The Multi-fidelity Multi-armed Bandit

Kirthevasan Kandasamy [♮], **Gautam Dasarathy** [◇], **Jeff Schneider** [♮], **Barnabás Póczos** [♮]
[♮] Carnegie Mellon University, [◇] Rice University
{kandasamy, schneide, bapoczos}@cs.cmu.edu, gautamd@rice.edu

Abstract

We study a variant of the classical stochastic K-armed bandit where observing the outcome of each arm is expensive, but cheap approximations to this outcome are available. For example, in online advertising the performance of an ad can be approximated by displaying it for shorter time periods or to narrower audiences. We formalise this task as a *multi-fidelity* bandit, where, at each time step, the forecaster may choose to play an arm at any one of M fidelities. The highest fidelity (desired outcome) expends cost $\lambda^{(M)}$. The m^{th} fidelity (an approximation) expends $\lambda^{(m)} < \lambda^{(M)}$ and returns a biased estimate of the highest fidelity. We develop MF-UCB, a novel upper confidence bound procedure for this setting and prove that it naturally adapts to the sequence of available approximations and costs thus attaining better regret than naive strategies which ignore the approximations. For instance, in the above online advertising example, MF-UCB would use the lower fidelities to quickly eliminate suboptimal ads and reserve the larger expensive experiments on a small set of promising candidates. We complement this result with a lower bound and show that MF-UCB is nearly optimal under certain conditions.

1 Introduction

Since the seminal work of Robbins [11], the multi-armed bandit has become an attractive framework for studying exploration-exploitation trade-offs inherent to tasks arising in online advertising, finance and other fields. In the most basic form of the K-armed bandit [9, 12], we have a set $\mathcal{K} = \{1, \ldots, K\}$ of K arms (e.g. K ads in online advertising). At each time step $t = 1, 2, \ldots$, an arm is played and a corresponding reward is realised. The goal is to design a strategy of plays that minimises the *regret* after n plays. The regret is the comparison, in expectation, of the realised reward against an oracle that always plays the best arm. The well known Upper Confidence Bound (UCB) algorithm [3], achieves regret $\mathcal{O}(K \log(n))$ after n plays (ignoring mean rewards) and is minimax optimal [9].

In this paper, we propose a new take on this important problem. In many practical scenarios of interest, one can associate a cost to playing each arm. Furthermore, in many of these scenarios, one might have access to cheaper approximations to the outcome of the arms. For instance, in online advertising the goal is to maximise the cumulative number of clicks over a given time period. Conventionally, an arm pull maybe thought of as the display of an ad for a specific time, say one hour. However, we may approximate its hourly performance by displaying the ad for shorter periods. This estimate is *biased* (and possibly noisy), as displaying an ad for longer intervals changes user behaviour. It can nonetheless be useful in gauging the long run click through rate. We can also obtain biased estimates of an ad by displaying it only to certain geographic regions or age groups. Similarly one might consider *algorithm selection* for machine learning problems [4], where the goal is to be competitive with the best among a set of learning algorithms for a task. Here, one might obtain cheaper approximate estimates of the performance of algorithm by cheaper versions using less data or computation. In this paper, we will refer to such approximations as fidelities. Consider a 2-fidelity problem where the cost at the low fidelity is $\lambda^{(1)}$ and the cost at the high fidelity is $\lambda^{(2)}$. We will present a cost weighted notion of regret for this setting for a strategy that expends a capital

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

of Λ units. A classical K-armed bandit strategy such as UCB, which only uses the highest fidelity, can obtain at best $\mathcal{O}(\lambda^{(2)}K\log(\Lambda/\lambda^{(2)}))$ regret [9]. In contrast, this paper will present multi-fidelity strategies that achieve $\mathcal{O}\left((\lambda^{(1)}K + \lambda^{(2)}|\mathcal{K}_g|)\log(\Lambda/\lambda^{(2)})\right)$ regret. Here $\mathcal{K}_g$ is a (typically) small subset of arms with high expected reward that can be identified using plays at the (cheaper) low fidelity. When $|\mathcal{K}_g| < K$ and $\lambda^{(1)} < \lambda^{(2)}$, such a strategy will outperform the more standard UCB algorithms. Intuitively, this is achieved by using the lower fidelities to eliminate several of "bad" arms and reserving expensive higher fidelity plays for a small subset of the most promising arms. We formalise the above intuitions in the sequel. Our main contributions are,

1. A novel **formalism** for studying bandit tasks when one has access to multiple fidelities for each arm, with each successive fidelity providing a better approximation to the most expensive one.
2. A new **algorithm** that we call Multi-Fidelity Upper Confidence Bound (MF-UCB) that adapts the classical Upper Confidence Bound (UCB) strategies to our multi-fidelity setting. Empirically, we demonstrate that our algorithm outperforms naive UCB on simulations.
3. A **theoretical characterisation** of the performance of MF-UCB that shows that the algorithm (a) uses the lower fidelities to explore all arms and eliminates arms with low expected reward, and (b) reserves the higher fidelity plays for arms with rewards close to the optimal value. We derive a lower bound on the regret and demonstrate that MF-UCB is near-optimal on this problem.

Related Work

The K-armed bandit has been studied extensively in the past [1, 9, 11]. There has been a flurry of work on upper confidence bound (UCB) methods [2, 3], which adopt the optimism in the face of uncertainty principle for bandits. For readers unfamiliar with UCB methods, we recommend Chapter 2 of Bubeck and Cesa-Bianchi [5]. Our work in this paper builds on UCB ideas, but the multi-fidelity framework poses significantly new algorithmic and theoretical challenges.

There has been some interest in multi-fidelity methods for optimisation in many applied domains of research [7, 10]. However, these works do not formalise or analyse notions of *regret* in the multi-fidelity setting. Multi-fidelity methods are used in the robotics community for reinforcement learning tasks by modeling each fidelity as a Markov decision process [6]. Zhang and Chaudhuri [16] study active learning with a cheap weak labeler and an expensive strong labeler. The objective of these papers however is not to handle the exploration-exploitation trade-off inherent to the bandit setting. A line of work on budgeted multi-armed bandits [13, 15] study a variant of the K-armed bandit where each arm has a random reward and cost and the goal is to play the arm with the highest reward/cost ratio as much as possible. This is different from our setting where each arm has multiple fidelities which serve as an approximation. Recently, in Kandasamy et al. [8] we extended ideas in this work to analyse multi-fidelity bandits with Gaussian process payoffs.

2 The Stochastic K-armed Multi-fidelity Bandit

In the classical K-armed bandit, each arm $k \in \mathcal{K} = \{1, \dots, K\}$ is associated with a real valued distribution θ_k with mean μ_k. Let $\mathcal{K}_\star = \mathrm{argmax}_{k\in\mathcal{K}}\,\mu_k$ be the set of optimal arms, $k_\star \in \mathcal{K}_\star$ be an optimal arm and $\mu_\star = \mu_{k_\star}$ denote the optimal mean value. A bandit strategy would *play* an arm $I_t \in \mathcal{K}$ at each time step t and observe a sample from θ_{I_t}. Its goal is to maximise the sum of expected rewards after n time steps $\sum_{t=1}^{n}\mu_{I_t}$, or equivalently minimise the cumulative pseudo-regret $\sum_{t=1}^{n}\mu_\star - \mu_{I_t}$ for *all values* of n. In other words, the objective is to be competitive, in expectation, against an oracle that plays an optimal arm all the time.

In this work we differ from the usual bandit setting in the following aspect. For each arm k, we have access to $M-1$ successively approximate distributions $\theta_k^{(1)}, \theta_k^{(2)}, \dots, \theta_k^{(M-1)}$ to the desired distribution $\theta_k^{(M)} = \theta_k$. We will refer to these approximations as fidelities. Clearly, these approximations are meaningful only if they give us some information about $\theta_k^{(M)}$. In what follows, we will assume that the m^{th} fidelity mean of an arm is within $\zeta^{(m)}$, a *known quantity*, of its highest fidelity mean, where $\zeta^{(m)}$, decreasing with m, characterise the successive approximations. That is, $|\mu_k^{(M)} - \mu_k^{(m)}| \leq \zeta^{(m)}$ for all $k \in \mathcal{K}$ and $m = 1, \dots, M$, where $\zeta^{(1)} > \zeta^{(2)} > \cdots > \zeta^{(M)} = 0$ and the $\zeta^{(m)}$'s are known. It is possible for the lower fidelities to be misleading under this assumption: there could exist an arm k with $\mu_k^{(M)} < \mu_\star = \mu_{k_\star}^{(M)}$ but with $\mu_k^{(m)} > \mu_\star$ and/or $\mu_k^{(m)} > \mu_{k_\star}^{(m)}$ for any $m < M$. In other words, we wish to explicitly account for the *biases* introduced by the lower fidelities, and not treat them as just a higher variance observation of an expensive experiment. This problem of course becomes interesting only when lower fidelities are more attractive than higher fidelities in terms of some notion of cost. Towards this end, we will assign a cost $\lambda^{(m)}$ (such as advertising time, money etc.) to playing an arm at fidelity m where $\lambda^{(1)} < \lambda^{(2)} \cdots < \lambda^{(M)}$.

Notation: $T_{k,t}^{(m)}$ denotes the number of plays at arm k, at fidelity m until t time steps. $T_{k,t}^{(>m)}$ is the number of plays at fidelities greater than m. $Q_t^{(m)} = \sum_{k\in\mathcal{K}} T_{k,t}^{(m)}$ is the number of fidelity m plays at all arms until time t. $\overline{X}_{k,s}^{(m)}$ denotes the mean of s samples drawn from $\theta_k^{(m)}$. Denote $\Delta_k^{(m)} = \mu_\star - \mu_k^{(m)} - \zeta^{(m)}$. When s refers to the number of plays of an arm, we will take $1/s = \infty$ if $s = 0$. $\overline{A}$ denotes the complement of a set $A \subset \mathcal{K}$. While discussing the intuitions in our proofs and theorems we will use $\asymp, \lesssim, \gtrsim$ to denote equality and inequalities ignoring constants.

Regret in the multi-fidelity setting: A strategy for a multi-fidelity bandit problem, at time t, produces an arm-fidelity pair (I_t, m_t), where $I_t \in \mathcal{K}$ and $m_t \in \{1,\dots,M\}$, and observes a sample X_t drawn (independently of everything else) from the distribution $\theta_{I_t}^{(m_t)}$. The choice of (I_t, m_t) could depend on previous arm-observation-fidelity tuples $\{(I_i, X_i, m_i)\}_{i=1}^{t-1}$. The multi-fidelity setting calls for a new notion of regret. For any strategy $\mathcal{A}$ that expends Λ units of the resource, we will define the pseudo-regret $R(\Lambda,\mathcal{A})$ as follows. Let q_t denote the *instantaneous pseudo-reward* at time t and $r_t = \mu_\star - q_t$ denote the instantaneous pseudo-regret. We will discuss choices for q_t shortly. Any notion of regret in the multi-fidelity setting needs to account for this instantaneous regret along with the cost of the fidelity at which we played at time t, i.e. $\lambda^{(m_t)}$. Moreover, we should receive no reward (maximum regret) for any unused capital. These observations lead to the following definition,

$$R(\Lambda,\mathcal{A}) \;=\; \Lambda\mu_\star - \sum_{t=1}^{N}\lambda^{(m_t)}q_t \;=\; \underbrace{\left(\Lambda - \sum_{t=1}^{N}\lambda^{(m_t)}\right)\mu_\star}_{\tilde{r}(\Lambda,\mathcal{A})} \;+\; \underbrace{\sum_{t=1}^{N}\lambda^{(m_t)}r_t}_{\tilde{R}(\Lambda,\mathcal{A})}. \qquad (1)$$

Above, N is the (random) number of plays within capital Λ by $\mathcal{A}$, i.e. the largest n such that $\sum_{t=1}^{n}\lambda^{(m_t)} \le \Lambda$. To motivate our choice of q_t we consider an online advertising example where $\lambda^{(m)}$ is the advertising time at fidelity m and $\mu_k^{(m)}$ is the expected number of clicks per unit time. While we observe from $\theta_{I_t}^{(m_t)}$ at time t, we wish to reward the strategy according to its highest fidelity distribution $\theta_{I_t}^{(M)}$. Therefore regardless of which fidelity we play we set $q_t = \mu_{I_t}^{(M)}$. Here, we are competing against an oracle which plays an optimal arm at any fidelity all the time. Note that we might have chosen q_t to be $\mu_{I_t}^{(m_t)}$. However, this does not reflect the motivating applications for the multi-fidelity setting that we consider. For instance, a clickbait ad might receive a high number of clicks in the short run, but its long term performance might be poor. Furthermore, for such a choice, we may as well ignore the rich structure inherent to the multi-fidelity setting and simply play the arm $\operatorname{argmax}_{m,k}\mu_k^{(m)}$ at each time. There are of course other choices for q_t that result in very different notions of regret; we discuss this briefly at the end of Section 7.

The distributions $\theta_k^{(m)}$ need to be well behaved for the problem to be tractable. We will assume that they satisfy concentration inequalities of the following form. For all $\epsilon > 0$,

$$\forall\, m,\, k, \qquad \mathbb{P}\big(\overline{X}_{k,s}^{(m)} - \mu_k^{(m)} > \epsilon\big) \;<\; \nu e^{-s\psi(\epsilon)}, \qquad \mathbb{P}\big(\overline{X}_{k,s}^{(m)} - \mu_k^{(m)} < -\epsilon\big) \;<\; \nu e^{-s\psi(\epsilon)}. \; (2)$$

Here $\nu > 0$ and ψ is an increasing function with $\psi(0) = 0$ and is at least increasing linearly $\psi(x) \in \Omega(x)$. For example, if the distributions are sub-Gaussian, then $\psi(x) \in \Theta(x^2)$.

The performance of a multi-fidelity strategy which switches from low to high fidelities can be worsened by artificially inserting fidelities. Consider a scenario where $\lambda^{(m+1)}$ is only slightly larger than $\lambda^{(m)}$ and $\zeta^{(m+1)}$ is only slightly smaller than $\zeta^{(m)}$. This situation is unfavourable since there isn't much that can be inferred from the $(m+1)^{\text{th}}$ fidelity that cannot already be inferred from the m^{th} by expending the same cost. We impose the following regularity condition to avoid such situations.

Assumption 1. *The $\zeta^{(m)}$'s decay fast enough such that $\sum_{i=1}^{m}\frac{1}{\psi(\zeta^{(i)})} \le \frac{1}{\psi(\zeta^{(m+1)})}$ for all $m < M$.*

Assumption 1 is not necessary to analyse our algorithm, however, the performance of MF-UCB when compared to UCB is most appealing when the above holds. In cases where M is small enough and can be treated as a constant, the assumption is not necessary. For sub-Gaussian distributions, the condition is satisfied for an exponentially decaying $(\zeta^{(1)}, \zeta^{(2)}, \dots)$ such as $(1/\sqrt{2}, 1/2, 1/2\sqrt{2}\dots)$.

Our goal is to design a strategy $\mathcal{A}_0$ that has low expected pseudo-regret $\mathbb{E}[R(\Lambda, \mathcal{A}_0)]$ for all values of (sufficiently large) Λ, i.e. the equivalent of an anytime strategy, as opposed to a fixed time horizon strategy, in the usual bandit setting. The expectation is over the observed rewards which also dictates the number of plays N. From now on, for simplicity we will write $R(\Lambda)$ when $\mathcal{A}$ is clear from context and refer to it just as regret.

3 The Multi-Fidelity Upper Confidence Bound (MF-UCB) Algorithm

As the name suggests, the MF-UCB algorithm maintains an upper confidence bound corresponding to $\mu_k^{(m)}$ for each $m \in \{1, \dots, M\}$ and $k \in \mathcal{K}$ based on its previous plays. Following UCB strategies [2, 3], we define the following set of upper confidence bounds,

$$\begin{aligned} \mathcal{B}_{k,t}^{(m)}(s) &= \overline{X}_{k,s}^{(m)} + \psi^{-1}\Big(\frac{\rho \log t}{s}\Big) + \zeta^{(m)}, \qquad \text{for all } m \in \{1, \dots, M\},\ k \in \mathcal{K} \\ \mathcal{B}_{k,t} &= \min_{m=1,\dots,M} \mathcal{B}_{k,t}^{(m)}(T_{k,t-1}^{(m)}). \end{aligned} \tag{3}$$

Here ρ is a parameter in our algorithm and ψ is from (2). Each $\mathcal{B}_{k,t}^{(m)}(T_{k,t-1}^{(m)})$ provides a high probability upper bound on $\mu_k^{(M)}$ with their minimum $\mathcal{B}_{k,t}$ giving the tightest bound (See Appendix A). Similar to UCB, at time t we play the arm I_t with the highest upper bound $I_t = \mathrm{argmax}_{k\in\mathcal{K}} \mathcal{B}_{k,t}$.

Since our setup has multiple fidelities associated with each arm, the algorithm needs to determine at each time t which fidelity (m_t) to play the chosen arm (I_t). For this consider an arbitrary fidelity $m < M$. The $\zeta^{(m)}$ conditions on $\mu_k^{(m)}$ imply a constraint on the value of $\mu_k^{(M)}$. If, at fidelity m, the uncertainty interval $\psi^{-1}(\rho \log(t)/T_{I_t,t-1}^{(m)})$ is large, then we have not constrained $\mu_{I_t}^{(M)}$ sufficiently well yet. There is more information to be gleaned about $\mu_{I_t}^{(M)}$ from playing the arm I_t at fidelity m. On the other hand, playing at fidelity m indefinitely will not help us much since the $\zeta^{(m)}$ elongation of the confidence band caps off how much we can learn about $\mu_{I_t}^{(M)}$ from fidelity m; i.e. even if we knew $\mu_{I_t}^{(m)}$, we will have only constrained $\mu_{I_t}^{(M)}$ to within a $\pm\zeta^{(m)}$ interval. Our algorithm captures this natural intuition. Having selected I_t, we begin checking at the first fidelity. If $\psi^{-1}(\rho \log(t)/T_{I_t,t-1}^{(1)})$ is smaller than a threshold $\gamma^{(1)}$ we proceed to check the second fidelity, continuing in a similar fashion. If at any point $\psi^{-1}(\rho \log(t)/T_{I_t,t-1}^{(m)}) \geq \gamma^{(m)}$, we play I_t at fidelity $m_t = m$. If we go all the way to fidelity M, we play at $m_t = M$. The resulting procedure is summarised below in Algorithm 1.

Algorithm 1 MF-UCB

```
• for $t = 1, 2, \dots$
    1. Choose $I_t \in \mathrm{argmax}_{k\in\mathcal{K}} \mathcal{B}_{k,t}$.      (See equation (3).)
    2. $m_t = \min_m \{ m \mid \psi^{-1}(\rho \log t / T_{I_t,t-1}^{(m)}) \geq \gamma^{(m)} \ \vee \ m = M \}$      (See equation (4).)
    3. Play $X \sim \theta_{I_t}^{(m_t)}$.
```

Choice of $\gamma^{(m)}$: In our algorithm, we choose

$$\gamma^{(m)} = \psi^{-1}\left(\frac{\lambda^{(m)}}{\lambda^{(m+1)}}\psi\big(\zeta^{(m)}\big)\right) \tag{4}$$

To motivate this choice, note that if $\Delta_k^{(m)} = \mu_\star - \mu_k^{(m)} - \zeta^{(m)} > 0$ then we can conclude that arm k is not optimal. Step 2 of the algorithm attempts to eliminate arms for which $\Delta_k^{(m)} \gtrsim \gamma^{(m)}$ from plays above the m^{th} fidelity. If $\gamma^{(m)}$ is too large, then we would not eliminate a sufficient number of arms whereas if it was too small we could end up playing a suboptimal arm k (for which $\mu_k^{(m)} > \mu_\star$) too many times at fidelity m. As will be revealed by our analysis, the given choice represents an optimal tradeoff under the given assumptions.

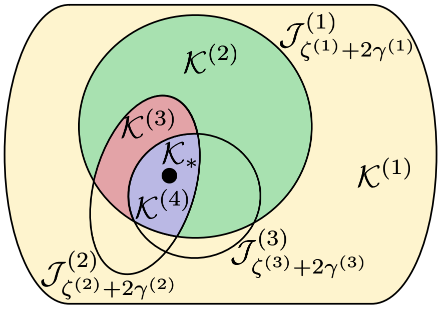
$\mathcal{J}^{(1)}_{\zeta^{(1)}+2\gamma^{(1)}}$
$\mathcal{K}^{(2)}$
$\mathcal{K}^{(3)}$
$\mathcal{K}_*$
$\mathcal{K}^{(4)}$
$\mathcal{K}^{(1)}$
$\mathcal{J}^{(3)}_{\zeta^{(3)}+2\gamma^{(3)}}$
$\mathcal{J}^{(2)}_{\zeta^{(2)}+2\gamma^{(2)}}$

Figure 1: Illustration of the partition $\mathcal{K}^{(m)}$'s for a $M = 4$ fidelity problem. The sets $\mathcal{J}^{(m)}_{\zeta^{(m)}+2\gamma^{(m)}}$ are indicated next to their boundaries. $\mathcal{K}^{(1)}, \mathcal{K}^{(2)}, \mathcal{K}^{(3)}, \mathcal{K}^{(4)}$ are shown in yellow, green, red and purple respectively. The optimal arms $\mathcal{K}_\star$ are shown as a black circle.

4 Analysis

We will be primarily concerned with the term $\tilde{R}(\Lambda, \mathcal{A}) = \tilde{R}(\Lambda)$ from (1). $\tilde{r}(\Lambda, \mathcal{A})$ is a residual term; it is an artefact of the fact that after the $N+1^{\text{th}}$ play, the spent capital would have exceeded Λ. For any algorithm that operates oblivious to a fixed capital, it can be bounded by $\lambda^{(M)}\mu_\star$ which is negligible compared to $\tilde{R}(\Lambda)$. According to the above, we have the following expressions for $\tilde{R}(\Lambda)$:

$$\tilde{R}(\Lambda) \;=\; \sum_{k\in\mathcal{K}} \Delta_k^{(M)} \left(\sum_{m=1}^{M} \lambda^{(m)} T_{k,N}^{(m)} \right), \tag{5}$$

Central to our analysis will be the following partitioning of $\mathcal{K}$. First denote the set of arms whose fidelity m mean is within η of $\mu_\star$ to be $\mathcal{J}_\eta^{(m)} = \{k \in \mathcal{K};\, \mu_\star - \mu_k^{(m)} \le \eta\}$. Define $\mathcal{K}^{(1)} \triangleq \overline{\mathcal{J}}^{(1)}_{\zeta^{(1)}+2\gamma^{(1)}} = \{k \in \mathcal{K};\, \Delta_k^{(1)} > 2\gamma^{(1)}\}$ to be the arms whose first fidelity mean $\mu_k^{(1)}$ is at least $\zeta^{(1)} + 2\gamma^{(1)}$ below the optimum $\mu_\star$. Then we recursively define,

$$\mathcal{K}^{(m)} \triangleq \overline{\mathcal{J}}^{(m)}_{\zeta^{(m)}+2\gamma^{(m)}} \cap \left(\bigcap_{\ell=1}^{m-1} \mathcal{J}^{(\ell)}_{\zeta^{(\ell)}+2\gamma^{(\ell)}} \right), \forall m {\le} M-1, \qquad \mathcal{K}^{(M)} \triangleq \overline{\mathcal{K}_\star} \cap \left(\bigcap_{\ell=1}^{M-1} \mathcal{J}^{(\ell)}_{\zeta^{(\ell)}+2\gamma^{(\ell)}} \right).$$

Observe that for all $k \in \mathcal{K}^{(m)}$, $\Delta_k^{(m)} > 2\gamma^{(m)}$ and $\Delta_k^{(\ell)} \le 2\gamma^{(\ell)}$ for all $\ell < m$. For what follows, for any $k \in \mathcal{K}$, $[\![k]\!]$ will denote the partition k belongs to, i.e. $[\![k]\!] = m$ s.t. $k \in \mathcal{K}^{(m)}$. We will see that $\mathcal{K}^{(m)}$ are the arms that will be played at the m^{th} fidelity but can be excluded from fidelities higher than m using information at fidelity m. See Fig. 1 for an illustration of these partitions.

4.1 Regret Bound for MF-UCB

Recall that $N = \sum_{m=1}^{M} Q_N^{(m)}$ is the total (random) number of plays by a multi-fidelity strategy within capital Λ. Let $n_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$ be the (non-random) number of plays by any strategy that operates only on the highest fidelity. Since $\lambda^{(m)} < \lambda^{(M)}$ for all $m < M$, N could be large for an arbitrary multi-fidelity method. However, our analysis reveals that for MF-UCB, $N \lesssim n_\Lambda$ with high probability. The following theorem bounds R for MF-UCB. The proof is given in Appendix A. For clarity, we ignore the constants but they are fleshed out in the proofs.

Theorem 2 (Regret Bound for MF-UCB)**.** *Let $\rho > 4$. There exists Λ_0 depending on $\lambda^{(m)}$'s such that for all $\Lambda > \Lambda_0$,* MF-UCB *satisfies,*

$$\frac{\mathbb{E}[R(\Lambda)]}{\log(n_\Lambda)} \;\lesssim\; \sum_{k\notin\mathcal{K}_\star} \Delta_k^{(M)} \cdot \frac{\lambda^{([\![k]\!])}}{\psi(\Delta_k^{([\![k]\!])})} \;\asymp\; \sum_{m=1}^{M} \sum_{k\in\mathcal{K}^{(m)}} \Delta_k^{(M)} \frac{\lambda^{(m)}}{\psi(\Delta_k^{(m)})}$$

Let us compare the above bound to UCB whose regret is $\frac{\mathbb{E}[R(\Lambda)]}{\log(n_\Lambda)} \asymp \sum_{k\notin\mathcal{K}_\star} \Delta_k^{(M)} \frac{\lambda^{(M)}}{\psi(\Delta_k^{(M)})}$. We will first argue that MF-UCB does not do significantly worse than UCB in the worst case. Modulo the $\Delta_k^{(M)} \log(n_\Lambda)$ terms, regret for MF-UCB due to arm k is $R_{k,\text{MF-UCB}} \asymp \lambda^{([\![k]\!])}/\psi(\Delta_k^{([\![k]\!])})$. Consider any $k \in \mathcal{K}^{(m)}$, $m < M$ for which $\Delta_k^{(m)} > 2\gamma^{(m)}$. Since

$$\Delta_k^{(M)} \;\le\; \Delta_k^{([\![k]\!])} + 2\zeta^{([\![k]\!])} \;\lesssim\; \psi^{-1}\Big(\frac{\lambda^{([\![k]\!]+1)}}{\lambda^{([\![k]\!])}} \psi(\Delta_k^{([\![k]\!])})\Big),$$

a (loose) lower bound for UCB for the same quantity is $R_{k,\mathsf{UCB}} \asymp \lambda^{(M)}/\psi(\Delta_k^{(M)}) \gtrsim \frac{\lambda^{(M)}}{\lambda^{(\llbracket k \rrbracket+1)}} R_{k,\mathsf{MF\text{-}UCB}}$. Therefore for any $k \in \mathcal{K}^{(m)}, m < M$, MF-UCB is at most a constant times worse than UCB. However, whenever $\Delta_k^{(\llbracket k \rrbracket)}$ is comparable to or larger than $\Delta_k^{(M)}$, MF-UCB outperforms UCB by a factor of $\lambda^{(\llbracket k \rrbracket)}/\lambda^{(M)}$ on arm k. As can be inferred from the theorem, most of the cost invested by MF-UCB on arm k is at the $\llbracket k \rrbracket^{\text{th}}$ fidelity. For example, in Fig. 1, MF-UCB would not play the yellow arms $\mathcal{K}^{(1)}$ beyond the first fidelity (more than a constant number of times). Similarly all green and red arms are played mostly at the second and third fidelities respectively. Only the blue arms are played at the fourth (most expensive) fidelity. On the other hand UCB plays all arms at the fourth fidelity. Since lower fidelities are cheaper MF-UCB achieves better regret than UCB.

It is essential to note here that $\Delta_k^{(M)}$ is small for arms in in $\mathcal{K}^{(M)}$. These arms are close to the optimum and require more effort to distinguish than arms that are far away. MF-UCB, like UCB , invests $\log(n_\Lambda)\lambda^{(M)}/\psi(\Delta_k^{(M)})$ capital in those arms. That is, the multi-fidelity setting does not help us significantly with the "hard-to-distinguish" arms. That said, in cases where K is very large and the sets $\mathcal{K}^{(M)}$ is small the bound for MF-UCB can be appreciably better than UCB.

4.2 Lower Bound

Since, $N \geq n_\Lambda = \lfloor \Lambda/\lambda^{(M)} \rfloor$, any multi-fidelity strategy which plays a suboptimal arm a polynomial number of times at any fidelity after n time steps, will have worse regret than MF-UCB (and UCB). Therefore, in our lower bound we will only consider strategies which satisfy the following condition.

Assumption 3. *Consider the strategy after n plays at any fidelity. For any arm with $\Delta_k^{(M)} > 0$, we have $\mathbb{E}[\sum_{m=1}^{M} T_{k,n}^{(m)}] \in o(n^a)$ for any $a > 0$.*

For our lower bound we will consider a set of Bernoulli distributions $\theta_k^{(m)}$ for each fidelity m and each arm k with mean $\mu_k^{(m)}$. It is known that for Bernoulli distributions $\psi(\epsilon) \in \Theta(\epsilon^2)$ [14]. To state our lower bound we will further partition the set $\mathcal{K}^{(m)}$ into two sets $\mathcal{K}_{\checkmark}^{(m)}, \mathcal{K}_{\text{✗}}^{(m)}$ as follows,

$$\mathcal{K}_{\checkmark}^{(m)} = \{k \in \mathcal{K}^{(m)} \ : \Delta_k^{(\ell)} \leq 0 \ \forall \ell < m\}, \qquad \mathcal{K}_{\text{✗}}^{(m)} = \{k \in \mathcal{K}^{(m)} \ : \exists\, \ell < m \ \text{s.t.}\ \Delta_k^{(\ell)} > 0\}.$$

For any $k \in \mathcal{K}^{(m)}$ our lower bound, given below, is different depending on which set k belongs to.

Theorem 4 (Lower bound for $R(\Lambda)$)**.** *Consider any set of Bernoulli reward distributions with $\mu_\star \in (1/2, 1)$ and $\zeta^{(1)} < 1/2$. Then, for any strategy satisfying Assumption 3 the following holds.*

$$\liminf_{\Lambda \to \infty} \frac{\mathbb{E}[R(\Lambda)]}{\log(n_\Lambda)} \;\geq\; c \cdot \sum_{m=1}^{M} \left[\sum_{k \in \mathcal{K}_{\checkmark}^{(m)}} \Delta_k^{(M)} \frac{\lambda^{(m)}}{{\Delta_k^{(m)}}^2} + \sum_{k \in \mathcal{K}_{\text{✗}}^{(m)}} \Delta_k^{(M)} \min_{\ell \in \mathcal{L}_m(k)} \frac{\lambda^{(\ell)}}{{\Delta_k^{(\ell)}}^2} \right] \tag{6}$$

Here c is a problem dependent constant. $\mathcal{L}_m(k) = \{\ell < m : \Delta_k^{(\ell)} > 0\} \cup \{m\}$ is the union of the m^{th} fidelity and all fidelities smaller than m for which $\Delta_k^{(\ell)} > 0$.

Comparing this with Theorem 2 we find that MF-UCB meets the lower bound on all arms $k \in \mathcal{K}_{\checkmark}^{(m)}$, $\forall m$. However, it may be loose on any $k \in \mathcal{K}_{\text{✗}}^{(m)}$. The gap can be explained as follows. For $k \in \mathcal{K}_{\text{✗}}^{(m)}$, there exists some $\ell < m$ such that $0 < \Delta_k^{(\ell)} < 2\gamma^{(\ell)}$. As explained previously, the switching criterion of MF-UCB ensures that we do not invest too much effort trying to distinguish whether $\Delta_k^{(\ell)} < 0$ since $\Delta_k^{(\ell)}$ could be very small. That is, we proceed to the next fidelity only if we cannot conclude $\Delta_k^{(\ell)} \lesssim \gamma^{(\ell)}$. However, since $\lambda^{(m)} > \lambda^{(\ell)}$ it might be the case that $\lambda^{(\ell)}/{\Delta_k^{(\ell)}}^2 < \lambda^{(m)}/{\Delta_k^{(m)}}^2$ even though $\Delta_k^{(m)} > 2\gamma^{(m)}$. Consider for example a two fidelity problem where $\Delta = \Delta_k^{(1)} = \Delta_k^{(2)} < 2\sqrt{\lambda^{(1)}/\lambda^{(2)}}\zeta^{(1)}$. Here it makes sense to distinguish the arm as being suboptimal at the first fidelity with $\lambda^{(1)}\log(n_\Lambda)/\Delta^2$ capital instead of $\lambda^{(2)}\log(n_\Lambda)/\Delta^2$ at the second fidelity. However, MF-UCB distinguishes this arm at the higher fidelity as $\Delta < 2\gamma^{(m)}$ and therefore does not meet the lower bound on this arm. While it might seem tempting to switch based on estimates for $\Delta_k^{(1)}, \Delta_k^{(2)}$, this idea is not desirable as estimating $\Delta_k^{(2)}$ for an arm requires $\log(n_\Lambda)/\psi(\Delta_k^{(2)})$ samples at the second fidelity; this is is exactly what we are trying to avoid for the majority of the arms via the multi-fidelity setting. We leave it as an open problem to resolve this gap.

	$\mathcal{K}^{(1)}$	$\mathcal{K}^{(2)}$	$\mathcal{K}^{(m)}$	$\mathcal{K}^{(M)}$	$\mathcal{K}_\star$
$\mathbb{E}[T_{k,n}^{(1)}]$	$\frac{\log(n)}{\psi(\Delta_k^{(1)})}$	$\frac{\log(n)}{\psi(\gamma^{(1)})}$	$\dots$ $\frac{\log(n)}{\psi(\gamma^{(1)})}$ $\dots$	$\frac{\log(n)}{\psi(\gamma^{(1)})}$	$\frac{\log(n)}{\psi(\gamma^{(1)})}$
$\mathbb{E}[T_{k,n}^{(2)}]$	$\mathcal{O}(1)$	$\frac{\log(n)}{\psi(\Delta_k^{(2)})}$	$\dots$ $\frac{\log(n)}{\psi(\gamma^{(2)})}$ $\dots$	$\frac{\log(n)}{\psi(\gamma^{(2)})}$	$\frac{\log(n)}{\psi(\gamma^{(2)})}$
$\vdots$ $\mathbb{E}[T_{k,n}^{(m)}]$ $\vdots$		$\mathcal{O}(1)$	$\dots$ $\frac{\log(n)}{\psi(\Delta_k^{(m)})}$ $\dots$	$\frac{\log(n)}{\psi(\gamma^{(m)})}$	$\frac{\log(n)}{\psi(\gamma^{(m)})}$
$\mathbb{E}[T_{k,n}^{(M)}]$			$\mathcal{O}(1)$	$\frac{\log(n)}{\psi(\Delta_k^{(M)})}$	$\Omega(n)$

Table 1: Bounds on the expected number of plays for each $k \in \mathcal{K}^{(m)}$ (columns) at each fidelity (rows) after n time steps (i.e. n plays at any fidelity) in MF-UCB.

5 Proof Sketches

5.1 Theorem 2

First we analyse MF-UCB after n plays (at any fidelity) and control the number of plays of an arm at various fidelities depending on which $\mathcal{K}^{(m)}$ it belongs to. To that end we prove the following.

Lemma 5. (Bounding $\mathbb{E}[T_{k,n}^{(m)}]$ – Informal) *After n time steps of* MF-UCB *for any $k \in \mathcal{K}$,*

$$T_{k,n}^{(\ell)} \lesssim \frac{\log(n)}{\psi(\gamma^{(m)})}, \;\; \forall\, \ell < [\![k]\!], \qquad \mathbb{E}[T_{k,n}^{([\![k]\!])}] \lesssim \frac{\log(n)}{\psi(\Delta_k^{([\![k]\!])}/2)}, \qquad \mathbb{E}[T_{k,n}^{(>[\![k]\!])}] \le \mathcal{O}(1).$$

The bounds above are illustrated in Table 1. Let $\tilde{R}_k(\Lambda) = \sum_{m=1}^{M} \lambda^{(m)} \Delta_k^{(M)} T_{k,N}^{(m)}$ be the regret incurred due to arm k and $\tilde{R}_{kn} = \mathbb{E}[\tilde{R}_k(\Lambda) | N = n]$. Using Lemma 5 we have,

$$\frac{\tilde{R}_{kn}}{\Delta_k^{(M)} \log(n)} \;\lesssim\; \sum_{\ell=1}^{[\![k]\!]-1} \frac{\lambda^{(\ell)}}{\psi(\gamma^{(m)})} \;+\; \frac{\lambda^{([\![k]\!])}}{\psi(\Delta_k^{([\![k]\!])}/2)} \;+\; o(1) \tag{7}$$

The next step will be to control the number of plays N within capital Λ which will bound $\mathbb{E}[\log(N)]$. While $\Lambda/\lambda^{(1)}$ is an easy bound, we will see that for MF-UCB, N will be on the order of $n_\Lambda = \Lambda/\lambda^{(M)}$. For this we will use the following high probability bounds on $T_{k,n}^{(m)}$.

Lemma 6. (Bounding $\mathbb{P}(T_{k,n}^{(m)} > \cdot\,)$ – Informal) *After n time steps of* MF-UCB *for any $k \in \mathcal{K}$,*

$$\mathbb{P}\left(T_{k,n}^{([\![k]\!])} \;\gtrsim\; x \cdot \frac{\log(n)}{\psi(\Delta_k^{([\![k]\!])}/2)} \right) \;\lesssim\; \frac{1}{n^{x\rho - 1}}, \qquad \mathbb{P}\left(T_{k,n}^{(>[\![k]\!])} > x \right) \;\lesssim\; \frac{1}{x^{\rho-2}}.$$

We bound the number of plays at fidelities less than M via Lemma 6 and obtain $n/2 > \sum_{m=1}^{M-1} Q_n^{(m)}$ with probability greater than, say δ, for all $n \ge n_0$. By setting $\delta = 1/\log(\Lambda/\lambda^{(1)})$, we get $\mathbb{E}[\log(N)] \lesssim \log(n_\Lambda)$. The actual argument is somewhat delicate since δ depends on Λ.

This gives as an expression for the regret due to arm k to be of the form (7) where n is replaced by n_Λ. Then we we argue that the regret incurred by an arm k at fidelities less than $[\![k]\!]$ (first term in the RHS of (7)) is dominated by $\lambda^{([\![k]\!])}/\psi(\Delta_k^{([\![k]\!])})$ (second term). This is possible due to the design of the sets $\mathcal{K}^{(m)}$ and Assumption 1. While Lemmas 5, 6 require only $\rho > 2$, we need $\rho > 4$ to ensure that $\sum_{m=1}^{M-1} Q_n^{(m)}$ remains sublinear when we plug-in the probabilities from Lemma 6. $\rho > 2$ is attainable with a more careful design of the sets $\mathcal{K}^{(m)}$. The $\Lambda > \Lambda_0$ condition is needed because initially MF-UCB is playing at lower fidelities and for small Λ, N could be much larger than n_Λ.

5.2 Theorem 4

First we show that for an arm k with $\Delta_k^{(p)} > 0$ and $\Delta_k^{(\ell)} \le 0$ for all $\ell < p$, any strategy should satisfy

$$R_k(\Lambda) \;\gtrsim\; \log(n_\Lambda)\, \Delta_k^{(M)} \left[\min_{\ell \ge p, \Delta_k^{(\ell)} > 0} \frac{\lambda^{(\ell)}}{{\Delta_k^{(\ell)}}^2} \right]$$

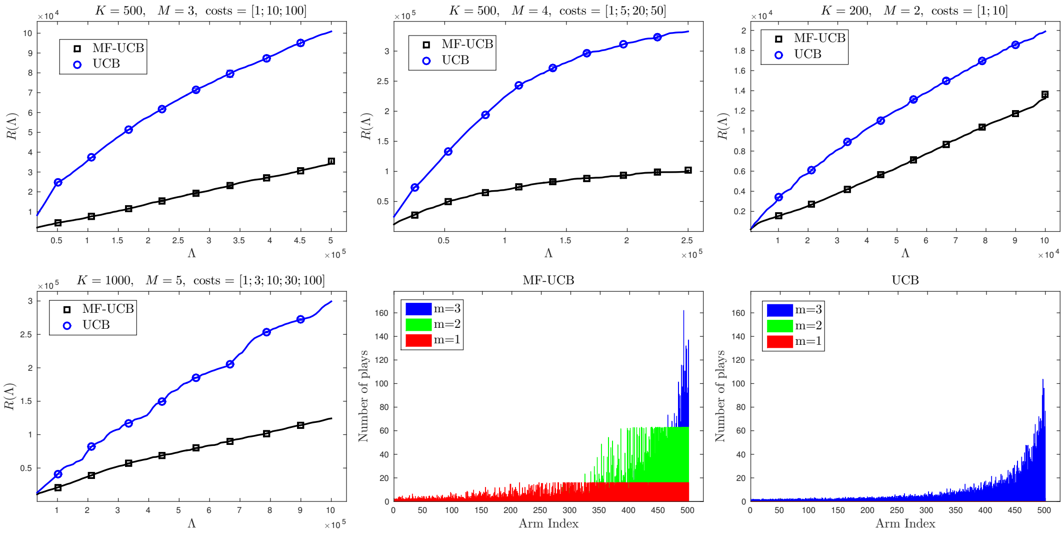

Figure 2: Simulations results on the synthetic problems. The first four figures compares UCB against MF-UCB on four synthetic problems. The title states K, M and the costs $\lambda^{(1)}, \ldots, \lambda^{(M)}$. The first two used Gaussian rewards and the last two used Bernoulli rewards. The last two figures show the number of plays by UCB and MF-UCB on a $K = 500, M = 3$ problem with Gaussian observations (corresponding to the first figure).

where R_k is the regret incurred due to arm k. The proof uses a change of measure argument. The modification has Bernoulli distributions with mean $\tilde{\mu}_k^{(\ell)}, \ell = 1, \ldots, M$ where $\tilde{\mu}_k^{(\ell)} = \mu_k^{(\ell)}$ for all $\ell < m$. Then we push $\tilde{\mu}_k^{(\ell)}$ slightly above $\mu_\star - \zeta^{(\ell)}$ from $\ell = m$ all the way to M where $\tilde{\mu}_k^{(M)} > \mu_\star$. To control the probabilities after changing to $\tilde{\mu}_k^{(\ell)}$ we use the conditions in Assumption 3. Then for $k \in \mathcal{K}^{(m)}$ we argue that $\lambda^{(\ell)} {\Delta_k^{(\ell)}}^2 \gtrsim \lambda^{(m)} / {\Delta_k^{(m)}}^2$ using, once again the design of the sets $\mathcal{K}^{(m)}$. This yields the separate results for $k \in \mathcal{K}_{\checkmark}^{(m)}, \mathcal{K}_{\times}^{(m)}$.

6 Some Simulations on Synthetic Problems

We compare UCB against MF-UCB on a series of synthetic problems. The results are given in Figure 2. Due to space constraints, the details on these experiments are given in Appendix C. Note that MF-UCB outperforms UCB on all these problems. Critically, note that the gradient of the curve is also smaller than that for UCB – corroborating our theoretical insights. We have also illustrated the number of plays by MF-UCB and UCB at each fidelity for one of these problems. The arms are arranged in increasing order of $\mu_k^{(M)}$ values. As predicted by our analysis, most of the very suboptimal arms are only played at the lower fidelities. As lower fidelities are cheaper, MF-UCB is able to use more higher fidelity plays at arms close to the optimum than UCB.

7 Conclusion

We studied a novel framework for studying exploration exploitation trade-offs when cheaper approximations to a desired experiment are available. We propose an algorithm for this setting, MF-UCB, based on upper confidence bound techniques. It uses the cheap lower fidelity plays to eliminate several bad arms and reserves the expensive high fidelity queries for a small set of arms with high expected reward, hence achieving better regret than strategies which ignore multi-fidelity information. We complement this result with a lower bound which demonstrates that MF-UCB is near optimal.

Other settings for bandit problems with multi-fidelity evaluations might warrant different definitions for the regret. For example, consider a gold mining robot where each high fidelity play is a real world experiment of the robot and incurs cost $\lambda^{(2)}$. However, a vastly cheaper computer simulation which incurs $\lambda^{(1)}$ approximate a robot's real world behaviour. In applications like this $\lambda^{(1)} \ll \lambda^{(2)}$. However, unlike our setting lower fidelity plays may not have any rewards (as simulations do not yield actual gold). Similarly, in clinical trials the regret due to a bad treatment at the high fidelity, would be, say, a dead patient. However, a bad treatment at a lower fidelity may not warrant a large penalty. These settings are quite challenging and we wish to work on them going forward.

References

[1] Rajeev Agrawal. Sample Mean Based Index Policies with O(log n) Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*, 1995.

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation Tradeoff Using Variance Estimates in Multi-armed Bandits. *Theor. Comput. Sci.*, 2009.

[3] Peter Auer. Using Confidence Bounds for Exploitation-exploration Trade-offs. *J. Mach. Learn. Res.*, 2003.

[4] Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *The Journal of Machine Learning Research*, 5:255–291, 2004.

[5] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

[6] Mark Cutler, Thomas J. Walsh, and Jonathan P. How. Reinforcement Learning with Multi-Fidelity Simulators. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[7] D. Huang, T.T. Allen, W.I. Notz, and R.A. Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 2006.

[8] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schenider, and Barnabás Póczos. Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations. In *Advances in Neural Information Processing Systems*, 2016.

[9] T. L. Lai and Herbert Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 1985.

[10] Dev Rajnarayan, Alex Haas, and Ilan Kroo. A multifidelity gradient-free optimization method and application to aerodynamic design. In *AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, Etats-Unis*, 2008.

[11] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

[12] W. R. Thompson. On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 1933.

[13] Long Tran-Thanh, Lampros C. Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R. Jennings, and Peter Key. Efficient Regret Bounds for Online Bid Optimisation in Budget-Limited Sponsored Search Auctions. In *UAI*, 2014.

[14] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

[15] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson Sampling for Budgeted Multi-Armed Bandits. In *IJCAI*, 2015.

[16] Chicheng Zhang and Kamalika Chaudhuri. Active Learning from Weak and Strong Labelers. In *Advances in Neural Information Processing Systems*, 2015.