Journal of Machine Learning Research 19 (2018) 1-32 Submitted 9/17; Revised 7/18; Published 7/18

Refining the Confidence Level for Optimistic Bandit Strategies

Tor Lattimore TOR.LATTIMORE@GMAIL.COM
DeepMind
5 New Street
London, United Kingdom

Editor: Peter Auer

Abstract

This paper introduces the first strategy for stochastic bandits with unit variance Gaussian noise that is simultaneously minimax optimal up to constant factors, asymptotically optimal, and never worse than the classical upper confidence bound strategy up to universal constant factors. Preliminary empirical evidence is also promising. Besides this, a conjecture on the optimal form of the regret is shown to be false and a finite-time lower bound on the regret of any strategy is presented that very nearly matches the finite-time upper bound of the newly proposed strategy.

Keywords Stochastic bandits, sequential decision making, regret minimisation.

1. Introduction

Let $k > 1$ be the number of bandits (or arms) and $\mu \in \mathbb{R}^k$ be the unknown vector of mean payoffs so that $\mu_i \in \mathbb{R}$ is the expected payoff when playing the ith bandit (or arm). In each round $t \in [n] = \{1, 2, \ldots, n\}$ the player chooses an arm $A_t \in [k]$ based on past observations and (optionally) an independent source of randomness. After making her choice, the player observes a payoff $X_t = \mu_{A_t} + \eta_t$ where $\eta_1, \eta_2, \ldots, \eta_n$ is a sequence of independent standard Gaussian random variables. It is standard to minimise the expected pseudo-regret (from now on, just the regret). Let $\Delta_i(\mu) = \max_j \mu_j - \mu_i$ be the suboptimality gap for the ith arm. The regret over n rounds is

$$\mathcal{R}_n(\mu) = \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}(\mu)\right] .$$

Because the regret depends on the unknown payoff vector, no strategy can hope to make the regret small for all μ simultaneously. There are a number of performance metrics in the literature, two of which are described below along with a new one. To spoil the surprise, the strategy introduced in the present article is simultaneously optimal with respect to all of them.

Worst-case optimality The *worst-case* regret of a strategy is the value of the regret it suffers when faced with the worst possible μ.

$$\mathcal{R}_n^{\text{WC}} = \sup_{\mu \in \mathbb{R}^k : \Delta(\mu) \in [0,1]^k} \mathcal{R}_n(\mu) .$$

The restriction to bounded suboptimality gaps is necessary to allow an algorithm to choose each arm at least once without suffering arbitrarily large regret. Generally problems with small suboptimality gaps are the most interesting. Provided that $n \geq k$ it is known that all algorithms suffer $\mathcal{R}_n^{\text{WC}} = \Omega(\sqrt{kn})$ (Auer et al., 1995).

Asymptotic optimality The worst-case regret obscures interesting structure in the problem that becomes relevant in practice. This motivates the study of a problem-dependent metric, which demands that strategies have smaller regret on 'easier' bandit instances. A strategy is called *asymptotically optimal* if

$$\lim_{n\to\infty} \frac{\mathcal{R}_n(\mu)}{\log(n)} = \sum_{i:\Delta_i(\mu)>0} \frac{2}{\Delta_i(\mu)} \qquad \text{for all } \mu \in \mathbb{R}^d \,.$$

The name is justified by the existence of policies satisfying the definition and lower bounds by Lai and Robbins (1985) and Burnetas and Katehakis (1996) showing that consistent policies (those with sub-polynomial regret on all μ) cannot do better.

The sub-UCB criteria While asymptotic analysis is quite insightful, the ultimate quantity of interest is the finite-time regret. To make a stab at quantifying this I say an algorithm is sub-UCB if there exist universal constants $C_1, C_2 > 0$ such that for all k, n and μ it holds that

$$\mathcal{R}_n(\mu) \leq C_1 \sum_{i=1}^{k} \Delta_i(\mu) + C_2 \sum_{i:\Delta_i(\mu)>0} \frac{\log(n)}{\Delta_i(\mu)} \,. \tag{1}$$

Of course UCB (Auer et al., 2002) satisfies Eq. (1), along with many other policies as shown in Table 2 in Appendix E, which outlines the long history of algorithms for stochastic finite-armed bandits. The study of this new metric can be justified in several ways. First, it provides a forgiving finite-time analogue of asymptotic optimality. Lai and Robbins (1985) derived asymptotic optimality by making a restriction on policies (the consistent ones). Consistency is an asymptotic notion, so it is not surprising that the resulting lower bound is also asymptotic. The sub-UCB notion is suggested by making a finite-time restriction on the worst case regret. Precisely, for any strategy the finite-time instance-dependent regret can be bounded in terms of the worst case regret by

$$\mathcal{R}_n(\mu) \geq \sup_{\varepsilon\in(0,1]} \sum_{i:\Delta_i(\mu)>0} \max\left\{0, \frac{2\log\left(\frac{n}{\mathcal{R}_n^{\text{WC}}}\right) + 2\log\left(\frac{\varepsilon\Delta_i(\mu)}{8}\right)}{(1+\varepsilon)\Delta_i(\mu)}\right\} \tag{2}$$

for all μ with $\max_i \Delta_i(\mu) \leq 1/2$ (Lattimore and Szepesvári, 2018). This means that if you demand a reasonable worst case bound, then the instance-dependent regret cannot be *much* better than sub-UCB. Note that the first sum in Eq. (1) is unavoidable for policies that always choose each arm at least once, which is also necessary for any algorithm to have reasonable worst case regret. The finite-time world is not as clean as the asymptotic and it is not easy to decide how tight Eq. (2) might be, which justifies the additional constant-factor allowance in Eq. (1) and the removal of the (typically negative) second logarithm term. The second justification for using Eq. (1) as a yardstick is that it is forgiving and yet recent policies that are minimax optimal up to constant factors do not satisfy it. One of the core contributions of this article is to correct these deficiencies. Note that Eq. (2) depends quite weakly on the worst case regret and is meaningful as long as $\mathcal{R}_n^{\text{WC}} = O(n^p)$ for p not too close to 1.

None of these criteria are perfect by themselves. Asymptotic optimality is achievable by policies with outrageous burn-in time and/or large minimax regret, minimax optimal policies may be unreasonably conservative on easy problems and sub-UCB policies may be far from asymptotically optimal.

Contributions The main contribution is a new strategy called ADA-UCB ('adaptive UCB') and analysis showing it is asymptotically optimal, minimax optimal and sub-UCB. No other algorithm is simultaneously minimax optimal and sub-UCB (see Table 2). Results are specialised to the Gaussian case with unit variance, but upper bounds can be generalised to subgaussian noise with known subgaussian constant at the price of increased constants (without losing asymptotic optimality) and longer proofs. The latter justifies the specialisation because it allows for an elegant concentration analysis via an embedding of Gaussian random walks into Brownian motion. Also included:

(a) Finite-time lower bounds showing the new strategy is close to optimal.

(b) A conjecture by Bubeck and Cesa-Bianchi (2012) is proven false.

(c) A generic analysis for a large class of strategies simplifying the analysis for existing strategies.

Beyond the concrete results, the approach used for deriving ADA-UCB by examining lower bounds will likely generalise to other noise models, and indeed, other sequential optimisation problems with an exploration/exploitation flavour. The contents of this article combines the best parts of two technical reports with improved results, intuition and analysis (Lattimore, 2015a, 2016b).

Notation For natural number n let $[n] = \{1, 2, \ldots, n\}$. Binary minimums and maximums are abbreviated by $\wedge$ and $\vee$ respectively. The complement of event A is A^c. Except where otherwise stated, it is assumed without loss of generality that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k$. None of the proposed strategies depend on the labelling of the arms, so if this is not the case the indices can simply be re-ordered. The dependence of the suboptimality gap on the mean vector will usually be omitted when the context is clear: $\Delta_i = \Delta_i(\mu) = \max_j \mu_j - \mu_i$. Occasionally it is convenient to define $\mu_{k+1} = -\infty$ and $\Delta_{k+1} = \infty$. Let $T_i(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s = i\}}$ be the number of times arm i has been chosen after round t and $\hat{\mu}_i(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s=i\}} X_s / T_i(t)$ be the corresponding empirical estimate of its return. Let $\hat{\mu}_{i,s}$ be the empirical estimate of the mean of arm i after s samples from that arm so that $\hat{\mu}_{i,T_i(t)} = \hat{\mu}_i(t)$. Define σ-algebra $\mathcal{F}_t = \sigma(\xi, X_1, \ldots, X_t)$ to contain the information available to the strategy after round t, where ξ is an independent source of randomness that allows for randomness in the strategy. This means that formally a strategy is a sequence of random variables $(A_t)_t$ such that A_t is $\mathcal{F}_{t-1}$-measurable. It is assumed throughout that $n \geq k$. Finally, let $\overline{\log}(x) = \log((x + e)\log^{1/2}(x + e))$. A table of notation is available in Table 3 in the appendix.

2. The strategy

The ADA-UCB strategy chooses each arm once in arbitrary order for the first k rounds and subsequently $A_t = \arg\max_{i \in [k]} \gamma_i(t)$ where the index of arm i in round t is:

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \overline{\log}\left(\frac{n}{H_i(t-1)}\right)},$$

$$\text{with} \quad H_i(t) = T_i(t) K_i(t) \quad \text{and} \quad K_i(t) = \sum_{j=1}^{k} \min\left\{1, \sqrt{\frac{T_j(t)}{T_i(t)}}\right\}.$$

At first sight the new index seems overly complicated. After the statement of the main regret guarantee I show how the strategy is derived in a principled fashion from *lower bounds* obtained

from information theoretic limits of the problem. Similar approaches have been used before, for example, by Agrawal et al. (1989) for reinforcement learning, and by Garivier and Kaufmann (2016) for pure exploration in bandits. One interesting consequence of this approach is that ADA-UCB is not a true index strategy in the sense that $\gamma_i(t)$ depends on random variables associated with other arms. An intriguing open question is whether or not there exists an index strategy for which all three performance criteria are met. A minor observation is that it is not clear whether or not a true index strategy should be allowed to depend on t or just on $n - t$ and the samples from the given arm.

Relation to other algorithms The index is the same as that used by Katehakis and Robbins (1995) except that $\log(t)$ has been replaced by $\overline{\log}(n/H_i(t-1))$. The change from $\log(\cdot)$ to $\overline{\log}(\cdot)$ is quite minor. Such inflations of the logarithmic term are typical for algorithms with finite-time guarantees. The main difference between ADA-UCB and previous work is the term inside the logarithm, often called the confidence level. The most common choice is the current round t, which is used by various versions of UCB, KL-UCB and BAYES-UCB (Katehakis and Robbins, 1995; Burnetas and Katehakis, 1996; Agrawal, 1995; Auer et al., 2002; Kaufmann et al., 2012; Cappé et al., 2013). Already in the early work by Lai (1987) there appeared an unnamed variant of UCB for which the confidence level was $n/T_i(t-1)$. Due to its similarity to KL-UCB (Cappé et al., 2013) this algorithm will be called KL-UCB* from now on. A variety of other choices have been used as shown in Table 1.

majority	t
Lai (1987)	n/T_i
Honda and Takemura (2010)	t/T_i
Audibert and Bubeck (2009)	$n/(kT_i)$
Degenne and Perchet (2016)	$t/(kT_i)$
Lattimore (2015a)	n/t

Table 1: Confidence levels

Computation A naive implementation of ADA-UCB requires a computation time that is quadratic in the number of arms in each round. Fortunately an incremental implementation leads to an algorithm with linear computation time by noting that:

(a) If $T_i(t-1) \leq T_{A_t}(t-1)$, then $\gamma_i(t+1) = \gamma_i(t)$.

(b) For arms i with $T_i(t-1) > T_{A_t}(t-1)$ the value of $K_i(t-1)$ may be computed incrementally by $K_i(t) = K_i(t-1) - \sqrt{(T_{A_t}(t)-1)/T_i(t)} + \sqrt{T_{A_t}(t)/T_i(t)}$.

(c) The index of A_t can be computed trivially in order k time.

The algorithm follows by maintaining a list of arms sorted by $T_i(t)$ and applying the above observations to incrementally update the indices. If all details are addressed carefully, then the computation required in round t is $O(\sum_{i=1}^{k} \mathbb{1}\{T_i(t-1) \geq T_{A_t}(t-1)\})$, which in the worst case is $O(k)$, but can be much smaller when a single arm is played significantly more often than any other.

Regret bound The theorem statement has a more complicated form than previous regret bounds for finite-armed bandits, mainly because it correctly deals with the case where there are many near-optimal arms that cannot be statistically identified within the time horizon. Define k_i and λ_i by

$$k_i = \sum_{j=1}^{k} \min\left\{1, \frac{\Delta_i}{\Delta_j}\right\} \qquad \text{and} \qquad \lambda_i = 1 + \frac{1}{\Delta_i^2} \overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right).$$

The main theorem is below, which gives the best known finite-time guarantee for any strategy, as well as all three optimality criteria defined in the introduction.

Theorem 1 *Assume that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k$. Then there exists a universal constant $C > 0$ such that the regret of* ADA-UCB *is bounded by*

$$\mathcal{R}_n(\mu) \leq C \min_{i \in [k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_m \lambda_m \right), \qquad \text{where } \bar{\Delta}_i = \frac{1}{i} \sum_{m=1}^{i} \Delta_m \,. \tag{3}$$

Furthermore:

(a) ADA-UCB *is minimax optimal up to constant factors: $\mathcal{R}_n^{\text{WC}} \leq C' \sqrt{kn}$.*

(b) ADA-UCB *is sub-UCB: $\mathcal{R}_n(\mu) \leq C \sum_{m:\Delta_m > 0} \left(\Delta_m + \frac{\log(n)}{\Delta_m} \right)$.*

(c) ADA-UCB *is asymptotically optimal: $\lim_{n\to\infty} \frac{\mathcal{R}_n(\mu)}{\log(n)} = \sum_{i:\Delta_i > 0} \frac{2}{\Delta_i}$.*

Remark 2 *The assumption on the order of the arms is purely for cosmetic purposes. The algorithm does not need this ordering and treats all arms symmetrically.*

Intuition for bound and strategy Let $\mu \in [0,1]^k$ and i be a suboptimal arm so that $\Delta_i = \Delta_i(\mu) > 0$. Set $\mu' \in [0,2]^k$ equal to μ except for $\mu'_i = \mu_i + 2\Delta_i$, which means that arm i has the largest mean for the bandit determined by μ'. Provided that n is sufficiently large, a sub-UCB strategy should play arm i logarithmically often if the mean payoff vector is μ, and linearly often for μ'. Let $\mathbb{E}$ and $\mathbb{E}'$ and $\mathbb{P}$ and $\mathbb{P}'$ denote the measures on the outcomes $A_1, X_1, \ldots, A_n, X_n$ when the strategy interacts with the bandits determined by μ and μ' respectively. Let $\delta \in (0,1]$ be such that

$$\mathbb{E}[T_i(n)] = \frac{2\log(1/\delta)}{(2\Delta_i)^2} \,. \tag{4}$$

Since μ and μ' are only different in the ith coordinate, the problem of minimising the regret is essentially equivalent to a hypothesis test on the mean of the ith arm, which satisfies $|\mu'_i - \mu_i| = 2\Delta_i$. Using this idea, a standard information-theoretic argument (see the section on lower bounds for formal details) shows that $\mathbb{P}'(T_i(n) \leq n/2) \gtrsim \delta$. Abbreviate $\Delta'_i = \Delta_i(\mu')$. Since $\Delta'_j = \mu'_i - \mu'_j \geq \Delta_i$ for all $j \neq i$ it holds that

$$\mathcal{R}_n(\mu') \geq \frac{n\Delta_i}{2} \mathbb{P}'(T_i(n) \leq n/2) \gtrsim n\Delta_i \delta / 2 \,. \tag{5}$$

Assuming the strategy is sub-UCB, then there exists a (hopefully small) constant $C > 0$ such that

$$\mathcal{R}_n(\mu') \leq C \sum_{j:\Delta'_j > 0} \left(\Delta'_j + \frac{\log(n)}{\Delta'_j} \right) \approx C \sum_{j:\Delta'_j > 0} \frac{\log(n)}{\Delta'_j} \,, \tag{6}$$

where the approximation follows because $\Delta'_j \in [0,2]$ has been assumed. By Eqs. (5) and (6):

$$\delta \lesssim \frac{2C\log(n)}{n\Delta_i} \sum_{j:\Delta'_j > 0} \frac{1}{\Delta'_j} = \frac{2C\log(n)}{n\Delta_i^2} \sum_{j \neq i} \frac{\Delta_i}{\Delta_j + \Delta_i} \leq \frac{2Ck_i \log(n)}{n\Delta_i^2} \,.$$

The regret guarantee given in Theorem 1 is now justified up to constant factors and an extraneous additive $\log\log(\cdot)$ term by substituting the above display into Eq. (4) and writing the regret as $\mathcal{R}_n(\mu) = \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)]$. The idea behind ADA-UCB is to use the approximation $\Delta_i^{-2} \approx T_i(t-1)$. The approximation is poor when t is small, but becomes reasonable at the critical time when arm i should no longer be played. Specifically, if $T_i(t-1) \approx \Delta_i^{-2}$, then we should expect $T_j(t-1) \approx \min\{T_i(t-1), \Delta_j^{-2}\} \approx \min\{\Delta_i^{-2}, \Delta_j^{-2}\}$. Then

$$\frac{n}{H_i(t-1)} = \frac{n}{\sum_{j=1}^k \min\left\{T_i(t-1), \sqrt{T_i(t-1)T_j(t-1)}\right\}} \approx \frac{n}{\sum_{j=1}^k \min\left\{\frac{1}{\Delta_i^2}, \frac{1}{\Delta_i\Delta_j}\right\}} = \frac{n\Delta_i^2}{k_i} \,.$$

The implication is that the index dynamically tunes its confidence level using the pull counts to loosely estimate the gaps. The ideas in this section are made formal in the proof of Theorem 1 or the lower bound (§5).

3. Asymptotic analysis

The primary purpose of this section is to prove part (c) of Theorem 1. Along the way, a finite-time regret bound for a whole class of strategies is derived, including slightly modified versions of KL-UCB* (Lai, 1987) and the MOSS (**M**inimax **O**ptimal in the **S**tochastic **S**etting, Audibert and Bubeck 2009). The analysis leads to an optimal worst-case analysis of MOSS and KL-UCB*, but not ADA-UCB. The following theorem holds for the class of index strategies that choose $A_t = t$ for $1 \leq t \leq k$ and subsequently maximise

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \operatorname{\overline{log}}\left(\frac{n}{J_i(t-1)T_i(t-1)}\right)}, \tag{7}$$

where $J_i(t-1)$ is $\mathcal{F}_{t-1}$-measurable and $J_i(t-1) \in [a, b]$ almost surely for constants $0 < a \leq b$. Except for minor differences in the leading constant and the logarithmic term, this index is the same as MOSS if $J_i(t-1) = k$, KL-UCB* if $J_i(t-1) = 1$ and ADA-UCB if $J_i(t-1) = K_i(t-1)$.

Theorem 3 *For any $\varepsilon \in (0, 1/2)$ and $1 \leq \ell \leq k$, the regret of the strategy in Eq. (7) is at most*

$$\mathcal{R}_n(\mu) \leq n\Delta_\ell + \frac{2c_1 b}{\varepsilon^2 \Delta_{\ell+1}} + \sum_{i>\ell} \left(2\Delta_i + \frac{1}{\Delta_i}\left(1 + \frac{1}{\varepsilon^2} + \frac{2\operatorname{\overline{log}}\left(\frac{n\Delta_i^2}{a}\right)}{(1-2\varepsilon)^2}\right)\right) .$$

Furthermore:

(a) $\lim_{n\to\infty} \mathcal{R}_n(\mu)/\log(n) = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$.

(b) If $b \leq k$, then $\mathcal{R}_n^{\text{WC}} \leq C\sqrt{nk\left(1 + \log\left(\frac{k}{a}\right)\right)}$ where $C > 0$ is a universal constant.

Before the proof a little more notation is required. For each i and $\Delta > 0$ let $\zeta_i(\Delta)$ be a random variable given by

$$\zeta_i(\Delta) = 1 + \max\left\{s : \hat{\mu}_{i,s} > \mu_i + \Delta\right\} . \tag{8}$$

Clearly, $\zeta_i(\Delta)$ is surely monotone non-increasing in Δ and may be upper bounded in expectation using Lemma 13 in the appendix.

Proof [of Theorem 3] Let $\Delta \in \mathbb{R}$ be the smallest value such that

$$\hat{\mu}_{1,s} + \sqrt{\frac{2}{s} \overline{\log}\left(\frac{n}{bs}\right)} \leq \mu_1 - \Delta \qquad \text{for all } 1 \leq s \leq n\,,$$

which is chosen so that $\gamma_1(t) \geq \mu_1 - \Delta$ for all t. By part (a) of Lemma 12 with $\alpha = n/b$ and $d = 1$ and $\lambda_1 = \infty$ we have for any $x > 0$ that

$$\mathbb{P}\left(\Delta \geq x\right) \leq \frac{c_1 b h_\lambda(1/x^2)}{n} = \frac{c_1 b}{n x^2}\,. \tag{9}$$

Define random variable

$$\Lambda_i = 1 + \max\left\{\frac{1}{\Delta_i^2},\, \zeta_i(\varepsilon\Delta_i),\, \frac{2}{(1-2\varepsilon)^2\Delta_i^2} \overline{\log}\left(\frac{n\Delta_i^2}{a}\right)\right\}.$$

The definitions of the policy, Λ_i and Δ ensure that if $\Delta_i > \Delta/\varepsilon$, then $T_i(n) \leq \Lambda_i$ and by Lemma 13,

$$\mathbb{E}[\Lambda_i] \leq 2 + \frac{1}{\Delta_i^2}\left(1 + \frac{1}{\varepsilon^2} + \frac{2\,\overline{\log}\left(\frac{n\Delta_i^2}{a}\right)}{(1-2\varepsilon)^2}\right). \tag{10}$$

Hence the regret of the strategy maximising the index in Eq. (7) is

$$\begin{aligned}
\mathcal{R}_n(\mu) &= \mathbb{E}\left[\sum_{i=1}^k \Delta_i T_i(n)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\ell} \Delta_i T_i(n)\right] + \mathbb{E}\left[\sum_{i=\ell+1}^{k} \mathbb{1}\left\{\Delta_i \leq \frac{\Delta}{\varepsilon}\right\} \Delta_i T_i(n)\right] + \mathbb{E}\left[\sum_{i=\ell+1}^{k} \mathbb{1}\left\{\Delta_i > \frac{\Delta}{\varepsilon}\right\} \Delta_i T_i(n)\right] \\
&\leq n\Delta_\ell + \mathbb{E}\left[\frac{n\Delta}{\varepsilon}\mathbb{1}\left\{\Delta \geq \varepsilon\Delta_{\ell+1}\right\}\right] + \sum_{i>\ell} \Delta_i \mathbb{E}\left[\Lambda_i\right]\,.
\end{aligned} \tag{11}$$

The last expectation in Eq. (11) is bounded using Eq. (10),

$$\sum_{i>\ell} \Delta_i \mathbb{E}\left[\Lambda_i\right] \leq \sum_{i>\ell}\left(2\Delta_i + \frac{1}{\Delta_i}\left(1 + \frac{1}{\varepsilon^2} + \frac{2\,\overline{\log}\left(\frac{n\Delta_i^2}{a}\right)}{(1-2\varepsilon)^2}\right)\right). \tag{12}$$

Given an arbitrary random variable X and constant $b \in \mathbb{R}$ it holds that

$$\mathbb{E}[X] \leq b\mathbb{P}\left(X \geq b\right) + \int_b^\infty \mathbb{P}\left(X \geq x\right) dx\,.$$

The first expectation in Eq. (11) is bounded by combining the above display with Eq. (9),

$$\begin{aligned}
\frac{n}{\varepsilon}\mathbb{E}\left[\Delta\mathbb{1}\left\{\Delta \geq \varepsilon\Delta_{\ell+1}\right\}\right] &\leq n\Delta_{\ell+1}\mathbb{P}\left(\Delta \geq \varepsilon\Delta_{\ell+1}\right) + \frac{n}{\varepsilon}\int_{\varepsilon\Delta_{\ell+1}}^\infty \mathbb{P}\left(\Delta \geq x\right) dx \\
&\leq \frac{c_1 b}{\varepsilon^2\Delta_{\ell+1}} + \frac{c_1 b}{\varepsilon}\int_{\varepsilon\Delta_{\ell+1}}^\infty \frac{dx}{x^2} = \frac{2c_1 b}{\varepsilon^2\Delta_{\ell+1}}\,,
\end{aligned}$$

which together with Eq. (12) and Eq. (11) completes the proof of the first part. The asymptotic result follows by choosing $\ell = \max\{i : \Delta_i = 0\}$ and $\varepsilon = \log^{-1/4}(n)$. The equality follows by the lower bound of Lai and Robbins (1985). The worst-case bound follows by choosing $\varepsilon = 1/4$ and tuning the cut-off ℓ. ■

The failure of MOSS Notice that if $J_i(t-1) = k$, then $a = b = k$ and except for a smaller inflated logarithm the resulting policy is the same as the variant proposed by Ménard and Garivier (2017). The troublesome term in the regret is the second term, which when $\ell = 1$ is approximately k/Δ_2. By contrast, for more conservative strategies this term is approximately $1/\Delta_2$, which is negligible. An especially challenging regime is when $\Delta_2 = 1/k$ and $\Delta_i = 1$ for $i > 2$. Suppose now that $n = k^3$ and k is large, then the regret of UCB on this problem should be

$$\mathcal{R}_n = O\left(k \log(k)\right) .$$

For MOSS, however, the regret on this problem is $\Omega(\sqrt{nk}) = \Omega(k^2)$, which for large k is arbitrarily worse than UCB. A vague explanation of the poor performance is that although there are k arms, all but two of them are so suboptimal that *effectively* it is a two-armed bandit. And yet MOSS is heavily tuned for the k-armed case and suffers as a consequence. A more precise argument is that distinguishing the first and second arms requires $T_2(t) \approx T_1(t) \approx 1/\Delta_2^2 = k^2$. But after playing these arms roughly k^2 times each the confidence level is $n/(kT_i(t)) \approx 1$ and the likelihood of misidentification is large enough that the regret is $\Omega(n\Delta_2) = \Omega(k^2) = \Omega(\sqrt{nk})$. Note that for ADA-UCB we would expect $K_i(t) \approx 2$ and the confidence level is approximately $n/k^2 = k$, which is exactly as large as necessary.

Remark 4 *An empirical study in this problem was given in a previous technical report (Lattimore, 2015a). In practice the failure does not become extreme until k is very large (approximately 1000).*

4. Finite-time analysis

In this section the remainder of Theorem 1 is proven. The argument is quite long, but never terribly complicated. The main novel challenge is to deal with the dependence of the index of one arm on the number of plays of other arms. The usual program for analysing strategies based on upper confidence bounds has two parts:

(a) Show that with high probability the index of the optimal arm is never much smaller than its mean.

(b) Show that the index of each suboptimal arm drops below the mean of the optimal arm after not too many plays with high probability.

The proof starts with (b), the main component of which is showing for suboptimal arms i that $H_i(t)$ grows at a reasonable rate. As discussed, this means showing that other arms are played sufficiently often. The following definitions spell out *which* arms will be played reasonably often. For each arm i

define a deterministic set of arms $V_i \subset [k]$ and random subset $W_i \subseteq V_i$ by

$$V_i = \{j \in [k] : j \text{ is even or } j \geq i\} \ . \tag{13}$$

$$W_i = \{i\} \cup \left\{ j \in V_i : \min_{1 \leq s \leq \Delta_i^{-2}} \hat{\mu}_{j,s} + \sqrt{\frac{2}{s} \overline{\log}\left(\frac{n\Delta_i}{k_i\sqrt{s}}\right)} - \sqrt{\frac{2}{s}} \geq \mu_j \right\} . \tag{14}$$

The set W_i is a subset of V_i that includes arm i and those arms for which the empirical mean is always nearly as large as the true mean. We'll soon see that arms $j \in W_i$ will be played sufficiently often to ensure that $H_i(t)$ grows at the right rate. The exclusion of odd arms with $j < i$ from V_i is for technical reasons. In order to show that arm i is not played too often we need to show that its index drops sufficiently fast and that the index of some other arm is sufficiently large. The separation of the arms allows us to exploit the independence between the arms. Those arms not in V_i will be used to show that some index is large enough. Define

$$\delta_i = c_2 \sqrt{\frac{k_i}{n\Delta_i^2}} \qquad \text{and} \qquad \ell = \max\{i : \delta_i > 1/4\} \ . \tag{15}$$

It is shown in Lemma 21 in the appendix that there exists a universal constant $C > 0$ such that ℓ satisfies

$$n\bar{\Delta}_\ell + \sum_{m>\ell} \lambda_m \Delta_m \leq C \min_{i \in [k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_m \lambda_m \right) , \tag{16}$$

and so the rest of the proof is devoted to bounding the regret of ADA-UCB in terms of the left-hand-side of Eq. (16). Define F_i to be the event that the 'mass' of W_i is sufficiently large.

$$F_i = \mathbb{1}\left\{ \sum_{m \in W_i} \min\{1, \Delta_i/\Delta_m\} \geq k_i/8 \right\} . \tag{17}$$

Recall that $k_i = \sum_{m=1}^k \min\{1, \Delta_i/\Delta_m\}$. So F_i holds if the sum over the restricted set W_i is at most a factor of 8 smaller. The point is that if $j \in V_i$, then Lemma 12 implies $\mathbb{P}(j \notin W_i) \leq \delta_i \leq 1/4$. Later this will be combined with Hoeffding's bound to show that F_i occurs with high probability. And now the lemmas begin. First up is to show that arms $j \in W_i$ are played sufficiently often relative to arm i. This will then be used to show that $H_i(t)$ grows at the right rate, which leads to the conclusion of the proof of part (b) in the outline in Lemma 7.

Lemma 5 *If $A_t = i$ and $H_i(t-1) \leq k_i/\Delta_i^2$ and $T_i(t-1) \geq \zeta_i(\Delta_i) \vee 1/\Delta_i^2$, then for all $j \in W_i$,*

$$T_j(t-1) \geq \min\left\{ \frac{1}{2\Delta_j^2}, \frac{T_i(t-1)}{4\,\overline{\log}\left(\frac{n}{H_i(t-1)}\right)} \right\} .$$

Proof If $i = j$ or $T_j(t-1) \geq T_i(t-1)$ or $T_j(t-1) \geq 1/(2\Delta_j^2)$, then we are done, so assume from now on that none of these are true.

$$\begin{aligned} \frac{k_i}{\Delta_i^2} \geq H_i(t-1) &= \sqrt{T_j(t-1)} \sum_{m=1}^k \min\left\{ \frac{T_i(t-1)}{\sqrt{T_j(t-1)}}, \sqrt{\frac{T_i(t-1)T_m(t-1)}{T_j(t-1)}} \right\} \\ &\geq \frac{\sqrt{T_j(t-1)}}{\Delta_i} \sum_{m=1}^k \min\left\{ 1, \sqrt{\frac{T_m(t-1)}{T_j(t-1)}} \right\} = \sqrt{T_j(t-1)} \frac{K_j(t-1)}{\Delta_i} , \end{aligned} \tag{18}$$

where the first inequality is assumed in the lemma statement and the second because $T_i(t-1) \geq T_j(t-1) \vee (1/\Delta_i^2)$. Therefore if $j \in W_i$, then

$$\begin{aligned}
\mu_1 + \sqrt{\frac{2}{T_i(t-1)} \overline{\log}\left(\frac{n}{H_i(t-1)}\right)} &\geq \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \overline{\log}\left(\frac{n}{H_i(t-1)}\right)} \\
&= \gamma_i(t) \geq \gamma_j(t) = \hat{\mu}_j(t-1) + \sqrt{\frac{2}{T_j(t-1)} \overline{\log}\left(\frac{n}{T_j(t-1)K_j(t-1)}\right)} \\
&\geq \hat{\mu}_j(t-1) + \sqrt{\frac{2}{T_j(t-1)} \overline{\log}\left(\frac{n\Delta_i}{k_i\sqrt{T_j(t-1)}}\right)} \\
&\geq \mu_j + \sqrt{\frac{2}{T_j(t-1)}} \geq \mu_1 + \sqrt{\frac{1}{2T_j(t-1)}} \qquad (19)
\end{aligned}$$

where the first inequality follows from the assumption that $T_i(t-1) \geq \zeta_i(\Delta_i)$, which ensures that $\mu_1 = \mu_i + \Delta_i \geq \hat{\mu}_i(t-1)$. The second follows from Eq. (18). The inequalities in Eq. (19) from the definition of $j \in W_i$ and the assumption that $T_j(t-1) \geq 1/(2\Delta_j^2)$. Therefore

$$T_j(t-1) \geq \frac{T_i(t-1)}{4\,\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}\,. \qquad \blacksquare$$

The next lemma uses the previous result to show that if W_i is large enough (F_i holds), then $H_i(t-1)$ is reasonably large at the critical point when $T_i(t-1) \approx 1/\Delta_i^2$.

Lemma 6 *If F_i holds and $T_i(t-1) \geq \zeta_i(\Delta_i) \vee 128/\Delta_i^2$ and $A_t = i$, then*

$$\overline{\log}\left(\frac{n}{H_i(t-1)}\right) \leq 2\,\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right)\,.$$

Proof If $H_i(t-1) > k_i/\Delta_i^2$, then there is nothing more to do. Otherwise, by Lemma 5

$$\begin{aligned}
H_i(t-1) &= \sum_{m=1}^{k} \min\left\{T_i(t-1), \sqrt{T_i(t-1)T_m(t-1)}\right\} \\
&\geq \sum_{m \in W_i} \min\left\{T_i(t-1), \sqrt{T_i(t-1)T_m(t-1)}\right\} \\
&\geq \sum_{m \in W_i} \min\left\{T_i(t-1), \frac{T_i(t-1)}{2\,\overline{\log}^{\frac{1}{2}}\left(\frac{n}{H_i(t-1)}\right)}, \frac{\sqrt{T_i(t-1)/2}}{\Delta_m}\right\} \qquad (20) \\
&\geq \frac{8\sum_{m \in W_i} \min\left\{\frac{1}{\Delta_i^2}, \frac{1}{\Delta_i\Delta_m}\right\}}{\overline{\log}^{\frac{1}{2}}\left(\frac{n}{H_i(t-1)}\right)} \geq \frac{k_i}{\Delta_i^2\,\overline{\log}^{\frac{1}{2}}\left(\frac{n}{H_i(t-1)}\right)}\,, \qquad (21)
\end{aligned}$$

where Eq. (20) follows from Lemma 5. The first inequality in Eq. (21) follows because $\overline{\log}(x) \geq 1$ and the assumption $T_i(t-1) \geq 128/\Delta_i^2$, and the second because F_i holds. The result follows via

re-arrangement and some algebraic trickery with the $\overline{\log}(\cdot)$ function using Part (v) of Lemma 20 in the appendix. ■

For each arm i define random variable Λ_i that will be shown to be a high probability bound on $T_i(n)$ and approximately equal to λ_i in expectation.

$$\Lambda_i = 1 + \max\left\{ \frac{128}{\Delta_i^2},\ \zeta_i(\Delta_i/3),\ \frac{36}{\Delta_i^2}\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right),\ \frac{181\!\!1_{\{F_i^c\}}}{\Delta_i^2}\overline{\log}\left(n\Delta_i^2\right) \right\},$$

where $\zeta_i(\cdot)$ is defined in Eq. (8). The next lemma is a simple consequence of the previous two and shows that if $T_i(t-1)+1 \geq \Lambda_i$, then either arm i is not played or its index is smaller than the mean of the optimal arm by a margin of at least $\Delta_i/3$.

Lemma 7 *If $T_i(t-1)+1 \geq \Lambda_i$, then either $A_t \neq i$ or $\gamma_i(t) \leq \mu_1 - \Delta_i/3$.*

Proof If F_i does not occur, then $H_i(t-1) \geq T_i(t-1) \geq 1/\Delta_i^2$ and so

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}{T_i(t-1)}} \leq \mu_i + \frac{\Delta_i}{3} + \sqrt{\frac{2\overline{\log}\left(n\Delta_i^2\right)}{T_i(t-1)}} \leq \mu_1 - \frac{\Delta_i}{3},$$

where the first inequality follows because $T_i(t-1)+1 \geq \Lambda_i \geq 1 + \zeta_i(\Delta_i/3)$ and the second because $H_i(t-1) = T_i(t-1)K_i(t-1) \geq T_i(t-1) \geq 1/\Delta_i^2$ and the definition of Λ_i and because $\mu_i + \Delta_i/3 = \mu_1 - 2\Delta_i/3$. Next suppose that F_i occurs, then either $A_t \neq i$ and the result is true, or $A_t = i$ and so by Lemma 6

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}{T_i(t-1)}} \leq \mu_i + \frac{\Delta_i}{3} + \sqrt{\frac{4\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right)}{T_i(t-1)}} \leq \mu_1 - \frac{\Delta_i}{3}. \qquad \blacksquare$$

This essentially completes the first part of the proof by showing that if $T_i(t-1)+1 \geq \Lambda_i$, then arm i is either not played or its index is not too large. The next step is to show that with reasonable probability there is some near-optimal arm for which the index is big enough to prevent arm i from being played. The value of Λ_i has been carefully chosen to bound the number of times arm i can be played provided the index of some other arm is always larger than $\mu_1 - \Delta_i/3$. But more than this, Λ_i does not depend on the rewards for odd-index arms $j < i$ so is measurable with respect to the σ-algebra $\sigma(\hat{\mu}_{j,s} : j \in V_i, 1 \leq s \leq n)$. Define random variable $I \in [k]$ to be the arm with the largest mean such that there exists an odd $j < I+1$ with $\Delta_j \leq \Delta_{I+1}/6$ and

$$\min_{1 \leq s \leq n} \hat{\mu}_{j,s} + \sqrt{\frac{2}{s}\overline{\log}\left(\frac{n}{sI + \sum_{m>I}\min\left\{s, \sqrt{s\Lambda_m}\right\}}\right)} > \mu_j - \frac{\Delta_{I+1}}{6}. \tag{22}$$

The definition of I implies that arms $i > I$ will not be played once their index drops far enough below the mean of the optimal arm. We should hope that I is small with reasonably high probability, with the best case being $I = 1$, which occurs when the optimal arm is always optimistic. Notice that I is well-defined because of the convention that $\Delta_{k+1} = \infty$. Let E_1 be the event that $I \leq \ell$, where ℓ is given in Eq. (15).

Lemma 8 *The regret of* ADA-UCB *is bounded by*

$$\mathcal{R}_n(\mu) \leq \mathbb{E}\left[\sum_{i>\ell} \Delta_i \Lambda_i\right] + n\mathbb{E}\left[\mathbb{1}_{\{E_1^c\}} \Delta_I\right] + \mathbb{E}\left[\mathbb{1}_{\{E_1\}} \sum_{i=1}^{\ell} \Delta_i T_i(n)\right] .$$

Proof The first task is to show that $T_i(n) < \Lambda_i$ for all $i > I$, which follows by induction over rounds $k+1 \leq t \leq n$. Starting with the base case, note that when $t = k+1$ we have $T_i(t-1) = 1 < \Lambda_i$ for all i. Now suppose for $t \geq k+1$ that $T_i(t-1) < \Lambda_i$ for all $i > I$. By the definition of I there must exist an odd j with $\Delta_j \leq \Delta_{I+1}/6$ that satisfies Eq. (22). For this arm we have

$$\begin{aligned} H_j(t-1) &= \sum_{m=1}^{k} \min\left\{T_j(t-1), \sqrt{T_j(t-1)T_m(t-1)}\right\} \\ &< T_j(t-1)I + \sum_{m>I} \min\left\{T_j(t-1), \sqrt{T_j(t-1)\Lambda_m}\right\} . \end{aligned}$$

Therefore

$$\gamma_j(t) = \hat{\mu}_j(t-1) + \sqrt{\frac{2}{T_j(t-1)} \overline{\log}\left(\frac{n}{H_j(t-1)}\right)} > \mu_j - \Delta_{I+1}/6 \geq \mu_1 - \Delta_{I+1}/3 .$$

It follows that if $i > I$ and $T_i(t-1) + 1 \geq \Lambda_i$, then Lemma 7 implies $A_t \neq i$. Therefore $T_i(t) < \Lambda_i$ for all $i > I$, which completes the induction. Finally, since $\sum_{i=1}^{k} T_i(n) = n$ and using the definition of E_1 as the event that $I \leq \ell$, the regret may be bounded by

$$\mathcal{R}_n(\mu) = \mathbb{E}\left[\sum_{i=1}^{k} \Delta_i T_i(n)\right] \leq \mathbb{E}\left[\sum_{i>\ell} \Delta_i \Lambda_i\right] + n\mathbb{E}[\mathbb{1}_{\{E_1^c\}} \Delta_I] + \mathbb{E}\left[\mathbb{1}_{\{E_1\}} \sum_{i=1}^{\ell} \Delta_i T_i(n)\right] . \quad \blacksquare$$

The last step in the proof of Theorem 1 is to bound each of the expectations in Lemma 8.

Lemma 9 *There exists a universal $C > 0$ such that:*

$$\textit{(i)}\ \ \mathbb{E}\left[\sum_{i>\ell} \Delta_i \Lambda_i\right] \leq C \sum_{i>\ell} \Delta_i \lambda_i \qquad \textit{(ii)}\ \ \mathbb{E}[\mathbb{1}_{\{E_1^c\}} \Delta_I] \leq C\left(n\bar{\Delta}_\ell + \sum_{i>\ell} \Delta_i \lambda_i\right)$$

$$\textit{(iii)}\ \ \mathbb{E}\left[\mathbb{1}_{\{E_1\}} \sum_{i=1}^{\ell} \Delta_i T_i(n)\right] \leq C\left(n\bar{\Delta}_\ell + \sum_{i>\ell} \Delta_i \lambda_i\right) .$$

The proof of part *(i)* follows shortly. The proof of part *(ii)* is deferred to Appendix B. Very briefly, it follows somewhat directly from part (a) of Lemma 12 (notice the similarity between the lemma and Eq. (22) that defines Δ_I). Part *(iii)* would be trivial if one assumed that $\mathbb{E}[T_i(n)]$ is monotone non-increasing in i. This seems likely, but despite significant effort I was only able to show that this is approximately true using a complicated proof (details are in Appendix C).

Proof [of Lemma 9 (*i*)] The proof follows by bounding $\mathbb{E}[\Lambda_i]$ for each arm $i > \ell$. Naively bounding the max in the definition of Λ_i by a sum shows that

$$\mathbb{E}[\Lambda_i] \leq 1 + \frac{128}{\Delta_i^2} + \frac{36}{\Delta_i^2} \overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right) + \mathbb{E}\left[\zeta_i\left(\frac{\Delta_i}{3}\right)\right] + \frac{18\mathbb{P}\left(F_i^c\right)}{\Delta_i^2} \overline{\log}\left(n\Delta_i^2\right) . \tag{23}$$

The first three terms are non-random. The second last term is bounded using Lemma 13 by $\mathbb{E}[\zeta_i(\Delta_i/3)] \leq 1 + 18/\Delta_i^2$. For the last term we need to upper bound $\mathbb{P}\left(F_i^c\right)$, where F_i is the event defined in Eq. (17). In order to do this we need to show that W_i is reasonably large with high probability. Let $\chi_j = \mathbb{1}\left\{j \notin W_i\right\}$, which for $j \in V_i$ satisfies $\mathbb{E}[\chi_j] \leq \delta_i \leq 1/4$. Then

$$\mathbb{P}\left(F_i^c\right) = \mathbb{P}\left(\sum_{j \in W_i} \min\left\{1, \frac{\Delta_i}{\Delta_j}\right\} < \frac{k_i}{8}\right) \leq \mathbb{P}\left(\sum_{j \in V_i} (\chi_j - \mathbb{E}[\chi_j]) \min\left\{1, \frac{\Delta_i}{\Delta_j}\right\} > \frac{k_i}{4}\right)$$
$$\leq \exp\left(-\frac{k_i^2}{8\sum_{j \in V_i} \min\left\{1, \frac{\Delta_i}{\Delta_j}\right\}^2}\right) \leq \exp\left(-\frac{k_i}{8}\right) ,$$

where the first inequality follows from the facts that $\mathbb{E}[\chi_j] \leq 1/4$ and $\sum_{j \in V_i} \min\{1, \Delta_i/\Delta_j\} \geq k_i/2$ and the second inequality follows from Hoeffding's bound and the fact that χ_j are independent for $j \in V_i$. Therefore

$$\mathbb{P}\left(F_i^c\right) \overline{\log}\left(n\Delta_i^2\right) \leq \overline{\log}\left(n\Delta_i^2\right) \exp\left(-k_i/8\right) \leq 4\,\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right) ,$$

which holds because $k_i \geq 2$ is guaranteed for all suboptimal arms i. The proof is completed by substituting the above display into Eq. (23) and using the definition of $\lambda_i = 1 + \frac{1}{\Delta_i^2}\,\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right)$. ■

Proof [of Theorem 1] The finite-time bound Eq. (3) follows by substituting the bounds given in Lemma 9 into Lemma 8 and Eq. (16). Minimax and sub-UCB results are derived as corollaries of the finite-time bound Eq. (3) via Parts (iii) and (iv) of Lemma 21 in the appendix. The asymptotic analysis has been given already in §3. ■

5. A lower bound

I now formalise the intuitive argument for the regret guarantee given in §2. The results show that in a certain sense the upper bound in Theorem 1 is very close to optimal. The following lower bound holds for all strategies, but does not give a lower bound for all μ simultaneously. Related results have been proven by a variety of authors (Kulkarni and Lugosi, 2000; Bubeck et al., 2013; Salomon et al., 2013; Lattimore, 2015b), with the most related by Garivier et al. (2016b). The most significant difference between that work and the present article is that the lower-order terms are more carefully considered here, and besides this, the assumptions, and so also results, are different.

Theorem 10 *Fix a strategy and let $\mu \in \mathbb{R}^k$ be such that $n\Delta_i^2 \geq 2k_i \log(n)$ and $\Delta_i \leq 1$ for all i with $\Delta_i > 0$. Then one of the following holds:*

(a) $\mathcal{R}_n(\mu) \geq \frac{1}{2} \sum_{i:\Delta_i > 0} \frac{1}{\Delta_i} \log\left(\frac{n\Delta_i^2}{2k_i \log(n)}\right)$.

(b) There exists a $\mu' \in \mathbb{R}^k$ and i with $\Delta_i > 0$ such that $\mathcal{R}_n(\mu') \geq \frac{1}{4} \sum_{i:\Delta_i' > 0} \frac{1}{\Delta_i'} \log(n)$,

where $\mu_i' = \mu_i + 2\Delta_i$ and $\mu_j' = \mu_j$ for $j \neq i$ and $\Delta_i' = \Delta_i(\mu')$.

Proof Suppose that (a) does not hold, then there exists a suboptimal arm i such that

$$\mathbb{E}[T_i(n)] \leq \frac{1}{2\Delta_i^2} \log\left(\frac{n\Delta_i^2}{2k_i \log(n)}\right) . \tag{24}$$

Let μ' be as defined in the second part of the lemma and write $\mathbb{P}'$ and $\mathbb{E}'$ for expectation when rewards are sampled from μ'. Then by Lemmas 18 and 19 in Appendix D we have

$$\mathbb{P}\left(T_i(n) \geq n/2\right) + \mathbb{P}'(T_i(n) < n/2) \geq \frac{k_i \log(n)}{n\Delta_i^2} \triangleq 2\delta\,.$$

By Markov's inequality and Eq. (24) and the fact that $k_i \geq 2$,

$$\mathbb{P}\left(T_i(n) \geq n/2\right) \leq \frac{2\mathbb{E}[T_i(n)]}{n} \leq \frac{1}{n\Delta_i^2} \log\left(\frac{n\Delta_i^2}{2k_i \log(n)}\right) \leq \frac{k_i \log(n)}{2n\Delta_i^2} = \delta\,.$$

Therefore $\mathbb{P}'(T_i(n) < n/2) \geq \delta$, which implies that

$$\mathcal{R}_n(\mu') \geq \frac{\delta n \Delta_i}{2} = \frac{1}{4}\sum_{j=1}^{k} \min\left\{\frac{1}{\Delta_i}, \frac{1}{\Delta_j}\right\} \log(n) \geq \frac{1}{4} \sum_{j:\Delta'_j > 0} \frac{1}{\Delta'_j} \log(n)\,. \qquad \blacksquare$$

A conjecture is false It was conjectured by Bubeck and Cesa-Bianchi (2012) that the optimal regret might have approximately the following form.

$$\mathcal{R}_n(\mu) \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \overline{\log}\left(\frac{n}{H}\right)\right) \qquad \text{for all } \mu \text{ and } n \text{ and } k\,, \tag{25}$$

where $C > 0$ is a universal constant and $H = \sum_{i:\Delta_i > 0} \Delta_i^{-2}$ is a quantity that appears in the best-arm identification literature (Bubeck et al., 2009; Audibert and Bubeck, 2010; Jamieson et al., 2014).

Theorem 11 *There does not exist a strategy for which Eq.* (25) *holds.*

Proof Let $k \geq 2$ and $\mu_1 = 0$ and $\mu_2 = -1/k$ and $\mu_i = -1$ for $i > 2$, which implies that $H = k^2 + k - 2 \geq n$. For the rest of the proof we view the horizon $n = k^2$ to be a function of k. Suppose that $\mathcal{R}_n(\mu) = o(k \log k)$, which must be true for any strategy witnessing Eq. (25). Then $\min_{i>2} \mathbb{E}\left[T_i(n)\right] = o(\log k)$. Let $i = \arg\min_{i>2} \mathbb{E}\left[T_i(n)\right]$ and define μ' to be equal to μ except for the ith coordinate, which has $\mu'_i = 1$. Let A be the event that $T_i(n) \geq n/2$ and let $\mathbb{P}$ and $\mathbb{P}'$ be measures on the space of outcomes induced by the interaction between the fixed strategy and the bandits determined by μ and μ' respectively. Then for all $\varepsilon > 0$,

$$\begin{aligned}
\mathcal{R}_n(\mu) + \mathcal{R}_n(\mu') &\geq \frac{n}{2}\left(\mathbb{P}\left(A\right) + \mathbb{P}'(A^c)\right) \geq \frac{n}{4}\exp\left(-\operatorname{KL}(\mathbb{P}, \mathbb{P}')\right) \\
&= \frac{k^2}{4}\exp\left(-2\mathbb{E}\left[T_i(n)\right]\right) = \omega(k^{2-\varepsilon})\,,
\end{aligned}$$

By the assumption on $\mathcal{R}_n(\mu)$ and for suitably small ε we have $\mathcal{R}_n(\mu') = \omega(k^{2-\varepsilon})$. But as the number of arms $k \to \infty$ (and so also the horizon), this cannot be true for any policy satisfying Eq. (25), or even Eq. (1). Therefore the conjecture is not true. $\blacksquare$

6. Empirical evaluation

ADA-UCB is compared to UCB (Katehakis and Robbins, 1995), MOSS (Ménard and Garivier, 2017), THOMPSON SAMPLING (Agrawal and Goyal, 2012) and IMED (Honda and Takemura, 2015),[1] where the reference indicates the source of the algorithm. All algorithms choose each arm once and subsequently:

$$A_t^{\text{IMED}} = \operatorname*{arg\,min}_{i \in [k]} \frac{T_i(t-1)}{2} (\hat{\mu}_i(t-1) - \max_{j \in [k]} \hat{\mu}_j(t-1))^2 + \log(T_i(t-1)) \,.$$

$$A_t^{\text{UCB}} = \operatorname*{arg\,max}_{i \in [k]} \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t)}{T_i(t-1)}} \,.$$

$$A_t^{\text{MOSS}} = \operatorname*{arg\,max}_{i \in [k]} \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \log_+ \left(\frac{n}{k T_i(t-1)} \log_+^2 \left(1 + \frac{n}{k T_i(t-1)} \right) \right)} \,.$$

$$A_t^{\text{TS}} = \operatorname*{arg\,max}_{i \in [k]} \theta_i(t) \qquad \text{with } \theta_i(t) \sim \mathcal{N}(\hat{\mu}_i(t-1), 1/T_i(t-1)) \,.$$

The logarithmic term used by Ménard and Garivier (2017) in their version of MOSS is larger than $\overline{\log}(\cdot)$ and this negatively affects its performance. If the variant proposed in Section 3 is used instead, then it becomes comparable to ADA-UCB on the experiments described below, but still fails on the computationally expensive experiment given in the previous technical report (Lattimore, 2015a). For all other algorithms I have chosen the variant for which (a) guarantees exist and (b) the empirical performance is best. In all plots N indicates the number of independent samples per data point and confidence bands are calculated at a 95% level. The first three plots in Figure 1 show the regret in the worst case regime where all suboptimal arms have the same suboptimality gap. Unsurprisingly the relative advantage of policies with well-tuned confidence levels increases with the number of arms. At its worst, UCB suffers about three times the regret of ADA-UCB. Coincidentally, $\sqrt{\log(10^4)} \approx 3.03$. Figure 2 shows the regret as a function of the horizon n on a fixed bandit with $k = 20$ arms (see caption of figure for means). The regret of ADA-UCB is again a little better than the alternatives.

7. Discussion

Anytime strategies The ADA-UCB strategy depends on the horizon n, which may sometimes be unknown. The natural idea is to replace n by t, which indeed leads to a reasonable strategy that enjoys the same guarantees as ADA-UCB provided the $\overline{\log}(\cdot)$ function is replaced by something fractionally larger. The analysis, however, is significantly longer and is not included. Interested readers may refer to the technical report (Lattimore, 2016b) for the core ideas, but more work is required to find a clean proof.

Multiple optimal arms The finite-time bound in Theorem 1 is the first that demonstrates an improvement when there are multiple (near-)optimal arms. The gain in terms of the expected regret is not very large because k_i (which grows as optimal arms are added) appears only in the denominator of the logarithm. There is, however, a more significant advantage when there are many optimal arms, which (up to a point) is an exponential decrease in the variance of most strategies. This can be

1. IMED is usually defined for bandits where the rewards have (semi-)bounded support, but Junya Honda kindly provided unpublished details of the adaptation to the Gaussian case.

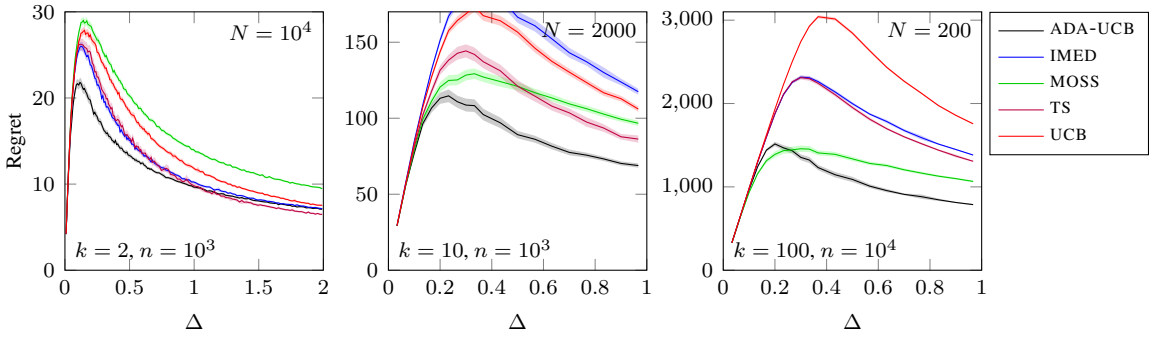

Figure 1: The regret of various algorithms as a function of Δ when $\mu = (\Delta, 0, \ldots, 0)$ and the number of arms is 2, 10 and 100 respectively. The y-axis shows the regret averaged over N independent samples for each data point.

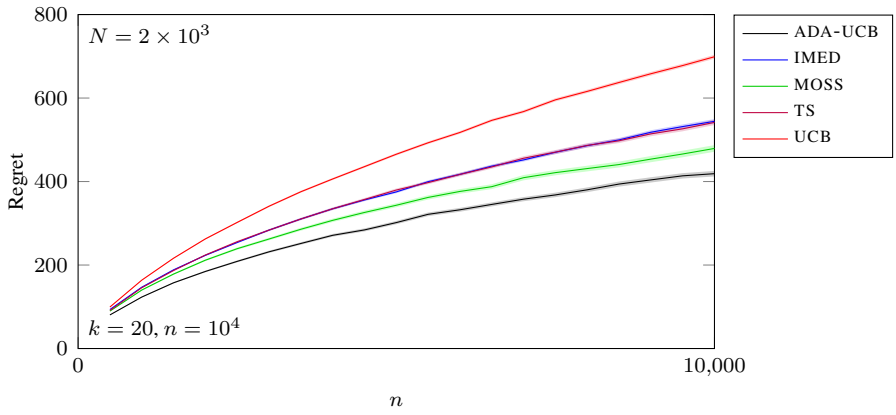

Figure 2: The regret of various algorithms as a function of the horizon for the Gaussian bandit with $k = 20$ arms and payoff vector $\mu =$ (0, -0.03, -0.03, -0.07, -0.07, -0.07, -0.15, -0.15, -0.15, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -1, -1). ADA-UCB is again outperforming the competitors. Note that IMED and THOMPSON SAMPLING are so similar they cannot be distinguished in the plot.

extracted from the analysis by observing that the high variance is caused by the possibility that an optimal arm is not sufficiently optimistic, but the probability of this occurring drops exponentially as the number of optimal arms increases.

Alternative noise models and other extensions The most obvious open question is how to generalise the results to a broader class of noise models and setups. I am quite hopeful that this is possible for noise from exponential families, though the analysis will necessarily become more complicated because the divergences become more cumbersome to work with than the squared distance that is the divergence in the Gaussian case. An alternative direction is to consider the

situation where the variance is also unknown, which has seen surprisingly little attention, but is now understood reasonably well (Honda and Takemura, 2014; Cowan et al., 2015). At the very least, the concentration analysis using Brownian motion could be applied, but I expect an adaptive confidence level will also yield theoretical and practical improvements. Another potential application of the ideas presented here would be to try and port them into other bandit strategies that depend on a confidence level such as BAYES-UCB (Kaufmann, 2016), or even linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Lattimore and Szepesvari, 2017).

References

Yasin Abbasi-Yadkori, Csaba Szepesvári, and David Tax. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.

Rajeev Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Transaction on Automatic Control*, 34:258–267, 1989.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of Conference on Learning Theory (COLT)*, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of Conference on Learning Theory (COLT)*, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory (ALT)*, pages 150–165. Springer, 2007.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

Peter Auer, Nicoló Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated, 2012. ISBN 9781601986269.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory (ALT)*, pages 23–37. Springer, 2009.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. *arXiv preprint arXiv:1302.1611*, 2013.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

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):1516–1541, 2013.

Nicolo Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *ICML*, pages 100–108. Citeseer, 1998.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems*, pages 6284–6293, 2017.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances: Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *arXiv preprint arXiv:1504.05823*, 2015.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2016.

Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. *arXiv preprint arXiv:1309.3376*, 2013.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory (COLT)*, 2016.

Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. On explore-then-commit strategies. In *Neural Information Processing Systems (NIPS)*, 2016a.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016b.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *arXiv preprint arXiv:1605.07416*, 2016.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

Eli Gutin and Vivek Farias. Optimistic gittins indices. In *Advances in Neural Information Processing Systems*, pages 3153–3161, 2016.

Takeyuki Hida. Brownian motion. In *Brownian Motion*, pages 44–113. Springer, 1980.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of Conference on Learning Theory (COLT)*, pages 67–79, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *Artificial Intelligence and Statistics*, pages 375–383, 2014.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *The Journal of Machine Learning Research*, 16(1):3721–3756, 2015.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of Conference on Learning Theory (COLT)*, 2014.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

Emilie Kaufmann. On Bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 592–600, 2012.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1448–1456, 2013.

Sanjeev R Kulkarni and Gábor Lugosi. Finite-time lower bounds for the two-armed bandit problem. *Automatic Control, IEEE Transactions on*, 45(4):711–714, 2000.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

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

Tor Lattimore. Optimally confident UCB: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015a.

Tor Lattimore. The pareto regret frontier for bandits. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS)*, 2015b.

Tor Lattimore. Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. In *Proceedings of Conference on Learning Theory (COLT)*, 2016a.

Tor Lattimore. Regret analysis of the anytime optimally confident UCB algorithm. Technical report, 2016b.

Tor Lattimore and Csaba Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. 2018.

Hans R Lerche. *Boundary crossing of Brownian motion: Its relation to the law of the iterated logarithm and to sequential analysis*. Springer, 1986.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. *arXiv preprint arXiv:1702.07211*, 2017.

Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

Antoine Salomon, Jean-Yves Audibert, and Issam El Alaoui. Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 14(Jan):187–207, 2013.

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

Walter Vogel. An asymptotic minimax theorem for the two armed bandit problem. *The Annals of Mathematical Statistics*, 31(2):444–451, 1960.

Appendix A. Boundary crossing for Gaussian random walks

Let $Z_1, Z_2, \ldots$ be an infinite sequence of independent standard Gaussian random variables and $S_n = \sum_{t=1}^n Z_t$. The proof of Theorem 1 relies on a precise understanding of the behaviour of the random walk $(S_n)_n$. More specifically, what is the hitting probability that S_n ever crosses above a carefully chosen concave boundary. The following lemma is an easy consequence of the elegant analysis of boundary crossing probabilities for Brownian motion by Lerche (1986).

Lemma 12 *Let $d \in \{1, 2, \ldots\}$ and $\Delta > 0$, $\alpha > 0$, $\lambda \in [0, \infty]^d$ and $h_\lambda(t) = \sum_{i=1}^d \min\{t, \sqrt{t\lambda_i}\}$, then there exist constants $c_1 = 4$ and $c_2 = 12$ such that*

(a) $$\mathbb{P}\left(\text{exists } t \geq 0 : S_t \geq \sqrt{2t \operatorname{log}\left(\frac{\alpha}{h_\lambda(t)}\right)} + t\Delta\right) \leq \frac{c_1 h_\lambda(1/\Delta^2)}{\alpha} \ .$$

(b) $$\mathbb{P}\left(\text{exists } t \leq \frac{1}{\Delta^2} : S_t \geq \sqrt{2t \operatorname{log}\left(\frac{\alpha}{t^{1/2}}\right)} - \sqrt{2t}\right) \leq \frac{c_2}{\sqrt{\alpha\Delta}} .$$

The second lemma is a bound on the expected number of samples required before the empirical mean after t samples is close to its true value. The result is relatively standard in the literature, except that here the proofs are simplified by using the properties of Brownian motion.

Lemma 13 *If $\Delta > 0$ and $\zeta = 1 + \max\left\{t : \frac{S_t}{t} \geq \Delta\right\}$, then $\mathbb{E}[\zeta] \leq 1 + \frac{2}{\Delta^2}$.*

Subgaussian case A common relaxation of the Gaussian noise assumption is to assume the noise is 1-subgaussian, which means the reward X_t is chosen so that $\mathbb{E}[\exp(c(X_t - \mu_{A_t})) \mid \mathcal{F}_{t-1}] \leq \exp(c^2/2)$ almost surely for all $c \in \mathbb{R}$. Brownian motion cannot be used to analyse this situation, but Lemma 12 can still be proven for martingale subgaussian noise using the peeling trick on a carefully optimised grid (as used by Garivier (2013) and others). Besides a messier proof, the price is that the $\overline{\log}$ function must be increased slightly (but not so much that Theorem 1 needs to change). Lemma 13 is also easily adapted to the subgaussian setting. The only other lemma that needs modification is Lemma 15 in the appendix, which has the same flavour as Lemma 12 and is adaptable via a peeling trick.

The tangent approximation The connection to Brownian motion is made by noting the discrete time random walk S_t can be embedded in Brownian motion, which means that if B_t is a standard Brownian motion, then for any function $f : \mathbb{R} \to \mathbb{R}$ we have $\mathbb{P}(\text{exists } n \geq 0 : S_n \geq f(n)) \leq \mathbb{P}(\text{exists } n \geq 0 : B_n \geq f(n))$. The main tool of the analysis is called the tangent approximation, which was developed in a beautiful book by Lerche (1986) and is summarised in the following lemma.

Lemma 14 (§3 of Lerche (1986)) *Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function with $f(x) \geq 0$ for all $x \geq 0$ and $\Lambda(t) = f(t) - tf'(t)$ be the intersection of the tangent to f at t with the y-axis, where if f is non-differentiable, then f' denotes any 'super-derivative' (gradient such that the tangent does not intersect the curve). Then*

$$\mathbb{P}(\textit{exists } t \geq 0 : B_t \geq f(t)) \leq \int_0^\infty \frac{\Lambda(t)}{\sqrt{2\pi t^3}} \exp\left(-\frac{f(t)^2}{2t}\right) dt\,.$$

Proof [of Lemma 12] Let B_t be a standard Brownian motion. Each part of the lemma will follow by analysing the probability that the Brownian motion hits the relevant boundary. For the first part let $f(t) = \sqrt{2t\,\overline{\log}(\alpha/h_\lambda(t))}$, which by simple calculus is monotone non-decreasing. Therefore the intersection of the tangent to $f(t) + t\Delta$ at t with the y-axis is $\Lambda(t) = f(t) - tf'(t) \leq f(t)$. By the tangent approximation:

$$\mathbb{P}(\exists t \geq 0 : B_t \geq f(t)) \leq \int_0^\infty \frac{f(t)}{\sqrt{2\pi t^3}} \exp\left(-\frac{(f(t)+t\Delta)^2}{2t}\right) dt \leq \int_0^\infty \frac{3h_\lambda(t)}{2\alpha\sqrt{\pi t}} \exp\left(-\frac{t\Delta^2}{2}\right) dt\,,$$

where the first inequality follows from Lemma 14 and the second from (iv) of Lemma 20 and since for positive $x, y \geq 0$ it holds that $(x+y)^2 \geq x^2 + y^2$.

$$\begin{aligned}
\int_0^\infty \frac{h_\lambda(t)}{t} \exp\left(-\frac{t\Delta^2}{2}\right) dt &= \sum_{i=1}^d \int_0^\infty \min\left\{1, \sqrt{\lambda_i/t}\right\} \exp\left(-\frac{t\Delta^2}{2}\right) dt \\
&= \sum_{i=1}^d \left(\frac{2 - 2\exp\left(-\frac{\lambda_i\Delta^2}{2}\right)}{\Delta^2} + \frac{\sqrt{2\pi\lambda_i}}{\Delta} \operatorname{erfc}\left(\Delta\sqrt{\frac{\lambda_i}{2}}\right)\right) \\
&\leq \sum_{i=1}^d \min\left\{ \frac{2+\sqrt{\pi}}{\Delta^2}, (1+\sqrt{2\pi})\frac{\sqrt{\lambda_i}}{\Delta} \right\} \leq (2+\sqrt{\pi}) h_\lambda(1/\Delta^2)\,,
\end{aligned}$$

where first inequality follows from the fact that $1 - e^{-x} \leq x$ and $\operatorname{erfc}(x) \leq \min\{1, 1/(2x)\}$. The result is completed by naively bounding the constants. For the second part, let $f(t) =$

$\sqrt{2t\log(\alpha/t^{1/2})} - \sqrt{2t}$, which is concave and monotone non-decreasing so that $\Lambda(t) = f(t) - tf'(t) \leq f(t)$. By Lemma 14,

$$\begin{aligned}
\mathbb{P}\left(\exists t \leq \frac{1}{\Delta^2} : B_t \geq \sqrt{2t\log\left(\frac{\alpha}{s^{1/2}}\right)} - \sqrt{2t}\right) &\leq \int_0^{\frac{1}{\Delta^2}} \frac{f(t)}{\sqrt{2\pi t^3}} \exp\left(-\frac{f(t)^2}{2t}\right) dt \\
&\leq \int_0^{\frac{1}{\Delta^2}} \sqrt{\frac{\log(\alpha/t^{1/2})}{\pi t^2}} \exp\left(-1 - \log\left(\frac{\alpha}{t^{1/2}}\right) + 2\sqrt{\log\left(\frac{\alpha}{t^{1/2}}\right)}\right) dt \\
&\leq \int_0^{\frac{1}{\Delta^2}} \frac{5dt}{\sqrt{\pi}\alpha^{1/2}t^{3/4}} \leq \frac{12}{\sqrt{\alpha\Delta}}\,,
\end{aligned}$$

where the second last inequality follows from part (vi) of Lemma 20. ■

Proof [of Lemma 13] The time inversion formula and reflection principle will imply the result (Hida, 1980, for example). Let B_s be a standard Brownian motion. Then for $t > 0$ we have

$$\mathbb{P}\left(\text{exists } s \geq t : B_s/s \geq \Delta\right) = \mathbb{P}\left(\text{exists } s \leq 1/t : B_s \geq \Delta\right) = 2\mathbb{P}\left(B_{1/t} \geq \Delta\right) \leq \exp\left(-\frac{t\Delta^2}{2}\right)\,,$$

where the first equality follows from the time inversion formula and the second from the reflection principle. The inequality is a standard Gaussian tail bound (Boucheron et al., 2013, Chap. 2). Therefore $\mathbb{E}[\zeta] \leq 1 + \int_0^\infty \exp(-t\Delta^2/2)dt = 1 + \frac{2}{\Delta^2}$, where the additive constant is due to the embedding of the discrete random walk in the continuous Brownian motion. ■

Appendix B. Proof of Lemma 9 *(ii)*

Some of the steps in this proof are simplified by using $C > 0$ for a universal positive constant that occasionally has a different value from one equation to the next. When these changes occur they are indicated by the $\overset{\times}{\leq}$ symbol. Let $i \in [k]$ and $j \notin V_i$ be such that $\Delta_j \leq \Delta_i/6$. By part (a) of Lemma 12,

$$\mathbb{P}\left(\min_{1\leq t\leq n} \gamma_j(t) \leq \mu_j - \frac{\Delta_i}{6}\middle|\Lambda_i, \ldots, \Lambda_k\right) \leq \frac{36c_1}{n}\left(\frac{i-1}{\Delta_i^2} + \sum_{m\geq i}\frac{\sqrt{\Lambda_m}}{\Delta_i}\right)\,.$$

It is important to note here that Lemma 12 could only be applied to control the conditional probability above because the random variables $\Lambda_i, \ldots, \Lambda_k$ are independent of $\hat{\mu}_{j,s}$ for all $1 \leq s \leq n$. Let

$$\Psi_i = \min\left\{1,\ \frac{36c_1}{n}\left(\frac{i-1}{\Delta_i^2} + \sum_{m\geq i}\frac{\sqrt{\Lambda_m}}{\Delta_i}\right)\right\}\,.$$

Let $m_i = \sum_{j\notin V_i} \mathbb{1}\left\{\Delta_j \leq \Delta_i/6\right\}$ be the number of arms that might satisfy Eq. (22) in the definition of I. Then by the previous display and the definition of I, $\mathbb{P}\left(I \geq i|\Lambda_i, \ldots, \Lambda_k\right) \leq \Psi_i^{m_i}$. It would be tempting to try and bound the expectation of Δ_I by taking a union bound over all arms, but this is not tight when many arms have nearly the same mean. Let $\mathcal{I} \subset [k]$ be empty if $i_1 = \ell + 1 > k$.

Otherwise $\mathcal{I} = \{i_1, i_2, \ldots, i_b\}$ where and $i_{j+1} = \min\left\{i : \Delta_i > 6\Delta_{i_j}\right\}$ and b is as large as possible so that $i_b \leq k$.

$$\mathbb{E}[\Delta_I] \leq 6\sum_{i\in\mathcal{I}} \mathbb{P}\left(I \geq i\right)\Delta_i = 6\sum_{i\in\mathcal{I}} \mathbb{E}\left[\mathbb{P}\left(I \geq i \middle| \Lambda_i, \ldots, \Lambda_k\right)\Delta_i\right]$$
$$\leq 6\sum_{i\in\mathcal{I}} \mathbb{E}\left[\Psi_i^{m_i}\Delta_i\right] = 6\mathbb{E}\left[\sum_{i\in\mathcal{I}} \Psi_i^{m_i}\Delta_i\right] \overset{\times}{\leq} C\mathbb{E}\left[\max_{i\in\mathcal{I}} \Psi_i^{m_i}\Delta_i\right].$$

Only the last step above is non-trivial. It follows by choosing a as the smallest value such that $\Psi_{i_a} < 1/2$ (or $a = b$ if such a choice does not exist). Then the contribution of $\Psi_i^{m_i}\Delta_i$ is decreasing exponentially in both directions away from i_a by the definition of $\mathcal{I}$ and the fact that $m_{i_{a+2}} \geq m_{i_a} + 1$.

Case 1 ($\Psi_i \geq 1/2$) By the definition of ℓ and the fact that $i > \ell$ we have $\delta_i \leq 1/4$ and so by Eq. (15), $k_i \leq n\Delta_i^2/(16c_2^2) \leq n\Delta_i^2(144c_1)$. Therefore

$$\frac{1}{2} \leq \frac{36c_1}{n}\left(\frac{i-1}{\Delta_i^2} + \sum_{m\geq i}\frac{\sqrt{\Lambda_m}}{\Delta_i}\right) \leq \frac{36c_1}{n}\left(\frac{k_i}{\Delta_i^2} + \sum_{m\geq i}\frac{\sqrt{\Lambda_m}}{\Delta_i}\right) \leq \frac{1}{4} + \frac{36c_1}{n}\sum_{m\geq i}\frac{\sqrt{\Lambda_m}}{\Delta_i}.$$

Rearranging and using the fact that $\Psi_i \leq 1$ and $\sqrt{\Lambda_m} \leq \Delta_m\Lambda_m$ shows that

$$\Psi_i^{m_i}\Delta_i \leq \Delta_i \leq \frac{C}{n}\sum_{m\geq i}\sqrt{\Lambda_m} \leq \frac{C}{n}\sum_{m>\ell}\Delta_m\Lambda_m.$$

Case 2 ($\Psi_i < 1/2$) By the definition of Ψ_i and the assumption that $\Psi_i < 1/2$ and $i > \ell$,

$$\Psi_i^{m_i}\Delta_i \leq \frac{36c_1 2^{1-m_i}}{n}\left(\frac{i-1}{\Delta_i} + \sum_{m\geq i}\sqrt{\Lambda_m}\right) \overset{\times}{\leq} \frac{C\ell 2^{1-m_{\ell+1}}}{n\Delta_{\ell+1}} + \frac{C}{n}\sum_{m>\ell}\Delta_m\Lambda_m.$$

The second term is already in the right form. For the first,

$$\frac{\ell 2^{1-m_{\ell+1}}}{n\Delta_{\ell+1}} \overset{\times}{\leq} \frac{C}{n\Delta_{\ell+1}} + \frac{\ell}{n\Delta_{\ell+1}}\mathbb{1}\left\{m_{\ell+1} < \ell/4\right\}$$
$$\overset{\times}{\leq} \frac{C}{n}\sum_{m>\ell}\Delta_m\Lambda_m + C\Delta_{\ell+1}\mathbb{1}\left\{m_{\ell+1} < \ell/4\right\} \overset{\times}{\leq} C\left(\bar{\Delta}_\ell + \frac{1}{n}\sum_{m>\ell}\Delta_m\Lambda_m\right),$$

where the last inequality follows since if $m_{\ell+1} < \ell/4$, then many arms $i \leq \ell$ must have means nearly as small as $\ell + 1$. The result is completed by part *(i)* of the lemma.

Appendix C. Proof of Lemma 9 *(iii)*

The proof relies on another concentration result.

Lemma 15 *There exists an $\varepsilon > 0$ such that for any arm j*

$$\mathbb{P}\left(\textit{exists } s \leq \frac{8n}{\ell} : \hat{\mu}_{j,s} + \sqrt{\frac{2}{s}\overline{\log}\left(\sqrt{\frac{n}{2\ell s}}\right)} \leq \mu_j + 2\sqrt{\frac{\varepsilon}{s}}\right) \leq \frac{1}{2}.$$

Proof The result follows by rescaling the time horizon and noting that if B_s is a Brownian motion, then for sufficiently small ε (for example, $1/200$).

$$\mathbb{P}\left(\text{exists } s \leq 8 : B_s \geq \sqrt{2s \, \overline{\log}\left(\frac{1}{\sqrt{2s}}\right)} - 2\sqrt{\varepsilon s}\right) \leq \frac{1}{2}\,.$$

The above bound does not depend on any variables and can be verified numerically (either by simulating Brownian motion or numerically solving the heat equation that characterises the density of the paths of Brownian motion). An analytical proof is also possible, but requires a modest increase in the definition $\overline{\log}(\cdot)$ if the tangent approximation is to yield a sufficiently tight bound. ■

We need a little more notation. First up is another set of 'usually optimistic' arms, $U \subset [k]$ defined by

$$U = \left\{ j : \Delta_j \leq 4\bar{\Delta}_\ell \text{ and } \hat{\mu}_{j,s} + \sqrt{\frac{2}{s} \, \overline{\log}\sqrt{\frac{n}{2\ell s}}} \geq \mu_j + 2\sqrt{\frac{\varepsilon}{s}} \text{ for all } s \leq \frac{8n}{\ell} \right\}.$$

Let $\underline{\Delta} = \sqrt{\varepsilon\ell/(8n)}$ with $\varepsilon > 0$ as given in Lemma 15. Finally we need two more events E_2 and E_3 given by

$$E_2 = \left\{ \sqrt{n\ell} + \sum_{m>\ell} \sqrt{\Lambda_m} \leq \sqrt{2n\ell} \right\} \qquad \text{and} \qquad E_3 = \left\{ |U| \geq \frac{\ell}{8} \right\}. \tag{26}$$

Lemma 16 *If E_1, E_2, E_3 and $\Delta_i > 4\bar{\Delta}_\ell$, then $T_i(n) \leq \left\lceil \zeta_i(\underline{\Delta}) + \frac{24n}{\varepsilon\ell} \right\rceil$.*

Proof Proceeding by contradiction. Suppose the claim is not true, then there exists a round $t-1 < n$ such that $A_t = i$ and

$$T_i(t-1) = \left\lceil \zeta_i(\underline{\Delta}) + \frac{24n}{\varepsilon\ell} \right\rceil. \tag{27}$$

By the definition of $j \in U$,

$$H_j(t-1) \leq \sum_{m=1}^{k} \sqrt{T_j(t-1)T_m(t-1)} \leq \sqrt{T_j(t-1)} \left(\sqrt{\ell n} + \sum_{m>\ell} \sqrt{\Lambda_m} \right) \leq \sqrt{2T_j(t-1)\ell n}\,, \tag{28}$$

where the first inequality follows from the definition of $H_j(t-1)$, the second by splitting the sum and because E_1 holds (so that $T_m(n) \leq \Lambda_m$ for $m > \ell$) and Cauchy-Schwarz, the third follows because E_2 holds. Therefore arms $j \in U$ with $T_j(t-1) \leq 8n/\ell$ satisfy

$$\gamma_j(t) = \hat{\mu}_j(t-1) + \sqrt{\frac{2 \, \overline{\log}\left(\frac{n}{H_j(t-1)}\right)}{T_j(t-1)}} \geq \mu_j + 2\sqrt{\frac{\varepsilon}{T_j(t-1)}}\,. \tag{29}$$

Furthermore, since $T_i(t-1) \geq \zeta_i(\underline{\Delta})$ and $A_t = i$ it holds that $\gamma_i(t) \geq \gamma_j(t)$ and so using Eq. (29) and the same argument as in Lemma 5 leads to

$$\begin{aligned}
&\mu_i + \underline{\Delta} + \sqrt{\frac{2\,\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}{T_i(t-1)}} \geq \hat{\mu}_i(t-1) + \sqrt{\frac{2\,\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}{T_i(t-1)}} \\
&\geq \hat{\mu}_j(t-1) + \sqrt{\frac{2\,\overline{\log}\left(\frac{n}{H_j(t-1)}\right)}{T_j(t-1)}} \geq \hat{\mu}_j(t-1) + \sqrt{\frac{2\,\overline{\log}\left(\sqrt{\frac{n}{2\ell T_j(t-1)}}\right)}{T_j(t-1)}} \\
&\geq \mu_j + 2\sqrt{\frac{\varepsilon}{T_j(t-1)}} \geq \mu_i + \underline{\Delta} + \sqrt{\frac{\varepsilon}{T_j(t-1)}}\,.
\end{aligned}$$

And by rearranging $\dfrac{2}{T_i(t-1)}\,\overline{\log}\left(\dfrac{n}{H_i(t-1)}\right) \geq \dfrac{\varepsilon}{T_j(t-1)}$. Therefore for all $j \in U$ we have

$$T_j(t-1) \geq \min\left\{\frac{8n}{\ell}, \frac{\varepsilon T_i(t-1)}{\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}\right\}. \tag{30}$$

Since $T_i(t-1) \geq 8n/(\varepsilon\ell)$, the definition of $H_i(t-1)$ implies that

$$\begin{aligned}
H_i(t-1) &= \sum_{j=1}^{k} \min\left\{T_i(t-1),\, \sqrt{T_i(t-1)T_j(t-1)}\right\} \\
&\geq \sum_{j\in U} \min\left\{T_i(t-1),\, T_j(t-1)\right\} \geq \frac{8|U|n}{\ell\,\overline{\log}\left(\frac{n}{H_i(t-1)}\right)} \geq \frac{n}{\overline{\log}\left(\frac{n}{H_i(t-1)}\right)}\,.
\end{aligned}$$

Then since E_3 holds, by Lemma 20(vii) we have $\overline{\log}(n/H_i(t-1)) \leq 3$. Therefore if $T_i(t-1) \geq 24n/(\varepsilon\ell)$, then another application of Eq. (30) shows that $T_j(t-1) \geq 8n/\ell$ and so $n > t-1 = \sum_{j=1}^{k} T_j(t-1) \geq \sum_{j\in U} T_j(t-1) \geq 8n|U|/\ell \geq n$, which is a contradiction. Therefore there does not exist a round $t-1$ where Eq. (27) holds and $A_t = i$ and the lemma follows. ■

Lemma 17 $\mathbb{P}\left(E_3^c\right) \leq 3/\ell$.

Proof By Markov's inequality, $m = \sum_{j\leq\ell} \mathbb{1}\left\{\Delta_j \leq 4\bar{\Delta}_\ell\right\} \geq \frac{3\ell}{4}$. Let $\chi_1, \ldots, \chi_m$ be a sequence of independent Bernoulli events given by $\chi_j = \mathbb{1}\left\{j \notin U\right\}$. Then by Lemma 15, $\mathbb{P}\left(\chi_j = 1\right) \leq 1/2$ and so Chebyshev's inequality implies that

$$\begin{aligned}
\mathbb{P}\left(E_3^c\right) &= \mathbb{P}\left(|U| < \frac{\ell}{8}\right) = \mathbb{P}\left(\sum_{j=1}^{m}(1-\chi_j) < \frac{\ell}{8}\right) = \mathbb{P}\left(\sum_{j=1}^{m}\chi_j > m - \frac{\ell}{8}\right) \\
&\leq \mathbb{P}\left(\sum_{j=1}^{m}(\chi_j - \mathbb{E}[\chi_j]) \geq \frac{m}{2} - \frac{\ell}{8}\right) \leq \frac{m/4}{\left(\frac{m}{2} - \frac{\ell}{8}\right)^2} \leq \frac{3}{\ell}\,.
\end{aligned}$$

■

At last all the tools are available to prove part (iii) of Lemma 9.

Proof [of Lemma 9 *(iii)*] The regret due to arms $i \leq \ell$ is decomposed

$$\mathbb{E}\left[\mathbb{1}_{\{E_1\}} \sum_{i \leq \ell} \Delta_i T_i(n)\right] \leq \mathbb{E}\left[\mathbb{1}_{\{E_2^c\}} n \Delta_\ell\right] + \mathbb{E}\left[\mathbb{1}_{\{E_1, E_2\}} \sum_{i \leq \ell} \Delta_i T_i(n)\right] . \tag{31}$$

The first term is bounded easily using the definition of E_2, Lemma 21(i) and Lemma 9(i).

$$\mathbb{E}\left[\mathbb{1}_{\{E_2^c\}} n \Delta_\ell\right] \leq C \sum_{m > \ell} \mathbb{E}\left[\sqrt{\Lambda_m}\right] \leq C \sum_{m > \ell} \Delta_m \lambda_m \, .$$

The second term in Eq. (31) is bounded using Lemmas 16 and 17. By noting that the contribution to the regret of arms i with $\Delta_i \leq 4\bar{\Delta}_\ell$ is at most $4n\bar{\Delta}_\ell$ it follows that

$$\mathbb{E}\left[\mathbb{1}_{\{E_1, E_2\}} \sum_{i \leq \ell} \Delta_i T_i(n)\right] \leq 4n\bar{\Delta}_\ell + n\Delta_\ell \mathbb{P}\left(E_3^c\right) + \mathbb{E}\left[\mathbb{1}_{\{E_1, E_2, E_3\}} \sum_{i \leq \ell : \Delta_i > 4\bar{\Delta}_\ell} \Delta_i \left(\zeta_i(\underline{\Delta}) + \left\lceil \frac{24n}{\varepsilon \ell} \right\rceil\right)\right] .$$

The proof is completed since Lemma 17 implies that $n\Delta_\ell \mathbb{P}\left(E_3^c\right) \leq 3n\Delta_\ell/\ell \leq 3n\bar{\Delta}_\ell$ and Lemma 13 implies that

$$\mathbb{E}\left[\mathbb{1}_{\{E_1, E_2, E_3\}} \sum_{i \leq \ell : \Delta_i > 4\bar{\Delta}_\ell} \Delta_i \left(\zeta_i(\underline{\Delta}) + \left\lceil \frac{24n}{\varepsilon \ell} \right\rceil\right)\right] \leq \mathbb{E}\left[\sum_{i \leq \ell} \Delta_i \left(\zeta_i(\underline{\Delta}) + \left\lceil \frac{24n}{\varepsilon \ell} \right\rceil\right)\right] \leq \sum_{i \leq \ell} \Delta_i \left(2 + \frac{40n}{\varepsilon \ell}\right) \leq Cn\bar{\Delta}_\ell \, . \qquad \blacksquare$$

Appendix D. Technical results

Here some lemmas that are either known or follow from uninteresting calculations. The first two are used for the lower bounds have have been seen before.

Lemma 18 (See Lemma 2.6 in Tsybakov 2008) *Let $\mathbb{P}$ and $\mathbb{P}'$ be measures on the same probability space and assume $\mathbb{P}'$ is absolutely continuous with respect to $\mathbb{P}$. Then for any event A, $\mathbb{P}(A) + \mathbb{P}'(A^c) \geq \exp(-\operatorname{KL}(\mathbb{P}, \mathbb{P}'))/2$, where $\operatorname{KL}(\mathbb{P}, \mathbb{P}')$ is the relative entropy between $\mathbb{P}$ and $\mathbb{P}'$.*

The next lemma has also been seen before. For example in the articles by Auer et al. (1995) or Gerchinovitz and Lattimore (2016) where the formalities are described in great detail.

Lemma 19 *Fix a strategy. Let $1 \leq i \leq k$ and $\mu \in \mathbb{R}^k$ and $\mu' \in \mathbb{R}^k$ be such that $\mu_j = \mu'_j$ for all $j \neq i$ and $\mu_i - \mu'_i = \Delta$. Then let $\mathbb{P}$ be the measure on $A_1, X_1, A_2, X_2, \ldots, A_n, X_n$ induced by the interaction of the strategy with rewards sampled using mean vector μ and $\mathbb{P}'$ be the same but with rewards sampled with means from μ'. Then $\operatorname{KL}(\mathbb{P}, \mathbb{P}') = \mathbb{E}[T_i(n)]\Delta^2/2$, where the expectation is taken with respect to $\mathbb{P}$.*

Recall that $\overline{\log}(x) = \log((e+x)\log^{\frac{1}{2}}(e+x))$. Here are a few simple facts that make manipulating this unusual function a little easier.

Lemma 20 *The following hold:*

(i) $\overline{\log}$ *is concave and monotone increasing on* $[0, \infty)$.

(ii) $\overline{\log}(0) \geq 1$.

(iii) $\lim_{x\to\infty} \overline{\log}(x)/\log(x) = 1$.

(iv) $(\overline{\log}(x))^{1/2}\exp(-\overline{\log}(x)) \leq 3/(2x)$.

(v) If $x\overline{\log}^{\frac{1}{2}}(b/x) \geq a$, *then* $\overline{\log}(b/x) \leq 2\overline{\log}(b/a)$ *for all* $a, b > 0$.

(vi) $\sqrt{\overline{\log}(x)}\exp(-1-\overline{\log}(x)+2\sqrt{\overline{\log}(x)}) \leq 5\sqrt{1/x}$.

(vii) If $\overline{\log}(x) \geq x$, *then* $\overline{\log}(x) \leq 2$.

Proof (i) is proven by checking derivatives:

$$\frac{d}{dx}\overline{\log}(x) = \frac{\frac{1}{2\sqrt{\log(e+x)}} + \sqrt{\log(e+x)}}{(e+x)\sqrt{\log(e+x)}} \qquad \frac{d^2}{dx^2}\overline{\log}(x) = -\frac{1+\log(e+x)+2\log^2(e+x)}{2(e+x)^2\log^2(e+x)}\,.$$

The former is clearly positive and the latter negative, which shows that $\overline{\log}$ is monotone increasing and concave. Parts (ii) and (iii) are trivial. For (iv),

$$\begin{aligned}
&\overline{\log}^{\frac{1}{2}}(x)\exp(-\overline{\log}(x)) = \overline{\log}^{\frac{1}{2}}(x)\exp(-\log((e+x)\log^{\frac{1}{2}}(e+x))) \\
&= \frac{\overline{\log}^{\frac{1}{2}}(x)}{(e+x)\log^{\frac{1}{2}}(e+x)} = \frac{1}{e+x}\sqrt{\frac{\log(e+x)+\log\log^{\frac{1}{2}}(e+x)}{\log(e+x)}} \leq \frac{1}{e+x}\sqrt{1+\frac{1}{2e}} \leq \frac{3}{2x}\,.
\end{aligned}$$

For (v),

$$\overline{\log}\left(\frac{b}{x}\right) \leq \overline{\log}\left(\frac{b\,\overline{\log}^{\frac{1}{2}}(b/x)}{a}\right) \leq \overline{\log}\left(\frac{b}{a}\,\overline{\log}^{\frac{1}{2}}\left(\frac{b}{a}\,\overline{\log}^{\frac{1}{2}}\left(\frac{b}{a}\,\overline{\log}^{\frac{1}{2}}\left(\frac{b}{a}\cdots\right.\right.\right.\right. = z\,,$$

where the final equality serves as the definition of z. If $z \leq 2$, then $\overline{\log}(b/x) \leq z \leq 2 \leq 2\overline{\log}(b/a)$. Suppose now that $z \geq 2$. Let $u, v \geq 0$, then $\overline{\log}(uv) \leq \overline{\log}(u) + \overline{\log}(v)$ and if $u \geq 2$, then $u^2 - \overline{\log}(u) \geq u^2/2$. Therefore $z^2/2 \leq z^2 - \overline{\log}(z) \leq \overline{\log}(zb/a) - \overline{\log}(z) \leq \overline{\log}(b/a)$. Therefore $\overline{\log}(b/x) \leq z^2 \leq 2\overline{\log}(b/a)$ as required. For (vi), using a similar reasoning as (iv),

$$\begin{aligned}
\sqrt{x\,\overline{\log}(x)}\exp\left(-1-\overline{\log}(x)+2\,\overline{\log}^{\frac{1}{2}}(x)\right) &= \frac{\overline{\log}^{\frac{1}{2}}(x)\sqrt{x}\exp\left(2\,\overline{\log}^{\frac{1}{2}}(x)\right)}{e(e+x)\log^{\frac{1}{2}}(e+x)} \\
&\leq \sqrt{\frac{\overline{\log}(x)}{\log(e+x)}}(x+e)^{-\frac{1}{2}}\exp\left(2\,\overline{\log}^{\frac{1}{2}}(x)\right)\,.
\end{aligned}$$

Simple calculus shows that $\overline{\log}(x)/\log(e+x) \leq (1/2+e)/e$. Let $g(x) = (x+e)^{-\frac{1}{2}} \exp(2\overline{\log}^{\frac{1}{2}}(x))$. Then $\max_{x \geq 0} g(x) \leq 10.34 \leq 11$ by numerical calculation, which is valid by the following argument: First, the function g is twice differentiable, satisfies $g(0) > 0$, $g'(0) > 0$ and $\lim_{x\to\infty} g(x) = 0$. By taking the first derivative it is easy to see that g has a unique maximum in $x^* \in (0, \infty)$ with $g(x^*) > 0$. Therefore g is monotone increasing for $x < x^*$ and monotone decreasing afterwards for $x > x^*$. This means the maximiser may be found by a binary search with arbitrary precision. Therefore $\sqrt{x\,\overline{\log}(x)} \exp(-1 - \overline{\log}(x) + 2\,\overline{\log}^{\frac{1}{2}}(x)) \leq \frac{11}{e}\sqrt{(1/2+e)/e} \leq 5$. For (vii), if $x \geq 0$, then $\frac{d}{dx}\,\overline{\log}(x) \leq 3/(2e) \leq 1$. Therefore $\overline{\log}(x) - x$ is monotone non-increasing and the result follows by checking that $\overline{\log}(2) \leq 2$. ■

The second technical lemma provides some useful results relating to the optimisation problem appearing in Eq. (3) and the definition of ℓ in Eq. (15).

Lemma 21 *There exists a universal constant $C > 0$ such that:*

(i) $\Delta_\ell \leq C\sqrt{\dfrac{\ell}{n}}$ *or* $n\Delta_\ell \leq C \displaystyle\sum_{m>\ell} \frac{1}{\Delta_m}$.

(ii) $n\bar{\Delta}_\ell + \displaystyle\sum_{m>\ell} \Delta_m \lambda_m \leq C \min_{i\in[k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_m \lambda_m \right)$.

(iii) $\displaystyle\min_{i\in[k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_i \lambda_i \right) \leq C\sqrt{kn} + \sum_{m=1}^{k} \Delta_m$.

(iv) $\displaystyle\min_{i\in[k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_i \lambda_i \right) \leq C \sum_{m:\Delta_m>0} \left(\Delta_m + \frac{\log(n)}{\Delta_m} \right)$.

Proof For part *(i)*, by the definition of ℓ we have

$$\Delta_\ell < 4c_2\sqrt{\frac{k_\ell}{n}} = 4c_2\sqrt{\frac{\ell}{n} + \frac{\Delta_\ell}{n}\sum_{i>\ell}\frac{1}{\Delta_i}} \leq 4c_2\sqrt{\max\left\{\frac{\ell}{n}, \frac{\Delta_\ell}{n}\sum_{i>\ell}\frac{1}{\Delta_i}\right\}}.$$

The result follows by simplifying each of the two cases in the maximum. For part *(ii)*, let $i = \arg\min_j n\bar{\Delta}_j + \sum_{m>j} \Delta_m\lambda_m$.

Case 1 ($\ell > i$) Using the fact that for $m \leq \ell$ we have $\Delta_m \leq 16c_2^2 k_m/\Delta_m$ leads to

$$\begin{aligned} n\bar{\Delta}_\ell + \sum_{m>\ell} \Delta_m\lambda_m &\leq n\bar{\Delta}_i + \frac{n}{\ell}\sum_{m=i+1}^{\ell} \Delta_m + \sum_{m>\ell} \Delta_m\lambda_m \\ &\leq n\bar{\Delta}_i + \frac{16c_2^2 n}{\ell}\sum_{m=i+1}^{\ell}\frac{k_m}{\Delta_m} + \sum_{m>\ell}\Delta_m\lambda_m \leq (1+32c_2^2)\left(n\bar{\Delta}_i + \sum_{m>i}\Delta_m\lambda_m\right). \end{aligned}$$

Case 2 ($\ell < i$) Using the fact that $\overline{\log}(x)/x \leq 3/2$ for $x \geq 1$ leads to

$$\sum_{m=\ell+1}^{i} \Delta_m \lambda_m = \sum_{m=\ell+1}^{i} \frac{n\Delta_m}{k_m} \cdot \frac{k_m}{n\Delta_m^2} \overline{\log}\left(\frac{n\Delta_m^2}{k_m}\right) + \sum_{m=\ell+1}^{i} \Delta_m$$
$$\leq \frac{3n}{2} \sum_{m=1}^{i} \frac{\Delta_m}{k_m} + \frac{n}{i} \sum_{m=\ell+1}^{i} \Delta_m \leq 7n\bar{\Delta}_i \,,$$

where the first inequality is true since $i/n \leq 1$ and by the definition of ℓ, if $m \geq \ell$, then $n\Delta_m^2/k_m \geq 16c_2^2 \geq 1$. The second inequality follows by letting $k(x) = \sum_{m=1}^{k} \min\{1, x/\Delta_m\}$ and noting that $x/k(x)$ is monotone increasing and by Markov's inequality $k(2\bar{\Delta}_i) \geq i/2$. Then for $\Delta_m \leq 2\bar{\Delta}_i$ it holds that $\Delta_m/k_m \leq 2\bar{\Delta}_i/k(2\bar{\Delta}_i) \leq 4\bar{\Delta}_i/i$ while for $\Delta_m > 2\bar{\Delta}_i$ we have $\Delta_m/k_m \leq 2\Delta_m/i$. Therefore

$$n\bar{\Delta}_\ell + \sum_{m>\ell} \Delta_m \lambda_m \leq n\bar{\Delta}_i + \sum_{m=\ell+1}^{i} \Delta_m \lambda_m + \sum_{m>i} \Delta_m \lambda_m \leq 8\left(n\bar{\Delta}_i + \sum_{m>i} \Delta_m \lambda_m\right)$$

The last two parts are straightforward. For *(iii)*, let $j = \max\{m : \Delta_m \leq 3\sqrt{k/n}\}$, which means that for $m > j$ it holds that $\overline{\log}(n\Delta_m^2/m) \leq 2\log(n\Delta_m^2/m)$. Then

$$\min_{i \in [k]} \left(n\bar{\Delta}_i + \sum_{m>i} \Delta_m \lambda_m\right) - \sum_{m=1}^{k} \Delta_m \leq n\bar{\Delta}_j + \sum_{m>j} \Delta_m(\lambda_m - 1)$$
$$\leq 3\sqrt{kn} + \sum_{m>j} \frac{1}{\Delta_m} \overline{\log}\left(\frac{n\Delta_m^2}{k_m}\right) \leq 3\sqrt{kn} + \sum_{m>j} \frac{2}{\Delta_m} \log\left(\frac{n\Delta_m^2}{m}\right)$$
$$\leq 3\sqrt{kn} + \frac{2}{3}\sqrt{\frac{n}{k}} \int_1^k \log\left(\frac{9k}{x}\right) dx = 3\sqrt{kn} + \frac{2}{3}\sqrt{nk}\,(1 + \log(9)) \leq 6\sqrt{kn}\,.$$

Rearranging completes the proof of *(iii)*. For part *(iv)*, first note that

$$\lambda_m = 1 + \frac{1}{\Delta_m^2} \overline{\log}\left(\frac{n\Delta_m^2}{k_m}\right) \leq 1 + \frac{\overline{\log}(n)}{\Delta_m^2} + \frac{\overline{\log}(\Delta_m^2)}{\Delta_m^2} \leq \frac{5}{2} + \frac{3/2 + \log(e+n)}{\Delta_m^2}\,.$$

The result follows by choosing $i = \max\{i : \Delta_i = 0\}$. ■

Appendix E. History

Table 2 outlines the long history of finite-armed stochastic bandits. It indicates which algorithms are asymptotically optimal and/or sub-UCB and the ratio (up to constant factors) by which they are minimax suboptimal. Empty cells represent results unknown at the time. Most papers do not provide minimax bounds, but they can be derived easily from finite-time bounds using the argument given by Bubeck and Cesa-Bianchi (2012), which I have done where possible. In some cases the finite-time bound cannot be used to derive the minimax bound and these results are marked as conjectures. Algorithms were omitted from the list if (a) I could not straightforwardly adapt their analysis to the

Gaussian noise model and/or frequentist regret (Honda and Takemura, 2011; Russo and Van Roy, 2014; Gutin and Farias, 2016), or (b) the algorithm depends on μ-dependent tuning such as SOFTMIX (Cesa-Bianchi and Fischer, 1998), ε-GREEDY (Auer et al., 2002), EXPLORE-THEN-COMMIT (Garivier et al., 2016a) and BOLTZMANN EXPLORATION (Cesa-Bianchi et al., 2017). Also omitted are algorithms designed for adversarial bandits, few of which are suitable for unbounded rewards and none are competitive with UCB for stochastic problems, but a nice survey of these algorithms is by Bubeck and Cesa-Bianchi (2012). The vast majority of the Bayesian literature is also omitted since it deals with discounted rewards. See the recent book by Gittins et al. (2011) for an overview of Bayesian algorithms.

Remark 22 *It must be emphasised that many of the algorithms in the table were designed for settings more general than Gaussian and the core contribution was actually this generality.*

Date	Algorithm	Sub-UCB	Asy. opt.	Minimax ratio	Anytime
1960	'explore-then-commit' Vogel (1960)			1*	no
1985	'forcing' Lai and Robbins (1985)		yes		yes
1987	KL-UCB* Lai (1987)		yes		no
1995	UCB Katehakis and Robbins (1995), Agrawal (1995)		yes		yes
2002	UCB† Auer et al. (2002)	yes	no	$\sqrt{\log(n)}$	yes
2002	UCB2† Auer et al. (2002)	yes	yes	$\sqrt{\log(n)}$	yes
2007	UCB-V Audibert et al. (2007)	yes	no	$\sqrt{\log(n)}$	yes
2009	MOSS† Audibert and Bubeck (2009)	no	no	1	no
2010	IMPROVED UCB† Auer and Ortner (2010)	yes	no	$\sqrt{\log(k)}$	yes
2010	DMED† Honda and Takemura (2010)		yes	$\sqrt{\log(n)}$ ♭	yes
2010	DMED+† Honda and Takemura (2010)		yes	$\sqrt{\log(k)}$ ♭	yes
2011	KL-UCB Cappé et al. (2013)	yes	yes	$\sqrt{\log(n)}$	yes
2011	KL-UCB+‡ Cappé et al. (2013)	yes	yes	$\sqrt{\log(k)}$	yes
2012	BAYES-UCB† Kaufmann et al. (2012)	yes	yes	$\sqrt{\log(n)}$	yes
2012	THOMPSON SAMPLING♮ Agrawal and Goyal (2012)	yes	yes	$\sqrt{\log(k)}$	yes
2015	IMED† Honda and Takemura (2015)	yes	yes	$\sqrt{\log(k)}$ ♭	yes
2016	BAYES-UCB+ Kaufmann (2016)	yes	yes	$\sqrt{\log(k)}$ ♭	yes
2016	FH-GITTINS Lattimore (2016a)	yes		$\sqrt{\log(n)}$	no
2016	MOSS-ANYTIME Degenne and Perchet (2016)	no	no	1	yes
2017	KL-UCB++ Ménard and Garivier (2017)	no	yes	1	no
2018	KL-UCB*	yes	yes	$\sqrt{\log(k)}$	no
2018	ADA-UCB	yes	yes	1	no

*Results given for two-armed Bernoulli bandits only.
†Results given for bounded and/or Bernoulli rewards, but algorithm/proof is easily adapted.
‡No known reference. Can be shown using tools of this paper combined with those by Kaufmann (2016); Lattimore (2016b).
♮Results are given for bounded rewards, but the same technique works for Gaussian rewards. See also the article by Korda et al. (2013).
♭A conjectured result.

Table 2: History of bandit algorithms

$\overline{\log}(\cdot)$	$\overline{\log}(x) = \log((e+x)\log^{\frac{1}{2}}(e+x))$
k	number of arms
n	time horizon
A_t	action chosen in round t
μ	k-dimensional vector of mean payoffs
$\Delta_i(\mu)$	suboptimality gap, $\Delta_i = \max_j \mu_j - \mu_i$
k_i	$\sum_{j=1}^k \min\left\{1, \frac{\Delta_i}{\Delta_j}\right\}$
λ_i	$1 + \frac{1}{\Delta_i}\overline{\log}\left(\frac{n\Delta_i^2}{k_i}\right)$
η_t	noise in round t
X_t	reward in the tth round, $X_t = \mu_{A_t} + \eta_t$
$T_i(t)$	number of plays of arm i after t rounds
$K_i(t)$	$\sum_{m=1}^k \min\left\{1, \sqrt{\frac{T_m(t-1)}{T_i(t-1)}}\right\}$
$H_i(t)$	$T_i(t)K_i(t)$
Λ_i	see display before Lemma 7
$\hat{\mu}_i(t)$	empirical mean of arm i after t rounds
$\hat{\mu}_{i,s}$	empirical mean of arm i after s plays
$\bar{\Delta}_i$	$\sum_{m=1}^i \Delta_m / i$
V_i, W_i	sets of arms defined in Eq. (14)
δ_i	see Eq. (15)
ℓ	see Eq. (15)
F_i	Event that enough arms are optimistic, see Eq. (17)
$\zeta_i(\Delta)$	$1 + \max_s \{s : \hat{\mu}_{i,s} > \mu_i + \Delta\}$
c_1, c_2	constants $c_1 = 4$ and $c_2 = 12$

Table 3: Table of notation