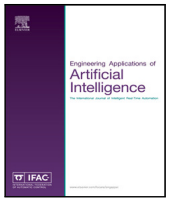
Engineering Applications of Artificial Intelligence 131 (2024) 107815

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Adapting bandit algorithms for settings with sequentially available arms

Marco Gabrielli [a,*], Manuela Antonelli [a], Francesco Trovò [b]

[a] Dipartimento di Ingegneria Civile e Ambientale (DICA), Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy
[b] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy

ARTICLE INFO

Keywords:
Online learning
Multi-armed Bandit
Regret minimization
Best-arm identification

ABSTRACT

Many real-world applications involve a sequential decision-making process where the options presented simultaneously. However, other applications, such as, Internet campaign management and environmental monitoring, the available options are presented sequentially to the decision-maker who, at each time, is asked to select the proposed option or not. This scenario is defined as the *Sequential Pull/No-Pull* setting The present study aims at developing a meta-algorithm, namely Sequential Pull/No-pull for MAB (Seq), to adapt any classical MAB (Multi-Armed Bandit) policy for this setting both in the case of regret minimization (RM) and best-arm identification (BAI) problems. This is achieved by exploting the sequential nature of the these settings allowing to select multiple arms and gather more information compared to classical policies. The proposed Seq meta-algorithm provides the same theoretical guarantees as the MAB policy employed, but was shown to provide improved performance compared to several classical MAB policies in RM and BAI problems employing real-world data. In particular, in the RM scenario regarding Internet advertising optimization, Seq-adapted algorithm resulted, on average, in $\approx 10\%$ lower regret during the whole time horizon than using classical MAB policies. When tested in a BAI problem involving the identification of the time of the day characterized by the highest concentration of pollutants in a water monitoring scenario, Seq identified the correct time in less than 4 days and 28 measurement.

1. Introduction

Many common applications involve sequential decision making among several options throughout a given time horizon, ranging from recommendation systems (Kawale et al., 2015) to online advertising (Nuara et al., 2018), from networking (Maghsudi and Stańczak, 2014) to dynamic pricing (Trovò et al., 2018). In such problems, an agent (i.e., the decision maker), either human or artificial, is faced with the problem of selecting among a set of options in a sequential manner over a finite time horizon. The strategy used for the selection determines the amount of reward the agent is gaining. For instance, in online advertising the options are the different advertisement one might display at a specific time on a webpage and the reward is the number of purchases that an advertising campaign is producing over time. The design of an effective strategy able to properly explore the different options and, at the same time, identify the most profitable one is critical for the success/failure of an advertising campaign.

Even if in most of these applications, the options are presented simultaneously to the decision-maker, this is not true for a large class of applications. In these cases, instead, the options are presented sequentially to the decision-maker within a time span, e.g., a day, and are repeated throughout the time horizon, e.g., months. The decision maker's task at a specific time is to either select the single proposed option or refuse to choose it for the current time. In this work, this scenario is defined as the *Sequential Pull/No-Pull (SPNP)* setting, and algorithms crafted explicitly to speed up the learning in such scenarios are provided.

A significant example for the SPNP setting occurs when an Internet campaign manager has to allocate the advertising budget over the day (Gasparini et al., 2018). In this setting, the advertiser divides the day into a finite number of time slots, representing the time steps of our sequential decision-making process, and sequentially chooses if it is worth allocating some advertising budget to that time slot or not. This process is repeated every day throughout the entire advertising campaign. The objective is to allocate the budget to the single time slots providing the largest revenue, e.g., clicks or conversions while minimizing the loss incurred due to the learning process due to suboptimal choices. The goal of minimizing such a loss is commonly called the Regret Minimization (RM) task. Instead, in other applicative settings, the goal of the sequential decision-making process is to identify the best option with the largest possible confidence, commonly addressed by the Multi-Armed Bandit (MAB) literature as the Best-Arm Identification (BAI) task. This is the case for environmental processes, which

* Corresponding author.
E-mail address: marco.gabrielli@polimi.it (M. Gabrielli).

https://doi.org/10.1016/j.engappai.2023.107815
Received 7 February 2022; Received in revised form 23 July 2023; Accepted 26 December 2023
Available online 10 January 2024

need to be monitored to identify the time of the day during which the most critical condition, e.g., in terms of pollution concentration, occurs to identify the presence of possible issues (i.e., the occurrence of contamination events) and react timely in order to reduce their impact on the environment and human health (Stravs et al., 2021; Besmer et al., 2017b). At each time, e.g., hours, during the day, the monitoring agent chooses if they want to perform a measurement or not. In this application, the objective is to determine, with the highest probability and as soon as possible, the time of the day during which the highest pollutants concentrations occur. Other than a correct and timely identification, another critical aspect for this application is the fact that each measurement corresponds to a non-negligible cost due to the reagents consumed and the manual labour required (Favere et al., 2021) which are one of the main hurdles which limit the adoption of more widespread use of monitoring campaigns (Besmer et al., 2017a). Hence, it is not possible to naively perform measurements at every instant.

In the classical sequential decision-making framework, an agent (i.e., the decision-maker) is presented at each time (i.e., at each round) with a finite set of available options (referred to as arms) over a finite time horizon, and they are asked to select one of them to maximize a specific objective. For this purpose, a wide range of algorithms have been designed in the MAB field (Bubeck et al., 2012), either resorting to frequentist (Auer et al., 2002; Audibert et al., 2010; Garivier and Cappé, 2011) or Bayesian (Kaufmann et al., 2012b; Agrawal and Goyal, 2013) approaches. Their adoption has been revealed to be effective in a wide range of practical problems (Kawale et al., 2015; Nuara et al., 2018; Maghsudi and Stańczak, 2014; Trovò et al., 2018). However, classical algorithms assume that the agent has to select one or more options at the beginning of each round. Even if a viable option consists of mapping the SPNP setting to this classical scenario, this modeling choice will limit the possibility of exploiting the information collected during each round to perform the successive decision as soon as new feedback is received. In addition, all the above mentioned MAB algorithms perform a fixed number of selections during each round. However, the possibility to select a variable number of arms per round can allow gathering information rapidly during rounds characterized by high uncertainty while converging to the optimal arm when enough information is collected. To the best of our knowledge, the design of specifically-crafted MAB algorithms able to exploit the temporal dependency offered by the SPNP setting is not known in the literature.

In this work, a meta-algorithm, namely Sequential Pull/No-Pull for MAB (Seq), was designed to improve the performance of classical bandit algorithms in the SPNP setting, exploiting the temporal ordering of the arms present in this scenario. More specifically:

- The SPNP problem is casted in the Multi-Armed Bandit framework, for both the RM and BAI tasks;
- A meta-algorithm, namely Seq, is designed to transform any classical MAB algorithm into one for the SPNP setting;
- It is shown that Seq has $\mathcal{O}(\log(T))$ regret, T being the time horizon of the learning process, when applied to classical MAB algorithms designed for RM, thus, maintaining the guarantees of such algorithms also in this setting;
- It is shown that applying the Seq to a generic classical BAI algorithm still provides the same guarantees;
- Extensive experimental analysis is provided on real-world data coming from advertisement management and water contaminant monitoring problems to compare the performance of state-of-the-art algorithms with the ones provided by the Seq meta-algorithm.

2. Related works

For a comprehensive review of the MAB setting we refer the interested reader to Bubeck et al. (2012) and Lattimore and Szepesvári (2020). In the following, is presented the related works present in the MAB literature from an application point of view and, after that, the ones which have more in common with the SPNP setting from a methodological point of view.[1]

Table 1
List of the acronyms and algorithms used in the paper.

MAB	Multi-Armed Bandit
RM	Regret Minimization
BAI	Best Arm Identification
SPNP	Sequential Pull/No-Pull
PAC	Probably Approximately Correct
Seq	Sequential Pull/No-Pull for MAB (algorithm)
UCB1	Upper Confidence Bound 1 (algorithm) by Auer et al. (2002)
Bayes-UCB	Bayes Upper Confidence Bound (algorithm) by Kaufmann et al. (2012a)
TS	Thompson Sampling (algorithm) by Kaufmann et al. (2012b)
UCBrev	UCB revisited (algorithm) by Auer and Ortner (2010)
UCBE	Upper Confidence Bound for Exploitation (algorithm) by Audibert et al. (2010)
SR	Successive Reject (algorithm) by Audibert et al. (2010)

2.1. Application-related works

Only a few works are present in the bandit literature to deal with RM and BAI for specific SPNP scenarios. In the Internet advertising management field, ads are selected in real-time to target potential customers (Estrada-Jiménez et al., 2019). In SPNP settings, a method to select the most profitable time slot during the day has been presented in Gasparini et al. (2018). Nonetheless, this method provides suggestions in an offline fashion, exploiting the information provided from historical data, not including any procedure to include a newly discovered piece of information. Moreover, Geng et al. (2020) uses an online method to partition the audience of an advertising campaign. Instead, Avadhanula et al. (2021) and Nuara et al. (2018) use a combinatorial bandit approach to optimize the budget spending over a multi-channel advertising campaign. However, in both these applications the options the learner can select do not show a temporal ordering. Also, in the environmental monitoring field, the most commonly applied strategies either do not seek to optimize the monitoring strategy online, e.g., setting an *a priori* sampling frequency, use external variables as proxy (Besmer et al., 2017a), or follow the explore-then-exploit principle (Gabrielli et al., 2021). In fact, while no methodologies involving online learning are known to the authors, other approaches exploiting previously available data have been proposed to improve environmental monitoring (e.g., Bottarelli et al. (2019), Pool and Seibert (2021), Cheng et al. (2003) and Russo et al. (2020)). However, these approaches suffer from different drawbacks as proxy variables are not necessarily available in all scenarios, and setting an *a priori* frequency or separating the exploration and exploitation phases is sub-optimal.

2.2. Methodologically-related works

From a methodological point of view, the possibility to select multiple arms during each round is usually tackled by Multiple Plays MAB (Bubeck et al., 2013; Komiyama et al., 2015) or Combinatorial MABs (Chen et al., 2013). While such approaches generalize the traditional framework by allowing the selection of multiple arms per round, they differ from the one presented here as the available arms (or superarms in the Combinatorial MAB setting) during each time step is fixed. Similar to the presented setting, in the Scaling MAB (Fouché et al., 2019; Wang and Masoud, 2021; Lesage-Landry and Taylor, 2018) a learner is allowed to pull a variable number of arms depending an efficiency parameter balancing the cost and the reward of each arm. However, differently from the presented work, all the arms are

[1] To facilitate the consultation of the article, we report the acronyms and the explanation of the algorithms' names in Table 1.

available at the beginning of each time step. Directly applying the sleeping bandit framework (Kleinberg et al., 2010), assuming to have an auxiliary action corresponding to "do nothing" and getting a null reward, cannot be applied in our setting. Indeed, such a framework would require knowing the reward at each time point, while in ours, the feedback received is only a proxy for the reward.

We remark that the present work aims to propose a meta-algorithm which, in contrast to previous applications, allows to optimize online and without resorting to external proxies advertising and environmental monitoring campaigns. Compared to previous studies, such meta-algorithm has been designed to face the settings in which the arms are presented sequentially to the learner and not at the same time, potentially allowing them to vary the number of arms pulled depending on the confidence of the arms rewards.

3. Problem formulation

In what follows, the SPNP setting are defined and two running examples are provided for the regret minimization and best arm identification tasks.[2] Assume a problem in which a learner is allowed to select among a finite set of $K \in \mathbb{N}$ arms $\{a_1, \dots, a_K\}$ over a finite time horizon of T time steps. At each time step $t \in \{1, \dots, T\}$, the learner is allowed to either select the arm a_i with $i = mod(t, K) + 1$ or decide not to pull it. The ith round ro_i is defined as a tuple of K consecutive time steps during which the learner is presented in a sequence all the available arms, formally $ro_i := (t_{(i-1)K+1}, \dots, t_{iK})$. During the time horizon T a total of $\tau = \lfloor \frac{T}{K} \rfloor$ rounds is available.[3] Each arm a_i at time step t is characterized with a value $x_{i,t}$ of the feedback provided to the learner. The feedback $x_{i,t}$ is modeled as a realization of a random variable $X_{i,t}$ drawn from a distribution D_i, whose expected value is $\mu_i := \mathbb{E}[D_i]$. As commonly done in the bandit literature, Bernoulli distributions are used to model the feedbacks, i.e., $D_i \sim Be(\mu_i)$. It is remarked that, in SPNP setting, the feedback might not correspond to the reward, depending on the specific arm that it is pulled (see Section 3.1). The expected value of the feedback of the optimal arm $a^* = \arg\max_i \mu_i$ is denoted with $\mu^* = \max_i \mu_i$ and with $X_{*,t}$ the random variable associated with the optimal arm. An algorithm $\mathfrak{U}$ is a sequential decision-making policy selecting, at each time step t, an arm a_t to pull, where the possible options at time t are $a_t = \emptyset$ or $a_t = a_{mod(t,K)+1}$. Depending on the setting, an algorithm $\mathfrak{U}$ might have different objectives to optimize: minimize the regret or identify the optimal arm.

3.1. Regret minimization

In the Regret Minimization (RM) framework, the learner's objective is to minimize the loss incurred over time due to the learning process. More specifically, if the arm to pull for the round is suboptimal, i.e., $a_i \neq a^*$, and the policy $\mathfrak{U}_{RM}$ opts to pull it, the learner gains a reward of $X_{i,t} - X_{*,t}$, i.e., equal to the difference between the values associated to currently considered arm a_i and the optimal one a^*. Instead, if the arm to pull at the current time step is the optimal one a^*, the learner gains $X_{*,t}$ reward if they opted to pull it. Finally, if the learner decides not to pull an arm, they get no reward. Formally, the instantaneous reward Z_t gained pulling arm a_t is defined as follows:

$$Z_t := \begin{cases} X_{i,t} - X_{*,t} & \text{if } a_t = a_i \neq a^* \\ X_{*,t} & \text{if } a_t = a^* \\ 0 & \text{if } a_t = \emptyset. \end{cases} \tag{1}$$

The loss incurred by an algorithm $\mathfrak{U}_{RM}$, commonly called *pseudo-regret*, is defined as:

$$R_T(\mathfrak{U}_{RM}) := \frac{T}{K}\mathbb{E}[Z_t^*] - \sum_{t=1}^{T} \mathbb{E}[Z_t], \tag{2}$$

[2] We recap the symbols used in this section and the following in Table 2.

[3] For the sake of simplicity, from now on, it is assumed that the time horizon T is a multiple of K, i.e., $T = \tau K$.

Table 2
List of the Symbols used in the paper.

$\{a_1, \dots, a_K\}$	Arm set
K	Number of arms
T	Time Horizon
ro_i	ith round
τ	Number of rounds
$X_{i,t}$	Random variable providing feedback for arm a_i at time t
$X_{*,t}$	Random variable providing feedback for the optimal arm at time t
$x_{i,t}$	Realization of the feedback for arm a_i at time t
D_i	Distribution of the feedback $X_{i,t}$ for all $t \in \{1, \dots, T\}$
Z_t	Random variable providing the reward to the learner at time t
μ_i	Expected value of the random variables $X_{i,t}$ for all $t \in \{1, \dots, T\}$
μ^*	Expected value of the optimal arm
Δ_i	Gap between the expected value of the optimal arm μ^* and of the ith arm μ_i
$\Delta_{(i)}$	Gap between the expected value of the optimal arm and of other arms, sorted in ascending value
$\mathfrak{U}_{RM}$	Arm selection algorithm for regret minimization
$R_T(\mathfrak{U}_{RM})$	Regret over a time horizon of T for algorithm $\mathfrak{U}_{RM}$
$\mathfrak{U}_{BAI}$	Arm selection algorithm for best-arm identification
$\mathcal{G}$	Stopping rule for best-arm identification
$\mathfrak{G}$	Final selection rule for best-arm identification
δ_i	Confidence level for a best-arm identification algorithm at time t

where the expected value is taken w.r.t. the stochasticity of the reward Z_t and of the algorithm used $\mathfrak{U}_{RM}$. The first element of the r.h.s. of Eq. (2) is the expected reward of the optimal strategy that pulls at each round only the optimal arm.[4]

Notice that, in this setting, the standard definition of reward ($Z_t = X_t$) is not meaningful for this problem. Indeed, the pseudo-regret corresponding to the optimal strategy mentioned above would be linear in T while the one corresponding to the naïve strategy always opting to pull the available arm at each time step t would be null.

In what follows, as customary in the bandit literature, the goal is to design algorithms $\mathfrak{U}_{RM}$ for which the regret $R_T(\mathfrak{U}_{RM})$ grows sublinearly over time, meaning that the cost per round of the learning process $\frac{R_T(\mathfrak{U}_{RM})}{T} \to 0$ as $T \to +\infty$.

Example 1 (*Internet Advertising Campaign Optimization*). The first running example models the decision process of an Internet campaign manager, who faces the problem to allocate the advertising budget over the day (Gasparini et al., 2018). In this setting, time slots partitioning each day are modeled as different arms a_i, and the manager has to sequentially decide if they want to allocate a given budget on them. Indeed, the advertising campaign splits each day into different slots as the users' interest in given products is not constant throughout the day but can peak during certain hours. The reward $X_{i,t}$ consists of the number of conversions/leads provided by the advertisement over the chosen time slot. The final objective is to allocate the budget to a single time slot a^* providing the largest expected revenue (conversion/lead) while minimizing $R_T(\mathfrak{U}_{RM})$, i.e., the loss incurred due to the choice of suboptimal arms during the learning process.

3.2. Best-arm identification

In the BAI framework, the learner's objective is to identify the arm providing the largest expected reward, minimizing the probability of selecting a different arm. In this setting, it is required to provide an algorithm $\mathfrak{U}_{BAI}$, a.k.a. sampling strategy, a stopping rule $\mathfrak{S}$, providing the learner a time t at which the algorithm has finished the process, and a procedure $\mathfrak{G}$ selecting the final guess, i.e., providing a guess of

[4] With a slight abuse of notation, when the context is clear, we will refer to a_i as a generic arm from the arm set and a_i as the arm selected at time t.

the optimal arm $\hat{a}_t^*$ at time t. In this setting, Probably Approximately Correct (PAC) guarantees for a given tuple $(\mathfrak{U}_{BAI}, \mathfrak{S}, \mathfrak{G})$ are provided so that:

$$\mathbb{P}(\hat{a}_t^* \neq a^*) \leq \delta_t, \tag{3}$$

where $\delta_t \in (0, 1)$ is a given confidence level. Depending on if either the stopping time of the algorithm t or its confidence δ_t are fixed in advance, these settings are called *fixed-budget* or *fixed-confidence*, respectively. See Audibert et al. (2010) for more details.

Example 2 (*Environmental Monitoring*)**.** The second running example provided is about monitoring an environmental process. More specifically, the problem consists of identifying the time of the day during which the most critical condition, e.g., in terms of pollution concentration, occurs. This procedure is crucial to identify the presence of possible issues (i.e., the occurrence of contamination events) and react promptly to reduce their impact on the environment and human health (Stravs et al., 2021; Besmer et al., 2017b). At each time, e.g., hours, during the day, the monitoring agent chooses if they want to perform a measurement or not, corresponding to the currently available arm a_t. In this application, the objective is to determine the time instant a^*, the time of the day during which the highest pollutants concentrations occur. Depending on the specific application, one may want to have guarantees on the probability $1 - \delta_t$ that the selected arm is the one providing the largest value, e.g., if statistical guarantees on the chosen arm are desired. Conversely, if the measurements have non-negligible costs, e.g., due to the reagents consumed and the manual labour required (Favere et al., 2021), the best guess over a fixed amount of samples t is requested, corresponding to the cost allowed for the monitoring action.

Remark 1. A naïve approach to solve the SPNP problem is to model it as a stochastic MAB setting in which the available K arms consists in selecting a single arm at each round ro_i. A pseudo-code describing such an approach is provided by Algorithm S1 present in Appendix. Therefore, the standard MAB setting is played over a time horizon of $\tau = \frac{T}{K}$ time steps since it is allowed to pull one arm per round. However, this approach is suboptimal since it is missing the chance of selecting multiple arms per round. This, instead, allows to perform multiple exploratory pulls per round and gather more information, a critical aspect in the first stages of the learning process during which the estimated values for the arms are uncertain.

Remark 2. A specific class of MAB algorithms commonly referred to as *elimination algorithms* are such that they present favorable characteristics to be applied to the SPNP setting for both the RM and BAI settings. Indeed, these algorithms iteratively exclude one or more arms that are likely not to be optimal during the learning process, and since the arms selection process occurs in a round-robin fashion, they can be applied to the setting so that at each round ro_i they can pull at most K arms. Therefore, in what follows, it is shown that slightly modifying their definition provides significant theoretical improvement over other classical MAB approaches in the SPNP setting.

In the following sections, the theoretical guarantees and the empirical performance of the approaches mentioned above and the proposed Seq meta-algorithm, crafted explicitly for the SPNP setting, will be analyzed.

4. The sequential Pull/No-pull MAB algorithm

In what follows, a meta-algorithm applicable to any classical MAB algorithm, either for the RM or BAI tasks, which is better suited to the SPNP setting is proposed. The overall idea is that the learner should pull an arm a_i any time they are allowed to do that, and a classical MAB algorithm would pull it, instead of waiting for the next round ro_i to pull it as a classical MAB algorithm would do. The pseudo-code of the proposed approach, namely Sequential Pull/No-pull MAB (Seq) is presented in Algorithm 1. It requires as input a classical MAB policy $\mathfrak{U}_{MAB}$, either for RM or BAI, and an ordered set of arms to choose from $\{a_1, \dots, a_K\}$. At first, it initializes the policy $\mathfrak{U}_{MAB}$ and a counter for the current number of pulls n.[5] At each time step t, the algorithm computes the arm a_{MAB} the MAB algorithm $\mathfrak{U}_{MAB}$ would pull as if a total number of pulls equal to n to select it would be available. If this arm is the one, the learner is allowed to pull it within the current round, i.e., $a_{MAB} = a_{mod(t,K)+1}$ they pull the arm, and collect the feedback $x_{MAB,t}$ from the selected arm a_{MAB}. Otherwise, the learner opts not to pull anything and proceed to the next round $ro_i + 1$ without any update. Differently from the naïve application of $\mathfrak{U}_{MAB}$ described in the previous section, the Seq approach allows to perform multiple pulls per round if this is advised by the strategy $\mathfrak{U}_{MAB}$.

Algorithm 1 Seq($\mathfrak{U}_{MAB}$)

```
Input: MAB algorithm 𝔘_MAB, arm set {a_1, ..., a_K}, time horizon T
Initialize 𝔘_MAB
n ← 0
a_MAB ← 1
for t ∈ {1, ..., T} do
    if a_MAB = a_mod(t,K)+1 then
        Pull arm a_MAB
        Collect feedback x_MAB,t
        n ← n + 1
        Update 𝔘_MAB
        a_MAB ← 𝔘_MAB(n)
    end if
end for
```

Notice that the developed Seq($\mathfrak{U}_{MAB}$) meta-algorithm can be applied to either RM or BAI problems. In what follows, its properties in both scenarios are described.

5. Theoretical analysis for the regret minimization algorithms in SPNP setting

In this section, the pseudo-regret upper bounds for the classical algorithms applied directly to the SPNP setting, as presented before in Remark 1, and for the Seq meta-algorithm are derived.

5.1. Regret analysis of classical MAB algorithms

In the case a classical algorithm for RM, e.g., UCB1 (Auer et al., 2002), Bayes-UCB (Kaufmann et al., 2012a), or Thompson Sampling (Thompson, 1933), is applied to the SPNP setting, as specified in Algorithm S1 (provided in Appendix) the pseudo-regret is:

Theorem 1. *Using a classical RM algorithm $\mathfrak{U}_{RM}$, with guarantees on the expected number of pulls of the suboptimal arms of $\mathbb{E}[T_i(t)] \leq C_i \log(t) + A_i$, where C_i is $o(1)$ and A_i is $o(\log(t))$, over a time horizon of t, on the SPNP setting it suffers a pseudo-regret of:*

$$R_T(\mathfrak{U}_{RM}) \leq \sum_{a_i \neq a^*} (\mu^* + \Delta_i)[C_i \log(T) + A_i - C_i K], \tag{4}$$

where $\Delta_i := \mu^* - \mu_i$ is the gap between the expected reward of the optimal arm a^* and a suboptimal arm a_i.[6]

[5] If the policy requires a number of steps for the initialization, we should remove them from the main loop and perform such a procedure at this step. In this case, the counter n of a number of rounds corresponding to the ones used for the initialization should also be increased.

[6] Where not specified the expected value $\mathbb{E}[\cdot]$ is w.r.t the stochasticity of the reward and the algorithm used. Moreover. we use $\sum_{a_i \neq a^*}$ as a concise version for $\sum_{a \in \{a_1, \dots, a_K\} \mid a \neq a^*}$.

The full proof of Theorem 1, as well as those of the following theorems, is deferred to Appendix for the sake of presentation.[7] For the UCB1 algorithm, for which the bound on the expected number of pulls is bounded by the constants $C_i = \frac{8}{\Delta_i^2}$ and $A_i = (1 + \frac{\pi^2}{3})$ (see Auer et al. (2002) for details), we have a bound on the pseudo-regret of:

$$R_T(UCB1) \leq \sum_{a_i \neq a^*} \frac{8(\mu^* + \Delta_i)}{\Delta_i^2} \log(\tau) + \sum_{a_i \neq a^*} \left(1 + \frac{\pi^2}{3} - \frac{8K}{\Delta_i^2}\right)(\mu^* + \Delta_i).$$

Conversely, for the Bayes-UCB algorithm we have that $C_i := \frac{1+\epsilon}{KL(\mu_i, \mu^*)}$ and $A_i := \frac{c \log(\log(t))}{KL(\mu_i, \mu^*)} + K_c(\log(\log(t)))^2 + o(1)$, for any $\epsilon > 0$, $c > 5$ and $K_c > 0$, providing a bound of:

$$R_T(Bayes - UCB) \leq \sum_{a_i \neq a^*} \frac{(1+\epsilon)(\mu^* + \Delta_i)}{KL(\mu_i, \mu^*)} \log(\tau) + o((\log(\log(t)))^2),$$

where $KL(a, b)$ is the Kullback–Leibler divergence of two Bernoulli variable with expected values a and b. Thanks to the Pinsker's inequality stating that $\frac{1}{KL(\mu_i, \mu^*)} \leq \frac{1}{2\Delta_i^2}$ the bounds can also be written as:

$$R_T(Bayes - UCB) \leq \sum_{a_i \neq a^*} \frac{(1+\epsilon)(\mu^* + \Delta_i)}{2\Delta_i^2} \log(\tau) + o((\log(\log(t)))^2).$$

Similarly, for the Thompson Sampling (TS) algorithm we have:

$$R_T(TS) \leq \sum_{a_i \neq a^*} \frac{(1+\epsilon)(\mu^* + \Delta_i)}{KL(\mu_i, \mu^*)} \log(\tau) + o((\log(\log(t)))),$$

since $C_i := \frac{1+\epsilon}{KL(\mu_i, \mu^*)}$ and $A_i := o(\log(\log(t)))$.

The use of the so-called *elimination algorithms* in the SPNP setting, due to their round-robin arm selection approach, allow their application in a more efficient way, and this reflects in a better regret bound. For instance, the UCBrev algorithm (Auer and Ortner, 2010) operates as follows: pulls all the arms in a round-robin fashion until all the arms have a given number of pulls; after that, it uses Hoeffding's bounds to exclude those arms which are likely to be suboptimal, and iterates until the total number of pulls reached the time horizon. The modification of the UCBrev algorithm that selects multiple arms per round, from now on denoted with UCBrev+, is detailed by Algorithm S2 in Appendix. Even if UCBrev+ exploits the temporal dependency better than the other RM approaches in the SPNP setting, a specifically crafted analysis on its regret fails in providing a better regret bound than the one provided in Theorem 1. See Appendix A for details.

5.2. Regret analysis of the Seq meta-algorithm

The use of a generic RM algorithm in the Seq meta-algorithm provides an upper bound on the pseudo-regret of the same order of using a generic RM algorithm in the SPNP setting:

Theorem 2. *Given a classical RM algorithm $\mathfrak{U}_{RM}$, with guarantees on the expected number of pulls of the suboptimal arms of $\mathbb{E}[T_i(t)] \leq C \log(t) + A$ over a time horizon of t, the Seq($\mathfrak{U}_{RM}$) algorithm on the SPNP setting over a time horizon of T rounds suffers from a pseudo-regret of:*

$$R_T(Seq(\mathfrak{U}_{RM})) \leq \sum_{a_i \neq a^*} (\Delta_i + \mu^*)[C_i \log(T) + A_i]. \tag{5}$$

Remark 3. Even if in principle the design of the Seq algorithm allows for pulling an arm at each time step t, it suffers from a regret of the same order of the one in Theorem 1. In the experimental section, the empirical improvement of the Seq approach will be analyzed.

[7] With $f(t) = o(g(t))$ we denote two functions for which as $t \to \infty$ if for every positive constant ε there exists a constant N such that $|f(t)| \leq \varepsilon g(x)$ for all $t \geq N$.

6. Theoretical analysis for best-arm identification in SPNP setting

6.1. PAC analysis of classical BAI algorithms

The focus in the BAI problem is to select with high probability, at the end of an exploration procedure, the optimal arm. In the SPNP scenario, one might straightforwardly apply a generic BAI algorithm by selecting the arm to pull once for each round ro_i, therefore selecting a single arm to pull every K time steps. This approach is exemplified again by Algorithm S1 provided in Appendix. This approach has the following guarantees:

Theorem 3. *A classical BAI algorithm $\mathfrak{U}_{BAI}$, with guarantees of $\delta_t(\mathfrak{U}_{BAI}) \leq C_1 t K$ on the classical MAB setting, on the SPNP setting, provides a confidence of:*

$$\delta_t(\mathfrak{U}_{BAI}) \leq C_1 t K^2. \tag{6}$$

Notice that depending on if we are in the fixed confidence or the fixed budget BAI setting, we set δ_t or t, respectively, and compute the corresponding t or δ_t, respectively. The additional linear dependence on K w.r.t. the standard BAI setting is due to the fact that the learner is allowed to pull a single arm at each round ro_i, performing a total of τ pulls, while the potentially available pulls are t in total. For instance, this result states that if the UCBE algorithm (Audibert et al., 2010) is chosen, it provides a guarantee of:

$$\delta_t(UCBE) \leq 2tK^2 \exp\left(\frac{2\sum_{i=1}^K 1/\Delta_i^2}{25}\right),$$

and choosing the SR algorithm (Audibert et al., 2010) with a budget of τ pulls results in:

$$\delta_T(SR) \leq \frac{K(K-1)}{2} \exp\left(-\frac{T - K^2}{K\overline{\log}(K) H_2}\right), \tag{7}$$

where $\overline{\log}(K) := \frac{1}{2} + \sum_{i=2}^K \frac{1}{i}$, $H_2 := \max_{i \in [K]} \frac{i}{\Delta_{(i)}^2}$, and the sequence $\{\Delta_{(i)}\}_{i=1}^K$ is the ordering of the gaps Δ_i in increasing order, i.e., formally $\min_i \Delta_i = \Delta_{(1)} \leq \Delta_{(2)} \leq \cdots \leq \Delta_{(K)} = \max_i \Delta_i$.

Even in the BAI setting, a slightly different use of an elimination algorithm provides some improvement w.r.t. the approach mentioned above. For instance, consider the SR algorithm (Audibert et al., 2010) that works as follows: it divides the total pulls into phases, during which all the available arms are pulled the same number of times, it eliminates a single arm at the end of each phase and repeats the process until a single arm remains. This procedure allows selecting multiple arms per round ro_i. The definition of the algorithm derived from SR and selecting multiple arms per round, denoted from now on with SR+, is provided by Algorithm S3 in Appendix. Using the SR+ algorithm in the SPNP setting, it is shown show that:

Theorem 4. *The SR+ algorithm with a budget of $n = \frac{(2T-1)\overline{\log}(K)}{2K} + K$ on the SPNP setting, provides a confidence of:*

$$\delta_T(SR+) \leq \frac{K(K-1)}{2} \exp\left(-\frac{2T-1}{2H_2}\right). \tag{8}$$

Notice that this result has a better scaling factor of $\approx \frac{K}{\overline{\log}(K)} \geq 2$ w.r.t. the one obtained by the SR algorithm. This improvement is due to the fact that this modified version can pull multiple arms per round, packing as much as possible the exploratory phases over the time horizon T.

6.2. PAC analysis of the Seq meta-algorithm

If the Seq meta-algorithm is applied to any BAI algorithm, it is trivial to show that the guarantees are the same as the ones provided by a generic BAI algorithm in Theorem 3, formally:

Corollary 1. *Consider a classical BAI algorithm $\mathfrak{U}_{BAI}$, with guarantees of $\delta_t(\mathfrak{U}_{BAI}) \leq C_1 t K$ on the classical MAB setting. The Seq($\mathfrak{U}_{BAI}$) algorithm, on the SPNP setting, provides a confidence of:*

$$\delta_t(Seq(\mathfrak{U}_{BAI})) \leq C_1 t K^2. \tag{9}$$

Even in the BAI setting, tighter result is not possible since no strong guarantees that this approach selects more than one arm at each round ro_i are available. Nonetheless, it will be shown in the next section how this approach provides better empirical performance w.r.t. the straightforward application of such techniques to the SPNP problem.

7. Experimental results

Numerical simulations have been conducted to assess the experimental performance of Seq with the ones of classical MAB in for RM and BAI settings, and the newly-introduced UCBrev+ and SR+. Two real-world datasets coming from advertising and environmental monitoring applications have been used for the experiments.[8] In the provided figures, solid lines, bars and dots show the estimated mean values, while shaded areas and confidence bars provide the estimated 95% confidence interval of the mean. The Python implementation of Seq and the newly-introduced UCBrev+ and SR+ used in the experiments can be found at https://github.com/mgabriell1/SeqMAB.

7.1. Regret minimization

An instance of Example 1 was simulated using the Yahoo! Front Page Today Module User Click Log Dataset (Li et al., 2011). Such dataset contains a user click log for the articles displayed in the Featured Tab of the front page of Today Module on Yahoo! of a few days in May 2009. The goal is to display (i.e., advertise) the article during the proper slot of time over the day. Similarly to what has been done in Liu et al. (2018) and Re et al. (2021), the clicks during the day have been divided into 10 slots $\{A, \dots, J\}$, evenly splitting the number of accesses over the day into 10 parts. For each slot i, the average click-though rates $\rho_i \in [0, 1]$ using the data from the corresponding slot i. Finally, an arm for each slot has been modeled corresponding to the hours during the day the ads have been displayed, i.e., at least one data was present in the dataset. For instance, in the setting represented in Fig. 1, the MAB has been set over the slots from D to J due to the fact that, in the A, B, and C slots, no samples were available. As in this scenario, each round represents one day in which the articles are to be displayed, and the entire experiment has been conducted in a time horizon T of 2 years ($\tau = 720$ days). The obtained click-through rates distributions were used to generate 100 independent simulations for each article present in the dataset.

The adapted version of the UCBrev (Auer and Ortner, 2010) algorithm (UCBrev+), Bayes-UCB (bUCB) (Kaufmann et al., 2012a), and TS (Thompson, 1933) applied in the SPNP setting have been compared with the application of our meta-algorithm to bUCB and TS, i.e., Seq(bUCB) and Seq(TS), respectively. The performance of each RM algorithm $\mathfrak{U}$ were evaluated in terms of:

- $\hat{R}_T$ the empirical pseudo-regret, formally $\hat{R}_T(\mathfrak{U}) = \frac{T\mu^*}{K} - \sum_{t=1}^{T} \mu_{i(t)}$, where $i(t)$ is the index of the arm chosen by $\mathfrak{U}$;
- NPR the number of pulls per round, formally $NPR(\mathfrak{U}, ro_i) = \sum_{t \in ro_i} \mathbb{1}\{a_{i(t)} = a_{mod(t,K)+1}\}$;
- Opt^* the percentage of pulls of the optimal arm a^* over the number of pulls, formally $Opt^*(\mathfrak{U}) := \frac{\sum_{t \in \{1,\dots,T\} \mid a^* = a_{mod(t,K)+1}} \mathbb{1}\{a_{i(t)} = a_{mod(t,K)+1}\}}{\sum_{tt \in \{1,\dots,T\}} \mathbb{1}\{a_{i(t)} = a_{mod(t,K)+1}\}}$;

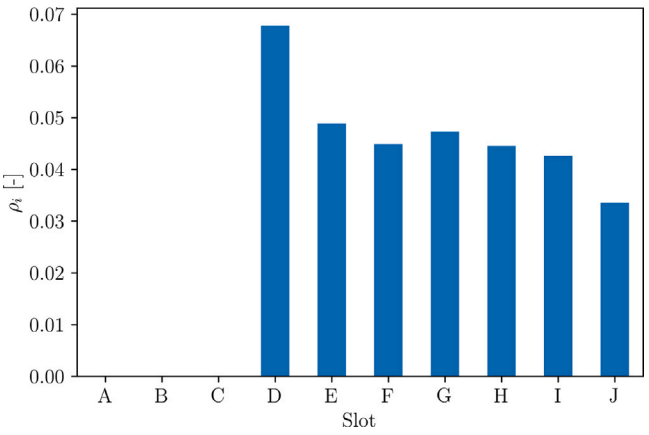


Fig. 1. Mean click-through rates calculated for a selected article.

- $Opti^*$ the percentage of the rounds for which the algorithm $\mathfrak{U}$ pulled the optimal arm, formally $Opti^*(\mathfrak{U}) := \frac{\sum_{t \in \{1,\dots,T\} \mid a^* = a_{mod(t,K)+1}} \mathbb{1}\{a_{i(t)} = a_{mod(t,K)+1}\}}{\tau}$;
- $\widehat{RR}_t$ the empirical pseudo-regret ratio obtained by the Seq-adapted algorithm and its traditional counterpart, formally $\widehat{RR}_t(\mathfrak{U}) := \frac{\hat{R}_t(Seq(\mathfrak{U}))}{\hat{R}_t(\mathfrak{U})}$.

The values reported in the following figures are the averaged values over 100 independent runs.

Results. Fig. 2 shows the empirical results obtained for the article whose arms are presented in Fig. 1. In Fig. 2(a), the empirical pseudo-regret $\hat{R}_T$ of the Seq meta-algorithm are comparable to the ones of their traditional counterparts, while UCBrev+ has a significantly larger regret. Conversely, Fig. 2(b) shows that the proposed methodology is capable of increasing the number of pulls per round.[9] However, it seems that the Seq meta-algorithms have not yet converged to the optimal arm. Indeed, Fig. 2(b) shows that even though the value of the NPR of Seq(bUCB) and Seq(TS) decreases during the simulations, it does not converge yet to 1 as one would have expected, meaning that the algorithms are still pulling multiple arms per round as the uncertainty regarding the arms' rewards has not been reduced sufficiently. However, looking at Fig. 2(c), both the Seq meta-algorithms select the optimal arm a larger number of times and spend a slightly higher fraction of the total pulls on the optimal arm a^* than their counterparts, indicating a better selection of the optimal arm (i.e., the optimal article) during the simulations. The large regret obtained by UCBrev+ can be explained by the fact that even though it pulled the optimal arm with a large probability ($Opti^*(UCBrev+) \approx 1$) over the rounds, the fraction of pulls spent on a^* is lower than the other algorithms. Indeed, Fig. 2(b) shows that UCBrev+ pulls all the available arms at every round as the confidence intervals of all the arms are still overlapping.

Fig. 3 shows the ratio of the regret of the application of the Seq($\mathfrak{U}$) on algorithm $\mathfrak{U}$ and the regret of the algorithm $\mathfrak{U}$ itself ($\widehat{RR}_t(\mathfrak{U})$), averaged over the different advertisement setting. A value of this ratio smaller than 1 provides evidence that the adoption of the Seq meta-algorithm provides a smaller regret than the traditional algorithm. Looking at the mean (solid lines) and 95% confidence intervals (semi-transparent areas) of $\widehat{RR}_t(bUCB)$ and $\widehat{RR}_t(TS)$, both ratios are below 1 over the entire time horizon T. Even if the application of Seq on $bUCB$ seems to lead to a better improvement with respect to the original algorithm during an initial period than on TS, both adaptations achieve a $\widehat{RR}_t$ of ≈ 0.9 by the end of the time horizon. Such result indicates

[8] A set of results on synthetically generated data are provided in Appendix. They are in line with the real-world ones.

[9] The number of pulls for bUCB and TS are not shown in Fig. 2(b) since these algorithms are allowed to pull a single arm per round deterministically.

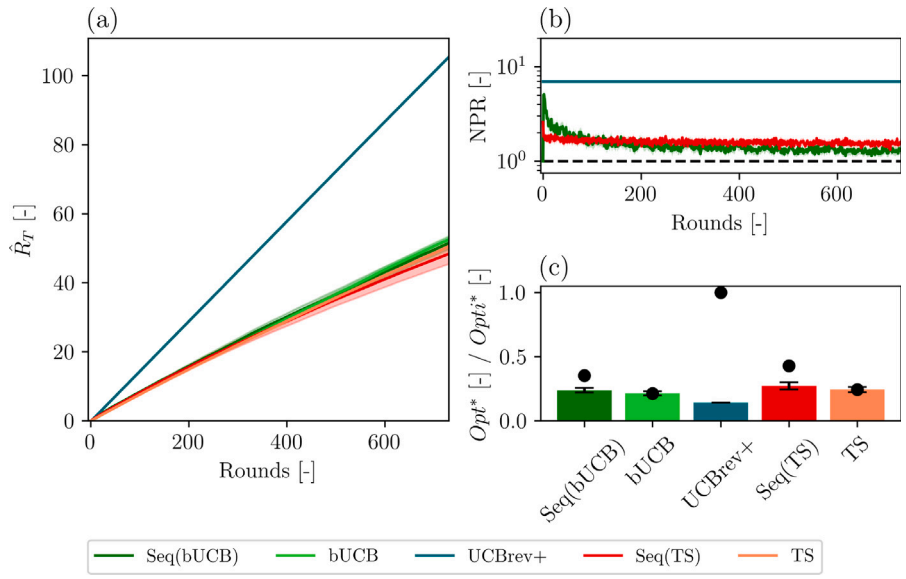


Fig. 2. Results for the advertising setting: (a) $\hat{R}_T$, (b) NPR, (c) Opt^* (shown as bars) and $Opti^*$ (shown as points).

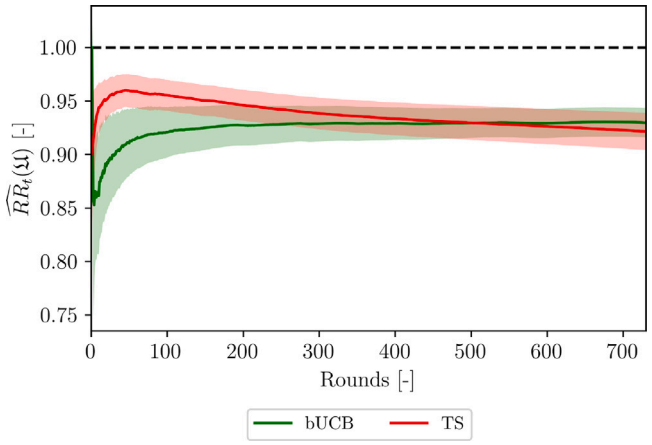


Fig. 3. $\widehat{RR}_t$ in the tested advertising setting.

how both Seq-adapted algorithms, on average, lead to a 10% reduction of the $\hat{R}_T$ with respect to their traditional counterparts, and with high probability they are providing better results than their traditional counterparts. Moreover, while $\widehat{RR}_t(bUCB)$ seems to have converged, this is not true for $\widehat{RR}_t(TS)$ which still presents a downward trend, indicating how an even lower ratio could be obtained over longer time horizons. Indeed, such results highlight that, regardless of the complexity of the advertising problem and the length of the campaign, the application of the Seq meta-algorithm provides a benefit compared to the naïve adaptation of classical RM algorithms.

7.2. Best-arm identification

An instance of Example 2 was simulated using the bacterial concentration measured during a high-frequency monitoring campaign of a drinking water distribution system, described in detail in Gabrielli et al. (2021), where bacterial concentrations have been monitored every 2 h, i.e., we have $K = 12$. The goal is to detect the sampling time, among the available ones, for which the bacterial concentrations overcomes a warning threshold $\Gamma = 60\ \frac{\text{cells}}{\mu\text{L}}$ with the largest probability. The samples collected over time are Bernoulli realizations of the measurement, stating if at a specific time the threshold Γ has been exceeded or not. The data corresponding to this phenomenon over a period of 26 days was analyzed, during which the data can be considered stationary over time. The selected stationary period was iterated to provide a sufficiently long time horizon, i.e., $T = 2796$, to allow all the algorithms converged steadily to $\hat{\delta} = 0$, which occurred for $\tau > 233$ days. The UCBE (Audibert et al., 2010), SR, and SR+ algorithms have been compared with the proposed Seq(UCBE).[10] Seq(UCBE) and UCBE were tested considering different values of the parameter $c \in \{1, 2, 4, 8\}$. Since real-world data was directly used to evaluate the performance of the algorithms in the BAI settings, it was not possible to test multiple realizations of the problem. For this reason, the minimum number of rounds $\hat{\tau}$ (i.e., days) and pulls $\hat{n}$ (i.e., measurements) required before steadily converging to the identification of the correct sampling time was used to assess the performance of the algorithms, formally:

- $\hat{\tau} := \min_{\tau \in \mathbb{N}} \{\hat{a}_t^* = a^*, \forall t \in [\tau K, \dots, +\infty)\}$;
- $\hat{n} := \sum_{t=1}^{\hat{\tau}K} \mathbb{1}\{a_t = a_{mod(t,K)+1}\}$.

Results. The results for the above-described experiment are presented in Table 3. The naïve adaptation to our setting of UCBE and SR provide worse performance than the Seq-adapted algorithm and SR+. The improvement provided by the latter approaches is due to that they are allowed to pull multiple arms per round, therefore increasing the number of pulls performed during the first rounds characterized by high uncertainty.[11] Moreover, the results of Seq(UCBE) are less influenced by the parameter c than UCBE, suggesting its behavior is robust to misspecifications of such a parameter. Furthermore, Seq(UCBE) resulted in better performance than SR+ for most of the values of the parameter c, similarly to what has been reported by Audibert et al. (2010) in the classical MAB settings. Indeed, the definition of phases to guide the selection of arms performed by SR+ limits the possibility of focusing on the most promising arms and does not allow to discard the ones with much worse performance quickly. A final remark is that the results in Table 3 show that the adoption of the proposed meta-algorithm is of paramount importance for this application. Indeed, commonly the

[10] Details regarding SR+ are provided in Appendix.

[11] Synthetic experiments, shown in Appendix, allowed us to highlight how Seq(UCBE) obtained a lower $\hat{\delta}_t$ than UCBE when the stopping criterion was based on the number of rounds elapsed Conversely, when the stopping criterion was set based on the number of pulls performed, both algorithms provided comparable $\hat{\delta}_t$, but Seq(UCBE) obtained such result in a lower number of rounds.

Table 3
Minimum number of rounds $\hat{\tau}$ and pulls $\hat{n}$ required by the algorithms to converge to $\hat{\delta} = 1$. Results corresponding to SR and SR+ have been repeated for all values of c as they are independent of such parameter.

Algorithm	$c = 1$		$c = 2$		$c = 4$		$c = 8$	
	$\hat{\tau}$	$\hat{n}$	$\hat{\tau}$	$\hat{n}$	$\hat{\tau}$	$\hat{n}$	$\hat{\tau}$	$\hat{n}$
Seq(UCBE)	8	42	**4**	**28**	**4**	**28**	**4**	**28**
UCBE	191	191	95	95	126	126	233	233
SR	223	223	223	223	223	223	223	223
SR+	**5**	**30**	5	30	5	30	5	30

time horizon in which the stationary assumption is met in this kind of applications are of the order of months (Gabrielli et al., 2021), which is a comparable time to the one required by Seq(UCBE) and SR+. Conversely, seasonality and unexpected changes occur over longer periods, jeopardizing the selection of the optimal arm provided by UCBE and SR.

8. Discussion

The Seq meta-algorithm developed allowed to successfully improve the empirical performances on RM and BAI problems of classical MAB algorithms. Indeed, even though such algorithm does not provide theoretical performance improvements, its experimental testing on both synthetic and real-world data derived from Internet advertising and environmental monitoring scenarios revealed its advantages over a naïve adaptation of classical algorithms.

In the Internet advertising RM scenario the possibility to select multiple arms throughout each day provided by the use of Seq (Fig. 2(b)) allowed to achieve lower regret compared to the traditional algorithms (Fig. 3). The fact that improved performances were achieved regardless of the complexity of the advertising problem and throughout the entire time horizon tested highlights how the developed meta-algorithm is suitable for most advertising management applications. Still, the combination of Seq to different classical algorithms might further improve performances in specific cases. For example, considering the two classical MAB algorithms tested, advertising campaigns of short duration might prefer the combination of Seq with bUCB, given the lower $\widehat{RR}_t$ values during the first round of the time horizon tested, while longer campaigns might be favored by the use of TS, as suggested by the decrease of $\widehat{RR}_t$ values at longer time horizons.

Significant improvements over the naïve use of classical algorithms were also obtained in the environmental monitoring BAI scenario. Indeed, the use of properly adapted algorithms (i.e., SR+ and Seq(UCBE)) allowed to identify the correct sampling hour in shorter times. We remark that the time required for this identification is a critical aspect of this application. Indeed, the required assumption of stationarity of the monitored quantities holds only over short period of times, and, therefore, delays in the identification of the most suitable sampling hour might reflect in the delayed identifications of contamination events, threatening public health. Modeling such a setting using SPNP allows the design of more efficient exploration strategies than the ones available in the classical MAB settings, taking advantage of the temporal structure of these settings.

It is possible to envision the direct application of the developed algorithm to such scenarios improving over current solutions used in society and industry. For example, with respect to the real-world applications tested, the developed Seq meta-algorithm could be used by advertising agencies to optimize the revenues from advertising campaigns. Instead, as online water monitoring instruments are already available, Seq could be used to guide sampling times in order to find the best sampling time throughout the day, rather than sampling at fixed times throughout the day. Given these results, it is likely that Seq will provide improved performances also in other scenarios characterized by sequential decision making processes described by SPNP settings, such as the detection of the presence of deteriorated industrial equipment, news/content recommendation or the selection of the most promising channel in networking problems.

8.1. Future developments

Even if the proposed method is applicable to any MAB stochastic setting, there are still some scenarios in which the proposed method does not apply. Indeed, the Seq meta-algorithm can only be applied for classical MAB setting, while the variety of existing MAB setting is yet to be explored. For instance, two interesting line of development are the one including contextual information on the MAB and on the non-stationarity of the reward provided by the arms. The former extension applies to the Internet advertising settings in which the user information are provided and the users' behavior is influencing the outcome of the advertising campaign. In this case, the information (context) can be exploited to effectively learn in a joint way the behavior of all the users. Instead, as mentioned before, the latter applies the environmental monitoring setting in which non-stationarity are present over long time period. In this case, one should include in the learning process mechanisms to deal with the change of the optimal option over time.

Given the previous limitations of the current framework, the most promising extensions are the following:

- Apply the sequential approach to the contextual bandit setting (Lu et al., 2010; Abbasi-Yadkori et al., 2011), allowing to leverage also the presence of external information in SPNP settings.
- Introduce techniques to handle the non-stationarity of the process generating the rewards, either in a passive (Garivier and Moulines, 2011; Trovo et al., 2020) or active (Garivier and Moulines, 2011; Re et al., 2021) way. Moreover, using the latter approach we would also provide a method to retrieve the time the process was affected by a change.

9. Concluding remarks

This paper shows the advantages of applying a novel meta-algorithm, namely Seq, to sequential decision-making problems involving options that are sequentially proposed to the decision-maker. Such problems were casted in a novel MAB setting, namely Sequential Pull/No-pull Bandit, in which one may exploit the ordering of the arms presented to the learner. The dedicated meta-algorithm adapts classical MAB algorithms aimed at problems regarding RM or BAI, such as Internet advertising and environmental monitoring, respectively, to the SPNP settings. Regarding Internet advertising, the task of selecting the optimal temporal slots during each day which provides a higher CTR was simulated starting from real-world data. The developed meta-algorithm selected the optimal option with a higher probability compared to traditional algorithms, leading to a lower regret. Instead, in an environmental monitoring application, the capability of the developed meta-algorithm to identify the time of the day in which the maximum concentration occurs was tested, based on the data derived from a previous monitoring campaign. By exploiting the possibility to intensify sampling during the initial rounds characterized by high uncertainty, the traditional algorithm adapted through the use of Seq allowed a faster correct identification than its traditional counterpart and state-of-the-art algorithms.

CRediT authorship contribution statement

Marco Gabrielli: Conceptualization, Software, Investigation, Writing – original draft, Visualization. **Manuela Antonelli:** Writing – review & editing, Supervision, Funding acquisition. **Francesco Trovò:** Methodology, Formal analysis, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.engappai.2023.107815.

References

Abbasi-Yadkori, Y., Pál, D., Szepesvári, C., 2011. Improved algorithms for linear stochastic bandits. In: Proceedings of the Neural Information Processing Systems conference (NeurIPS), Vol. 24.

Agrawal, S., Goyal, N., 2013. Further optimal regret bounds for thompson sampling. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 31. pp. 99–107.

Audibert, J.Y., Bubeck, S., Munos, R., 2010. Best arm identification in multi-armed bandits. In: Proceedings of the Conference on Learning Theory (COLT). pp. 41–53.

Auer, P., Cesa-Bianchi, N., Fischer, P., 2002. Finite-time analysis of the multiarmed bandit problem. Mach. Learn. 47 (2–3), 235–256.

Auer, P., Ortner, R., 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. Period. Math. Hungar. 61 (1–2), 55–65.

Avadhanula, V., Colini Baldeschi, R., Leonardi, S., Sankararaman, K.A., Schrijvers, O., 2021. Stochastic bandits for multi-platform budget optimization in online advertising. In: Proceedings of the Web Conference (WWW). pp. 2805–2817.

Besmer, M.D., Hammes, F., Sigrist, J.A., Ort, C., 2017a. Evaluating monitoring strategies to detect precipitation-induced microbial contamination events in karstic springs used for drinking water. Front. Microbiol. 8, 2229.

Besmer, M.D., Sigrist, J.A., Props, R., Buysschaert, B., Mao, G., Boon, N., Hammes, F., 2017b. Laboratory-scale simulation and real-time tracking of a microbial contamination event and subsequent shock-chlorination in drinking water. Front. Microbiol. 8, 1900.

Bottarelli, L., Bicego, M., Blum, J., Farinelli, A., 2019. Orienteering-based informative path planning for environmental monitoring. Eng. Appl. Artif. Intell. 77, 46–58.

Bubeck, S., Cesa-Bianchi, N., et al., 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found. Trends Mach. Learn. 5 (1), 1–122.

Bubeck, S., Wang, T., Viswanathan, N., 2013. Multiple identifications in multi-armed bandits. In: Proceedings of the International Conference on International Conference on Machine Learning (ICML). pp. 258–265.

Chen, W., Wang, Y., Yuan, Y., 2013. Combinatorial multi-armed bandit: General framework and applications. In: Proceedings of the International Conference on Machine Learning (ICML), Vol. 28. pp. 151–159.

Cheng, H., Yang, Z., Chan, C.W., 2003. An expert system for decision support of municipal water pollution control. Eng. Appl. Artif. Intell. 16 (2), 159–166.

Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A., Forné, J., 2019. On the regulation of personal data distribution in online advertising platforms. Eng. Appl. Artif. Intell. 82, 13–29.

Favere, J., Waegenaar, F., Boon, N., De Gusseme, B., 2021. Online microbial monitoring of drinking water: How do different techniques respond to contaminations in practice? Water Res. 202, 117387.

Fouché, E., Komiyama, J., Böhm, K., 2019. Scaling multi-armed bandit algorithms. In: Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD). pp. 1449–1459.

Gabrielli, M., Turolla, A., Antonelli, M., 2021. Bacterial dynamics in drinking water distribution systems and flow cytometry monitoring scheme optimization. J. Environ. Manag. 286, 112151.

Garivier, A., Cappé, O., 2011. The KL-UCB algorithm for bounded stochastic bandits and beyond. In: Proceedings of the Conference on Learning Theory (COLT), Vol. 19. pp. 359–376.

Garivier, A., Moulines, E., 2011. On upper-confidence bound policies for switching bandit problems. In: Proceedings of the International Conference on Algorithmic Learning Theory (ALT). Springer, pp. 174–188.

Gasparini, M., Nuara, A., Trovò, F., Gatti, N., Restelli, M., 2018. Targeting optimization for internet advertising by learning from logged bandit feedback. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN). pp. 1–8.

Geng, T., Lin, X., Nair, H.S., 2020. Online evaluation of audiences for targeted advertising via bandit experiments. In: Proceedings of the Conference on Artificial Intelligence (AAAI), Vol. 34. pp. 13273–13279.

Kaufmann, E., Cappe, O., Garivier, A., 2012a. On Bayesian upper confidence bounds for bandit problems. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 22. pp. 592–600.

Kaufmann, E., Korda, N., Munos, R., 2012b. Thompson sampling: An asymptotically optimal finite-time analysis. In: Proceedings of the International Conference on Algorithmic Learning Theory (ALT). pp. 199–213.

Kawale, J., Bui, H.H., Kveton, B., Tran-Thanh, L., Chawla, S., 2015. Efficient thompson sampling for online matrix-factorization recommendation. In: Proceedings of the Neural Information Processing Systems Conference (NeurIPS), Vol. 28. pp. 1297–1305.

Kleinberg, R., Niculescu-Mizil, A., Sharma, Y., 2010. Regret bounds for sleeping experts and bandits. Mach. Learn. 80 (2), 245–272.

Komiyama, J., Honda, J., Nakagawa, H., 2015. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In: Proceedings of the International Conference on International Conference on Machine Learning (ICML). pp. 1152–1161.

Lattimore, T., Szepesvári, C., 2020. Bandit Algorithms. Cambridge University Press.

Lesage-Landry, A., Taylor, J.A., 2018. The multi-armed bandit with stochastic plays. IEEE Trans. Automat. Control 63 (7), 2280–2286.

Li, L., Chu, W., Langford, J., Wang, X., 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In: Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM). pp. 297–306.

Liu, F., Lee, J., Shroff, N., 2018. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In: Proceedings of the Conference on Artificial (AAAI). pp. 3651–3658.

Lu, T., Pál, D., Pál, M., 2010. Contextual multi-armed bandits. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 485–492.

Maghsudi, S., Stańczak, S., 2014. Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. Trans. Wirel. Commun. 14 (3), 1309–1322.

Nuara, A., Trovò, F., Gatti, N., Restelli, M., 2018. A combinatorial-bandit algorithm for the online joint bid / budget optimization of pay-per-click advertising campaigns. In: Proceedings of the Conference on Artificial Intelligence (AAAI). pp. 1–8.

Pool, S., Seibert, J., 2021. Gauging ungauged catchments – Active learning for the timing of point discharge observations in combination with continuous water level measurements. J. Hydrol. 598, 126448.

Re, G., Chiusano, F., Trovò, F., Carrera, D., Boracchi, G., Restelli, M., 2021. Exploiting history data for nonstationary multi-armed bandit. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). pp. 51–66.

Russo, S., Lürig, M., Hao, W., Matthews, B., Villez, K., 2020. Active learning for anomaly detection in environmental data. Environ. Model. Softw. 134, 104869.

Stravs, M.A., Stamm, C., Ort, C., Singer, H., 2021. Transportable automated HRMS platform "MS2field" enables insights into water-quality dynamics in real time. Environ. Sci. Technol. Lett. 8 (5), 373–380.

Thompson, W.R., 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25 (3/4), 285–294.

Trovò, F., Paladino, S., Restelli, M., Gatti, N., 2018. Improving multi-armed bandit algorithms in online pricing settings. Internat. J. Approx. Reason. 98, 196–235.

Trovo, F., Paladino, S., Restelli, M., Gatti, N., 2020. Sliding-window thompson sampling for non-stationary settings. J. Artificial Intelligence Res. 68, 311–364.

Wang, Y., Masoud, N., 2021. Adversarial online learning with variable plays in the pursuit-evasion game: Theoretical foundations and application in connected and automated vehicle cybersecurity. IEEE Access 9, 142475–142488.