UKRI Centre for Doctoral Training in
Artificial Intelligence for Healthcare

Imperial College London

Paper Summary and Discussion

Bayesian Design Principles for Frequentist Sequential Learning
Xu and Zeevi, 2023

Barbora Barancikova

February 6, 2024

1 Paper Summary

The paper by Xu and Zeevi[9], "Bayesian Design Principles for Frequentist Sequential Learning", presents a general approach to solving a wide class of sequential decision-making problems. It introduces a novel optimization setting where an agent holds "algorithmic beliefs" and aims to minimize a generalized notion of regret. The authors propose a new loss function that allows the agent to update beliefs using standard Bayesian principles without committing to a specific prior. The framework is applicable to various multi-armed bandit and reinforcement learning settings and achieves superior empirical performance in stochastic, adversarial, and non-stationary bandit environments.

This work was published in The International Conference for Machine Learning (ICML) 2023 and has received the "Outstanding Paper Award" for its contribution.

1.1 Background

Sequential learning problems with partial feedback are ubiquitous in machine learning. The agent performs a set of decisions under uncertain environmental conditions, while only being able to observe feedback for its chosen actions. The agent's goal is to maximize the cumulative reward over a finite time horizon. This setting naturally arises in many real-world applications, such as clinical trials [4], game playing [6], and recommender systems [5].

The central challenge introduced by the partial feedback setting is the exploration and exploitation tradeoff. That is, the agent needs to balance between exploring new actions to gain information and exploiting "good" actions already known to provide high rewards. Algorithms have been proposed to solve this problem in different variants of the famous multi-armed bandit (MAB) setting, where the agent seeks to maximize the cumulative reward of sequentially choosing from a set of actions (arms), each with an unknown reward distribution. The UCB [1] and Thompson Sampling [8] algorithms have often been applied to solve the stochastic MAB problem, EXP3 [2] solves the adversarial MAB problem, and different approaches have been proposed to handle the non-stationary MAB [3] problem with a variable reward distribution. More general sequential learning problems are formulated using reinforcement learning (RL) [7], where the agent's view of the environment is also subject to a state transition function, i.e. the setup of a chess board changes after each move.

1.2 Theoretical Contributions

The authors generalize the problems mentioned above by introducing the following notation. The agent has a decision space Π and an observation space $\mathcal{O}$, where both are compact sets. $\mathcal{M}$ is a model class where each model is a function $M : \Pi \to \mathcal{O}$ with a corresponding reward function f_M. At each round t, we have a probability distribution p_t over Π and a model M_t selected by the environment. The agent samples a decision $\pi_t \sim p_t$ and observes $o_t \sim M_t(\pi_t)$. We can define the *regret* at time t as the difference between the reward obtained at each previous round minus the reward that would have been obtained by choosing the optimal decision π^*:

$$\mathcal{R}_t = \sup_{\pi^* \in \Pi} \exp\left[\sum_{t=1}^{T} f_{M_t}(\pi^*) - \sum_{t=1}^{T} f_{M_t}(\pi_t)\right], \tag{1}$$

where T is the time horizon and the expectation is taken over the randomness of the environment. The exploration and exploitation dilemma is now implicitly included in the agent's effort to minimize regret.

With this formulation at hand, the authors define a probability measure ν of a joint random variable $(M, \pi^*) \in \mathcal{M} \times \Pi$, which will serve as a prior over the decision and model space. We

can then define a marginal prior distribution over the agent's actions as

$$\nu_{\pi^*}(\cdot) = \int_{\mathcal{M}} \nu(M, \cdot) dM. \tag{2}$$

After drawing a new decision π and observing $o \sim M(\pi)$, we obtain a marginal posterior distribution over $\pi*$ conditioned on π and o as $\nu_{\pi*|\pi,o}(\cdot)$. As a central notion of the paper, the authors define an Algorithmic Information Ratio (AIR) loss function as

$$\mathrm{AIR}_{q,\eta}(p, \nu) = \exp_{p,\nu}\left[f_M(\pi^*) - f_M(\pi) - \frac{1}{\eta} KL(\nu_{\pi*|\pi,o}, q)\right], \tag{3}$$

where η is a learning rate and q is a reference distribution that can be defined according to the specifics of the problem. The first and second terms of AIR correspond to the expected regret, and the third term represents the expected information gain of a decision. The authors propose several algorithms that use AIR optimization in a sequential learning environment, the most notable of which is the Adaptive Posterior Sampling (APS), shown in Figure 1. APS assumes that the reference distribution q_t at each step t is the marginal posterior distribution $(\nu_t)_{\pi*|\pi_t,o_t}$.

Algorithm 2 Adaptive Posterior Sampling (APS)

Input learning rate $\eta > 0$.
Initialize $p_1 = \mathrm{Unif}(\Pi)$.

1: **for** round $t = 1, 2, \cdots, T$ **do**
2: Find a distribution ν_t of (M, π^*) that solves

$$\sup_{\nu \in \Delta(\mathcal{M} \times \Pi)} \mathtt{AIR}_{p_t,\eta}(p_t, \nu).$$

3: Sample decision $\pi_t \sim p_t$ and observe $o_t \sim M_t(\pi_t)$.
4: Update $p_{t+1} = (\nu_t)_{\pi^*|\pi_t,o_t}$.
5: **end for**

Figure 1: Adaptive Posterior Sampling (APS) algorithm[9].

The authors also show a simplification of APS into a series of closed-form updates in the case of a Bernoulli stochastic MAB setting, shown in Figure 2, where APS behaves as a more robust version of Thompson Sampling.

The authors also prove several regret bounds and convergence properties of the proposed algorithms in the stochastic, adversarial, and non-stationary MAB environments, and a reinforcement learning setting.

1.3 Methodology and Experiments

The authors evaluate the performance of the simplified APS algorithm for Bernoulli Multi-Armed Bandits described in Figure 2 in four different synthetic experiment settings. To analyze the results, they plot the average regret defined in Equation 1 over 100 independent runs. The learning rate η and other hyperparameters necessary for the baseline algorithms are varied randomly within a range of values. In each setting, APS is compared to the standard widely-known algorithms that solve the corresponding problem, such as UCB, Thompson Sampling, and EXP3. A plot of the accumulating regret for each algorithm over time is referred to as

Algorithm 4 Simplified APS for Bernoulli MAB

Input learning rate $\eta > 0$.

Initialize $p_1 = \text{Unif}(\Pi)$.

1: **for** round $t = 1, 2, \cdots, T$ **do**

2: Sample action $\pi_t \sim p_t$ and receives r_t.

3: Update p_{t+1} by

$$p_{t+1}(\pi_t) = \begin{cases} \frac{1-\exp(-\eta)}{1-\exp(-\eta/p_t(\pi_t))}, & \text{if } r_t = 1 \\ \frac{1-\exp(\eta)}{1-\exp(\eta/p_t(\pi_t))}, & \text{if } r_t = 0 \end{cases}, \text{ and}$$

$$p_{t+1}(\pi) = p_t(\pi) \cdot \frac{1 - p_{t+1}(\pi_t)}{1 - p_t(\pi_t)}, \quad \forall \pi \neq \pi_t.$$

4: **end for**

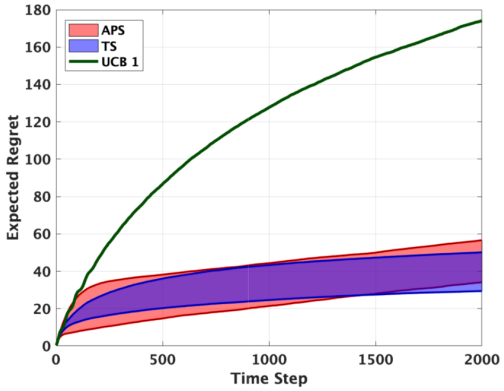
Figure 2: Simplified APS for Bernoulli MAB[9].

"sensitivity analysis," and an algorithm is declared to be beating a baseline if its regret curve is strictly lower than that of a baseline. To show robustness, the colored areas in the sensitivity analysis figures correspond to a range of hyperparameters used in each algorithm.

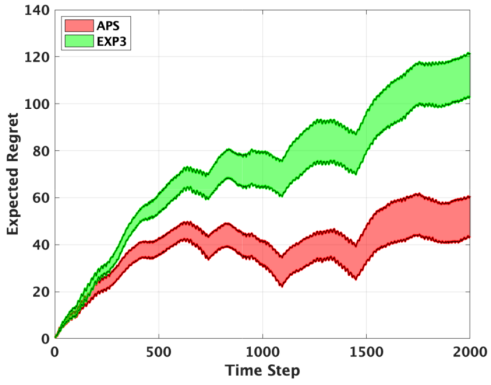
The **Stochastic Bernoulli MAB** problem was modeled by a 16-armed bandit where the rewards are sampled from Bernoulli distributions. APS shows comparable performance to Thompson Sampling with randomly initialized Beta$(c, 1)$ priors with $c \in [0.05, 5]$ and outperforms UCB1 (see Figure 3a). In the **Adversarial Bernoulli MAB** setting, 16 random sequences of rewards are generated with no additional statistical assumptions. APS is compared to the EXP3 algorithm for adversarial bandits and outperforms it regardless of the choice of hyperparameters (see Figure 3b).

For non-stationary environments where the reward distribution changes over time, a modified notion of regret, called the dynamic regret [3], is used. Since the optimal decision in hindsight π^* (previously mentioned in Equation 1) can no longer be characterized by always pulling the single best arm, the dynamic regret compares the agent's cumulative reward to the cumulative reward of the best possible non-stationary policy, i.e., π^* is allowed to change over time. In a **Change Points** environment, the algorithm runs for 4000 time steps, and the parameters of the Bernoulli reward distributions are reset every 1000 time steps. In this setting, APS dramatically outperforms EXP3 (see Figure 4a). The authors also compare APS to UCB1, Thompson Sampling, and EXP3, where the baseline algorithms are restarted when a change point occurs, effectively allowing them to "see the future." APS still outperforms all of them except restarted Thompson sampling, to which it has comparable performance (see Figure 5). Hence, in this setting, APS dramatically outperforms all the baseline algorithms. When the non-stationary environment is modeled using **"Sine Curve" Reward Sequences**, APS is tested in a 4-armed bandit setting with the parameters of the four Bernoulli reward distributions continuously changing according to a sine curve. As shown in Figure 4b, APS outperforms all the baseline algorithms. In particular, Thompson Sampling, which has been a strong competitor to APS in the other experimental settings, fails to learn meaningful information and accumulates the highest dynamic regret.

The results of the conducted experiments show that APS is a truly robust algorithm and is especially superior in use cases where one is uncertain about the modeling assumptions and the stationarity of the reward function.

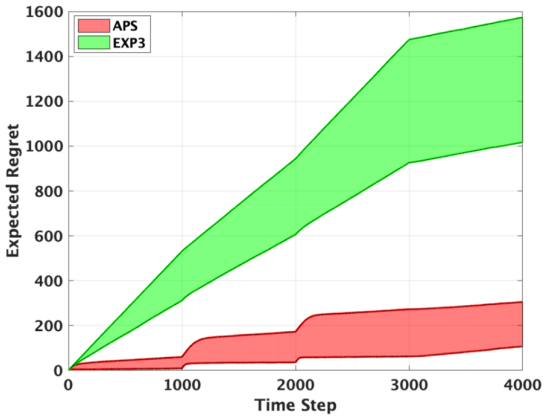


(a) Stochastic bandit problem[9].

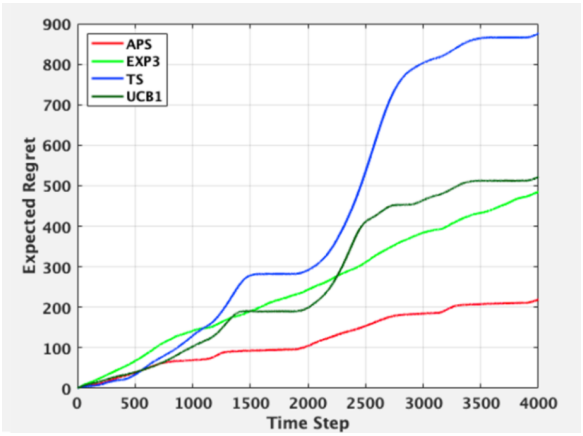


(b) Adversarial bandit problem[9].

Figure 3: Sensitivity analysis in stationary environments.



(a) Change points environment[9].



(b) Sine curve reward sequences[9].

Figure 4: Sensitivity analysis in non-stationary environments.

2 Discussion and Critique

The authors demonstrate, both theoretically and experimentally, that the derived methods achieve superior performance in non-stationary and adversarial bandit environments. The proposed solutions allow for a prior-free implementation but still come with the advantages of Bayesian belief updates. In many cases, the algorithms reduce to simple-to-implement update rules. While no experiments were performed on real-world data, many industry problems can be formulated as multi-armed bandit variants, and there certainly is potential for more applied research to be conducted based on this work.

The paper also illustrates how the derived principles can be used in linear bandit, convex bandit, and reinforcement learning problems, but no experiments are conducted for these use cases. It would be interesting to see how the proposed algorithms perform in these settings and whether there is a tractable way to implement them in practice. For example, in many standard reinforcement learning problems, the joint probability measure ν of the model and decision space will be a very complex object. It is also reasonable to assume that the performance of such methods will vary with respect to the choice of model parametrization. This potentially introduces several new hyperparameters that need to be tuned, which may reduce the robustness of the proposed algorithms. While the paper provides a useful way to reason about general sequential learning problems, further research has yet to show if the proposed methods also work in more complicated settings, such as when the environment contains state transitions.

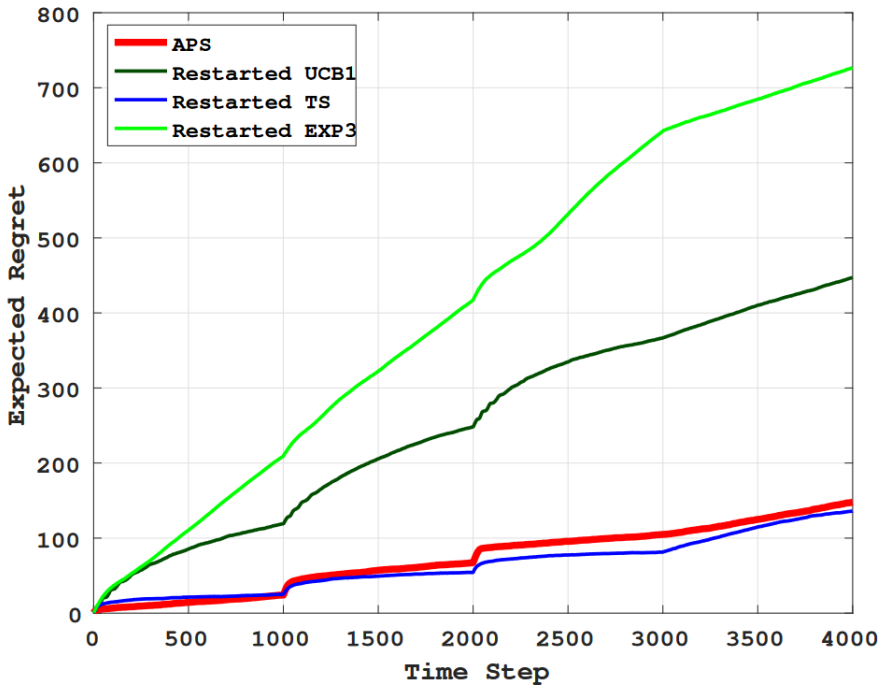


Figure 5: Sensivity analysis of the change points environment with "clairvoyant" restarted algorithms[9].

The problem setting formulation and the corresponding proofs are written with strong mathematical rigor, which makes the paper very convincing but also difficult to follow for a reader without a strong mathematical background. It would be helpful to include more intuitive explanations of the key concepts and theorems and motivate the abstract results with simpler examples.

3 Conclusion

The authors introduce a new perspective on sequential learning algorithm design and provide a generalized perspective on a class of problems that have previously only been addressed separately. Overall, this paper has a lot of potential to redefine how the industry thinks about sequential learning problems. The performance results for non-stationary and adversarial MAB environments are very convincing and will likely be applied in many real-world settings. The paper also provides a useful framework for reasoning about more complex sequential learning problems, such as reinforcement learning, but it is yet unclear if the proposed methods truly work in practice.

References

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

[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[3] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

[4] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

[5] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[6] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[7] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[8] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[9] Yunbei Xu and Assaf Zeevi. Bayesian design principles for frequentist sequential learning. In *International Conference on Machine Learning*, pages 38768–38800. PMLR, 2023.