

---

# Sub-sampling for Efficient Non-Parametric Bandit Exploration

---

**Dorian Baudry**  
dorian.baudry@inria.fr

**Emilie Kaufmann**  
emilie.kaufmann@univ-lille.fr

**Odalric-Ambrym Maillard**  
odalric.maillard@inria.fr

Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9198-CRISTAL, F-59000 Lille, France

## Abstract

In this paper we propose the first multi-armed bandit algorithm based on *re-sampling* that achieves asymptotically optimal regret simultaneously for different families of arms (namely Bernoulli, Gaussian and Poisson distributions). Unlike Thompson Sampling which requires to specify a different prior to be optimal in each case, our proposal RB-SDA does not need any distribution-dependent tuning. RB-SDA belongs to the family of Sub-sampling Duelling Algorithms (SDA) which combines the *sub-sampling* idea first used by the BESA [1] and SSMC [2] algorithms with different sub-sampling schemes. In particular, RB-SDA uses *Random Block* sampling. We perform an experimental study assessing the flexibility and robustness of this promising novel approach for exploration in bandit models.

## 1 Introduction

A  $K$ -armed bandit problem is a sequential decision-making problem in which a learner sequentially samples from  $K$  unknown distributions called arms. In each round the learner chooses an arm  $A_t \in \{1, \dots, K\}$  and obtains a random reward  $X_t$  drawn from the distribution of the chosen arm, that has mean  $\mu_{A_t}$ . The learner should adjust her sequential sampling strategy  $\mathcal{A} = (A_t)_{t \in \mathbb{N}}$  (or bandit algorithm) in order to maximize the expected sum of rewards obtained after  $T$  selections. This is equivalent to minimizing the *regret*, defined as the difference between the expected total reward of an oracle strategy always selecting the arm with largest mean  $\mu_*$  and that of the algorithm:

$$\mathcal{R}_T(\mathcal{A}) = \mu_* T - \mathbb{E} \left[ \sum_{t=1}^T X_t \right] = \mathbb{E} \left[ \sum_{t=1}^T (\mu_* - \mu_{A_t}) \right].$$

An algorithm with small regret needs to balance exploration (gain information about arms that have not been sampled a lot) and exploitation (select arms that look promising based on the available information). Many approaches have been proposed to solve this exploration-exploitation dilemma (see [3] for a survey), the most popular being Upper Confidence Bounds (UCB) algorithms [4, 5, 6] and Thompson Sampling (TS) [7, 8]. TS is a randomized Bayesian algorithm that selects arms according to their posterior probability of being optimal. These algorithms enjoy logarithmic regret under some assumptions on the arms, and some of them are even *asymptotically optimal* in that they attain the smallest possible asymptotic regret given by the lower bound of Lai & Robbins [9], for some parametric families of distributions. For distributions that are continuously parameterized by their means, this lower bound states that under any uniformly efficient algorithm,

$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}_T(\mathcal{A})}{\log(T)} \geq \sum_{k: \mu_k < \mu_*} \frac{(\mu_* - \mu_k)}{\text{kl}(\mu_k, \mu_*)}, \quad (1)$$

where  $\text{kl}(\mu, \mu')$  is the Kullback-Leibler divergence between the distribution of mean  $\mu$  and that of mean  $\mu'$  in the considered family of distributions. For arms that belong to a one-parameter exponential

family (e.g. Bernoulli, Gaussian, Poisson arms)  $\text{kl-UCB}$  using an *appropriate divergence function* [6] and Thompson Sampling using an *appropriate prior distribution* [10, 11, 12] are both *asymptotically optimal* in the sense that their regret matches that prescribed by the lower bound (1), for large values of  $T$ . Yet, a major drawback of these algorithms is that their optimal tuning requires the knowledge of the families of distributions they operate on. In this paper, we overcome this issue and propose an algorithm that is simultaneously asymptotically optimal for several families of distributions.

In the past years, there has been a surge of interest in the design of non-parametric algorithms that directly use the empirical distribution of the data instead of trying to fit it in an already defined model, and are therefore good candidates to meet our goal. In [13], the authors propose the General Randomized Exploration (GRE) framework in which each arm  $k$  is assigned an index  $\hat{\mu}_{k,t}$  sampled from a distribution  $p(\mathcal{H}_{k,t})$  that depends on the history of past observed rewards for this arm  $\mathcal{H}_{k,t}$ , and the arm with largest index is selected. GRE includes Thompson Sampling (for which  $p(\mathcal{H}_{k,t})$  is the posterior distribution given a specified prior) but also allows for more general non-parametric re-sampling schemes. However, the authors of [13, 14] show that setting  $p(\mathcal{H}_{k,t})$  to be the non-parametric Bootstrap [15] leads to linear regret. They propose variants called GIRO and PHE which perturb the history by augmenting it with fake samples. History perturbation was already suggested by [16] and is also used by Reboot [17], with a slightly more complicated bootstrapping scheme. Finally, the recently proposed Non Parametric TS [18] does not use history perturbation but instead sets  $\hat{\mu}_{k,t}$  as a weighted combination of all observations in  $\mathcal{H}_{k,t}$  and the upper bound of the support, where the weights are chosen uniformly at random in the simplex of dimension  $|\mathcal{H}_{k,t}|$ .

Besides Reboot [17], which has been analyzed only for Gaussian distributions, all other algorithms have been analyzed for distributions with known bounded support, for which they are proved to have logarithmic regret. Among them, Non Parametric TS has strong optimality property as its regret is proved to match the lower bound of Burnetas and Katehakis [19] for (non-parametric) distribution that are bounded in  $[0,1]$ . In this paper, we propose the first re-sampling based algorithm that is asymptotically optimal for several classes of possibly un-bounded parametric distributions. We introduce a new family of algorithms called Sub-Sampling Duelling Algorithms, and provide a regret analysis for RB-SDA, an algorithm based on *Random Block* sub-sampling. In Theorem 3.1, we show that RB-SDA has logarithmic regret under some general conditions on the arms distributions. These conditions are in particular satisfied for Gaussian, Bernoulli and Poisson distribution, for which we further prove in Corollary 3.1.1 that RB-SDA is asymptotically optimal.

The general SDA framework that we introduce is inspired by two ideas first developed for the BESA algorithm by [1] and for the SSMC algorithm by [2]: 1) the arms pulled are chosen according to the outcome of pairwise comparison (*duels*) between arms, instead of choosing the maximum of some index computed for each arm as GRE algorithms do, and 2) the use of *sub-sampling*: the algorithm penalizes arms that have been pulled a lot by making them compete with the other arms with only a fraction of their history. More precisely, in a duel between two arms  $A$  and  $B$  selected  $n_A$  and  $n_B$  times respectively, with  $n_A < n_B$ , the empirical mean of arm  $A$  is compared to the empirical mean of a sub-sample of size  $n_A$  of the history of arm  $B$ . In BESA the duels are organized in a tournament and only the winner is sampled, while SSMC uses rounds of  $K-1$  duels between an arm called *leader* and all other arms. Then the leader is pulled only if it wins all the duels, otherwise all the winning *challengers* are pulled. Second difference is that in BESA the sub-sample of the leader's history is obtained with *Sampling Without Replacement*, whereas SSMC selects this sub-sample as the block of consecutive observations with smallest empirical mean. Hence BESA uses randomization while SSMC does not. Finally, SSMC also uses some forced exploration (i.e. selects any arm drawn less than  $\sqrt{\log r}$  times in round  $r$ ). In SDA, we propose to combine the round structure for the duels used by SSMC with the use of a sub-sampling scheme assumed to be independent of the observations in the history (this generalizes the BESA duels), and we get rid of the use of forced exploration.

The rest of the paper is structured as follows. In Section 2 we introduce the SDA framework and present different instances that correspond to the choice of different sub-sampling algorithms, in particular RB-SDA. In Section 3 we present upper bounds on the regret of RB-SDA, showing in particular that the algorithm is asymptotically optimal for different exponential families. We sketch the proof of Theorem 3.1 in Section 4, highlighting two important tools: First, a new concentration lemma for random sub-samples (Lemma 4.2). Second, an upper bound on the probability that the optimal arm is under-sampled, which decouples the properties of the sub-sampling algorithm used, and that of the arms' distributions (Lemma 4.3). Finally, Section 5 presents the results of an empirical study comparing several instances of SDA to asymptotically optimal parametric algorithms and other

algorithms based on re-sampling or sub-sampling. These experiments reveal the robustness of the SDA approaches, which match the performance of Thompson Sampling, without exploiting the knowledge of the distribution.

## 2 Sub-sampling Duelling Algorithms

In this section, we introduce the notion of Sub-sampling Duelling Algorithm (SDA). We first introduce a few notation. For every integer  $n$ , we let  $[n] = \{1, \dots, n\}$ . We denote by  $(Y_{k,s})_{s \in \mathbb{N}}$  the i.i.d. sequence of successive rewards from arm  $k$ , that are i.i.d. under a distribution  $\nu_k$  with mean  $\mu_k$ . For every finite subset  $\mathcal{S}$  of  $\mathbb{N}$ , we denote by  $\hat{Y}_{k,\mathcal{S}}$  the empirical mean of the observations of arm  $k$  indexed by  $\mathcal{S}$ : if  $|\mathcal{S}| > 1$ ,  $\hat{Y}_{k,\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} Y_{k,i}$ . We also let  $\hat{Y}_{k,n}$  as a shorthand notation for  $\hat{Y}_{k,[n]}$ .

**A round-based algorithm** Unlike index policies, a SDA algorithm relies on *rounds*, in which several arms can be played (at most once). In each round  $r$  the learner selects a subset of arms  $\mathcal{A}_r = \{k_1, \dots, k_{i_r}\} \subseteq \{1, \dots, K\}$ , and receives the rewards  $\mathcal{X}_r = \{Y_{k_1, N_{k_1}(r)}, \dots, Y_{k_{i_r}, N_{k_{i_r}}(r)}\}$  associated to the chosen arms, where  $N_k(r) := \sum_{s=1}^r \mathbb{1}(k \in \mathcal{A}_s)$  denotes the number of times arm  $k$  was selected up to round  $r$ . Letting  $\hat{r}_T \leq T$  be the (random) number of rounds used by algorithm  $\mathcal{A}$  before the  $T$ -th arm selection, the regret of a round-based algorithm can be upper bounded as follows:

$$\begin{aligned} \mathcal{R}_T(\mathcal{A}) &= \mathbb{E} \left[ \sum_{t=1}^T (\mu_\star - \mu_{A_t}) \right] \leq \mathbb{E} \left[ \sum_{s=1}^{\hat{r}_T} \sum_{k=1}^K (\mu_\star - \mu_k) \mathbb{1}(k \in \mathcal{A}_s) \right] \\ &\leq \mathbb{E} \left[ \sum_{s=1}^T \sum_{k=1}^K (\mu_\star - \mu_k) \mathbb{1}(k \in \mathcal{A}_s) \right] = \sum_{k=1}^K (\mu_\star - \mu_k) \mathbb{E}[N_k(T)]. \end{aligned} \quad (2)$$

Hence upper bounding  $\mathbb{E}[N_k(T)]$  for each sub-optimal arm provides a regret upper bound.

**Sub-sampling Duelling Algorithms** A SDA algorithm takes as input a *sub-sampling algorithm*  $\text{SP}(m, n, r)$  that depends on three parameters: two integers  $m \geq n$  and a round  $r$ . A call to  $\text{SP}(m, n, r)$  at round  $r$  produces a subset of  $[m]$  that has size  $n$ , modeled as a random variable that is further assumed to be independent of the rewards generated from the arms,  $(Y_{k,s})_{k \in [K], s \in \mathbb{N}^*}$ .

In the first round, a SDA algorithm selects  $\mathcal{A}_1 = [K]$  in order to initialize the history of all arms. For  $r \geq 1$ , at round  $r + 1$ , a SDA algorithm based on a sampler  $\text{SP}$ , that we refer to as  $\text{SP}$ -SDA, first computes the *leader*, defined as the arm being selected the most in the first  $r$  round:  $\ell(r) = \arg\max_k N_k(r)$ . Ties are broken in favor of the arm with the largest mean, and if several arms share this mean then the previous leader is kept or one of these arms is chosen randomly. Then the set  $\mathcal{A}_{r+1}$  is initialized to the empty set and  $K - 1$  *duels* are performed. For each "challenger" arm  $k \neq \ell(r)$ , a subset  $\hat{\mathcal{S}}_k^r$  of  $[N_{\ell(r)}(r)]$  of size  $N_k(r)$  is obtained from  $\text{SP}(N_{\ell(r)}(r), N_k(r), r)$  and arm  $k$  wins the duels if its empirical mean is larger than the empirical mean of the sub-sampled history of the leader. That is

$$\hat{Y}_{k, N_k(r)} > \hat{Y}_{\ell(r), \hat{\mathcal{S}}_k^r} \implies \mathcal{A}_{r+1} = \mathcal{A}_{r+1} \cup \{k\}.$$

If the leader wins all the duels, that is if  $\mathcal{A}_{r+1}$  is still empty after the  $K - 1$  duels, we set  $\mathcal{A}_{r+1} = \{\ell(r)\}$ . Arms in  $\mathcal{A}_{r+1}$  are then selected by the learner in a random order and are pulled if the total budget of pulls remains smaller than  $T$ . The pseudo-code of  $\text{SP}$ -SDA is given in Algorithm 1.

To properly define the random variable  $\hat{\mathcal{S}}_k^r$  used in the algorithm, we introduce the following probabilistic modeling: for each round  $r$ , each arm  $k$ , we define a family  $(S_k^r(m, n))_{m \geq n}$  of independent random variables such that  $S_k^r(m, n) \sim \text{SP}(m, n, r)$ . In words,  $S_k^r(m, n)$  is the subset of the leader history used should arm  $k$  be a challenger drawn  $n$  times up to round  $r$  dueling against a leader that has been drawn  $m$  times. With this notation, for each arm  $k \neq \ell(r)$  one has  $\hat{\mathcal{S}}_k^r = S_k^r(N_{\ell(r)}(r), N_k(r), r)$ . We recall that in the SDA framework, it is crucial that those random variables are independent from the reward streams  $(Y_{k,s})$  of all arms  $k$ . We call such sub-sampling algorithms *independent sampler*.

**Particular instances** We now present a few sub-sampling algorithms that we believe are interesting to use within the SDA framework. Intuitively, these algorithms should ensure enough *diversity* in the output subsets when called in different rounds, so that the leader cannot always look good,

---

**Algorithm 1** SP-SDA

---

**Require:**  $K$  arms, horizon  $T$ , Sampler SP

```
 $t \leftarrow K, r \leftarrow 1, \forall k, N_k \leftarrow 1, \mathcal{H}_k \leftarrow \{Y_{k,1}\}$  (Each arm is drawn once)
while  $t < T$  do
   $r \leftarrow r + 1, \mathcal{A} \leftarrow \{\}, \ell \leftarrow \text{leader}(N, \mathcal{H}, \ell)$  (Initialize the round)
  for  $k \neq \ell \in 1, \dots, K$  do
    Draw  $\hat{S}_k^r \sim \text{SP}(N_\ell, N_k, r)$  (Choice of the sub-sample of  $\ell$  used for the duel with  $k$ )
    if  $\hat{Y}_{k, N_k} > \hat{Y}_{\ell, \hat{S}_k^r}$  then
       $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$  (Duel outcome)
    end if
  end for
  if  $|\mathcal{A}| = 0$  then
     $\mathcal{A} \leftarrow \{\ell\}$ 
  end if
  if  $|\mathcal{A}| > T - t$  then
     $\mathcal{A} \leftarrow \text{choose}(\mathcal{A}, T - t)$  (Randomly selects a number of arm that does not exceed the budget)
  end if
  for  $a \in \mathcal{A}$  do
    Pull arm  $a$ , observe reward  $Y_{a, N_a+1}$ 
     $t \leftarrow t + 1, N_a \leftarrow N_a + 1, \mathcal{H}_a \leftarrow \mathcal{H}_a \cup \{Y_{a, N_a}\}$  (Update step)
  end for
end while
```

---

and challengers may win and be explored from time to time. The most intuitive candidates are random samplers like *Sampling Without Replacement* (WR) and *Random Block Sampling* (RB): the first one returns a subset of size  $n$  selected uniformly at random in  $[m]$ , while the second draws an element  $n_0$  uniformly at random in  $[m - n]$  and returns  $\{n_0 + 1, \dots, n_0 + n\}$ . But we also propose two deterministic sub-sampling: *Last Block* (LB) which returns  $\{m - n + 1, \dots, m\}$ , and *Low Discrepancy Sampling* (LDS) that is similar to RB with the first element  $n_0$  of the block at a round  $r$  defined as  $\lceil u_r(m - n) \rceil$  with  $u_r$  a predefined low discrepancy sequence [20] (Halton [21], Sobol [22]). We believe that these last two samplers may ensure enough diversity without the need for random sampling. These four variants of SDA will be compared in Section 5 in terms of empirical efficiency and numerical complexity. For RB-SDA, we provide a regret analysis in the next sections, highlighting what parts may or may not be extended to other sampling algorithms.

**Links with existing algorithms** The BESA algorithm [1] with  $K = 2$  coincides with WR-SDA. However beyond  $K > 2$ , the authors of [1] rather suggest a tournament approach, without giving a regret analysis. WR-SDA can therefore be seen as an alternative generalization of BESA beyond 2 arms, which performs much better than the tournament, as can be seen in Section 5. While the structure of SDA is close to that of SSMC [2], SSMC is not a SP-SDA algorithm, as its sub-sampling algorithm heavily relies on the rewards, and is therefore not an independent sampler. Indeed, it outputs the set  $\mathcal{S} = \{n_0 + 1, \dots, n_0 + n\}$  for which  $\hat{Y}_{\ell(r), \mathcal{S}}$  is the smallest. The philosophy of SSMC is a bit different than that of SDA: while the former tries to disadvantage the leader as much as possible, the latter only tries to make the leader use different parts of its history. Our experiments reveal that the SSMC approach seems to lead to a slightly larger regret, due to a bit more exploration in the beginning. Finally, we emphasize that alternative algorithms based on re-sampling (PHE, Reboot, Non-Parametric TS) are fundamentally different to SDA as they do not perform *sub-sampling*.

**On the use of forced exploration** In [2], SSMC additionally requires some *forced exploration*: each arm  $k$  such that  $N_k(r)$  is smaller than some value  $f_r$  is added to  $\mathcal{A}_{r+1}$ . SSMC is proved to be asymptotically optimal for exponential families provided that  $f_r = o(\log r)$  and  $\log \log r = o(f_r)$ . In the next section, we show that RB-SDA does not need forced exploration to be asymptotically optimal for Bernoulli, Gaussian and Poisson distributions. However, we show in Appendix H that adding forced exploration to RB-SDA is sufficient to prove its optimality for any exponential family.

### 3 Regret Upper Bounds for RB-SDA

In this section, we present upper bounds on the expected number of selections of each sub-optimal arm  $k$ ,  $\mathbb{E}[N_k(T)]$ , for the RB-SDA algorithm. They directly yield an upper bound on the regret via (2). To ease the presentation, we assume that there is a unique optimal arm<sup>1</sup>, and denote it by 1.

In Theorem 3.1, we first identify some conditions on the arms distribution under which RB-SDA has a regret that is provably logarithmic in  $T$ . In order to introduce these conditions, we recall the definition of the following *balance function*, first introduced by [1].  $\alpha_k(M, j)$  is equal to the probability that arm 1 loses a certain amount  $M$  of successive duels against  $M$  sub-samples from arm  $k$  that have non-overlapping support, when arm 1 has been sampled  $j$  times.

**Definition.** Letting  $\nu_{k,j}$  denote the distribution of the sum of  $j$  independent variables drawn from  $\nu_k$ , and  $F_{\nu_{k,j}}$  its corresponding CDF, the balance function of arm  $k$  is

$$\alpha_k(M, j) = \mathbb{E}_{X \sim \nu_{1,j}} \left( (1 - F_{\nu_{k,j}}(X))^M \right).$$

**Theorem 3.1** (Logarithmic Regret for RB-SDA). *If the arms distributions  $\nu_1, \dots, \nu_k$  are such that*

1. *the empirical mean of each arm  $k$  has exponential concentration given by a certain rate function  $I_k(x)$  which is continuous and satisfies  $I_k(x) = 0$  if and only if  $x = \mu_k$ :*

$$\forall x > \mu_k, \mathbb{P} \left( \hat{Y}_{k,n} \geq x \right) \leq e^{-nI_k(x)} \text{ and } \forall x < \mu_k, \mathbb{P} \left( \hat{Y}_{k,n} \leq x \right) \leq e^{-nI_k(x)},$$

2. *the balance function of each sub-optimal arm  $k$  satisfies*

$$\forall \beta \in (0, 1), \sum_{t=1}^T \sum_{j=1}^{\lfloor (\log t)^2 \rfloor} \alpha_k(\lfloor \beta t / (\log t)^2 \rfloor, j) = o(\log T).$$

Then, for all sub-optimal arm  $k$ , for all  $\varepsilon > 0$ , under RB-SDA

$$\mathbb{E}[N_k(T)] \leq \frac{1 + \varepsilon}{I_k(\mu_1)} \log(T) + o(\log T).$$

If the distributions belong to the same one-dimensional exponential family (see e.g. [6] for a presentation of some of their important properties), the Chernoff inequality tells us that the concentration condition 1. is satisfied with a rate function equal to  $I_k(x) = \text{kl}(x, \mu_k)$  where  $\text{kl}(x, y)$  is the Kullback-Leibler divergence between the distribution of mean  $x$  and the distribution of mean  $y$  in that exponential family. In Appendix G, we prove that Gaussian distribution with known variance, Bernoulli and Poisson distribution also satisfy the balance condition 2., which yields the following.

**Corollary 3.1.1.** *Assume that the distribution of all arms belong to the family of Gaussian distributions with a known variance, Bernoulli or Poisson distributions. Then under RB-SDA for all  $\varepsilon > 0$ , for all sub-optimal arm  $k$ ,*

$$\mathbb{E}[N_k(T)] \leq \frac{1 + \varepsilon}{\text{kl}(\mu_k, \mu_1)} \log(T) + o_{\varepsilon, \mu}(\log(T)).$$

Corollary 3.1.1 permits to prove that  $\limsup_{T \rightarrow \infty} \frac{\mathcal{R}_T(\text{RB-SDA})}{\log(T)} \leq \sum_{k=2}^K \frac{(\mu_1 - \mu_k)}{\text{kl}(\mu_k, \mu_1)}$ , which is matching the Lai & Robbins lower bound (1) in each of these exponential families. In particular, RB-SDA is simultaneously asymptotically optimal for different examples of bounded (Bernoulli) and un-bounded (Poisson, Gaussian) distributions. In contrast, Non-Parametric TS is asymptotically optimal for any bounded distributions, but cannot be used for Gaussian or Poisson distributions. Note that the guarantees of Corollary 3.1.1 also hold for the SSMC algorithm [2], but we prove that RB-SDA can be asymptotically optimal *without forced exploration* for some distributions. Moreover, as will be seen in Section 5, algorithms based on randomized history-independent sub-sampling such as RB-SDA tend to perform better than deterministic algorithms such as SSMC.

Theorem 3.1 also shows that RB-SDA may have logarithmic regret for a wider range of distributions. For example, we conjecture that a truncated Gaussian distribution also satisfy the balance condition

<sup>1</sup>as can be seen in the analysis of SSMC [2], treating the general case only requires some additional notation.

2.. On the other hand, condition 2. does not hold for Exponential distribution, as discussed in Appendix G.4. But we show in Appendix H.2 that any distribution that belongs to a one-dimensional exponential family satisfies a slightly modified version of this condition, which permits to establish the asymptotic optimality of a variant of RB-SDA using forced exploration.

Finally, we note that it is possible to build on RB-SDA to propose a bandit algorithm that has logarithmic regret for any distribution that is bounded in  $[0, 1]$ . To do so, we can use the binarization trick already proposed by [11] for Thompson Sampling, and run RB-SDA on top of a binarized history  $\mathcal{H}'_k$  for each arm  $k$  in which a reward  $Y_{k,s}$  is replaced by a binary pseudo-reward is  $Y'_{k,s}$  generated from a Bernoulli distribution with mean  $Y_{k,s}$ . The resulting algorithm inherits the regret guarantees of RB-SDA applied to Bernoulli distributions.

Characterizing the set of distributions for which the vanilla RB-SDA algorithm has logarithmic regret (without forced exploration or a binarization trick) is left as an interesting future work.

## 4 Sketch of Proof

In this section, we provide elements of proof for Theorem 3.1, postponing the proof of some lemmas to the appendix. The first step is given by the following lemma, which is proved in Appendix D.

**Lemma 4.1.** *Under condition 1., for any SP-SSDA algorithm (using an independent sampler), for every  $\varepsilon > 0$ , there exists a constant  $C_k(\boldsymbol{\nu}, \varepsilon)$  with  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_k)$  such that*

$$\mathbb{E}[N_k(T)] \leq \frac{1 + \varepsilon}{I_1(\mu_k)} \log(T) + 32 \sum_{r=1}^T \mathbb{P}(N_1(r) \leq (\log(r))^2) + C_k(\boldsymbol{\nu}, \varepsilon).$$

The proof of this result follows essentially the same decomposition as the one proposed by [2] for the analysis of SSMC. However, it departs from this analysis in two significant ways. First, instead of using properties of forced exploration (that is absent in RB-SDA), we distinguish whether or not arm 1 has been selected a lot, which yields the middle term in the upper bound. Then, the argument relies on a new concentration result for sub-samples averages, that we state below. Lemma 4.2, proved in Appendix C, crucially exploits the fact that a SP-SDA algorithm is based on an independent sampler. Using condition 1. allows to further upper bound the right-hand side of the two inequalities in Lemma 4.2 by terms that decay exponentially and contribute to the constant  $C_k(\boldsymbol{\nu}, \varepsilon)$ .

**Lemma 4.2** (concentration of a sub-sample). *For all  $(a, b)$  such that  $\mu_a < \mu_b$ , for all  $\xi \in (\mu_a, \mu_b)$  and  $n_0 \in \mathbb{N}$ , under any instance of SP-SDA using an independent sampler, it holds that*

$$\begin{aligned} \sum_{s=1}^r \mathbb{P}\left(\hat{Y}_{a, N_a(s)} \geq \hat{Y}_{b, \hat{S}_b^s(N_b(s), N_a(s))}, N_b(s) \geq N_a(s), N_a(s) \geq n_0\right) &\leq \sum_{j=n_0}^r \mathbb{P}\left(\hat{Y}_{a, j} \geq \xi\right) + r \sum_{j=n_0}^r \mathbb{P}(Y_{b, j} \leq \xi), \\ \sum_{s=1}^r \mathbb{P}\left(\hat{Y}_{b, N_b(s)} \leq \hat{Y}_{a, \hat{S}_a^s(N_a(s), N_b(s))}, N_a(s) \geq N_b(s), N_b(s) \geq n_0\right) &\leq \sum_{j=n_0}^r \mathbb{P}\left(\hat{Y}_{b, j} \leq \xi\right) + r \sum_{j=n_0}^r \mathbb{P}(Y_{a, j} \geq \xi). \end{aligned}$$

So far, we note that the analysis has *not* been specific to RB-SDA but applies to any instance of SDA. Then, we provide in Lemma 4.3 an upper bound on  $\sum_{t=1}^T \mathbb{P}(N_1(t) \leq (\log t)^2)$  which is specific to RB-SDA. This sampler is randomized and independent of  $r$ , hence we use the notation  $\text{RB}(m, n) = \text{RB}(m, n, r)$ . The strength of this upper bound is that it decouples the properties of the sub-sampling algorithm and that of the arm distributions (through the balance function  $\alpha_k$ ).

**Lemma 4.3.** *Let  $X_{m, H, j}$  be a random variable giving the number of non-overlapping sub-samples of size  $j$  obtained in  $m$  i.i.d. samples from  $\text{RB}(H, j)$  and define  $c_r = \lfloor \frac{r/(\log r)^2 - 1}{2K} \rfloor - 1$ . There exists  $\gamma \in (0, 1)$  and a constant  $r_K$  such that with  $\beta_{r, j} = \lfloor \gamma r / j (\log r)^2 \rfloor$ ,*

$$\sum_{r=1}^T \mathbb{P}(N_1(r) \leq (\log r)^2) \leq r_K + \sum_{r=r_K}^T \sum_{j=1}^{\lfloor \log r^2 \rfloor} \left[ (K-1) \mathbb{P}(X_{c_r, c_r, j} < \beta_{r, j}) + \sum_{k=2}^K \alpha_k(\beta_{r, j}, j) \right].$$

To prove Lemma 4.3 (see Appendix E), we extend the proof technique introduced by [1] for the analysis of BESA to handle more than 2 arms. The rationale is that if  $N_1(r) \leq (\log r)^2$  then arm 1 is not the leader and has lost "many" duels, more precisely *at least* a number of *successive duels*

proportional to  $r/(\log r)^2$ . A fraction of these duels necessarily involves sub-samples of the leader history that have non-overlapping support. Exploiting the independence of these sub-samples brings in the balance function  $\alpha_k$ .

In order to conclude the proof, it remains to upper bound the right hand side of Lemma 4.3. Using condition 2. of balanced distributions the terms depending on  $\alpha_k$  sum in  $o(\log T)$  and negligibly contribute to the regret. Upper bounding the term featuring  $X_{m,H,j}$  amounts to establishing the following diversity property of the random block sampler.

**Definition** (Diversity Property). *Let  $X_{m,H,j}$  be the random variable defined in Lemma 4.3 for a randomized sampler SP. SP satisfies the Diversity Property for a sequence  $N_r$  of integers if*

$$\sum_{r=1}^T \sum_{j=1}^{(\log r)^2} \mathbb{P}(X_{N_r, N_r, j} < \gamma r / (\log r)^2) = o(\log T).$$

We prove in Appendix F that the RB sampler satisfies the diversity property for the sequence  $c_r$ , which leads to  $\sum_{t=1}^T \mathbb{P}(N_1(t) \leq (\log t)^2) = o(\log(T))$  and concludes the proof of Theorem 3.1.

We believe that the WR sampler also satisfies the diversity property (as conjectured by [1]). While Lemma 4.3 should apply to WR-SDA as well, a different path has to be found for analyzing the LDS-SDA and LB-SDA algorithms, that are based on deterministic samplers and also perform well in practice. This is left for future work.

## 5 Experiments

In this section, we perform experiments on simulated data in order to illustrate the good performance of the four instances of SDA algorithms introduced in Section 2 for various distributions. The Python code used to perform these experiments is available on Github.

**Exponential families** First, in order to illustrate Corollary 3.1.1, we investigate the performance of RB-SDA for both Bernoulli and Gaussian distributions (with known variance 1). Our first objective is to check that for a finite horizon the regret of RB-SDA is comparable with the regret of Thompson Sampling (with respectively a beta and improper uniform prior), which efficiently uses the knowledge of the distribution. Our second objective is to empirically compare different variants of SDA to other non-parametric approaches based on sub-sampling (BESA, SSMC) or on re-sampling. For Bernoulli and Gaussian distribution, Non-Parametric TS coincides with Thompson Sampling, so we focus our study on algorithms based on history perturbation. We experiment with PHE [14] for Bernoulli bandits and ReBoot [17] for Gaussian bandits, as those two algorithms are guaranteed to have logarithmic regret in each of these settings. As advised by the authors, we use a parameter  $a = 1.1$  for PHE and  $\sigma = 1.5$  for ReBoot.

We ran experiments on 4 different Bernoulli bandit models: 1)  $K = 2$ ,  $\mu = [0.8, 0.9]$ , 2)  $K = 2$ ,  $\mu = [0.5, 0.6]$ , 3)  $K = 10$ ,  $\mu_1 = 0.1, \mu_{2,3,4} = 0.01, \mu_{5,6,7} = 0.03, \mu_{8,9,10} = 0.05$ , 4)  $K = 8$ ,  $\mu = [0.9, 0.85, \dots, 0.85]$  and 3 different bandits models with  $\mathcal{N}(\mu_k, 1)$  arms with means: 1)  $K = 2$ ,  $\mu = [0.5, 0]$ , 2)  $K = 4$ ,  $\mu = [0.5, 0, 0, 0]$ , 3)  $K = 4$ ,  $\mu = [1.5, 1, 0.5, 0]$ . For each experiment, Table 1 and Table 2 report an estimate of the regret at time  $T = 20000$  based on 5000 independent runs (extended tables with standard deviations can be found in Appendix A.1). The best performing algorithms are highlighted in bold. In Figure 1 and Figure 2 we plot the regret of several algorithms as a function of time (in log scale) for  $t \in [15000; 20000]$  for one Bernoulli and one Gaussian experiment respectively. We also add the Lai and Robbins lower bound  $t \mapsto [\sum_k (\mu^* - \mu_k) / \text{kl}(\mu_k, \mu^*)] \log(t)$ .

Table 1: Regret at  $T = 20000$  for Bernoulli arms

xp	Benchmark				SDA			
	TS	PHE	BESA	SSMC	RB	WR	LB	LDS
1	<b>11.2</b>	25.9	<b>11.7</b>	<b>12.3</b>	<b>11.5</b>	<b>11.6</b>	<b>12.2</b>	<b>11.4</b>
2	<b>22.9</b>	<b>24.0</b>	<b>22.1</b>	<b>24.3</b>	<b>22.0</b>	<b>21.5</b>	<b>24.0</b>	<b>21.8</b>
3	<b>94.2</b>	248.1	<b>88.1</b>	100.1	<b>89.0</b>	<b>86.9</b>	100.7	<b>89.2</b>
4	<b>108.1</b>	216.5	147.5	119.9	<b>105.1</b>	<b>106.9</b>	119.6	<b>106.8</b>

Figure 1: Regret as a function of time for Bernoulli experiment 3

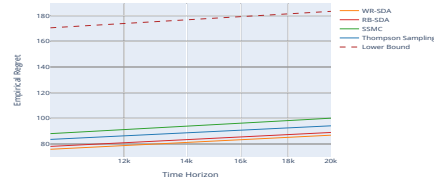
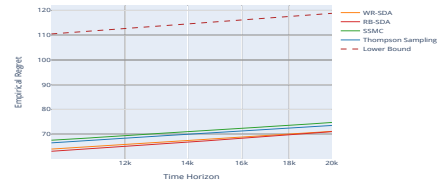


Table 2: Regret at  $T = 20000$  for Gaussian arms

xp	Benchmark				SDA			
	TS	ReBoot	BESA	SSMC	RB	WR	LB	LDS
1	<b>24.4</b>	92.2	<b>25.3</b>	<b>26.9</b>	<b>25.6</b>	<b>24.7</b>	<b>25.1</b>	<b>26.5</b>
2	<b>73.5</b>	277.1	122.5	<b>74.8</b>	<b>71.0</b>	<b>71.1</b>	<b>74.6</b>	<b>69.0</b>
3	<b>49.7</b>	190.9	72.1	<b>51.3</b>	<b>50.4</b>	<b>50.0</b>	<b>51.2</b>	<b>48.6</b>

Figure 2: Regret as a function of time for Gaussian experiment 2



In all of these experiments, we notice that SDA algorithms are indeed strong competitors to Thompson Sampling (with appropriate prior) for both Bernoulli and Gaussian bandits. Figures 1 and 2 further show that RB-SDA is empirically matching the Lai and Robbins’ lower bound on two instances, just like SSMC and Thompson Sampling, which can be seen from the parallel straight lines with the  $x$  axis in log scale. The fact that the lower bound is above shows that it is really asymptotic and only captures the right first order term. The same observation was made for all experiments, but is not reported due to space limitation. Even if we only established the asymptotic optimality of RB-SDA, these results suggest that the other SDA algorithms considered in this paper may also be asymptotically optimal. Compared to SDA, re-sampling algorithms based on history perturbation seem to be much less robust. Indeed, in the Bernoulli case, PHE performs very well for experiment 2, but is significantly worse than Thompson Sampling on the three other instances. In the Gaussian case, ReBoot always performs significantly worse than other algorithms. This lack of robustness is also corroborated by additional experiments reported in Appendix ?? in which we average the performance of these algorithms over a large number of randomly chosen instances.

Turning our attention to algorithms based on sub-sampling, we first notice that WR-SDA seems to be a better generalization of BESA with 2 arms than the tournament approach currently proposed, as in experiments with  $K > 2$ , WR-SDA often performs significantly better than BESA. Then we observe that SSMC and SDA algorithms have similar performance. Looking a bit closer, we see that the performance of SSMC is very close to that of LB-SDA, whereas SDA algorithms based on “randomized” (or pseudo-randomized for LDS-SDA) samplers tend to perform slightly better.

**Truncated Gaussian** Theorem 3.1 suggests that RB-SDA may attain logarithmic regret beyond exponential families. As an illustration, we present the results of experiments performed with Truncated Gaussian distributions (in which the distribution of arm  $k$  is that of  $Y_k = 0 \vee (X_k \wedge 1)$  where  $X_k \sim \mathcal{N}(\mu_k, \sigma^2)$ ). We report in Table 8 the regret at time  $T = 20000$  (estimated over 5000 runs) of various algorithms on four different problem instances: 1)  $\mu = [0.5, 0.6]$ ,  $\sigma = 0.1$ , 2)  $\mu = [0, 0.2]$ ,  $\sigma = 0.3$ , 3)  $\mu = [1.5, 2]$ ,  $\sigma = 1$  4)  $\mu = [0.4, 0.5, 0.6, 0.7]$ ,  $\sigma = 1$ . We include Non-Parametric TS which is known to be asymptotically optimal in this setting (while TS which uses a Beta prior and a binarization trick is not), PHE, and all algorithms based on sub-sampling. We again observe the good performance of SSMC and SDA algorithms across all experiments. They even outperform NP-TS in some experiments, which suggests SDA algorithms may be asymptotically optimal for a wider class of parametric distributions.

Table 3: Regret at  $T = 20000$  for Truncated Gaussian arms

xp	Benchmark				SDA				
	TS	NP-TS	PHE	BESA	SSMC	RB	WR	LB	LDS
1	21.9	4.2	22.3	<b>1.4</b>	<b>1.5</b>	<b>1.4</b>	<b>1.4</b>	<b>1.5</b>	<b>1.4</b>
2	13.3	8	19.5	<b>4.6</b>	<b>4.7</b>	<b>4.4</b>	<b>4.5</b>	<b>4.6</b>	<b>4.3</b>
3	9.7	<b>7.8</b>	48.5	<b>7.8</b>	<b>7.6</b>	<b>7.1</b>	<b>7.7</b>	8.2	<b>7.1</b>
4	86.6	<b>70</b>	86	76.5	<b>69.5</b>	<b>64.9</b>	<b>64.8</b>	<b>68.7</b>	<b>63.2</b>

**Bayesian Experiments** So far we tried our algorithms on specific instances of the distributions we considered. It is also interesting to check the robustness of the algorithms when the means of the arms are drawn at random according to some distribution. In this section we consider two examples:



Bernoulli bandits where the arms are drawn uniformly at random in  $[0, 1]$ , and Gaussian distributions with the mean parameter of each arm itself drawn from a gaussian distribution  $\mu_k \sim \mathcal{N}(0, 1)$ . In both cases we draw 10000 random problems with  $K = 10$  arms and run the algorithms for a time horizon  $T = 20000$ . We experiment with TS, SSMC, RB-SDA and WR-SDA and also add the IMED algorithm ([23]) which is an asymptotically optimal algorithm that uses the knowledge of the distribution. We do not add LDS-SDA and LB-SDA as they are similar to RB-SDA and SSMC, respectively. In the Bernoulli case, we also run the PHE algorithm, which fails to compete with the other algorithms. This is not in contradiction with the results of [14] as in the Bayesian experiments of this paper, arms are drawn uniformly in  $[0.25, 0.75]$  instead of  $[0, 1]$ . Actually, we noticed that PHE with parameter  $a = 1.1$  has some difficulties when several arms are close to 1.

Table 4: Average Regret on 10000 random experiments with Bernoulli Arms

T	TS	IMED	PHE	SSMC	RB	WR
100	13.8	15.1	16.7	16.5	14.8	14.3
1000	27.8	31.9	39.5	34.2	31.8	30.9
10000	45.8	51.2	72.3	55.0	51.1	50.6
20000	52.2	57.6	85.6	61.9	57.7	57.3

Table 5: Average Regret on 10000 random experiments with Gaussian Arms

T	TS	IMED	WR	RB	SSMC
100	41.2	45.1	38.3	38.1	40.6
1000	76.4	82.1	72.7	70.4	76.2
10000	118.5	124.0	115.8	111.8	120.1
20000	132.6	138.1	130.2	125.7	135.1

Results reported in Tables 4 and 5 show that RB-SDA and WR-SDA are strong competitors to TS and IMED for both Bernoulli and Gaussian bandits. Recall that these algorithm operate without the need for a specific tuning for each distribution, unlike TS and IMED. Moreover, observe that in the Bernoulli case, TS further uses the same prior as that from which the means are drawn.

**Computational aspects** To choose a sub-sampling based algorithm, numerical consideration can be taken into account. First, compared to Thompson Sampling, all sub-sampling based algorithm require to store the history of the observation. But then, the cost of sub-sampling varies across algorithms: in the general case RB-SDA is more efficient than WR-SDA as the latter requires to draw a random subset while the former only needs to draw the random integer starting the block. However, for distributions with finite supports WR-SDA can be made as efficient as TS using multivariate geometric distributions, just like PHE does. If one does not want to use randomization then LDS-SDA could be preferred to RB-SDA as it uses a deterministic sequence. Finally, LB-SDA has the smallest computational cost in the SDA family and while its performance is very close to that of SSMC, it can avoid the cost of scanning all the sub-sample means in this algorithm. The computational cost of these two algorithms is difficult to evaluate precisely. Indeed, they can be made very efficient when the leader does not change, but each change of leader is costly, in particular for SSMC. The expected number of such changes is proved to be finite, but for experiments with a finite time horizon the resulting constant can be big. Finally, Non-Parametric TS has a good performance for Truncated Gaussian, but the cost of drawing a random probability vector over a large history is very high.

**More experiments** In Appendix A we enhance this empirical study: we show some limitations of SDA for exponential distributions and propose a fix using forced exploration as in SSMC.

## 6 Conclusion

We introduced the SDA framework for exploration in bandits models. We proved that one particular instance, RB-SDA, combines both optimal theoretical guarantees and good empirical performance for several distributions, possibly with unbounded support. Moreover, SDA can be associated with other samplers that seem to achieve similar performance, with their own specificity in terms of computation time. The empirical study presented in the paper also shows the robustness of *sub-sampling* approach over other types of *re-sampling* algorithms. This new approach to exploration may be generalized in many directions, for example to contextual bandits or reinforcement learning, where UCB and Thompson Sampling are still the dominant approaches. It is also particularly promising to develop new algorithm for non-stationary bandit, as such algorithms already store the full history of rewards.

## Broader Impact

This work is advertising a new way to do non-parametric exploration in bandit models, that enjoy good empirical performance and strong theoretical guarantees. First, bandit problems are at the heart of numerous applications to online content recommendation, hence the good performance of SDA algorithms may inspire new algorithms for more realistic models used for these applications, such as contextual bandits. Then, exploration is a central question in the broader field of reinforcement learning, hence new ideas for bandits may lead to new ideas for reinforcement learning.

## Acknowledgments and Disclosure of Funding

The PhD of Dorian Baudry is funded by a CNRS80 grant. The authors acknowledge the funding of the French National Research Agency under projects BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04).

Experiments presented in this paper were carried out using the Grid’5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see <https://www.grid5000.fr>).

## References

- [1] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014. Proceedings, Part I*, 2014.
- [2] Hock Peng Chan. The multi-armed bandit problem: An efficient nonparametric solution. *The Annals of Statistics*, 48(1), Feb 2020.
- [3] Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.
- [4] R. Agrawal. Sample mean based index policies with  $O(\log n)$  regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4), 1995.
- [5] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3), 2002.
- [6] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3), Jun 2013.
- [7] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4), 12 1933.
- [8] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT 2012 - The 25th Annual Conference on Learning Theory*, 2012.
- [9] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1), 1985.
- [10] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012*, 2012.
- [11] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013*, 2013.
- [12] N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-dimensional Exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [13] Branislav Kveton, Csaba Szepesvari, Zheng Wen, Mohammad Ghavamzadeh, and Tor Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *ICML*, 2019.

- [14] Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 2019.
- [15] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- [16] Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *CoRR*, abs/1507.00300, 2015.
- [17] Chi-Hua Wang, Yang Yu, Botao Hao, and Guang Cheng. Residual bootstrap exploration for bandit algorithms. *CoRR*, abs/2002.08436, 2020.
- [18] Charles Riou and Junya Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.
- [19] A.N Burnetas and M. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2), 1996.
- [20] Michael Drmota and Robert Tichy. *Sequences, discrepancies and applications*, volume 1651 of *Lecture Notes in Mathematics*. Springer Verlag, Deutschland, 1 edition, 1997.
- [21] J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12), December 1964.
- [22] I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4), 1967.
- [23] Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16(113), 2015.
- [24] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *24th Annual Conference on Learning Theory (COLT)*, 2011.
- [25] Alex Mendelson, Maria Zuluaga, Brian Hutton, and Sébastien Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *CoRR*, abs/1602.05822, 2016.
- [26] B.C. Rennie and A.J. Dobson. On stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2), 1969.
- [27] Anthony W. Ledford and Jonathan A. Tawn. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), May 1997.

## A Complement of Experiments

### A.1 Additional Figures for Bernoulli, Gaussian and Truncated Gaussian arms

We enhance the tables of Section 5 with the standard deviation (reported in parenthesis) of the regret at time  $T = 20000$  on the 5000 trajectories.

Table 6: Regret and at  $T = 20000$  for Bernoulli arms, with standard deviation

xp	Benchmark				SSDA			
	TS	PHE	BESA	SSMC	RB	WR	LB	LDS
1	<b>11.2</b> (10.)	25.9 (87.9)	<b>11.7</b> (12.1)	<b>12.3</b> (7.3)	<b>11.5</b> (10.1)	<b>11.6</b> (10.2)	<b>12.2</b> (7.4)	<b>11.4</b> (9.0)
2	<b>22.9</b> (29.2)	<b>24.0</b> (22.0)	<b>22.1</b> (25.2)	<b>24.3</b> (38.2)	<b>22.0</b> (34.5)	<b>21.5</b> (17.3)	<b>24.0</b> (24.6)	<b>21.8</b> (24.5)
3	<b>94.2</b> (15.8)	248.1 (25.5)	<b>88.1</b> (89.2)	100.1 (20.0)	<b>89.0</b> (19.8)	<b>86.9</b> (21.7)	100.7 (21.3)	<b>89.2</b> (21.8)
4	<b>108.1</b> (45.1)	216.5 (89.8)	147.5 (209.8)	119.9 (40.8)	<b>105.1</b> (41.1)	<b>106.9</b> (42.1)	119.6 (42.7)	<b>106.8</b> (47.7)

Table 7: Regret and at  $T = 20000$  for Gaussian arms, with standard deviation

xp	Benchmark				SDA			
	TS	ReBoot	BESA	SSMC	RB	WR	LB	LDS
1	<b>24.4</b> (17.1)	92.2 (23.4)	<b>25.3</b> (27.1)	<b>26.9</b> (52.8)	<b>25.6</b> (62.8)	<b>24.7</b> (20.6)	<b>25.1</b> (17.9)	<b>26.5</b> (140.2)
2	<b>73.5</b> (107.8)	277.1 (41.3)	122.5 (585.5)	<b>74.8</b> (34.7)	<b>71.0</b> (152.2)	<b>71.1</b> (50.2)	<b>74.6</b> (35.1)	<b>69.0</b> (50.4)
3	<b>49.7</b> (26.9)	190.9 (29.6)	72.1 (410.3)	<b>51.3</b> (23.7)	<b>50.4</b> (156.5)	<b>50.0</b> (33.3)	<b>51.2</b> (22.4)	<b>48.6</b> (41.6)

Table 8: Regret at  $T = 20000$  for Truncated Gaussian arms

xp	Benchmark				SDA				
	TS	NP-TS	PHE	BESA	SSMC	RB	WR	LB	LDS
1	21.9 (20.4)	4.2 (0.6)	22.3 (2.6)	<b>1.4</b> (1.7)	<b>1.5</b> (0.7)	<b>1.4</b> (1.1)	<b>1.4</b> (0.8)	<b>1.5</b> (0.7)	<b>1.4</b> (0.8)
2	13.3 (7)	8 (1.8)	19.5 (3.8)	<b>4.6</b> (3.3)	<b>4.7</b> (2.3)	<b>4.4</b> (4.6)	<b>4.5</b> (3.1)	<b>4.6</b> (2.4)	<b>4.3</b> (2.9)
3	9.7 (10.1)	<b>7.8</b> (4.5)	48.5 (217.8)	<b>7.8</b> (9.4)	<b>7.6</b> (5)	<b>7.1</b> (10)	<b>7.7</b> (13.4)	8.2 (27.5)	<b>7.1</b> (5.8)
4	86.6 (57.8)	<b>70</b> (39.4)	86 (53.7)	76.5 (113.9)	<b>69.5</b> (40.9)	<b>64.9</b> (60.5)	<b>64.8</b> (43.9)	<b>68.7</b> (39.1)	<b>63.2</b> (51.1)

For Bernoulli arms and Truncated Gaussian, the standard deviations of SDA are very similar to that of Thompson Sampling, while the trajectories of PHE and BESA have much more variance in experiment 1 and 4, and on experiments 3 and 4 respectively. For Gaussian arms we remark the low variability of ReBoot, but at the cost of a non-competitive regret. SDA are less homogeneous in this

case: some algorithms have large variance for some instances (LDS-SDA on experiment 1, RB-SDA on experiments 2 and 3). Note that TS also has a high variability in experiment 2.

We believe that this is due to the nature of the Gaussian distribution, and in particular to its balance function: in Appendix G we prove that  $\alpha_k(M, j)$  does satisfy Assumption 2. of Theorem 3.1, however the upper bound derived for  $\alpha_k(M, j)$  is much larger than the one for Bernoulli distribution, which justifies that “bad runs” in which a good arm loses many duels are more likely to happen in that case, and can explain the larger variance. If one wants to reduce the variance of the regret of SDA we recommend the use of some asymptotically negligible forced exploration, as presented for exponential distribution in Appendix A.2, and for which we prove that the algorithm remains asymptotically optimal in Appendix H.

Finally, as in Section 5, we plot the regret of several algorithms as a function of time (in log scale) for  $t \in [10000, 20000]$ , this time for the Truncated Gaussian distributions. These plots illustrate the fact that some SDA algorithms may achieve asymptotic optimality for this distribution too, even if it does not belong to a one-parameter exponential family. Indeed, the rate of the regret of all SDA seem too match both the rate of the regret of Non-Parametric TS, which is optimal for this family, and the Burnetas and Katehakis lower bound whose expression is  $\left(\sum_{k \neq k^*} \frac{\mathbb{E}_{X \sim \nu_{k^*}}[X] - \mathbb{E}_{X \sim \nu_k}[X]}{\text{KL}(\nu_k, \nu_{k^*})}\right) \log(T)$  in this particular case, with  $\text{KL}(\nu_k, \nu_{k^*}) = p_{0,k} \log\left(\frac{p_{0,k}}{p_{0,*}}\right) + (1 - p_{1,k}) \log\left(\frac{1 - p_{1,k}}{1 - p_{1,*}}\right) + \int_0^1 f_k(x) \log\left(\frac{f_k(x)}{f_*(x)}\right) dx$ .  $p_{x,k}$  is the value of the CDF of the underlying Gaussian random variable associated with  $\nu_k$  in  $x$ , and  $f_k(x)$  the density of this variable in  $x$ .

Figure 3: SDA vs NP-TS on TG expe 2

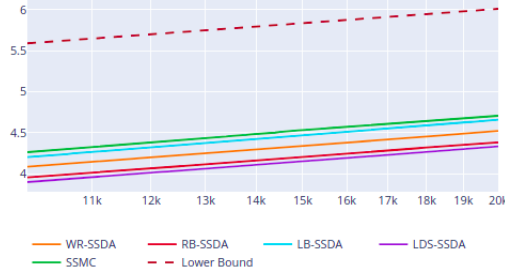
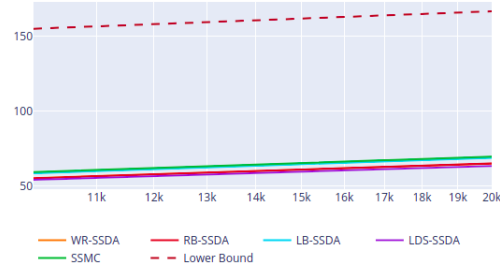


Figure 4: SDA vs NP-TS on TG expe 4



## A.2 Experiments with Exponential Arms

In Appendix G, we prove that exponential distributions are *not* balanced (i.e. do not satisfy Assumption 2. of Theorem 3.1), so our theoretical results on the regret of RB-SDA do not apply. However, it is still interesting to test our algorithms for these distributions in order to see if it still achieves a good performance. We performed 6 experiments, with the following mean parameters: 1)  $\mu = [1.5, 1]$ , 2)  $\mu = [0.2, 0.1]$ , 3)  $\mu = [11, 10]$ , 4)  $\mu = [4, 3, 2, 1]$ , 5)  $\mu = [0.4, 0.3, 0.2, 0.1]$ , 6)  $\mu = [5, 4, 4, 4]$ . It is interesting to remark that the standard deviation of an exponential distribution is equal to its mean, so with similar gaps problems are harder when the means are high.

Table 9: Average Regret with Exponential Arms (with std)

xp	TS	IMED	BESA	SSMC	RB	WR	LB	LDS
1	48.2 (191.8)	<b>40.0</b> (78.4)	45.7 (114.1)	<b>41.9</b> (84.2)	44.8 (121.4)	45.4 (134.4)	46.6 (176.8)	45.5 (109.7)
2	3.8 (9.9)	<b>3.4</b> (3.6)	4.2 (25.1)	<b>3.6</b> (41.9)	4.1 (14.3)	3.9 (13.4)	3.9 (8.7)	5.4 (49.5)
3	832.8 (1065.1)	<b>779.9</b> (896.9)	820.5 (1304.6)	856.9 (1111.0)	848.4 (1533.3)	<b>778.4</b> (1118.7)	846.7 (1150.1)	877.7 (1708.7)
4	258.3 (519.6)	<b>234.6</b> (126.6)	525.4 (2115.1)	<b>251.3</b> (328.3)	272.6 (692.2)	262.1 (524.4)	263.8 (477.9)	258.4 (599.0)
5	<b>25.6</b> (51.2)	<b>24.0</b> (33.6)	55.7 (219.9)	<b>25.6</b> (23.6)	<b>25.5</b> (46.7)	<b>25.0</b> (24.0)	26.5 (36.8)	<b>24.7</b> (37.6)
6	<b>618.7</b> (672.3)	<b>603.6</b> (576.8)	1184.2 (3096.4)	<b>627.9</b> (755.6)	<b>595.7</b> (790.7)	<b>616.0</b> (780.2)	652.6 (685.3)	<b>605.9</b> (871.4)

First, we notice that the performance of the SDA in terms of the average regret are reasonable, although less impressive than with the other distributions we tested. IMED is almost always the best algorithm in these experiments, and SSMC performs pretty well on many examples (which is not surprising as SSMC is proved to be asymptotically optimal for exponential distributions). We remark that there is much more variability in the results of RB-SDA, WR-SDA and LDS-SDA than before, where they performed quite similarly. For instance, we notice that on example 3, LDS-SDA and RB-SDA are much worse than WR-SDA. A look at the quantile table for this experiment, which displays the empirical quantiles of  $R_T$  estimated over 5000 runs, shows that this is due to a small number of "bad" trajectories for these algorithms:

Table 10: Quantile Table for Experiment 3 with Exponential Arms

% of runs	TS	IMED	SSMC	RB	WR	LB	LDS
20%	319.8	336.0	335.0	261.0	290.0	326.0	261.8
50%	626.0	650.0	661.0	532.0	568.5	642.0	536.0
80%	1122.0	1080.0	1142.0	1006.0	1019.0	1143.2	1020.2
95%	1924.1	1704.0	1846.0	<b>2199.0</b>	1817.2	1869.1	<b>2134.1</b>
99%	4209.4	2632.9	3536.8	<b>6813.1</b>	4146.0	3762.3	<b>7396.7</b>

We see that up to the 80% quantile, RB-SDA and LDS-SDA are even significantly better than IMED. This is very different when we look at the 95% and 99% quantiles, which are much greater for our 2 algorithms (even 2.5 times greater for the 99% quantile).

We believe that this very high variability prevents RB-SDA to have a logarithmic regret for exponential arms. Still, the regret is not as bad as being linear, as using the fact that the balance function  $\alpha_k(M, j)$  is of order  $\exp(-jC)/M$  permits to prove that  $\sum_{r=1}^T \mathbb{P}(N_1(r) < \log^2(r)) = \mathcal{O}(\log^2(T))$  (which requires to choose a different  $\beta_{r,j}$  in Lemma 4.3). But we also found a solution to obtain (asymptotically optimal) linear regret, which consists in adding an asymptotically negligible amount of *forced exploration* as the SSMC algorithm does. This exploration in  $o(\log T)$  avoids trajectories where the optimal arm has a very bad first observation and is not drawn for a very long time. In Appendix H, we prove the asymptotic optimality of RB-SDA with forced exploration  $f_r = \sqrt{\log r}$  for any one-dimensional exponential family. In practice, adding this amount of forced exploration to SDA algorithms leads to the following results:

Table 11: Average Regret with Exponential Arms: SDA with forced exploration

xp	RB	WR	LB	LDS
1	44.9 (167.3)	<b>42.5</b> (107.4)	<b>42.4</b> (60.5)	45.0 (176.0)
2	<b>3.6</b> (9.2)	<b>3.4</b> (2.2)	4.0 (27.9)	<b>3.6</b> (11.2)
3	837.5 (1466.1)	<b>788.5</b> (1222.1)	827.7 (1055.3)	832.3 (1514.6)
4	<b>244.8</b> (403.3)	<b>238.9</b> (250.8)	251.7 (248.5)	<b>246.0</b> (323.4)
5	<b>23.6</b> (33.4)	<b>25.1</b> (41.0)	<b>25.4</b> (23.4)	<b>24.9</b> (42.2)
6	<b>578.9</b> (651.9)	<b>595.1</b> (561.3)	631.2 (446.4)	<b>577.8</b> (652.7)

Hence, adding forced exploration results in a noticeable improvement for SDA algorithms, with RB-SDA, WR-SDA and LDS-SDA becoming competitive with IMED (or even slightly better) on most examples. Observe that LB-SDA has again comparable performance with SSMC with this new feature. This is not surprising as we implemented the SSMC algorithm with the same amount of forced exploration  $f_r = \sqrt{\log r}$ .

## B Notation for the Proof

General notations:

- $K$  number of arms
- $\nu_k$  distribution of the arm  $k$ , with mean  $\mu_k$
- we assume that  $\mu_1 = \max_{k \in [K]} \mu_k$  so we call the (unique) optimal arm "arm 1"
- $I_k(x)$  some rate function of the arm  $k$ , evaluated in  $x$ . For 1-parameter exponential families this function will always be the KL-divergence between  $\nu_k$  and the distribution from the same family with mean  $x$ .
- $N_k(r)$  number of pull of arm  $k$  up to (and including) round  $r$ .
- $Y_{k,i}$  reward obtained at the  $i$ -th pull of arm  $k$ .
- $\hat{Y}_{k,i}$  mean of the  $i$ -th first reward of arm  $k$ ,  $\hat{Y}_{k,\mathcal{S}}$  mean of the rewards of  $k$  on a subset of indices  $\mathcal{S} \subset [N_k(r)]$ :  $\hat{Y}_{k,\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} Y_{k,s}$ . If  $|\mathcal{S}| = i$ , then  $\hat{Y}_{k,i}$  and  $Y_{k,\mathcal{S}}$  have the same distribution.
- $\ell(r)$  leader at round  $r + 1$ ,  $\ell(r) = \operatorname{argmax}_{k \in [K]} N_k(r)$ .
- $\text{SP}(m, n, r)$  sub-sampling algorithm, or Sampler, which returns a sequence of  $n$  unique elements out of  $[m]$ .
- $(S_k^r(m, n))_{m \geq n}$  a family of independent random variables such that  $S_k^r(m, n) \sim \text{SP}(m, n, r)$ .
- $\mathcal{A}_r$  set of arms pulled at a round  $r$ .
- $\mathcal{R}_r$  regret at the *end* of round  $r$ .

Notations for the regret analysis, part relying on concentration:

- $\mathcal{G}_k^r = \cup_{s=1}^{r-1} \{\ell(s) = 1\} \cap \{k \in \mathcal{A}_{s+1}\} \cap \{N_k(s) \geq (1 + \varepsilon)\xi_k \log r\}$
- $\mathcal{H}_k^r = \cup_{s=1}^{r-1} \{\ell(s) = 1\} \cap \{k \in \mathcal{A}_{s+1}\} \cap \{N_k(s) \geq J_k \log r\}$
- $\mathcal{Z}^r = \{\ell(r) \neq 1\}$ , the leader used for the duels in round  $r + 1$  is sub-optimal

- $\mathcal{D}^r = \{\exists u \in [\lfloor r/4 \rfloor, r] \text{ such that } \ell(u-1) = 1\}$ , the leader has been optimal at least once between  $\lfloor r/4 \rfloor$  and  $r$
- $\mathcal{B}^u = \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1 \text{ for some arm } k\}$ , the optimal arm is leader in  $u$  but loses its duel against some arm  $k$ , that have been pulled enough to possibly take over the leadership at next round
- $\mathcal{C}^u = \{\exists k \neq 1, N_k(u) \geq N_1(u), \hat{Y}_{k, S_k^u(N_k(u), N_1(u))} \geq \hat{Y}_{1, N_1(u)}\}$ , the optimal arm is not the leader and has lost its duel against the sub-optimal leader.
- $\mathcal{L}^r = \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{1}_{\mathcal{C}^u}$

Notations for the regret analysis, control of the number of pulls of the optimal arm:

- $r_j$  round of the  $j$ -th play of the optimal arm
- $\tau_j = r_{j+1} - r_j$
- $\mathcal{E}_j^r := \{\tau_j \geq r/\log r^2 - 1\}$
- $\mathcal{M}_{j,r}^1 = \left[ r_j + 1, r_j + \left\lceil \frac{r/\log r^2 - 1}{2} \right\rceil \right]$
- $\mathcal{M}_{j,r}^2 = \left[ t_j + \left\lceil \frac{r/\log r^2 - 1}{2} \right\rceil, r_j + \lfloor r/\log r^2 \rfloor - 1 \right]$
- $\mathcal{I}_{j,r}^k = \{s \in \mathcal{M}_{j,r}^2 : \ell(s-1) = k\}$
- $\mathcal{W}_{s,j}^k = \left\{ \left\{ \hat{Y}_{1,j} < \hat{Y}_{k, S_k^s(N_k(s), j)} \right\}, N_k(s) \geq c_{r,K}, N_1(s) = j \right\}$
- $\mathcal{F}_{j,M}^{k,r} = \left\{ \exists i_1, \dots, i_M \in \mathcal{I}_{j,r}^k : \forall m < m' \in [M], S_1^{i_m}(N_k(i_m), j) \cap S_1^{i_{m'}}(N_k(i_{m'}), j) = \emptyset \right\}$
- CDF: Cumulative Distribution Function, PDF: Probability Density Function and PMF: Probability Mass Function.

## C Concentration Result: Proof of Lemma 4.2

We first recall the probabilistic model introduced in Section 2: for each round  $r$ , each arm  $k$ , we define a family  $(S_k^r(n, m))_{n > m}$  of independent random variables such that  $S_k^r(n, m) \sim \text{SP}(n, m, r)$ . Those random variables are also independent from the reward streams  $(Y_{k,s})_{s \geq 0}$  of all arms  $k$ .

$S_k^r(n, m)$  is the subset of the leader history that is used should arm  $k$  be a challenger drawn  $m$  times up to round  $r$  duelling against a leader that has been drawn  $n$  times. With this notation, letting  $\ell(r)$  be the leader after  $r$  rounds, at round  $r+1$ , for all  $k \neq \ell(r)$ ,

$$(k \in \mathcal{A}_{r+1}) \Leftrightarrow \left( \hat{Y}_{k, N_k(r)} > \hat{Y}_{\ell(r), S_{k, S_k^r(N_{\ell(r)}(r), N_k(r))}} \right).$$

Let  $k$  be an arm such that  $\mu_k < \mu_1$ . We denote by  $[n_1, n_k]$  the set of subset of  $\{1, \dots, n_1\}$  of size  $n_k$ . We define an event

$$\mathcal{Q}_k^s = \{N_k(s) \geq n_0, \ell(s) = 1, \hat{Y}_{k, N_k(s)} > \hat{Y}_{\ell(s), S_{k, S_k^s(N_1(s), N_k(s))}}\}.$$

Noting that  $\{\hat{Y}_{k, N_k(s)} > \hat{Y}_{\ell(s), S_{k, S_k^s(N_1(s), N_k(s))}}\} \subset \{\hat{Y}_{k, N_k(s)} \geq \xi\} \cup \{\hat{Y}_{\ell(s), S_{k, S_k^s(N_1(s), N_k(s))}} \leq \xi\}$  for all  $\xi \in \mathbb{R}$ , we can write  $\mathcal{Q}_k^s \subset \mathcal{Q}_k^{s,1} \cup \mathcal{Q}_k^{s,2}$  where

$$\begin{aligned} \mathcal{Q}_k^{s,1} &= \{N_k(s) \geq n_0, \ell(s) = 1, \hat{Y}_{k, N_k(s)} > \xi\} \\ \text{and } \mathcal{Q}_k^{s,2} &= \{N_k(s) \geq n_0, \ell(s) = 1, \hat{Y}_{\ell(s), S_{k, S_k^s(N_1(s), N_k(s))}} \leq \xi\}. \end{aligned}$$



This yields  $\sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^s) \leq \sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{r,1}) + \sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{r,2})$ , which will later provide the two terms in the bound of the lemma. The first one does not involve sub-sampling and can be upper bounded as:

$$\begin{aligned}
\sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{s,1}) &\leq \mathbb{E} \sum_{s=1}^r \mathbb{1}(N_k(s) \geq n_0) \mathbb{1}(N_1(s) > N_k(s)) \mathbb{1}(\hat{Y}_{k,N_k(s)} \geq \xi) \mathbb{1}(k \in \mathcal{A}_{s+1}) \\
&\leq \mathbb{E} \sum_{s=n_0}^r \sum_{n_k=n_0}^r \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) \mathbb{1}(\hat{Y}_{k,n_k} \geq \xi) \\
&\leq \mathbb{E} \sum_{n_k=n_0}^r \mathbb{1}(\hat{Y}_{k,n_k} \geq \xi) \underbrace{\sum_{s=n_0}^r \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1})}_{\leq 1} \\
&\leq \sum_{n_k=n_0}^r \mathbb{P}(\hat{Y}_{k,n_k} \geq \xi),
\end{aligned}$$

where in the last inequality we use that the event  $(N_k(s) = n) \cap (k \in \mathcal{A}_{s+1})$  can happen at most once for  $s \in \{n_0, \dots, r\}$  (a similar trick was used for example in the analysis of kl-UCB [24]).

Upper bounding the second term  $B_r = \sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{r,2})$  is more intricate as it involves both  $N_k(s)$  and  $N_1(s)$ . With a similar method we get:

$$\begin{aligned}
B_r &\leq \mathbb{E} \sum_{s=n_0}^r \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{S \in [n_1, n_k]} \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) \mathbb{1}(N_1(s) = n_1) \mathbb{1}(S_k^s(n_1, n_k) = S) \mathbb{1}(\hat{Y}_{\ell, S} \leq \xi) \\
&\leq \mathbb{E} \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{S \in [n_1, n_k]} \mathbb{1}(\hat{Y}_{\ell, S} \leq \xi) \sum_{s=n_0}^r \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) \mathbb{1}(S_k^s(n_1, n_k) = S) \\
&= \mathbb{E} \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{S \in [n_1, n_k]} \mathbb{1}(\hat{Y}_{\ell, S} \leq \xi) \sum_{s=n_0}^r \mathbb{E}[\mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) \mathbb{1}(S_k^s(n_1, n_k) = S) | \mathcal{F}],
\end{aligned}$$

where  $\mathcal{F}$  is the filtration generated by the reward streams.  $N_k(s)$  may have a complicated distribution with respect to this filtration but this is not a problem here. Indeed,  $S_k^s(n_1, n_k)$  is by design independent of this filtration, and one can write

$$\begin{aligned}
B_r &\leq \mathbb{E} \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{\mathcal{S} \in [n_1, n_k]} \mathbb{1}(\hat{Y}_{1, \mathcal{S}} \leq \xi) \sum_{s=n_0}^r \mathbb{P}(S_k^s(n_1, n_k) = \mathcal{S}) \mathbb{E}[\mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) | \mathcal{F}] \\
&= \mathbb{E} \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{\mathcal{S} \in [n_1, n_k]} \mathbb{1}(\hat{Y}_{1, \mathcal{S}} \leq \xi) \sum_{s=n_0}^r \mathbb{P}(S_k^s(n_1, n_k) = \mathcal{S}) \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1}) \\
&= \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \sum_{\mathcal{S} \in [n_1, n_k]} \mathbb{P}(\hat{Y}_{1, \mathcal{S}} \leq \xi) \sum_{s=n_0}^r \mathbb{P}(S_k^s(n_1, n_k) = \mathcal{S}) \mathbb{E}(\mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1})) \\
&= \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \mathbb{P}(\hat{Y}_{1, n_1} \leq \xi) \sum_{s=n_0}^r \left( \sum_{\mathcal{S} \in [n_1, n_k]} \mathbb{P}(S_k^s(n_1, n_k) = \mathcal{S}) \right) \mathbb{E}(\mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1})) \\
&= \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \mathbb{P}(\hat{Y}_{1, n_1} \leq \xi) \underbrace{\mathbb{E} \sum_{s=n_0}^r \mathbb{1}(N_k(s) = n_k, k \in \mathcal{A}_{s+1})}_{\leq 1} \\
&\leq \sum_{n_k=n_0}^r \sum_{n_1=n_k}^r \mathbb{P}(\hat{Y}_{1, n_1} \leq \xi) \\
&\leq r \sum_{n_1=n_0}^r \mathbb{P}(\hat{Y}_{1, n_1} \leq \xi).
\end{aligned}$$

Here we have used the independence of the  $S_k^s(m, n)$  from the reward streams and the fact that for every subset  $\mathcal{S}$  of size  $n_1$ ,  $\hat{Y}_{k, n_1}$  and  $\hat{Y}_{k, \mathcal{S}}$  have the same distribution. We can conclude as follows, proving the lemma:

$$\begin{aligned}
\sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^s) &\leq \sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{r,1}) + \sum_{s=1}^r \mathbb{P}(\mathcal{Q}_k^{r,2}) \\
&\leq \sum_{n_k=n_0}^r \mathbb{P}(\hat{Y}_{k, n_k} \geq \xi) + r \sum_{n_1=n_0}^r \mathbb{P}(\hat{Y}_{1, n_1} \leq \xi).
\end{aligned}$$

## D Regret Decomposition: Proof of Lemma 4.1

We recall that we assume that arm 1 is the only optimal arm:  $\mu_1 = \max_{k \in [K]} \mu_k$ . The proof in this section follows the path of the proof in [2] for SSMC, but hinges on the new concentration result of Lemma 4.2. Moreover, some parts need to be adapted to handle the properties of an independent sampler instead of the duelling rule used in SSMC. As in [2], we introduce the following events:

- $\mathcal{G}_k^T = \cup_{r=1}^{T-1} \{\ell(r) = 1\} \cap \{k \in \mathcal{A}_{r+1}\} \cap \{N_k(r) \geq (1 + \varepsilon)\xi_k \log T\}$
- $\mathcal{H}_k^T = \cup_{r=1}^{T-1} \{\ell(r) = 1\} \cap \{k \in \mathcal{A}_{r+1}\} \cap \{N_k(r) \geq J_k \log T\}$
- $\mathcal{Z}^r = \{\ell(r) \neq 1\}$ , the leader used at round  $r + 1$  is sub-optimal.

These events directly provide an upper bound of the number of pulls of a sub-optimal arm  $k$ :

$$\begin{aligned}
\mathbb{E}[N_k(T)] &= \mathbb{E}[N_k(T) \mathbb{1}_{\mathcal{H}_k^T}] + \mathbb{E}[N_k(T) \mathbb{1}_{\mathcal{G}_k^T} \mathbb{1}_{\bar{\mathcal{H}}_k^T}] + \mathbb{E}[N_k(T) \mathbb{1}_{\bar{\mathcal{G}}_k^T}] \\
&\leq T \mathbb{P}(\mathcal{H}_k^T) + (1 + J_k \log T) \mathbb{P}(\mathcal{G}_k^T) + 1 + (1 + \varepsilon)\xi_k \log T + 2 \sum_{r=1}^{T-1} \mathbb{P}(\mathcal{Z}^r) \quad (3)
\end{aligned}$$

Indeed, due to the definition of each event we have:

$$\begin{aligned}
N_k(T) \mathbb{1}_{\bar{\mathcal{G}}_k^T} &\leq 1 + \sum_{r=1}^{T-1} \mathbb{1}_{(k \in \mathcal{A}_{r+1})} \mathbb{1}_{(\ell(r) \neq 1) \cup (N_k(r) < (1+\varepsilon)\xi_k \log(T))} \\
&\leq 1 + \sum_{r=1}^{T-1} \mathbb{1}_{(k \in \mathcal{A}_{r+1})} \mathbb{1}_{(N_k(r) < (1+\varepsilon)\xi_k \log(T))} + \sum_{r=1}^{T-1} \mathbb{1}_{(\ell(r) \neq 1)} \\
&\leq 1 + (1 + \varepsilon)\xi_k \log T + \sum_{r=1}^{T-1} \mathbb{1}_{\mathcal{Z}^r}
\end{aligned}$$

and similarly

$$\begin{aligned}
N_k(T) \mathbb{1}_{\mathcal{G}_k^T} \mathbb{1}_{\bar{\mathcal{H}}_k^T} &\leq \left( 1 + J_k \log T + \sum_{r=1}^{T-1} \mathbb{1}_{\mathcal{Z}^r} \right) \mathbb{1}_{\mathcal{G}_k^T} \\
&\leq (1 + J_k \log T) \mathbb{1}_{\mathcal{G}_k^T} + \sum_{r=1}^{T-1} \mathbb{1}_{\mathcal{Z}^r}
\end{aligned}$$

Choosing  $\xi_k = 1/I_1(\mu_k)$  the bound in (3) exhibits the term in  $\frac{1+\varepsilon}{I_1(\mu_k)} \log T$  in Lemma 4.1. To obtain the result, it remains to upper bound

$$T\mathbb{P}(\mathcal{H}_k^T) + (1 + J_k \log T)\mathbb{P}(\mathcal{G}_k^T) + 2 \sum_{r=1}^{T-1} \mathbb{P}(\mathcal{Z}^r)$$

for an appropriate choice of  $J_k$ . To do so, we shall first use the concentration inequality Lemma 4.2 to upper bound the terms involving  $\mathcal{G}_k^T$  and  $\mathcal{H}_k^T$  by problem-dependent constants (Appendix D.1), and then we carefully handle the terms in  $\mathcal{Z}^r$  (Appendix D.2).

### D.1 Upper Bounds on $\mathbb{P}(\mathcal{G}_k^T)$ and $\mathbb{P}(\mathcal{H}_k^T)$

We first fix some real numbers  $J_k$  and  $\omega, \omega_k$  in  $(\mu_k, \mu_1)$  to be specified later. We also fix  $\xi_k = 1/I_1(\mu_k)$ . Starting with  $\mathcal{G}_k^T$ , we apply the second statement in Lemma 4.2 for arm  $k$  and arm 1 with  $n_0 = (1 + \varepsilon)\xi_k \log r$  and  $\xi = \omega_k$ :

$$\begin{aligned}
\mathbb{P}(\mathcal{G}_k^T) &\leq \sum_{r=1}^{T-1} \mathbb{P}(\ell(r) = 1, k \in \mathcal{A}_{r+1}, N_k(r) \geq (1 + \varepsilon)\xi_k \log T) \\
&\leq \sum_{r=1}^{T-1} \mathbb{P}\left(N_1(r) \geq N_k(r), \hat{Y}_{k, N_k(r)} > \hat{Y}_{1, S_1^r(N_1(r), N_k(r))}, N_k(r) \geq (1 + \varepsilon)\xi_k \log T\right) \\
&\leq \frac{T}{1 - e^{-I_1(\omega_k)}} e^{-(1+\varepsilon)\xi_k I_1(\omega_k) \log T} + \frac{1}{1 - e^{-I_k(\omega_k)}} e^{-(1+\varepsilon)\xi_k I_k(\omega_k) \log T}.
\end{aligned}$$

Similarly, we obtain

$$\mathbb{P}(\mathcal{H}_k^T) \leq \frac{T}{1 - e^{-I_1(\omega)}} e^{-J_k I_1(\omega) \log T} + \frac{1}{1 - e^{-I_k(\omega)}} e^{-J_k I_k(\omega) \log T}.$$

Our objective is to bound  $T\mathbb{P}(\mathcal{H}_k^T)$  and  $\log(T)\mathbb{P}(\mathcal{G}_k^T)$  by constants. This is achieved for instance if  $T\mathbb{P}(\mathcal{H}_k^T) \xrightarrow{T \rightarrow +\infty} 0$  and  $\log(T)\mathbb{P}(\mathcal{G}_k^T) \xrightarrow{T \rightarrow +\infty} 0$ . The following conditions are sufficient to ensure these properties:

- $(1 + \varepsilon)\xi_k I_1(\omega_k) > 1$
- $(1 + \varepsilon)\xi_k I_k(\omega_k) > 0$
- $J_k I_1(\omega) > 2$
- $J_k I_k(\omega) > 1$

These conditions are met with the following values:

- $\omega = \frac{1}{2}(\mu_1 + \max_{k \neq 1} \mu_k)$
- $J_k > \max(\frac{1}{I_k(\omega)}, \frac{2}{I_1(\omega)})$
- $\mu_k < \omega_k < \mu_1$  chosen such that  $(1 + \varepsilon)I_1(\omega_k) > I_k(\omega_k)$ . We are sure that such value exists if we choose  $\omega_k$  close enough to  $\mu_k$ , thanks to the continuity of the rate functions and the fact that  $I_k(\mu_k) = 0$  and  $I_1(\mu_k) \neq 0$  (assumed in Assumption 1. of Theorem 3.1).

Choosing these values, both the terms in  $\mathcal{G}_k^T$  and  $\mathcal{H}_k^T$  in (3) are part of the constant  $C_k(\nu, \varepsilon)$  in Lemma 4.1. We can now focus on upper bounding  $\sum_{r=1}^{T-1} \mathbb{P}(\mathcal{Z}^r)$ , which is more challenging.

## D.2 Upper Bound on $\sum_{r=1}^{T-1} \mathbb{P}(\mathcal{Z}^r)$

The first steps of this part of the proof are again similar to [2]. The definition of the leader as the arm with the largest history gives the following property, that will be very useful for the analysis:

$$\ell(r) = k \Rightarrow N_k(r) \geq \left\lfloor \frac{r}{K} \right\rfloor - 1$$

So if an arm  $k$  is the leader at a given round it has been drawn a linear amount of time at this round. Intuitively, this will provide very interesting concentration guarantees for the leader after a reasonable amount of rounds, that we are going to use in this section. For every  $r \geq 8$ , we define  $a_r = \lfloor \frac{r}{4} \rfloor$  and use the decomposition

$$\mathbb{P}(\mathcal{Z}^r) = \mathbb{P}(\mathcal{Z}^r \cap \mathcal{D}^r) + \mathbb{P}(\mathcal{Z}^r \cap \bar{\mathcal{D}}^r), \quad (4)$$

where  $\mathcal{D}^r$  is the event that the optimal has been leader at least once in  $[a_r, r]$ :

$$\mathcal{D}^r = \{\exists u \in [a_r, r] \text{ such that } \ell(u) = 1\}.$$

We now explain how to upper bound the sum of the two terms in the left hand side of (4).

### D.2.1 Controlling $\mathbb{P}(\mathcal{Z}^r \cap \mathcal{D}^r)$ : arm 1 has been leader between $\lfloor r/4 \rfloor$ and $r$

We introduce a new event

$$\mathcal{B}^u = \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1 \text{ for some arm } k\}$$

If  $\mathcal{D}^r$  happens, then the event  $\mathcal{Z}^r$  can be true only if the leadership has been taken over by a sub-optimal arm at some round between  $a_r$  and  $r$ , that is

$$\mathcal{Z}^r \cap \mathcal{D}^r \subset \cup_{u=a_r}^r \{\bar{\mathcal{Z}}_u, \mathcal{Z}_{u+1}\} \subset \cup_{u=a_r}^r \mathcal{B}^u$$

We now upper bound  $\sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{P}(\mathcal{B}^u)$ . We use the notation  $b_r = \lfloor a_r/K \rfloor$ , where we recall  $a_r = \lfloor r/4 \rfloor$ . Then we write  $\mathcal{B}^u = \cup_{k=2}^K \mathcal{B}_k^u := \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1\}$ , which fixes a specific suboptimal arm. For any  $w_k$  in  $(\mu_k, \mu_1)$ , one can write

$$\begin{aligned} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{P}(\mathcal{B}_k^u) &= \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{1}(\ell(u) = 1) \mathbb{1}(k \in \mathcal{A}_{u+1}) \mathbb{1}(N_1(u) = N_k(u) + 1) \\ &\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{1}(N_1(u) \geq b_r) \mathbb{1}(\bar{Y}_{k, N_k(u)} \geq \bar{Y}_{1, S_k^u(N_1(u), N_k(u))}) \mathbb{1}(N_1(u) = N_k(u) + 1) \mathbb{1}(k \in \mathcal{A}_{u+1}) \\ &\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{1}(N_1(u) \geq b_r) \mathbb{1}(\bar{Y}_{k, N_k(u)} < w_k) \mathbb{1}(N_1(u) = N_k(u) + 1) \mathbb{1}(k \in \mathcal{A}_{u+1}) \end{aligned} \quad (5)$$

$$+ \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{1}(N_1(u) \geq b_r) \mathbb{1}(\bar{Y}_{1, S_k^u(N_1(u), N_k(u))} > w_k) \mathbb{1}(N_1(u) = N_k(u) + 1) \mathbb{1}(k \in \mathcal{A}_{u+1}) \quad (6)$$

We now separately upper bound each of these two terms. First,

$$\begin{aligned}
(5) &\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \sum_{n_k=b_r-1}^r \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1}) \mathbb{1}(\bar{Y}_{k,n_k} < w_k) \\
&\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r \mathbb{1}(\bar{Y}_{k,n_k} < w_k) \underbrace{\sum_{u=a_r}^r \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1})}_{\leq 1} \\
&\leq \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r \mathbb{P}(\bar{Y}_{k,n_k} < w_k) \\
&\leq \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r \exp(-n_k I_k(w_k)) \\
&\leq \frac{e^{(2+1/K)I_k(\omega_k)}}{(1 - e^{-I_k(\omega_k)})(1 - e^{-I_k(\omega_k)/4K})}
\end{aligned}$$

Then, letting  $[m, n]$  denote the set of subset of  $[n]$  of size  $m$ ,

$$\begin{aligned}
(6) &\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \sum_{n_k=b_r-1}^r \mathbb{1}(\bar{Y}_{1,S_k^u(n_k+1,n_k)} > w_k) \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1}) \\
&\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \sum_{n_k=b_r-1}^r \sum_{\mathcal{S} \in [n_k, n_k+1]} \mathbb{1}(\bar{Y}_{1,\mathcal{S}} > w_k) \mathbb{1}(S_k^u(n_k+1, n_k) = \mathcal{S}) \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1}) \\
&\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{u=a_r}^r \sum_{n_k=b_r-1}^r \sum_{\mathcal{S} \in [n_k, n_k+1]} \mathbb{1}(\bar{Y}_{1,\mathcal{S}} > w_k) \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1}) \\
&\leq \mathbb{E} \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r \sum_{\mathcal{S} \in [n_k, n_k+1]} \mathbb{1}(\bar{Y}_{1,\mathcal{S}} > w_k) \underbrace{\sum_{u=a_r}^r \mathbb{1}(N_k(u) = n_k) \mathbb{1}(k \in \mathcal{A}_{u+1})}_{\leq 1} \\
&\leq \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r \sum_{\mathcal{S} \in [n_k, n_k+1]} \mathbb{P}(\bar{Y}_{1,\mathcal{S}} > w_k) \\
&\leq \sum_{r=8}^{T-1} \sum_{n_k=b_r-1}^r (n_k + 1) \mathbb{P}(\bar{Y}_{1,n_k} > w_k) \\
&\leq \sum_{r=8}^{T-1} (r+1) \sum_{n_k=b_r-1}^r \exp(-n_k I_1(w_k)) \\
&\leq \frac{e^{(2+1/K)I_1(\omega_k)}}{(1 - e^{-I_1(\omega_k)})(1 - e^{-I_1(\omega_k)/4K})^2}
\end{aligned}$$

Here we have used that there are  $n_k + 1$  subsets in  $[n_k, n_k + 1]$  and that  $\mathbb{P}(\bar{Y}_{1,\mathcal{S}} > w_k) = \mathbb{P}(\bar{Y}_{1,n_k} > w_k)$  for all such subsets. Choosing  $\omega_k$  such that  $I_1(\omega_k) = I_k(\omega_k)$  (which is possible given the continuity assumptions on the two rate functions), we obtain

$$\sum_{r=8}^{T-1} \mathbb{P}(\mathcal{Z}^r \cap \mathcal{D}^r) \leq \sum_{r=8}^{T-1} \sum_{u=a_r}^r \mathbb{P}(\mathcal{B}^u) \leq \sum_{k=2}^K \frac{2e^{(2+1/K)I_1(\omega_k)}}{(1 - e^{-I_1(\omega_k)})(1 - e^{-I_1(\omega_k)/4K})^2}. \quad (7)$$

## D.2.2 Controlling $\mathbb{P}(\mathcal{Z}^r \cap \mathcal{D}^r)$ : arm 1 has not been leader between $\lfloor r/4 \rfloor$ and $r$

The idea in this part is to leverage the fact that if the optimal arm is not leader between  $\lfloor s/4 \rfloor$  and  $s$ , then it has necessarily lost a lot of duels against the current leader at each round. We then use the

fact that when the leader has been drawn "enough", concentration prevents this situation with large probability. We introduce

$$\mathcal{L}^r = \sum_{u=s_0}^r \mathbb{1}_{\mathcal{C}^u}$$

for the event  $\mathcal{C}^u = \{\exists a \neq 1, N_a(u) \geq N_1(u), \hat{Y}_{a, S_1^u(N_a(u), N_1(u))} \geq \hat{Y}_{1, N_1(u)}\}$ . One can prove the following inequality:

$$\mathbb{P}(\mathcal{Z}^r \cap \bar{\mathcal{D}}^r) \leq \mathbb{P}(\mathcal{L}^r \geq r/4).$$

*Proof.* Under  $\bar{\mathcal{D}}^r$  arm 1 is a challenger for every round  $u \in [a_r, r]$ . Then, each time  $\mathcal{C}^u$  is not true arm 1 wins its duel against the current leader and is pulled. Hence, if  $\{\mathcal{L}_r < r/4\}$  then we necessarily have  $\{N_1(r) > r/2\}$  and arm 1 is leader in round  $r$ . Hence,  $\{\mathcal{Z}^r \cap \bar{\mathcal{D}}^r\} \cap \{\mathcal{L}_r < r/4\} = \emptyset$ , which justifies the inequality.  $\square$

Now, as in [2] we use the Markov inequality to get:

$$\mathbb{P}(\mathcal{L}^r \geq r/4) \leq \frac{\mathbb{E}(\mathcal{L}^r)}{r/4} = \frac{4}{r} \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(\mathcal{C}^u).$$

By further decomposing the probability of  $\mathbb{P}(\mathcal{C}^u)$  in two parts depending on the value of the number of selections of arm 1, we obtain the upper bound

$$\mathbb{P}(\mathcal{Z}^r \cap \bar{\mathcal{D}}^r) \leq \frac{4}{r} \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(N_1(u) \leq (\log u)^2) + \underbrace{\frac{4}{r} \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(\mathcal{C}^u, N_1(u) \geq (\log u)^2)}_{B_r}.$$

We now upper bound the quantity  $B_r$  defined above by using Lemma 4.2. For each  $a$ , for any  $\omega_a$  such that  $\omega_a \in (\mu_a, \mu_1)$ , one can write

$$\begin{aligned} B_r &\leq \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(\mathcal{C}^u, N_1(u) \geq (\log \lfloor r/4 \rfloor)^2) \\ &\leq \sum_{a=2}^K \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(Y_{a, S_1^u(N_a(u), N_1(u))} > \hat{Y}_{1, N_1(u)}, N_1(u) \geq \log(\lfloor r/4 \rfloor)^2, N_a(u) > N_1(u)) \\ &\leq \sum_{a=2}^K \left( \frac{1}{1 - e^{-I_1(\omega_a)}} e^{-(\log \lfloor r/4 \rfloor)^2 I_1(\omega_a)} + \frac{r}{1 - e^{-I_k(\omega_a)}} e^{-(\log \lfloor r/4 \rfloor)^2 I_a(\omega_a)} \right). \end{aligned}$$

Choosing each  $\omega_a$  such that  $I_1(\omega_a) = I_a(\omega_a)$ , we obtain:

$$\begin{aligned} \frac{4}{r} \sum_{r=8}^T B_r &\leq \sum_{r=8}^T \sum_{a=2}^K \frac{4(r+1)}{r(1 - e^{-I_1(\omega_a)})} e^{-(\log \lfloor r/4 \rfloor)^2 I_a(\omega_a)} \\ &\leq \sum_{a=2}^K \sum_{r=8}^T \frac{6}{1 - e^{-I_1(\omega_a)}} e^{-(\log \lfloor r/4 \rfloor)^2 I_a(\omega_a)}, \end{aligned}$$

and for each  $a$  the series in  $r$  is convergent as for any constant  $C$ ,  $C \log(r) \leq (\log \lfloor r/4 \rfloor)^2$  for  $r$  large enough. Hence, there exists some constant  $D(\nu)$  where  $\nu = (\nu_1, \dots, \nu_K)$  such that  $\frac{4}{r} \sum_{r=8}^T B_r \leq D(\nu)$ . It follows that

$$\sum_{r=8}^T \mathbb{P}(\mathcal{Z}^r \cap \bar{\mathcal{D}}^r) \leq \sum_{r=8}^T \frac{4}{r} \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(N_1(u) \leq (\log u)^2) + D(\nu).$$

We now transform the double sum in the right-hand side into a simple sum by counting the number of times each term appears in the double sum:

$$\sum_{r=8}^T \frac{4}{r} \sum_{u=\lfloor r/4 \rfloor}^r \mathbb{P}(N_1(u) \leq (\log u)^2) = \sum_{r=8}^T \left( \sum_{t=1}^r \frac{4}{t} \mathbb{1}(t \in [r, 4r]) \right) \mathbb{P}(N_1(r) \leq (\log r)^2).$$

Noting that  $\sum_{t=1}^r \frac{4}{t} \mathbb{1}(t \in [s, 4s]) \leq (4s - s + 1) \times \frac{4}{s} \leq 16$ , we finally obtain:

$$\sum_{r=8}^T \mathbb{P}(\mathcal{Z}^r \cap \overline{\mathcal{D}}^r) \leq 16 \sum_{r=1}^T \mathbb{P}(N_1(r) \leq (\log(r))^2) + D(\boldsymbol{\nu}). \quad (8)$$

Combining (7) and (8) yields

$$\sum_{r=1}^T \mathbb{P}(\mathcal{Z}^r) \leq 16 \sum_{r=1}^T \mathbb{P}(N_1(r) \leq (\log(r))^2) + D'_k(\boldsymbol{\nu})$$

for some constant  $D'_k(\boldsymbol{\nu})$  that depends on  $k$  and  $\boldsymbol{\nu}$ , which contributes to the final constant  $C_k(\boldsymbol{\nu}, \varepsilon)$  in Lemma 4.1. Plugging this inequality in Equation (3) concludes the proof of Lemma 4.1.

## E Probability that the Optimal Arm is not Drawn Enough: Proof of Lemma 4.3

We start with a decomposition that follows the steps of [1] for BESA with 2 arms that we generalize for  $K$  arms.

We first denote by  $r_j$  the round of the  $j^{\text{th}}$  play of arm 1 with  $r_0 = 0$  and let  $\tau_j = r_{j+1} - r_j$ . We notice that  $\tau_0 \leq K$  as all arms are initialized once. Then:

$$\begin{aligned} \mathbb{P}(N_1(r) \leq (\log r)^2) &\leq \mathbb{P}(\exists j \in \{1, \dots, \log r^2\} : \tau_j \geq r/(\log r)^2 - 1) \\ &\leq \sum_{j=1}^{(\log r)^2} \mathbb{P}(\tau_j \geq r/(\log r)^2 - 1) \end{aligned}$$

*Proof.* If we assume that  $\forall j \tau_j \leq r/(\log r)^2 - 1$  then  $t_{\log r^2} = \sum_{j=0}^{\log r^2} \tau_j < r$ , which yields  $N_{\ell}(r) > \log r^2 + 1$ .  $\square$

We now fix  $j \leq (\log r)^2$  and upper bound the probability of the event

$$\mathcal{E}_j := \{\tau_j \geq r/\log r^2 - 1\}.$$

On this event arm 1 lost at least  $r/\log r^2$  consecutive duels between  $r_j + 1$  and  $r_{j+1}$  (either as a challenger or as the leader) which yields

$$\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}(\forall s \in \{r_j + 1, \dots, r_j + \lfloor r/\log r^2 - 1 \rfloor\} : \{\hat{Y}_{1,j} \leq \hat{Y}_{\ell(s), S_1^*(N_{\ell(s)}(s), j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\} \cup \{\ell(s) = 1, N_1(s) = j\})$$

The important change compared to the proof of [1] is that with  $K > 2$ , 1) we don't know the identity of the leader and 2) the leader is not necessarily pulled if it wins its duel against 1.

Now we notice that when  $r$  is large, the time range considered in  $\mathcal{E}_j$  is large. By looking at the second half of this time range only, we can ensure that the leader has been drawn a large number of times. More precisely, introducing the two intervals

$$\begin{aligned} \mathcal{M}_{j,r}^1 &= \left[ r_j + 1, r_j + \left\lfloor \frac{r/\log r^2 - 1}{2} \right\rfloor \right] \\ \mathcal{M}_{j,r}^2 &= \left[ t_j + \left\lfloor \frac{r/\log r^2 - 1}{2} \right\rfloor, t_j + \lfloor r/\log r^2 \rfloor - 1 \right] \end{aligned}$$

it holds that

$$\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}(\forall s \in \mathcal{M}_{j,r}^2 : \{\hat{Y}_{1,j} \leq \hat{Y}_{\ell(s), S_1^*(N_{\ell(s)}(s), j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\} \cup \{\ell(s) = 1, N_1(s) = j\}).$$

But we know that on  $\mathcal{M}_{j,r}^2$  the leader must have been selected at least  $\frac{1}{K} \left( j + \left\lfloor \frac{r/\log r^2 - 1}{2} \right\rfloor \right)$  times.

Let  $r_K$  be the first integer such that  $\log^2(r) < \frac{1}{K-1} \left\lfloor \frac{r/\log r^2 - 1}{2} \right\rfloor$ , for every  $r \geq r_K$ , as  $j \leq \log^2(r)$ ,

the leader has been selected strictly more than  $j$  times, which prevents arm 1 from being the leader for any round in  $\mathcal{M}_{j,r}^2$ . Hence, for  $r \geq r_K$ , for all  $j \leq \log^2(r)$ ,

$$\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}\left(\forall s \in \mathcal{M}_{j,r}^2 : \{\hat{Y}_{1,j} \leq \hat{Y}_{\ell(s), S_1^s(N_{\ell(s)}(s), j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\}\right).$$

To remove the problem of the identity of the leader we would like to find a way to fix our attention on one arm. To this extent, we notice that during an interval of length  $|\mathcal{M}_{j,r}^2|$ , if there are only  $K - 1$  candidates for the leader then one of them must have been leader at least  $m_r := |\mathcal{M}_{j,r}^2|/(K - 1) - 1$  times during this range. We also know that at any round in  $\mathcal{M}_{j,r}^2$ , the leader satisfies  $N_{\ell(s)}(s) \geq (t_j + \lfloor \frac{r/\log r^2 - 1}{2} \rfloor)/K - 1 \geq (\lfloor \frac{r/\log r^2 - 1}{2} \rfloor)/K - 1 = \frac{|\mathcal{M}_{j,r}^1|}{K} - 1 := c_r$ . Observe that  $m_r > c_r$ . Finally, we introduce the notation

$$I_{j,r}^k = \{s \in \mathcal{M}_{j,r}^2 : \ell(s) = k\}$$

for the set of rounds in  $\mathcal{M}_{j,r}^2$  in which a particular arm  $k$  is leader. From the above discussion, we know that there exists an arm  $k$  such that  $|I_{j,r}^k| \geq m_r$ .

To ease the notation, we introduce the event

$$\mathcal{W}_{s,j}^k = \left\{ \left\{ \hat{Y}_{1,j} < \hat{Y}_{k, S_1^s(N_k(s), j)} \right\}, N_k(s) \geq c_r, N_1(s) = j \right\}$$

and write

$$\begin{aligned} \mathbb{P}(\mathcal{E}_j) &\leq \mathbb{P}\left(\bigcap_{s \in \mathcal{M}_{j,r}^2} \bigcup_{k=2}^K \{\ell(s) = k, 1 \notin \mathcal{A}_s\}\right) \\ &\leq \mathbb{P}\left(\bigcap_{k=2}^K \bigcap_{s \in I_{j,r}^k} \mathcal{W}_{s,j}^k\right) \\ &\leq \mathbb{P}\left(\bigcup_{k=2}^K \left\{ |I_{j,r}^k| > m_r, \bigcap_{s \in I_{j,r}^k} \mathcal{W}_{s,j}^k \right\}\right) \\ &\leq \sum_{k=2}^K \mathbb{P}\left(|I_{j,r}^k| > m_r, \bigcap_{s \in I_{j,r}^k} \mathcal{W}_{s,j}^k\right). \end{aligned}$$

Finally, we define for any integer  $M$  the event that we can find  $M$  pairwise non-overlapping sub-samples in the set of the sub-samples of arm  $k$  drawn in rounds  $s \in I_{j,r}^k$ :

$$\mathcal{F}_{j,M}^{k,r} = \left\{ \exists i_1, \dots, i_M \in I_{j,r}^k : \forall m < m' \in [M], S_1^{i_m}(N_k(i_m), j) \cap S_1^{i_{m'}}(N_k(i_{m'}), j) = \emptyset \right\}$$

Introducing  $H_{j,r}^k = \min_{s \in I_{j,r}^k} N_k(s)$ , the minimal size of the history of arm  $k$  during rounds in  $I_{j,r}^k$  (which is known to be larger than  $c_r$  as  $k$  is leader in these rounds), one has

$$\begin{aligned} \mathbb{P}(\mathcal{E}_j) &\leq \sum_{k=2}^K \mathbb{P}\left(|I_{j,r}^k| > m_r, \bigcap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \{\mathcal{F}_{j,M}^{k,r} \cup \bar{\mathcal{F}}_{j,M}^{k,r}\}\right) \\ &\leq \sum_{k=2}^K \mathbb{P}\left(|I_{j,r}^k| \geq m_r, H_{j,r}^k \geq c_r, \bar{\mathcal{F}}_{j,M}^{k,r}\right) + \sum_{k=2}^K \mathbb{P}\left(|I_{j,r}^k| > m_r, \bigcap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \mathcal{F}_{j,M}^{k,r}\right) \quad (9) \end{aligned}$$

**Upper bound on the first term in (9)** The probability  $\mathbb{P}\left(|I_{j,r}^k| \geq m_r, H_{j,r}^k \geq c_r, \bar{\mathcal{F}}_{j,M}^{k,r}\right)$  can be upper bounded by

$$\mathbb{P}\left(\#\left\{\text{pairwise non-overlapping subsets in } (S_1^s(N_k(s), j))_{s \in I_{j,r}^k}\right\} < M \mid \{|I_{j,r}^k| > m_r, H_{j,r}^k \geq c_r\}\right).$$

This probability can be related to some intrinsic properties of the sampler  $\text{SP}(H, j)$ . To formalize this, we introduce the following definition.



**Definition.** For every integers  $N, H, j$  such that  $H > j$ ,  $X_{N,H,j}$  is a random variable which counts the maximum number of non-overlapping subsets among  $N$  i.i.d. samples from  $\text{SP}(H, j)$ .

Letting  $H_1, \dots, H_{m_r}$  be integers that are all larger than  $c_r$ , and letting  $S_1, \dots, S_{m_r}$  be independent subsets such that  $S_i \sim \text{SP}(H_i, j)$ , the above probability is upper bounded by

$$\mathbb{P}(\#\{\text{pairwise non-overlapping subsets in } (S_i)_{i=1}^{m_r}\} < M)$$

which is itself upper bounded by  $\mathbb{P}(X_{m_r, c_r, j} < M)$ .

This last inequality is quite intuitive: if one draws subsets of size  $j$  from histories that may be larger than  $c_r$ , there is more ‘‘room’’ for non-overlapping subsets than if we always draw them from the same history of size  $c_r$ . For Random Block sampling, where the drawn subset is fully determined by the random position of its first element, to formalize this intuition it is sufficient to prove that if  $X_i, Y_i$  are two sequences of random variables such that  $X_i$  is uniform in  $[H_i - j]$  and  $Y_i$  is uniform in  $[H - j]$ , where  $H_i \geq H$ , the random variable that counts the maximal number of elements in the sequence  $(Y_i)$  whose pairwise distance are larger than  $j$  is stochastically dominated by that the same random variable but for the sequence  $(X_i)$ . We performed numerical experiments that confirm that this last condition holds.

**Upper bound on the second term in (9)** On the event  $(|I_{j,r}^k| > m_r, \cap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \mathcal{F}_{j,M}^{k,r})$ , one can define  $\tilde{i}_1, \dots, \tilde{i}_M$  the first  $M$  rounds in  $I_{j,r}^k$  for which the subsets  $\tilde{S}_m := S^{\tilde{i}_m}(N_k(\tilde{i}_m), j)$  are pairwise non-overlapping and we get

$$\mathbb{P}\left(|I_{j,r}^k| > m_r, \cap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \mathcal{F}_{j,M}^{k,r}\right) \leq \mathbb{P}\left(\forall m \in [M], \hat{Y}_{1,j} \leq \hat{Y}_{k, \tilde{S}_m}\right).$$

By definition the subsets  $\tilde{S}_m$  are pairwise non-overlapping, hence the sub-samples  $\hat{Y}_{k, \tilde{S}_m}$  are independent. We prove that this probability can be in fact upper bound by the *balance function* we defined in section 3.

Indeed, introducing  $X \sim \nu_{1,j}$  and an independent i.i.d. sequence  $Z_i \sim \nu_{k,j}$ , one can write

$$\begin{aligned} \mathbb{P}\left(|I_{j,r}^k| > m_r, \cap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \mathcal{F}_{j,M}^{k,r}\right) &\leq \mathbb{P}(X < \min_{i \in [M]} Z_i) \\ &= \mathbb{E}_{X \sim \nu_{1,j}} \left[ \prod_{i=1}^M \mathbb{1}_{X \leq Z_i} \right] \\ &= \mathbb{E}_{X \sim \nu_{1,j}} \left[ \mathbb{E}_{Z \sim \nu_{k,j}^{\otimes j}} \left[ \prod_i \mathbb{1}_{X \leq Z_i} \middle| X \right] \right] \\ &= \mathbb{E}_{X \sim \nu_{1,j}} \left[ (1 - F_{k,j}(X))^M \right] \\ &= \alpha_k(M, j). \end{aligned}$$

**Conclusion** Putting things together, we have proved that

$$\mathbb{P}(\mathcal{E}_j) \leq (K - 1)\mathbb{P}(X_{m_r, c_r, j} < M) + \sum_{k=2}^K \alpha_k(M, j),$$

where  $X_{N,H,j}$  and  $\alpha_k(M, j)$  are introduced in Definition E and 3 respectively. If we replace  $M$  by the sequence  $\beta_{r,j}$  we have

$$\begin{aligned} \sum_{r=1}^T \mathbb{P}(N_1(r) \leq \log r^2) &\leq r_K + \sum_{r=r_K}^T \sum_{j=1}^{\log r^2} \left[ (K - 1)\mathbb{P}(X_{m_r, c_r, j} < \beta_{r,j}) + \sum_{k=2}^K \alpha_k(\beta_{r,j}, j) \right] \\ &\leq r_K + \sum_{r=r_K}^T \sum_{j=1}^{\log r^2} \left[ (K - 1)\mathbb{P}(X_{c_r, c_r, j} < \beta_{r,j}) + \sum_{k=2}^K \alpha_k(\beta_{r,j}, j) \right] \end{aligned}$$

as  $c_r \leq m_r$ , which proves Lemma 4.3.

This definition allows to analyze separately the properties of the sub-sampling algorithms and the properties of the distribution family for randomized samplers.

## F Proof that RB-SDA Satisfies the Diversity Property

We recall that  $X_{m,H,j}$  denotes the maximal number of pairwise non-overlapping subsets obtained in  $m$  i.i.d. samples from  $\text{RB}(H, j)$ . In this section we aim at upper bounding the probability of

$$\mathbb{P}(X_{m,H,j} \leq \gamma r / (\log r)^2)$$

for some values of  $m, H, j$ , that will be fixed later. This probability depends on several parameters, with straightforward effects:

- The probability decreases with the length of the history size  $H$ .
- The probability increases with the size  $j$  of each sub-sample.
- The probability decreases with the total number of sub-samples we draw  $m$ . Intuitively if  $m$  is large enough every sample of size  $j$  in the history will be drawn.

**First step with  $j = 1$ :** in this case the distribution of the  $m$  subsets of size 1 is actually the distribution of sampling with replacement in  $H$ . The question of the number of different items drawn with sampling without replacement has been studied in [25], from which we use the following result:

**Result 1:** for any  $k \in [H]$ ,  $\mathbb{P}(X_{m,H,1} = k) = \frac{H!}{(H-k)! \times H^m} \times S_{k,m}$ , where  $S_{k,m}$  is the Stirling number of the second kind for  $k, m$ .

We use this result with further assumptions that are specific to our problem and will ease the computation:  $H = m = O(r/(\log r)^2)$ . To ease the notation we continue to use  $H$ , and write  $\gamma t / (\log t)^2 = \alpha H$  for some  $\alpha \in (0, 1)$ .

We first look at  $\mathbb{P}(X_{H,H,1} = \alpha H)$ . According to [26] the following inequality holds

$$S_{k,H} \leq \frac{1}{2} \binom{H}{k} k^{H-k},$$

This allows to upper bound the expression in result 1:

$$\mathbb{P}(X_{H,H,1} = k) \leq \frac{1}{2} \left(\frac{k}{H}\right)^{H-k} \binom{H}{k}$$

We now want to bound  $\binom{H}{k}$ . As  $k$  is small compared with  $H$ , it is natural to use

$$\binom{H}{k} \leq \frac{H^k}{k!}$$

We then bound  $1/k!$  by its Stirling approximation and add a multiplicative constant  $c$  along the way:

$$\binom{H}{k} \leq c \frac{H^k}{\sqrt{2\pi k} \times k^k} e^k$$

Refactoring provides

$$\mathbb{P}(X_{H,H,1} = k) \leq \frac{c}{2} \left(\frac{k}{H}\right)^{H-2k} \frac{e^k}{\sqrt{2\pi k}}$$

Then we notice that if  $k \leq H - 2k$  we get:

$$\mathbb{P}(X_{H,H,1} = k) \leq \frac{c}{2\sqrt{2\pi k}} \left(\frac{ke}{H}\right)^{H-2k}$$

Now we can replace  $k$  by  $\alpha H$  (assume it's an integer for the simplicity of notations), such that  $1) \alpha \leq \frac{1}{3} \Rightarrow H(1 - 3\alpha) > 0$  and  $\alpha e < 1$ .

If  $k \leq \alpha H$ ,  $\alpha e \leq 1$  and  $(1 - 2\alpha) > 0$  then:  $(\frac{ke}{H})^{H-2k} \leq (\alpha e)^{H-2k}$ . We have:

$$\begin{aligned}
\mathbb{P}(X_{H,H,1} \leq \alpha H) &\leq \frac{c}{2\sqrt{2\pi}} \sum_{k=0}^{\lfloor \alpha H \rfloor} (\alpha e)^{H-2k} \\
&\leq \frac{c}{2\sqrt{2\pi}} \sum_{k=0}^{\lfloor \alpha H \rfloor} (\alpha e)^{H-2(\lfloor \alpha H \rfloor - k)} \\
&\leq \frac{c}{2\sqrt{2\pi}} (\alpha e)^{H-2\lfloor \alpha H \rfloor} \frac{1}{1 - (\alpha e)^2} \\
&\leq \frac{c}{2\sqrt{2\pi}} \frac{1}{1 - (\alpha e)^2} \exp(-(1 - 2\alpha)H \log(1/(\alpha e)))
\end{aligned} \tag{10}$$

**From  $X_{H,H,1}$  to  $X_{H,H,j}$**  This result is enough to get general properties for Random Block Sampling. Indeed, as the process of RBS consists in only drawing the first element of the block used in the duel we can see that the previous bound also applies to the number of unique starting points. With this property, the Random Block Sampler satisfies for all  $x > 0$ :

$$\mathbb{P}\left(X_{m,H,j} \leq \left\lfloor \frac{x}{j} \right\rfloor\right) \leq \mathbb{P}(X_{m,H,1} \leq x)$$

*Proof.* Assume that the Random Block Sampler provides  $x$  blocks with different starting points. Let's further assume that  $x$  is an integer and try to identify the sequence of starting times  $t_i = (t_1, \dots, t_x)$  that minimizes the number of mutually non-overlapping samples: the value of  $t_1$  is not important due to the symmetry of the problem. Then if we want to reduce the possibilities to get non-overlapping sample we want to choose a value for  $t_2$  that 1) makes the blocks  $[t_1, t_1 + j]$  and  $[t_2, t_2 + k]$  non-overlapping and 2) makes things easier to continue this process for  $t_3, \dots, t_m$ . It seems intuitive to choose either the block starting at  $t_1 + 1$  or at  $t_1 - 1$  as we cover the minimum amount of space with the constraint that  $t_2 \neq t_1$ . If we repeat this choice until  $m$  blocks are chosen and reorder the blocks properly, we get a sequence of starting points  $[t_1, t_1 + 1, \dots, t_1 + m]$  that are all different and minimize the total amount of space covered by the block. Even in this setup, we can find exactly  $\lfloor \frac{m}{j} \rfloor$  mutually non-overlapping blocks as for instance all  $[t_1 + kj, t_1 + (k+1)j - 1]$ ,  $[t_1 + k'j, t_1 + (k'+1)j - 1]$  blocks are non-overlapping for  $k \neq k'$  and  $(k, k') \in [0, \lfloor \frac{m}{j} \rfloor - 1]$ .  $\square$

We can finally prove the following for Random Block sampling.

**Lemma F.1** (Diversity Property for Random Block Sampling). *If we choose a constant  $\gamma \leq 1/3 \times \frac{1}{2K}$  then Random Block Sampling satisfies the diversity property.*

*Proof.* For  $\gamma \leq \lfloor 1/3 \times \frac{1}{2K} \rfloor$ , there exists  $\alpha > 0$  such that:

$$\begin{aligned}
\mathbb{P}(X_{c_r, c_r, j} \leq \gamma/j(r/(\log r)^2)) &\leq \mathbb{P}(X_{c_r, c_r, j} \leq \alpha/jc_r) \\
&\leq \mathbb{P}(X_{c_r, c_r, 1} \leq \alpha c_r) \\
&= o(r^{-2})
\end{aligned}$$

The last line comes from the expression obtained in Equation (10), and allows to conclude that  $\sum_{r=1}^T \sum_{j=1}^{(\log r)^2} \mathbb{P}(X_{c_r, c_r, j} \leq \alpha/j(r/(\log r)^2)) = o(\log T)$   $\square$

## G Analysis of the Balance Function for Some Distributions

For the simplicity of the notation we write the balance function  $\alpha(M, j)$  for any distribution and any instance of these distributions. The family of distributions and the notation for their parameter will always be mentioned at the beginning of the corresponding sub-section.

In the next parts we use the notation  $G(x) = 1 - F(x)$  where  $F$  denotes the CDF of some distribution. For some arm distribution  $\nu_i$  the distribution of the sum of  $j$  independent observations drawn from  $\nu_i$  is denoted by  $\nu_{i,j}$ . With this notation, for two arms 1 and 2 we write:

$$\alpha(M, j) = \mathbb{E}_{Z \sim \nu_{1,j}}(G_{2,j}(Z)^M)$$

### G.1 The Bernoulli Distribution is Balanced

We prove the following lemma, which bears strong similarity with an upper bound given by [10] for a similar quantity in their analysis of Thompson Sampling.

**Lemma G.1** (Bound on  $\alpha(M, j)$ ). *For two Binomial distributions  $\nu_1 \sim \mathcal{B}(j, \mu_1)$  and  $\nu_2 \sim \mathcal{B}(j, \mu_2)$  such that  $\mu_1 > \mu_2$  and for any integer  $M > 1$ :  $\exists \lambda > 1$  such as*

$$\mathbb{E}_{X \sim \nu_{1,j}} \left( (1 - F_{j,\mu_2}(X))^M \right) \leq C_{\lambda_0, \lambda} \frac{1}{M^\lambda} e^{-j d_{\lambda, \mu_1, \mu_2}} + \left( \frac{1}{2} \right)^M$$

Where  $C_{\lambda, \mu_1, \mu_2} > 0$ , and  $F_{j, \mu_2}$  is the CDF of a Binomial  $\mathcal{B}(j, \mu_2)$ .

*Proof.* We use the same notation as before:  $G_2(k) = 1 - F_2(k)$  and  $f_1, f_2$  as the PMF of  $\nu_1, \nu_2$ .

We first use a common property of Binomial distributions,  $\forall k > \lceil j\mu_2 \rceil$ :  $G(k) \leq \frac{1}{2}$ . So we can directly write:

$$\mathbb{E}_{X \sim \nu_1} \left( (1 - F_{j,\mu_2}(X))^M \right) \leq \left( \frac{1}{2} \right)^M + \underbrace{\sum_{k=0}^{\lfloor j\mu_2 \rfloor} f_1(k) G_2(k)^M}_{(A)}$$

Using convexity we get:  $G(k)^M \leq \exp(-MF_2(k))$ , hence

$$(A) \leq \sum_{k=0}^{\lfloor j\mu_2 \rfloor} f_1(k) \exp(-MF_2(k))$$

Then we use that for  $\lambda > 1, \forall x > 0$ :  $x^\lambda e^{-x} \leq \left( \frac{\lambda}{e} \right)^\lambda = C_\lambda$ , so:

$$(A) \leq \frac{C_\lambda}{M^\lambda} \sum_{k=0}^{\lfloor j\mu_2 \rfloor} \frac{f_1(k)}{F_2(k)^\lambda} \leq \frac{C_\lambda}{M^\lambda} \sum_{k=0}^{\lfloor j\mu_2 \rfloor} \frac{f_1(k)}{f_2(k)^\lambda}$$

As in [10], we compute:

$$\begin{aligned} \frac{f_1(k)}{f_2(k)^\lambda} &\leq \frac{\mu_1^k (1 - \mu_1)^{j-k}}{\mu_2^{\lambda k} (1 - \mu_2)^{\lambda(j-k)}} \\ &\leq \left( \frac{1 - \mu_1}{(1 - \mu_2)^\lambda} \right)^j \left( \frac{\mu_1 (1 - \mu_2)^\lambda}{\mu_2^\lambda (1 - \mu_1)} \right)^k \\ &= \left( \frac{1 - \mu_1}{(1 - \mu_2)^\lambda} \right)^j R_\lambda(\mu_1, \mu_2)^k \end{aligned}$$

with  $R_\lambda(\mu_1, \mu_2) = \frac{\mu_1(1-\mu_2)^\lambda}{\mu_2^\lambda(1-\mu_1)}$ . We then notice that we can choose  $\lambda > 1$  such that  $R_\lambda(\mu_1, \mu_2) > 1$ . It is true for any  $\lambda > 1$  if  $\mu_2 \leq 0.5$ , and for  $1 < \lambda < \log \frac{\mu_1}{1-\mu_1} / \log \frac{\mu_2}{1-\mu_2}$  if  $\mu_2 > 0.5$ .

Plugging that expression into the sum gives:

$$\begin{aligned}
(A) &\leq \frac{C_\lambda}{M^\lambda} \sum_{k=0}^{\lceil j\mu_2 \rceil} \frac{f_1(k)}{f_2(k)^\lambda} \\
&\leq \frac{C_\lambda}{M^\lambda} \left( \frac{1-\mu_1}{(1-\mu_2)^\lambda} \right)^j \sum_{k=0}^{\lceil j\mu_2 \rceil} R_\lambda(\mu_1, \mu_2)^k \\
&= \frac{C_\lambda}{M^\lambda} \left( \frac{1-\mu_1}{(1-\mu_2)^\lambda} \right)^j \frac{R_\lambda(\mu_1, \mu_2)^{\lfloor j\mu_2 \rfloor + 1} - 1}{R_\lambda(\mu_1, \mu_2) - 1} \\
&\leq \frac{C_\lambda}{M^\lambda} \left( \frac{1-\mu_1}{(1-\mu_2)^\lambda} \right)^j \frac{R_\lambda(\mu_1, \mu_2)}{R_\lambda(\mu_1, \mu_2) - 1} R_\lambda(\mu_1, \mu_2)^{j\mu_2} \\
&= \frac{C_\lambda}{M^\lambda} \frac{R_\lambda(\mu_1, \mu_2)}{R_\lambda(\mu_1, \mu_2) - 1} \left( \frac{1-\mu_1}{(1-\mu_2)^\lambda} \right)^{j(1-\mu_2)} \left( \frac{\mu_1}{\mu_2^\lambda} \right)^{j\mu_2} \\
&= \frac{C_\lambda}{M^\lambda} \frac{R_\lambda(\mu_1, \mu_2)}{R_\lambda(\mu_1, \mu_2) - 1} e^{-jd_\lambda(\mu_2, \mu_1)} \\
&= \frac{C_{\lambda, \mu_1, \mu_2}}{M^\lambda} e^{-jd_\lambda(\mu_2, \mu_1)}
\end{aligned}$$

where  $d_\lambda(\mu_2, \mu_1) = \lambda(\mu_2 \log \mu_2 + (1-\mu_2) \log(1-\mu_2)) - (\mu_2 \log \mu_1 + (1-\mu_2) \log(1-\mu_1)) = \text{KL}(\mu_2, \mu_1) - (\lambda-1)\text{H}(\mu_2)$ ,  $\text{KL}(\mu_2, \mu_1)$  denotes the KL-divergence between  $\nu_2$  and  $\nu_1$ , and  $\text{H}(\mu_2) = \mathbb{E}_{X \sim \nu_2, j}(\log f_2(X))$ . We need to choose  $\lambda$  as:

$$\lambda < 1 + \frac{\text{KL}(\mu_2, \mu_1)}{\text{H}(\mu_2)} = \lambda_0(\mu_1, \mu_2)$$

Note that those quantities correspond to the Bernoulli distributions, the  $j$  is not involved here. In [10], the authors explain that this condition is more restrictive than the previous one so we can state that  $\forall \lambda < \lambda_0(\mu_1, \mu_2)$ :

$$\mathbb{E}_{X \sim \nu_{1,j}} \left( (1 - F_{j, \mu_2}(X))^M \right) \leq \left( \frac{1}{2} \right)^M + \frac{C_{\lambda, \mu_1, \mu_2}}{M^\lambda} e^{-jd_\lambda(\mu_2, \mu_1)}$$

□

This is enough to prove that the Bernoulli distribution is balanced by replacing  $M$  by  $\lfloor \beta t / (\log t)^2 \rfloor$  in the expression in Lemma G.1 and summing on  $t$  and  $j$ . The power term is in  $o(t(\log t)^2)$ , while the other term is the term of a convergent geometric series in  $j$  multiplied by a term in  $o(1/t)$  in  $t$ , which is enough to get the result.

## G.2 The Poisson Distribution is Balanced

We can actually use the same sketch of proof as for Bernoulli distributions, using that for 2 Poisson random variables:

$$\frac{p_{1,j}(k)}{p_{2,j}(k)^\lambda} = e^{-j(\theta_1 - \lambda\theta_2)} \binom{k!}{n^k}^\lambda \left( \frac{\theta_1}{\theta_2^\lambda} \right)^k \leq e^{-j(\theta_1 - \lambda\theta_2)} \left( \frac{\theta_1}{\theta_2^\lambda} \right)^k$$

So:

$$\begin{aligned}
\sum_{k=0}^{d_{0,j}} \frac{p_{1,j}(k)}{p_{2,j}(k)^\lambda} &\leq e^{-j(\theta_1 - \lambda\theta_2)} \sum_{k=0}^{d_{0,j}} \left( \frac{\theta_1}{\theta_2^\lambda} \right)^k \\
&\leq \frac{\theta_2^\lambda}{|\theta_1 - \theta_2^\lambda|} e^{-j(\theta_1 - \lambda\theta_2)} \times \max \left\{ 1, \left( \frac{\theta_1}{\theta_2^\lambda} \right)^{d_{0,j}} \right\},
\end{aligned}$$

where  $d_{0,j} = \theta_2 j - 1$ . Now we remark that if we choose  $\lambda \in (1, \theta_1/\theta_2)$  we have 2 possibilities: 1) we can choose  $\lambda$  such that the second term equals one, hence we can bound the whole term by a constant without further conditions, or 2)  $\forall \lambda \in (0, \theta_1, \theta_2)$ :  $\left(\frac{\theta_1}{\theta_2^\lambda}\right) > 1$ . Let us focus on the second case, we study the term  $e^{-j((\theta_1 - \lambda\theta_2) - \theta_2(\log \theta_1 - \lambda \log \theta_2))}$ . As for Bernoulli distributions, we identify the KL-divergence between  $\nu_2$  and  $\nu_1$  and write:

$$(\theta_1 - \lambda\theta_2) - \theta_2(\log \theta_1 - \lambda \log \theta_2) = \text{KL}(\nu_2, \nu_1) - (\lambda - 1)\theta_2(1 - \log \theta_2)$$

So if  $\log \theta_2 > 1$  we can choose any  $\lambda > 1$ . In the other case we have to restrict our choice of  $\lambda$  to get:

$$\lambda < 1 + \frac{\text{KL}(\nu_2, \nu_1)}{\theta_2(1 - \log(\theta_2))}$$

So with an appropriate choice for  $\lambda$  Poisson distributions are balanced with the same argument that makes Bernoulli distributions balanced.

### G.3 The Gaussian Distribution is Balanced

For the Gaussian distribution we leverage the fact that both the PDF and CDF of any Gaussian distribution can be expressed with the PDF and CDF of the standard normal distribution. With such decomposition, we can express  $\alpha(M, j)$  as a function of these CDF/PDF and use some properties of the normal distribution.

We use the notations  $f$  and  $G$  for the PDF and CDF of the  $\mathcal{N}(0, 1)$  distribution,  $\Delta$  for the gap between the two arms, and compute the expectation:

$$\begin{aligned} \alpha(M, j) &= \int_{-\infty}^{+\infty} f_{1,j}(x) G_{2,j}(x)^M dx \\ &\leq \int_{-\infty}^z f_{1,j}(x) G_{2,j}(x)^M dx + G_{2,j}(z)^M, \forall z \in \mathbb{R} \\ &\leq \int_{-\infty}^z f\left(\frac{x - \mu_{1,j}}{\sqrt{j}}\right) G\left(\frac{x - \mu_{2,j}}{\sqrt{j}}\right)^M dx + G_{2,j}(z)^M \\ &\leq \int_{-\infty}^{\frac{z - \mu_{2,j}}{\sqrt{j}}} f(y - \sqrt{j}\Delta) G(y)^M dy + G_{2,j}(z)^M \end{aligned}$$

At this step we use two things: 1) the normal distribution satisfies  $f(x - a) = e^{-a^2 + 2ax} f(x)$  for all  $a, x$ , and 2)  $h : x \rightarrow (M + 1)f(x)G(x)^M$  is a probability distribution of CDF  $x \rightarrow 1 - G(x)^{M+1}$ . We continue the computation with:

$$\begin{aligned} \alpha(M, j) &\leq \frac{e^{-j\Delta^2}}{M + 1} \int_{-\infty}^{\frac{z - \mu_{2,j}}{\sqrt{j}}} e^{\sqrt{j}\Delta y} h(y) dy + G_{2,j}(z)^M \\ &\leq \frac{e^{-j\Delta^2}}{M + 1} e^{\sqrt{j}\Delta \frac{z - \mu_{2,j}}{\sqrt{j}}} (1 - G\left(\frac{z - \mu_{2,j}}{\sqrt{j}}\right)^{M+1}) + G\left(\frac{z - \mu_{2,j}}{\sqrt{j}}\right)^M \\ &\leq \frac{e^{-j\Delta^2}}{M + 1} e^{\sqrt{j}\Delta \frac{z - \mu_{2,j}}{\sqrt{j}}} + G\left(\frac{z - \mu_{2,j}}{\sqrt{j}}\right)^M \end{aligned}$$

As the inequality is true for all  $z \in \mathbb{R}$ , it holds that,

$$\forall y \in \mathbb{R}, \quad \alpha(M, j) \leq \frac{e^{-j\Delta^2}}{M + 1} e^{\sqrt{j}\Delta y} + G(y)^M .$$

Now let  $y_M$  be such as  $G(y_M) = 1 - \frac{1}{\sqrt{M}}$ . This value ensures that the second term satisfies  $G(y_M)^M \leq e^{-\sqrt{M}} = o(M^{-2})$ . Observe that  $y_M = F^{-1}\left(\frac{1}{\sqrt{M+1}}\right)$ . Using the following equivalent of the quantile function of the normal distribution when the quantile is small (see for instance [27]):

$$F^{-1}(p) = -\sqrt{\log \frac{1}{p^2} - \log \log \frac{1}{p} + \log 2\pi} + o_{p \rightarrow 0}(1) ,$$

there exists a constant  $C \in \mathbb{R}$  such that  $y_M \leq -C\sqrt{\log M - \log \log M + \log 4\pi}$ . This yields

$$\alpha(M, j) \leq \frac{e^{-j\Delta^2}}{M+1} e^{-C\sqrt{j}\Delta\sqrt{\log M - \log \log M + \log 4\pi}} + e^{-\sqrt{M}}$$

Noting that for all  $k \in \mathbb{N}^*$ ,

$$k \log \log M = o(C\sqrt{j}\Delta\sqrt{\log M - \log \log M + \log 4\pi})$$

we get that  $\forall k \in \mathbb{N}^*$ :

$$\alpha(M, j) = o\left(\frac{e^{-j\Delta^2}}{(M+1)(\log M)^k}\right)$$

This is sufficient to prove that the Gaussian distribution is balanced. Indeed, as for the Bernoulli distribution this term sums as a convergent geometric series in  $j$ , and with  $M = O(t/(\log t)^2)$  we can make the sum in  $t$  a convergent Bertrand Series.

#### G.4 The Exponential Distribution is Not Balanced

For  $j = 1$ , a direct calculation yields

$$\alpha(M, 1) = \frac{1}{1 + \left(\frac{\mu_1}{\mu_2}\right) M}.$$

Using this, we now prove that the series in Assumption 2. of Theorem 3.1 is in  $\Omega(\log(T))$ , hence the balance condition is not satisfied. As all the  $\alpha_k(M, j)$  are positive, one can write

$$\begin{aligned} \sum_{t=1}^T \sum_{j=1}^{\lfloor (\log t)^2 \rfloor} \alpha_k(\lfloor \beta t / (\log t)^2 \rfloor, j) &\geq \sum_{t=1}^T \alpha_k(\lfloor \beta t / (\log t)^2 \rfloor, 1) \\ &= \sum_{t=2}^T \frac{1}{1 + \left(\frac{\mu_1}{\mu_k}\right) \lfloor \beta t / (\log t)^2 \rfloor} \\ &\geq \sum_{t=2}^T \frac{1}{1 + \left(\frac{\mu_1}{\mu_k}\right) \beta t / (\log t)^2} \\ &\geq C \sum_{t=2}^T \frac{1}{t} = \mathcal{O}(\log(T)), \end{aligned}$$

where  $C$  is some small enough constant that depend on  $\mu_1, \mu_k$  and  $\beta$ .

## H Sketch of Proof with Forced Exploration

In this section, we explain how the proof of Theorem 3.1 is modified when we add forced exploration with  $f_r = \sqrt{\log r}$ , that is when in every round  $r+1$  we add to  $\mathcal{A}_{r+1}$  every arm  $k$  such that  $N_k(r) \leq f_r$ . It is easy to verify that the proof of Lemma 4.1 remains unchanged, as it is inspired by the analysis of SSMC which also uses forced exploration. We now explain how forced exploration modifies the proof of Lemma 4.3 and how we upper bound the resulting new terms for any exponential family.

### H.1 Handling Forced Exploration in Lemma 4.3

The idea is to use the same proof sketch as without forced exploration. We note  $f(r) = \sqrt{\log r}$  the forced exploration rate and  $f^{-1}(r) = \exp(r^2)$  its inverse function.

Let us consider the round  $a_r = f^{-1}(f(r) - 1)$ . At this round, the value of exploration function is  $f(a_r) - 1 = \sqrt{\log a_r} - 1$ , which means that the number of pulls of arm 1 is at least  $\lfloor \sqrt{\log a_r} - 1 \rfloor \geq \sqrt{\log a_r} - 2$ .

Now we look at the length of the interval  $r - a_r$ :

$$\begin{aligned}
r - a_r &= r - f^{-1}(f(r) - 1) \\
&= r - \exp((\sqrt{\log r} - 1)^2) \\
&= r - \exp(\log r - 2\sqrt{\log r} + 1) \\
&= r(1 - \exp(-2\sqrt{\log r} + 1)) \\
&\sim r \text{ when } r \rightarrow +\infty
\end{aligned}$$

As  $r - a_r$  is equivalent to  $r$  when  $r$  is large, for any constant  $\gamma > 0$  there exists a round  $r_\gamma$  such that for  $r > r_\gamma$ :  $r - a_r > \gamma r$ . This means that after the round  $a_r$  arm 1 faces a linear amount of duels, and has an history of at least  $j = \lfloor \sqrt{\log r} - 1 \rfloor$  samples. Introducing  $b_r$  the random variable giving the first time when  $N_1(b_r) = \lfloor \sqrt{\log r} - 1 \rfloor$ , we necessarily have  $b_r \leq a_r$ . We now use that  $N_1(r) \leq (\log r)^2 \Rightarrow \sum_{j=1}^{\lfloor (\log r)^2 \rfloor - 1} \tau_j \leq r$ , which further implies

$$b_r + \sum_{j=\sqrt{\log r}-1}^{\lfloor (\log r)^2 \rfloor - 1} \tau_j \leq r \Rightarrow \sum_{j=\sqrt{\log r}-1}^{\lfloor (\log r)^2 \rfloor - 1} \tau_j \leq r - b_r$$

We can then use the same proof as in Appendix E:

$$\begin{aligned}
\mathbb{P}(N_1(r) \leq (\log r)^2) &\leq \mathbb{P}\left(\exists j \in \{\lfloor \sqrt{\log r} - 1 \rfloor, \dots, \lfloor (\log r)^2 \rfloor - 1\} : \tau_j \geq \frac{r - b_r}{(\log r)^2 - \lfloor \sqrt{\log r} - 1 \rfloor}\right) \\
&\leq \mathbb{P}\left(\exists j \in \{\lfloor \sqrt{\log r} - 1 \rfloor, \dots, \lfloor (\log r)^2 \rfloor - 1\} : \tau_j \geq \frac{r - a_r}{(\log r)^2 - \lfloor \sqrt{\log r} - 1 \rfloor}\right) \\
&\leq \mathbb{P}\left(\exists j \in \{\lfloor \sqrt{\log r} - 1 \rfloor, \dots, \lfloor (\log r)^2 \rfloor - 1\} : \tau_j \geq \frac{r - a_r}{(\log r)^2}\right) \\
&\leq \mathbb{P}\left(\exists j \in \{\lfloor \sqrt{\log r} - 1 \rfloor, \dots, \lfloor (\log r)^2 \rfloor - 1\} : \tau_j \geq \frac{\gamma r}{(\log r)^2}\right)
\end{aligned}$$

The constant  $\gamma$  does not change the sketch of proof, and we finally have:

$$\sum_{r=1}^T \mathbb{P}(N_1(r) \leq (\log r)^2) \leq r'_K + \sum_{r=r'_K}^T \sum_{j=\lfloor f_r \rfloor - 1}^{(\log r)^2} \left[ (K-1) \mathbb{P}(X_{c_r, c_r, j} < M_{r,j}) + \sum_{k=2}^K \alpha_k(M_{r,j}, j) \right] \quad (11)$$

for any sequence  $M_{r,j}$ , and a new constant  $r'_K$ . Observe that the sum in  $j$  does not start in 1 as it does in the statement of Lemma 4.3 in the absence of forced exploration. This justifies the introduction of the *generalized balance condition* in Appendix H.2

## H.2 Exponential Families Satisfy a Generalized Balanced Condition

To conclude the proof as in Theorem 3.1, as Random Block Sampling satisfies the Diversity Property, from (11) (with the choice  $M_{r,j} = \lfloor \beta r / (\log r)^2 \rfloor$ ) it is sufficient to prove that one-dimensional exponential families satisfy the following generalized balance condition.

**Definition** (generalized balance condition). *If SDA is defined with a forced exploration rate  $f_r$  then the generalized balance condition for the rate  $f_r$  is:*

$$\forall \beta \in (0, 1), \sum_{r=1}^T \sum_{j=f_r}^{\lfloor (\log r)^2 \rfloor} \alpha_k(\lfloor \beta r / (\log r)^2 \rfloor, j) = o(\log T).$$

The following lemma proves that this holds in particular for the choice  $f_r = \sqrt{\log(r)}$ , which permits to prove that RB-SDA with this forced exploration sequence is asymptotically optimal for distributions that belong to any one-dimensional exponential family.

**Lemma H.1** (Generalized balance condition on exponential families). *If the exploration rate  $f_r$  satisfies  $\frac{f_r}{\log \log r} \rightarrow +\infty$  then any exponential family of distributions with one parameter satisfies the generalized balance condition.*



*Proof.* A distribution that belong to a one-dimensional exponential family has a density  $f_\theta(y) = f(x, 0)e^{\eta(\theta)y - \psi(\theta)}$  for some natural parameter  $\theta \in \mathbb{R}$ .

We observe that for any  $y_1, \dots, y_j \in \mathbb{R}^j$ , if  $\sum_{i=1}^j y_i \leq \mu_k$ :

$$\prod_{u=1}^j f_{\theta_1}(y_u) = \prod_{u=1}^j e^{(\eta(\theta_1) - \eta(\theta_k))y_u - (\psi(\theta_1) - \psi(\theta_k))} f_{\theta_k}(y_u) \leq e^{-jI_1(\mu_k)} \prod_{u=1}^j f_{\theta_k}(y_u)$$

This inequality ensures that for all  $x, u \in \mathbb{R}$ , if  $F_{k,j}^{-1}(u) \leq \mu_k$ :

$$F_{1,j}(x) \leq e^{-jI_1(\mu_k)} F_{k,j}(x) \Rightarrow F_{1,j}(F_{k,j}^{-1}(u)) \leq e^{-jI_1(\mu_k)} u$$

So for exponential families a strictly positive gap between two distributions leads to an exponential decrease of the ratio of the CDF of the sum. If we use the fact that for all  $u \in \mathbb{R}$ :

$$\begin{aligned} \alpha_k(M, j) &= \int_{-\infty}^{+\infty} f_{1,j}(x) G_{2,j}(x)^M d\mathbb{P}(x) \\ &\leq \int_{-\infty}^u f_{1,j}(x) G_{2,j}(x)^M + \int_u^{+\infty} f_{1,j}(x) G_{2,j}(x)^M d\mathbb{P}(x) \\ &\leq F_{1,j}(u) + G_{2,j}(u)^M \end{aligned}$$

Then  $\forall \beta \in (0, 1)$  and for all sequence  $u_r$ :

$$\begin{aligned} \sum_{r=1}^T \sum_{j=\mathbf{f}_r}^{\lfloor (\log r)^2 \rfloor} \alpha_k(\lfloor \beta r / (\log r)^2 \rfloor, j) &\leq \sum_{r=1}^T \sum_{j=\mathbf{f}_r}^{\lfloor (\log r)^2 \rfloor} (1 - u_r)^{\lfloor \beta r / (\log r)^2 \rfloor} + e^{-jI_1(\mu_k)} u_r \\ &\leq \sum_{r=1}^T \log(r)^2 (1 - u_r)^{\lfloor \beta r / (\log r)^2 \rfloor} + \sum_{r=1}^T \frac{e^{-\mathbf{f}_r I_1(\mu_k)}}{1 - e^{-I_1(\mu_k)}} u_r \end{aligned}$$

We now choose  $u_r$  of the form  $u_r = \frac{(\log r)^k}{r}$ . Indeed, for the first term we get:

$$\begin{aligned} \log(r)^2 (1 - u_r)^{\lfloor \beta r / (\log r)^2 \rfloor} &\leq \log(r)^2 \exp\left(-\lfloor \beta r / (\log r)^2 \rfloor \frac{(\log r)^k}{r}\right) \\ &\leq \log(r)^2 \exp\left(-(\beta r / (\log r)^2 - 1) \frac{(\log r)^k}{r}\right) \\ &\leq \zeta_k (\log r)^2 \exp(-\beta (\log r)^{k-2}) \\ &= o(r^{-1}) \text{ for } k > 2 \end{aligned}$$

where  $\zeta_k$  is an upper bound for  $\exp(\frac{(\log r)^k}{r})$ . From now on we work with  $k = 3$ . For the second term we have to study  $u_r e^{-I_1(\mu_k) f_r}$ :

$$\begin{aligned} u_r e^{-I_1(\mu_k) f_r} &= \exp(\log u_r - I_1(\mu_k) f_r) \\ &= \exp(3 \log \log r - \log r - I_1(\mu_k) f_r) \end{aligned}$$

We see that  $u_r e^{-I_1(\mu_k) f_r} = o(r^{-1})$  if  $3 \log \log r - I_1(\mu_k) f_r \rightarrow 0$ . This condition is achieved if  $f_r / \log \log r \rightarrow +\infty$ , hence an exploration rate satisfying this condition ensures the generalized balance condition for any exponential family of distributions with one parameter for this rate.  $\square$

We point out the fact that this forced exploration is not necessary in SDA, as we proved that some distributions (Bernoulli, Gaussian, Poisson) directly satisfy the balance condition defined in Assumption 2. of Theorem 3.1. We leave for future research an in-depth analysis of the properties of different families of distribution that could exhibit general conditions for the use of forced exploration (or not) in the SDA family.