The Annals of Statistics
2020, Vol. 48, No. 1, 346–373
https://doi.org/10.1214/19-AOS1809
© Institute of Mathematical Statistics, 2020

THE MULTI-ARMED BANDIT PROBLEM: AN EFFICIENT NONPARAMETRIC SOLUTION

BY HOCK PENG CHAN

Department of Statistics and Applied Probability, National University of Singapore, stachp@nus.edu.sg

Lai and Robbins (*Adv. in Appl. Math.* **6** (1985) 4–22) and Lai (*Ann. Statist.* **15** (1987) 1091–1114) provided efficient parametric solutions to the multi-armed bandit problem, showing that arm allocation via upper confidence bounds (UCB) achieves minimum regret. These bounds are constructed from the Kullback–Leibler information of the reward distributions, estimated from specified parametric families. In recent years, there has been renewed interest in the multi-armed bandit problem due to new applications in machine learning algorithms and data analytics. Nonparametric arm allocation procedures like ϵ-greedy, Boltzmann exploration and BESA were studied, and modified versions of the UCB procedure were also analyzed under nonparametric settings. However, unlike UCB these nonparametric procedures are not efficient under general parametric settings. In this paper, we propose efficient nonparametric procedures.

1. Introduction. Lai and Robbins (1985) provided an asymptotic lower bound for the regret in the multi-armed bandit problem, and proposed an index strategy that is efficient, that is, it achieves this bound. Lai (1987) showed that allocation to the arm having the highest upper confidence bound (UCB), constructed from the Kullback–Leibler (KL) information between the estimated reward distributions of the arms, is efficient when the distributions belong to a specified exponential family. Agrawal (1995) proposed a modified UCB procedure that is efficient despite not having to know in advance the total sample size. Cappé et al. (2013) provided explicit, nonasymptotic bounds on the regret of a KL-UCB procedure that is efficient on a larger class of distribution families.

Burnetas and Katehakis (1996) extended UCB to multi-parameter families, almost showing efficiency in the natural setting of normal rewards with unequal variances. Yakowitz and Lowe (1991) proposed nonparametric procedures that do not make use of KL-information, suggesting logarithmic and polynomial rates of regret under finite exponential moment and moment conditions, respectively.

Auer, Cesa-Bianchi and Fischer (2002) proposed a UCB1 procedure that achieves logarithmic regret when the reward distributions are supported on [0, 1]. They also studied the ϵ-greedy algorithm of Sutton and Barto (1998) and provided finite-time upper bounds of its regret. Both UCB1 and ϵ-greedy are nonparametric in their applications and, unlike UCB-Lai or UCB-Agrawal, are not expected to be efficient under a general exponential family setting. Other nonparametric methods that have been proposed include reinforcement comparison, Boltzmann exploration (Sutton and Barto (1998)) and pursuit (Thathacher and Sastry (1985)). Kuleshov and Precup (2014) provided numerical comparisons between UCB and these methods. For a description of applications to recommender systems and clinical trials, see Shivaswamy and Joachims (2012). Burtini, Loeppky and Lawrence (2015) provided a comprehensive survey of the methods, results and applications of the multi-armed bandit problem, developed over the past 30 years.

Received September 2017; revised December 2018.
MSC2010 subject classifications. 62L05.
Key words and phrases. Efficiency, KL-UCB, subsampling, Thompson sampling, UCB.

A strong competitor to UCB under the parametric setting is the Bayesian method; see, for example, Fabius and van Zwet (1970) and Berry (1972). There is also a well-developed literature on optimization under an infinite-time discounted window setting, in which allocation is to the arm maximizing a dynamic allocation (or Gittins) index, see the seminal papers Gittins (1979) and Gittins and Jones (1979), and also Berry and Fristedt (1985), Chang and Lai (1987), Brezzi and Lai (2002). Recently, there has been renewed interest in the Bayesian method due to the developments of UCB-Bayes [see Kaufmann, Cappé and Garivier (2012)] and Thompson sampling [see, e.g., Korda, Kaufmann and Munos (2013)].

In this paper, we propose an arm allocation procedure subsample-mean comparison (SSMC), that though nonparametric, is nevertheless efficient when the reward distributions are from an *unspecified* one-dimensional exponential family. It achieves this by comparing subsample means of the leading arm with the sample means of its competitors. It is empirical in its approach, using more informative subsample means rather than full-sample means alone, for better decision-making. The subsampling strategy was first employed by Baransi, Maillard and Mannor (2014) in their best empirical sampled average (BESA) procedure. However, there are key differences in their implementation of subsampling from ours, as will be elaborated in Section 2.2. Though efficiency has been attained for various one-dimensional exponential families by say UCB-Agrawal or KL-UCB, SSMC is the first to achieve efficiency without having to know the specific distribution family. In addition, we propose in Section 2.4 a related subsample-t comparison (SSTC) procedure, applying t-statistic comparisons in place of mean comparisons, that is, efficient for normal distributions with unknown and unequal variances.

The layout of the paper is as follows. In Section 2, we describe the subsample comparison strategy for allocating arms. In Section 3, we show that the strategy is efficient for exponential families, including the setting of normal rewards with unknown and unequal variances. In Section 4, we show logarthmic regret for Markovian rewards. In Section 5, we provide numerical comparisons against existing methods. In Section 6, we provide a concluding discussion. In Section 7, we prove the results of Sections 3 and 4.

2. Subsample comparisons.

Let $Y_{k1}, Y_{k2}, \dots$, $1 \le k \le K$, be the observations (or rewards) from a population (or arm) Π_k. We assume here and in Section 3 that the rewards are independent and identically distributed (i.i.d.) within each arm. We extend to Markovian rewards in Section 4. Let $\mu_k = EY_{kt}$ and $\mu_* = \max_{1 \le k \le K} \mu_k$.

Consider a sequential procedure for selecting the population to be sampled, with the decision based on past rewards. Let N_k be the number of observations from Π_k when there are N total observations, hence $N = \sum_{k=1}^K N_k$. The objective is to minimize the *regret*

$$R_N := \sum_{k=1}^{K} (\mu_* - \mu_k) E N_k.$$

The Kullback–Leibler information number between two densities f and g, with respect to a common (σ-finite) measure, is

$$D(f|g) = E_f\left[\log \frac{f(Y)}{g(Y)}\right], \tag{2.1}$$

where E_f denotes expectation with respect to $Y \sim f$. An arm allocation procedure is said to be uniformly good if

$$R_N = o(N^{\epsilon}) \quad \text{for all } \epsilon > 0, \tag{2.2}$$

over all reward distributions lying within a specified parametric family.

Let f_k be the density of Y_{kt} and let $f_* = f_k$ for k such that $\mu_k = \mu_*$ (assuming f_* is unique). The celebrated result of Lai and Robbins (1985) is that under (2.2) and additional regularity conditions,

$$\liminf_{N\to\infty} \frac{R_N}{\log N} \geq \sum_{k:\mu_k<\mu_*} \frac{\mu_* - \mu_k}{D(f_k|f_*)}. \tag{2.3}$$

Lai and Robbins (1985) and Lai (1987) went on to propose arm allocation procedures that have regrets achieving the lower bound in (2.3), and are hence *efficient*.

2.1. *Review of existing methods.* In the setting of normal rewards with unit variances, UCB-Lai can be described as the selection, for sampling, Π_k maximizing

$$\bar{Y}_{kn_k} + \sqrt{\frac{2\log(N/n)}{n}}, \tag{2.4}$$

where $\bar{Y}_{kt} = \frac{1}{t}\sum_{u=1}^t Y_{ku}$, n is the current number of observations from the K populations, and n_k is the current number of observations from Π_k. Agrawal (1995) proposed a modified version of UCB-Lai that does not involve the total sample size N, with the selection instead of the population Π_k maximizing

$$\bar{Y}_{kn_k} + \sqrt{\frac{2(\log n + \log\log n + b_n)}{n_k}}, \tag{2.5}$$

with $b_n \to \infty$ and $b_n = o(\log n)$. Efficiency holds for (2.4) and (2.5), and there are corresponding versions of (2.4) and (2.5) that are efficient for other one-parameter exponential families. Cappé et al. (2013) proposed a more general KL-UCB procedure that is also efficient for distributions with given finite support.

Auer, Cesa-Bianchi and Fischer (2002) simplified UCB-Agrawal to UCB1, proposing that Π_k maximizing

$$\bar{Y}_{kn_k} + \sqrt{\frac{2\log n}{n_k}} \tag{2.6}$$

be selected. They showed that under UCB1, logarithmic regret $R_N = O(\log N)$ is achieved when the reward distributions are supported on [0, 1]. In the setting of normal rewards with unequal and unknown variances, Auer et al. suggested applying a variant of UCB1 which they called UCB1-Normal, and showed logarithmic regret. Under UCB1-Normal, an observation is taken from any population Π_k with $n_k < 8\log n$. If such a population does not exist, then an observation is taken from Π_k maximizing

$$\bar{Y}_{kn_k} + 4\widehat{\sigma}_{kn_k}\sqrt{\frac{\log n}{n_k}},$$

where $\widehat{\sigma}_{kt}^2 = \frac{1}{t-1}\sum_{u=1}^t (Y_{ku} - \bar{Y}_{kt})^2$.

Auer et al. provided an excellent study of various nonparametric arm allocation procedures, for example, the ϵ-greedy procedure proposed by Sutton and Barto (1998), in which an observation is taken from the population with the largest sample mean with probability $1-\epsilon$, and randomly with probability ϵ. Auer et al. suggested replacing the fixed ϵ at every stage by a stage-dependent

$$\epsilon_n = \min\left(1, \frac{cK}{d^2 n}\right), \tag{2.7}$$

with c user-specified and $0 < d \le \min_{k:\mu_k<\mu^*}(\mu_* - \mu_k)$. They showed that if $c > 5$, then logarithmic regret is achieved for reward distributions supported on [0, 1]. A more recent numerical study by Kuleshov and Precup (2014) considered additional nonparametric procedures, for example, Boltzmann exploration in which an observation is taken from Π_k with probability proportional to $e^{\bar{Y}_{kn_k}/\tau}$, for some $\tau > 0$.

2.2. *Subsample-mean comparisons.* A common characteristic of the procedures described in Section 2.1 is that allocation is based solely on a comparison of the sample means $\bar{Y}_{kn_k}$, with the exception of UCB1-Normal in which $\widehat{\sigma}_{kn_k}$ is also utilized. As we shall illustrate in Section 2.3, we can utilize subsample-mean information from the leading arm to estimate the confidence bounds for selecting from the other arms. In contrast, UCB-based procedures like KL-UCB discard subsample information and rely on parametric information to estimate these bounds. Even though subsample-mean and KL-UCB are both efficient for exponential families, the advantage of subsample-mean is that the underlying family need not be specified.

In SSMC a leader is chosen in each round of play to compete against all the other arms. Let r denote the round number. In round 1, we sample all K arms. In round r for $r > 1$, we set up a challenge between the leading arm (to be defined below) and each of the other arms. An arm is sampled only if it wins all its challenges in that round. Hence, for round $r > 1$ we sample either the leading arm or a nonempty subset of the challengers. Let n $(= n^r)$ be the total number of observations from all K arms at the beginning of round r, let n_k $(= n_k^r)$ be the corresponding number from Π_k. Hence, $n_k^1 = 0$ and $n_k^2 = 1$ for all k, and $K + (r-2) \le n^r \le K + (K-1)(r-2)$ for $r \ge 2$.

Let c_n be a nonnegative monotone increasing sampling threshold in SSMC and SSTC, with

$$c_n = o(\log n) \quad \text{and} \quad \frac{c_n}{\log\log n} \to \infty \quad \text{as } n \to \infty. \tag{2.8}$$

For example in our implementation of SSMC and SSTC in Section 5, we select $c_n = (\log n)^{\frac{1}{2}}$. An explanation of why (2.8) is required for efficiency of SSMC is given in the beginning of Section 7.1. Let $\bar{Y}_{k,t:u} = \frac{1}{u-t+1}\sum_{v=t}^{u} Y_{kv}$, hence $\bar{Y}_{kt} = \bar{Y}_{k,1:t}$.

Subsample-mean comparison (SSMC).

1. $r = 1$. Sample each Π_k exactly once.
2. $r = 2, 3, \ldots$.
   (a) Let the leader ζ $(= \zeta^r)$ be the population with the most observations, with ties resolved by (in order):
      i. the population with the larger sample mean,
      ii. the leader of the previous round,
      iii. randomization.
   (b) For all $k \ne \zeta$ set up a challenge between Π_ζ and Π_k in the following manner:
      i. If $n_k = n_\zeta$, then Π_k loses the challenge automatically.
      ii. If $n_k < n_\zeta$ and $n_k < c_n$, then Π_k wins the challenge automatically.
      iii. If $c_n \le n_k < n_\zeta$, then Π_k wins the challenge when

$$\bar{Y}_{kn_k} \ge \bar{Y}_{\zeta,t:(t+n_k-1)} \quad \text{for some } 1 \le t \le n_\zeta - n_k + 1. \tag{2.9}$$

   (c) For all $k \ne \zeta$, sample from Π_k if Π_k wins its challenge against Π_ζ. Sample from Π_ζ if Π_ζ wins all its challenges. Hence, either Π_ζ is sampled, or a nonempty subset of $\{\Pi_k : k \ne \zeta\}$ is sampled.

SSMC may recommend more than one populations to be sampled in a single round when $K > 2$. In the event that $n^r < N < n^{r+1}$ for some r, we select $N - n^r$ populations randomly from among the $n^{r+1} - n^r$ recommended by SSMC in the rth round, in order to make up exactly N observations.

If Π_ζ wins all its challenges, then ζ and $(n_k : k \neq \zeta)$ are unchanged, and in the next round it suffices to perform the comparison in (2.9) at the largest t instead of at every t. The computational cost is thus $O(1)$. The computational cost is $O(r)$ if at least one $k \neq \zeta$ wins its challenge. Hence, when there is only one optimal arm and SSMC achieves logarithmic regret, the total computational cost is $O(r \log r)$ for r rounds of the algorithm.

In step 2(b)ii, we force the exploration of arms with less than c_n rewards. By (2.8) we select c_n small compared to $\log n$, so that the cost of such forced explorations is asymptotically negligible. In contrast the forced exploration in the greedy algorithm (2.7) is more substantial, of order $\log n$ for n rewards.

BESA, proposed by Baransi, Maillard and Mannor (2014), also applies subsample-mean comparisons. We describe BESA for $K = 2$ below, noting that tournament-style elimination is applied for $K > 2$. Unlike SSMC, exactly one population is sampled in each round $r > 1$ even when $K > 2$.

Best Empirical Sampled Average (BESA).

1. $r = 1$. Sample both Π_1 and Π_2.
2. $r = 2, 3, \dots$.

   (a) Let the leader ζ be the population with more observations, and let $k \neq \zeta$.

   (b) Sample randomly without replacement n_k of the n_ζ observations from Π_ζ, and let $\bar{Y}^*_{\zeta n_k}$ be the mean of the n_k observations.

   (c) If $\bar{Y}_{k n_k} \geq \bar{Y}^*_{\zeta n_k}$, then sample from Π_k. Otherwise sample from Π_ζ.

As can be seen from the descriptions of SSMC and BESA, the mechanism of choosing the arm to be played in SSMC clearly promotes exploration of nonleading arms, relative to BESA. Whereas Baransi et al. demonstrated logarithmic regret of BESA for rewards bounded on [0, 1] (though BESA can of course be applied on more general settings but with no such guarantees), we show in Section 3 that SSMC is able to extend BESA's subsampling idea to achieve asymptotic optimality, that is efficiency, on a wider set of distributions. Tables 4 and 5 in Section 5 show that SSMC controls the oversampling of inferior arms better relative to BESA, due to its added explorations.

2.3. *Comparison of SSMC with UCB methods.*

Lai and Robbins (1985) proposed a UCB strategy in which the arms take turns to challenge a leader with order n observations. Let us restrict to the setting of exponential families. Denote the leader by ζ and the challenger by k. Lai and Robbins proposed, in their (3.1), upper confidence bounds $U^n_{kt} = U^n_k(Y_{k1}, \dots, Y_{kt})$ satisfying

$$P\Big(\min_{1 \leq t \leq n} U^n_{kt} \geq \mu_k - \epsilon\Big) = 1 - o\big(n^{-1}\big) \quad \text{for all } \epsilon > 0.$$

The decision is to sample from arm k if

$$U^n_{k n_k} \geq \bar{Y}_{\zeta n_\zeta} \quad (\doteq \mu_\zeta),$$

otherwise arm ζ is sampled. By doing this we ensure that if $\mu_k > \mu_\zeta$, then the probability that arm k is sampled is $1 - o(n^{-1})$.

We next consider SSMC. Let $L_{\zeta n_k} = \min_{1 \le t \le n_\zeta - n_k + 1} \bar{Y}_{\zeta, t:(t+n_k-1)}$. Since n_ζ is of order n, it follows that if $\mu_k > \mu_\zeta$, then as Y_{kt} is stochastically larger than $Y_{\zeta t}$,

$$P(L_{\zeta n_k} \le \bar{Y}_{kn_k}) = 1 - o(n^{-1}).$$

In SSMC, we sample from arm k if $L_{\zeta n_k} \le \bar{Y}_{kn_k}$, ensuring, as in Lai and Robbins, that an optimal arm is sampled with probability $1 - o(n^{-1})$ when the leading arm is inferior.

In summary, SSMC differs from UCB in that it compares $\bar{Y}_{kn_k}$ against a lower confidence bound $L_{\zeta n_k}$ of the leading arm, computed from subsample-means instead of parametrically. Nevertheless the critical values that SSMC and UCB-based methods employ for allocating arms are asymptotically the same, as we shall next show.

For simplicity, let us consider unit variance normal densities with $K = 2$. Consider first unbalanced sample sizes with say $n_2 = O(\log n)$ and note, see Appendix A, that

$$\min_{1 \le t \le n_1 - n_2 + 1} \bar{Y}_{1,t:(t+n_2-1)} = \mu_1 - [1 + o_p(1)] \sqrt{\frac{2 \log n}{n_2}}. \tag{2.10}$$

Hence, arm 2 winning the challenge requires

$$\bar{Y}_{2n_2} \ge \mu_1 - [1 + o_p(1)] \sqrt{\frac{2 \log n}{n_2}}. \tag{2.11}$$

By (2.5) and (2.6), UCB-Agrawal, KL-UCB and UCB1 also select arm 2 when (2.11) holds, since $\bar{Y}_{1n_1} + \sqrt{\frac{2 \log n}{n_1}} = \mu_1 + o_p(1)$. Hence, what SSMC does is to estimate the critical value $\mu_1 - [1 + o_p(1)] \sqrt{\frac{2 \log n}{n_2}}$, empirically by using the minimum of the running averages $\bar{Y}_{1,t:(t+n_2-1)}$. In the case of n_1, n_2 both large compared to $\log n$, $\sqrt{\frac{2 \log n}{n_1}} + \sqrt{\frac{2 \log n}{n_2}} \to 0$, and SSMC, UCB-Agrawal, KL-UCB and UCB1 essentially select the population with the larger sample mean.

2.4. *Subsample-t comparisons.* For efficiency outside, one-parameter exponential families, we need to work with test statistics beyond sample means. For example, to achieve efficiency for normal rewards with unknown and unequal variances, the analogue of mean comparisons is t-statistic comparisons

$$\frac{\bar{Y}_{kn_k} - \mu_\zeta}{\hat{\sigma}_{kn_k}} \ge \frac{\bar{Y}_{\zeta, t:(t+n_k-1)} - \mu_\zeta}{\hat{\sigma}_{\zeta, t:(t+n_k-1)}},$$

where $\widehat{\sigma}^2_{k,t:u} = \frac{1}{u-t} \sum_{v=t}^{u} (Y_{kv} - \bar{Y}_{k,t:u})^2$ and $\widehat{\sigma}_{kt} = \widehat{\sigma}_{k,1:t}$. Since μ_ζ is unknown, we estimate it by $\bar{Y}_{\zeta n_\zeta}$.

Subsample-t comparison (SSTC). Proceed as in SSMC, with step 2(b)iii′ below replacing step 2(b)iii.

iii.′ If $c_n \le n_k < n_\zeta$, then Π_k wins the challenge when either $\bar{Y}_{kn_k} \ge \bar{Y}_{\zeta n_\zeta}$ or

$$\frac{\bar{Y}_{kn_k} - \bar{Y}_{\zeta n_\zeta}}{\hat{\sigma}_{kn_k}} \ge \frac{\bar{Y}_{\zeta, t:(t+n_k-1)} - \bar{Y}_{\zeta n_\zeta}}{\hat{\sigma}_{\zeta, t:(t+n_k-1)}} \quad \text{for some } 1 \le t \le n_\zeta - n_k + 1. \tag{2.12}$$

As in SSMC only $O(r \log r)$ computations are needed for r rounds when there is only one optimal arm and the regret is logarithmic. This is because it suffices to record the range of $\bar{Y}_{\zeta n_\zeta}$ that satisfies (2.12) for each $k \ne \zeta$, and the actual value of $\bar{Y}_{\zeta n_\zeta}$. The updating of these requires $O(1)$ computations when both ζ and $(n_k : k \ne \zeta)$ are unchanged.

3. Efficiency. Consider first an exponential family of density functions

$$f(x;\theta)=e^{\theta x-\psi(\theta)}f(x;0),\quad \theta\in\Theta, \tag{3.1}$$

with respect to some measure ν, where $\psi(\theta)=\log[\int e^{\theta x}f(x;0)\nu(dx)]$ is the log moment generating function and $\Theta=\{\theta:\psi(\theta)<\infty\}$. For example the Bernoulli family satisfies (3.1) with ν the counting measure on $\{0,1\}$ and $f(0;0)=f(1;0)=\frac{1}{2}$. The family of normal densities with variance σ^2 satisfies (3.1) with ν the Lebesgue measure and $f(x;0)=\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/(2\sigma^2)}$.

Let $f_k=f(\cdot;\theta_k)$ for some $\theta_k\in\Theta$, $1\le k\le K$. Let $\theta_*=\max_{1\le k\le K}\theta_k$ and $f_*=f(\cdot;\theta_*)$. By (2.1) and (3.1), the KL-information in (2.3),

$$\begin{aligned}D(f_k|f_*)&=\int\{(\theta_k-\theta_*)x-[\psi(\theta_k)-\psi(\theta_*)]\}f(x;\theta_k)\nu(dx)\\&=(\theta_k-\theta_*)\mu_k-[\psi(\theta_k)-\psi(\theta_*)]=I_*(\mu_k),\end{aligned}$$

where I_* is the large deviations rate function of f_*. Let $\Xi=\{\ell:\mu_\ell=\mu_*\}$ be the set of optimal arms.

THEOREM 1. *For the exponential family* (3.1), *SSMC satisfies*

$$\limsup_{r\to\infty}\frac{En_k^r}{\log r}\le\frac{1}{D(f_k|f_*)},\quad k\notin\Xi, \tag{3.2}$$

and is thus efficient.

UCB-Agrawal and KL-UCB are efficient as well for (3.1), see Agrawal (1995) and Cappé et al. (2013), SSMC is unique in that it achieves efficiency by being adaptive to the exponential family, whereas UCB-Agrawal and KL-UCB achieve efficiency by having selection procedures that are specific to the exponential family. On the other hand UCB-based methods require less storage space, and more informative finite-time bounds have been obtained. Specifically for UCB-based methods in exponential families we need only store the sample mean for each arm, and the numerical complexity is of the same order as the sample size. For SSMC as given in Section 2.3, all observations are stored (more of this in Section 6) and the numerical complexity for a sample of size N is $N\log N$ when we have efficiency and exactly one optimal arm.

We next consider normal rewards with unequal and unknown variances, that is with densities

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.3}$$

with respect to Lebesgue measure. Let $M(g)=\frac{1}{2}\log(1+g^2)$. Burnetas and Katehakis (1996) showed that if $f_k=f(\cdot;\mu_k,\sigma_k^2)$, then under uniformly fast convergence and additional regularity conditions, an arm allocation procedure must have regret R_N satisfying

$$\liminf_{N\to\infty}\frac{R_N}{\log N}\ge\sum_{k:\mu_k<\mu_*}\frac{\mu_*-\mu_k}{M(\frac{\mu_*-\mu_k}{\sigma_k})}.$$

They proposed an extension of UCB-Lai but needed the verification of a technical condition to show efficiency. In the case of UCB1-Normal, logarithmic regret also depended on tail bounds of the χ^2- and t-distributions that were only shown to hold numerically by Auer, Cesa-Bianchi and Fischer (2002). In Theorem 2, we show that SSTC achieves efficiency.

THEOREM 2. *For normal densities* (3.3) *with unequal and unknown variances, SSTC satisfies*

$$\limsup_{r\to\infty} \frac{En_k^r}{\log r} \le \frac{1}{M(\frac{\mu_*-\mu_k}{\sigma_k})}, \quad k \notin \Xi,$$

and is thus efficient.

4. Logarithmic regret.

We show here that logarithmic regret can be achieved by SSMC under Markovian assumptions. This is possible because in SSMC we compare blocks of observations that retain the Markovian structure.

For $1 \le k \le K$, let $X_{k1}, X_{k2}, \dots$ be a potentially unobserved $\mathcal{X}$-valued Markov chain, with σ-field $\mathcal{A}$ and transition kernel

$$P_k(x, A) = P(X_{kt} \in A | X_{k,t-1} = x), \quad x \in \mathcal{X}, A \in \mathcal{A}. \tag{4.1}$$

We shall assume for convenience that $(X_{kt})_{t\ge1}$ is stationary. Let $Y_{k1}, Y_{k2}, \dots$ be real-valued and conditionally independent given $(X_{kt})_{t\ge1}$, and having conditional densities $\{f_k(\cdot|x) : 1 \le k \le K, x \in \mathcal{X}\}$, with respect to some measure ν, such that

$$P(Y_{kt} \in B | X_{k1} = x_1, X_{k2} = x_2, \dots) = \int_B f_k(y|x_t)\nu(dy).$$

We assume that the K Markov chains are independent, and that the following Doeblin-type condition holds.

(C1) For $1 \le k \le K$, there exists a nontrivial measure λ_k on $(\mathcal{X}, \mathcal{A})$ such that

$$P_k(x, A) \ge \lambda_k(A), \quad x \in \mathcal{X}, A \in \mathcal{A}.$$

As before let $\mu_k = EY_{kt}$, $\mu_* = \max_{1\le k\le K} \mu_k$ and the regret

$$R_N = \sum_{k:\mu_k<\mu_*} (\mu_* - \mu_k)EN_k.$$

In addition to (C1), we assume the following sample mean large deviations.

(C2) For any $\epsilon > 0$, there exists $b(= b_\epsilon) > 0$ and $Q(= Q_\epsilon) > 0$ such that for $1 \le k \le K$ and $t \ge 1$,

$$P(|\bar{Y}_{kt} - \mu_k| \ge \epsilon) \le Qe^{-tb}. \tag{4.2}$$

(C3) For k such that $\mu_k < \mu_*$ and ℓ such that $\mu_\ell = \mu_*$, there exists $b_1 > 0$, $Q_1 > 0$ and $t_1 \ge 1$ such that for $\omega \le \mu_k$ and $t \ge t_1$,

$$P(\bar{Y}_{\ell t} < \omega) \le Q_1 e^{-tb_1} P(\bar{Y}_{kt} < \omega). \tag{4.3}$$

THEOREM 3. *For Markovian rewards satisfying* (C1)–(C3), *SSMC achieves* $En_k^r = O(\log r)$ *for* $k \notin \Xi$, *hence* $R_N = O(\log N)$.

Agrawal, Teneketzis and Anantharam (1989) and Graves and Lai (1997) considered control problems in which, instead of (4.1) with K Markov chains, there are K arms with each arm representing a distinct Markov transition kernel acting on the same chain. Tekin and Liu (2010) on the other hand considered (4.1), with the constraints that $\mathcal{X}$ is finite and $f_k(\cdot|x)$ is a point mass function for all k and x. They provided a UCB algorithm that achieves logarithmic regret.

We can apply Theorem 3 to show logarithmic regret for i.i.d. rewards on nonexponential parametric families. Lai and Robbins (1985) showed that for the double exponential (DE) densities

$$f_k(y) = \frac{1}{2\tau} e^{-|y-\mu_k|/\tau}, \tag{4.4}$$

with $\tau > 0$, efficiency is achieved by a UCB strategy involving KL-information of the DE densities, hence implementation requires knowledge that the family is DE, including knowing τ. In Example 1 below, we state logarithmic regret, rather than efficiency, for SSMC. The advantage of SSMC is that we do not assume knowledge of (4.4) in its implementation. Verifications of (C1)–(C3) under (4.4) is given in Appendix B.

EXAMPLE 1. For the double exponential densities (4.4), conditions (C1)–(C3) hold, hence under SSMC, $En_k^r = O(\log r)$ for $k \notin \Xi$.

5. Numerical studies. We compare SSMC and SSTC against procedures described in Section 2.1, as well as more modern procedures like BESA, KL-UCB, UCB-Bayes and Thompson sampling. The reader can refer to Chapters 1–3 of Kaufmann (2014) for a description of these procedures. In Examples 2 and 3, we consider normal rewards and the comparisons are against procedures in which either efficiency or logarithmic regret has been established. In Example 4, we consider double exponential rewards and there the comparisons are against procedures that have been shown to perform well numerically. In Examples 5–7, we perform comparisons under the settings of Baransi, Maillard and Mannor (2014).

In the simulations done here, $J = 10{,}000$ datasets are generated for each N, and the regret of a procedure is estimated by averaging over $\sum_{k=1}^{K} (\mu_* - \mu_k) N_k$. Standard errors are located after the ± sign. In Examples 5–7, we reproduce simulation results from Baransi, Maillard and Mannor (2014). Though no standard errors are provided, they are likely to be small given that a larger $J = 50{,}000$ number of datasets are generated there.

EXAMPLE 2. Consider $Y_{kt} \sim \mathrm{N}(\mu_k, 1)$, $1 \le k \le 10$. In Table 1 we see that SSMC improves upon UCB1 and outperforms UCB-Agrawal [setting $b_n = \log\log\log n$ in (2.5)]. Here we generate $\mu_k \sim \mathrm{N}(0, 1)$ in each dataset.

EXAMPLE 3. Consider $Y_{kt} \sim \mathrm{N}(\mu_k, \sigma_k^2)$, $1 \le k \le 10$. We compare SSTC against UCB1-tuned and UCB1-Normal. UCB1-tuned was suggested by Auer et al. and shown to perform well numerically. Under UCB1-tuned the population Π_k maximizing

$$\bar{Y}_{kn_k} + \sqrt{\frac{\log n}{n_k} \min\left(\frac{1}{4}, V_{kn}\right)},$$

TABLE 1

The regrets of SSMC, UCB1 and UCB-Agrawal. The rewards have normal distributions with unit variances. For each N we generate $\mu_k \sim N(0, 1)$ for $1 \le k \le 10$ a total of $J = 10{,}000$ times

	Regret	
	$N = 1000$	$N = 10{,}000$
SSMC	88.4 ± 0.2	137.0 ± 0.5
UCB1	90.2 ± 0.3	154.4 ± 0.7
UCB-Agrawal	113.0 ± 0.3	195.7 ± 0.8

TABLE 2
The regrets of SSTC, UCB1-tuned and UCB1-Normal. The rewards have normal distributions with unequal and unknown variances. For each N we generate $\mu_k \sim \mathrm{N}(0, 1)$ and $\sigma_k^{-2} \sim \mathrm{Exp}(1)$ for $1 \le k \le 10$ a total of $J = 10{,}000$ times

	Regret	
	$N = 1000$	$N = 10{,}000$
SSTC	239 ± 1	492 ± 5
UCB1-tuned	130 ± 2	847 ± 23
UCB1-Normal	1536 ± 5	4911 ± 31

where $V_{kn} = \widehat{\sigma}_{kn_k}^2 + \sqrt{\frac{2\log n}{n_k}}$, is selected. In Table 2, we see that UCB1-tuned is significantly better at $N = 1000$ whereas SSTC is better at $N = 10{,}000$. UCB1-Normal performs quite poorly. Here we generate $\mu_k \sim \mathrm{N}(0, 1)$ and $\sigma_k^{-2} \sim \mathrm{Exp}(1)$ in each dataset.

Kaufmann, Cappé and Garivier (2012) performed simulations under the setting of normal rewards with unequal variances, with $(\mu_1, \sigma_1) = (1.8, 0.5)$, $(\mu_2, \sigma_2) = (2, 0.7)$, $(\mu_3, \sigma_3) = (1.5, 0.5)$ and $(\mu_4, \sigma_4) = (2.2, 0.3)$. They showed that UCB-Bayes achieves regret of about 28 at $N = 1000$ and about 47 at $N = 10{,}000$. We apply SSTC on this setting, achieving regrets of 26.0 ± 0.1 at $N = 1000$ and 43.3 ± 0.2 at $N = 10{,}000$.

EXAMPLE 4. Consider double exponential rewards $Y_{kt} \sim f_k$, with densities

$$f_k(y) = \frac{1}{2\lambda} e^{-|y-\mu_k|/\lambda}, \quad 1 \le k \le 10.$$

We compare SSMC against UCB1-tuned, BESA, Boltzmann exploration and ϵ-greedy. For ϵ-greedy we consider $\epsilon_n = \min(1, \frac{3c}{n})$. We generate $\mu_k \sim \mathrm{N}(0, 1)$ in each dataset.

Table 3 shows that UCB1-tuned has the best performances at $N = 1000$, whereas SSMC has the best performances at $N = 10{,}000$. BESA does well for $\lambda = 2$ at $N = 1000$, and also for $\lambda = 5$ at $N = 10{,}000$. A properly-tuned Boltzmann exploration does well at $N = 1000$ for $\lambda = 2$, whereas a properly-tuned ϵ-greedy does well at $\lambda = 2$ and 5 for $N = 1000$ and at $\lambda = 5$ for $N = 10{,}000$.

In Tables 4 and 5, we tabulate the frequencies of the empirical regrets $\sum_{k=1}^K (\mu_* - \mu_k) N_k$ over the $J = 10{,}000$ simulation runs each for $N = 1000$ and 10,000, at $\lambda = 1$, for SSMC, BESA and UCB1-tuned. Tha tables show that SSMC has the best control of excessive sampling of inferior arms, the worst empirical regret being less than half that of BESA and UCB1-tuned.

EXAMPLE 5. Consider $N = 20{,}000$ Bernoulli rewards under the following scenarios:

1. $\mu_1 = 0.9$, $\mu_2 = 0.8$.
2. $\mu_1 = 0.81$, $\mu_2 = 0.8$.
3. $\mu_2 = 0.1$, $\mu_2 = \mu_3 = \mu_4 = 0.05$, $\mu_5 = \mu_6 = \mu_7 = 0.02$, $\mu_8 = \mu_9 = \mu_{10} = 0.01$.
4. $\mu_1 = 0.51$, $\mu_2 = \cdots = \mu_{10} = 0.5$.

When comparing the simulated regrets in Table 6, it is useful to remember that BESA and SSMC are nonparametric, using the same procedures even when the rewards are not Bernoulli, whereas KL-UCB and Thompson sampling utilize information on the Bernoulli family. SSMC* is a variant of SSMC, see Section 6, with more moderate levels of explorations.

TABLE 3

Regret comparisons for double exponential density rewards. For each N and λ we generate $\mu_k \sim \mathrm{N}(0,1)$ for $1 \le k \le 10$ a total of $J = 10{,}000$ times

	Regret			Regret (×10)		
	$N = 1000$			$N = 10{,}000$		
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$
SSMC	141.7 ± 0.4	330 ± 1	795 ± 3	23.6 ± 0.1	65.0 ± 0.3	236.9 ± 0.8
BESA	117 ± 1	265 ± 2	627 ± 3	28.9 ± 0.7	73 ± 1	215 ± 2
UCB1-tuned	101 ± 2	244 ± 3	608 ± 6	50 ± 1	183 ± 3	499 ± 6
Boltz $\tau = 0.1$	130 ± 2	294 ± 4	673 ± 7	84 ± 2	224 ± 4	557 ± 6
0.2	128 ± 2	264 ± 3	632 ± 6	80 ± 1	169 ± 3	465 ± 6
0.5	332 ± 1	387 ± 2	632 ± 5	310 ± 5	311 ± 2	428 ± 4
1	728 ± 2	737 ± 2	816 ± 4	731 ± 2	716 ± 2	712 ± 3
ϵ-greedy $c = 0.1$	170 ± 3	327 ± 4	681 ± 7	133 ± 3	283 ± 4	579 ± 7
0.2	162 ± 3	312 ± 4	653 ± 6	114 ± 2	251 ± 4	536 ± 6
0.5	150 ± 2	282 ± 3	604 ± 6	82 ± 2	189 ± 3	444 ± 5
1	159 ± 2	271 ± 3	569 ± 5	61 ± 1	146 ± 3	370 ± 5
2	200 ± 1	289 ± 2	559 ± 4	52.9 ± 0.9	113 ± 2	302 ± 4
5	334 ± 1	396 ± 2	617 ± 4	63.4 ± 0.5	101 ± 1	241 ± 3
10	524 ± 2	567 ± 2	742 ± 3	95.7 ± 0.4	119.5 ± 0.8	226 ± 2
20	811 ± 3	839 ± 3	951 ± 3	156.9 ± 0.5	172.1 ± 0.7	251 ± 2

TABLE 4

Number of simulations (out of 10,000) lying within a given empirical regret range, and the worst empirical regret, when $N = 1000$ and $\lambda = 1$

	Frequency of emp. regrets lying within a given range							
	0	200	400	600	800	1000	1200	
	to	to	to	to	to	to	to	Worst
	200	400	600	800	1000	1200	2100	emp. regret
SSMC	9134	845	16	5	0	0	0	770
BESA	9314	424	143	66	27	15	11	2089
UCB1-tuned	8830	625	301	132	64	32	16	1772

TABLE 5

Number of simulations (out of 10,000) lying within a given empirical regret range, and the worst empirical regret, when $N = 10{,}000$ and $\lambda = 1$

	Frequency of emp. regrets lying within a given range							
	0	1000	2000	3000	4000	5000	10,000	
	to	to	to	to	to	to	to	Worst
	1000	2000	3000	4000	5000	10,000	21,000	emp. regret
SSMC	9988	8	3	0	0	1	0	6192
BESA	9708	125	59	34	25	40	9	20,639
UCB1-tuned	8833	365	250	161	122	225	44	16,495

TABLE 6
Regret comparisons for Bernoulli rewards

	Scenario			
	1	2	3	4
SSMC	12.4 ± 0.1	43.1 ± 0.4	97.9 ± 0.2	165.3 ± 0.2
SSMC*	9.5 ± 0.2	48.5 ± 0.6	64.4 ± 0.3	156.0 ± 0.4
BESA	11.83	42.6	74.41	156.7
KL-UCB	17.48	52.34	121.21	170.82
KL-UCB+	11.54	41.71	72.84	165.28
Thompson	11.3	46.14	83.36	165.08

EXAMPLE 6. Consider truncated exponential and Poisson distributions with $N = 20{,}000$. For truncated exponential, we consider $Y_{kt} = \min(\frac{X_{kt}}{10}, 1)$, where $X_{kt} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda_k)$ (density $\lambda_k e^{-\lambda_k x}$) with $\lambda_k = \frac{1}{k}$, $1 \le k \le 5$. For truncated Poisson, we consider $Y_{kt} = \min(\frac{X_{kt}}{10}, 1)$, where $X_{kt} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda_k)$, with $\lambda_k = 0.5 + \frac{k}{3}$, $1 \le k \le 6$. The simulation results are given in Table 7. BESAT is a variation of BESA that starts with 10 observations from each population.

EXAMPLE 7. Consider $K = 2$ and $N = 20{,}000$ with $Y_{1t} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0.2, 0.4)$ and $Y_{2t} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$. Here SSMC underperforms with regret of 163 ± 7 compared to Thompson sampling, which has regret of 13.18. On the other hand SSTC, by normalizing the different scales of the two uniform distributions, is able to achieve the best regret of 2.9 ± 0.2.

6. Discussion. Together with BESA, the procedures SSMC and SSTC that we introduce here form a class of nonparametric procedures that differ from traditional nonparametric procedures, like ϵ-greedy and Boltzmann exploration, in their recognition that when deciding between which of two populations to be sampled, samples or subsamples of the same rather than different sizes should be compared. Among the parametric procedures, Thompson sampling fits most with this scheme.

As mentioned earlier, in SSMC (and SSTC), when the leading population Π_ζ in the previous round is sampled, essentially only one additional comparison is required in the current round between Π_ζ and Π_k for $k \neq \zeta$. On the other hand when there are n rewards, an order n comparisons may be required between Π_ζ and Π_k when Π_k wins in the previous round. It is

TABLE 7
Regret comparisons for truncated exponential and Poisson rewards

	Trunc. expo.	Trunc. Poisson
SSMC	33.8 ± 0.4	18.6 ± 0.1
SSMC*	29.6 ± 0.7	14.7 ± 0.2
BESA	53.26	19.37
BESAT	31.41	16.72
KL-UCB-expo	65.67	–
KL-UCB-Poisson	–	25.05

these added comparisons that, relative to BESA, allows for faster catching-up of a potentially undersampled optimal arm. Tables 4 and 5 show the benefits of such added explorations in minimizing the worst-case empirical regret.

To see if SSMC still works well if we moderate these added explorations, we experimented with the following variation of SSMC in Examples 6 and 7. The numerical results indicate improvements.

SSMC*. Proceed as in SSMC, with step 2(b)iii replaced by the following:

2(b)iii′ If $c_n \le n_k < n_\zeta$, then Π_k wins the challenge when

$$\bar{Y}_{kn_k} \ge \bar{Y}_{\zeta,t:(t+n_k-1)} \quad \text{for some } t = 1 + un_k, 0 \le u \le \left\lfloor \frac{n_\zeta}{n_k} \right\rfloor - 1.$$

In contrast to SSMC, in SSMC* we partition the rewards of the leading arm into groups of size n_k for comparisons instead of reusing the rewards in moving-averages. In principle, the members of the group need not be consecutive in time, thus allowing for the modifications of SSMC* to provide storage space savings when the support of the distributions is finite. That is, rather than to store the full sequence, we simply store the number of occurrences at each support point, and generate a new (permuted) sequence for comparisons whenever necessary. Likewise in BESA, there is substantial storage space savings for finite-support distributions by storing the number of occurrences at each support point.

7. Proofs of Theorems 1–3. Since SSMC and SSTC are index-blind, we may assume without loss of generality that $\mu_1 = \mu_*$. We provide here the statements and proofs of supporting Lemmas 1 and 2, and follow up with the proofs of Theorems 1–3 in Sections 7.1–7.3. We denote the complement of an event D by $\bar{D}$, let $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the greatest and least integer function respectively, and let $|A|$ denote the number of elements in a set A.

Let n_k^r $(= n_k)$ be the number of observations from Π_k at the beginning of round r. Let $n^r (= n) = \sum_{k=1}^K n_k^r$. Let $n_*^r = \max_{1\le k\le K} n_k^r$. Let

$$\Xi = \{\ell : \mu_\ell = \mu_*\} \quad \text{be the set of optimal arms,}$$
$$\zeta^r (= \zeta) \text{ the leader at the beginning of round } r (\ge 2).$$

More specifically, let

$$\mathcal{Z}^r = \{k : n_k^r = n_*^r\},$$
$$\mathcal{Z}_1^r = \{\ell \in \mathcal{Z}^r : \bar{Y}_{\ell n_\ell^r} \ge \bar{Y}_{kn_k^r} \text{ for all } k \in \mathcal{Z}^r\}.$$

If $\zeta^{r-1} \in \mathcal{Z}_1^r$, then $\zeta^r = \zeta^{r-1}$. Otherwise the leader ζ^r is selected randomly (uniformly) from $\mathcal{Z}_1^r$. In particular if $\mathcal{Z}_1^r$ has a single element, then that element must be ζ^r. For $r \ge 2$, let

$$A^r = \{\zeta^r \notin \Xi\} = \{\text{leader at round } r \text{ is inferior}\}.$$

We restrict to $r \ge 2$ because the leader is not defined at $r = 1$. Likewise in our subsequent notations on events B^r, C^r, D^r, G_k^r and H_k^r, we restrict to $r \ge 2$.

In Lemma 1 below, the key ingredient leading to (7.3) is condition (I) on the event G_k^r, which says that it is difficult for an inferior arm k with at least $(1+\epsilon)\xi_k \log r$ rewards to win against a leading optimal arm ζ. In the case of exponential families we show efficiency by verifying (I) with $\xi_k = \frac{1}{I_1(\mu_k)}$. Condition (II), on the event H_k^r, says that analogous winnings from an inferior arm k with at least $J_k \log r$ rewards, for J_k large, are asymptotically negligible. Condition (III) limits the number of times an inferior arm is leading. This condition is important because G_k^r and H_k^r refer to the winning of arm k when the leader is optimal, hence the need, in (III), to bound the event probability of an inferior leader.

LEMMA 1. *Let $k \notin \Xi$ (i.e., k is not an optimal arm) and define*

$$G_k^r = \{\zeta^s \in \Xi, n_k^{s+1} = n_k^s + 1, \\ n_k^s \geq (1+\epsilon)\xi_k \log r \textit{ for some } 2 \leq s \leq r-1\}, \tag{7.1}$$

$$H_k^r = \{\zeta^s \in \Xi, n_k^{s+1} = n_k^s + 1, \\ n_k^s \geq J_k \log r \textit{ for some } 2 \leq s \leq r-1\}, \tag{7.2}$$

for some $\epsilon > 0$, $\xi_k > 0$ and $J_k > 0$. Consider the following conditions:

(I) *There exists $\xi_k > 0$ such that for all $\epsilon > 0$, $P(G_k^r) \to 0$ as $r \to \infty$.*
(II) *There exists $J_k > 0$ such that $P(H_k^r) = O(r^{-1})$ as $r \to \infty$.*
(III) *$P(A^r) = o(r^{-1})$ as $r \to \infty$.*

Under (I)–(III),

$$\limsup_{r\to\infty} \frac{En_k^r}{\log r} \leq \xi_k. \tag{7.3}$$

PROOF. Consider $r \geq 3$. Let $b_r = 1 + (1+\epsilon)\xi_k \log r$ and $d_r = 1 + J_k \log r$. Under the event $\bar{G}_k^r$, arm k in round $s \in [2, r-1]$ is sampled to a size beyond b_r only when $\zeta^s \notin \Xi$ (i.e., under the event A^s). In view that $n_k^2 = 1$ $(< b_r)$, it follows that

$$n_k^r \leq b_r + \sum_{s=2}^{r-1} \mathbf{1}_{A^s}.$$

Hence

$$n_k^r \mathbf{1}_{\bar{G}_k^r} \leq b_r + \sum_{s=2}^{r-1} \mathbf{1}_{A^s}. \tag{7.4}$$

Similarly, under the event $\bar{H}_k^r$,

$$n_k^r \leq d_r + \sum_{s=2}^{r-1} \mathbf{1}_{A^s}.$$

Hence

$$n_k^r \mathbf{1}_{(G_k^r \setminus H_k^r)} \leq d_r \mathbf{1}_{G_k^r} + \sum_{s=2}^{r-1} \mathbf{1}_{A^s}. \tag{7.5}$$

Since $n_k^r \leq r$, by (7.4) and (7.5),

$$\begin{aligned} En_k^r &= E(n_k^r \mathbf{1}_{G_k^r \cap H_k^r}) + E(n_k^r \mathbf{1}_{(G_k^r \setminus H_k^r)}) + E(n_k^r \mathbf{1}_{\bar{G}_k^r}) \\ &\leq rP(H_k^r) + \left[d_r P(G_k^r) + \sum_{s=2}^{r-1} P(A^s)\right] + \left[b_r + \sum_{s=2}^{r-1} P(A^s)\right]. \end{aligned} \tag{7.6}$$

By (III), $\sum_{s=2}^{r} P(A^s) = o(\log r)$, therefore by (7.6), (I) and (II),

$$\limsup_{r\to\infty} \frac{En_k^r}{\log r} \leq (1+\epsilon)\xi_k.$$

We can thus conclude (7.3) by letting $\epsilon \to 0$. □

The verification of (III) is made easier by Lemma 2 below. To provide intuitions for the reader, we sketch its proof first before providing the details.

LEMMA 2. *Let*

$$B^s = \{\zeta^s \in \Xi, n_k^{s+1} = n_k^s + 1, n_k^s = n_\zeta^s - 1 \textit{ for some } k \notin \Xi\},$$

$$C^s = \{\zeta^s \notin \Xi, n_\ell^{s+1} = n_\ell^s \textit{ for some } \ell \in \Xi\}.$$

If as $s \to \infty$,

$$P(B^s) = o(s^{-2}), \tag{7.7}$$

$$P(C^s) = o(s^{-1}), \tag{7.8}$$

then $P(A^r) = o(r^{-1})$ *as* $r \to \infty$.

SKETCH OF PROOF. Note that (7.7) bounds the probability of an inferior arm taking the leadership from an optimal leader in round $s+1$, whereas (7.8) bounds the probability of an inferior leader winning against an optimal challenger in round s. Let $s_0 = \lfloor \frac{r}{4} \rfloor$ and for $r \geq 8$, let

$$\begin{aligned} D^r &= \{\zeta^s \in \Xi \text{ for some } s_0 \leq s \leq r-1\} \\ &= \{\text{the leader is optimal for some rounds between } s_0 \text{ to } r-1\}. \end{aligned}$$

Under $A^r \cap D^r$, there is a leadership takeover by an inferior arm at least once between rounds $s_0 + 1$ and r. More specifically, let s_1 be the largest $s \in [s_0, r-1]$ for which $\zeta^s \in \Xi$. If $s_1 < r-1$, then by the definition of s_1, $\zeta^{s_1+1} \notin \Xi$. If $s_1 = r-1$, then since we are under A^r, $\zeta^{s_1+1} = \zeta^r \notin \Xi$. In summary

$$\begin{aligned} A^r \cap D^r &= \{\xi^s \in \Xi \text{ for some } s_0 \leq s \leq r-1, \zeta^r \notin \Xi\} \\ &\subset \bigcup_{s=s_0}^{r-1} \{\zeta^s \in \Xi, \zeta^{s+1} \notin \Xi\}. \end{aligned} \tag{7.9}$$

By showing that

$$\{\zeta^s \in \Xi, \zeta^{s+1} \notin \Xi\} \subset B^s, \tag{7.10}$$

we can conclude from (7.7) and (7.9) that

$$P(A^r \cap D^r) \leq \sum_{s=s_0}^{r-1} P(B^s) = o(r s_0^{-2}) = o(r^{-1}). \tag{7.11}$$

To see (7.10), recall that by step 2(b)i of SSMC or SSTC, if the (optimal) leader and (inferior) challenger have the same sample size, then the challenger loses by default. The tie-breaking rule then ensures that the challenger is unable to take over leadership in the next round. Hence for ζ^s to lose leadership to an inferior arm k in round $s+1$, it has to lose to arm k when arm k has exactly $n_\zeta^s - 1$ observations.

What (7.11) says is that if at some previous round $s \geq s_0$ the leader is optimal, then (7.7) makes it difficult for an inferior arm to take over leadership during and after round s, so the leader is likely to be optimal all the way from rounds s to r. The only situation we need to guard against is $\bar{D}^r$, the event that leaders are inferior for all rounds between s_0 and $r-1$. Let $\#^r = \sum_{s=s_0}^{r-1} \mathbf{1}_{C^s}$ be the number of rounds an inferior leader wins against at least one optimal arm. In (7.13) we show that by (7.8), the optimal arms will, with high probability, lose less than $\frac{r}{4}$ times between rounds s_0 and $r-1$ when the leader is inferior.

We next show that

$$\bar{D}^r \subset \left\{\#^r \geq \frac{r}{4}\right\} \tag{7.12}$$

(or $\{\#^r < \frac{r}{4}\} \subset D^r$), that is, if the optimal arms lose this few times, then one of them has to be a leader at some round between s_0 to $r-1$. Lemma 2 follows from (7.11)–(7.13). □

PROOF OF LEMMA 2. Consider $r \geq 8$. By (7.8),

$$E(\#^r) = \sum_{s=s_0}^{r-1} P(C^s) = o(r s_0^{-1}) \to 0,$$

hence, by Markov's inequality,

$$P\left(\#^r \geq \frac{r}{4}\right) \leq \frac{E(\#^r)}{r/4} = o(r^{-1}). \tag{7.13}$$

It remains for us to show (7.12). Assume $\bar{D}^r$. Let $m^s = n^s_\zeta - \max_{\ell \in \Xi} n^s_\ell$. Observe that $n^{s+1}_\zeta = n^s_\zeta$ if $n^{s+1}_\ell = n^s_\ell + 1$ for some $\ell \neq \zeta^s$. This is because the leader ζ^s is not sampled if it loses at least one challenge. Moreover by step 2(b)i of SSMC or SSTC, all arms with the same number of observations as ζ^s are not sampled. Therefore if $\zeta^s \notin \Xi$ and $n^{s+1}_\ell = n^s_\ell + 1$ for all $\ell \in \Xi$, that is, if all optimal arms win against an inferior leader, then $m^{s+1} = m^s - 1$. In other words,

$$F^s := \{\zeta^s \notin \Xi, n^{s+1}_\ell = n^s_\ell + 1 \text{ for all } \ell \in \Xi\} \subset \{m^{s+1} = m^s - 1\}. \tag{7.14}$$

Since $m^{s+1} \leq m^s + 1$, it follows from (7.14) that $m^{s+1} \leq m^s + 1 - 2\mathbf{1}_{F^s}$. Therefore

$$m^r \leq m^{s_0} + (r - s_0) - 2\sum_{s=s_0}^{r-1} \mathbf{1}_{F^s},$$

and since $m^r \geq 0$ and $m^{s_0} \leq s_0$, we can conclude that

$$\sum_{s=s_0}^{r-1} \mathbf{1}_{F^s} \leq \frac{r}{2}. \tag{7.15}$$

Under $\bar{D}^r$, $\mathbf{1}_{C^s} = 1 - \mathbf{1}_{F^s}$ for $s_0 \leq s \leq r-1$, and it follows from (7.15) that

$$\#^r \geq (r - s_0) - \frac{r}{2} \geq \frac{r}{4},$$

and (7.12) indeed holds. □

7.1. *Proof of Theorem* 1. We consider here SSMC. Equation (7.7) follows from Lemma 4 below and $c_r = o(\log r)$ whereas (7.8) follows from Lemma 5 and $\frac{c_r}{\log\log r} \to \infty$. We can thus conclude $P(A^r) = o(r^{-1})$ from Lemma 2, and together with the verification in Lemma 6 of (I), see Lemma 1, for $\xi_k = 1/I_1(\mu_k)$ and (II) for J_k large, we can conclude Theorem 1.

The proofs of Lemmas 4–6 use large deviations Chernoff bounds that are given below in Lemma 3. They can be shown using change-of-measure arguments. Let I_k be the large deviations rate function of f_k.

LEMMA 3. *Under* (3.1), *if* $1 \leq k \leq K$, $t \geq 1$ *and* $\omega = \psi'(\theta)$ *for some* $\theta \in \Theta$, *then*

$$P(\bar{Y}_{kt} \geq \omega) \leq e^{-tI_k(\omega)} \quad \text{if } \omega > \mu_k, \tag{7.16}$$

$$P(\bar{Y}_{kt} \leq \omega) \leq e^{-tI_k(\omega)} \quad \text{if } \omega < \mu_k. \tag{7.17}$$

In Lemmas 4–6, we let $\omega = \frac{1}{2}(\mu_* + \max_{k:\mu_k<\mu_*} \mu_k)$ and $a = \min_{1 \leq k \leq K} I_k(\omega)$. Recall that the parameter c_r is a threshold for forced explorations, in step 2(b)ii of SSMC.

LEMMA 4. *Under* (3.1), $P(B^r) \le \frac{3K^2}{1-e^{-a}} e^{-a(\frac{r}{K}-1)}$ *when* $\frac{r}{K} - 1 \ge c_r$.

PROOF. Let r be such that $\frac{r}{K} - 1 \ge c_r$. The event B^r occurs if at round r the leading arm ℓ is optimal (i.e. $\ell \in \Xi$), and it loses to an inferior arm $k(\notin \Xi)$ with $n_k = u$ and $n_\ell = u+1$ for $u + 1 \ge \frac{r}{K}$ (since arm ℓ is leading). It follows from Lemma 3 that

$$P(\bar{Y}_{\ell,t:(t+u-1)} \le \omega \text{ for } t = 1 \text{ or } 2) \le 2e^{-uI_\ell(\omega)}, \quad \ell \in \Xi,$$
$$P(\bar{Y}_{ku} \ge \omega) \le e^{-uI_k(\omega)}, \quad k \notin \Xi.$$

Since arm ℓ loses to arm k when $\bar{Y}_{ku} \ge \min(\bar{Y}_{\ell,1:u}, \bar{Y}_{\ell,2:(u+1)})$, it follows that

$$P(B^r) \le \sum_{\ell \in \Xi} \sum_{k \notin \Xi} \sum_{u=\lceil \frac{r}{K} \rceil - 1}^{r} (2e^{-uI_\ell(\omega)} + e^{-uI_k(\omega)}),$$

and Lemma 4 holds. □

LEMMA 5. *Under* (3.1), $P(C^r) \le K^2 e^{-c_r a} \frac{(\log r)^6}{r} + o(r^{-1})$.

PROOF. The event C^r occurs if at round r the leading arm k is inferior (i.e. $k \notin \Xi$), and it wins a challenge against one or more optimal arms ℓ $(\in \Xi)$. By step 2(b)ii of SSMC, arm k loses automatically when $n_\ell < c_n$, hence we need only consider $n_\ell \ge c_n$. Note that when $n_k = n_\ell$, for arm k to be the leader, by the tie-breaking rule we require $\bar{Y}_{kn_\ell} \ge \bar{Y}_{\ell n_\ell}$. We shall consider $n_\ell > (\log r)^2$ in Case 1 and $n_\ell = v$ for $c_n \le v < (\log r)^2$ in Case 2.

Case 1: $n_\ell > (\log r)^2$. By Lemma 3,

$$P(\bar{Y}_{\ell n_\ell} \le \omega \text{ for some } n_\ell > (\log r)^2) \le \frac{1}{1-e^{-a}} e^{-a(\log r)^2}, \tag{7.18}$$

$$P(\bar{Y}_{kn_\ell} \ge \omega \text{ for some } n_\ell > (\log r)^2) \le \frac{1}{1-e^{-a}} e^{-a(\log r)^2}. \tag{7.19}$$

Case 2: $n_\ell = v$ for $(c_r \le) c_n \le v < (\log r)^2$. In view that $n_k \ge \frac{r}{K}$ when k is the leading arm, we shall show that for r large, for each such v there exists ξ $(= \xi_v)$ such that

$$P(\bar{Y}_{\ell v} < \xi) \le e^{-c_r a} \frac{(\log r)^4}{r}, \tag{7.20}$$

$$\begin{aligned} &P\left(\bar{Y}_{k,t:(t+v-1)} > \xi \text{ for } 1 \le t \le \frac{r}{K}\right) \\ &\quad [\le P(\bar{Y}_{kv} > \xi)^{\lfloor \frac{r}{Kv} \rfloor}] \le \exp\left[-\frac{(\log r)^2}{K} + 1\right]. \end{aligned} \tag{7.21}$$

The inequality within the brackets in (7.21) follows from partitioning $[1, \frac{r}{K}]$ into $\lfloor \frac{r}{Kv} \rfloor$ segments of length v, and applying independence of the sample on each segment.

Since $\theta_\ell > \theta_k$, if $\sum_{t=1}^{v} y_t \le v\mu_k$, then by (3.1),

$$\begin{aligned} \prod_{t=1}^{v} f(y_t; \theta_\ell) &= e^{(\theta_\ell - \theta_k)\sum_{t=1}^{v} y_t - v[\psi(\theta_\ell) - \psi(\theta_k)]} \prod_{t=1}^{v} f(y_t; \theta_k) \\ &\le e^{-vI_\ell(\mu_k)} \prod_{t=1}^{v} f(y_t; \theta_k). \end{aligned}$$

Hence if $\xi \le \mu_k$, then as $v \ge c_r$,

$$P(\bar{Y}_{\ell v} < \xi) \le e^{-vI_\ell(\mu_k)} P(\bar{Y}_{kv} < \xi) \le e^{-c_r a} P(\bar{Y}_{kv} < \xi). \tag{7.22}$$

Let $\xi (\leq \mu_k$ for large $r)$ be such that

$$P(\bar{Y}_{kv} < \xi) \leq \frac{(\log r)^4}{r} \leq P(\bar{Y}_{kv} \leq \xi). \tag{7.23}$$

Equation (7.20) follows from (7.22) and the first inequality in (7.23), whereas (7.21) follows from the second inequality in (7.23) and $v < (\log r)^2$. By (7.18)–(7.21),

$$\begin{aligned} P(C^r) \leq \sum_{\ell \in \Xi} \sum_{k \notin \Xi} \biggl\{ & \frac{2}{1 - e^{-a}} e^{-a(\log r)^2} \\ & + \sum_{v = \lceil c_r \rceil}^{\lfloor (\log r)^2 \rfloor} \biggl(e^{-c_r a} \frac{(\log r)^4}{r} + \exp\biggl[-\frac{(\log r)^2}{K} + 1 \biggr] \biggr) \biggr\}, \end{aligned}$$

and Lemma 5 holds. □

LEMMA 6. *Under* (3.1) *and* $c_r = o(\log r)$, (I) (*in the statement of Lemma* 1) *holds for* $\xi_k = 1/I_1(\mu_k)$ *and* (II) *holds for* $J_k > \max(\frac{1}{I_k(\omega)}, \frac{2}{I_1(\omega)})$, *where* $\omega = \frac{1}{2}(\mu_* + \max_{k:\mu_k < \mu_*} \mu_k)$.

PROOF. Let $k \notin \Xi$. Let $\mu_k < \omega_k < \mu_1$ be such that $(1+\epsilon) I_1(\omega_k) > I_1(\mu_k)$. Consider $n_k = u$ for $u \geq (1+\xi_k) \log r$ (in G_k^r) and $u \geq J_k \log r$ (in H_k^r). Since $I_\ell = I_1$ for $\ell \in \Xi$, it follows from Lemma 3 that

$$P(\bar{Y}_{\ell, t:(t+u-1)} \leq \omega_k \text{ for some } 1 \leq t \leq r) \leq r e^{-u I_1(\omega_k)}, \tag{7.24}$$

$$P(\bar{Y}_{ku} \geq \omega_k) \leq e^{-u I_k(\omega_k)}. \tag{7.25}$$

Since $c_r = o(\log r)$, we can consider r large enough such that $(1+\epsilon)\xi_k \log r \geq c_r$. Hence if in round $1 \leq s \leq r$ arm k has sample size of at least $(1+\epsilon)\xi_r \log r$, it wins against leading optimal arm ℓ only if

$$\bar{Y}_{ku} \geq \bar{Y}_{\ell, t:(t+u-1)} \quad \text{for some } 1 \leq t \leq n_\ell - u + 1 (\leq r).$$

By (7.1), (7.24), (7.25) and Bonferroni's inequality,

$$\begin{aligned} P(G_k^r) &\leq \sum_{u = \lceil (1+\epsilon)\xi_k \log r \rceil}^{r-1} P\{\bar{Y}_{ku} \geq \bar{Y}_{\ell, t:(t+u-1)} \text{ for some } 1 \leq t \leq r \text{ and } \ell \in \Xi\} \\ &\leq \sum_{u = \lceil (1+\epsilon)\xi_k \log r \rceil}^{r-1} \bigl(|\Xi| r e^{-u I_1(\omega_k)} + e^{-u I_k(\omega_k)} \bigr) \\ &\leq \frac{Kr}{1 - e^{-I_1(\omega_k)}} e^{-(1+\epsilon)\xi_k I_1(\omega_k) \log r} + \frac{1}{1 - e^{-I_k(\omega_k)}} e^{-(1+\epsilon)\xi_k I_k(\omega_k) \log r}, \end{aligned}$$

and (I) holds because $(1+\epsilon)\xi_k I_1(\omega_k) > 1$ and $(1+\epsilon)\xi_k I_k(\omega_k) > 0$.

Let $J_k > \max(\frac{1}{I_k(\omega)}, \frac{2}{I_1(\omega)})$. It follows from (7.2), (7.24), (7.25) and the arguments above that

$$\begin{aligned} P(H_k^r) &\leq \sum_{u = \lceil J_k \log r \rceil}^{r-1} \bigl(|\Xi| r e^{-u I_1(\omega)} + e^{-u I_k(\omega)} \bigr) \\ &\leq \frac{Kr}{1 - e^{-I_1(\omega)}} e^{-J_k I_1(\omega) \log r} + \frac{1}{1 - e^{-I_k(\omega)}} e^{-J_k I_k(\omega) \log r}, \end{aligned}$$

and (II) holds because $J_k I_1(\omega) > 2$ and $J_k I_k(\omega) > 1$. □

7.2. *Proof of Theorem* 2. We consider here SSTC. By Lemmas 1 and 2 it suffices, in Lemmas 8–11 below, to verify the conditions needed to show that (7.3) holds with $\xi_k = 1/M(\frac{\mu_* - \mu_k}{\sigma_k})$. Lemma 7 provides the underlying large deviations bounds for the standard error estimator. Let $\Phi(z) = P(Z \le z)$ and $\bar{\Phi}(z) = P(Z > z) (\le e^{-z^2/2}$ for $z \ge 0)$ for $Z \sim \mathrm{N}(0,1)$.

LEMMA 7. *For* $1 \le k \le K$ *and* $t \ge 2$,

$$P(\widehat{\sigma}_{kt}^2/\sigma_k^2 \ge x) \le \exp\left[\frac{(t-1)}{2}(\log x - x + 1)\right] \quad \textit{if } x > 1, \tag{7.26}$$

$$P(\widehat{\sigma}_{kt}^2/\sigma_k^2 \le x) \le \exp\left[\frac{(t-1)}{2}(\log x - x + 1)\right] \quad \textit{if } 0 < x < 1. \tag{7.27}$$

PROOF. We note that $\widehat{\sigma}_{kt}^2/\sigma_k^2 \overset{d}{=} \frac{1}{t-1}\sum_{s=1}^{t-1} U_s$, where $U_s \overset{\text{i.i.d.}}{\sim} \chi_1^2$, and that U_1 has large deviations rate function

$$\begin{aligned} I_U(x) &= \sup_{\theta < \frac{1}{2}} (\theta x - \log E e^{\theta U_1}) \\ &= \sup_{\theta < \frac{1}{2}} \left[\theta x - \frac{1}{2}\log\left(\frac{1}{1-2\theta}\right)\right] = \frac{1}{2}(x - 1 - \log x). \end{aligned}$$

The last equality holds because the supremum occurs when $\theta = \frac{x-1}{2x}$. We conclude (7.26) and (7.27) from (7.16) and (7.17) respectively. □

LEMMA 8. *Under* (3.3), $P(B^r) \le Q e^{-ar}$ *for some* $Q > 0$ *and* $a > 0$, *when* $\frac{r}{K} - 1 \ge c_r$.

PROOF. Let r be such that $\frac{r}{K} - 1 \ge c_r$. The event B^r occurs if at round r the leading arm ℓ is optimal, and it loses to an inferior arm k with $n_k = u$ and $n_\ell = u + 1$ for $u \ge \frac{r}{K} - 1$. Let $k \notin \Xi$, $\ell \in \Xi$ and let $\epsilon > 0$ be such that $\omega := \frac{\mu_k - \mu_\ell + \epsilon}{2\sigma_k} < 0$. Let $\tau_i(u)$, $1 \le i \le 3$, be quantities that we shall define below. Note that

$$\begin{aligned} \tau_1(u) &:= P\left(\frac{\bar{Y}_{ku} - \bar{Y}_{\ell,u+1}}{\hat{\sigma}_{ku}} \ge \omega\right) \\ &\le P\left(\frac{\bar{Y}_{ku} - \bar{Y}_{\ell,u+1}}{2\sigma_k} \ge \omega\right) + P(\widehat{\sigma}_{ku} \ge 2\sigma_k). \end{aligned} \tag{7.28}$$

Since $\bar{Y}_{ku} - \bar{Y}_{\ell,u+1} \sim \mathrm{N}(\mu_k - \mu_\ell, \frac{\sigma_\ell^2}{u+1} + \frac{\sigma_k^2}{u})$,

$$P\left(\frac{\bar{Y}_{ku} - \bar{Y}_{\ell,u+1}}{2\sigma_k} \ge \omega\right) \le \bar{\Phi}\left(\epsilon\sqrt{\frac{u}{\sigma_\ell^2 + \sigma_k^2}}\right) \le e^{-\frac{\epsilon^2 u}{2(\sigma_k^2 + \sigma_\ell^2)}}. \tag{7.29}$$

It follows from (7.26) and (7.27) that

$$P(\widehat{\sigma}_{ku} \ge 2\sigma_k) \le e^{-a_1(u-1)/2}, \tag{7.30}$$

$$P\left(\widehat{\sigma}_{\ell u} \le \frac{\sigma_\ell}{2}\right) \le e^{-a_2(u-1)/2}, \tag{7.31}$$

where $a_1 = 1 - \log 2\ (> 0)$ and $a_2 = \log 2 - \frac{1}{2}\ (> 0)$. By (7.28)–(7.30),

$$\tau_1(u) \le e^{-\frac{\epsilon^2 u}{2(\sigma_k^2 + \sigma_\ell^2)}} + e^{-\frac{a_1(u-1)}{2}}. \tag{7.32}$$

Since $\frac{\bar{Y}_{\ell u}-\bar{Y}_{\ell,u+1}}{\sigma_\ell/2} \sim \mathrm{N}(0,\lambda)$ for $\lambda \le 4(\frac{1}{u}+\frac{1}{u+1}) \le \frac{8}{u}$, it follows that

$$P\biggl(\frac{\bar{Y}_{\ell u}-\bar{Y}_{\ell,u+1}}{\sigma_\ell/2} \le \omega\biggr) \le \bar{\Phi}\biggl(|\omega|\sqrt{\frac{u}{8}}\biggr) \le e^{-\frac{\omega^2 u}{16}}.$$

Hence, by (7.31),

$$\begin{aligned}\tau_2(u) &:= P\biggl(\frac{\bar{Y}_{\ell,t:(t+u-1)}-\bar{Y}_{\ell,u+1}}{\hat{\sigma}_{\ell,t:(t+u-1)}} \le \omega \text{ for } t=1 \text{ or } 2\biggr)\\ &\le 2\biggl[P\biggl(\frac{\bar{Y}_{\ell u}-\bar{Y}_{\ell,u+1}}{\sigma_\ell/2} \le \omega\biggr)+P\biggl(\widehat{\sigma}_{\ell u} \le \frac{\sigma_\ell}{2}\biggr)\biggr]\\ &\le 2\bigl(e^{-\frac{\omega^2 u}{16}}+e^{-\frac{a_2(u-1)}{2}}\bigr).\end{aligned} \tag{7.33}$$

We check that for $\omega_k = \frac{\mu_k+\mu_\ell}{2}$,

$$\begin{aligned}\tau_3(u) &:= P(\bar{Y}_{ku} \ge \bar{Y}_{\ell,u+1})\\ &\le P(\bar{Y}_{ku} \ge \omega_k)+P(\bar{Y}_{\ell,u+1} \le \omega_k)\\ &\le e^{-u(\omega_k-\mu_k)^2/(2\sigma_k^2)}+e^{-(u+1)(\omega_k-\mu_\ell)^2/(2\sigma_\ell^2)}.\end{aligned} \tag{7.34}$$

By (7.32)–(7.34),

$$P(B^r) \le \sum_{k\notin\Xi}\sum_{\ell\in\Xi}\sum_{u=\lceil\frac{r}{K}\rceil-1}^{r}\bigl[\tau_1(u)+\tau_2(u)+\tau_3(u)\bigr],$$

and Lemma 8 indeed holds. □

LEMMA 9. *Under* (3.3), $P(C^r) \le K^2 e^{-c_r a\frac{(\log r)^6}{r}}+o(r^{-1})$ *for some* $a>0$.

PROOF. The event C^r occurs if at round r the leading arm k is inferior, and it wins a challenge against one or more optimal arms ℓ. By step 2(b)ii of SSTC, we need only consider $n_\ell \ge c_n$. Note that when $n_k = n_\ell$, for arm k to be leader, by the tie-breaking rule we require $\bar{Y}_{kn_k} \ge \bar{Y}_{\ell n_\ell}$. Consider n_k taking values u, n_ℓ taking values v and let $\tau_i(\cdot)$, $1 \le i \le 4$, be quantities that we shall define below.

Case 1. $n_\ell > (\log r)^2$. Let $\omega = \frac{\mu_\ell+\mu_k}{2}$ and check that

$$\begin{aligned}\tau_1(u,v) &:= P(\bar{Y}_{\ell v} \le \omega)+P(\bar{Y}_{ku} \ge \omega)\\ &\le e^{-v(\mu_\ell-\mu_k)^2/(8\sigma_\ell^2)}+e^{-u(\mu_\ell-\mu_k)^2/(8\sigma_k^2)}.\end{aligned} \tag{7.35}$$

Case 2. $(c_r \le) c_n \le n_\ell < (\log r)^2$. Let ω be such that

$$(p_\omega :=) P\biggl(\frac{\bar{Y}_{kv}-\mu_k+r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} \le \omega\biggr) = \frac{(\log r)^4}{r}. \tag{7.36}$$

Hence,

$$\begin{aligned}\tau_2(v) &:= P\biggl(\frac{\bar{Y}_{k,t:(t+v-1)}-\mu_k+r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} > \omega \text{ for } 1 \le t \le \frac{r}{K}\biggr)\\ &\bigl[\le (1-p_\omega)^{\lfloor\frac{r}{Kv}\rfloor}\bigr] \le \exp\biggl[-\frac{(\log r)^2}{K}+1\biggr].\end{aligned} \tag{7.37}$$

We shall show that there exists $a > 0$ such that for large r,

$$\tau_3(v) := P\left(\frac{\bar{Y}_{\ell v} - \mu_k - r^{-\frac{1}{3}}}{\hat{\sigma}_{\ell v}} \le \omega\right) \le \frac{e^{-av}(\log r)^4}{r} \left(\le \frac{e^{-c_r a}(\log r)^4}{r}\right). \tag{7.38}$$

For $u \ge \frac{r}{K}$,

$$\tau_4(u) := P\left(|\bar{Y}_{ku} - \mu_k| \ge r^{-\frac{1}{3}}\right) \le e^{-ur^{-1/3}/(2\sigma_k^2)} \le e^{-r^{2/3}/(2K\sigma_k^2)}. \tag{7.39}$$

Since (7.37) and (7.38) hold with "$-\bar{Y}_{ku}$" replacing "$-\mu_k + r^{-\frac{1}{3}}$" and "$-\mu_k - r^{-\frac{1}{3}}$" respectively, by adding $\tau_4(u)$ to the upper bounds,

$$\begin{aligned} P(C^r) \le \sum_{k \notin \Xi} \sum_{\ell \in \Xi} \Bigg(&\sum_{v=\lceil c_r \rceil}^{\lfloor (\log r)^2 \rfloor} [\tau_2(v) + \tau_3(v)] + \sum_{u=\lceil \frac{r}{K} \rceil}^{r} 2\tau_4(u) \\ &+ \sum_{u=\lceil \frac{r}{K} \rceil}^{r} \sum_{v=\lceil (\log r)^2 \rceil}^{r} \tau_1(u, v) \Bigg). \end{aligned}$$

We conclude Lemma 9 from (7.35) and (7.37)–(7.39).

We shall now show (7.38), noting first that for r large, the ω satisfying (7.36) is negative. This is because for $v < (\log r)^2$,

$$P\left(\frac{\bar{Y}_{kv} - \mu_k + r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} \le 0\right) = \Phi\left(-\frac{r^{-\frac{1}{3}}\sqrt{v}}{\sigma_k}\right) \to \frac{1}{2},$$

whereas $\frac{(\log r)^4}{r} \to 0$.

Let g_v be the common density function of $\widehat{\sigma}_{kv}/\sigma_k$ and $\widehat{\sigma}_{\ell v}/\sigma_\ell$. By the independence of $\bar{Y}_{kv}$ and $\widehat{\sigma}_{kv}$,

$$\begin{aligned} P\left(\frac{\bar{Y}_{kv} - \mu_k + r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} \le \omega\right) &= \int_0^\infty P\left(\frac{\bar{Y}_{kv} - \mu_k + r^{-\frac{1}{3}}}{\sigma_k} \le \omega x\right) g_v(x)\,dx \\ &= \int_0^\infty \Phi\left(\sqrt{v}\left(\omega x - \frac{r^{-\frac{1}{3}}}{\sigma_k}\right)\right) g_v(x)\,dx. \end{aligned} \tag{7.40}$$

By similar arguments,

$$P\left(\frac{\bar{Y}_{\ell v} - \mu_k - r^{-\frac{1}{3}}}{\hat{\sigma}_{\ell v}} \le \omega\right) = \int_0^\infty \Phi\left(\sqrt{v}\left(\omega x - \frac{\Delta - r^{-\frac{1}{3}}}{\sigma_\ell}\right)\right) g_v(x)\,dx, \tag{7.41}$$

where $\Delta := \mu_\ell - \mu_k$ (> 0). Let $\delta_1 = \frac{r^{-\frac{1}{3}}}{\sigma_k}$, $\delta_2 = \frac{\Delta - r^{-\frac{1}{3}}}{\sigma_\ell}$ and $b = -\omega x$. Since $b > 0$ and $\delta_2 > \delta_1 > 0$ for r large,

$$\Phi\left(\sqrt{v}(-b - \delta_2)\right) \le e^{-a_r v} \Phi\left(\sqrt{v}(-b - \delta_1)\right), \tag{7.42}$$

where $a_r = \frac{(\delta_2 - \delta_1)^2}{2}$ $(\to \frac{\Delta^2}{2\sigma_\ell^2}$ as $r \to \infty)$. Let $a = \frac{\Delta_2^2}{4\sigma_\ell^2}$. It follows from (7.40)–(7.42) that for r large,

$$P\left(\frac{\bar{Y}_{\ell v} - \mu_k - r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} \le \omega\right) \le e^{-av} P\left(\frac{\bar{Y}_{kv} - \mu_k + r^{-\frac{1}{3}}}{\hat{\sigma}_{kv}} \le \omega\right).$$

Hence, by (7.36), the inequality in (7.38) indeed holds. □

LEMMA 10. *Let $Z_s \sim \mathrm{N}(0, \frac{1}{s+1})$ and $W_s \sim \chi_s^2/s$ be independent. For any $g < 0$ and $0 < \delta < M(g)$, there exists $Q > 0$ such that for $s_1 \geq 1$,*

$$\sum_{s=s_1}^{\infty} P\left\{\frac{Z_s}{\sqrt{W_s}} \leq g\right\} \leq Qe^{-s_1[M(g)-\delta]}.$$

PROOF. Consider the domain $\Omega = \mathbf{R}^+ \times \mathbf{R}$, and the set

$$A = \{(w, z) \in \Omega : z \leq g\sqrt{\omega}\}.$$

Let $I(w, z) = \frac{1}{2}(z^2 + w - 1 - \log w)$, and check that

$$\begin{aligned}\inf_{(w,z)\in A} I(w, z) &= \inf_{w>0} I(w, g\sqrt{w}) \\ &= \inf_{w>0}\left[\frac{1}{2}(g^2 w + w - 1 - \log w)\right] = \frac{1}{2}\log(1 + g^2) \\ &= M(g),\end{aligned} \tag{7.43}$$

the second to last equality follows from the infimum occurring at $w = \frac{1}{g^2+1}$.

Let L_v, $1 \leq v \leq V$, be half-spaces constructed as follows. Let

$$\begin{aligned} L_1 = \{(w, z) : z \leq z_1, 0 < w < \infty\} \\ \text{with } g < z_1 < 0 \text{ and } I(1, z_1) \geq M(g) - \delta.\end{aligned} \tag{7.44}$$

The existence of z_1 satisfying second line of (7.44) follows from $I(1, g) = \frac{1}{2}g^2 > M(g)$. Since $(A \setminus L_1) \subset (0, 1) \times (z_1, 0)$, by (7.43), we can find half-spaces

$$\begin{aligned} L_v = \{(w, z) : 0 < w \leq w_v, z \leq z_v\} \quad \text{with } 0 < w_v < 1, \\ z_v \leq 0 \text{ and } I(w_v, z_v) \geq M(g) - \delta, 2 \leq v \leq V,\end{aligned} \tag{7.45}$$

such that $(A \setminus L_1) \subset \bigcup_{v=2}^{V} L_v$. Therefore $A \subset \bigcup_{v=1}^{V} L_v$, and so

$$\sum_{s=s_1}^{\infty} P\left\{\frac{Z_s}{\sqrt{W_s}} \leq g\right\} \leq \sum_{s=s_1}^{\infty}\sum_{v=1}^{V} P\{(W_s, Z_s) \in L_v\}. \tag{7.46}$$

It follows from (7.27), (7.44), (7.45) and the independence of Z_s and W_s, setting $w_1 = 1$, that

$$P\{(W_s, Z_s) \in L_v\} \leq e^{-sI(w_v,z_v)} \leq e^{-s[M(g)-\delta]}, \quad 1 \leq v \leq V. \tag{7.47}$$

Lemma 10, with $Q = \frac{V}{1-e^{-M(g)+\delta}}$, follows from substituting (7.47) into (7.46). □

LEMMA 11. *Under* (3.3) *and $c_r = o(\log r)$,* (I) (*in the statement of Lemma* 1) *holds for $\xi_k = 1/M(\frac{\mu_* - \mu_k}{\sigma_k})$ and* (II) *holds for J_k large.*

PROOF. By considering the rewards $Y_{kt} - \mu_*$, we may assume without loss of generality that $\mu_* = 0$. Let $k \notin \Xi$ (hence $\mu_k < 0$) and $\epsilon > 0$. Let $g_k = \frac{\mu_k}{\sigma_k}$ and let $g_\omega < 0$ and $\delta > 0$ be such that

$$0 > g_\omega - 3\delta > g_k \quad \text{and} \quad (1 + \epsilon)[M(g_\omega - \delta) - \delta] > M(g_k). \tag{7.48}$$

Let $m_r = \lceil(1 + \epsilon)(\log r)/M(g_k)\rceil$. Since $c_r = o(\log r)$, we can consider r large enough such that $m_r \geq c_r$. By (7.27),

$$\sum_{u=m_r}^{r} P\left(\widehat{\sigma}_{\ell u}^2/\sigma_\ell^2 \leq \frac{1}{4}\right) \to 0, \quad 1 \leq \ell \leq K. \tag{7.49}$$

Let $\sigma_0 = \min_{1 \leq \ell \leq K} \sigma_\ell$. For $\ell \in \Xi$,

$$\sum_{v=\lceil \frac{r}{K} \rceil}^{r} P\left(\frac{|\bar{Y}_{\ell v}|}{\sigma_0/2} \geq \delta\right) \leq \sum_{v=\lceil \frac{r}{K} \rceil}^{r} \exp\left(-\frac{\delta^2 \sigma_0^2 v}{8\sigma_\ell^2}\right) = O(r^{-1}), \tag{7.50}$$

$$\begin{aligned} \eta^r &:= P\left(\bar{Y}_{kn_k} \geq \bar{Y}_{\ell n_\ell} \text{ for some } n_k \geq m_r, n_\ell \geq \frac{r}{K}, \ell \in \Xi\right) \\ &\leq \sum_{u=m_r}^{r} \exp\left(-\frac{u\mu_k^2}{8\sigma_k^2}\right) + \sum_{\ell \in \Xi} \sum_{v=\lceil \frac{r}{K} \rceil}^{r} \exp\left(-\frac{v\mu_k^2}{8\sigma_\ell^2}\right) \to 0. \end{aligned} \tag{7.51}$$

By (7.1) and (7.48),

$$\begin{aligned} P(G_k^r) \leq P\biggl(&\frac{\bar{Y}_{kn_k} - \bar{Y}_{\ell n_\ell}}{\hat{\sigma}_{kn_k}} \geq \frac{\bar{Y}_{\ell,t:(t+n_k-1)} - \bar{Y}_{\ell n_\ell}}{\hat{\sigma}_{\ell n_k}} \\ &\text{for some } 1 \leq t \leq r, \ell \in \Xi, n_k \geq m_r, n_\ell \geq \frac{r}{K}\biggr) + \eta^r \\ \leq \sum_{u=m_r}^{r} \Biggl[& P\left(\frac{\bar{Y}_{ku}}{\hat{\sigma}_{ku}} \geq g_k + \delta\right) + r \sum_{\ell \in \Xi} P\left(\frac{\bar{Y}_{\ell u}}{\hat{\sigma}_{\ell u}} \leq g_\omega - \delta\right) \\ &+ \sum_{\ell=1}^{K} P\left(\widehat{\sigma}_{\ell u}^2/\sigma_\ell^2 \leq \frac{1}{4}\right)\Biggr] + \sum_{\ell \in \Xi} \sum_{v=\lceil \frac{r}{K} \rceil}^{r} P\left(\frac{|\bar{Y}_{\ell v}|}{\sigma_0/2} \geq \delta\right) + \eta^r. \end{aligned} \tag{7.52}$$

By (7.49)–(7.52), to show (I), it suffices to show that

$$\sum_{u=m_r}^{r} P\left(\frac{\bar{Y}_{ku}}{\hat{\sigma}_{ku}} \geq g_k + \delta\right) \to 0, \tag{7.53}$$

$$r \sum_{u=m_r}^{r} P\left(\frac{\bar{Y}_{\ell u}}{\hat{\sigma}_{\ell u}} \leq g_\omega - \delta\right) \to 0. \tag{7.54}$$

Keeping in mind that $g_k + \delta < 0$, let $w > 1$ be such that $\sqrt{w}(g_k + \delta) > g_k$. It follows from (7.26) and $g_k \sigma_k = \mu_k$ that

$$\begin{aligned} &\sum_{u=m_r}^{r} P\left(\frac{\bar{Y}_{ku}}{\hat{\sigma}_{ku}} \geq g_k + \delta\right) \\ &\quad \leq \sum_{u=m_r}^{r} [P(\bar{Y}_{ku} \geq \sqrt{w}(\mu_k + \delta\sigma_k)) + P(\widehat{\sigma}_{ku}^2/\sigma_k^2 \geq w)] \\ &\quad \leq \sum_{u=m_r}^{r} [e^{-u[\mu_k - \sqrt{w}(\mu_k + \delta\sigma_k)]^2/(2\sigma_k^2)} + e^{-(u-1)(w+1-\log w)/2}], \end{aligned}$$

and (7.53) indeed holds. Finally, by Lemma 10,

$$\sum_{u=m_r}^{r} P\left(\frac{\bar{Y}_{\ell u}}{\hat{\sigma}_{\ell u}} \leq g_\omega - \delta\right) \leq Qe^{-(m_r-1)[M(g_\omega - \delta) - \delta]},$$

for some $Q > 0$, and so (7.54) follows from (7.48).

To show (II), we consider $m_r = \lceil J_r \log r \rceil$. By (7.27), we can select J_k large enough to satisfy (7.49) with "$\to 0$" replaced by "$= O(r^{-1})$". We note that (7.52) holds with H_k^r in place of G_k^r for this m_r. Therefore to show (II), it suffices to note that for J_k large enough, (7.51), (7.53) and (7.54) hold with "$\to 0$" replaced by "$= O(r^{-1})$". □

7.3. *Proof of Theorem* 3. Assume (C1)–(C3) and let $\widetilde{\mu} = \max_{k:\mu_k<\mu^*} \mu_k$. By Lemmas 1 and 2 it suffices, in Lemmas 12–14 below, to verify the conditions needed for SSMC to satisfy (7.3) for some $\xi_k > 0$.

LEMMA 12. *Under* (C2), $P(B^r) \le \frac{3QK^2}{1-e^{-b}} e^{-b(\frac{r}{K}-1)}$ *for some* $b > 0$ *and* $Q > 0$, *when* $\frac{r}{K} - 1 \ge c_r$.

PROOF. Consider r such that $(n_k \ge) \frac{r}{K} - 1 \ge c_r$. Let $\epsilon = \frac{1}{2}(\mu_* - \widetilde{\mu})$ and let b and Q be the constants satisfying (C2). Lemma 12 follows from arguments similar to those in the proof of Lemma 4, setting $\omega = \frac{1}{2}(\mu_* + \widetilde{\mu})$. □

LEMMA 13. *Under* (C1)–(C3), $P(C^r) \le K^2 Q_1 e^{-c_r b_1 \frac{(\log r)^6}{r}} + o(r^{-1})$ *for some* $b_1 > 0$ *and* $Q_1 > 0$.

PROOF. The event C^r occurs if at round r the leading arm k is inferior, and it wins against one or more optimal arms ℓ. By step 2(b)ii of SSMC, we need only consider $n_\ell = v$ for $v \ge c_n$. Note that $n_k \ge \frac{r}{K}$ and $n_k \ge n_\ell$.

Case 1: $n_\ell > (\log r)^2$. Let ω and ϵ be as in the proof of Lemma 12. By (C2), there exists $b > 0$ and $Q > 0$ such that

$$P(\bar{Y}_{\ell v} \le \omega) + P(\bar{Y}_{kv} \ge \omega) \le 2Qe^{-vb}. \tag{7.55}$$

Case 2: $n_\ell = v$ for $(c_r \le) c_n \le v < (\log r)^2$. Select $\omega (\le \mu_k$ for r large) such that

$$P(\bar{Y}_{kv} < \omega) \le \frac{(\log r)^4}{r} \le P(\bar{Y}_{kv} \le \omega). \tag{7.56}$$

Let $p_\omega = P(\bar{Y}_{kv} > \omega)$ and let $d = \lceil 2(\log r)^2 \rceil$, $\eta = \lfloor \frac{r/K-1}{d} \rfloor$. By (C1) and the second inequality of (7.56),

$$\begin{aligned} \tau(v) &:= P\biggl(\bar{Y}_{k,t:(t+v-1)} > \omega \text{ for } 1 \le t \le \frac{r}{K}\biggr) \\ &\le P(\bar{Y}_{k,t:(t+v-1)} > \omega \text{ for } t = 1, d+1, \dots, \eta d + 1) \\ &\le p_\omega^{\eta+1} + \eta[1 - \lambda_k(\mathbf{R})]^{d-v+1} \\ &\le \exp\biggl(-\frac{(\eta+1)(\log r)^4}{r}\biggr) + \eta[1-\lambda_k(\mathbf{R})]^{(\log r)^2} \quad [= o(r^{-2})]. \end{aligned} \tag{7.57}$$

To see the second inequality of (7.57), let

$$D_m = \{\bar{Y}_{k,t:(t+v-1)} > \omega \text{ for } t = md+1\}, \quad 0 \le m \le \eta.$$

Note that the probability in the second line of (7.57) is $P(\bigcap_{m=0}^{\eta} D_m)$, and that by (7.56), $P(D_m) = p_\omega \le 1 - \frac{(\log r)^4}{r}$. By the triangular inequality and the convention $\prod_{m=\eta+1}^{\eta} = 1$,

$$\begin{aligned} &\biggl| P\biggl(\bigcap_{m=0}^{\eta} D_m\biggr) - \prod_{m=0}^{\eta} P(D_m) \biggr| \\ &\quad \le \sum_{u=1}^{\eta} \biggl| P\biggl(\bigcap_{m=0}^{u} D_m\biggr) \prod_{m=u+1}^{\eta} P(D_m) - P\biggl(\bigcap_{m=0}^{u-1} D_m\biggr) \prod_{m=u}^{\eta} P(D_m) \biggr| \\ &\quad \le \sum_{u=1}^{\eta} \biggl| P\biggl(\bigcap_{m=0}^{u} D_m\biggr) - P\biggl(\bigcap_{m=0}^{u-1} D_m\biggr) P(D_u) \biggr|. \end{aligned} \tag{7.58}$$

By (C1),

$$\left|P\left(\bigcap_{m=0}^{u} D_m\right) - P\left(\bigcap_{m=0}^{u-1} D_m\right)P(D_u)\right| \le [1-\lambda_k(\mathbf{R})]^{d-v+1}, \quad 1 \le u \le \eta, \tag{7.59}$$

since $\bigcap_{m=0}^{u-1} D_m$ depends on $(Y_{k1}, \dots, Y_{k,(u-1)d+v})$ whereas D_u depends on $(Y_{k,ud+1}, \dots, Y_{k,ud+v})$. Substituting (7.59) into (7.58) gives us the second inequality of (7.57).

It follows from (C3) and the first inequality of (7.56) that there exists $Q_1 > 0$, $b_1 > 0$ and $t_1 \ge 1$ such that for $v \ge t_1$,

$$P(\bar{Y}_{\ell v} < \omega) \le Q_1 e^{-b_1 v} \frac{(\log r)^4}{r}.$$

Hence, by (7.55) and (7.57), for r such that $c_r \ge t_1$,

$$P(C^r) \le \sum_{k \notin \Xi} \sum_{\ell \in \Xi} \left(\sum_{v=\lceil (\log r)^2 \rceil}^{r} 2Qe^{-vb} + \sum_{v=\lceil c_r \rceil}^{\lfloor (\log r)^2 \rfloor} \left[Q_1 e^{-b_1 c_r} \frac{(\log r)^4}{r} + \tau(v) \right] \right),$$

and Lemma 13 holds. □

LEMMA 14. *Under* (C2) *and* $c_r = o(\log r)$, *statement* (II) *in Lemma* 1 *holds.*

PROOF. Let ϵ and ω be as in the proof of Lemma 12, and let b and Q be the constants satisfying (C2). For an optimal arm ℓ,

$$P(\bar{Y}_{\ell,t:(t+u-1)} \le \omega \text{ for some } 1 \le t \le r) \le Qre^{-ub},$$

$$P(\bar{Y}_{ku} \ge \omega) \le Qe^{-ub}.$$

Let $J_k > \frac{2}{b}$. Since $c_r = o(\log r)$, for r large, $\lceil J_k \log r \rceil \ge c_r$ and therefore by Bonferroni's inequality,

$$P(H_k^r) \le \sum_{\ell \in \Xi} \sum_{u=\lceil J_k \log r \rceil}^{r} Q(r+1)e^{-ub},$$

and (II) holds. □

APPENDIX A: SHOWING (2.10)

Let $\Phi(z) = P(Z \le z)$ for $Z \sim N(0,1)$. It follows from $\Phi(-z) = [1+o(1)]\frac{1}{z\sqrt{2\pi}} e^{-z^2/2}$ as $z \to \infty$ that

$$\Phi(-\sqrt{2\log n}) = \frac{1+o(1)}{2n\sqrt{\pi \log n}}, \tag{A.1}$$

$$\Phi\left(-\sqrt{2\log\left(\frac{n}{(\log n)^2}\right)}\right) = [1+o(1)]\frac{(\log n)^{3/2}}{2n\sqrt{\pi}}. \tag{A.2}$$

Assume without loss of generality $\mu_1 = 0$ and consider $n_1 = u$ and $n_2 = v$ (hence $u + v = n$) with $v = O(\log n)$. By (A.1) and Bonferroni's inequality,

$$\begin{aligned} &P\left(\min_{1 \le t \le u-v+1} \bar{Y}_{1,t:(t+v-1)} \le -\sqrt{\frac{2\log n}{v}}\right) \\ &\quad \le \sum_{t=1}^{u-v+1} P\left(\bar{Y}_{1,t:(t+v-1)} \le -\sqrt{\frac{2\log n}{v}}\right) \\ &\quad = (u-v+1)\Phi(-\sqrt{2\log n}) \to 0. \end{aligned} \tag{A.3}$$

By (A.2) and independence of $\bar{Y}_{1,(sv+1):[(s+1)v]}$ for $0 \le s \le \frac{u-v}{v}$,

$$\begin{aligned} & P\biggl(\min_{1\le t\le u-v+1} \bar{Y}_{1,t:(t+v-1)} \ge -\sqrt{\frac{2\log(n/(\log n)^2)}{v}}\biggr) \\ & \quad \le P\biggl(\min_{0\le s\le (u-v)/v} \bar{Y}_{1,(sv+1):[(s+1)v]} \ge -\sqrt{\frac{2\log(n/(\log n)^2)}{v}}\biggr) \\ & \quad = \biggl[1-\Phi\biggl(-\sqrt{2\log\biggl(\frac{n}{(\log n)^2}\biggr)}\biggr)\biggr]^{\lfloor\frac{u-v}{v}\rfloor+1} \\ & \quad \le \exp\biggl[-\biggl(\biggl\lfloor\frac{u-v}{v}\biggr\rfloor+1\biggr)\Phi\bigl(-\sqrt{2\log(n/(\log n)^2)}\bigr)\biggr] \to 0. \end{aligned} \tag{A.4}$$

We conclude (2.10) from (A.3) and (A.4).

APPENDIX B: VERIFICATIONS OF (C1)–(C3) FOR DOUBLE EXPONENTIAL DENSITIES

By dividing Y_{kt} by τ if necessary, we may assume without loss of generality that $\tau = 1$. We check that (C1) holds for $\lambda_k(A) = \int_A f_k(y)\,dy$, whereas (C2) follows from the Chernoff bounds given in Lemma 3, that is, (4.2) holds for $Q = 2$ and $b = I(\epsilon)$, where $I(\mu) = \sup_{|\theta|<1}[\theta\mu - \log(1-\theta^2)]$ is the large deviations rate function of the double exponential density $f(y) = \frac{1}{2}e^{-|y|}$.

Let $S_t = \sum_{u=1}^t Y_u$ with $Y_u \overset{\text{i.i.d.}}{\sim} f$ and let $\Delta = \mu_\ell - \mu_k$. Since $\mu_k - Y_{kt} \sim f$, and similarly when k is replaced by ℓ, to show (C3), it suffices to show that for $z \ge 0$ and $t \ge 1$,

$$P(S_t > z + \Delta t) \le e^{-tb_1} P(S_t > z), \tag{B.1}$$

where $b_1 = \Delta - 2\log(1+\frac{\Delta}{2})$ (>0). By (B.1), (C3) holds for $Q_1 = 1$, $t_1 = 1$ and the above b_1.

Since $Y_u \overset{d}{=} Z_{u1} - Z_{u2}$, with Z_{u1} and Z_{u2} independent exponential random variables with mean 1, it follows that $S_t \overset{d}{=} S_{t1} - S_{t2}$ where S_{t1} and S_{t2} are independent Gamma random variables. Using this, Kotz, Kozubowski and Podgórski (2001) showed, see their (2.3.25), that the density f_t of S_t can be expressed as $f_t(x) = e^{-x}g_t(x)$ for $x \ge 0$, where

$$g_t(x) = \frac{1}{(t-1)!2^{2t-1}} \sum_{j=0}^{t-1} c_{tj}x^j \quad \text{with } c_{tj} = \frac{(2t-2-j)!2^j}{j!(t-1-j)!}. \tag{B.2}$$

We shall show that

$$g_t'(x)\biggl(1+\frac{x}{2t}\biggr) \le g_t(x). \tag{B.3}$$

By (B.3),

$$\frac{f_t'(x)}{f_t(x)} = \frac{e^{-x}[g_t'(x) - g_t(x)]}{e^{-x}g_t(x)} \le \frac{2t}{x+2t} - 1,$$

and therefore for $y \ge 0$,

$$\begin{aligned} \log\biggl[\frac{f_t(y+t\Delta)}{f_t(y)}\biggr] &= \int_y^{y+t\Delta} \frac{f_t'(x)}{f_t(x)}\,dx \\ &\le 2t\log\biggl(\frac{y+(2+\Delta)t}{y+2t}\biggr) - t\Delta \le -tb_1. \end{aligned}$$

Hence $f_t(y+t\Delta) \le e^{-tb_1} f_t(y)$. It follows that for $z \ge 0$,

$$P(S_t > z+t\Delta) = \int_z^\infty f_t(y+t\Delta)\,dy$$
$$\le e^{-tb_1} \int_z^\infty f_t(y)\,dy = e^{-tb_1} P(S_t > z),$$

and (C3) indeed holds.

We shall now show (B.3) by checking that after substituting (B.2) into (B.3), the coefficient of x^j in the left-hand side of (B.3) is not more than in the right-hand side, for $0 \le j \le t-1$. More specifically that (with $c_{tt}=0$),

$$\text{(B.4)} \qquad (j+1)c_{t,j+1} + \frac{j}{2t}c_{tj} \le c_{tj} \quad \left[\Leftrightarrow \quad c_{t,j+1} \le \frac{1}{j+1}\left(1-\frac{j}{2t}\right)c_{tj}\right].$$

Indeed by (B.2),

$$c_{t,j+1} = \frac{2(t-1-j)}{(j+1)(2t-2-j)}c_{tj} = \frac{1}{j+1}\left(1-\frac{j}{2t-2-j}\right)c_{tj}, \qquad 0 \le j \le t-1,$$

and the right-inequality of (B.4) holds.

Acknowledgments. The author would like to thank three referees and an Associate Editor for going over the manuscript carefully, and providing useful feedback. The changes made in response to their comments have resulted in a much better paper.

Thanks also to Shouri Hu for going over the proofs and performing some of the simulations in Examples 5 and 6.

The author was supported by MOE Grant R-155-000-158-112.

REFERENCES

AGRAWAL, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. in Appl. Probab.* **27** 1054–1078. MR1358906 https://doi.org/10.2307/1427934

AGRAWAL, R., TENEKETZIS, D. and ANANTHARAM, V. (1989). Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Trans. Automat. Control* **34** 1249–1259. MR1029375 https://doi.org/10.1109/9.40770

AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

BARANSI, A., MAILLARD, O. A. and MANNOR, S. (2014). Sub-sampling for multi-armed bandits. In *Proceedings of the European Conference on Machine Learning* 13.

BERRY, D. (1972). A Bernoulli two-armed bandit. *Ann. Math. Stat.* **43** 871–897.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems*: *Sequential Allocation of Experiments*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0813698 https://doi.org/10.1007/978-94-015-3711-7

BREZZI, M. and LAI, T. L. (2002). Optimal learning and experimentation in bandit problems. *J. Econom. Dynam. Control* **27** 87–108. MR1925627 https://doi.org/10.1016/S0165-1889(01)00028-8

BURNETAS, A. N. and KATEHAKIS, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Adv. in Appl. Math.* **17** 122–142. MR1390571 https://doi.org/10.1006/aama.1996.0007

BURTINI, G., LOEPPKY, J. and LAWRENCE, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. Available at arXiv:1510.00757.

CAPPÉ, O., GARIVIER, A., MAILLARD, O.-A., MUNOS, R. and STOLTZ, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* **41** 1516–1541. MR3113820 https://doi.org/10.1214/13-AOS1119

CHANG, F. and LAI, T. L. (1987). Optimal stopping and dynamic allocation. *Adv. in Appl. Probab.* **19** 829–853. MR0914595 https://doi.org/10.2307/1427104

FABIUS, J. and VAN ZWET, J. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Stat.* **41** 1906–1916.

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177. MR0547241

GITTINS, J. C. and JONES, D. M. (1979). A dynamic allocation index for the discounted multi-armed bandit problem. *Biometrika* **66** 561–565.

GRAVES, T. L. and LAI, T. L. (1997). Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM J. Control Optim*. **35** 715–743. MR1444336 https://doi.org/10.1137/S0363012994275440

KAUFMANN, E. (2014). Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources. Ph.D. thesis.

KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2012). On Bayesian upper confidence bounds for bandit problems. *Proc*. *Mach*. *Learn*. *Res*. **22** 592–600.

KORDA, N., KAUFMANN, E. and MUNOS, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. In *NIPS* 26 1448–1456.

KOTZ, S., KOZUBOWSKI, T. J. and PODGÓRSKI, K. (2001). *The Laplace Distribution and Generalizations*: *A Revisit with Applications to Communications*, *Economics*, *Engineering*, *and Finance*. Birkhäuser, Boston, MA. MR1935481 https://doi.org/10.1007/978-1-4612-0173-1

KULESHOV, V. and PRECUP, D. (2014). Algorithms for the multi-armed bandit problem. Available at arXiv:1402.6028.

LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann*. *Statist*. **15** 1091–1114. MR0902248 https://doi.org/10.1214/aos/1176350495

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl*. *Math*. **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

SHIVASWAMY, P. and JOACHIMS, T. (2012). Multi-armed bandit problems with history. *Proc*. *Mach*. *Learn*. *Res*. **22** 1046–1054.

SUTTON, B. and BARTO, A. (1998). *Reinforcement Learning*, *an Introduction*. MIT Press, Cambridge, MA.

TEKIN, C. and LIU, M. (2010). Online algorithms for the multi-armed bandit problem with Markovian rewards. In 48*th Annual Allerton Conference on Communication*, *Control and Computing* 1675–1682.

THATHACHER, V. and SASTRY, P. S. (1985). A class of rapidly converging algorithms for learning automata. *IEEE Trans*. *Syst*. *Man Cybern*. **16** 168–175.

YAKOWITZ, S. and LOWE, W. (1991). Nonparametric bandit methods. *Ann*. *Oper. Res*. **28** 297–312. MR1105179 https://doi.org/10.1007/BF02055587