Considering Heterogeneity and Lack of Distributional Side-Information in Finite-Arm Bandits

Alejandro Gomez-Leos

Dec 2023

Abstract

Motivated by a type of k-armed bandit problem with heterogeneous and unknown reward distributions, we review the fundamental works in the domain of variance-aware algorithms—those that use sample variances in computation of arm scores. We revisit works deemed capable of providing insight to the setting of motivation, particularly Auer et al. (2002), Audibert et al. (2009), and, more recently, Cowan et al. (2017). Along the way, we discover a tighter regret analysis of Auer et al.'s `UCB1-NORMAL`, eventually removing an unnecessarily large constant in front of the logarithmic term.

1 Introduction

In a typical formulation of the k-armed bandit problem, an agent is faced with a *horizon* of n rounds, and is required to decide upon one of k alternative *arms* each round $t \in [n]$. The agent strategically chooses via a *policy* π. Upon picking an arm $i \in [k]$, the agent receives a sample from a distribution P_i with mean μ_i. Denoting $\mu^* := \max_i \mu_i$ as the reward mean of an *optimal* arm, the agent subsequently suffers a per-round expected *regret* $\mu^* - \mu_i$. The performance of a policy π is summarized via the cumulative expected regret $R_n^\pi(\nu) := \mathbb{E}(\sum_{t=1}^n \mu^* - \mu_{a_t})$, where the expectation is over randomness in the policy as well as the *environment* ν, i.e. the particular set of distributions $\{P_i\}_{i\in[k]}$.

Although the P_i's often have the same distributional characteristics, e.g. same sub-Gaussian parameter, there are interesting settings in which they need not. In fact, a major innovation in *multi-fidelity* bandits (e.g. see [5]) is the algorithmic understanding of how one should optimally balance the trade-off between quality of samples, as well as the budget constrained by the costs of different modes of measurement, called *fidelities*. Here one can view the quality of samples—given by the particular fidelity used to collect them—as conceptually analogous to a particular scenario in which P_i's exhibit heterogeneous distributional characteristics. From here, we delineate our motivating model.

1.1 Motivation Despite diverging from the canonical *multi-fidelity* bandit, we imagine a system of $k = LH$ arms, where each arm corresponds to one of L *actions*, and one of H *channels*[1]. Upon playing the arm indexed by $(\ell, h) \in [L] \times [H]$, the agent receives a reward sampled from a mean μ_ℓ Gaussian with variance σ_H^2. So far, this is an alternative view of a (cost-less) multi-fidelity bandit. However, the analogy breaks down once we assume the variances—analogous to the fidelities—are unknown beforehand. In this case, the agent is incentivized to not only find $\arg\max_\ell \mu_\ell$, but also identify $\arg\min_h \sigma_H^2$, for anything else will lead to relatively unreliable mean reward estimation. Viewing this model from yet another angle, one can recognize it as a system characterized by unknown zero-mean distributions $(P_h)_{h\in[H]}$. Here, it seems that optimal algorithm design could subtly depend on the nature of heterogeneity and prior side-information on pertinent noise parameters. In the end, we wish to obtain lower and upper-bounds on primarily the following settings:

- $P_h \equiv \mathcal{N}(0, \beta_h)$ for all h, $(\beta_1, \dots, \beta_H) \in \mathbb{R}_+^H$ unknown.
- P_h has bounded support $[-b_h, b_h]$ for all h, $(b_1, \dots, b_H) \in \mathbb{R}_+^H$ unknown.

Fortunately, a good deal of machinery has been developed since Auer et al. (2002)—arguably the foundational paper for modern study of bandit algorithms. In this report, we review a small selection of works for potentially useful ideas towards our motivating problem. After returning to our initial k-armed bandit formulation, we describe the organization of this survey.

1.2 Formulation Recalling the notation above, we assume by default $k \in \mathbb{N}$ and $k \geq 2$. To each arm $i \in [k]$, we correspond the reward distribution P_i with unknown mean μ_i and unknown variance $\sigma_i^2 < \infty$. Let $\mu^* = \max_i \mu_i$, and $\Delta_i = \mu^* - \mu_i$ for each $i \in [k]$. Without loss of generality, assume $\mu_1 = \mu^*$, i.e. arm 1 is an optimal arm. For each round $t = 1, 2, \dots, n$ under policy π, let a_t^π denote the arm chosen by the policy π, and $X_t \sim P_{a_t^\pi}$ the random reward received. Under the same policy π, let $T_i^\pi(t) := \sum_{s=1}^t \mathbb{I}\{a_s^\pi = i\}$ denote the number of times arm $i \in [k]$ is played by and including round t, with $T_i^\pi(0) := 0$. Wherever clear from context, we'll drop the superscript π. We'll denote $\bar{\mu}_{i,t} := (1/T_i(t)) \sum_{s=1}^t X_s \cdot \mathbb{I}\{a_s = i\}$ as the sample mean available at the end of round t, constructed with all samples collected by the agent from arm i. For $a, b \in \mathbb{R}$, we denote $(a \vee b)$ as their maximum, and $(a \wedge b)$ as their minimum.

1.3 Organization We organize this survey into Section (2) for unbounded rewards, and Section (3) for bounded rewards. In Section 2 we revisit Auer et al.'s `UCB1-NORMAL`, which is less well-known, perhaps, due to its reliance on two (now one) unproven conjectures. We provide a complete, alternative[2] proof to the regret bound in the original paper, thereby removing a large constant scaling for the dominant logarithmic term. Following, we describe a more recent algorithm

[1] As a practical application, consider the setting in which an experimenter has access to a selection of H noisy probes, and seeks to efficiently test a black-box system at L select locations.

[2] Assuming the same conditions.

ISM-NORMAL2: an updated version of its ancestor ISM-NORMAL1 (c.f. [1]), which similarly required an unproven conjecture. However, as the recent work [6] indicates, not only is the conjecture false, but also a small modification is a sufficient condition for asymptotic optimality. In Section 3, we slightly shift focus to the case of bounded rewards, revisiting the celebrated work of Audibert et al., along with its variance-aware algorithm UCB-V. In Section 4, we finish with a discussion for potential future work.

2 Unbounded Rewards

2.1 UCB1-NORMAL We consider the setting in which $P_i \equiv \mathcal{N}(\mu_i, \sigma_i^2)$ for all $i \in [k]$, and σ_i^2 is unknown. Seemingly, the earliest study of this problem was initiated by the (original UCB1) authors in [2]. As a slight modification of the UCB1 algorithm, one adjusts the upper confidence bound with the respective plug-in estimator of the variance. For the arm i, let

$$S_{i,t}^2 := \frac{1}{T_i(t)-1} \sum_{s=1}^{t} (X_s - \bar{\mu}_{i,t})^2 \cdot \mathbb{I}\{a_s = i\} \tag{1}$$

denote its unbiased reward sample variance[3], which is available to the agent at the end of round t. In this work, the authors introduce the algorithm UCB1-NORMAL 1 with score

$$U_i(t) := \bar{\mu}_{i,t} + \sqrt{S_{i,t}^2 \frac{16 \ln(t)}{T_i(t)}}$$

defined for $T_i(t-1) > 0$. The algorithm is as follows.

Algorithm 1 UCB1-NORMAL

for each round $t = 1, 2, \ldots, n$ **do**
  If there is an arm i with $\leq \lceil 8 \log t \rceil$ plays, play arm i
  Otherwise, play arm $j = \arg\max_j U_j(t-1)$
  Update estimates and UCB scores
end for

Unfortunately, to make any regret guarantees on UCB1-NORMAL, the authors required some conjectures on tail bounds for Student and χ^2 r.v.'s, which they were only able to verify numerically.

Conjecture 1. *[2, 4] Let X be a Student r.v. with s degrees of freedom. Then, for all $a \in [0, \sqrt{2(s+1)}]$, $P(X \geq a) \leq e^{-a^2/4}$.*

Conjecture 2. *[2] Let X be a χ^2 r.v. with s degrees of freedom. Then, $P(X \geq 4s) \leq e^{-(s+1)/2}$.*

Conjecture 1 was actually proven in 2015 by [4], whereas Conjecture 2 is seemingly open. Assuming these conjectures, the authors proved an upper bound on the regret.

[3] Computed in the usual way, with Bessel's correction.

Theorem 1. *[2] Consider a k-armed bandit on (unknown) Gaussian reward distributions $\nu = (P_1 \dots P_k)$. For policy π described by* `UCB1-NORMAL`*, i.e. algorithm 1, the expected regret satisfies*

$$R_n^\pi(\nu) \leq \sum_{i:\Delta_i>0} \Delta_i \left(1 + \frac{\pi^2}{2} + 8\ln(n)\right) + \frac{256\,\sigma_i^2 \ln(n)}{\Delta_i} \tag{2}$$

if Conjecture 2 is true.

Originally, we wished to sketch the proof of this regret bound, but we realized we could perform a more modern analysis of this algorithm, drastically shaving down the gap and variance dependent logarithmic term (assuming the same conjectures). We could not find in the literature our upper bound on the regret of `UCB1-NORMAL`.

Theorem 2. *[This work] In the same setting as above,*

$$R_n^\pi(\nu) \leq \sum_{i:\Delta_i>0} \Delta_i \left(4 + 8\ln(n)\right) + \frac{8\,\sigma_i^2 \ln(n)}{\Delta_i} \tag{3}$$

if Conjecture 2 is true.

Proof. Let $\delta \in (0,1)$ to be decided upon later. Fix any $j \in [k]$ such that $\Delta_j > 0$. Denote the auxiliary quantity $U_i(t-1,\delta) := \bar{\mu}_{i,t-1} + \sqrt{S^2_{i,t-1}\frac{16\ln(1/\delta)}{T_i(t-1)}}$, where $U_i(0,\delta)$ is defined to be $+\infty$ for all $i \in [k]$. For j, consider the "good" event

$$G_j := \left\{\mu_1 \leq \min_{k+1\leq s\leq n} U_1(s,\delta)\right\} \bigcap \left\{\bar{\mu}_{j,u_j} + \sqrt{S^2_{j,u_j}\frac{16\ln(1/\delta)}{u_j}} < \mu_1\right\}$$

where $\bar{\mu}_{j,u_j}$ and S^2_{j,u_j} denote the sample mean and unbiased sample variance constructed with u_j IID samples from distribution P_j. The first claim is that if G_j occurs, then $T_j(n) \leq u_j$, for any $u_j > k$.

Supposing that G_j holds, but $T_j(n) > u_j$, then there exists a $t \leq n$ such that $a_t = j$ and $T_j(t-1) = u_j$. Since $k < u_j = T_j(t-1) \leq t-1$, it must be that $t-1 \geq k+1$. Thus, $U_j(t-1,\delta) = \bar{\mu}_{j,u_j} + \sqrt{S^2_{j,u_j}\frac{16\ln(1/\delta)}{u_j}} < \mu_1 \leq U_1(t-1,\delta)$, a contradiction, as the algorithm prefers arms with higher score. Applying this, it follows that for any integer $\ell \geq 1$, we have

$$\begin{aligned}
\mathbb{E}(T_j(n)) &\leq \mathbb{E}\big(1 + \sum_{t=k+1}^{n} \mathbb{I}\{a_t = j\}\big) \\
&\leq \ell + \mathbb{E}\big(\sum_{t=k+1}^{n} \mathbb{I}\{a_t = j, T_j(t-1) \geq \ell\}\big) \\
&= \ell + \mathbb{E}\bigg(\mathbb{I}\{G_j\} \sum_{t=k+1}^{n} \mathbb{I}\{a_t = j, T_j(t-1) \geq \ell\}\bigg) + \mathbb{E}\bigg(\mathbb{I}\{G_j^C\} \sum_{t=k+1}^{n} \mathbb{I}\{a_t = j, T_j(t-1) \geq \ell\}\bigg) \\
&\leq \ell + u_j + \mathbb{E}\bigg(\mathbb{I}\{G_j^C\} \sum_{t=k+1}^{n} \mathbb{I}\{a_t = j, T_j(t-1) \geq \ell\}\bigg) \\
&\leq \ell + u_j + n \cdot P(G_j^C \cap A_j)
\end{aligned}$$

where $A_j := \cap_{t=k+1}^{n} \{T_j(t-1) \geq \ell\}$[4]. We now bound the probability of $G_j^C \cap A_j$. By a union bound and "conversion of time to samples,"

$$\begin{aligned}
P\bigg(\mu_1 \geq \min_{k+1 \leq s \leq n} U_1(s, \delta),\, A_j\bigg) &\leq \sum_{r=k+1}^{n} P\bigg(\mu_1 \geq \bar{\mu}_{1,r} + \sqrt{S_{1,r}^2 \frac{16 \ln(1/\delta)}{r}},\, r \geq \ell\bigg) \\
&= \sum_{r=k+1}^{n} P\bigg(\frac{\mu_1 - \bar{\mu}_{1,r}}{\sqrt{\frac{S_{1,r}^2}{r}}} \geq 4\sqrt{\ln(1/\delta)},\, r \geq \ell\bigg) \\
&:= \sum_{r=k+1}^{n} P\bigg(Z_r \geq 4\sqrt{\ln(1/\delta)},\, r \geq \ell\bigg)
\end{aligned}$$

Each of the Z_r is a Student r.v. with $r-1$ degrees of freedom (c.f. [2]). For $\ell \geq 8\ln(1/\delta)$, we have $r \geq 8\ln(1/\delta)$ for all r under the summand—thus, we could apply conjecture 1 for $s = r-1$ and $a = 4\sqrt{\ln(1/\delta)}$, since $4\sqrt{\ln(1/\delta)} \leq \sqrt{2r}$. Under such an ℓ, by conjecture 1, we have that the RHS is at most $\sum_{r=k+1}^{n} \delta^4 \leq n\delta^4$. Moreover,

$$\begin{aligned}
P\bigg(\bar{\mu}_{j,u_j} + \sqrt{S_{j,u_j}^2 \frac{16\ln(1/\delta)}{u_j}} > \mu_1,\, A_j\bigg) &\leq P\bigg(\bar{\mu}_{j,u_j} + \sqrt{S_{j,u_j}^2 \frac{16\ln(1/\delta)}{u_j}} > \mu_1\bigg) \\
&= P\bigg(\sqrt{S_{j,u_j}^2 \frac{16\ln(1/\delta)}{u_j}} > \Delta_j + \mu_j - \bar{\mu}_{j,u_j}\bigg)
\end{aligned}$$

[4]Note that $T_j(\tau) \geq \ell \implies T_j(\tau+1) \geq \ell$.

Let F_j denote the event on the RHS. Denoting $B_j := \{\mu_j - \bar{\mu}_{j,u_j} \leq \Delta_j\}$, we have that

$$\begin{aligned} P(F_j) &\leq P(F_j \cap B_j) + e^{-\frac{u_j \Delta_j^2}{2\sigma_j^2}} \\ &\leq P\left(\sqrt{S_{j,u_j}^2 \frac{16\ln(1/\delta)}{u_j}} > 2\Delta_j\right) + e^{-\frac{u_j \Delta_j^2}{2\sigma_j^2}} \\ &= P\left((u_j - 1)\frac{S_{j,u_j}^2}{\sigma_j^2} > (u_j - 1)\frac{u_j\,\Delta_j^2}{4\,\sigma_j^2 \ln(1/\delta)}\right) + e^{-\frac{u_j \Delta_j^2}{2\sigma_j^2}} \end{aligned}$$

For $u_j \geq \left((16\,\sigma_j^2/\Delta_j^2) \vee 8\right)\,\ln(1/\delta)$, we can bound the RHS as

$$P\left((u_j - 1)\frac{S_{j,u_j}^2}{\sigma_j^2} > 4(u_j - 1)\right) + \delta^8$$

Noting the r.v. under $P(\cdot)$ is χ^2 with $u_j - 1$ degrees of freedom (c.f. [2]), an application of conjecture 2 yields that the RHS $\leq e^{-u_j/2} + \delta^8 \leq \delta^4 + \delta^8 \leq 2\delta^4$ under this regime for u_j. Thus, for $\ell \geq 8\ln(1/\delta)$ and $u_j \geq \left((16\,\sigma_j^2/\Delta_j^2) \vee 8\right)\,\ln(1/\delta)$, we have that $P(G_j^C \cap A_j) \leq (n+2)\delta^4$. Taking $\ell = \lceil 8\ln(1/\delta)\rceil$ and $u_j = \lceil\left((16\,\sigma_j^2/\Delta_j^2) \vee 8\right)\,\ln(1/\delta)\rceil$, and setting $\delta = 1/\sqrt{n}$ we have that

$$\mathbb{E}(T_j(n)) \leq \ell + u_j + n(n+2)\delta^4 \leq \max\left(\frac{8\,\sigma_j^2}{\Delta_j^2}, 8\right)\ln(n) + 4\ln(n) + 4 \leq \frac{8\,\sigma_j^2}{\Delta_j^2}\ln(n) + 8\ln(n) + 4$$

The final bound is immediately from bounding $R_n^\pi(\nu) = \sum_{j:\Delta_j > 0} \Delta_j \mathbb{E}(T_j(n))$ with the above. ■

In contrast, the original bound[5] for `UCB1` (known variances case) given by the authors is

$$R_n^\pi(\nu) \leq \sum_{i:\Delta_i > 0} \Delta_i\left(1 + \frac{\pi^2}{2}\right) + \frac{8\sigma_i^2\ln(n)}{\Delta_i} \tag{4}$$

Comparing equations (4) and (3), it seems that the lack of variance knowledge beforehand imposes a per-arm cost on the regret of order $\Delta_i \ln(n)$. Recalling that this additive term is required for the conjectured *Student* and χ^2 concentration, it raises the question whether this logarithmic term is necessary. Unfortunately, this work does not offer a discussion on the regret lower-bound of this problem, to which we turn to discussion in [6].

2.2 `ISM-NORMAL2` Usually, one can construct a lower-bound for a class of well-behaved policies termed α-*consistent* policies[6] A naturally stronger notion (c.f. [1]) is that of a *strongly consistent* policy—one that is α-consistent for all $\alpha > 0$. Specifically,

Definition 1. *A policy π is said to be α-consistent on the set of environments $\mathcal{E}$ if for any $\nu \in \mathcal{E}$ and any $\alpha > 0$, there exists a $\tau(\alpha, \nu)$ and constant $C > 0$ such that $R_n^\pi(\nu) < Cn^\alpha$ for all $n \geq \tau(\alpha, \nu)$.*

[5] Having taken the liberty of re-scaling the variance dependent term, when variances are known.
[6] Recall that such a policy suffers at most $o(n^\alpha)$ regret on all environments, for a fixed α.

For example, the upper bounds above demonstrate that UCB1-NORMAL is strongly consistent for the heterogeneous Gaussian reward setting, even when variances are unknown. For such policies, we paraphrase the following result in [1], which is a special case of Theorem 1 of the same paper (c.f. [6]).

Theorem 3. *[1] Let $\mathcal{E}$ denote the set of all environments ν, where each ν corresponds to some ordered tuple $\big((\mu_i, \sigma_i^2)\big)_{i=1}^k$. If π is strongly consistent for $\mathcal{E}$, then*

$$\liminf_{n\to\infty} \frac{R_n^\pi(\nu)}{\ln(n)} \geq \sum_{i:\Delta_i>0} \frac{2\Delta_i}{\ln(1+\frac{\Delta_i^2}{\sigma_i^2})}$$

To give a comparison with equation (3), see that UCB1-NORMAL plays each sub-optimal arm no more than $(8\Delta_i + \frac{8\sigma_i^2}{\Delta_i})\ln(n)$ times as $n\to\infty$. As indicated above, any strongly consistent algorithm must play each sub-optimal arm i at least $\frac{2\Delta_i}{\ln(1+(\Delta_i^2/\sigma_i^2))}\ln(n)$ times, asymptotically. The ratio of the former with the latter is no more than $4\big(1+(\sigma_i^2/\Delta_i^2)\big)\ln(1+(\Delta_i^2/\sigma_i^2))$. Letting $\rho_i := \sigma_i^2/\Delta_i^2$, perhaps the most interesting setting is when $\rho_i \geq 1$ which is roughly the setting where mean rewards are "most easily confused" with the optimal arm means. One can verify that $4\big(1+\rho_i\big)\ln(1+\rho_i^{-1}) \leq 6$ for $\rho_i \geq 1$, suggesting that our analysis nearly matches the lower-bound constant. Without any further assumptions, however, the authors of [6] offer an algorithm that achieves the lower-bound scaling for all strongly consistent policies—that is, asymptotically optimal over the set of all strongly consistent policies. They call their algorithm[7] ISM-NORMAL2, which utilizes the following score (recalling (1))

$$I_i(t) = \bar{\mu}_{i,t} + \sqrt{S_{i,t}^2\left(t^{\frac{2}{T_i(t)-2}} - 1\right)} \tag{5}$$

This admits the following regret upper bound.

Algorithm 2 ISM-NORMAL2

Play each arm three times
for each round $t = 3k+1, 2, \ldots, n$ **do**
  Play arm $j = \arg\max_j I_j(t-1)$
  Update estimates and score
end for

Theorem 4. *[6] Consider a k-armed bandit on (unknown) Gaussian reward distributions $\nu = (P_1 \ldots P_k)$. For policy π described by ISM-NORMAL2, i.e. algorithm 2, for all $\epsilon \in (0,1)$, the expected regret satisfies*

$$R_n^\pi(\nu) \leq \sum_{i:\Delta_i>0} \Delta_i \left[\frac{8}{\epsilon^2} + 3 + \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\epsilon)^2}{1+\epsilon}\right)}\right] + \sqrt{\frac{\pi}{2e}}\frac{8\min_j \sigma_j^3 \ln\ln(n)}{\Delta_i^3\epsilon^3} + \frac{8\sigma_i^2}{\Delta_i^2\epsilon^2} \tag{6}$$

[7] ISM stands for "inflated sample mean."

The proof follows the classical techniques in [2], primarily innovating through the following un-intuitive[8] lemma, which provides a scaling law for the probability that a χ^2 random variable exceeds a χ^2_d random variable above a certain threshold.

Lemma 1. *[6] Let Z and U be independent random variables, $Z \sim \mathcal{N}(0,1)$, and $U \sim \chi^2_d$, where $d \geq 2$. For $\delta > 0, p > 0$ the following holds for all $k \geq 1$:*

$$\frac{1}{2}P\big(\frac{1}{4}Z^2 \geq U \geq \delta^2\big)\, k^{-d/p} \leq P\big(\delta + \sqrt{U}\sqrt{k^{2/p}-1} < Z\big) \leq \frac{e^{-(1+\delta^2)/2}p}{2\delta^2\sqrt{d}}\frac{k^{(1-d)/p}}{\ln k} \tag{7}$$

The authors also make use of a useful, but less well-known Chernoff bound.

Lemma 2. *[6] For $Z \sim \chi^2_d$, $P(Z > d(1+\epsilon)) \leq \big(e^{-\epsilon}(1+\epsilon)\big)^{d/2}$.*

As an outline of the proof of Theorem 6, the authors follow the standard technique of bounding frequency of sub-optimal events. Defining

$$I_j(m,u) = \bar{\mu}_{j,u} + \sqrt{S^2_{j,u}\left(m^{\frac{2}{u-2}} - 1\right)} \tag{8}$$

The following events for any sub-optimal arm j are used to construct the bad events.

- $E_{j,1}(t-1) := \{I_j\big(t-1, T_j(t-1)\big) \geq \mu^* - \epsilon\frac{\Delta_i}{2}\}$ j^{th} score nearly exceeds μ^*
- $E_{j,2}(t-1) := \{\bar{\mu}_{j,T_j(t-1)} \leq \mu_j + \epsilon\frac{\Delta_i}{2}\}$ j^{th} sample mean is nearly below the true mean
- $E_{j,3}(t-1) := \{S^2_{j,T_j(t-1)} \leq \sigma^2_j(1+\epsilon)\}$ j^{th} sample var. is nearly below the true var.

Going on, the authors propose to control the following counts,

$$\begin{aligned}
n^1_j &= \sum_{t=3k+1}^{n} \mathbb{I}\{a_t = j \cap E_{j,1}(t-1) \cap E_{j,2}(t-1) \cap E_{j,3}(t-1)\} \\
n^2_j &= \sum_{t=3k+1}^{n} \mathbb{I}\{a_t = j \cap E_{j,1}(t-1) \cap E_{j,2}(t-1) \cap E_{j,3}(t-1)^C\} \quad \leq \sum_{t=3k+1}^{n} \mathbb{I}\{E_{j,3}(t-1)^C\} \\
n^3_j &= \sum_{t=3k+1}^{n} \mathbb{I}\{a_t = j \cap E_{j,1}(t-1) \cap E_{j,2}(t-1)^C\} \quad \leq \sum_{t=3k+1}^{n} \mathbb{I}\{E_{j,2}(t-1)^C\} \\
n^4_j &= \sum_{t=3k+1}^{n} \mathbb{I}\{a_t = j \cap E_{j,1}(t-1)^C\}
\end{aligned}$$

since $T_j(n) = 3 + n^1_j + n^2_j + n^3_j + n^4_j$. Upon re-arranging terms, n^1_j can be upper-bounded a.s., as

$$n^1_j \leq \sum_{t=3k+1}^{n} \mathbb{I}\{a_t = j \cap T_j(t-1) = \mathcal{O}\big(\frac{\ln n}{\ln(1 + \frac{\Delta^2_j}{\sigma^2_j}o(\epsilon))}\big)\}$$

[8] Author's own words.

which supplies the correct asymptotic scaling with $\ln n$. The terms n_j^2, n_j^3 are not tedious to bound, thanks to the normality of the rewards—applying lemma 2 to bound n_j^2 and the usual tail bound to n_j^3. These two yield the ϵ and σ_j^2 dependent constant terms in the overall regret upper bound (6).

The ultimate term contributes the iterated logarithmic term, arising as follows. Let j^* denote any optimal mean arm with minimum variance. Since arm $j \neq j^*$ is picked in the indicator, n_j^4 is at most

$$\sum_{t=3k+1}^{n} \mathbb{I}\{I_{j^*}\big(t-1, T_{j^*}(t-1)\big) \leq \mu^* - \epsilon\frac{\Delta_j}{2}\} \leq \sum_{t=3k+1}^{n} \mathbb{I}\{ \cup_{s=3}^{t-1} \{I_{j^*}(t-1,s) \leq \mu^* - \epsilon\frac{\Delta_j}{2}\}\}$$
$$:= \sum_{t=3k}^{n-1} \mathbb{I}\{A_{s,t}\}$$

For $Z \sim \mathcal{N}(0,1)$ and $U_{s-1} \sim \chi^2_{s-1}$ independent,

$$\begin{aligned} P(A_{s,t}) &= \left(\mu^* + Z\frac{\sigma_{j^*}}{\sqrt{s}} + \sigma_{j^*}\frac{\sqrt{U_{s-1}}}{\sqrt{s}}\sqrt{t^{\frac{2}{s-2}} - 1} < \mu^* - \epsilon\frac{\Delta_j}{2}\right) \\ &= P\left(Z + \sqrt{U_{s-1}}\sqrt{t^{\frac{2}{s-2}} - 1} < -\epsilon\frac{\Delta_j}{2}\frac{\sqrt{s}}{\sigma_{j^*}}\right) \\ &= P\left(\epsilon\frac{\Delta_j}{2}\frac{\sqrt{s}}{\sigma_{j^*}} + \sqrt{U_{s-1}}\sqrt{t^{\frac{2}{s-2}} - 1} < Z\right) \end{aligned}$$

Note crucially the exchange of estimates for Z and U_{s-1} depends on normality, for the usual sample mean and unbiased sample variance are independent only for Gaussian r.v.'s. Also, note the symmetry of Gaussians was used. At this point one can apply lemma 1, and the remainder of the proof is balancing terms.

3 Bounded Rewards

3.1 `UCB-V` Next we consider the generally stated sample-variance tuned algorithm of [3]. To preface, Auer et al. also studied the scenario where the reward distributions are limited to a bounded interval $[0, b]$ a.s., and that the agent has knowledge of the interval. Here, Auer et al.'s analysis of `UCB1` yields that the expected regret satisfies

$$R_n^{\pi}(\nu) \leq 8\Big(\sum_{j:\Delta_j>0} \frac{b^2}{\Delta_j}\Big) \ln(n) + \mathcal{O}(1) \tag{9}$$

If we interpret b^2 as the sub-Gaussian parameter of the reward distributions, then given the discussion in the previous section, one can see this is not really improvable under Gaussian rewards, at least for strongly consistent policies (c.f. Theorem 3). However, in practice, it may be that b is rather a conservative uninformed guess. For instance, there could be an optimal mean arm whose reward distribution lies within a sub-interval of $[0, b]$ of length $b_0 < b$. Clearly then, sequential esti-

mates of each arm's variance should be used to tune UCB1's score. In fact, Auer et al. was cognizant of this, providing the algorithm UCB-TUNED in their experimentation section, but could not prove a regret bound. This inspired the study of a parameterized variance-aware algorithm UCB-V in [3], which includes UCB-TUNED as a special case. To describe UCB-V, let $E_1, E_2, \dots$ be non-negative and non-increasing sequence, termed the *exploration function*. With a slight inconsistency with (1) in the previous section, re-define

$$S_{i,t}^2 := \frac{1}{T_i(t)} \sum_{s=1}^{t} (X_s - \bar{\mu}_{i,t})^2 \cdot \mathbb{I}\{a_s = i\} \tag{10}$$

as the sample variance with $T_i(t)$ samples[9]. Consider the arm score,

$$B_i(t) := \bar{\mu}_{i,t} + \sqrt{S_{i,t}^2 \frac{E_t}{T_i(t)}} + c\frac{3b\,E_t}{T_i(t)}$$

with the convention that $1/0 = +\infty$. Notably, this form of bonus arises from the now well-known *empirical Bernstein bound*, introduced in the same paper.

Lemma 3. *[3] Let $X_1 \dots X_t$ be IID r.v.'s with support in $[0, b]$. Let $\bar{X}_t$ be their empirical mean and $V_t = \frac{t-1}{t} S_t$ where S_t^2 is their unbiased sample variance. Then, for any $t \in \mathbb{N}_{\geq 0}$ and $x > 0$, with probability $\geq 1 - 3e^{-x}$,*

$$|\bar{X}_t - \mu| \leq \sqrt{\frac{2V_t x}{t}} + \frac{3bx}{t}$$

Moreover, for $\beta(x,t) := 3 \inf_{\alpha \in (1,3]} \left(\frac{\ln t}{\ln \alpha} \wedge t\right) e^{-x/\alpha}$, with probability $\geq 1 - \beta(x,t)$, holding simultaneously for all $s \in \{1, \dots, t\}$,

$$|\bar{X}_s - \mu| \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s}$$

Algorithm 3 UCB-V

Input: $c \geq 0$, $(E_t)_{t \geq 0}$, $b > 0$
for each round $t = 1, 2, \dots, n$ **do**
  Play arm $j = \arg\max_j B_j(t-1)$
  Update estimates and arm score.
end for

UCB-V enjoys the following regret guarantee.

Theorem 5. *[3] Consider a k-armed bandit on reward distributions $\nu = (P_1 \dots P_k)$, where P_i is bounded almost surely in $[0, b]$ for all $i \in [k]$. For policy π described by UCB-V, i.e. algorithm 3, the expected regret satisfies*

$$R_n^\pi(\nu) \leq \sum_{i:\Delta_i > 0} \Delta_i \left(1 + 8(c \vee 1)\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2b}{\Delta_i}\right) E_n + n e^{-E_n}\left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4b}{\Delta_i}\right) + \sum_{t=\lfloor 16 E_n \rfloor}^{n} \beta((c \vee 1)E_t, t)\right) \tag{11}$$

[9] We believe the authors forwent the unbiased estimator for brevity of notation.

One nicety of this regret bound is that it is particularly instructive as to the choice of $E_t = \ln(t)$, for otherwise a too large of an exploration function leads to polynomial regret. Unfortunately, a negative result provided in this work explains that b is un-removable from the bound, over the set of consistent policies.

4 Future Work

Re-examining our motivating problem in the case of Gaussian noise, it seems that one straightforward procedure could be the usage of `UCB1-NORMAL` on all $k = LH$ arms. However, this is likely wasteful for the following reason. Letting $a_{\ell,h}$ denote the arm corresponding to action ℓ and channel h, the reward distribution of this arm is $\mathcal{N}(\mu_\ell, \sigma_h^2)$. The sample variance of the arm $a_{\ell,h}$ is independent of the sample mean, by normality. Thus, a more accurate sample variance estimator can be constructed by aggregating over all action samples $\ell \in [L]$ for a fixed channel h. Of course, the number of samples used for each action should determine the averaging scheme. Although this leads to a modified `UCB1-NORMAL` algorithm for the action selection, it is unclear what kind of algorithm should be used for the channel selection.

In any case, the next step would be to extend such an algorithm to the case that each channel is non-Gaussian, presenting the challenge of circumventing the usage of independent sample mean and variance, as well as distributional symmetry. But this is more of a general challenge, for we saw how crucial this was for Theorems 2 and 4, primarily through conjectures 1, 2, and lemma 1, respectively. Indeed, for bounded rewards, one can say more via empirical Bernstein bounds. However, one may not want to assume, for their system model, precisely that boundedness. Beyond these considerations, and beyond our motivation, more general curiosities include the natural extensions to structured bandit scenarios, e.g. linear bandits, contextual bandits, etc.

References

[1] Burnetas, Apostolos N., and Michael N. Katehakis. "Optimal adaptive policies for sequential allocation problems." Advances in Applied Mathematics 17.2 (1996): 122-142.

[2] Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." Machine learning 47 (2002): 235-256.

[3] Audibert, Jean-Yves, Rémi Munos, and Csaba Szepesvári. "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits." Theoretical Computer Science 410.19 (2009): 1876-1902.

[4] Orabona, Francesco. "A Simple Expression for Mill's Ratio of the Student's t-Distribution." arXiv preprint arXiv:1502.01632 (2015).

[5] Kandasamy, Kirthevasan, et al. "The multi-fidelity multi-armed bandit." Advances in neural information processing systems 29 (2016).

[6] Cowan, Wesley, Junya Honda, and Michael N. Katehakis. "Normal bandits of unknown means and variances." The Journal of Machine Learning Research 18.1 (2017): 5638-5665.

[7] Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.