Regret Analysis of the Anytime Optimally Confident UCB Algorithm

Tor Lattimore
Department of Computing Science
University of Alberta, Canada
tor.lattimore@gmail.com

Abstract

I introduce and analyse an anytime version of the Optimally Confident UCB (OCUCB) algorithm designed for minimising the cumulative regret in finite-armed stochastic bandits with subgaussian noise. The new algorithm is simple, intuitive (in hindsight) and comes with the strongest finite-time regret guarantees for a horizon-free algorithm so far. I also show a finite-time lower bound that nearly matches the upper bound and has a trivial proof.

1 Introduction

A potential drawback of the Optimally Confident UCB algorithm for finite-armed bandits is that it depends on advance knowledge of the horizon [Lattimore, 2015]. I address this issue by analysing an anytime version that performs about as well empirically and for which the regret guarantee suffers only an additional $\log\log(n)$ additive term. The proof is also significantly easier – though still quite involved – and more importantly is quite intuitive. For the sake of brevity I will give neither a detailed introduction nor an exhaustive survey of the literature. Readers looking for a gentle primer on multi-armed bandits might enjoy the monograph by Bubeck and Cesa-Bianchi [2012] from which I borrow notation. Let K be the number of arms and $I_t \in \{1, \ldots, K\}$ be the arm chosen in round t. The reward is $X_t = \mu_{I_t} + \xi_t$ where $\mu \in \mathbb{R}^K$ is the unknown vector of means and the noise term ξ_t is assumed to be 1-subgaussian (therefore zero-mean). The n-step pseudo-regret of strategy π given mean vector μ with maximum mean $\mu^* = \max_i \mu_i$ is

$$R^\pi_\mu(n) = n\mu^* - \mathbb{E}\sum_{t=1}^n \mu_{I_t}\,,$$

where the expectation is taken with respect to the uncertainty in both the rewards and actions. In all analysis I make the standard notational assumption that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$. The new algorithm is called OCUCB-n and depends on two parameters $\eta > 1$ and $\rho \in [1/2, 1]$. The algorithm chooses $I_t = t$ in rounds $t \leq K$ and subsequently $I_t = \arg\max_i \gamma_i(t)$ with

$$\gamma_i(t) = \hat\mu_i(t-1) + \sqrt{\frac{2\eta\log(B_i(t-1))}{T_i(t-1)}}\,, \tag{1}$$

where $T_i(t-1)$ is the number of times arm i has been chosen after round $t-1$ and $\hat\mu_i(t-1)$ is the empirical estimate of its mean and

$$B_i(t-1) = \max\left\{e,\, \log(t),\, t\log(t)\left(\sum_{j=1}^K \min\left\{T_i(t-1),\, T_j(t-1)^\rho T_i(t-1)^{1-\rho}\right\}\right)^{-1}\right\}.$$

Besides the algorithm, the contribution of this article is a proof that OCUCB-n satisfies a nearly optimal regret bound.

Theorem 1. *If $\rho \in [1/2, 1]$ and $\eta > 1$, then*

$$R_\mu^{\text{OCUCB-}n}(n) \leq C_\eta \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \log \max\left\{\frac{n\Delta_i^2 \log(n)}{k_{i,\rho}}, \log(n)\right\}\right) ,$$

where $\Delta_i = \mu^* - \mu_i$ and $k_{i,\rho} = \sum_{j=1}^K \min\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\}$ and $C_\eta > 0$ is a constant that depends only on η. Furthermore, for all $\rho \in [0, 1]$ it holds that $\limsup_{n\to\infty} R_\mu^{\text{OCUCB-}n}(n)/\log(n) \leq \sum_{i:\Delta_i>0} \frac{2\eta}{\Delta_i}$.

Asymptotically the upper bound matches the lower bound given by Lai and Robbins [1985] except for a factor of η. In the non-asymptotic regime the additional terms inside the logarithm significantly improves on UCB. The bound in Theorem 1 corresponds to a worst-case regret that is suboptimal by a factor of just $\sqrt{\log \log n}$. Algorithms achieving the minimax rate are MOSS [Audibert and Bubeck, 2009] and OCUCB, but both require advance knowledge of the horizon. The quantity $k_{i,\rho} \in [1, K]$ may be interpreted as the number of "effective" arms with larger values leading to improved regret. A simple observation is that $k_{i,\rho}$ is always non-increasing in ρ, which makes $\rho = 1/2$ the canonical choice. In the special case that all suboptimal arms have the same expected payoff, then $k_{i,\rho} = K$ for all ρ. Interestingly I could not find a regime for which the algorithm is empirically sensitive to $\rho \in [1/2, 1]$. If $\rho = 1$, then except for $\log \log$ additive terms the problem dependent regret enjoyed by OCUCB-n is equivalent to OCUCB. Finally, if $\rho = 0$, then the asymptotic result above applies, but the algorithm in that case essentially reduces to MOSS, which is known to suffer suboptimal finite-time regret in certain regimes [Lattimore, 2015].

Intuition for regret bound. Let us fix a strategy π and mean vector $\mu \in \mathbb{R}^K$ and suboptimal arm i. Suppose that $\mathbb{E}[T_i(n)] \leq \Delta_i^{-2} \log(1/\delta)/2$ for some $\delta \in (0, 1)$. Now consider the alternative mean reward μ' with $\mu'_j = \mu_j$ for $j \neq i$ and $\mu'_i = \mu_i + 2\Delta_i$, which means that i is the optimal action for mean vector μ'. Standard information-theoretic analysis shows that μ and μ' are not statistically separable at confidence level δ and in particular, if Δ_i is large enough, then $R^\pi_{\mu'}(n) = \Omega(n\delta\Delta_i)$. For mean μ' we have $\Delta'_j = \mu'_i - \mu'_j \approx \max\{\Delta_i, \Delta_j\}$ and for any reasonable algorithm we would like

$$\sum_{j:\Delta'_j>0} \frac{\log(n)}{\Delta'_j} \gtrsim R^\pi_{\mu'}(n) = \Omega(n\delta\Delta_i) .$$

But this implies that δ should be chosen such that

$$\delta = O\left(\frac{\log(n)}{n} \sum_{j:\Delta'_j>0} \frac{1}{\Delta'_j \Delta_i}\right) = O\left(\frac{\log(n) k_{i,1/2}}{n\Delta_i^2}\right) ,$$

which up to $\log \log$ terms justifies the near-optimality of the regret guarantee given in Theorem 1 for ρ close to $1/2$. Of course Δ is not known in advance, so no algorithm can choose this confidence level. The trick is to notice that arms j with $\Delta_j \leq \Delta_i$ should be played about as often as arm i and arms j with $\Delta_j > \Delta_i$ should be played about as much as arm i until $T_j(t-1) \approx \Delta_j^{-2}$. This means that as $T_i(t-1)$ approaches the critical number of samples Δ_i^{-2} we can approximate

$$\sum_{j=1}^K \min\left\{T_i(t-1),\, T_j(t-1)^{\frac{1}{2}} T_i(t-1)^{\frac{1}{2}}\right\} \approx \sum_{j=1}^K \min\left\{\Delta_i^{-2},\, \Delta_j^{-1}\Delta_i^{-1}\right\} = \frac{k_{i,1/2}}{\Delta_i^2} .$$

Then the index used by OCUCB-n is justified by ignoring $\log \log$ terms and the usual $n \approx t$ used by UCB and other algorithms. Theorem 1 is proven by making the above approximation rigorous. The argument for this choice of confidence level is made concrete in Appendix A where I present a lower bound that matches the upper bound except for $\log \log(n)$ additive terms.

2 Concentration

The regret guarantees rely on a number of concentration inequalities. For this section only let $X_1, X_2, \ldots$ be i.i.d. 1-subgaussian and $S_n = \sum_{t=1}^n X_t$ and $\hat{\mu}_n = S_n/n$. The first lemma below is well known and follows trivially from the maximal inequality.

Important remark. For brevity I use $O_\eta(1)$ to indicate a constant that depends on η but not other variables such as n and μ. The dependence is never worse than polynomial in $1/(\eta - 1)$.

Lemma 2. *If $\varepsilon > 0$, then $\mathbb{P}\left\{\exists t \leq n : S_t \geq \varepsilon\right\} \leq \exp\left(-\frac{\varepsilon^2}{2n}\right)$.*

The following lemma analyses the likelihood that S_n ever exceeds $f(n) = \sqrt{2\eta n \log \max\left\{e, \log n\right\}}$ where $\eta > 1$. By the law of the iterated logarithm $\limsup_{n\to\infty} S_n / f(n) = \sqrt{1/\eta}$ a.s. and for small δ it has been shown by Garivier [2013] that

$$\mathbb{P}\left\{\exists n : S_n \geq \sqrt{2n \log\left(\frac{\log(n)}{\delta}\right)}\right\} = O(\delta)\,.$$

The case where $\delta = \Omega(1)$ seems not to have been analysed and relies on the usual peeling trick, but without the union bound.

Lemma 3. *There exists a monotone non-decreasing function $p : (1, \infty) \to (0, 1]$ such that for all $\eta > 1$ it holds that $\mathbb{P}\left\{\forall n : S_n \leq \sqrt{2\eta n \log \max\left\{e, \log n\right\}}\right\} \geq p(\eta)$.*

Lemma 4. *Let $b \geq 1$ and $\Delta > 0$ and $\tau = \min\left\{n : \sup_{t \geq n} \hat\mu_t + \sqrt{\frac{2\eta \log(b)}{t}} < \Delta\right\}$, then*

$$\mathbb{E}[\tau] \leq \sqrt{\mathbb{E}[\tau^2]} = O_\eta(1) \cdot \left(1 + \frac{1}{\Delta^2} \log_+(b)\right) \qquad \textit{where } \log_+(x) = \max\left\{1, \log(x)\right\}\,.$$

The final concentration lemma is quite powerful and forms the lynch-pin of the following analysis.

Lemma 5. *Let $\Delta > 0$ and $\rho \in [0, 1]$ and $d \in \{1, 2, \ldots\}$ and $\lambda_1, \ldots, \lambda_d \in [1, \infty]$ be constants. Furthermore, let α be the random variable given by*

$$\alpha = \inf\left\{\alpha \geq 0 : \inf_s \hat\mu_s + \sqrt{\frac{2\eta}{s} \log \max\left\{1, \frac{\alpha}{\sum_{i=1}^d \min\left\{s, \lambda_i^\rho s^{1-\rho}\right\}}\right\}} \geq -\Delta\right\}\,.$$

Finally let $\beta = \inf\left\{\beta \geq 0 : \beta \log(\beta) = \alpha\right\}$. Then

(a) If $\rho \in (1/2, 1]$, then $\Delta \mathbb{E}[\alpha] = O\left(\frac{1}{(2\rho - 1)(\eta - 1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\Delta^{-1}, \sqrt{\lambda_i}\right\}$

(b) If $\rho \in [1/2, 1]$, then $\Delta \mathbb{E}[\beta] = O\left(\frac{1}{(\eta - 1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\Delta^{-1}, \sqrt{\lambda_i}\right\}$

The proofs of Lemmas 3 to 5 may be found in Appendices B to D.

3 Analysis of the KL-UCB+ Algorithm

Let us warm up by analysing a simpler algorithm, which chooses the arm that maximises the following index.

$$\gamma_i(t) = \hat\mu_i(t-1) + \sqrt{\frac{2\eta}{T_i(t-1)} \log\left(\frac{t}{T_i(t-1)}\right)}\,. \qquad (2)$$

Strategies similar to this have been called KL-UCB+ and suggested as a heuristic by Garivier and Cappé [2011] (this version is specified to the subgaussian noise model). Recently Kaufmann [2016] has established the asymptotic optimality of strategies with approximately this form, but finite-time analysis has not been available until now. Bounding the regret will follow the standard path of bounding $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. Let $\hat\mu_{i,s}$ be the empirical estimate of the mean of the

ith arm having observed s samples. Define τ_i and τ_Δ by

$$\tau_i = \min\left\{t \geq 1/\Delta_i^2 : \sup_{s\geq t} \hat{\mu}_{i,s} + \sqrt{\frac{2\eta}{s}\log\max\left\{1, n\Delta_i^2\right\}} < \mu_i + \frac{\Delta_i}{2}\right\}$$

$$\tau_\Delta = \min\left\{t : \inf_{1\leq s\leq n} \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s}\log\max\left\{1, \frac{t}{s}\right\}} \geq \mu_1 - \frac{\Delta_i}{2}\right\}.$$

If $T_i(t-1) \geq \tau_i$ and $t \geq \tau_{\Delta_i}$, then by the definition of τ_{Δ_i} we have $\gamma_1(t) \geq \mu_i + \Delta_i/2$ and by the definition of τ_i

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\eta\log(t/T_i(t-1))}{T_i(t-1)}} \leq \hat{\mu}_i(t-1) + \sqrt{\frac{2\eta\log(n\Delta_i^2)}{T_i(t-1)}} < \mu_i + \frac{\Delta_i}{2},$$

which means that $I_t \neq i$. Therefore $T_i(n)$ may be bounded in terms of τ_i and τ_{Δ_i} as follows:

$$T_i(n) = \sum_{t=1}^{n} \mathbb{1}\{I_t = i\} \leq \tau_{\Delta_i} + \sum_{t=\tau_{\Delta_i}+1}^{n} \mathbb{1}\{I_t = i \text{ and } T_i(t-1) < \tau_i\} \leq \tau_i + \tau_{\Delta_i}.$$

It remains to bound the expectations of τ_i and τ_{Δ_i}. By Lemma 5a with $d = 1$ and $\rho = 1$ and $\lambda_1 = \infty$ it follows that $\mathbb{E}[\tau_{\Delta_i}] = O_\eta(1) \cdot \Delta_i^{-2}$ and by Lemma 4

$$\mathbb{E}[\tau_i] = O_\eta(1) \cdot \left(1 + \frac{1}{\Delta_i^2}\log_+(n\Delta_i^2)\right).$$

Therefore the strategy in Eq. (2) satisfies:

$$R_\mu^{\text{KL-UCB+}}(n) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] = O_\eta(1) \cdot \sum_{i:\Delta_i>0}\left(\Delta_i + \frac{1}{\Delta_i}\log_+(n\Delta_i^2)\right).$$

Remark 6. Without changing the algorithm and by optimising the constants in the proof it is possible to show that $\limsup_{n\to\infty} R_\mu^{\text{KL-UCB+}}(n)/\log(n) \leq \sum_{i:\Delta_i>0} 2\eta/\Delta_i$, which is just a factor of η away from the asymptotic lower bound of Lai and Robbins [1985].

4 Proof of Theorem 1

The proof follows along similar lines as the warm-up, but each step becomes more challenging, especially controlling τ_Δ.

Step 1: Setup and preliminary lemmas

Define Φ to be the random set of arms for which the empirical estimate never drops below the critical boundary given by the law of iterated logarithm.

$$\Phi = \left\{i > 2 : \hat{\mu}_{i,s} + \sqrt{\frac{2\eta_1 \log\max\left\{e, \log s\right\}}{s}} \geq \mu_i \text{ for all } s\right\}, \tag{3}$$

where $\eta_1 = (1+\eta)/2 \in (1, \eta)$. By Lemma 3, $\mathbb{P}\left\{i \in \Phi\right\} \geq p(\eta_1) > 0$. It will be important that Φ only includes arms $i > 2$ and that the events $i, j \in \Phi$ are independent for $i \neq j$. From the definition of the index γ and for $i \in \Phi$ it holds that $\gamma_i(t) > \mu_i$ for all t. The following lemma shows that the pull counts for optimistic arms "chase" those of other arms up the point that they become clearly suboptimal.

Lemma 7. *There exists a constant $c_\eta \in (0, 1)$ depending only on η such that if (a) $j \in \Phi$ and (b) $\hat{\mu}_i(t-1) \leq \mu_i + \Delta_i/2$ and (c) $T_j(t-1) \leq c_\eta \min\{T_i(t-1), \Delta_j^{-2}\}$, then $I_t \neq i$.*

Proof. First note that $T_j(t-1) \leq T_i(t-1)$ implies that $B_j(t-1) \geq B_i(t-1)$. Comparing the indices:

$$\gamma_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\eta\log B_i(t-1)}{T_i(t-1)}} \leq \mu_i + \sqrt{\frac{2\eta c_\eta \log B_j(t-1)}{T_j(t-1)}} + \frac{\Delta_i}{2}.$$

On the other hand, by choosing c_η small enough and by the definition of $j \in \Phi$:

$$\gamma_j(t) = \hat{\mu}_j(t-1) + \sqrt{\frac{2\eta \log B_j(t-1)}{T_j(t-1)}} \geq \mu_j + \sqrt{\frac{2\eta c_\eta \log B_j(t-1)}{T_j(t-1)}} + \sqrt{\frac{c_\eta}{T_j(t-1)}}$$
$$\geq \mu_1 + \sqrt{\frac{2\eta c_\eta \log B_j(t-1)}{T_j(t-1)}} > \gamma_i(t)\,,$$

which implies that $I_t \neq i$. □

Let $J = \min \Phi$ be the optimistic arm with the largest return where if $\Phi = \emptyset$ we define $J = K+1$ and $\Delta_J = \max_i \Delta_i$. By Lemma 3, $i \in \Phi$ with constant probability, which means that J is sub-exponentially distributed with rate dependent on η only. Define $K_{i,\rho}$ by

$$K_{i,\rho} = 1 + c_\eta \sum_{j \in \Phi, j \neq i} \min\left\{1,\, \frac{\Delta_i^{2\rho}}{\Delta_j^{2\rho}}\right\}, \tag{4}$$

where c_η is as chosen in Lemma 7. Since $\mathbb{P}\{i \in \Phi\} = \Omega_\eta(1)$ we will have $K_{i,\rho} = \Omega_\eta(k_{i,\rho})$ with high probability (this will be made formal later). Let

$$b_i = \max\left\{\frac{n\Delta_i^2 \log(n)}{k_{i,\rho}},\, \log(n),\, e\right\} \quad \text{and} \quad B_i = \max\left\{\frac{n\Delta_i^2 \log(n)}{K_{i,\rho}},\, \log(n),\, e\right\}$$
$$\tau_i = \min\left\{s \geq \frac{1}{\Delta_i^2} : \sup_{s' \geq s} \hat{\mu}_{i,s'} + \sqrt{\frac{2\eta}{s'}\log(B_i)} \leq \mu_i + \frac{\Delta_i}{2}\right\}. \tag{5}$$

The following lemma essentially follows from Lemma 4 and the fact that J is sub-exponentially distributed. Care must be taken because J and τ_i are not independent. The proof is found in Appendix E.

Lemma 8. $\mathbb{E}[\tau_i] \leq \mathbb{E}[J\tau_i] = O_\eta(1) \cdot \left(1 + \frac{1}{\Delta_i^2}\log(b_i)\right)$.

The last lemma in this section shows that if $T_i(t-1) \geq \tau_i$, then either i is not chosen or the index of the ith arm is not too large.

Lemma 9. *If $T_i(t-1) \geq \tau_i$, then $I_t \neq i$ or $\gamma_i(t) < \mu_i + \Delta_i/2$.*

Proof. By the definition of τ_i we have $\tau_i \geq \Delta_i^{-2}$ and $\hat{\mu}_i(t-1) \leq \mu_i + \Delta_i/2$. By Lemma 7, if $j \in \Phi$ and $T_j(t-1) \leq c_\eta \min\left\{\Delta_i^{-2}, \Delta_j^{-2}\right\}$, then $I_t \neq i$. Now suppose that $T_j(t-1) \geq c_\eta \min\left\{\Delta_i^{-2}, \Delta_j^{-2}\right\}$ for all $j \in \Phi$. Then

$$B_i(t-1) = \max\left\{e,\, \log(t),\, t\log(t)\left(\sum_{j=1}^K \min\left\{T_i(t-1),\, T_j(t-1)^\rho T_i(t-1)^{1-\rho}\right\}\right)^{-1}\right\}$$
$$\leq \max\left\{e,\, \log(n),\, \frac{n\Delta_i^2\log(n)}{K_{i,\rho}}\right\} = B_i\,.$$

Therefore from the definition of τ_i we have that $\gamma_i(t) < \mu_i + \Delta_i/2$. □

Step 2: Regret decomposition

By Lemma 9, if $T_i(n) \geq \tau_i$, then $I_t \neq i$ or $\gamma_i(t) < \mu_i + \Delta_i/2$. Now we must show there exists a j for which $\gamma_j(t) \geq \mu_i + \Delta_i/2$. This is true for arms i with $\Delta_i \geq 2\Delta_J$ since by definition $\gamma_J(t) > \mu_J \geq \mu_i + \Delta_i/2$ for all t. For the remaining arms we follow the idea used in Section 3 and define a random time for each $\Delta > 0$.

$$\tau_\Delta = \min\left\{t : \inf_{s \geq t} \sup_j \gamma_j(s) \geq \mu_1 - \frac{\Delta}{2}\right\}. \tag{6}$$

Then the regret is decomposed as follows

$$R_\mu^{\text{OCUCB-}n}(n) \leq \mathbb{E}\left[\sum_{i:\Delta_i>0} \Delta_i \tau_i + 2\Delta_J \tau_{\Delta_J/4} + \sum_{i:\Delta_i<\Delta_J/4} \Delta_i \tau_{\Delta_i}\right]. \tag{7}$$

The next step is to show that the first sum is dominant in the above decomposition, which will lead to the result via Lemma 8 to bound $\mathbb{E}[\Delta_i \tau_i]$.

Step 3: Bounding τ_Δ

This step is broken into two quite technical parts as summarised in the following lemma. The proofs of both results are quite similar, but the second is more intricate and is given in Appendix G.

Lemma 10. *The following hold:*

(a). $\mathbb{E}\left[\Delta_J \tau_{\Delta_J/4}\right] \leq O_\eta(1) \cdot \sum_{i:\Delta_i>0} \sqrt{1 + \frac{\log(b_i)}{\Delta_i^2}}$

(b). $\mathbb{E}\left[\sum_{i:\Delta_i<\Delta_J/4} \Delta_i \tau_{\Delta_i}\right] \leq O_\eta(1) \cdot \sum_{i:\Delta_i>0} \sqrt{1 + \frac{\log(b_i)}{\Delta_i^2}}$.

Proof of Lemma 10a. Preparing to use Lemma 5, let $\lambda \in (0,\infty]^K$ be given by $\lambda_i = \tau_i$ for i with $\Delta_i \geq 2\Delta_J$ and $\lambda_i = \infty$ otherwise. Now define random variable α by

$$\alpha = \inf\left\{\alpha \geq 0 : \inf_s \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s} \log\max\left\{1, \frac{\alpha}{\sum_{i=1}^K \min\left\{s, \lambda_i^\rho s^{1-\rho}\right\}}\right\}} \geq \mu_1 - \frac{\Delta_J}{8}\right\}$$

and $\beta = \min\left\{\beta \geq 0 : \beta\log(\beta) = \alpha\right\}$. Then for $t \geq \beta$ and abbreviating $s = T_1(t-1)$ we have

$$\begin{aligned}
\gamma_1(t) &= \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s}\log B_1(t-1)} \\
&= \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s}\log\left(\max\left\{e, \log(t), \frac{t\log(t)}{\sum_{i=1}^K \min\left\{s, T_i(t-1)^\rho s^{1-\rho}\right\}}\right\}\right)} \\
&\geq \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s}\log\max\left\{1, \frac{\alpha}{\sum_{i=1}^K \min\left\{s, T_i(t-1)^\rho s^{1-\rho}\right\}}\right\}} \\
&\geq \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s}\log\max\left\{1, \frac{\alpha}{\sum_{i=1}^K \min\left\{s, \lambda_i^\rho s^{1-\rho}\right\}}\right\}} \geq \mu_1 - \frac{\Delta_J}{8}\,,
\end{aligned}$$

where the second last inequality follows since for arms with $\Delta_i \geq 2\Delta_J$ we have $T_i(n) \leq \tau_i = \lambda_i$ and for other arms $\lambda_i = \infty$ by definition. The last inequality follows from the definition of α. Therefore $\tau_{\Delta_J/4} \leq \beta$ and so $\mathbb{E}[\Delta_J \tau_{\Delta_J/4}] \leq \mathbb{E}[\Delta_J \beta]$, which by Lemma 5b is bounded by

$$\begin{aligned}
\mathbb{E}[\Delta_J\beta] &= \mathbb{E}[\mathbb{E}[\Delta_J\beta|\lambda]] \leq O_\eta(1)\cdot\mathbb{E}\left[\mathbb{1}\{\Delta_J>0\}\sum_{i=1}^d \min\left\{\Delta_J^{-1}, \sqrt{\lambda_i}\right\}\right] \\
&\leq O_\eta(1)\cdot\mathbb{E}\left[\sum_{i:\lambda_i=\infty,\Delta_i=0} \frac{\mathbb{1}\{\Delta_J>0\}}{\Delta_{\text{min}}} + \sum_{i:\Delta_i>0}\sqrt{\tau_i}\right] \leq O_\eta(1)\cdot\mathbb{E}\left[\sum_{i:\Delta_i>0}\sqrt{\tau_i}\right],
\end{aligned} \tag{8}$$

where the last line follows since $\mathbb{E}[J] = O_\eta(1)$ and

$$\mathbb{E}\left[\sum_{i:\lambda_i=\infty,\Delta_i=0}\frac{\mathbb{1}\{\Delta_J>0\}}{\Delta_{\text{min}}}\right] \leq \mathbb{E}\left[\frac{J}{\Delta_{\text{min}}}\right] \leq O_\eta(1)\cdot\frac{1}{\Delta_{\text{min}}} \leq O_\eta(1)\max\left\{i : \sqrt{\mathbb{E}[\tau_i]}\right\}.$$

The resulting is completed substituting $\mathbb{E}[\sqrt{\tau_i}] \leq \sqrt{\mathbb{E}[\tau_i]}$ into Eq. (8) and applying Lemma 8 to show that $\mathbb{E}[\tau_i] \leq O_\eta(1) \cdot \left(1 + \frac{\log(b_i)}{\Delta_i^2}\right)$. □

Step 4: Putting it together

By substituting the bounds given in Lemma 10 into Eq. (7) and applying Lemma 8 we obtain

$$\begin{aligned} R_\mu^{\text{OCUCB-}n}(n) &\leq \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[\tau_i] + O_\eta(1) \cdot \sum_{i:\Delta_i>0} \sqrt{1 + \frac{\log(b_i)}{\Delta_i^2}} \\ &\leq O_\eta(1) \cdot \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{1}{\Delta_i} \log \max\left\{\frac{n\Delta_i^2 \log(n)}{k_{i,\rho}},\, \log(n),\, e\right\}\right) , \end{aligned}$$

which completes the proof of the finite-time bound.

Asymptotic analysis. Lemma 5 makes this straightforward. Let $\varepsilon_n = \min\{\frac{\Delta_{\min}}{2}, \log^{-\frac{1}{4}}(n)\}$ and

$$\alpha_n = \min\left\{\alpha : \inf_s \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s} \log\left(\frac{\alpha}{Ks}\right)} \geq \mu_1 - \varepsilon_n\right\} .$$

Then by Lemma 5a with $\rho = 1$ and $\lambda_1, \ldots, \lambda_K = \infty$ we have $\sup_n \mathbb{E}[\alpha_n]\varepsilon^2/K \leq O_\eta(1)$. Then we modify the definition of τ by

$$\tau_{i,n} = \min\left\{s : \sup_{s' \geq s} \hat{\mu}_{i,s} + \sqrt{\frac{2\eta}{s} \log(n \log(n))} \leq \mu_1 - \varepsilon_n\right\} ,$$

which is chosen such that if $T_i(t-1) \geq \tau_{i,n}$, then $\gamma_i(t) \leq \mu_1 - \varepsilon_n$. Therefore

$$R_\mu^{\text{OCUCB-}n}(n) \leq \Delta_{\max}\mathbb{E}[\alpha_n] + \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[\tau_{i,n}] \leq O_\eta(1) \cdot \frac{\Delta_{\max} K}{\varepsilon_n^2} + \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[\tau_{i,n}] .$$

Classical analysis (eg., by Garivier and Cappé [2011]) shows that $\limsup_{n\to\infty} \mathbb{E}[\tau_{i,n}]/\log(n) \leq 2\eta\Delta_i^{-2}$ and $\lim_{n\to\infty} \varepsilon_n^{-2}/\log(n) = 0$, which implies the asymptotic claim in Theorem 1.

$$\limsup_{n\to\infty} \frac{R_\mu^{\text{OCUCB-}n}(n)}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{2\eta}{\Delta_i} .$$

This naive calculation demonstrates a weakness of asymptotic results. The $\Delta_{\max} K \varepsilon_n^{-2}$ term in the regret will typically dominate the higher-order terms except when n is outrageously large. A more careful argument (similar to the derivation of the finite-time bound) would lead to the same asymptotic bound via a nicer finite-time bound, but the details are omitted for readability. Interestingly the result is not dependent on ρ and so applies also to the MOSS-type algorithm that is recovered by choosing $\rho = 0$.

5 Discussion

The UCB family has a new member. This one is tuned for subgaussian noise and roughly mimics the OCUCB algorithm, but without needing advance knowledge of the horizon. The introduction of $k_{i,\rho}$ is a minor refinement on previous measures of difficulty, with the main advantage being that it is very intuitive. The resulting algorithm is efficient and close to optimal theoretically. Of course there are open questions, some of which are detailed below.

Shrinking the confidence level. Empirically the algorithm improves significantly when the logarithmic terms in the definition of $B_i(t-1)$ are dropped. There are several arguments that theoretically justify this decision. First of all if $\rho > 1/2$, then it is possible to replace the $t\log(t)$ term in the definition of $B_i(t-1)$ with just t and use part (a) of Lemma 5 instead of part (b). The price is that the regret guarantee explodes as ρ tends to $1/2$ (also not observed in practice). The second improvement is to replace $\log(t)$ in the definition of $B_i(t-1)$ with

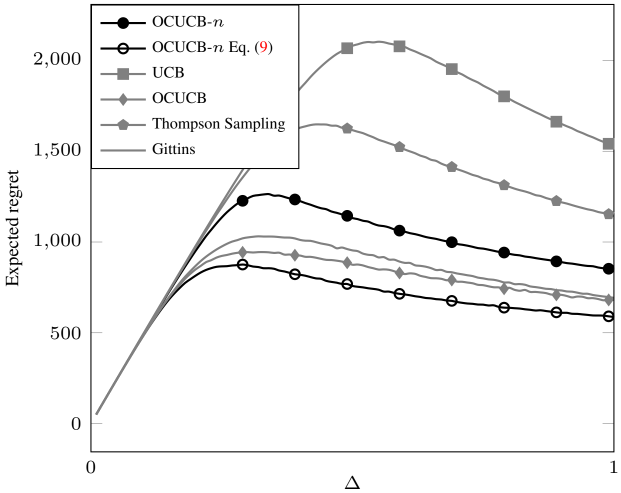
$$\max\left\{0,\, \log\left(t \cdot \left(\sum_{j=1}^{K} \min\left\{T_i(t-1), T_j(t-1)^{\rho} T_i(t-1)^{1-\rho}\right\}\right)^{-1}\right)\right\}, \tag{9}$$

which boosts empirical performance and rough sketches suggest minimax optimality is achieved. I leave details for a longer article.

Improving analysis and constants. Despite its simplicity relative to OCUCB, the current analysis is still significantly more involved than for other variants of UCB. A cleaner proof would obviously be desirable. In an ideal world we could choose $\eta = 1$ or (slightly worse) allow it to converge to 1 as t grows, which is the technique used in the KL-UCB algorithm [Cappé et al., 2013, and others]. I anticipate this would lead to an asymptotically optimal algorithm.

Informational confidence bounds. Speaking of KL-UCB, if the noise model is known more precisely (for example, it is bounded), then it is beneficial to use confidence bounds based on the KL divergence. Such bounds are available and could be substituted directly to improve performance without loss [Garivier, 2013, and others]. Repeating the above analysis, but exploiting the benefits of tighter confidence intervals would be an interesting (non-trivial) problem due to the need to exploit the non-symmetric KL divergences. It is worth remarking that confidence bounds based on the KL divergence are also *not* tight. For example, for Gaussian random variables they lead to the right exponential rate, but with the wrong leading factor, which in practice can improve performance as evidenced by the confidence bounds used by (near) Bayesian algorithms that exactly exploit the noise model (eg., Kaufmann et al. [2012], Lattimore [2016], Kaufmann [2016]). This is related to the "missing factor" in Hoeffding's bound studied by Talagrand [1995].

Precise lower bounds. Perhaps the most important remaining problem for the subgaussian noise model is the question of lower bounds. Besides the asymptotic results by Lai and Robbins [1985] and Burnetas and Katehakis [1997] and the finite-time analysis in the two-armed case by Kulkarni and Lugosi [2000], there has been some recent progress on finite-time lower bounds, both in the OCUCB paper and a recent article by Garivier et al. [2016]. Results formalising the intuition in the introduction are given in Appendix A, but still there are regimes where the bounds are not matching.



Empirical evaluation. I plot the regret in the worst-case regime where $K = 100$ and $n = 5000$ and all sub-optimal arms have $\Delta_i = \Delta > 0$. Error bars are too small to see and all code/data will be made available with any final version. Each data point is the average of 14113 i.i.d. samples. There are two versions of OCUCB-n, both with $\eta = 1$. The first using B_i as given after Eq. (1) and the second using Eq. (9). The performance is compared to UCB (as defined by Katehakis and Robbins [1995]), OCUCB ($\alpha = 3$, $\psi = 2$), Thompson sampling and the finite-horizon Gittins index strategy. The two Bayesian algorithms use a flat Gaussian prior. Besides OCUCB-n, only UCB and Thompson sampling do not depend on the horizon. A more extensive empirical evaluation would be quite interesting, especially when the gaps are unbalanced or even better with real data. Other synthetic evaluations were consistent with the results above, with OCUCB-n performing comparably with the state-of-the-art in all regimes, and sometimes better. These results are omitted due to space constraints.

References

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.

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.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

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.

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

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of Conference on Learning Theory (COLT)*, 2011.

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

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*, pages 592–600, 2012.

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 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*, 2015.

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

Michel Talagrand. The missing factor in Hoeffding's inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 689–702, 1995.

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

A Lower Bounds

I now prove a kind of lower bound showing that the form of the regret in Theorem 1 is approximately correct for ρ close to $1/2$. The result contains a lower order $-\log\log(n)$ term, which for large n dominates the improvements, but is meaningful in many regimes.

Theorem 11. *Assume a standard Gaussian noise model and let π be any strategy and $\mu \in [0,1]^K$ be such that $\frac{n\Delta_i^2}{k_{i,1/2}\log(n)} \geq 1$ for all i. Then one of the following holds:*

1. *$R^\pi_\mu(n) \geq \frac{1}{4} \sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \log\left(\frac{n\Delta_i^2}{2k_{i,1/2}\log(n)}\right)$.*
2. *There exists an i with $\Delta_i > 0$ such that*

$$R^\pi_{\mu'}(n) \geq \frac{1}{2} \sum_{i:\Delta'_i>0} \frac{1}{\Delta'_i} \log\left(\frac{n\Delta_i'^2}{2k'_{i,1/2}\log(n)}\right)$$

where $\mu'_i = \mu_i + 2\Delta_i$ and $\mu'_j = \mu_j$ for $j \neq i$ and Δ'_i and $k'_{i,\rho}$ are defined as Δ_i and $k_{i,\rho}$ but using μ'.

Proof. On our way to a contradiction, assume that neither of the items hold. Let i be a suboptimal arm and μ' be as in the second item above. I write $\mathbb{P}'$ and $\mathbb{E}'$ for expectation when when rewards are sampled from μ'. Suppose

$$\mathbb{E}[T_i(n)] \leq \frac{1}{4\Delta_i^2} \log\left(\frac{n\Delta_i^2}{2k_{i,1/2}\log(n)}\right) . \tag{10}$$

Then Lemma 2.6 in the book by Tsybakov [2008] and the same argument as used by Lattimore [2015] gives

$$\mathbb{P}\left\{T_i(n) \geq n/2\right\} + \mathbb{P}'\left\{T_i(n) < n/2\right\} \geq \frac{k_{i,1/2}\log(n)}{n\Delta_i^2} \equiv 2\delta\,.$$

By Markov's inequality

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

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

$$R^\pi_{\mu'}(n) \geq \frac{\delta n \Delta_i}{2} = \frac{1}{2}\sum_{j=1}^K \min\left\{\frac{1}{\Delta_i}, \frac{1}{\Delta_j}\right\} \log(n) \geq \frac{1}{2} \sum_{j:\Delta'_j>0} \frac{1}{\Delta'_j} \left(\frac{n\Delta'_j}{2k'_{j,1/2}\log(n)}\right) ,$$

which is a contradiction. Therefore Eq. (10) does not hold for all i with $\Delta_i > 0$, but this also leads immediately to a contradiction, since then

$$R^\pi_\mu(n) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] \geq \frac{1}{4} \sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \log\left(\frac{n\Delta_i^2}{2k_{i,1/2}\log(n)}\right) . \qquad \square$$

B Proof of Lemma 3

Monotonicity is obvious. Let $\varepsilon > 0$ be such that $\eta = 1 + 2\varepsilon$ and and $G_k = [(1+\varepsilon)^k, (1+\varepsilon)^{k+1}]$ and $F_k = \left\{\exists n \in G_k : S_n > \sqrt{2\eta n \log\max\left\{e, \log n\right\}}\right\}$. Then

$$\mathbb{P}\left\{\forall n : S_n \leq \sqrt{2\eta n \log\max\left\{e, \log n\right\}}\right\} = \mathbb{P}\left\{\forall k \geq 0 : \neg F_k\right\} = \prod_{k=0}^\infty \mathbb{P}\left\{\neg F_k | \neg F_1, \ldots, \neg F_{k-1}\right\} .$$

Now we analyse the failure event F_k.

$$\begin{aligned}
\mathbb{P}\left\{F_k | \neg F_1, \ldots, \neg F_{k-1}\right\} &\leq \mathbb{P}\left\{F_k\right\} \\
&= \mathbb{P}\left\{\exists n \in G_k : S_n > \sqrt{2\eta n \log\max\left\{e, \log n\right\}}\right\} \\
&\leq \exp\left(-\frac{2\eta(1+\varepsilon)^k \log_+ \log(1+\varepsilon)^k}{2(1+\varepsilon)^{k+1}}\right) \\
&= \left(\frac{1}{k\log(1+\varepsilon)}\right)^{1+\frac{\varepsilon}{1+\varepsilon}} .
\end{aligned}$$

Since this is vacuous when k is small we need also need a naive bound.

$$\mathbb{P}\left\{\exists n \in G_k : S_n \geq \sqrt{2\eta n \log \max\{e, \log n\}}\right\} \leq \exp(-\eta) < 1\,.$$

Combining these completes the results since for sufficiently large k_0 (depending only on η) we have that

$$p(\eta) \geq \exp(-\eta k_0) \prod_{k=k_0}^{\infty} (1 - \mathbb{P}\{F_k\}) \geq \exp(-\eta k_0) \prod_{k=k_0}^{\infty} \left(1 - \left(\frac{1}{k \log(1+\varepsilon)}\right)^{1+\frac{\varepsilon}{1+\varepsilon}}\right) > 0\,.$$

C Proof of Lemma 4

Let $\alpha \geq 1$ be fixed and $t_0 = \lceil 8\eta \log_+(b)/\Delta^2 \rceil$ and $t_k = t_0 2^k$. Then

$$\mathbb{P}\{\tau \geq \alpha t_0\} \leq \mathbb{P}\{\exists t \geq \alpha t_0 : \hat{\mu}_t \geq \Delta/2\} \leq \sum_{k=0}^{\infty} \mathbb{P}\left\{\exists t \leq t_k : S_t \geq \alpha 2^{k-1} t_0 \Delta/2\right\}$$

$$\leq \sum_{k=0}^{\infty} \exp\left(-\frac{\alpha^2 2^{2k-2} t_0^2 \Delta^2}{8\alpha 2^k t_0}\right) \leq \sum_{k=0}^{\infty} \exp\left(-\frac{\alpha 2^k}{4}\right) = O\left(\exp(-\alpha/4)\right)\,.$$

Therefore $\mathbb{E}\left[(\tau/t_0)^2\right] = O(1)$ and so the result follows.

D Proof of Lemma 5

Let $\eta_1 = (1+\eta)/2$ and $\eta_2 = \eta/\eta_1$ and

$$\Lambda = \sum_{i=1}^{d} \min\left\{\frac{1}{\Delta^2}, \frac{\lambda_i^\rho}{\Delta^{2-2\rho}} \log_+\left(\frac{1}{\lambda_i \Delta^2}\right)\right\}\,.$$

Let $x > 0$ be fixed and let $G_k = [\eta_1^k, \eta_1^{k+1}]$. We will use the peeling trick. First, by Lemma 2.

$$q_k = \mathbb{P}\left\{\inf_{s \in G_k} \hat{\mu}_s + \sqrt{\frac{2\eta}{s} \log \max\left\{1, \frac{x\Lambda}{\sum_{i=1}^d \min\{s,\, \lambda_i^\rho s^{1-\rho}\}}\right\}} \leq -\Delta\right\}$$

$$\leq \mathbb{P}\left\{\exists s \leq \eta_1^{k+1} : S_s + \sqrt{2\eta\eta_1^k \log\max\left\{1, \frac{x\Lambda}{\sum_{i=1}^d \min\left\{\eta_1^{k+1}, \lambda_i^\rho \eta_1^{(k+1)(1-\rho)}\right\}}\right\}} + \Delta \eta_1^k \leq 0\right\}$$

$$\overset{(a)}{\leq} \left(\frac{\sum_{i=1}^d \min\left\{\eta_1^{k+1}, \lambda_i^\rho \eta_1^{(k+1)(1-\rho)}\right\}}{x\Lambda}\right)^{\eta_2} \exp\left(-\frac{\Delta^2 \eta_1^{k-1}}{2}\right)$$

$$= \left(\frac{\sum_{i=1}^d \min\left\{\eta_1^{k+1}, \lambda_i^\rho \eta_1^{(k+1)(1-\rho)}\right\}}{x\Lambda} \exp\left(-\frac{\Delta^2 \eta_1^k}{2\eta}\right)\right)^{\eta_2}\,,$$

where (a) follows by Lemma 2. By the union bound

$$\mathbb{P}\left\{\inf_s \hat{\mu}_s + \sqrt{\frac{2\eta}{s} \log\max\left\{1, \frac{x\Lambda}{\sum_{i=1}^d \min\{s^\rho,\, \lambda_i s^{1-\rho}\}}\right\}} \leq -\Delta\right\} \leq \sum_{k=0}^{\infty} q_k$$

$$\leq \sum_{k=0}^{\infty} \left(\frac{\sum_{i=1}^d \min\left\{\eta_1^{k+1},\, \lambda_i^\rho \eta_1^{(k+1)(1-\rho)}\right\}}{x\Lambda} \exp\left(-\frac{\Delta^2 \eta_1^k}{2\eta}\right)\right)^{\eta_2}$$

$$\leq \left(\frac{1}{x\Lambda} \sum_{i=1}^{d} \sum_{k=0}^{\infty} \min\left\{\eta_1^{k+1},\, \lambda_i^\rho \eta_1^{(k+1)(1-\rho)}\right\} \exp\left(-\frac{\Delta^2 \eta_1^k}{2\eta}\right)\right)^{\eta_2}$$

$$= O\left(\frac{\eta}{\eta - 1}\right) \cdot x^{-\eta_2}\,,$$

where the last line follows from Lemma 12. Therefore $\mathbb{P}\left\{\alpha \geq x\Lambda\right\} \leq O\left(\frac{\eta}{\eta-1}\right) \cdot x^{-\eta_2}$.

Now the first part follows easily since $\mathbb{E}[\alpha] \leq \int_0^\infty \mathbb{P}\left\{\alpha \geq x\Lambda\right\} = O\left(\frac{\eta}{(\eta-1)^2}\right) \cdot \Lambda$. Therefore

$$\begin{aligned}\Delta\mathbb{E}[\alpha] &\leq O\left(\frac{\eta}{(\eta-1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\frac{1}{\Delta},\, \lambda_i^\rho \Delta^{2\rho-1} \log_+\left(\frac{1}{\lambda_i \Delta^2}\right)\right\} \\ &\leq O\left(\frac{\eta}{(2\rho-1)(\eta-1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\frac{1}{\Delta},\, \sqrt{\lambda_i}\right\} .\end{aligned}$$

For the second part let $x_0 = \Lambda/\operatorname{productlog}(\Lambda)$ where $\operatorname{productlog}$ is the inverse of the function $x \to x\exp(x)$.

$$\begin{aligned}\mathbb{E}[\beta] &\leq \int_0^\infty \mathbb{P}\left\{\beta \geq x\right\} dx \leq x_0 + \int_{x_0}^\infty \mathbb{P}\left\{\alpha \geq \frac{x}{\Lambda}\log(x)\right\} dx \\ &\leq x_0 + O\left(\frac{\eta}{\eta-1}\right) \cdot \int_{x_0}^\infty \left(\frac{\Lambda}{x\log(x)}\right)^{\eta_2} dx \\ &\leq x_0 + O\left(\frac{\eta}{\eta-1}\right) \cdot \left(\frac{\Lambda}{\log(x_0)}\right)^{\eta_2} \int_{x_0}^\infty x^{-\eta_2} dx \\ &\leq x_0 + O\left(\frac{\eta}{(\eta-1)^2}\right) \cdot \left(\frac{\Lambda}{\log(x_0)}\right)^{\eta_2} x_0^{1-\eta_2} = O\left(\frac{\eta}{(\eta-1)^2}\right) \cdot \frac{\Lambda}{\operatorname{productlog}(\Lambda)} .\end{aligned}$$

If $\Lambda < e$, then the result is trivial. For $\Lambda \geq e$ we have $\operatorname{productlog}(\Lambda) \geq 1$. Then

$$\begin{aligned}\Delta\mathbb{E}[\beta] &\leq O\left(\frac{1}{(\eta-1)^2}\right) \cdot \frac{\Delta\Lambda}{\operatorname{productlog}(\Lambda)} \\ &\leq O\left(\frac{1}{(\eta-1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\frac{1}{\Delta},\, \frac{\lambda_i^\rho \Delta^{2\rho-1}}{\operatorname{productlog}(\Lambda)} \log_+\left(\frac{1}{\lambda_i\Delta^2}\right)\right\} .\end{aligned}$$

By examining the inner minimum we see that if $\Delta \geq \lambda_i^{-\frac{1}{2}}$, then $1/\Delta \leq \lambda_i^{\frac{1}{2}}$. If $\Delta < \lambda_i^{-\frac{1}{2}}$, then

$$\begin{aligned}\min\left\{\frac{1}{\Delta},\, \frac{\lambda_i^\rho \Delta^{2\rho-1}}{\operatorname{productlog}(\Lambda)} \log_+\left(\frac{1}{\lambda_i\Delta^2}\right)\right\} &< \frac{\lambda_i^{\frac{1}{2}}}{\max\left\{1,\, \operatorname{productlog}(\Delta^{-2})\right\}} \log_+\left(\frac{1}{\lambda_i\Delta^2}\right) \\ &\leq 2\lambda_i^{\frac{1}{2}} .\end{aligned}$$

Therefore $\mathbb{E}[\Delta t] \leq O\left(\frac{\eta}{(\eta-1)^2}\right) \cdot \sum_{i=1}^d \min\left\{\Delta^{-1}, \sqrt{\lambda_i}\right\}$ as required.

E Proof of Lemma 8

Since J is sub-exponentially distributed with rate dependent only on η we have $\sqrt{\mathbb{E}[J^2]} = O(1)$. By using Lemma 4 we obtain

$$\begin{aligned}\sqrt{\mathbb{E}[\tau_i^2]} &= \sqrt{\mathbb{E}[\mathbb{E}\left[\tau_i^2 | K_{i,\rho}\right]]} \\ &= O_\eta(1) \cdot \sqrt{\mathbb{E}\left[\left(1 + \frac{1}{\Delta_i^2}\log(B_i)\right)^2\right]} = O_\eta(1) \cdot \left(1 + \frac{1}{\Delta_i^2}\log\left(b_i\right)\right) .\end{aligned}$$

The latter inequality follows by noting that $B_i \geq e$ and $(1 + c\log(x))^2$ is concave for $x \geq e$ and $c > 0$.

$$\begin{aligned}\sqrt{\mathbb{E}\left[\left(1 + \frac{1}{\Delta_i^2}\log(B_i)\right)^2\right]} &\leq 1 + \frac{1}{\Delta_i^2}\log(\mathbb{E}[B_i]) \\ &= 1 + \frac{1}{\Delta_i^2}\log\left(\mathbb{E}\left[\max\left\{\log(n),\, \frac{n\Delta_i^2\log(n)}{K_{i,\rho}}\right\}\right]\right) \\ &= O_\eta(1)\cdot\left(1 + \frac{1}{\Delta_i^2}\log\left(\max\left\{\log(n),\, \frac{n\Delta_i^2\log(n)}{k_{i,\rho}}\right\}\right)\right)\,,\end{aligned}$$

where the last inequality follows from (a) $K_{i,\rho} \geq 1$ and (b) Azuma's concentration inequality implies that $\mathbb{P}\left\{K_{i,\rho} \leq c_\eta\rho(\eta)k_{i,\rho}/2\right\} = O(k_{i,\rho}^{-1})$ as shown in the following appendix. Finally by Holder's inequality

$$\mathbb{E}[J\tau_i] \leq \sqrt{\mathbb{E}[J^2]\mathbb{E}[\tau_i^2]} \leq O_\eta(1)\cdot\left(1 + \frac{1}{\Delta_i^2}\log\left(b_i\right)\right)\,.$$

F Tail Bound on $K_{i,\rho}$

Recall that $K_{i,\rho} = 1 + c_\eta\sum_{j\in\Phi,j\neq i}\min\left\{1,\, \Delta_i^{2\rho}/\Delta_j^{2\rho}\right\}$ and $k_{i,\rho} = 1 + \sum_{j\neq i}^K \min\left\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\right\}$. Therefore by Azuma's inequality and naive simplification we have

$$\begin{aligned}\mathbb{P}\left\{K_{i,\rho} \leq c_\eta\rho(\eta)k_{i,\rho}/2\right\} &\leq \mathbb{P}\left\{\sum_{j\in\Phi,j\neq i}\min\left\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\right\} \leq \frac{\rho(\eta)}{2}\sum_{j\neq i}\min\left\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\right\}\right\} \\ &\overset{(a)}{\leq} \exp\left(-\frac{\left(\rho(\eta)\sum_{j\neq i}\min\left\{1, \Delta_i^{2\rho}/\Delta_J^{2\rho}\right\}\right)^2}{2\sum_{j\neq i}\min\left\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\right\}^2}\right) \\ &\overset{(b)}{\leq} \exp\left(-\frac{\rho(\eta)^2\sum_{j\neq i}\min\left\{1, \Delta_i^{2\rho}/\Delta_J^{2\rho}\right\}}{2}\right) \\ &\overset{(c)}{=} O(k_{i,\rho}^{-1})\,,\end{aligned}$$

where (a) follows from Azuma's inequality and (b) since $\min\{1, x\}^2 \leq \min\{1, x\}$ and (c) by $\exp(-x) \leq 1/x$ for all $x \geq 0$.

G Proof of Lemma 10b

Recall that we are trying to show that

$$\mathbb{E}\left[\sum_{i:\Delta_i < \Delta_J/4}\Delta_i\tau_{\Delta_i}\right] = O\left(\sum_{i:\Delta_i > 0}\Delta_i\mathbb{E}[J\tau_i]\right)\,. \tag{11}$$

Let E be the event that $\Delta_2 \leq \Delta_J/4$ and define random sets $A_1 = \{i : \Delta_i \in (2\Delta_J, \infty)\}$ and $A_2 = \{i : \Delta_i \in [\Delta_J, 2\Delta_J]\}$. For $i \in A_1$ we have $\Delta_i > 2\Delta_J$ and since $J \in \Phi$ we have $\gamma_J(t) \geq \mu_J \geq \mu_1 - \Delta_i/2$. Therefore $i \in A_1$ implies that $\tau_{\Delta_i} = 1$ and so $T_i(n) \leq \tau_i$. Let $\lambda \in (0, \infty]^K$ be given by $\lambda_i = \tau_i$ for $i \in A_1$ and $\lambda_i = \infty$ otherwise.

$$\alpha = \min\left\{\alpha : \inf_s \hat{\mu}_{2,s} + \sqrt{\frac{2\eta}{s}\log\max\left\{1,\, \frac{\alpha}{\sum_{i=1}^K \min\left\{s, \lambda_i^\rho s^{1-\rho}\right\}}\right\}} \geq \mu_2 - \frac{\Delta_J}{4}\right\}\,.$$

It is important to note that we have used $\hat{\mu}_{2,s}$ in the definition of α and not $\hat{\mu}_{1,s}$ that appeared in the proof of part (a) of this lemma. The reason is to preserve independence when samples from the first arm are used later. Let $\beta = \min\{\beta \geq 0 : \beta \log(\beta) = \alpha\}$. If E holds, then for $t \geq \beta$ we have $\gamma_2(t) \geq \mu_2 - \Delta_J/4 \geq \mu_1 - \Delta_J/2$, which implies that

$$\mathbb{1}\{E\} \sum_{i \in A_2} T_i(n) \leq t_{\Delta_J} + \sum_{i \in A_2} \tau_i \leq \beta + \sum_{i \in A_2} \tau_i \,.$$

Therefore for any $s, t \leq n$ the concavity of $\min\{s, \cdot\}$ and $x \to x^\rho$ combined with Jensen's inequality implies that

$$\mathbb{1}\{E\} \sum_{i \in A_2} \min\left\{s, T_i(t-1)^\rho s^{1-\rho}\right\} \leq \sum_{i \in A_2} \min\left\{s, \left(\frac{\beta + \sum_{i \in A_2} \tau_i}{|A_2|}\right)^\rho s^{1-\rho}\right\} .$$

We are getting close to an application of Lemma 5. Let $\omega \in (0, \infty]^K$ be given by

$$\omega_j = \begin{cases} \tau_j & \text{if } j \in A_1 \\ \beta/|A_2| + \sum_{j \in A_2} \tau_j/|A_2| & \text{if } j \in A_2 \\ \infty & \text{otherwise}\,, \end{cases}$$

which has been chosen such that for $T_1(t-1) = s$ and if E holds, then

$$\begin{aligned} B_1(t-1) &\geq \max\left\{1, \frac{t\log(t)}{\sum_{j=1}^K \min\left\{s, T_j(t-1)^\rho s^{1-\rho}\right\}}\right\} \\ &\geq \max\left\{1, \frac{t\log(t)}{\sum_{j=1}^K \min\left\{s, \omega_j^\rho s^{1-\rho}\right\}}\right\} . \end{aligned} \tag{12}$$

Now let i be the index of some arm for which $\Delta_i < \Delta_J/4$ and define

$$\alpha_i = \min\left\{\alpha : \inf_s \hat{\mu}_{1,s} + \sqrt{\frac{2\eta}{s} \log\max\left\{1, \frac{\alpha}{\sum_{j=1}^K \min\left\{s, \omega_j^\rho s^{1-\rho}\right\}}\right\}} \geq \mu_1 - \frac{\Delta_i}{2}\right\}$$

and $\beta_i = \min\{\beta \geq 0 : \beta\log(\beta) = \alpha_i\}$. Therefore by Eq. (12), if E holds and $t \geq \beta_i$, then $\gamma_1(t) \geq \mu_1 - \Delta_i/2$ and so $t_{\Delta_i} \leq \beta_i$. At last we are able to write t_{Δ_i} in terms of something for which the expectation can be controlled.

$$\begin{aligned} \mathbb{E}\left[\sum_{i:\Delta_i < \Delta_J/4} \Delta_i \tau_{\Delta_i}\right] &\leq \mathbb{E}\left[\sum_{i:\Delta_i < \Delta_J/4} \Delta_i \beta_i\right] \\ &\leq O_\eta(1) \cdot \mathbb{E}\left[\sum_{i:\Delta_i < \Delta_J/4} \sum_{j=1}^K \min\left\{\frac{1}{\Delta_i}, \sqrt{\omega_j}\right\}\right] \\ &\leq O_\eta(1) \cdot \mathbb{E}\left[\sum_{i:\Delta_i < \Delta_J/4} \left(\sum_{j \in A_1} \sqrt{\tau_j} + |A_2|\sqrt{\omega_j} + \frac{J}{\Delta_{\min}}\right)\right] \\ &\leq O_\eta(1) \cdot \mathbb{E}\left[\sum_{j \in A_1} J\sqrt{\tau_j} + \frac{J^2}{\Delta_{\min}} + J|A_2|\sqrt{\omega_j}\right] . \end{aligned} \tag{13}$$

The first two terms are easily bounded as we shall soon see. For the last we have

$$\begin{aligned} \mathbb{E}\left[J|A_2|\sqrt{\omega_j}\right] \leq O_\eta(1) \cdot \sqrt{\mathbb{E}\left[|A_2|^2 \omega_j\right]} &= O_\eta(1) \cdot \sqrt{\mathbb{E}\left[|A_2| \sum_{j \in A_2} \tau_j + |A_2|\beta\right]} \\ &\leq O_\eta(1) \cdot \left(\sqrt{\mathbb{E}\left[|A_2| \sum_{j \in A_2} \tau_j\right]} + \sqrt{\mathbb{E}\left[|A_2|\beta\right]}\right) \end{aligned} \tag{14}$$

Bounding each term separately. For the first, let $\tilde{A}_\ell = \left\{j : \Delta_j \in [2^\ell, 2^{\ell+2})\right\}$, which is chosen such that no matter the value of Δ_J there exists an $\ell \in \mathbb{Z}$ with $A_2 \subseteq A_\ell$.

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|A_2| \sum_{j \in A_2} \tau_j\right]} &\leq O(1) \cdot \sqrt{\sum_{\ell \in \mathbb{Z}} |\tilde{A}_\ell| \sum_{j \in \tilde{A}_\ell} \mathbb{E}[\tau_j]} \\
&\leq O(1) \cdot \sqrt{\sum_{\ell \in \mathbb{Z}} |\tilde{A}_\ell|^2 \max_{j \in A_\ell} \mathbb{E}[\tau_j]} \\
&\leq O_\eta(1) \cdot \sum_{j : \Delta_j > 0} \sqrt{1 + \frac{\log(b_j)}{\Delta_j^2}}\,, \qquad (15)
\end{aligned}$$

where the last inequality follows because $\sum_{\ell \in \mathbb{Z}} \mathbb{1}\{j \in \tilde{A}_\ell\} = 2$ for each j and from Lemma 8, which gives the same order-bound on $\mathbb{E}[\tau_j]$ for all $j \in \tilde{A}_\ell$ for fixed ℓ. For the second term in Eq. (14) we have

$$\begin{aligned}
\mathbb{E}\left[\sqrt{|A_2|\beta}\right] &\overset{(a)}{\leq} O_\eta(1) \cdot \mathbb{E}\left[\sqrt{|A_2|\left(\sum_{j : \lambda_j = \infty} \frac{1}{\Delta_J^2} + \frac{1}{\Delta_J} \sum_{j : \lambda_j < \infty} \sqrt{\tau_j}\right)}\right] \\
&\overset{(b)}{\leq} O_\eta(1) \cdot \left(\sqrt{\mathbb{E}\left[|A_2| \sum_{j : \lambda_j = \infty} \frac{1}{\Delta_J^2}\right]} + \mathbb{E}\left[\frac{|A_2|}{\Delta_J}\right] + \sum_{j : \Delta_j > 0} \mathbb{E}[\sqrt{\tau_j}]\right) \\
&\overset{(c)}{\leq} O_\eta(1) \sum_{j : \Delta_j > 0} \sqrt{1 + \frac{\log(b_j)}{\Delta_j^2}}\,,
\end{aligned}$$

where (a) follows from Lemma 5 and (b) since for all $x, y \geq 0$ it holds that $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \leq x + y$. To get (c) we bound the first term as in Eq. (15), the second by the fact that arms in $j \in A_2$ have $\Delta_j \leq 2\Delta_J$ and the third using Lemma 8. Finally by substituting this into Eq. (13) we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in A_3} \Delta_i \tau_{\Delta_i}\right] &\leq O_\eta(1) \cdot \left(\mathbb{E}\left[\sum_{j \in A_1} J\sqrt{\tau_j} + \frac{J^2}{\Delta_{\min}}\right] + \sum_{j : \Delta_j > 0} \left(1 + \frac{\log(b_j)}{\Delta_j^2}\right)\right) \\
&\leq O_\eta(1) \sum_{j : \Delta_j > 0} \left(1 + \frac{\log(b_j)}{\Delta_j^2}\right)\,,
\end{aligned}$$

where the last line follows since $\mathbb{E}[J^2/\Delta_{\min}] = O_\eta(1)\Delta_{\min}^{-1}$ and by Lemma 8

$$\mathbb{E}[J\sqrt{\tau_j}] \leq \sqrt{\mathbb{E}[J^2]\mathbb{E}[\tau_j]} = O_\eta(1) \cdot \sqrt{1 + \frac{\log(b_j)}{\Delta_j^2}}\,,$$

which completes the proof.

H Technical Lemmas

Lemma 12. *Let $\eta > 1$ and $\rho \in [0, 1]$ and $\lambda \in (0, \infty]$ and $x > 0$, then*

$$\begin{aligned}
\sum_{k=0}^{\infty} \min\left\{\eta^{k+1},\, \lambda^\rho \eta^{(1-\rho)(k+1)}\right\} \exp\left(-x\eta^k\right) &\leq \begin{cases} \frac{1}{x}\left(\frac{2}{e} + \frac{\eta}{\log(\eta)}\right) & \textit{if } x\lambda \geq 1 \\ \frac{\lambda^\rho x^{\rho-1}}{\log(\eta)}\left(1 + \frac{1}{e} + \log\left(\frac{1}{\lambda x}\right)\right) & \textit{otherwise}\,. \end{cases} \\
&= O\left(\frac{\eta}{\eta - 1}\right) \cdot \min\left\{\frac{1}{x},\, \lambda^\rho x^{\rho-1} \log_+\left(\frac{1}{\lambda x}\right)\right\}\,.
\end{aligned}$$

Proof. Let $f(k) = \min\left\{\eta^{k+1},\, \lambda^\rho \eta^{(k+1)(1-\rho)}\right\}\exp(-x\eta^k)$, which is unimodal and so $\sum_{k=0}^\infty f(k) \leq 2\sup_k f(k) + \int_0^\infty f(k)dk$. If $x\lambda \geq 1$, then

$$\int_0^\infty f(k)dk \leq \eta \int_0^\infty \eta^k \exp\left(-\eta^k x\right) dk = \frac{\eta}{x\log(\eta)}\,.$$

If $x\lambda < 1$, then let k_λ be such that $\eta^{k_\lambda} = \lambda^\rho \eta^{k_\lambda(1-\rho)}$ and k_x be such that $\eta^k = 1/x$.

$$\begin{aligned}
\int_0^\infty f(k)dk &\leq \eta \int_0^{k_\lambda} \eta^k dk + \eta \int_{k_\lambda}^{k_x} \lambda^\rho \eta^{k_x(1-\rho)} dk + \eta \int_{k_x}^\infty \lambda^\rho \eta^{k(1-\rho)} \exp\left(-x\eta^k\right) dk \\
&= \frac{\lambda - 1}{\log(\eta)} + \eta\left(k_x - k_\lambda\right)\lambda^\rho x^{\rho-1} + \eta\lambda^\rho x^{\rho-1} \int_{k_x}^\infty \eta^{(k-k_x)(1-\rho)} \exp\left(-\eta^{k_x-k}\right) dk \\
&\leq \frac{\lambda-1}{\log(\eta)} + \frac{\eta\lambda^\rho x^{\rho-1}\log\left(\frac{1}{\lambda x}\right)}{\log(\eta)} + \frac{\eta\lambda^\rho x^{\rho-1}}{e\log(\eta)}
\end{aligned}$$

Finally

$$\sup_k f(k) \leq \min\left\{\frac{1}{ex},\, \eta\lambda^\rho x^{\rho-1}\right\}\,.$$

Therefore

$$\sum_{k=0}^\infty \min\left\{\eta^{k+1},\, \lambda^\rho \eta^{(1-\rho)(k+1)}\right\}\exp\left(-x\eta^k\right) \leq \begin{cases} \frac{1}{x}\left(\frac{2}{e} + \frac{\eta}{\log(\eta)}\right) & \text{if } x\lambda \geq 1 \\ \frac{\lambda^\rho x^{\rho-1}}{\log(\eta)}\left(1 + \frac{1}{e} + \log\left(\frac{1}{\lambda x}\right)\right) & \text{otherwise}\,. \end{cases}$$

□

I Table of Notation

K	number of arms
n	horizon
t	current time step
ξ_t	noise in time step t
η	constant parameter greater than 1 determining width of confidence interval
ρ	constant parameter in $[1/2, 1]$
η_1, η_2	$\eta_1 = (1+\eta)/2$ and $\eta_2 = \eta/\eta_1$
μ_i	expected return of arm i
$\hat{\mu}_{i,s}$	empirical estimate of return of arm i based on s samples
$\hat{\mu}_i(t)$	empirical estimate of return of arm i after time step t
Δ_i	gap between the expected returns of the best arm and the ith arm
$\Delta_{\min}$	minimal non-zero gap $\Delta_{\min} = \min\{\Delta_i : \Delta_i > 0\}$
$\Delta_{\max}$	maximum gap $\Delta_{\max} = \max_i \Delta_i$
$\log_+(x)$	$\max\{1, \log(x)\}$
B_i and b_i	see Eq. (5)
$k_{i,\rho}$	$\sum_{j=1}^{K} \min\{1, \Delta_i^{2\rho}/\Delta_j^{2\rho}\}$
$K_{i,\rho}$	see Eq. (4)
τ_i	see Eq. (5)
τ_Δ	see Eq. (6)
$p(\eta)$	see Lemma 3
Φ	set of optimistic arms Eq. (3)
J	$J = \min \Phi$