Selection of Learning Experts

Robin Allesiardo[1,2] and Raphaël Féraud[2]
[1] TAO - Laboratoire de Recherche en Informatique
[2] PROF - Orange Labs

***Abstract*—The contextual bandits can be viewed as a generalization of online classification models, where only the chosen class is observed. The selection of learning experts allows to find the best parametrization of an expert during its learning, within a set of predefined parameters, and reduces the bias of the hypothesis space, and hence improves the performances. As the contextual bandits learn, their performances tend to increase during time, and hence the choice of the best one implies to solve a non-stationary problem. We provide a theoretical framework to solve this difficult problem. A first approach to handle the selection of learning experts problem is to reduce it to stochastic problems: the experts are learned in parallel during a first phase, then they are explored and exploited in a second phase. A second approach models the selection of learning experts as an adversarial problem with a finite budget of *contaminated rewards*. We call this setting *adversarial bandits with budget*. Here, experts are learned, selected and exploited at the same time. When the budget of *contaminated rewards* is known, we propose and analyze a randomized variant of the algorithm** SUCCESSIVE ELIMINATION. **When this budget is unknown, we analyze the algorithm** EXP3 **for the proposed setting. We illustrate both approaches in the case where the learning experts are based on** BANDIT FORESTS **initialized with different sets of parameters.**

I. INTRODUCTION

The *contextual bandit problem* [1] is a repeating game where a player must select actions after observing a context. At each step, it receives a reward which depends on the played action and the associated context. The rewards of not chosen actions are never revealed. The goal is to minimize the *regret* expressed as the difference between rewards that could be acquired by an *optimal policy* knowing hidden parameters of the problem.

There are two popular approaches to deal with the contextual bandit problem. The first is based on policy selection algorithms ([2], [3], [4]). At each step, a player chooses a policy within a known set of policies. Then, the selected policy chooses the action to play. The modeling of the dependence between actions, rewards and contexts is delegated to the policies. The policy selection algorithms do not have to manage the environment. The second approach to handle the contextual bandit problem is based on the learning of a model from the feedback obtained through the game. Here, the algorithms interact directly with the environment by choosing the player action. At each step, the available information (context, played action, reward) is used to improve the model. Such approach can be performed with linear models (e.g., [5], [6]), neural networks [7] or random forest [8]. A methodology for using any batch learning algorithm in a contextual bandit setting has also been proposed by [1] with the epoch-greedy algorithm.

However, both approaches suffer from weaknesses. The policy selection algorithms have strong theoretical guarantees but in practice their performances depend of the availability of a good policy within the set of policies. Moreover, their computational time cost in $O(\text{poly(T)})$ could be an issue when the time horizon T is large. Learning a policy directly from the data could be very effective in practice and in theory but still has an approximation error due to the bias of the hypothesis space considered by the algorithm: the best linear model, the best random forest of depth D and size L, a reasonably good multi-layer perceptron with h hidden neurons...

In this paper, we consider an hybrid approach, the selection of learning experts. We call a learning expert the tuple characterizing an instance of a contextual bandit algorithm (parameters, random seed, ...). At each round, a player chooses an expert and this expert selects an action after observing a context vector generated by the environment. Both receive a feedback and the expert can modify its policy with this new observation. The player has to estimate the performances of experts and to optimize their selection. Each expert learns in order to minimize the estimation error within its hypothesis space. The player minimizes the approximation error among the experts. This hybrid approach tackles the main weakness of the two previous approaches for the contextual bandit problem. Firstly, the exploration of the hypothesis space (i.e. the set of policies) is done by efficient algorithms with strong theoretical guarantees such as LINUCB [6], or BANDIT FOREST [8]. Secondly, the bias of these algorithms is reduced by selecting the expert with the highest performances. Selection of learning experts is a non-trivial problem. Indeed, as experts learn and can modify their policies over time, the optimal algorithms for stochastic multi-armed bandit problem, such as UCB [9] or Thompson Sampling (e.g., [10], [11]), cannot be used directly. Moreover, the adversarial multi-armed bandit algorithms such as EXP3 or EXP4 [12], [13] do not take advantage of the fact that the rewards of learning experts tend to increase over time.

Our contribution: Contrarily to offline learning, where the learner has beforehand a set of samples from the joint distribution of contexts and rewards $D_{x,y}$, in online learning, $D_{x,y}$ is unknown at the beginning of the optimization. As we cannot try several learning algorithms and parameters before the deployment, to control the risk of deploying models, we need worst case theoretical guarantees. We provide a theoretical framework and two selection algorithms that can be applied to the cases of bandit information (the rewards are partially observed), and full information (the rewards of not chosen actions are not observed). In the full information

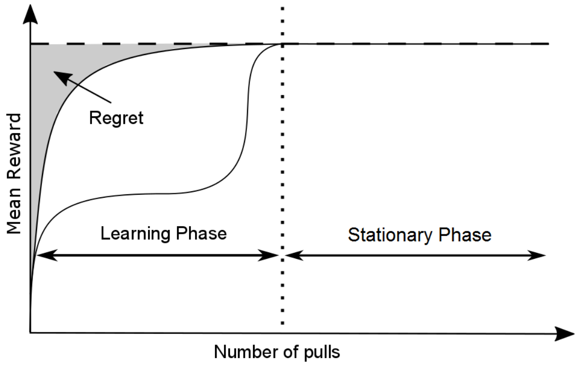

Fig. 1: Two examples of mean reward obtained by a learning algorithm over time.

setting the dependence in K of the lower and upper bounds are removed, and the vector of rewards y is fully observed by experts.

A promising approach for learning experts selection is to weaken the adversarial hypotheses and to include information about the non-stationarity of the learning phase. Based on the shape of the regret curves of contextual bandit algorithms (see Figure 1), we can make the following observations:

- during the learning phase, the mean reward of experts tend to increase over time,
- after the learning phase, the mean reward can be considered to be constant over time.

The shape of the expert learning curves depends on the algorithm and the values of parameters. The expert can quickly converge to a good solution and then take a long time to reach the optimum, or the learning curve can have a plateau for a long time and then quickly converge to the optimum (see Figure 1). Two quantities characterize the learning curves:

- The sample-complexity: the number of pulls needed to obtain the best policy within the set of experts with high probability,
- the regret accumulated over time versus the best policy within the set of experts: the area above the learning curve.

Two natural settings appear from these observations. The first is based on the PAC-setting [14]. The experts are learned, and after enough time, mean rewards of experts become stationary. Thus, after the learning phase, a stochastic MAB algorithm can be used to select the best expert. This setting is presented in Section III. Using SUCCESSIVE ELIMINATION algorithm [15] to select the best expert, we provide an analysis of the selection experts after then learning: we showed that the proposed approach is optimal up to logarithmic factors when near optimal experts such as BANDIT FOREST [8] are used. This approach takes advantages of the exploration phase, where actions are sequentially played, to share observations between experts, and thus to learn them in parallel.

The second setting, presented in Section IV uses a similar approach than *contaminated rewards* introduced by [16], where the rewards are mainly drawn from stationary distributions except for a minority of rewards of means chosen in advance by an adversary. We adapted this setting to the selection of learning experts. Here, as the performances over time of each learning expert are bounded by those of the best expert of each hypothesis space, the maximal mean of *contaminated rewards* is bounded by the mean of the best one. Moreover, to take into account that the performances of learning experts tend to increase and their regret is bounded, we bound the amount of contamination available to the adversary. The time steps of contaminated rewards and the means of their distributions are arbitrary chosen by an oblivious adversary before the beginning of the run. The other rewards are drawn from the distribution with highest mean reward within the hypothesis space of the expert. We call this setting *adversarial bandits with budget*. In Section III, we propose a randomized version of the algorithm SUCCESSIVE ELIMINATION for the case where the budget B is known. We take advantage of this randomized version to share the observations of played experts with the others: the reward used to update each expert is divided by the probability to select the action. In comparison to the first setting, the strength of this approach is to play more and more the best experts over time. The weakness of this approach is that it necessitates to know the budget of contaminated rewards. When the budget is unknown, we analyzed the adversarial bandit algorithm EXP3 for *adversarial bandits with budget*. In Section V, we illustrate experimentally our results using the BANDIT FOREST algorithm to learn the experts.

II. SELECTION OF LEARNING EXPERT PROBLEM

Let $A = \{1, ..., K\}$ be a set of actions. Let x_t be a context vector describing the environment at time t. Let y_t be a vector of rewards at time t and $y_k(t) \in [0, 1]$ the reward of the action k at time t. Let $D_{x,y}$ be a joint distribution on (x, y). Let $\pi : X \to A$ be a policy. Let $S = \{1, ..., M\}$ be the set of indexes of experts. Let Π_m be the set of policies reachable during the learning of the expert $m \in S$ and $\Pi = \bigcup_{m=0}^{M} \Pi_m$ be the set of policies. The policy used by the expert m at time t is denoted $\pi_{m,t}$. Its mean reward is defined by:

$$\mu_m(t) = \mathbb{E}_{D_{x,y}}[y_{\pi_{m,t}(x)}].$$

The optimal policy of the set Π_m is defined by:

$$\pi_m^* = \arg \max_{\pi \in \Pi_m} \mathbb{E}_{D_{x,y}}[y_{\pi(x)}].$$

Let μ_m be the mean reward of the optimal policy of the set Π_m. The optimal policy is defined by:

$$\pi^* = \arg \max_m \pi_m \, .$$

The expected regret of the learning experts selection task is defined by:

$$\mathbb{E}_{D_{x,y}}[R(T)] = \mathbb{E}_{D_{x,y}} \left[\sum_{t=1}^{T} \left(y_{\pi^*(x_t)}(t) - y_{\pi_{m_t,t}(x_t)}(t) \right) \right],$$

where m_t is the expert played at time t.

Let m^* be the index of the subset containing π^*.

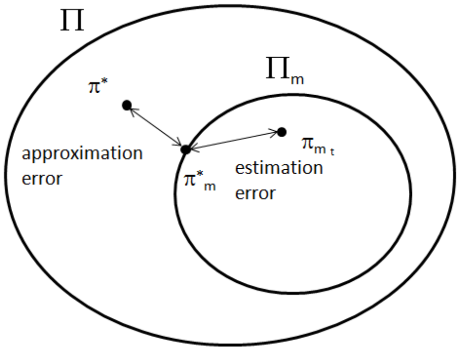

Fig. 2: The learning algorithm task minimizes the estimation error in the set Π_m. The learning experts selection task reduces the approximation error.

The decomposition of the expected regret of the expert m in terms of approximation and estimation (see Figure II) is given below:

$$\begin{aligned}\mathbb{E}_{D_{x,y}}[R_m(T)] &= \sum_{t=0}^{T} (\mu_{m^*} - \mu_m(t)) \\ &= \sum_{t=0}^{T} (\mu_{m^*} - \mu_m) + \sum_{t=0}^{T} (\mu_m - \mu_m(t)) \,.\end{aligned}$$

Notice that the expected regret of the best expert m^* has no approximation error. Now, the expected regret of an algorithm of selection of learning experts can be defined by:

$$\mathbb{E}_{D_{x,y}}[R(T)] = \sum_{t=0}^{T} (\mu_{m^*} - \mu_{m_t}) + \sum_{t=0}^{T} (\mu_{m_t} - \mu_{m_t}(t)) \ ,$$

where m_t denotes the selected expert at time t.

With contextual bandit algorithms that learn the dependency between contexts, rewards and actions, the player minimizes the estimation error. By performing a selection within a set of several contextual bandit algorithms, the learning experts selection task reduces the approximation error.

III. Selection of Learning Experts as a Learn Then Explore and Exploit approach

A first approach to handle selection of learning algorithms is to use a Learn Then Explore and Exploit approach (LTEE):

- M contextual bandit problems are allocated and solved during the learning phase,
- one multi-armed bandit problem is allocated and optimizes the choice of experts during the exploration and exploitation phases.

Learn Then Explore and Exploit

Learning Phase: In order to obtain an unbiased learning of each expert, during the learning phase the actions are played sequentially. To ensure that each expert has ended its learning, we set the size of the learning phase as the maximum sample complexity of experts. We consider here bandit algorithms, where the sample complexity is known such as BANDIT FOREST [8] or linear bandits [17].

Definition 1: The sample-complexity n_m of the expert m is the number of samples in $D_{x,y}$ needed to find π^*_m with a probability $1 - \delta$.

Proposition 1: Let $[M]$ be a set of M experts able to learn from the uniform sampling of $D_{x,y}$. After sampling n times in $D_{x,y}$, the mean reward of each expert $m \in [M]$ is μ^*_m with a probability $1 - M\delta$, where $n = \max_m n_m$.

Proof: Using the sample-complexity definition, the proof is a direct application of the union bound. □

Proposition 1 is now used to set the size of the learning phase. Then, when the experts are learned, any bandit algorithm such as UCB [9] or SUCCESSIVE ELIMINATION [15] can handle the usual trade-off between exploration and exploitation.

Exploration and Exploitation Phase: At the end of the Exploration phase, experts have learned and their performances can be considered as stationary. Now, one needs to find and play the best one. To handle the trade-off between evaluation and exploitation, we choose SUCCESSIVE ELIMINATION algorithm [15] to provide a consistent end-to-end analysis of the selection of learning expert problem.

SUCCESSIVE ELIMINATION plays each reamining expert sequentially. The empirical mean of each expert $\hat{\mu}_m(t)$ after the time-step n is maintained through the run using past observations:

$$\hat{\mu}_m(t) = \frac{1}{t-n} \sum_{i=n+1}^{t} [\![m_i = m]\!] y_{\pi_{m_i}(x_i)}(i).$$

When the gap between the empirical mean rewards of an expert m and the estimated best expert is high enough, then the expert is considered as suboptimal and removed from the set.

Algorithm 1 LTEE

```
**input:** $\delta \in (0,1]$, $\epsilon \in [0,1)$
**output:** an $\epsilon$-approximation of the best expert
**For** $t = 0, ..., n$
Play sequentially an action $k_t \in A$
Update each expert in $S$ with the tuple $(x_t, k_t, y_{k_t}(t))$
**End for**
$S_t = S$, $r = 0$
**While** $t \leq T$
 **For each** $m \in S_t$
    Play the action $k_t = \pi_m(x_t)$
    Update $\hat{\mu}_m(t+1)$
    $t = t + 1$ [1]
  **End for**
  $r = r + 1$
  $m_{\max} = \arg\max_{m \in S_t} \hat{\mu}_m(t)$
  Remove from $S_t$ all $m$ such as:
        $\hat{\mu}_{\max}(t) - \hat{\mu}_m(t) + \epsilon \geq 2\sqrt{\log(4r^2 M/\delta)/2r}$
**End while**
```

Theorem 1: For $M > 0$, and $\delta > 0$, and $\epsilon = 0$, the sample complexity of LTEE is upper bounded by:

$$O\left(\frac{M}{\Delta^2}\log\frac{M}{\delta\Delta} + n\right),$$

where Δ is the difference of mean rewards between the two best experts, n the sample complexity needed to learn the set of experts.

The proof of Theorem 1 is provided in appencides.

Theorem 2: There exists a distribution $D_{x,y}$ such that any algorithm that learns M experts and then finds the best has a sample-complexity of at least:

$$\Omega\left(\frac{M}{\Delta^2}\log\frac{1}{\delta} + \mathcal{N}\right),$$

where δ the probability of finding the optimal expert within the set of experts and $\mathcal{N}$ the lower-bound of the sample complexity needed to learn the experts.

The proof of Theorem 2 is provided in appencides. The LTEE algorithm is optimal up to logarithmic factors when the sample complexity needed to learn the set of experts is optimal, i.e. when $n = \mathcal{N}$.

IV. Selection of Learning Experts as an Adversarial Problem with Budget

In previous approach, the experts are learned at the beginning of the run and then selected when their performance becomes stationary. LTEE algorithm obtains strong theoretical results and can be very efficient in practice when the sample complexities needed to learn the experts are close to each other. When one algorithm needs a large sample complexity and the others do not, LTEE algorithm spends a lot of steps playing sequentially the actions to learn the last expert. In this section, we present a new algorithm Learn, Explore and Exploit (LEE), which handles the parallel learning of experts, and which explores and exploits at the same time the experts. The exploration and exploitation of learning experts is modeled by a new problem setting, which we called *adversarial bandit with budget*. We describe this setting in the next section. We will analyze EXP3 for this setting. Then, we will explain the methodology used to learn experts in parallel and we propose a new algorithm, Randomized Successive Elimination with Budget, based on a randomized version of Successive Elimination. Finally, we will present LEE, which uses Randomized Successive Elimination with Budget and an unbiased estimation of rewards to simultaneously learn, explore, and exploit experts.

Setting of Adversarial Bandit with Budget

Let $S = 1, ..., M$ be a set of experts. The reward $y_m(t) \in [0, 1]$, obtained by the player after selecting the expert m, is drawn from a distribution $D_m(t)$ of mean $\mu_m(t) \in [0, \mu_m]$, with $\mu_m \in [0, 1]$. The values of $\mu_m(t)$ are arbitrary chosen by an oblivious adversary before the beginning of the run. The adversary is constrained by a budget B such as $\forall m$, $\sum_{t=1}^{T} \mu_m - \mu_m(t) \leq B$. We abstract the learning experts selection problem by considering steps where $\mu_m(t) \neq \mu_m$ as chosen by an adversary with a fixed budget. This budget constraint is the main innovation compared to *contaminated rewards* proposed in [16]. Notice that unlike *contaminated rewards* this setting models different behaviors of learning curves: an high amount of the contaminated budget can be spent in a short period to model quick learner, or the budget can be spent equally over iterations to model slow learner.

EXP3 for Adversarial Bandit with Budget

When the knowledge of the budget B is not available and the time horizon T is known, EXP3 can be successfully applied to *adversarial bandit with budget* (see Theorem 3).

Theorem 3: For any $M > 0$, the expected regret of EXP3 for learning experts selection run with input parameter $\gamma = \sqrt{\frac{M\log(M)}{(e-1)T}}$ is:

$$\mathbb{E}\left[R(T)\right] \leq 2\sqrt{(e-1)MT\log(M)} + B\,, \tag{1}$$

where $\mathbb{E}$ denotes the expectation with respect to the joint distribution of experts $(m_1, ..., m_T)$ and rewards $(y_1, ..., y_T)$.

The proof of Theorem 3 is provided in appendices.

Unbiased Learning of Experts

In order to update each expert m in parallel with any tuple $(x_t, k_t, y_{k_t}(t))$, where k_t is chosen by another randomized policy, the observed reward $y_{k_t}(t)$ is replaced by an unbiased estimator of y_{k_t}:

$$r_{k_t}(t) = y_{k_t}(t)/P(k_t)(t).$$

Indeed, at any time t we have:

$$\mathbb{E}[r_{k_t}] = P(k_t).y_{k_t}/P(k_t) = y_{k_t}.$$

This approach, called *Inverse Propensity Scoring* [18], can be used to evaluate the gain of policies only when no action has a zero probability (see Theorem 1 in [19]). Lemma 1 shows that using LEE (see Algorithm 3), the probability of any remaining actions for any experts cannot be equal to zero.

Lemma 1: When LEE algorithm uses randomized experts, which draw actions from their set of remaining actions $A_{m,t}$ according to an uniform distribution, the probability to draw any remaining action k is $P(k) \geq \frac{1}{MK}$.

The proof of Lemma 1 is provided in appendices.

Randomized Successive Elimination with Budget

In this section, we present an algorithm for the best expert identification problem which is able to perform on distributions of large support and reward contaminated by an adversary with a budget B (see Algorithm 2). This algorithm serves two purposes, firstly as an algorithm for selecting learning experts and secondly as an elementary block to build Randomized Bandit Forests.

[1] For the sake of clarity, the condition $t \leq T$ is tested only before the round-robin phase. To ensure the end of the algorithm at T, another test is necessary at this mark.

We introduce three significant differences: choices of experts are randomized, rewards are unbiased by the probability of playing the expert and the contaminated budget is taken into account when eliminating experts. Notice that this formulation is equivalent to a full information problem where the played expert has a reward of $|S_t|.y_m(t)$ and other experts a reward of zero.

Algorithm 2 SUCCESSIVE ELIMINATION WITH BUDGET

```
input: $\delta \in (0,1]$, $\epsilon \in [0,1)$, $B \geq 0$
output: an $\epsilon$-approximation of the best expert
$S_1 = S$, $\forall m$, $\hat{\mu}_m(0) = 0$, $Z(0) = |S|^2$, $t = 1$
While $|S_t| > 1$
  $Z(t) = Z(t-1) + |S_t|^2$
  Sample an expert $m_t \sim 1/|S_t|$
  For each $m \in S_t$ do
    $\hat{\mu}_m(t) = \frac{t-1}{t}\hat{\mu}_m(t-1) + \frac{|S_t|.y_m(t)}{t}[\![m = m_t]\!]$
  End for
  $m_{\max} = \arg\max_{m \in S_t} \hat{\mu}_m(t)$
  Remove from $S_t$ all $m$ such as:
  $$\hat{\mu}_{\max}(t) - \hat{\mu}_m(t) + \epsilon \geq B/t + 2\sqrt{\frac{Z(t)}{2t^2}\log\left(\frac{4Mt^2}{\delta}\right)}$$
  $t = t + 1$
End while
```

When $B = 0$, SUCCESSIVE ELIMINATION WITH BUDGET can be applied to best action identification problem.

Theorem 4: For $M > 0$, and $\delta > 0$, and $\epsilon = 0$, the sample-complexity of RANDOMIZED SUCCESSIVE ELIMINATION WITH BUDGET needed to find the best expert is upper bounded by:

$$O\left(\frac{M^2}{\Delta^2}\left(\log(\frac{M}{\delta\Delta}) + B\right)\right).$$

where Δ is the difference of mean rewards between the two best experts.

Corollary 1: For $M > 0$, $\delta > 0$, and $\epsilon = 0$, the expected regret of RANDOMIZED SUCCESSIVE ELIMINATION WITH BUDGET is upper bounded by:

$$O\left(\frac{M^2}{\Delta}\left(\log(\frac{MT}{\Delta}) + B\right)\right).$$

The proofs of Theorem 4 and Corollary 1 are provided respectively in appendices.

This upper-bound, including a factor M^2, may seem high but is unavoidable when dealing with rewards debiased by the probability of playing the experts. We provide a lower-bound on the best expert identification problem with full information and rewards distribution using a support of $[0, M]$, instead of $[0, 1]$ for the case where $B = 0$.

Theorem 5: For $B = 0$ and $\epsilon = 0$, there exists a distribution D_Y such that any algorithm that observes at each step all experts' rewards $0 \leq Y_m(t) \leq M$ and then finds the best one has a sample-complexity of at least:

$$\Omega\left(\frac{M^2}{\Delta^2}\log\frac{1}{\delta}\right),$$

where Δ is the difference of mean rewards between the two best experts.

The proof of Theorem 5 is provided in appendices. This lower-bound shows the optimality of RANDOMIZED SUCCESSIVE ELIMINATION WITH BUDGET up to a logarithmic factor when the support of the reward distribution is $[0, M]$ for $B = 0$.

Algorithm 3 LEARN, EXPLORE AND EXPLOIT (LEE)

```
input: $\delta \in (0,1]$, $\epsilon \in [0,1)$, $B \geq 0$
output: an $\epsilon$-approximation of the best expert
$S_1 = S$, $\forall m$, $\hat{\mu}_m(0) = 0$, $Z(0) = |S|^2$
For $t = 1, ..., T$
  $Z(t) = Z(t-1) + |S_t|^2$
  Sample an expert $m_t \sim 1/|S_t|$
  Sample an action $k_t \sim A_{m,t}$
  $P_{k_t} = 0$
  For each $m \in S_t$ do
    $P_{k_t} = P_{k_t} + \frac{1}{|A_{m,t}|.|S_t|}$
  End for
  For each $m \in S_t$ do
    $\hat{\mu}_m(t) = \frac{t-1}{t}\hat{\mu}_m(t-1) + \frac{|S_t|.y_{k_t}(t)}{t}[\![m = m_t]\!]$
    Update the expert $m$ with the tuple $\left(x_t, k_t, \frac{y_{k_t}(t)}{P_{k_t}}\right)$
  End for
  $m_{\max} = \arg\max_{m \in S_t} \hat{\mu}_m(t)$
  Remove from $S_t$ all $m$ such as:
  $$\hat{\mu}_{\max}(t) - \hat{\mu}_m(t) + \epsilon \geq B/t + 2\sqrt{\frac{Z(t)}{2t^2}\log\left(\frac{4Mt^2}{\delta}\right)}.$$
End for
```

V. APPLICATION OF THE SELECTION OF LEARNING EXPERTS METHODOLOGY TO BANDIT FORESTS

In this section, we illustrate the application of the LTEE and the LEE approaches, using BANDIT FORESTS [8] as experts and illustrate their efficiency with numerical simulations. Due to length constraint, we only provide a short insight.

Parametrization of the experts' learning algorithms have an high incidence on the final experts' performances. For instance, on the *Forest Cover Type* dataset from the *UCI Machine Learning Repository*, a poorly parametrized expert may achieve a final classification rate of 45% whereas the expert with the most effective parametrization within the pool achieves a classification rate of 65.4%. As Figure 3 shows, if the parameters n or B are underestimated, LEE and LTEE do not find the best expert in the set. Indeed, with $n = 0$ or $B = 0$, LTEE and LEE are equivalent to a stationary multi-armed bandit algorithm.

On a wide range of parameter values, LTEE and LEE find the best parametrization and achieve the classification rate of 65.4%, the same than the optimal forest, but at the cost of a slightly higher cumulative regret. In contrast, EXP3.S also finds the best parametrization but suffers from the constant exploration of the others experts inherent to this algorithm. While an underestimation of n or B can prevent the convergence of the algorithms to the best expert, an overestimation has a low impact on the cumulative regret. The practical use of LTEE and LEE, for which the optimal values of n and B are unknown, makes appear a trade-off between the better performances of LTEE at its point of best parametrization and the low sensitivity of LEE to an overestimation of B. Indeed, LTEE outperforms LEE only on a small range of values and shows an high degradation of performances when the parameter n is overestimated (LTEE is outperformed by EXP3.S when $n = 2 \times 10^6$) .

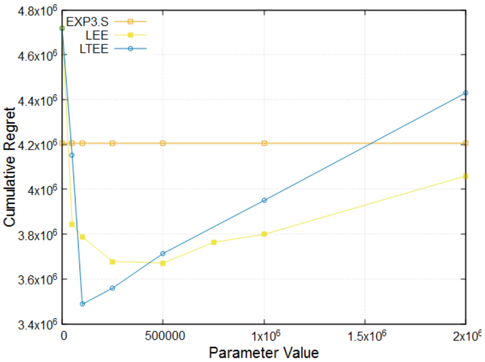

Fig. 3: The cumulative regrets of EXP3.S, LTEE and LEE initialized with different parameters (respectively n and B) on Forest Cover Type.

TABLE I: Forest Cover Type dataset where each continuous variable is recoded using *equal frequencies* into 5 binary variables, and each categorical variable is recoded into disjunctive binary variables. The 7 targeted classes are used as the set of actions. Cumulative regrets are given for $n = 100000$ and $B = 500000$. Classification rates are computed on the 100000 last contexts.

Algorithm	Regret	Classification Rate
Forest Cover Type, Target: Cover Type (7 classes)		
LTEE	$3.56\ 10^6 \pm 10^6$	65.4%
LEE	$3.68\ 10^6 \pm 2.10^6$	65.4%
EXP3.S	$4.21\ 10^6 \pm 5.10^5$	62.9%
OPTIMAL FOREST	$3.55\ 10^6 \pm 5.10^4$	65.4%

VI. Conclusion

In this paper, we reduced the selection of learning experts to stochastic MAB problems. Taking advantage of this reduction, we proposed the LTEE algorithm. We showed that the proposed algorithm is optimal up to logarithmic factors. Despite this strong theoretical result, when the sample complexities of experts are not close to each other, in practice LTEE can spend a lot of steps playing sequentially the actions. We proposed a second algorithm LEE based on an unbiased estimation of rewards to share observations between experts and on a new problem setting, which we called *adversarial with budget*. We proposed a randomized variant of SUCCESSIVE ELIMINATION taking advantage of the proposed setting. Even if the obtained upper bound of the sample complexity is not as tight as the one obtained by LTEE, we experimentally demonstrated the efficiency of LEE. Further works may involve the analysis of the lower-bound of the *adversary bandit problem with budget*. An adaptation of these algorithms to handle switches in the distribution generating contexts and rewards also could be considered.

References

[1] J. Langford and T. Zhang, “The epoch-greedy algorithm for contextual multi-armed bandits,” in *NIPS*, 2007.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[3] M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang, “Efficient optimal learning for contextual bandits,” *CoRR*, 2011.

[4] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire, “Taming the monster: A fast and simple algorithm for contextual bandits,” in *The 31st International Conference on Machine Learning (ICML 2014)*, 2014.

[5] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, “Efficient bandit algorithms for online multiclass prediction,” in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML, 2008.

[6] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” *CoRR*, 2010.

[7] R. Allesiardo, R. Féraud, and D. Bouneffouf, “A neural networks committee for the contextual bandit problem,” in *ICONIP*, 2014, pp. 374–381.

[8] R. Féraud, R. Allesiardo, T. Urvoy, and F. Clérot, “Random forest for the contextual bandit problem,” *AISTATS*, 2016.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[10] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, June 2012.

[11] E. Kaufmann, N. Korda, and R. Munos, “Thompson sampling: An asymptotically optimal finite-time analysis,” in *Algorithmic Learning Theory*, 2012, pp. 199–213.

[12] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The nonstochastic multiarmed bandit problem,” *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[13] R. Allesiardo and R. Féraud, “Exp3 with drift detection for the switching bandit problem,” in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)*, 2015.

[14] L. Vailant, “A theory of learnable,” *Communications of the ACM*, 1984.

[15] E. Even-Dar, S. Mannor, and Y. Mansour, “Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems.” *Journal of Machine Learning Research*, vol. 7, 2006.

[16] Y. Seldin and A. Slivkins, “One practical algorithm for both stochastic and adversarial bandits,” in *31th Intl. Conf. on Machine Learning (ICML)*, 2014.

[17] M. Soare, A. Lazaric, and R. Munos, “Best-arm identification in linear bandits,” in *NIPS*, 2014.

[18] D. G. Horvitz and D. J. Thompson, “A generalization of sampling without replacement from a finite universe,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[19] J. Langford and A. Strehl, “Exploration scavenging,” in *ICML*, 2008.

[20] S. Mannor, J. N. Tsitsiklis, K. Bennett, and N. Cesa-bianchi, “The sample complexity of exploration in the multi-armed bandit problem,” *Journal of Machine Learning Research*, vol. 5, p. 2004, 2004.

[21] E. V. SSlud, “Distribution inequalities for the binomial law,” *Ann. Probab.*, 1977.

[22] N. Mousavi, “How tight is chernoff bound ?” *https://ece.uwaterloo.ca/ nmousavi/Papers/Chernoff-Tightness.pdf*, 2010.

[23] J. T. Chu, “On bounds for the normal integral,” *Biometrika 42*, 1955.

APPENDIX

Proof of Theorem 1

Proof: The sample-complexity of LTEE is upper bounded by the sum of the sample complexity of SUCCESSIVE ELIMINATION (see Theorem 8 in [15]) and n (see Proposition 1). □

Proof of Theorem 2

Proof: The lower bound of the sample-complexity is the sum of the lower bound of best arm identification problem (see Theorem 1 in [20]) and the lower bound $\mathcal{N}$ of learning experts. Indeed, in worst case all the experts end their learning phase at the same time, and no observation coming from the learning phase can be used to evaluate the performances of experts. □

Proof of Theorem 3

Proof: The proof is trivial by adding B to both sides of the upper-bound on the cumulative regret of EXP3 provided in [12]:

$$\mathbb{E}\left[R(T) - B\right] + B \leq 2\sqrt{(e-1)MT\log(M)} + B \,.$$

□

Proof of Lemma 1

Proof: Let S_t be the set of remaining experts at time t and $A_{m,t}$ be the set of remaining actions of the expert m. Let A_t bet the set of remaining actions, $A_t = \bigcup_m A_{m,t}$. In worst case, there exists an action $k \in A_t$ which is played by only one expert. Each expert has a probability $1/|S_t|$ to be drawn, which does not depend on the action. The probability that the action k be drawn is:

$$P(k) = \sum_m P(k|\pi_m).P(\pi_m) = \frac{1}{MK}.\sum_m \frac{K}{|A_{m,t}|} \geq \frac{1}{MK}$$

□

Proof of Theorem 4

Proof: From Hoeffding's inequality, at time t for any m we have:

$$P\left(|\hat{\mu}_m - \mathbb{E}[\hat{\mu}_m]| \geq \epsilon_t\right) \leq 2\exp\left(-\frac{2\epsilon_t^2 t^2}{\sum_{i=1}^t |S_t|^2}\right)\,,$$

where $\mathbb{E}$ denotes the expectation with respect to the joint distribution $D_{y,m}$, where $p_m(t) = 1/|S_t$.

We upper-bound $\sum_{i=1}^t |S_t|^2$ by tM^2.

By setting $\epsilon_t = \sqrt{\frac{M^2}{2t}\log\left(\frac{4Mt^2}{\delta}\right)}$, we have:

$$P\left(|\hat{\mu}_m - \mathbb{E}[\hat{\mu}_m]| \geq \epsilon_t\right) \leq 2\exp\left(\frac{-2\sqrt{\frac{M^2}{2t}\log\left(\frac{4Mt^2}{\delta}\right)}^2 t}{M^2}\right) = \frac{\delta}{2Mt^2}\,. \quad (2)$$

Using Hoeffding's inequality on each time-step t, applying the union bound and then $\sum 1/t^2 = \pi^2/6$, the following inequality holds for any time t with a probability at least $1 - \frac{\delta\pi^2}{12M}$:

$$\hat{\mu}_m - \epsilon_t \leq \mathbb{E}[\hat{\mu}_m] \leq \hat{\mu}_m + \epsilon_t\,. \quad (3)$$

An expert m' remains in the set S as long as for each $m \in S - \{m'\}$:

$$\hat{\mu}_m - \epsilon_t < \hat{\mu}_{m'} + \frac{B}{t} + \epsilon_t\,. \quad (4)$$

Using (3) in (4):

$$\mathbb{E}[\hat{\mu}_m] - 2\epsilon_t < \mathbb{E}[\hat{\mu}_{m'}] + \frac{B}{t} + 2\epsilon_t\,. \quad (5)$$

For each m we have:

$$\hat{\mu}_m = \frac{1}{t}\sum_{i=0}^t \frac{y_m(i)}{p_m(i)} [\![m = m_i]\!]\,,\ \text{where } p_m(i) = \frac{1}{|S_t|}.$$

Taking the expectation with respect to the reward distribution D_y we have:

$$\mathbb{E}_{D_y}[\hat{\mu}_m] = \frac{1}{t}\sum_{i=0}^t \frac{\mu_m - \mu_m + \mu_m(i)}{p_m(i)} [\![m = m_i]\!]\,,$$

$$\mathbb{E}_{D_y}[\hat{\mu}_m] = \frac{1}{t}\sum_{i=0}^t \left(\frac{\mu_m}{p_m(i)} - \frac{\mu_m - \mu_m(i)}{p_m(i)}\right) [\![m = m_i]\!]\,. \quad (6)$$

Taking the expectation of both sides of equation (6) with respect to the distribution of experts $\langle m_1, ..., m_T\rangle$, with $p_m(t) = 1/|S_t|$ and using $B \geq 0$ we have:

$$\mu_m - \frac{B}{t} \leq \mathbb{E}[\hat{\mu}_m] \leq \mu_m\,. \quad (7)$$

If (5) holds for m and m' by using (7) :

$$\mu_m - 2\epsilon_t - \frac{B}{t} \leq \mu_{m'} + 2\epsilon_t + \frac{B}{t}\,.$$

$$\Delta_{m,m'} < 4\epsilon_t + \frac{2B}{t}\,. \quad (8)$$

Replacing the value of ϵ_t in (8) and taking the square:

$$\Delta^2_{m,m'} < \frac{8M^2}{t}\log(\frac{4Mt^2}{\delta}) + \frac{4B^2}{t^2}\,. \quad (9)$$

For $m' = m^*$ and $m \neq m^*$ (9) is always true, involving that the optimal expert will always remain in the set with high probability for any t. We assume $t \geq B$.

$$\Delta^2_{m,m'} < \frac{8M^2}{t}\log(\frac{4Mt^2}{\delta}) + \frac{4B}{t}\,. \quad (10)$$

The expert m' is or has been eliminated if inequality (10) is false. Let $\Delta = \min_{i \neq m^*} \Delta_{m^*,i}$ and

$$t_1^* = \frac{64^2}{\Delta^2}M^2\log\left(\frac{16M}{\delta\Delta}\right)\,.$$

We denote

$$C_1(t) = \frac{8M^2}{t}\log(\frac{4Mt^2}{\delta})\,.$$

For $t = t_1^*$,

$$C_1(t_1^*) = \frac{8\Delta^2}{64^2 \log \frac{16M}{\delta\Delta}} \left(\log \frac{4M}{\delta} + 4 \log \frac{64M}{\Delta} + 2 \log\log \frac{16M}{\delta\Delta} \right),$$

$$C_1(t_1^*) \leq \frac{8\Delta^2}{64^2 \log \frac{16M}{\delta\Delta}} \left(8 \log \frac{16M}{\delta\Delta} + 24 \log 2 + 2 \log\log \frac{16M}{\delta\Delta} \right).$$

For $X > 8$ we have

$$24 \log 2 + 2 \log\log X < 8 \log X \,.$$

Hence, we have:

$$C_1(t_1^*) \leq \frac{8\Delta^2}{64^2 \log \frac{16M}{\delta\Delta}} \left(16 \log \frac{16M}{\delta\Delta} \right),$$

$$C_1(t_1^*) \leq \frac{\Delta^2}{512} \,. \tag{11}$$

As $C_1(t_1^*)$ is strictly decreasing with regard to t, (11) is true for all $t > t_1^*$. When $t > t_1^*$, it exists $C_2(t)$ such as:

$$\Delta^2 = C_1(t) + C_2(t) \,.$$

For invalidating (10) we need to find a value of $t_2^* > t_1^*$ for which:

$$t \geq \frac{4B}{C_2(t)} \,.$$

As $C_2(t) = \Delta^2 - C_1(t)$ we have $C_2(t) \geq \Delta^2 - \frac{\Delta^2}{512}$,

$$t \geq \frac{2048B}{511\Delta^2} \,. \tag{12}$$

For $t = t_2^*$ with

$$t_2^* = \frac{64^2}{\Delta^2} M^2 \log \left(\frac{16M}{\delta\Delta} \right) + \frac{5B}{\Delta^2} \,.$$

(12) is true, invalidating (10) and involving the elimination of all suboptimal experts with a probability at least $1 - \delta$, concluding the proof. □

Proof of Corollary 1

Proof: Let t^* be the number of time steps needed to find with high probability the best expert. The expected cumulative regret at time T is:

$$\mathbb{E}_{D_{x,y}}[R(T)] = \mathbb{E}_{D_{x,y}} \left[\sum_{t=1}^{t^*} r(t) + \sum_{t=t^*+1}^{T} r(t) \right].$$

Then, for each time step of the exploration phase (i.e. $t \leq t^*$), we bound the expected instantaneous regret $\mathbb{E}_{D_{x,y}}[r(t)]$ by $\Delta = \min_{m \neq m^*}(\mu_{m^*} - \mu_m)$, we obtain:

$$\begin{aligned}\mathbb{E}_{D_{x,y}}[R(T)] &\leq t^*.\Delta + (T - t^*)\mathbb{E}_{D_{x,y}}[r(t)] \\ &\leq t^*.\Delta + (T - t^*)\left[1.\mathbb{P}(k_t \neq k_t^*) + 0.\mathbb{P}(k_t = k_t^*)\right].\end{aligned}$$

From the union bound, we have:

$$\begin{aligned}\mathbb{E}_{D_{x,y}}[R(T)] &\leq t^*.\Delta + (T - t^*)\left[\mathbb{P}(i \neq i^*) + \mathbb{P}(k \neq k^*)\right] \\ &\leq t^*.\Delta + 2(T - t^*)\delta \,.\end{aligned}$$

The sample complexity t^* is given by Theorem 4. If we choose $\delta = \frac{1}{T}$, we obtain:

$$\begin{aligned}&\mathbb{E}_{D_{x,y}}[R(T)] \\ &\qquad \leq t^*.\Delta + 2\frac{T - t^*}{T} \\ &\qquad \leq O\left(\frac{M^2}{\Delta} \left(\log(\frac{MT}{\Delta}) + B \right) \right).\end{aligned}$$

□

Proof of Theorem 5

Proof: Let $0 \leq Y_i(t) \leq M$ be a bounded random variable corresponding to the rewards of the expert i. By dividing $Y_i(t)$ by M we have:

$$0 \leq \frac{Y_i(t)}{M} \leq 1 \tag{13}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{t^*} \frac{Y_i(t)}{M} \right] = \frac{\mu_i}{M} \,. \tag{14}$$

Let Θ be the sum of the binary random variables $\theta_1, ..., \theta_t, ...\theta_{t^*}$ where

$$\theta_t = \left[\!\left[\frac{Y_i(t)}{M} \geq \frac{Y_j(t)}{M} \right]\!\right]. \tag{15}$$

Let $p_{i,j}$ be the probability that the use of expert i leads to more rewards than the use of action j. We have

$$p_{i,j} = \frac{1}{2} - \Delta_{i,j}, \text{where } \Delta_{i,j} = \frac{\mu_i - \mu_j}{M} \,. \tag{16}$$

We now follow the proof of Lemma 2 in [8]. Slud's inequality [21] states that when $p \leq 1/2$ and $t_k^* \leq x \leq t_k^*.(1-p)$, we have:

$$P(\Theta \geq x) \geq P\left(Z \geq \frac{x - t_k^*.p}{\sqrt{t_k^*.p(1-p)}} \right), \tag{17}$$

where Z is a normal $\mathcal{N}(0,1)$ random variable. To choose the best expert between i and j, one needs to find the time t^* where $P(\Theta \geq t^*/2) \geq \delta$. To state the number of trials t^* needed to estimate Δ_{ij}, we recall and adapt the arguments developed in [22]. Using Slud's inequality (see equation 17), we have:

$$P(\Theta \geq t_k^*/2) \geq P\left(Z \geq \frac{t^*.\Delta_{ij}}{\sqrt{t^*.p_{ij}(1 - p_{ij})}} \right), \tag{18}$$

Then, we use the lower bound of the error function [23]:

$$P(Z \geq z) \geq 1 - \sqrt{1 - \exp\left(-\frac{z^2}{2} \right)}$$

Therefore, we have:

$$\begin{aligned}P(\Theta \geq t^*/2) &\geq 1 - \sqrt{1 - \exp\left(-\frac{t^*.\Delta_{ij}^2}{2p_{ij}(1 - p_{ij})} \right)} \\ &\geq \frac{1}{2} \exp\left(-\frac{t^*.\Delta_{ij}^2}{p_{ij}} \right)\end{aligned}$$

As $p_{ij} = 1/2 - \Delta_{ij}$, we have:

$$\log \delta = \log \frac{1}{2} - \frac{t^*.\Delta_{ij}^2}{1/2 - \Delta_{ij}} \geq \log \frac{1}{2} - 2t^*.\Delta_{ij}^2$$

Hence, we have:

$$t^* = \Omega\left(\frac{1}{\Delta_{ij}^2} \log \frac{1}{\delta}\right)$$

In the worst case, μ_m is the same for all suboptimal experts $m \neq m^*$ and $\min_{ij} \Delta_{ij} = \min_j \Delta_{i^*j} = \Delta'$. Using the fact that θ_t is know at each round for all actions and that $\Delta' = \frac{\Delta}{\mu_{m^*} - \mu_m}$ the sample complexity is lower bounded by

$$\Omega\left(\frac{M^2}{(\mu_{m^*} - \mu_m)^2} \log \frac{1}{\delta}\right) . \tag{19}$$

□