for action a has mean μ_a and variance σ_a^2. Let $X_{a,i} \sim r_a$ denote the i-th reward sampled from the distribution of action a. All actions and samples are independent. The bandit

*hossein.aboutalebi@mail.mcgill.ca

problem is described as an iterative game where, on each step (round) t, the player (an algorithm) selects action (arm) a_t and observes sample $X_{a,N_{a,t}}$, where $N_{a,t} = \sum_{s=1}^{t} \mathbb{I}\{a_s = a\}$ denotes the number of samples observed for action a up to time t (inclusively). A policy is a distribution over $\mathcal{A}$. In general, stochastic distributions are necessary during the learning stage, in order to identify the best arm. We discuss the exact notion of "best" below.

We define $I_S(m, j)$ as the set obtained by sub-sampling without replacement j elements form the set S of size m. Let $\mathcal{X}_{a,t}$ denote the history of observations (records) obtained from action (arm) a up to time t (inclusively), such that $|\mathcal{X}_{a,t}| = N_{a,t}$. The notation $\mathcal{X}_{a,t}(\mathcal{I})$ indicates the set of sub-samples from $\mathcal{X}_{a,t}$, where sub-sample $\mathcal{I} \subset \{1, 2, \ldots, N_{a,t}\}$.

The multi-armed bandit was first presented in the seminal work of Robbins [Robbins, 1985]. It has been shown that under certain conditions [Burnetas and Katehakis, 1996; Lai and Robbins, 1985], a policy can have logarithmic cumulative regret:

$$\lim_{t\to\infty} \inf \frac{R_t}{\log(t)} \geqslant \sum_{a:\mu_a<\mu_\star} \frac{\mu_\star - \mu_a}{K_{\inf}(r_a; r_\star)}$$

where $K_{\inf}(r_a; r_\star)$ is the Kullback-Leibler divergence between the reward distributions of the respective arms. Policies for which this bound holds are called *admissible*.

Several algorithms have been shown to produce admissible policies, including UCB1 [Auer, Cesa-Bianchi, and Fischer, 2002], Thompson sampling [Chapelle and Li, 2011; Agrawal and Goyal, 2013] and BESA [Baransi, Maillard, and Mannor, 2014]. However, theoretical bounds are not always matched by empirical results. For example, it has been shown in [Kuleshov and Precup, 2014] that two algorithms which do not produce admissible policies, ε-greedy and Boltzmann exploration [Sutton and Barto, 1998], behave better than UCB1 on certain problems. Both BESA and Thompson sampling were shown to have comparable performance with Softmax and ε-greedy.

While the expected regret is a natural and popular measure of performance which allows the development of theoretical results, recently, some papers have explored other definitions for regret. For example, [Sani, Lazaric, and Munos, 2012b] consider a linear combination of variance and mean as the definition of regret for a learning algorithm A:

$$\widehat{MV}_t(A) = \widehat{\sigma}_t^2(A) - \rho\widehat{\mu}_t(A),$$

where $\widehat{\mu}_t$ is the estimate of the average of observed rewards up to time step t and $\widehat{\sigma}_t$ is a biased estimate of the variance of rewards up to time step t. The regret is then defined as:

$$R_t(A) = \widehat{MV}_t(A) - \widehat{MV}_{\star,t}(A),$$

where $\star$ is the optimal arm. According to [Maillard, 2013], however, this definition is going to penalize the algorithm if it switches between optimal arms. Instead, in [Maillard, 2013], the authors devise a new definition of regret which controls the lower tail of the reward distribution. However, the algorithm to solve the corresponding objective function seems time-consuming, and the optimization to be performed may be intricate. Finally, in [Galichet, Sebag, and Teytaud, 2013], the authors use the notion of conditional value at risk in order to define the regret.

Measure of regret

In this section, we will present our definition of regret which considers both the expected value and the variability of the reward of arms, but unlike [Sani, Lazaric, and Munos, 2012b], it does not penalize the algorithm if it switches between optimal arms.

Definition 0.1. ***Optimal arm (action):*** *$\star$ is an optimal arm if it maximizes the trade-off between the expected outcome and the variance:*

$$\star \in \arg\max_{a\in\mathcal{A}} (\mu_a - \rho\sigma_a^2), \tag{1}$$

for some fixed $\rho \geqslant 0$.

Definition 0.2. ***Consistency-aware regret:*** *The consistency-aware regret for a bandit algorithm $\mathcal{B}$ is defined as:*

$$\mathfrak{R}_T^{\mathcal{B}}(\rho) = \sum_{a\in\mathcal{A}} (\mu_\star - \rho\sigma_\star^2 - \mu_a + \rho\sigma_a^2)\mathbb{E}[N_{a,T}] \tag{2}$$

Note that when $\rho = 0$, the consistency-aware regret corresponds to the well-known expected regret.

$$\mathfrak{R}_T^{\mathcal{B}} = \sum_{a\in\mathcal{A}} (\mu_\star - \mu_a)\mathbb{E}[N_{a,T}]. \tag{3}$$

Note that when the context is clear, we will just use $\mathfrak{R}_T$.

It is clear that computing the consistency-aware regret is not feasible in a real environment, as we do not have access to the underlying distributions of arms. Hence, we define the following empirical mean and variance which will be used to estimate this regret in our algorithm:

Definition 0.3. ***Empirical mean and variance:*** *For an algorithm $\mathcal{B}$, the empirical mean and empirical variance of arm a up to time t is:*

$$\widehat{\mu}_{a,t} = \frac{1}{N_{a,t}} \sum_{i=1}^{N_{a,t}} r_{a,i} \tag{4}$$

$$\widehat{\sigma}_{a,t}^2 = \frac{1}{N_{a,t} - 1} \sum_{i=1}^{N_{a,t}} (r_{a,i} - \widehat{\mu}_{a,t})^2 \tag{5}$$

where $r_{a,i}$ is the ith reward obtained from pulling arm a.

Note that unlike in [Sani, Lazaric, and Munos, 2012b], the empirical estimation of the variance of an arm here is unbiased. We will exploit this feature later in our proofs.

For ease of notation, we define the value function for the set of records of an arm as follows:

Definition 0.4. ***Consistency-aware value function:*** *For a given record set $\mathcal{X}_{a,t}(\mathcal{I})$ of an arm a up to time step t, the corresponding value function is defined as:*

$$\widehat{v}(\mathcal{X}_{a,t}(\mathcal{I})) = \widehat{\mu}(\mathcal{X}_{a,t}(\mathcal{I})) - \rho\widehat{\sigma}^2(\mathcal{X}_{a,t}(\mathcal{I}_{a,t})). \tag{6}$$

In the next section, we are going to develop an algorithm which optimizes the consistency-aware regret using the quantities defined above in its estimation.

Proposed Algorithm

In order to optimize the consistency-aware regret, we build on the BESA algorithm, which we will now briefly review. As discussed in [Baransi, Maillard, and Mannor, 2014], BESA is a non-parametric approach for finding the optimal arm according to the expected mean regret criterion. Consider a two-armed bandit with actions a and $\star$,where $\mu_\star > \mu_a$, and assume that $N_{a,t} < N_{\star,t}$ at time step t. In order to select the next arm for time step $t+1$, BESA first sub-samples $s_\star = I_\star(N_{\star,t}, N_{a,t})$ from the observation history (records) of the arm $\star$ and similarly sub-sample $s_a = I_a(N_{a,t}, N_{a,t}) = \mathcal{X}_{a,t}$ from the records the arm a. If $\widehat{\mu}_{s_a} > \widehat{\mu}_{s_\star}$, BESA chooses arm a, otherwise it chooses arm $\star$.

The main reason behind the sub-sampling is that it gives a similar opportunity to both arms. Consequently, the effect of having a small sample size, which may cause bias in the estimates, diminishes. When there are more than two arms, BESA runs a tournament algorithm on the arms [Baransi, Maillard, and Mannor, 2014].

Finally, it is worth mentioning that the proof of the regret bound of BESA uses a non-trivial lemma for which authors did not provide any formal proof. In this paper, we will avoid using this lemma to prove the soundness of our proposed algorithm.

We are now ready to outline our proposed approach, which we call BESA+. As in [Baransi, Maillard, and Mannor, 2014], we focus on the two-arm bandit. For more than two arms, a tournament can be set up in our case as well.

Algorithm BESA+ two action case

Parameters: current time step t, actions a and b. Initially $N_{a,0} = 0, N_{b,0} = 0$

Shuffle the arms a and b with a function $\mathcal{M}$ to get a', b'.

if $N_{a',t-1} = 0 \vee N_{a',t-1} < \log(t)$ **then**
  $a_t = a'$
else if $N_{b',t-1} = 0 \vee N_{b',t-1} < \log(t)$ **then**
  $a_t = b'$
else
  $n_{t-1} = \min\{N_{a',t-1}, N_{b',t-1}\}$
  $\mathcal{I}_{a',t-1} \leftarrow I_{a'}(N_{a',t-1}, n_{t-1})$
  $\mathcal{I}_{b',t-1} \leftarrow I_{b'}(N_{b',t-1}, n_{t-1})$
  Calculate $\tilde{v}_{a',t} = \widehat{v}(\mathcal{X}_{a',t-1}(\mathcal{I}_{a',t-1}))$ and $\tilde{v}_{b',t} = \widehat{v}(\mathcal{X}_{b',t-1}(\mathcal{I}_{b',t-1}))$
  $a_t = \arg\max_{i \in \{a',b'\}} \tilde{v}_{i,t}$ (break ties by choosing arm with fewer tries)
end if
return $\mathcal{M}^{-1}(a_t)$

The first major difference between BESA+ and BESA is the use of the consistency-aware value function instead of the simple regret. A second important change is that BESA+ selects the arm which has been tried less up to time step t if the arm has been chosen less than $\log(t)$ times up to t. Essentially, this change in the algorithm is negligible in terms of establishing the total expected regret, as we cannot achieve any better bound than $\log(T)$, as shown in Robbins' lemma [Lai and Robbins, 1985]. This tweak also turns out to be vital in proving that the expected regret of the BESA+ algorithm is bounded by $\log(T)$ (a result which we present shortly).

To better understand why this modification is necessary, consider a two arms scenario. The first arm gives a deterministic reward of $r \in [0, 0.5)$ and the second arm has a uniform distribution in the interval [0,1] with the expected reward of 0.5. If we are only interested in the expected reward ($\rho = 0$), the algorithm should ultimately favor the second arm. On the other hand, there exists a probability of r that the BESA algorithm is going to constantly choose the first arm if the second arm gives a value less than r on its first pull. In contrast, BESA+ evades this problem by letting the second arm be selected enough times such that it eventually becomes distinguishable from the first arm.

We are now ready to state the main theoretical result of our proposed algorithm.

Theorem 0.1. *Let $\mathcal{A} = \{a, \star\}$ be a two-armed bandit with bounded rewards $\in [0, 1]$, and the value gap $\Delta = v_\star - v_a$. Given the value ρ, the expected consistency-aware regret of the Algorithm BESA+ up to time T is upper bounded as follows:*

$$\mathfrak{R}_T = C_{\Delta,\rho} + O(\log(T)) \tag{7}$$

where in (7)*, $C_{\Delta,\rho}$ is a constant which is dependent on the value of ρ, Δ.*

Interested reader can visit **here** to see the full proof.

Empirical results

Empirical comparison of BESA and BESA+

As discussed in the previous section, BESA+ has some advantages over BESA. We illustrate the example we discussed in the previous section through the results in Figures 1-6, for $r \in \{0.2, 0.3, 0.4\}$. Each experiment has been repeated 200 times. Note that while BESA has an almost a linear regret behavior, BESA+ can learn the optimal arm within the given the time horizon and its expected accumulated regret is upper bounded by a log function. It is also easy to notice that BESA+ has a faster convergence rate compared with BESA. As r gets closer to 0.5, the problem becomes harder. This phenomenon is a direct illustration of our theoretical result.

Statistical dispersion estimate via sub-sampling without replacement

In this subsection, we are going to explore the effect of sample size and sub-sample size on the consistency-aware value function error of BESA+. We have studied different distributions to find out their effect on consistency-aware value function error as well. Average results and standard error bars are computed over 200 independent experiments in all graphs.

Based on our experiments, changing the ρ value in the consistency-aware value function definition does not have much impact on the convergence rate. Moreover, as one would expect, the distribution type does not have a noticeable influence on the convergence rate either, although some small differences can be observed. Due to the space limit, we only included two figures (figures 7, 8) to illustrate our claim. As it can be seen in both figures, as we increase the sub-sample

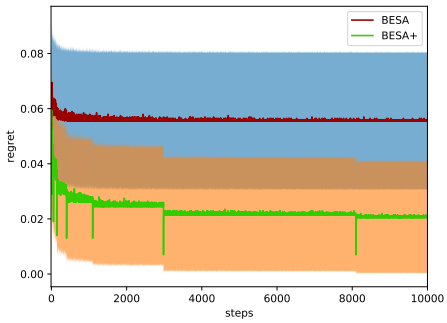

Figure 1: Result of expected regret per step for $r = 0.4, \rho = 0$

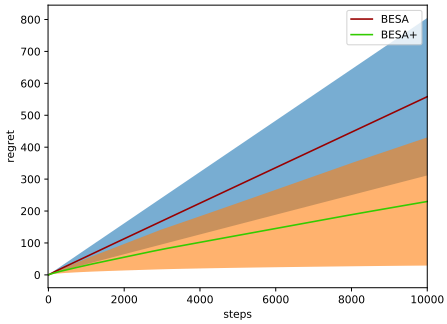

Figure 2: Result of accumulated expected regret for $r = 0.4, \rho = 0$

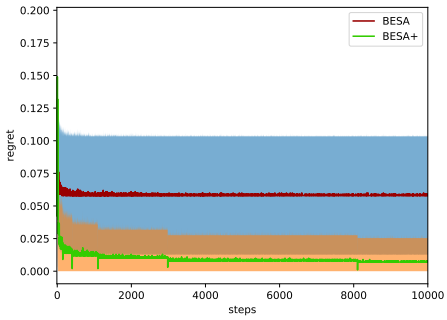

Figure 3: Result of expected regret per step for $r = 0.3, \rho = 0$

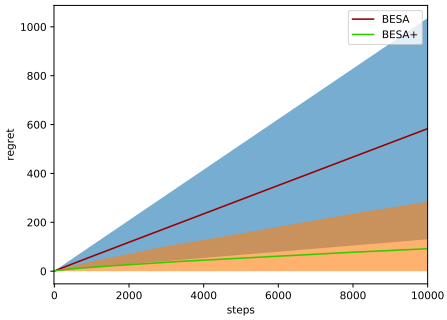

Figure 4: Result of accumulated expected regret for $r = 0.3, \rho = 0$

size, the expected error decreases significantly. It is also in-

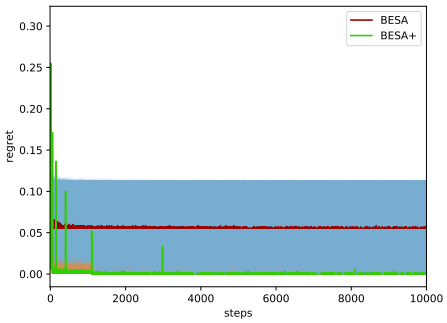

Figure 5: Result of expected regret per step for $r = 0.2, \rho = 0$

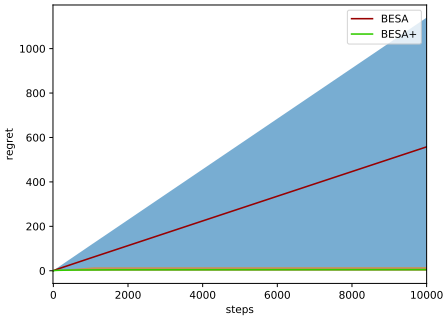

Figure 6: Result of accumulated expected regret for $r = 0.2, \rho = 0$

teresting to note that the expected error is almost independent of the sample size given the sub-sample size.

Algorithm BESA+ performance

We also evaluated the performance of BESA+ with consistency-aware regret in a simulated environment. Our work here is similar to the work [Sani, Lazaric, and Munos, 2012b] in which the authors depicted the performance of MV-LCB and ExpExp with a synthetic environment. Here, we considered a two-arm environment with arm 1 having mean 1 and variance in the range $[0.1, 1]$ and arm 2 having the mean in the range $[0.1, 1]$ and variance 1. It is clear that under any positive value of ρ, arm 1 should be preferred over arm 2. We have examined different values of ρ and studied their corresponding effects on the expected consistency-aware regret of the Algorithm BESA+. (figures 9, 10, 11). In the figures, n stands for time step. These figures uncover three important aspects of BESA+. First, we can observe the kind of problems which are difficult for BESA+. It appears that as the difference between the mean or variance of two arms shrinks, BESA+ usually suffers a higher amount of regret. This fact can also be inferred from Theorem 0.1. Second, we can see the importance of ρ value in diminishing the effect of variance or mean. In figure 9, where $\rho = 1$, we can observe a bump near the squares where both mean and variance gaps are small. In figure 10, when $\rho = 10$, the effect of the mean gap almost vanishes and we see that as we go from figure 9 to figure 10, the regret graph orients itself toward the smaller variance gap. The same thing happens as we go from figure 10 to 11. In this regard, in figure 11 (when $\rho = 0.1$), the

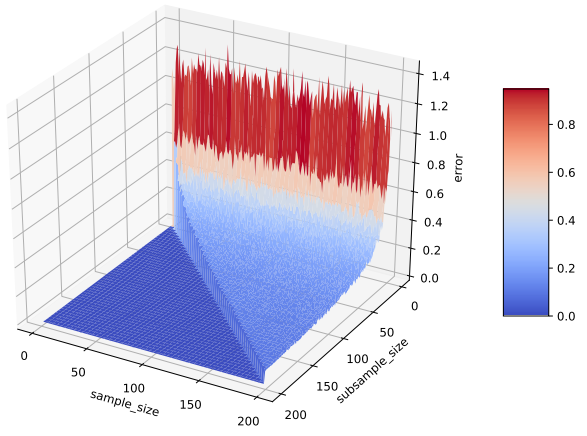

Figure 7: Result of expected error for normal distribution ($\mu = 0$ and $\sigma = 1$)

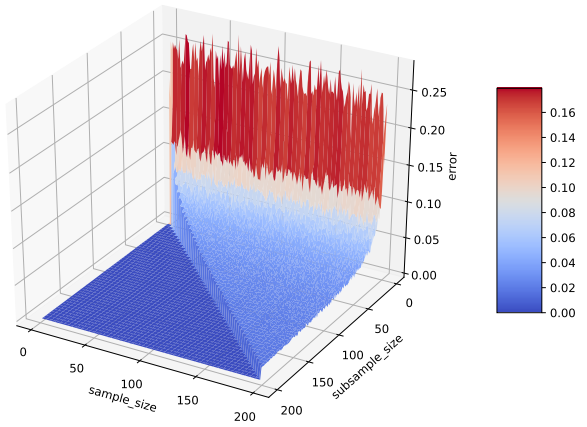

Figure 8: Result of expected error for uniform distribution of the interval $[0, 1]$

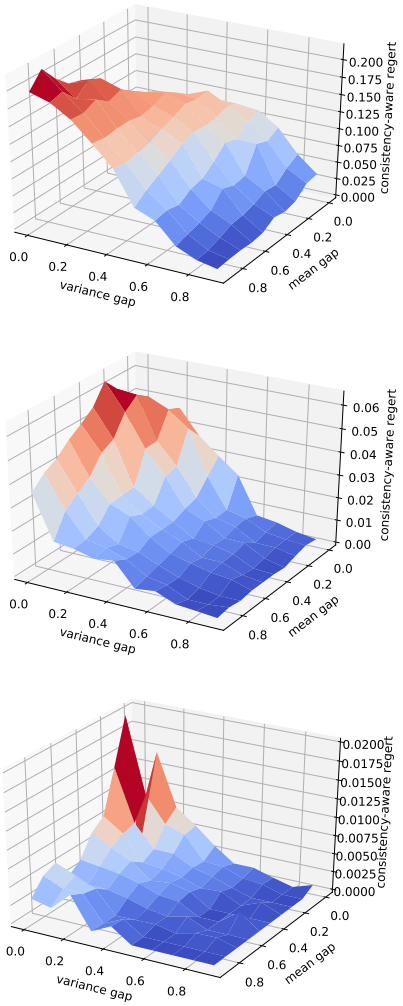

Figure 9: Case: $\rho = 1$, Top figure: $n = 20$. Middle figure: $n = 200$. Bottom figure: $n = 2000$.

regret graph has oriented itself toward the smaller mean gap size. Finally, these figures depict the speed of convergence of BESA+ Algorithm which seems faster than MV-LCB and ExpExp.

Real Clinical Trial Dataset

Finally, we examined the performance of BESA+ against other methods (BESA, UCB1 , Thompson sampling, MV-LCB, and ExpExp) based on a real clinical dataset. This dataset includes the survival times of patients who were suffering from lung cancer [Ripley et al., 2013]. Two different kinds of treatments (standard treatment and test treatment) were applied to them and the results are based on the number of days the patient survived after receiving one of the treatments. For the purpose of illustration and simplicity, we assumed non-informative censoring and equal follow-up times in both treatment groups. As the experiment has already been conducted, to apply bandit algorithms, each time a treatment is selected by a bandit algorithm, we sampled uniformly from the recorded results of the patients whom received that selected treatment and used the survival time as the reward signal. Figure 12 shows the distribution of treatment 1 and 2. We categorized the survival time into ten categories (category 1 showing the minimum survival time). It is interesting to notice that while treatment 2 has a higher mean than treatment 1 due to the effect of outliers, it has a higher level of variance compared to treatment 1. From figure 12 it is easy to deduce that treatment 1 has a more consistent behavior than treatment 2 and a higher number of patients who received treatment 2 died early. That is why treatment 1 may be preferred over treatment 2 if we use the consistency-aware regret. In this regard, by setting $\rho = 1$, treatment 1 has less expected mean-variance regret than treatment 2, and it should be ultimately favored by the learning algorithm. Figure 13 illustrates the performance of different bandit algorithms. It is easy to notice that BESA+ has relatively better performance than all the other ones.

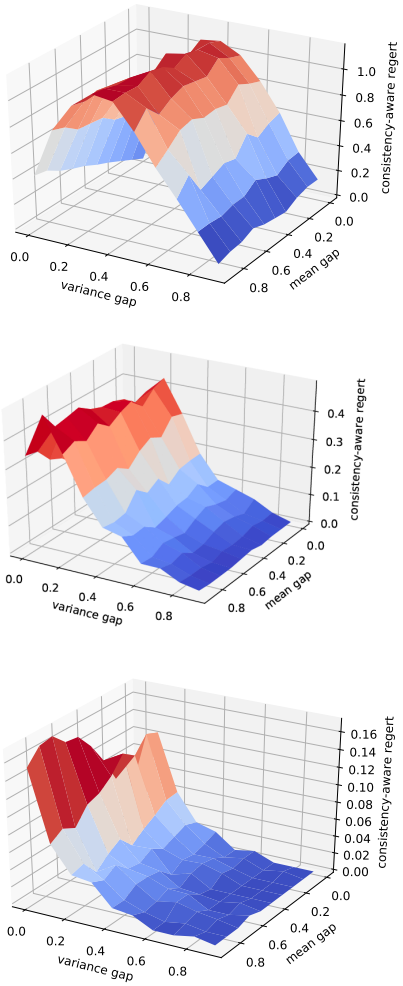

Figure 10: Case: $\rho = 10$, Top figure:$n = 20$. Middle figure: $n = 200$. Bottom figure: $n = 2000$.

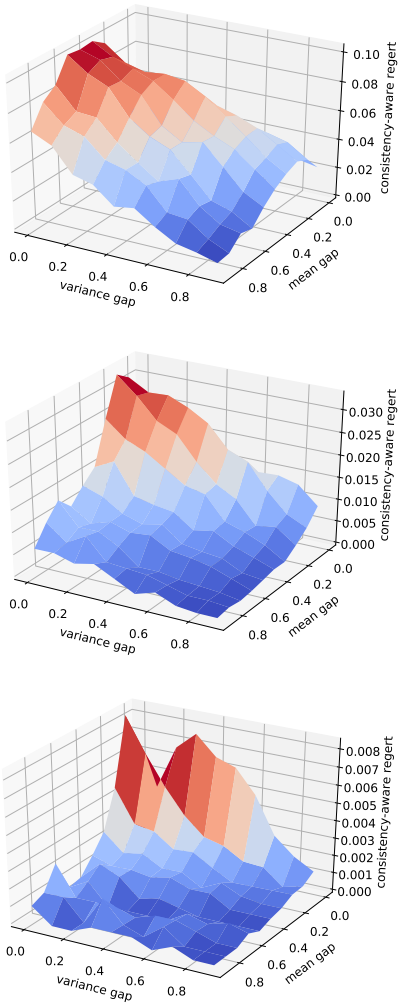

Figure 11: Case: $\rho = 0.1$, Top figure: $n = 20$. Middle figure: $n = 200$. Bottom figure: $n = 2000$.

Conclusion and future work

In this paper, we developed a new definition of regret (called consistency-aware regret) which is sensitive to the variability in rewards of the different arms by considering the variance of the rewards. We extended and modified the BESA algorithm to optimize consistency-aware regret and provided a bound on its performance. Finally, we illustrated the utility of our proposed algorithm on a real clinical dataset and studied its behaviour on some synthetic datasets.

We believe still there exist a noticeable gap between clinical trial problems, which inspired our work, and the nature of multi-armed bandit problems. Considering other ways to incorporate reward variability and providing some bounds on the confidence interval of the arm chosen by a bandit learning algorithm are promising for the future studies. It is also interesting to extend other bandit algorithms like Thompson sampling to consistency-aware regret and study their properties. Finally, utilizing BESA+ in the acquisition of real data would be an important future validation step.

Acknowledgment

We would like to thank Audrey Durand for her comments and insight on this project. We also thank department of family medicine of McGill University and CIHR for their generous support during this project.

References

[Agrawal and Goyal, 2013] Agrawal, S., and Goyal, N. 2013. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, 99–107.

[Auer, Cesa-Bianchi, and Fischer, 2002] Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

[Austin, 2011] Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

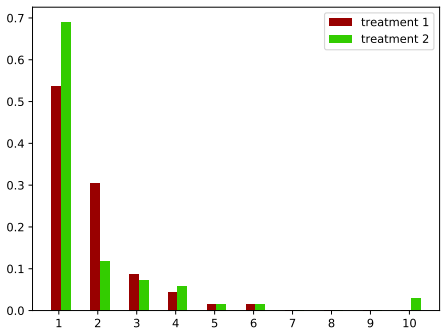

Figure 12: Distribution graph

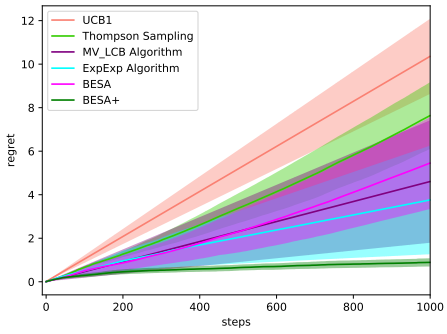

Figure 13: Accumulated consistency-aware regret

[Baransi, Maillard, and Mannor, 2014] Baransi, A.; Maillard, O.-A.; and Mannor, S. 2014. Sub-sampling for multi-armed bandits. In *ECML-KDD*, 115–131.

[Burnetas and Katehakis, 1996] Burnetas, A. N., and Katehakis, M. N. 1996. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17(2):122–142.

[Carandini, 2004] Carandini, M. 2004. Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS biology* 2(9):e264.

[Chapelle and Li, 2011] Chapelle, O., and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

[Corbin-Berrigan et al., 2018] Corbin-Berrigan, L.-A.; Kowalski, K.; Faubert, J.; Christie, B.; and Gagnon, I. 2018. Three-dimensional multiple object tracking in the pediatric population: the neurotracker and its promising role in the management of mild traumatic brain injury. *NeuroReport* 29(7):559–563.

[Galichet, Sebag, and Teytaud, 2013] Galichet, N.; Sebag, M.; and Teytaud, O. 2013. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, 245–260.

[Garcıa and Fernández, 2015] Garcıa, J., and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16(1):1437–1480.

[Kuleshov and Precup, 2014] Kuleshov, V., and Precup, D. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

[Lai and Robbins, 1985] Lai, T. L., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

[Maillard, 2013] Maillard, O.-A. 2013. Robust risk-averse stochastic multi-armed bandits. In *ICML*, 218–233.

[Ripley et al., 2013] Ripley, B.; Venables, B.; Bates, D. M.; Hornik, K.; Gebhardt, A.; Firth, D.; and Ripley, M. B. 2013. Package mass. *Cran R*.

[Robbins, 1985] Robbins, H. 1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer. 169–177.

[Sani, Lazaric, and Munos, 2012a] Sani, A.; Lazaric, A.; and Munos, R. 2012a. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 3275–3283.

[Sani, Lazaric, and Munos, 2012b] Sani, A.; Lazaric, A.; and Munos, R. 2012b. Risk-aversion in multi-armed bandits. In *NIPS*, 3275–3283.

[Sutton and Barto, 1998] Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.