Best Model Identification: A Rested Bandit Formulation

Leonardo Cella [1] **Massimiliano Pontil** [1 2] **Claudio Gentile** [3]

Abstract

We introduce and analyze a best arm identification problem in the rested bandit setting, wherein arms are themselves learning algorithms whose expected losses decrease with the number of times the arm has been played. The shape of the expected loss functions is similar across arms, and is assumed to be available up to unknown parameters that have to be learned on the fly. We define a novel notion of regret for this problem, where we compare to the policy that always plays the arm having the smallest expected loss at the end of the game. We analyze an arm elimination algorithm whose regret vanishes as the time horizon increases. The actual rate of convergence depends in a detailed way on the postulated functional form of the expected losses. We complement our analysis with lower bounds, indicating strengths and limitations of the proposed solution.

1. Introduction

Multi-armed bandits are a mathematical framework of sequential decision problems that have played a fundamental role in machine learning and statistics (see e.g. Bubeck et al., 2012; Cesa-Bianchi, 2016; Lattimore & Szepesvári, 2020; Siegmund, 2003, and references therein). This framework consists of a sequence of T interactions (or *rounds*) between a learning agent and an unknown environment. During each round the learner picks an action from a set of options $\mathcal{K}$, usually referred to as *arms*, and the environment consequently generates a feedback (e.g., in the form of a loss value) associated with the chosen action/pulled arm. Multi-armed bandits have applications to a wide variety of domains including clinical trials, online advertising, and marketing.

In the standard i.i.d. stochastic bandit setting (Auer et al., 2002), the feedback generated when pulling an arm is modeled as a random variable sampled from a prescribed distribution associated with the selected arm. In contrast, in this paper we are interested in a *non-stationary* stochastic bandit setting called *rested bandits* (Allesiardo et al., 2017; Besbes et al., 2014; Cella & Cesa-Bianchi, 2020; Kleinberg & Immorlica, 2018; Levine et al., 2017; Seznec et al., 2018), whereby the feedback/losses received upon pulling arms are not i.i.d. anymore. That is, the distribution of losses changes as a function of the number of times each arm has been pulled so far. As a relevant example, assume the expected loss of action $i \in \mathcal{K}$ at a given round is decreasing with time and takes the parametric form

$$\frac{\alpha_i}{\sqrt{\tau}} + \beta_i \,, \tag{1}$$

where τ is the number of times arm i has been pull up to that round, and α_i and β_i are unknown parameters.

Considering decreasing expected losses is reasonable whenever the properties of the chosen arm improve as we allocate resources to them. For example, this is the case in scenarios where the goal is to find the best talent in a pool of candidates, say, the most valuable worker to train in an online labor platform having limited training time.

Overall, we interpret this problem as an *algorithm selection* or *selective training* problem. In this scenario, an arm represents a learning device that satisfies Equation (1) (like a specific neural network architecture) and the goal is to keep training the learner that will be the best at the end of the game. The parameters α_i and β_i in (1) may therefore quantify relevant properties of such models. For instance, in a standard statistical learning setting, parameter α_i can quantify the complexity (which may or may not be known) of the i-th model class, β_i might encode the representational power of that class in the form of the statistical risk of the best-in-class hypothesis (which is typically unknown), while the dependence on $1/\sqrt{\tau}$ is meant to suggest a plausible behavior of the generalization error of the i-th algorithm as a function of the training set size τ. Within this setting, an arm $i \in \mathcal{K}$ with small α_i and large β_i may represent an empirical risk minimizer (ERM) operating on a simple model class where the ERM has an estimation error getting small with few samples, but which only underfits the data without effectively minimizing the approximation error. Conversely, an arm $i \in \mathcal{K}$ with large α_i and small β_i may correspond

[1]Italian Institute of Technology, Genoa, Italy [2]University College London, United Kingdom [3]Google Research, New York, USA. Correspondence to: Leonardo Cella <leonardo-cella@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

to an ERM operating on a complex model class with large estimation error (where overfitting is likely to occur) and small approximation error.

Given a budget of T training samples, our specific goal is to design a strategy for online *selective training*, whereby at each round we have to decide which algorithm the next training example has to be fed to. This problem is of fundamental importance since, in many practical situations, performing a batch model selection (or model training) might be too computationally demanding. Thus, the goal is to design a strategy (a learning policy) that interacts with different learning algorithms with the goal of spending the budget of T samples on the algorithm/model that is likely to perform best after training. Pulling an arm corresponds to feeding the current sample to the associated algorithm, while observing the feedback corresponds to being able to estimate in an approximate manner (e.g., on a separate test set) the generalization error of the trained algorithm for that arm, this error being a decreasing function of the number of samples the chosen algorithm has so far been trained over. Because the algorithms to select from may originate from different modeling assumptions, we also view this as a *best model identification* problem.

Contributions. We first propose a novel notion of regret which is suited to the online learning problem we consider here. This regret criterion frames our problem as a best arm identification problem within a rested bandit scenario. We then characterize the structure of the problem by proving a non-asymptotic lower bound restricted to the 2-arm case. Finally, we propose two action elimination algorithms, and show for one of the two algorithms a regret upper bound that essentially matches the above-mentioned lower bound.

Notation. For a integer $N > 0$, we abbreviate the set $\{1, \dots, N\}$ by $[N]$. We use $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$, to denote expected value and probability measure, respectively. Moreover, for a given σ-algebra $\mathcal{F}$, $\mathbb{E}_{\mathcal{F}}[\cdot]$ and $\mathbb{P}_{\mathcal{F}}[\cdot]$ denote their conditional counterparts.

2. Related works

Our problem can be seen as a (rested variant of) the *best arm identification problem*, in that our metric reminds the simple-regret that was previously designed for the best-arm identification problem in the standard (stationary) stochastic multi-armed bandit setting (e.g. Even-Dar et al., 2006; Audibert & Bubeck, 2010; Gabillon et al., 2012; Kaufmann et al., 2016). The best arm identification problem has been investigated from two slightly different viewpoints. In the so called *fixed-confidence* variant, the goal is to minimize the sample complexity (that is, the number of pulls) needed to guarantee that, with some fixed confidence level, the selected arm is the one with smallest expected loss. In the *fixed-budget* variant, the goal is to find the best-arm within a fixed number of rounds (budget), while minimizing the probability of error.
In our case, we want the learning algorithm to single out with high probability (fixed confidence) the best base learner but, due to the non-stationary nature of the expected loss of base learners, we also want to do so with as few pulls as possible. Hence, we are in a sense combining the two criteria of fixed confidence and fixed budget.

A problem which is similar in spirit to ours is that of online model selection in bandit settings (that is, the case where the base learners are themselves bandit algorithms). This has been investigated in a number of papers in recent years, e.g., (Agarwal et al., 2017; Foster et al., 2019; 2020; Pacchiano et al., 2020; Cutkosky et al., 2021). In particular Agarwal et al. (2017) consider a very general class of base learners which have to satisfy reasonable stability assumptions. In order to deal with bandit information, importance weighted feedback is given to the bandit learners. In (Foster et al., 2019; 2020) the emphasis is specifically on linear bandit model selection problems, where model selection operates on the input dimension (Foster et al., 2019) or the amount of misspecification (Foster et al., 2020). Similar to (Agarwal et al., 2017), in (Pacchiano et al., 2020) the authors investigate the problem of algorithm selection in contextual bandits where contexts are stochastic. In order to bypass the stability assumption in (Agarwal et al., 2017), an additional smoothed transformation is introduced. The positive side effect induced by this additional step is the ability to feed the base learners with the original feedback with no re-weighting.

Unlike all the above works, we assume the expected loss of the base learners (which are not limited to bandit policies) depends in specific ways on the number of times each base learner is selected, this dependence being known up unknown parameters that have to be estimated. More importantly, we investigate a performance metric that is different from the standard cumulative regret incurred with respect to the best allocation policy, as studied in (Agarwal et al., 2017; Foster et al., 2020; 2019; Pacchiano et al., 2020). Further remarks on this comparison is given in the next section.

A different stream of literature, which is loosely related to our work, is hyperparameter optimization; see (Li et al., 2017) for a representative example. The main difference with our setting is that, besides the standard exploration-exploitation trade-off, here we also have to deal with a trade-off induced by non-stationarity. Hyperparameter optimization algorithms like the one in (Li et al., 2017) adaptively searches in the space of hyperparameters, and the goal is akin to best arm identification. Yet, the feedback is assumed to be *stationary*, since hyperparameter values do not correspond to stateful objects (as in the case of our

base learners), and hyperparameter configurations are usually evaluated on a separate validation set. An adversarial variant of hyperparameter optimization was considered in (Jamieson & Talwalkar, 2016), but their notion of regret is different from ours.

In the bandits literature, there are two standard ways of modeling non-stationarity: *restless* (Whittle, 1988; Tekin & Liu, 2012; Ortner et al., 2014; Russac et al., 2019) and *rested* (Levine et al., 2017; Mintz et al., 2017; Kleinberg & Immorlica, 2018; Seznec et al., 2018; Pike-Burke & Grunewalder, 2019; Cella & Cesa-Bianchi, 2020; Kolobov et al., 2020) bandits. In the restless case, the non-stationary nature of the feedback is determined only by the environment, and the learning policies either try to detect changes in the payoff distribution in order to restart the learning model, or to apply a weight-decay scheme to the collected observations. On the contrary, in rested bandits, the non-stationarity depends on the learning policy itself. For instance, in (Cella & Cesa-Bianchi, 2020; Kleinberg & Immorlica, 2018; Kolobov et al., 2020; Pike-Burke & Grunewalder, 2019), an arm expected payoff distribution is parametrized by the elapsed time since that arm was last pulled. The main leverage given to the proposed solutions is the possibility of observing more unbiased samples corresponding to a fixed arm-delay pair. This simplifies the parameter estimation problem. Similar to the setting we are proposing, in (Levine et al., 2017; Seznec et al., 2018) the authors assume the expected loss of an arm to be monotonically *increasing* in the number of times the arm was pulled. The striking difference is that, in their variant, a simple greedy solution which at each round selects the currently-best arm is actually an optimal solution. Therefore, their learning problem reduces to estimating for each arm the expected loss corresponding to its next pull and always select the most promising one. In our setting (see Section 3 below), because expected losses are *decreasing*, a similar solution would be far from optimal, since our objective is to identify the arm minimizing the resulting loss at the end of the game.

3. Learning setting

We consider a set of K arms (or learning agents) $\mathcal{K} = [K] = \{1, \dots, K\}$, whose average performance improves as we play them. At each round $t \in [T]$, the learner picks an arm $I_t \in \mathcal{K}$ and observes the realization $X_{I_t,t}$ of a loss random variable whose (conditional) expectation $\mu_{I_t,t}$ is a decreasing function of the number of times arm I_t has been pulled so far. Specifically, for any $i \in \mathcal{K}$ and $t \in [T]$, denote by $\tau(i,t)$ the number of times arm i has been pulled up to time t, and by $\mathcal{F}_t$ the σ-algebra generated by the past history of pulls and loss random variables $I_1, X_{I_1,1}, \dots, X_{I_{t-1},t-1}$. Given a time horizon T, a learning policy π is a function that maps at each time $t \in [T]$ the observed history $I_1, X_{I_1,1}, \dots, I_{t-1}, X_{I_{t-1},t-1}$ to the next action $I_t \in \mathcal{K}$. At the end of round T, policy π has to commit to (or to output) a given action $i_{\text{out}} \in \mathcal{K}$. We define

$$\mu_{i,t} \equiv \mathbb{E}_{\mathcal{F}_t}[X_{i,t}] = \frac{\alpha_i}{\big(1+\tau(i,t-1)\big)^{\rho}} + \beta_i \ , \qquad (2)$$

where exponent $\rho \in (0,1]$ is a known parameter common to all arms while, for all arms $i \in \mathcal{K}$, scaling parameter α_i and position parameter β_i are assumed to be non-negative but *unknown* to the learning algorithm. We assume $\alpha_i \in [0,U]$ and $\beta_i \in [0,1]$, where the upper extreme U is a known quantity. Hence, $\mu_{i,t}$ is the expected loss of arm i at round t, conditioned on the fact that i has already been played $\tau(i,t-1)$ times during the previous $t-1$ rounds.
As a shorthand, from now on we will use $\mu_i(\tau)$ to denote the expected loss of arm $i \in \mathcal{K}$ if pulled so far $\tau \in [T]$ times. Notice that when $\alpha_i = 0$ for all $i \in \mathcal{K}$ our setting reduces to the standard stochastic multi-armed bandit setting (e.g. (Auer et al., 2002)).[1] It is the decaying component $\frac{\alpha_i}{(1+\tau(i,t-1))^{\rho}}$ that makes this setting an instance of the *rested* bandit setting, where the stochastic behavior of the arms depends on the actual policy $I_1, \dots, I_{t-1}$ that has so far been deployed during the game.

We compare a learning policy π to the optimal policy that knows all parameters $\{\alpha_i, \beta_i\}_{i \in \mathcal{K}}$ in advance, and pulls from beginning to end the arm i_T^* whose expected loss at time T is smallest, i.e.,

$$i_T^* = \arg\min_{i \in \mathcal{K}} \left(\frac{\alpha_i}{T^{\rho}} + \beta_i \right) \ .$$

We define the *pseudo regret* of π after T rounds as

$$R_T^{\pi}(\underline{\mu}) = \mu_{i_{\text{out}}}\big(\tau_{\text{out}}\big) - \mu_{i_T^*}\big(T\big) \ , \qquad (3)$$

where $\tau_{\text{out}} = \tau(i_{\text{out}}, T)$ is the random variable counting the number of pulls of arm $i_{\text{out}} \in \mathcal{K}$ after T rounds. In the above, $\underline{\mu} \in \{\mu_i : [T] \to [0,1]\}_{i \in \mathcal{K}}$ collectively denotes the non-stationary environment generating the observed losses. Our goal is to bound pseudo-regret $R_T^{\pi}(\underline{\mu})$ with high probability, where the probability is w.r.t. the random draw of variables $X_{i,t}$ (and possibly the random choice of $I_1, \dots, I_T$, and i_{out}). Notice that $R_T^{\pi}(\underline{\mu})$ is always non-negative.

A closer inspection of Eq. (3) reveals that, unlike standard best-arm identification problems (e.g., (Even-Dar et al., 2006; Audibert & Bubeck, 2010; Gabillon et al., 2012; Kaufmann et al., 2016)), our objective here is not limited to predicting which arm is best at the end of the game, but also to pull it as much as we can, that is, to single it out as early as possible. This also entails that if the arm our policy π pulls the most throughout the T rounds is $i \neq i_T^*$,

[1] Observe that the stationary case can equivalently be recovered by setting $\rho = 0$, which is therefore redundant and ruled out by the condition $\rho \in (0,1]$.

then it may be better for π to output $i_{\text{out}} = i$ rather than i_T^* itself, even if π gets to know at some point the identity of i_T^* and starts pulling it from that time onward. This is because if, say, for some t_0 close to T we have $\tau(i, t_0) = t_0$ and $\tau(i_T^*, t_0) = 0$, then we may well have $\mu_{i,t_0} < \mu_{i_T^*, T-t_0}$, so that (3) is smaller for $i_{\text{out}} = i$ than for $i_{\text{out}} = i_T^*$. In order to gather further insights, it is also worth considering the simple policy π which selects all arms T/K times, and then outputs the best arm i_T^*. According to Eq. (3), π will still suffer significant regret, since it did not play i_T^* often enough throughout the T rounds (that is, π has explored "too much" on sub-optimal arms). We can thus claim that, thanks to the presence of the τ_{out} variable, our regret in (3) is only seemingly non-cumulative.

Finally, observe that the average loss $\mu_{i,t}$ in (2) can be expressed as the linear combination $\mu_{i,t} = x_t^\top \theta_i^*$, where $\theta_i^* = [\alpha_i, \beta_i]^\top$ is the unknown vector associated with arm i, and $x_t = [1/\tau(i, t-1)^\rho, 1]^\top$ is the "context" vector at time t. This might give the impression of some kind of linear contextual bandit (e.g., (Soare et al., 2014)) in the best arm identification regime. Yet, this impression is erroneous, since in our problem x_t is itself generated by the learning policy during its online functioning.

In the sequel, we also adopt the notion of state $\underline{\tau} = (\tau_1, \tau_2, \dots \tau_K) \in [T]^K$ to encode the case where, for all $i \in \mathcal{K}$, arm i has been pulled τ_i times. Notice that when the learning policy is at state $(\tau, \dots, \tau)$, keep sampling all arms in a round-robin fashion (exploring) entails observing K many samples with expected value $\mu_1(\tau), \dots, \mu_K(\tau)$ respectively, and ending up into state $(\tau + 1, \dots, \tau + 1) \in [T]^K$. Conversely, when the learning policy is at state $(\tau, \dots, \tau)$, then keep pulling the same arm $i \in \mathcal{K}$ for the remaining $T - K\tau$ rounds (exploiting) corresponds to reaching the furthest still reachable state where arm i (which will then be the most pulled one) will have expected loss $\mu_i(T - (K-1)\tau)$.

On the comparison to Bandit Model Selection. The reader may wonder to what extent our task is similar to the bandit model selection problem (Agarwal et al., 2017; Foster et al., 2019; 2020; Pacchiano et al., 2020; Cutkosky et al., 2021) we alluded to in the related works section. At a high-level, this question is similar to the difference between best-arm identification (BAI) and regret minimization (RM). We can coarsely claim that the cited papers correspond to the generalization to stateful arms of RM, while here we generalize BAI to stateful arms having a specific parametric form of their losses. In fact, we cannot easily compare to bandit model selection, since our problem is substantially different. To see why, consider two base learners (say, two students) whose expected loss curves intersect just at time $T - 1$. The first one has $\alpha_1 = 0$ and $\beta_1 > 0$ (not improving over time but looking promising at the beginning), the second one has $\alpha_2 > 0$ and $\beta_2 = 0$ (lagging behind at the beginning, but able to ramp up faster over time). In our BAI setting, the optimal policy would stick to arm 2 since $\mu_2(T) < \mu_1(T)$, while an RM policy would clearly seek to play arm 1. This also helps elucidate the key role played by T: As $T \to \infty$ arm 2 becomes more attractive even in the RM sense, while if we stop earlier, say $T/2$ (so that the two loss curves no longer intersect), then arm 1 becomes better also in our BAI setting. All in all, in our setting one has to depart from the general idea, quite common in RM, of studying regret "for large T" (or even "for large K"), since the complex interplay among problem parameters makes these investigations less meaningful than in RM. Observe that, as $T \to \infty$ our setting turns to standard stationary BAI, since $\mu_{i,t} \to \beta_i$ in Eq. (2). Hence, what matters here is the case when T is *not* large.

4. Main trade-offs and lower bound

In this section we provide a distribution-dependent lower bound for the proposed setting. This will also give us the chance to comment on the specific features of our learning task in terms of the main trade-offs a learning policy has to face. We start by defining the class of *arm-elimination policies* as those which periodically remove sub-optimal arms and keep sampling in a round-robin fashion[2] the remaining arms across the rounds. The following simple fact holds.

Fact 1. *The regret incurred by an arbitrary policy π is invariant to permutations of the chronological order of its actions. In fact,* (3) *only depends on the arm i_{out} selected at the end, and the number of times τ_{out} that arm has been chosen during the T rounds. Hence, for any π, there exists an arm-elimination policy π' whose regret is not worse (i.e., having the same pair $i_{\text{out}}, \tau_{\text{out}}$).*

Proof. Let for simplicity $K = 2$: (i) Since expected losses (2) decrease with the noumber of pulls, it is easy to see that the best thing a policy can do so as to minimize regret (3) is to decide as early as possible which arm to commit to, pull that arm from that point on, and output it as i_{out}. Any deviation from this results in higher regret. (ii) In the initial stage before commitment, the best thing a policy can do is to *equalize* the number of pulls of the two arms (hence the round-robin sampling of an arm-elimination policy). Any deviation from this equalization strategy can be penalized by the adversary generating parameters (α_i, β_i): if we pull arm 1 more than arm 2 the adversary may have made 2 the optimal arm at the time we commit, and vice versa. □

We can therefore restrict our lower bound investigation to arm-elimination policies. An advantage of this restriction is a more convenient characterization of the state space $\{\underline{\tau}\}$

[2]For simplicity, we restrict here to deterministic policies.

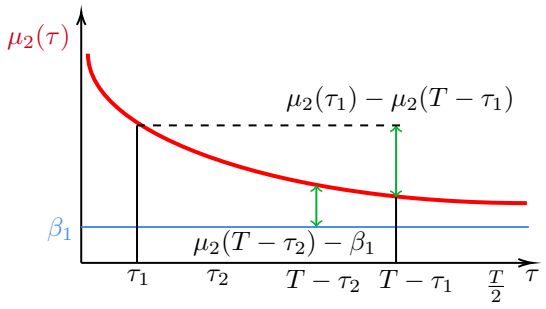

Figure 1. Expected losses associated with the arms in Example 1.

associated with the learning problem. The size of the state space is clearly of the form T^K.

Another relevant aspect of our problem is that, based on (2) and (3), for each given state $(\tau_1, \dots, \tau_K) \in [T]^K$, there are at most K many candidate optimal and still reachable states that any policy could end up to. These are specifically the K alternative states that the learning policy at hand would reach by committing to one of the K arms for all the remaining rounds. All other states (which are exponential many) can easily be seen to be sub-optimal.

Before moving to the main result of this section (the regret lower bound), we would like to give an additional characterization of the considered class of policies. The missing component which gives a well-specified policy is the condition governing the arm elimination. Since expected losses (2) are non-increasing, and given the regret criterion (3), once a policy is confident that sticking to an arm would give a smaller expected loss than the one associated with the last pull, this policy might be tempted to eliminate all the other arms. In the next example we show that operating this way can be sub-optimal.

Example 1. *Let us consider the specific instance of our problem with $K = 2$ arms whose expected losses are sketched in Figure 1. Whereas the first arm is stationary $\mu_1(\tau) = \beta_1$, the second is not, $\mu_2(\tau) = \frac{\alpha_2}{\tau^\rho} + \beta_1$. At state $\underline{\tau}_1 = (\tau_1, \tau_1)$ it may occur that the τ_1 observations associated with arm 2 are enough to realize that $\mu_2(T - \tau_1) < \mu_2(\tau_1)$. Hence the learning policy knows that if it kept sampling arm 2 for the remaining $T - 2\tau_1$ pulls it would achieve a smaller (expected) loss compared to $\mu_2(\tau_1)$. The same would not hold for the other arm, as it is stationary.*
Let us now denote by τ_2 the number of pulls it takes to figure out that $\beta_1 < \mu_2(T - \tau_2)$. It could be the case that $\tau_2 > \tau_1$ (that is, as in Figure 1, we have $\mu_2(T - \tau_2) - \beta_1 < \mu_2(\tau_1) - \mu_2(T - \tau_1)$). In order to maximize τ_{out} (so as to minimize regret (3)) a naive policy might eliminate arm 1 after τ_1 observations. This would translate into choosing the wrong value of i_{out}, hence clearly incurring a regret. Conversely, a smarter policy that keeps exploring up to state (τ_2, τ_2) would return $i_{\text{out}} = 1$ and yield $\tau_{\text{out}} = T - \tau_2$. Notice that, thanks to the stationary nature of the optimal arm, the regret incurred by this policy is indeed zero.

All in all, the above observations help better understand the structure of our problem, which will be useful in all technical proofs (see the appendix). We can now turn our attention to the lower bound. In doing so, we generalize the results in (Bubeck et al., 2013), which in turn adopts a hypothesis testing argument that hinges on a lower bound for the minimax risk of hypothesis testing (ee e.g. Tsybakov, 2008, Ch. 2). Notice that the classical lower bound result for stationary bandits (Lai & Robbins, 1985) cannot easily be adapted here since, being asymptotic in nature, that result tends to lose the non-stationary component of our expected losses (2), and thus the cumulated effect of this non-stationarity on the τ_{out} variable.

As done in (Bubeck et al., 2013), for all arms and all possible number of pulls, we consider all families of loss distributions $\{\mathbb{P}_\mu\}$, indexed by their expected value μ, and such that $KL(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) = C(\mu - \mu')^2$ for some absolute constant $C > 0$ (e.g., in the case of normal distributions, $KL(\mathcal{N}(\mu, \sigma), \mathcal{N}(\mu', \sigma)) = \frac{1}{2}(\mu - \mu')^2$). In the sequel, we use $\tau_{\text{sub}} = T - \tau_{\text{out}}$ to denote the number of rounds spent by pulling all arms different from i_{out}. Additionally, we denote by $\mathbb{P}_{\underline{\mu}(\tau)} = \mathbb{P}_{\mu_1(\tau)} \otimes \dots \otimes \mathbb{P}_{\mu_K(\tau)}$ the product distribution that generates the losses from $\mathbb{P}_{\mu_i(\tau)}$ when pulling arm $i \in \mathcal{K}$ for the τ-th time. The result that follows restricts to the two arm case,[3] and delivers a bound on the regret that holds in expectation over the loss random draws.

Theorem 1. *Let $\mathbb{P}_{\underline{\mu}(\tau)} = \mathbb{P}_{\mu_1(\tau)} \otimes \mathbb{P}_{\mu_2(\tau)}$ be defined by distributions whose expected values are $\mu_1(\tau) = \frac{\alpha_1}{\tau^\rho} + \beta$ and $\mu_2(\tau) = \frac{\alpha_2}{\tau^\rho} + \beta + \Delta$, respectively, where $\Delta > 0$ is an unknown but fixed constant. Additionally, let $\mathbb{P}_{\underline{\mu}'(\tau)} = \mathbb{P}_{\mu_1'(\tau)} \otimes \mathbb{P}_{\mu_2'(\tau)}$ be another product distribution, whose expected value components are $\mu_1'(\tau) = \mu_1(\tau)$ and $\mu_2'(\tau) = \mu_1(\tau) - \Delta$. Then, for any policy π, and any horizon $T \geq 1$, the quantity $max\Big\{\mathbb{E}\big[R_T^\pi(\underline{\mu})\big], \mathbb{E}\big[R_T^\pi(\underline{\mu}')\big]\Big\}$ can be lower bounded by*

$$\begin{cases} \mu_2(T - \tau_{\text{sub}}) - \mu_{i_T^*}(T) & \text{if } T - \tau_{\text{sub}} \leq \lceil \overline{\tau} \rceil,\ \alpha_1 > \alpha_2 \\ \mu_1(T - \tau_{\text{sub}}) - \mu_{i_T^*}(T) & \text{if } T - \tau_{\text{sub}} \geq \lceil \overline{\tau} \rceil,\ \alpha_1 > \alpha_2 \\ \alpha_1 \left(\frac{1}{(T - \tau_{\text{sub}})^\rho} - \frac{1}{T^\rho} \right) & \text{if } \alpha_1 \leq \alpha_2\ , \end{cases}$$

where $\overline{\tau} = \left(\frac{\alpha_1 - \alpha_2}{\Delta}\right)^{\frac{1}{\rho}}$ satisfies $\mu_1(\overline{\tau}) = \mu_2(\overline{\tau})$. We also have $i_T^* = \{1\}$ if $\overline{\tau} < T$ and $i_T^* = \{2\}$ otherwise. Finally,

[3] We believe that restricting to the two arm case helps better elucidate the nature and trade-offs in our problem. We conjecture that a similar but more involved result can be shown for K arms.

$\tau_{\rm sub}$ is the smallest $\tau \in [T/2]$ which is strictly larger than

$$\min\Bigg\{ \frac{1}{8C\Delta_\tau^2}\log\Big(C\Delta_\tau^2\tau/4\Big), \frac{1}{8C\widetilde{\Delta}_{1,\tau}^2}\log\Big(C\widetilde{\Delta}_{1,\tau}^2\tau/4\Big), \frac{1}{8C\widetilde{\Delta}_{2,\tau}^2}\log\Big(C\widetilde{\Delta}_{2,\tau}^2\tau/4\Big)\Bigg\},$$

where $\Delta_\tau = \mu_1(\tau) - \mu_2(\tau)$ and $\widetilde{\Delta}_{i,\tau} = \mu_i(T-\tau-1) - \mu_i(T-\tau)$.

The proof is given in the appendix. The main idea behind it is that under the assumptions of Theorem 1, we have two quantities characterizing the lower bound on the number of sub-optimal pulls $\tau_{\rm sub}$. The first one is associated with $i_{\rm out}$, and is of order $1/\Delta_\tau^2$. The second one is induced by the objective of minimizing the incurred expected loss at $\tau_{\rm out}$, and is of order $1/\widetilde{\Delta}_{i,\tau}^2$. The main point here is that exploring towards arm $i_{\rm out}$ is worthwhile only if it does not cause a higher incurred loss $\mu_{i_{\rm out}}(\tau_{\rm out})$. The second ingredient is specified by relation between α_1 and α_2. In fact, if $\alpha_1 \leq \alpha_2$ we have that the second arm's expected loss $\mu_2(\tau)$ is always worse than the one of the first, if considered at the same number of pulls τ. On the other hand, if $\alpha_1 > \alpha_2$ their order relation depends on $\overline{\tau}$.

We would like to emphasize that even when $\alpha_1 = \alpha_2$, the horizon T is large enough, and $\Delta \to 0$ (that is, the two arms are less and less statistically distinguishable), an optimal strategy for our regret minimization problem is by no means to pull both arms an equal $(T/2)$ number of times. Rather, an optimal strategy would commit to one of the two arms as soon as it is confident enough on which of them has the smaller loss (at any reachable state), unless trying to determine the best arm causes a bigger regret than immediately committing to any of the two. The quantity $\tau_{\rm sub}$ (at least in the two-arm case) will play a key role in characterizing the statistical complexity of our learning problem.

5. Estimation of parameters

In order to minimize the regret $R_T^\pi(\underline{\mu})$, any reasonable policy π has to be able to estimate, for all arms $i \in \mathcal{K}$, the associated expected loss $\mu_i(\cdot)$, and it has to do so at any still reachable state where arm i will be pulled $\tau_{\rm out}$ times. To this effect, we now introduce two statistically independent estimators. Upon pulling arm $i \in \mathcal{K}$ for 2τ times, we define

$$\widehat{X}_{i,\tau} = \frac{1}{\tau}\sum_{s=1}^{\tau} X_i(s)\,, \quad \widetilde{X}_{i,\tau} = \frac{1}{\tau}\sum_{s=\tau+1}^{2\tau} X_i(s)\,, \quad (4)$$

where $X_{i,s}$ denotes the loss incurred by arm i after having pulled it s times ($\mathbb{E}[X_{i,s}] = \mu_{i,s}$). Notice, that due to the way we have defined $\mu_{i,s}$, the above estimators are empirical averages of *independent* but non-identically distributed random variables, the independence deriving from the fact that pulling one arm does not influence the distribution of the others. Moreover, because of the time decay, the expectation of $\widehat{X}_{i,\tau}$ cannot be smaller than the one of $\widetilde{X}_{i,\tau}$.
We combine these two estimators together with standard concentration inequalities to derive a joint estimator for (α_i, β_i). Since the two estimators are non-redundant, this allows us to come up with estimators for α_i and β_i individually. Using Bernstein's inequality,[4] we can derive confidence bounds around $\widehat{X}_{i,\tau}$ and $\widetilde{X}_{i,\tau}$ as functions of β_i and α_i. Specifically, for each arm $i \in \mathcal{K}$, number of pulls $2\tau \in [T]$, the expectation $\mathbb{E}[\widehat{X}_{i,\tau}]$ is contained with probability at least $1-\delta$ in the interval $[\widehat{X}_{i,\tau} - \mathrm{CB}_{\widehat{X},\tau}(\delta), \widehat{X}_{i,\tau} + \mathrm{CB}_{\widehat{X},\tau}(\delta)]$ where

$$\mathrm{CB}_{\widehat{X},\tau}(\delta) = \Big(\sqrt{U}+1\Big)^2\sqrt{\frac{2}{\tau}\log\frac{1}{\delta}} + \frac{(U+1)\log\frac{1}{\delta}}{\tau}\,.$$

A similar argument follows for $\mathbb{E}[\widetilde{X}_{i,\tau}]$ and $\mathrm{CB}_{\widetilde{X},\tau}(\delta)$. We defer to the appendix the details of the exact expression for the confidence intervals. Starting from these definitions we can build the following set of inequalities

$$\mathbb{E}[\widehat{X}_{i,\tau}] - \mathrm{CB}_{\widehat{X},\tau}(\delta) \leq \widehat{X}_{i,\tau} \leq \mathbb{E}[\widehat{X}_{i,\tau}] + \mathrm{CB}_{\widehat{X},\tau}(\delta)$$
$$\mathbb{E}[\widetilde{X}_{i,\tau}] - \mathrm{CB}_{\widetilde{X},\tau}(\delta) \leq \widetilde{X}_{i,\tau} \leq \mathbb{E}[\widetilde{X}_{i,\tau}] + \mathrm{CB}_{\widetilde{X},\tau}(\delta)$$

which can be solved for α_i and β_i individually. As shown in the appendix, this gives rise to the following confidence intervals for α_i :

$$\alpha_i \in \overbrace{\frac{\tau\Delta\widehat{X}_{i,\tau}}{\sum_{s=1}^{\tau}\frac{1}{s^\rho} - \sum_{s=\tau+1}^{2\tau}\frac{1}{s^\rho}}}^{\widehat{\alpha}_{i,\tau}} \pm \pm \frac{5\tau^\rho\big(\sqrt{U}+1\big)^2}{\rho}\Big[\frac{\log 1/\delta}{\tau} + \sqrt{\frac{1}{\tau}\log\frac{1}{\delta}}\Big]\,, \quad (5)$$

where $\Delta\widehat{X}_{i,\tau} = \widehat{X}_{i,\tau} - \widetilde{X}_{i,\tau}$. For brevity, the confidence interval centroid will be denoted by $\widehat{\alpha}_{i,\tau}$. Similarly, β_i can be shown to satisfy

$$\beta_i \in \overbrace{\widehat{X}_{i,\tau} - \frac{\widehat{\alpha}_{i,\tau}}{\tau}\sum_{s=1}^{\tau}\frac{1}{s^\rho}}^{\widehat{\beta}_{i,\tau}} \pm \frac{5\big(\sqrt{U}+1\big)^2}{(1-\rho)\rho}\left[\frac{\log 1/\delta}{\tau} + \sqrt{\frac{1}{\tau}\log\frac{1}{\delta}}\right], \quad (6)$$

where $\widehat{\beta}_{i,\tau}$ denotes the centroid of confidence interval (6). Despite we have provided separate estimators for α_i and β_i, it is important to stress that our interest here is not to estimate them separately. Combining these estimators gives

[4] It is worth mentioning in passing that the standard Hoeffding inequality delivers vacuous estimators here.

$$\widehat{\mu}_{i,\tau}(\tau_{\text{out}}) = \frac{\widehat{\alpha}_{i,\tau}}{\tau_{\text{out}}^{\rho}} + \widehat{\beta}_{i,\tau} \ ,$$

an estimate of the expected loss incurred by arm $i \in \mathcal{K}$ as if we had pulled it τ_{out} times after having observed only 2τ realizations of $X_{i,t}$. All the above can be summarized by the following theorem.

Theorem 2. *After observing $X_{i,1}, \dots, X_{i,2\tau}$ loss realizations of arm $i \in \mathcal{K}$, we can predict the expected loss $\mu_{i,\tau_{\text{out}}}$ of arm i as it were pulled τ_{out}-many times (with $\tau_{\text{out}} > \tau$). In particular, we have that with probability at least $1-\delta$ jointly over $i \in \mathcal{K}$, $\tau \in [T]$ and $\tau_{\text{out}} \in [T]$,*

$$\widehat{\mu}_{i,\tau}(\tau_{\text{out}}) - \boldsymbol{CB}_{\mu,\tau}(\delta) \le \mu_{i,\tau_{\text{out}}} \le \widehat{\mu}_{i,\tau}(\tau_{\text{out}}) + \boldsymbol{CB}_{\mu,\tau}(\delta) \ ,$$

where

$$\text{CB}_{\mu,\tau}(\delta) = \frac{10\left(\sqrt{U}+1\right)^2}{(1-\rho)\rho}\left[\frac{\log\frac{\tau KT}{\delta}}{\tau} + \sqrt{\frac{1}{\tau}\log\frac{\tau KT}{\delta}}\right] .$$

Hence, the approach contained in Theorem 2 allows us to obtain confidence intervals for $\mu_{i,\tau_{\text{out}}}$ shrinking with τ as $\frac{1}{\sqrt{\tau}}$ up to a numerical constant depending on ρ and U.
Finally, observe that these confidence intervals are non-vacuous only when $\rho \in (0,1)$, that is, excluding the extreme cases $\rho = 0$ and $\rho = 1$. The case $\rho = 0$ is indeed uninteresting, since it yields a stationary case which is equivalent to the one achieved by the setting $\alpha_i = 0$ for all i. In fact, due to the specific nature of the empirical averages in (4), when $\rho = 0$ the centroid $\widehat{\alpha}_{i,\tau}$ occurring in (5) is not well defined, independent of the number of observed samples τ. On the other hand, because our derivations rely on approximations of the form $\sum_{s=1}^{\tau} \frac{1}{s^{\rho}} \approx \frac{s^{1-\rho}}{1-\rho}$, which only hold for $\rho \neq 1$, the case $\rho = 1$ should be treated separately via standard approximations of the form $\sum_{s=1}^{\tau} \frac{1}{s} \approx \log\tau$.
The above estimators will be the building blocks of our learning algorithms, presented in the next section. In particular, the definition of $CB_{\mu,\tau}(\delta)$ given in Theorem 2 above will be repeatedly used throughout the rest of the paper.

6. Regret minimization

In this section we present two learning policies. We first describe as a warm-up a simple explore-then-commit strategy, then we present a more sophisticated strategy inspired by the Successive Reject algorithm (Audibert & Bubeck, 2010). For both policies, we set the confidence parameter δ to $\frac{1}{T}$.

The first solution we propose is a rested bandit variant of the standard explore-then-commit (ETC) policy (e.g. Lattimore & Szepesvári, 2020, Ch. 6)). In its original formulation, ETC requires as input a parameter $n \in [T]$ specifying the number of initial pulls associated with each arm. Once all the arms have been pulled n times, the exploratory phase finishes. The original ETC algorithm then sticks to the most promising arm according to the estimates computed during the exploration. Hence the two phases of exploration and exploitation are kept separate. This strategy has a clear limitation. Since the exploration parameter n is an input to the algorithm, the original ETC algorithm does not adapt the length of the exploration phase to the actual samples, so that understanding how to best set n is not a simple task. One thing that is worth noticing is that in the 2-arm bandit case, this parameter n takes values in the range $[T/2]$. If τ_{sub} in our lower bound of Theorem 1 equals $T/2$ (that is, when $T/2$ is smaller than $\log T/\Delta_{\tau}^2$, $\log T/\widetilde{\Delta}_{1,\tau}^2$ and $\log T/\widetilde{\Delta}_{2,\tau}^2$), we cannot commit to any specific arm, and the ETC algorithm results in a solo-exploration strategy which is indeed optimal in this case (up to the choice of i_{out}). Algorithm 1 describes a variant of the standard ETC policy adapted to our rested bandit scenario. At a generic round $t = Kn$, this algorithm starts committing to an arm $i \in \mathcal{K}$ only when we are confident with probability at least $1-\delta$ that i is the arm with lowest expected loss if pulled for the remaining $T - Kn$ times (Line 4). Hence, unlike the original ETC algorithm, this algorithm implicitly computes n on the fly based on the observed samples. Finally, upon committing to an arm, our algorithm does not reconsider its decision based on the newly collected samples (Line 9). We have the following result, that help elucidate the benefit of adaptively inferring n.

Algorithm 1 Explore-Then-Commit (ETC)

Require: Confidence parameter $\delta = 1/T$
1: **for** $n \in 1, \dots, \lfloor T/K \rfloor$ **do**
2: pull each arm once
3: $\tau_{\text{out}} = T - n(K-1)$
4: **if** $\exists i \in \mathcal{K} : \ \widehat{\mu}_{i,n}(\tau_{\text{out}}) < \min_{j \in \mathcal{K}\setminus\{i\}} \widehat{\mu}_{j,n}(\tau_{\text{out}}) - 2CB_{\mu,n}(\delta)$ **then**
5: $i_{\text{out}} = \arg\min_{i \in \mathcal{K}} \widehat{\mu}_{i,n}(\tau_{\text{out}})$
6: break; {The exploration phase terminates}
7: **end if**
8: **end for**
9: Play i_{out} until round T {Commit}
10: Output i_{out}

Theorem 3. *Consider the same two-arm setting $\mathbb{P}_{\underline{\mu}(\tau)} = \mathbb{P}_{\mu_1(\tau)} \otimes \mathbb{P}_{\mu_2(\tau)}$ contained in Theorem 1 and the notation introduced therein. Running Algorithm 1 with $T \ge 1$ yields*

$$R_T^{ETC}(\underline{\mu}) \le \mu_{i_{\text{out}}}(T - n_0) - \mu_{i_T^*}(T)$$

with probability at least $1 - \frac{1}{T}$, where $n_0 = \min\left\{\frac{T}{2}, \frac{c_\rho}{\Delta_{n_0}^2}\right\}$ and $c_\rho = \frac{1600(\sqrt{U}+1)^4}{\rho^2(1-\rho)^2}\log(4n_0T^2)$.
This result is optimal up to a logarithmic factor (namely, $i_{\text{out}} \in \arg\min_{i \in \mathcal{K}} \mu_i(T - n_0)$ and $n_0 = \tau_{\text{sub}}$ up to a logarithmic factor) whenever $n_0 = \frac{c_\rho}{\Delta_{n_0}^2}$ and $\tau_{\text{sub}} = \frac{1}{8C\Delta_\tau^2}\log\left(C\Delta_\tau^2\tau/4\right)$. Conversely, when at least one of the

above conditions is not met, the bound contains an additional $\tilde{O}\left(1/\sqrt{n_0}\right)$ term if compared to the result of Theorem 1, where $\tilde{O}(\cdot)$ hides $\log T$ factors.

Notice that finding the commitment condition for ETC cannot be readily obtained by available results in the bandit literature, as this requires a specific understanding of the interplay among the loss curves. Even in the two arm case of Theorem 3, it is not possible to avoid the unfriendly implicit form of the regret bound coming from the definition of n_0 therein.

Starting from ETC, in the next section we present our final learning policy, which will be analyzed both in the general K-armed case and in the specific setting contained in the lower bound of Section 4.

6.1. Towards an Optimal Policy

The first limitation of the ETC strategy in Algorithm 1 becomes clear when considering more than 2 arms. Let us consider an instance with $K = 3$ arms where there exist two values n_2, n_3 satisfying:

$$\widehat{\mu}_1(n') < \widehat{\mu}_2(n') - 2CB_{\mu,n_2}(n_2) \quad \forall n' > n_2$$
$$\widehat{\mu}_1(n') < \widehat{\mu}_3(n') - 2CB_{\mu,n_3}(n_3) \quad \forall n' > n_3 \, .$$

The ETC policy in Algorithm 1 has a single counter n that has to satisfy at the same time $K - 1 = 2$ arm elimination conditions (line 4 of Algorithm 1). The best this algorithm can do in order not to commit to the wrong arm is to keep exploring up to $n = \max\{n_2, n_3\}$. The obvious drawback of this solution is that ETC would then waste $|n_2 - n_3|$ pulls on the sub-optimal arms 2 and 3, rather than selecting $i_{\text{out}} = 1$. We now present in Algorithm 2 the strategy REST-SURE (RESTed SUccessive REject), a rested version of the Successive Reject algorithm from (Even-Dar et al., 2006; Audibert & Bubeck, 2010). As for its stationary counterpart, REST-SURE keeps sampling all the active arms in a round-robin fashion, and then periodically removes arms once it is confident about their sub-optimality (line 12). The key adaptation to our rested bandit scenario is that one arm is deemed sub-optimal when there is a better arm in *any* of the still reachable states.

Going into some details of the pseudocode, the stopping condition in lines 4 of Algorithm 2 is inspired by the same reasoning governing the commitment in the stationary bandit problem. This condition tells us that exploration has provided enough information to identify (with high probability) arm $i_{\text{out}} = \arg\min_{i \in \mathcal{K}} \mu_i(\tau_{\text{out}})$ at the best reachable state. The second stopping condition (line 8) is due to the non-stationary component in the expected loss (2). This condition controls the trade-off between the estimation of $i_{\text{out}} = \arg\min_{i \in \mathcal{K}} \mu_i(\tau_{\text{out}})$ and the minimization of the incurred expected loss, namely the impact on the value of τ_{out}. In particular, this condition stops the policy in its exploration towards the identity of i_{out} as soon as this would cause an increased regret due to a reduced valued of τ_{out}.

Algorithm 2 REST-SURE

Require: Confidence parameter $\delta = 1/T$
1: Initialize: $\mathcal{A}_0 = \mathcal{K}, n = 0, \tau_{\text{out}} = T$, and $t = 0$.
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad \tau_{\text{out}} = T - t + n$
4: $\quad$ **if** $\exists i \in \mathcal{A}_n : \widehat{\mu}_i(\tau_{\text{out}}) < \min_{j \in \mathcal{A}_n \setminus \{i\}} \widehat{\mu}_j(\tau_{\text{out}}) - 2CB_{\mu,n}(\delta)$ **then**
5: $\quad\quad i_{\text{out}} = \arg\min_{i \in \mathcal{A}_n} \widehat{\mu}_i(\tau_{\text{out}})$
6: $\quad\quad$ break; {Found $i^*_{\tau_{\text{out}}}$ w.h.p.}
7: $\quad$ **end if**
8: $\quad$ **if** $\min_{i \in \mathcal{A}_n} \widehat{\mu}_{i,n}(\tau_{\text{out}} - |\mathcal{A}_n| + 1) - 2CB_{\mu,n}(\delta) > \min_{i \in \mathcal{A}_n} \widehat{\mu}_{i,n}(\tau_{\text{out}})$ **then**
9: $\quad\quad i_{\text{out}}$ randomly chosen in $\mathcal{A}_n$
10: $\quad\quad$ break; {No advantage in learning $i^*_{\tau_{\text{out}}}$}
11: $\quad$ **end if**
12: $\quad \mathcal{A}_{n+1} = \mathcal{A}_n$
13: $\quad$ **for each** arm $i \in \mathcal{A}_{n+1}$ such that $\forall m \in [n, \tau_{\text{out}}] : \exists j \in \mathcal{A}_{n+1} : \widehat{\mu}_{i,n}(m) - \widehat{\mu}_{j,n}(m) > 2CB_{\mu,n}(\delta)$ **do**
14: $\quad\quad \mathcal{A}_{n+1} = \mathcal{A}_{n+1} \setminus \{i\}$ {Arm elimination}
15: $\quad$ **end for**
16: $\quad$ Pull once each active arm $i \in \mathcal{A}_{n+1}$
17: $\quad t = t + |\mathcal{A}_{n+1}|; \quad n = n + 1$
18: **end for**
19: Play i_{out} until round T {Commit}
20: Output i_{out}

We need the following additional notation. We set for brevity $\Delta_{j,i}(\tau) = \mu_j(\tau) - \mu_i(\tau)$ for any $\tau \in [T]$, $K_n = K - n$, and $\mu^*(\tau) = \min_{i \in \mathcal{K}} \mu(\tau)$ denotes the smallest expected loss over all arms after each one of them has been pulled τ times. The following is the main result of this section.

Theorem 4. *For all $K > 1$, if* REST-SURE is run on K arms having arbitrary non-stationary loss distributions $\mathbb{P}_{\underline{\mu}(m)} = \mathbb{P}_{\mu_1(m)}, \otimes, \ldots, \otimes, \mathbb{P}_{\mu_K(m)}$ with support in $[0, 1]$ and expected value parameterized according to (2), then with probability at least $1 - \frac{1}{T}$ the pseudo-regret of REST-SURE after T interactions satisfies

$$R_T^{\text{REST-SURE}}(\underline{\mu}) \leq \mu_{i_{\text{out}}}(T - \overline{n}) - \mu_{i^*_T}(T),$$

where $\overline{n} = \sum_{s \in [K-1]} n_{\sigma(s)}$, and $n_{\sigma(s)}$ is defined as the

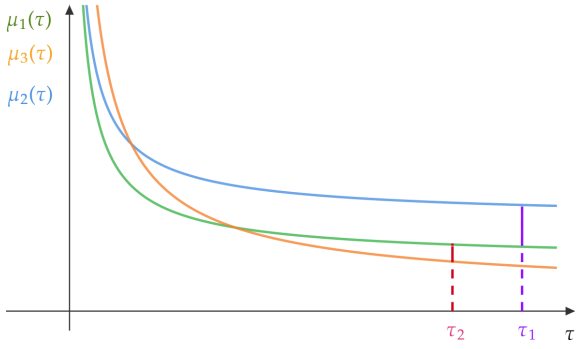

Figure 2. Expected losses associated with the arms in Example 2.

smallest $n \in [T]$ which is greater than

$$\min \left\{ \frac{T - \sum_{j=1}^{s-1} n_{\sigma(j)}}{K_{s-1}}, \frac{c_\rho \log(nK^2T^2)}{\min_{j \in \mathcal{A}^*_{s-1}, m \in [n_{\sigma(s)}, \tau_{\text{out}}(s)]} \Delta^2_{\sigma(s),j}(m)}, \frac{c_\rho \log(nK^2T^2)}{\left(\mu^*\left(\tau_{\text{out}}(s) - K_{s+1}\right) - \mu^*\left(\tau_{\text{out}}(s)\right)\right)^2}, \frac{c_\rho \log(nK^2T^2)}{\left(\min_{j \in \mathcal{K}} \left(\Delta_{\sigma(s),j}(\tau_{\text{out}}(s))\right)\right)^2} \right\}.$$

In the above, $\mathcal{A}^*_s = \mathcal{K} \setminus \{\sigma(1), \dots, \sigma(s-1)\}$, $\tau_{\text{out}}(n) = T - \sum_{s=1}^{n} K_{s+1} n_{\sigma(s)}$ and $c_\rho = \frac{1600(\sqrt{U}+1)^2}{\rho^2(1-\rho)^2}$. Notice that $\sigma(s) = \arg\min_{j \in \mathcal{A}^*_{s-1}} n_{\sigma(j)}$. Finally, $i_{\text{out}} \in \arg\min_{i \in \mathcal{K}} \mu_i(T - \overline{n})$ only if $\min_{j \in \mathcal{A}^*_{s-1}} \Delta_{\sigma(s),j}(\tau_{\text{out}}(s))$ is greater than $\mu^*\left(\tau_{\text{out}}(s) - K_{s+1}\right) - \mu^*\left(\tau_{\text{out}}(s)\right)$. Conversely, when the latter condition is not met we can only guarantee that $\mu_{i_{\text{out}}}(T - \overline{n}) \leq \mu^*(T - \overline{n}) + \tilde{O}\left(1/\sqrt{n_{i_{\text{out}}}}\right)$.

Notice that $\overline{n}$ is solely a function of the problem parameters $\{\alpha_i, \beta_i\}_{i \in \mathcal{K}}, \rho, K$, U, and T. This is because so are the involved quantities $\sigma(s)$ and $n_{\sigma(s)}$. The exact expression for $\overline{n}$ might be somewhat hard to interpret. The first term in the min plays the same role as term $T/2$ in Theorem 1, and guarantees the total number of pulls is most T. The second term is obtained from the arm-elimination condition of line 12. The third term in the min is obtained by analyzing the condition at lines 7-8. Finally, the commitment to arm i_{out} yields the fourth term. In order to clarify the heavy statement of Theorem 4, we now present a symbolic example.

Example 2. *For the sake of illustration, let us consider the specific instance of the ahove theorem with $K = 3$ arms whose expected losses are sketched in Figure 2. In that figure, τ_1 is the number of times REST-SURE needs to pull each arm before eliminating arm 2. Hence, we have $\sigma(1) = 2$ and $n_{\sigma(1)} = \tau_1$. Similarly, τ_2 is the number of pulls associated with the remaining arms $\mathcal{A}_{\tau_2} = \{1, 3\}$ before committing to arm 3. Hence, with probability at least $1 - 1/T$, we have $\sigma(2) = 1$, $i_{\text{out}} = 3$ and $\overline{n} = \tau_1 + \tau_2$.*

The proof of Theorem 4 is an extension of the proof of Theorem 3, and is given in the appendix. To conclude, we now show that the bound of REST-SURE matches the lower bound given in Theorem 1 up to logarithmic factors.

Corollary 1. *Let us consider the same two-arm setting $\mathbb{P}_{\underline{\mu}(\tau)} = \mathbb{P}_{\mu_1(\tau)} \otimes \mathbb{P}_{\mu_2(\tau)}$ as in Theorem 1 and the notation introduced therein. Running Algorithm 2 with $T \geq 1$ yields*

$$R_T^{\text{REST}-\text{SURE}}(\underline{\mu}) \leq \mu_{i_{\text{out}}}(T - n_0) - \mu_{i^*_T}(T)$$

with probability at least $1 - \frac{1}{T}$. In the above, $n_0 = \min\left\{\frac{c_\rho}{\Delta^2_{n_0}}, \frac{c_\rho}{\tilde{\Delta}^2_{n_0}}, \frac{T}{2}\right\}$ and $c_\rho = 25600(\sqrt{U} + 1)^4 \log(4n_0T^2)$. This result is optimal up to a logarithmic factor (namely, $i_{\text{out}} \in \arg\min_{i \in \mathcal{K}} \mu_i(T - n_0)$ and $n_0 = \tau_{\text{sub}}$ up to a logarithmic factor) whenever $n_0 = \frac{c_\rho}{\Delta^2_{n_0}}$ and $\tau_{\text{sub}} = \frac{1}{8C\Delta^2_\tau} \log\left(C\Delta^2_\tau \tau/4\right)$. Conversely, when at least one of the above conditions is not met, the bound contains an additional $\tilde{O}\left(1/\sqrt{n_0}\right)$ term if compared to the result of Theorem 1, where $\tilde{O}(\cdot)$ hides $\log T$ factors.

Notice that, differently from Theorem 3, the definition of n_0 now matches the one of τ_{sub} in Theorem 1 up to a logarithmic factor. Furthermore, the extra term $\tilde{O}\left(\frac{1}{\sqrt{n_0}}\right)$ is always smaller than the one mentioned in Theorem 3.

7. Conclusions and Ongoing Research

In this work we have proposed an online algorithm selection problem formulated as a best arm identification within a specific rested bandit scenario. Here, each arm represents a candidate learning model and each pull corresponds to giving the associated learner more i.i.d training samples, thus allowing the learner to reduce its generalization error. We formulated an ad hoc notion of regret, provided a lower bound for the learning problem, and analyzed two alternative strategies, one of which we have shown to be optimal in the cases covered by the lower bound.

We considered losses of the parametric form $f(\tau; (\alpha, \beta)) = \frac{\alpha}{\tau^\rho} + \beta$ due to their relevance when considering the typical behavior of generalization error as a function of training set size. Yet, our analysis can be generalized to any parametric family of non-increasing functions $f(\tau; \theta)$, where θ is a vector of parameters.

At last we note that, while the focus of this paper has been theoretical, future work may be devoted to study the empirical performance of our methods and the underlying bounds. In Appendix D we included simple preliminary experiments on synthetic data that help corroborate our theoretical findings. A deeper investigation on real-world data is left to the future.

References

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38, 2017.

Allesiardo, R., Féraud, R., and Maillard, O.-A. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4): 267–283, 2017.

Audibert, J.-Y. and Bubeck, S. Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, pp. 13 p., Haifa, Israel, June 2010. URL https://hal-enpc.archives-ouvertes.fr/hal-00654404.

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

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Bubeck, S., Perchet, V., and Rigollet, P. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pp. 122–134, 2013.

Cella, L. and Cesa-Bianchi, N. Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1168–1177, 2020.

Cesa-Bianchi, N. *Multi-armed Bandit Problem*. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4.

Cutkosky, A., Dann, C., Das, A., Gentile, C., Pacchiano, A., and Purohit, M. Dynamic balancing for model selection in bandits and rl. In *ICML*, 2021.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Foster, D., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in linear contextual bandits and beyond. In *Advances in Neural Information Processing Systems*, 2020.

Foster, D. J., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 14714–14725, 2019.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Jamieson, K. and Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pp. 240–248, 2016.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kleinberg, R. and Immorlica, N. Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 309–319. IEEE, 2018.

Kolobov, A., Bubeck, S., and Zimmert, J. Online learning for active cache synchronization. *arXiv preprint arXiv:2002.12014*, 2020.

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

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Levine, N., Crammer, K., and Mannor, S. Rotting bandits. In *Advances in neural information processing systems*, pp. 3074–3083, 2017.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y. Non-stationary bandits with habituation and recovery dynamics. *arXiv preprint arXiv:1707.08423*, 2017.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 558:62–76, Nov 2014. ISSN 0304-3975. doi: 10.1016/j.tcs.2014.09.026. URL http://dx.doi.org/10.1016/j.tcs.2014.09.026.

Pacchiano, A., Phan, M., Abbasi-Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020.

Pike-Burke, C. and Grunewalder, S. Recovering bandits. In *Advances in Neural Information Processing Systems*, pp. 14122–14131, 2019.

Russac, Y., Vernade, C., and Cappé, O. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pp. 12017–12026, 2019.

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. Rotting bandits are no harder than stochastic ones. *arXiv preprint arXiv:1811.11043*, 2018.

Siegmund, D. Herbert Robbins and sequential analysis. *Annals of statistics*, pp. 349–365, 2003.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Tekin, C. and Liu, M. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58 (8):5588–5611, 2012.

Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Whittle, P. Restless bandits: Activity allocation in a changing world. *booktitle of applied probability*, 25(A):287–298, 1988.