Learning Modular Safe 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 dynamic changes in the treatment allocation probabilities of patients. 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 paper, we define and study a new robustness criterion for bandit problems. Specifically, we consider optimizing a function of the distribution of returns as a regret measure. This provides practitioners more flexibility to define an appropriate regret measure. The learning algorithm we propose to solve this type of problem is a modification of the BESA algorithm [Baransi *et al.*, 2014], which considers a more general version of regret. We present a regret bound for our approach and evaluate it empirically both on synthetic problems as well as on a dataset from 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 *et al.*, 2014; Auer *et al.*, 2002; Sani *et al.*, 2012a; Chapelle and Li, 2011; Sutton and Barto, 1998]. One of the main advantage of multi-armed bandit problems is its simplicity that allows for a higher level of 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. Consider a series of mulitple arm pulls (or steps) $t = 1, ..., T$ and selecting a specific arm $a \in \mathcal{A}$ at each step i.e. $a(t) = a_t$. The standard goal in the multi-armed bandit setting is to find the arm $\star$ which has the maximum expected reward $\mu_\star$ (or equivalently, minimum expected regret). The expected regret after T steps R_T is defined as the sum of the expected difference between the mean reward under $\{a_t\}$ and the reward expected under the optimal arm $\star$:

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

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]. The consistency of treatments 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. Thus, the idea of consistency and safety seems to some extent subjective and problem dependent. As a result, it might be necessary to develop an algorithm which can work with an arbitrary definition of consistency for a distribution.

This kind of system design which allows the separation of different parts of a system (here regret function and learning algorithm) has already been explored in modular programming. In modular programming, we emphasize on splitting the entire system into independent modules which at the end, the composite of these modules builds our system. This design trick is necessary when we are dealing with the change of customer demands and we require our system to adapt with the new demands. Here, we follow the same paradigm by making regret definition independent of the learning algorithm. As a result, we allow more flexibility in defining the regret function which is capable of incorporating problem specific demands.

Measure of regret

Definition 0.1. ***safety value function:*** *Let $\mathcal{D}$ denotes the set of all possible reward distributions for a given interval. The*

*hossein.aboutalebi@mail.mcgill.ca
†dprecup@cs.mcgill.ca
‡tibor.schuster@mcgill.ca

safety value function $v : \mathcal{D} \to \mathcal{R}$ provides a score for a given distribution.

The optimal arm $\star$ under this value function is defined as

$$\star \in \arg\max_{a\in\mathcal{A}}(v(\varphi_a)) \tag{1}$$

The regret corresponding to the safety value function up to time T is defined as:

$$\mathfrak{R}_{T,v} = \mathbb{E}\left[\sum_{t=1}^{T}(v(\varphi_\star) - v(\varphi_{a_t}))\right] \tag{2}$$

We call (2)*, safety-aware regret.*

When the context is clear, we usually drop the subscript v and use only $\mathfrak{R}_T$ for the ease of notation.

Definition 0.2. ***Well-behaved safety value function:*** *Given a reward distribution φ_a over the interval $[0, 1]$, a safety value function v for this distribution is called well-behaved if there exists an unbiased estimator $\widehat{v}$ of v such that for any set of observation $\{x_1, x_2, \dots, x_n\}$ sampled from φ_a, and for some constant γ we have:*

$$\sup_{x_1,\dots,x_n,\widehat{x}_i} |\widehat{v}(x_1,\dots,x_i,\dots,x_n) - \widehat{v}(x_1,\dots,\widehat{x}_i,\dots,x_n)| < \frac{\gamma}{n} \tag{3}$$

If (3) *holds for any reward distribution φ over the interval* $[0, 1]$*, we call the safety value function v, a well-behaved safety value function.*

Other types of well-behaved safety function can be defined as a function of standard deviation or conditional value at risk similar to the previous example. In the next section, we are going to develop an algorithm which can optimize the safety-aware regret.

Proposed Algorithm

In order to optimize the safety-aware regret, we build on the BESA algorithm, which we will now briefly review. As discussed in [Baransi *et al.*, 2014], BESA is a non-parametric (without hyperparameter) 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 of arm a. If $\widehat{\mu}_{s_a} > \widehat{\mu}_{s_\star}$, BESA chooses arm a, otherwise it chooses arm $\star$.

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

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

$$\mathfrak{R}_T \leqslant \zeta_{\Delta,\gamma}\log(T) + \theta_{\Delta,\gamma} \tag{4}$$

where in (4)*, $\zeta_{\Delta,\gamma}, \theta_{\Delta,\gamma}$ are constants which are dependent on the value of γ, Δ.*

Algorithm BESA+ two action case

```
Input: Safety aware value function $v$ and its estimate $\widehat{v}$
Parameters: current time step $t$, actions $a$ and $b$. Initially $N_{a,0} = 0, N_{b,0} = 0$
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 $a_t$
```

Empirical results

Conditional value at risk safety value function

As discussed in [Galichet *et al.*, 2013], in some situations, we need to limit the exploration of risky arms. Examples include financial investment where inverters may tend to choose risk-averse kind of strategy. Using conditional value at risk as a risk measure is one of the approaches to achieve this goal. Informally, conditional value at risk level α is defined as the expected values of the quantiles of reward distribution where the probability of the occurrence of values inside this quantile is less than or equal to α. More formally:

$$CVaR_\alpha = \mathbb{E}[X|X < v_\alpha] \tag{5}$$

where in (5), $v_\alpha = \arg\max_\beta\{\mathbb{P}(X < \beta) \leqslant \alpha\}$. To estimate (5), we have used the estimation measure introduced by [Chen, 2007]. This estimation is also employed in [Galichet *et al.*, 2013] work to derive their MARAB algorithm. Here, we have used this estimation for the Conditional value at risk safety value function which is the regret measure for this problem. Our environment consists of 20 arms where each arm reward distribution is the truncated Gaussian mixture consisting of four Gaussian distribution with equal probability. The reward of arms are restricted to the interval $[0, 1]$. To make the environment more complex, the mean and standard deviation of arms are sampled uniformly from the interval $[0, 1]$ and $[0.5, 1]$ respectively. The experiments are carried out for $\alpha = 10\%$. For MARAB algorithm, we have used grid search and set the value $C = 1$. The figures 4, 5 depict the results of the run for ten experiments. It is noticeable that in both figures BESA+ has a lower variance in experiments.

Mean-variance safety value function

Next, we evaluated the performance of BESA+ with the regret definition provided by [Sani *et al.*, 2012a]. Here, we used the same 20 arms Gaussian mixture environment described in the previous section. We evaluated the experiments with $\rho = 1$ which is the trade off factor between variance and the mean. The results of this experiment is depicted in figures 6, 7. The

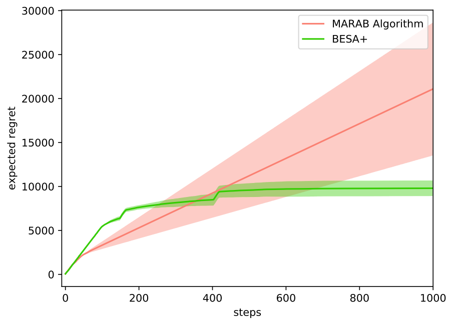

Figure 1: Accumulated regret figure. The safety value function here is conditional value at risk.

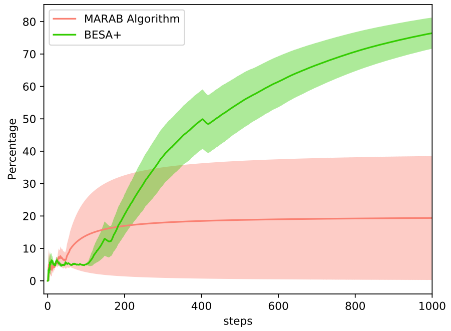

Figure 2: Percentage of optimal arm play figure. The safety value function here is conditional value at risk.

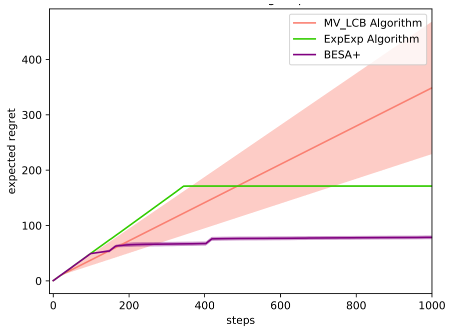

Figure 3: Accumulated regret figure. The safety value function here is mean-variance.

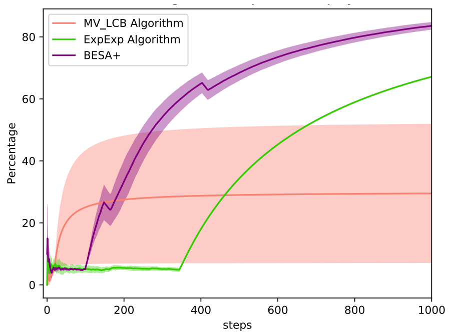

Figure 4: Percentage of optimal arm play figure. The safety value function here is mean-variance.

hyper-parameters used here for algorithms MV-LCB and ExpExp are based on what [Sani *et al.*, 2012a] suggests using. Again, we can see that BESA+ has a relatively small variance over 10 experiments.

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 8 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 8 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 safety value function described in Example 1. 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 9 illustrates the performance of different bandit algorithms. It is easy to notice that BESA+ has relatively better performance than all the other ones.

Conclusion and future work

In this paper, we developed a modular safety-aware regret definition which can be used to define the function of interest as a safety measure. We also modified the BESA algorithm and equipped it with new features to solve modular safety-aware regret bandit problems. We then computed the asymptotic regret of BESA+ and showed that it can perform like an admissible policy if the safety value function satisfies a mild assumption. Finally, we depicted the performance of BESA+ on the regret definition of previous works and showed that it can have better performance in most cases.

It is still interesting to investigate whether we can find better bounds for BESA+ algorithm with modular safety-aware regret definition. Another interesting path would be to research if we can define similar safety-aware regret definition for broader reinforcement learning problems including MDP environments.

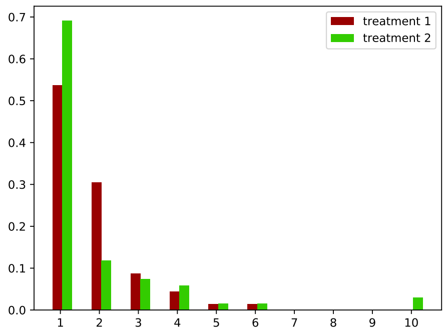

Figure 5: Distribution graph

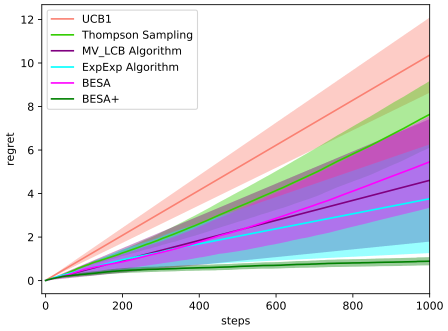

Figure 6: Accumulated consistency-aware regret

References

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

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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

[Baransi *et al.*, 2014] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *ECML-KDD*, pages 115–131, 2014.

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

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

[Chen, 2007] Song Xi Chen. Nonparametric estimation of expected shortfall. *Journal of financial econometrics*, 6(1):87–107, 2007.

[El-Yaniv and Pechyony, 2009] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35(1):193, 2009.

[Galichet *et al.*, 2013] Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

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

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

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

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

[Ripley *et al.*, 2013] Brian Ripley, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. Package 'mass'. *Cran R*, 2013.

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

[Sani *et al.*, 2012a] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[Sani *et al.*, 2012b] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *NIPS*, pages 3275–3283, 2012.

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

[Tolstikhin, 2017] IO Tolstikhin. Concentration inequalities for samples without replacement. *Theory of Probability & Its Applications*, 61(3):462–481, 2017.