Statistica Sinica Preprint No: SS-2020-0242	
Title	Infinite Arms Bandit: Optimality via Confidence Bounds
Manuscript ID	SS-2020-0242
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0242
Complete List of Authors	Hock Peng Chan and Shouri Hu
Corresponding Author	Hock Peng Chan
E-mail	stachp@nus.edu.sg
Notice: Accepted version subject to English editing.	

INFINITE ARMS BANDIT: OPTIMALITY VIA CONFIDENCE BOUNDS

Hock Peng Chan and Shouri Hu

National University of Singapore

Berry et al. (1997) initiated the development of the infinite arms bandit problem. They derived a regret lower bound of all allocation strategies for Bernoulli rewards with uniform priors, and proposed strategies based on success runs. Bonald and Proutière (2013) proposed a two-target algorithm that achieves the regret lower bound, and extended optimality to Bernoulli rewards with general priors. We present here a confidence bound target (CBT) algorithm that achieves optimality for rewards that are bounded above. For each arm we construct a confidence bound and compare it against each other and a target value to determine if the arm should be sampled further. The target value depends on the assumed priors of the arm means. In the absence of information on the prior, the target value is determined empirically. Numerical studies here show that CBT is versatile and outperforms its competitors.

Key words and phrases: MAB, optimal allocation, sequential analysis, UCB.

1 Introduction

Berry, Chen, Zame, Heath and Shepp (1997) initiated the development of the infinite arms bandit problem. They showed in the case of Bernoulli rewards with uniform prior a $\sqrt{2n}$ regret lower bound for n rewards, and provided algorithms based on success runs that achieve no more than $2\sqrt{n}$ regret. Bonald and Proutière (2013) provided a two-target stopping-time algorithm that can get arbitrarily close to Berry et al.'s lower bound, and is also optimal on Bernoulli rewards with general priors. Wang, Audibert and Munos (2008) considered bounded rewards and showed that their confidence bound algorithm has regret bounds that are $\log n$ times the optimal regret. Vermorel and Mohri (2005) proposed a POKER algorithm for general reward distributions and priors.

The confidence bound method is arguably the most influential approach over the past thirty years for the fixed arm-size bandit problem. Lai and Robbins (1985) derived the smallest asymptotic regret that can be achieved by any algorithm. Lai (1987) showed that by constructing an upper confidence bound (UCB) for each arm, playing the arm with the largest UCB, this smallest regret is achieved in exponential families. The UCB approach was subsequently extended to unknown time-horizons and other parametric families in Agrawal (1995a), Auer, Cesa-Bianchi and Fischer (2002), Burnetas and Katehakis (1996), Cappé, Garivier, Maillard, Munos and Stoltz (2013) and Kaufmann, Cappé and Garivier (2012), and it has been shown to perform well in practice, achieving optimality beyond exponential families. Chan (2020) modified the subsampling approach of Baransi, Maillard and Mannor (2014) and showed that optimality is achieved in exponential families, despite not applying parametric information in the selection of arms. The method can be considered to be applying confidence bounds that are computed empirically from subsample information, which substitutes for the missing parametric information. A related problem is the study of the multi-armed bandit with irreversible constraints, initiated by Hu and Wei (1989).

The Bayesian approach has also enjoyed considerable success, see Berry and Fridstedt (1985), Gittins (1989) and Thompson (1933) for early groundwork and Korda, Kaufmann and Munos (2013) for more recent advances.

We show here how the confidence bound method can be applied on infinite arms. We call this new procedure confidence bound target (CBT). Like UCB, in CBT a confidence bound is computed for each arm. The difference is that in CBT we specify an additional target value. We compare the confidence bound of an arm against this target to decide whether to play an arm further, or to discard it and play a new arm.

We derive a regret lower bound that applies to all bandit algorithms, and proceed to show how the target in CBT is to be chosen to achieve this lower bound. This optimal target depends only on the prior distribution of the arm means and not the reward distributions. That is the reward distributions need not be specified for optimality to be achieved.

To handle the situation in which the prior is not available, we provide an empirical version of CBT in which the target value is computed empirically. Numerical studies on Bernoulli rewards and on a URL dataset show that CBT and empirical CBT outperform their competitors.

In a related continuum-armed bandit problem, there are uncountably infinite number of arms. Each arm is indexed by a known parameter θ and has rewards with mean $f(\theta)$, where f is an unknown continuous function. For solutions to the problem of maximizing the expected sum of rewards, see Agrawal (1995b), Auer, Ortner and Szepesvari (2007), Cope (2009), Kleinberg (2004) and Tyagi and Gärtner (2013).

The layout of this paper is as follows. In Section 2 we describe the infinite arms bandit problem. In Section 3 we review the literature on this problem. In Section 4 we describe CBT. In Section 5 we motivate why the chosen target of CBT leads to the regret lower bound and state the optimality of CBT. In Section 6 we introduce an empirical version of CBT to tackle unknown priors and explain why it works. In Section 7 we perform numerical studies. In Section 8 we provide a short conclusion.

2 Problem set-up

Let $Y_{k1}, Y_{k2}, \ldots$ be i.i.d. rewards from arm k. In the classical multi-armed bandit problem, there are finitely many arms and the objective is to sequentially select the arms so as to maximize expected sum of rewards. Equivalently we minimize the regret, which is the expected cumulative differences between the best arm mean and the mean of the arm played.

In the infinite arms bandit problem that we consider here, there are infinitely many arms and rewards are bounded above by a value that we shall assume for simplicity to be 1. We assume in addition that it is possible for an arm to have reward mean arbitrarily close to 1.

The regret of a bandit algorithm, after n trials, is defined to be

$$R_n = E\Big(\sum_{k=1}^{\infty}\sum_{t=1}^{n_k} X_{kt}\Big), \text{ where } X_{kt} = 1 - Y_{kt} \ (\geq 0) \tag{2.1}$$

is the loss associated with reward Y_{kt}, and n_k is the number of times arm k has been played (hence $n = \sum_{k=1}^{\infty} n_k$). The expectation in (2.1) is with respect to the following Bayesian framework.

Let g be a prior on $(0, \infty)$. For each μ in which $g(\mu) > 0$, let F_μ be a non-negative distribution with mean μ. The expectation in (2.1) is with respect to

$$\mu_k \overset{\text{i.i.d.}}{\sim} g \text{ for } k \geq 1 \text{ and } X_{kt} \overset{\text{i.i.d.}}{\sim} F_{\mu_k} \text{ for } t \geq 1. \tag{2.2}$$

The minimization of the regret under (2.2) for finite arms, known as the *stochastic bandit problem*, has been studied in Agrawal and Goyal (2012), Bubeck and Liu (2013) and Russo and van Roy (2014).

In the infinite arms bandit problem a key decision to be made at each trial is whether to sample a new arm or to play a previously played arm. The Bayesian framework in (2.2) provides useful information on the new arms.

3 Preliminary background

Let $a \wedge b$ denote $\min(a, b)$, $\lfloor \cdot \rfloor$ ($\lceil \cdot \rceil$) denote the greatest (least) integer function and a^+ denote $\max(0, a)$. We say that $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$, $a_n = o(b_n)$ if $\lim_{n\to\infty}(a_n/b_n) = 0$, and $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$.

Berry et al. (1997) showed that for Bernoulli rewards with g uniform on $(0, 1)$, a regret lower bound

$$\liminf_{n\to\infty} \tfrac{R_n}{\sqrt{n}} \geq \sqrt{2} \tag{3.1}$$

is unavoidable. They proposed the following bandit strategies.

1. f-failure strategy. We play the same arm until f failures are encountered. When this happens we switch to a new arm. We do not go back to a previously played arm, that is the strategy is *non-recalling*.

2. s-run strategy. We restrict ourselves to no more than s arms, following the 1-failure strategy in each, until a success run of length s is observed in an arm. When this happens we play the arm for the remaining trials. If no success run of length s is observed in all s arms, then the arm with the highest proportion of success is played for the remaining trials.

3. Non-recalling s-run strategy. We follow the 1-failure strategy until an arm produces a success run of length s. When this happens we play the arm for the remaining trials. If no arm produces a success run of length s, then the 1-failure strategy is used in all n trials.

4. m-learning strategy. We follow the 1-failure strategy for the first m trials, with the arm at trial m played until it yields a failure. Thereafter we play, for the remaining trials, the arm with the highest proportion of successes.

Berry et al. showed that $R_n \sim n/(\log n)$ for the f-failure strategy for any $f \geq 1$, whereas for the $\sqrt{n}$-run strategy, the $\sqrt{n} \log n$-learning strategy and the non-recalling $\sqrt{n}$-run strategy,

$$\limsup_{n\to\infty} \tfrac{R_n}{\sqrt{n}} \leq 2.$$

Bonald and Proutière (2013) proposed a two-target algorithm with target values $s_1 = \lfloor \sqrt[3]{\frac{n}{2}} \rfloor$ and $s_f = \lfloor f\sqrt{\frac{n}{2}} \rfloor$, where $f \geq 2$ is user-defined. An arm is discarded if it has fewer than s_1 successes when it encounters its first failure, or fewer than s_f successes when it encounters its fth failure. If both targets are met, then the arm is accepted and played for the remaining trials. Bonald and Proutière showed that for the uniform prior, the two-target algorithm satisfies, for $n \geq \frac{f^2}{2}$,

$$R_n \leq f + (\tfrac{s_f+1}{f})(\tfrac{s_f-f+2}{s_f-s_1-f+2})^f(2 + \tfrac{1}{f} + \tfrac{2(f+1)}{s_1+1}),$$

from which they conclude that

$$\limsup_{n\to\infty} \tfrac{R_n}{\sqrt{n}} \leq \sqrt{2} + \tfrac{1}{f\sqrt{2}}.$$

Thus for f and n large, the regret is close to the asymptotic lower bound $\sqrt{2n}$.

Bonald and Proutière extended their optimality on Bernoulli rewards to non-uniform priors. They showed that when $g(\mu) \sim \alpha\mu^{\beta-1}$ for some $\alpha > 0$ and $\beta > 0$ as $\mu \to 0$, the regret lower bound of Berry et al. is extended to

$$\liminf_{n\to\infty}(n^{-\frac{\beta}{\beta+1}} R_n) \geq C_0, \text{ where } C_0 = (\tfrac{\beta(\beta+1)}{\alpha})^{\frac{1}{\beta+1}}. \tag{3.2}$$

They also showed that their two-target algorithm with $s_1 = \lfloor n^{\frac{1}{\beta+2}} C_0^{-\frac{\beta+1}{\beta+2}} \rfloor$ and $s_f = \lfloor f n^{\frac{1}{\beta+1}} C_0^{-1} \rfloor$ satisfies

$$\limsup_{f\to\infty} [\limsup_{n\to\infty} (n^{-\frac{\beta}{\beta+1}} R_n)] \le C_0.$$

Wang, Audibert and Munos (2008) proposed a UCB-F algorithm for rewards taking values in $[0,1]$ and showed that under suitable regularity conditions, $R_n = O(n^{\frac{\beta}{\beta+1}} \log n)$. In UCB-F an order $n^{\frac{\beta}{\beta+1}}$ arms are chosen, and confidence bounds are computed on these arms to determine which arm to play. UCB-F is different from CBT in that it pre-selects the number of arms, and it also does not have a mechanism to reject weak arms quickly. Carpentier and Valko (2015) also considered rewards taking values in [0,1] but their interest in maximizing the selection of a good arm differs from the aims here and in the papers above.

4 Proposed methodology

We propose here a new bandit algorithm CBT in which a confidence bound is constructed for each arm and compared against a target value. Let $S_{kt} = \sum_{u=1}^t X_{ku}$, $\bar{X}_{kt} = t^{-1} S_{kt}$ and $\widehat{\sigma}_{kt}^2 = t^{-1} \sum_{u=1}^t (X_{ku} - \bar{X}_{kt})^2$. Let b_n and

c_n be positive confidence coefficients satisfying

$$b_n \to \infty \text{ and } c_n \to \infty \text{ with } b_n + c_n = o(n^\delta) \text{ for all } \delta > 0. \tag{4.1}$$

In our numerical studies we select $b_n = c_n = \log\log n$. We define the confidence bound of arm k, after it has been played t times, to be

$$L_{kt} = \max\Big(\frac{\bar{X}_{kt}}{b_n}, \bar{X}_{kt} - \frac{c_n \widehat{\sigma}_{kt}}{\sqrt{t}}\Big). \tag{4.2}$$

Let $\zeta > 0$ be a specified target value. In CBT the arms are played sequentially. Arm k is played until L_{kt} goes above ζ and it is discarded when that happens. We discuss in Section 5 how ζ should be selected to achieve optimality. It suffices to mention here that the optimal ζ decreases at a polynomial rate with respect to n.

Confidence bound target (CBT)

1. Play arm 1 at trial 1.
2. For $m = 1, \ldots, n-1$: Let k be the arm played at trial m, and let t be the number of times arm k has been played up to trial m.
   (a) If $L_{kt} \le \zeta$, then play arm k at trial $m+1$.
   (b) If $L_{kt} > \zeta$, then play arm $k+1$ at trial $m+1$.

Let K be the number of arms played after n trials, and let n_k be the number of times arm k has been played after n trials. Hence $n = \sum_{k=1}^{K} n_k$.

There are three types of arms that we need to take care of, and that explains the design of L_{kt}. The first type are arms with μ_k (mean of loss X_{kt}) significantly larger than ζ. We would like to reject these arms quickly. The decision to reject arm k when $\bar{X}_{kt}/b_n$ exceeds ζ is key to achieving this.

The second type are arms with μ_k larger than ζ but not by as much as those of the first type. We are unlikely to reject these arms quickly as it is difficult to determine whether μ_k is smaller or larger than ζ based on a small sample. Rejecting arm k when $\bar{X}_{kt} - c_n \widehat{\sigma}_{kt}/\sqrt{t}$ exceeds ζ ensures that arm k is rejected only when it is statistically significant that μ_k is larger than ζ. Though there may be large number of rewards from these arms, their contributions to the regret are small because these arms have small μ_k, as ζ is chosen small when n is large.

The third type of arms are those with μ_k smaller than ζ. For these arms the best strategy (when ζ is chosen correctly) is to play them for the remaining trials. Selecting $b_n \to \infty$ and $c_n \to \infty$ in (4.2) ensures that the probabilities of rejecting these arms are small.

For Bernoulli rewards the first target s_1 of the two-target algorithm is designed for quick rejection of the first type of arms, and the second target s_f is designed for rejection of the second type. What is different is that whereas two-target monitors an arm for rejection only when there are 1 and f failures, with f chosen large for optimality, CBT checks for rejection each time a failure occurs. The frequent monitoring of CBT is a positive feature that results in significantly better performances in the numerical experiments in Section 7.

5 Optimality

We state the regret lower bound in Section 5.1 and show that CBT achieves this bound in Section 5.2.

5.1 Regret lower bound

In Lemma 1 below we motivate the choice of ζ. Let P_μ denote probability and E_μ denote expectation, with respect to $X \stackrel{d}{\sim} F_\mu$. Let $P_g(\cdot) = \int_0^\infty P_\mu(\cdot)g(\mu)d\mu$ and $E_g(\cdot) = \int_0^\infty E_\mu(\cdot)g(\mu)d\mu$. Let $\lambda = \int_0^\infty E_\mu(X|X > 0)g(\mu)d\mu [= E_g(X|X > 0)]$ be the mean of the first positive loss of a random arm. We assume that $\lambda < \infty$. The value λ is the unavoidable cost of exploring a new arm. We consider $E_\mu(X|X > 0)$ instead of μ because it makes sense to reject an arm only after observing a positive loss. For Bernoulli rewards $\lambda = 1$. Let $p(\zeta) = P_g(\mu_1 \leq \zeta)$ and $v(\zeta) = E_g(\zeta - \mu_1)^+$.

Consider an idealized algorithm which plays arm k until a positive loss is observed, and μ_k is revealed when that happens. If $\mu_k > \zeta$, then arm k is rejected and a new arm is played next. If $\mu_k \leq \zeta$, then we stop exploring and play arm k for the remaining trials.

Let

$$r_n(\zeta) = \tfrac{\lambda}{p(\zeta)} + nE_g(\mu_1|\mu_1 \leq \zeta). \tag{5.1}$$

Assuming that the exploration stage of the idealized algorithm uses $o(n)$ trials and ζ is small, its regret is asymptotically $r_n(\zeta)$. Let K be the total number of arms played. The first term in the expansion of $r_n(\zeta)$ approximates $E(\sum_{k=1}^{K-1} \sum_{t=1}^{n_k} X_{kt})$ whereas the second term approximates $E(\sum_{t=1}^{n_K} X_{Kt})$.

Lemma 1. *Let ζ_n be such that $v(\zeta_n) = \frac{\lambda}{n}$. We have*

$$\inf_{\zeta>0} r_n(\zeta) = r_n(\zeta_n) = n\zeta_n.$$

Proof. Since $E_g(\zeta - \mu_1|\mu_1 \leq \zeta) = \frac{v(\zeta)}{p(\zeta)}$, it follows from (5.1) that

$$r_n(\zeta) = \tfrac{\lambda}{p(\zeta)} + n\zeta - \tfrac{nv(\zeta)}{p(\zeta)}. \tag{5.2}$$

It follows from $\frac{d}{d\zeta}v(\zeta) = p(\zeta)$ and $\frac{d}{d\zeta}p(\zeta) = g(\zeta)$ that

$$\tfrac{d}{d\zeta}r_n(\zeta) = \tfrac{g(\zeta)[nv(\zeta)-\lambda]}{p^2(\zeta)}.$$

Since v is continuous and strictly increasing when it is positive, the root to $v(\zeta) = \frac{\lambda}{n}$ exists, and Lemma 1 follows from solving $\frac{d}{d\zeta} r_n(\zeta) = 0$. $\square$

Consider:

(A1) There exists $\alpha > 0$ and $\beta > 0$ such that $g(\mu) \sim \alpha\mu^{\beta-1}$ as $\mu \to 0$.

Under (A1), $p(\zeta) = \int_0^\zeta g(\mu)d\mu \sim \frac{\alpha}{\beta}\zeta^\beta$ and $v(\zeta) = \int_0^\zeta p(\mu)d\mu \sim \frac{\alpha}{\beta(\beta+1)}\zeta^{\beta+1}$ as $\zeta \to 0$, hence $v(\zeta_n) \sim \frac{\lambda}{n}$ for

$$\zeta_n \sim Cn^{-\frac{1}{\beta+1}}, \text{ where } C = (\tfrac{\lambda\beta(\beta+1)}{\alpha})^{\frac{1}{\beta+1}}. \tag{5.3}$$

In Lemma 2 below we state the regret lower bound. We assume there that:

(A2) There exists $a_1 > 0$ such that $P_\mu(X > 0) \geq a_1 \min(\mu, 1)$ for all μ.

We need this assumption to avoid having bad arms that are played a large number of times because their losses are mostly zeros but can be very big when positive.

Lemma 2. *Under* (A1) *and* (A2), *all infinite arms bandit algorithms have regret satisfying*

$$R_n \geq [1 + o(1)]n\zeta_n \sim Cn^{\frac{\beta}{\beta+1}} \text{ as } n \to \infty. \tag{5.4}$$

Lemma 2 is proved in the supplementary document.

Example 1. Consider $X \overset{d}{\sim}$ Bernoulli(μ). Condition (A2) holds with $a_1 = 1$. If g is uniform on (0,1), then (A1) holds with $\alpha = \beta = 1$. Since

$\lambda = 1$, by (5.3), $\zeta_n \sim \sqrt{\frac{2}{n}}$. Lemma 2 says that $R_n \geq [1+o(1)]\sqrt{2n}$, agreeing with Theorem 3 of Berry et al. (1997).

Bonald and Proutière (2013) showed (5.4) in their Lemma 3 for Bernoulli rewards under (A1), and showed that their two-target algorithm gets close to the regret lower bound when f is large. It will be shown in Theorem 1 that the lower bound in (5.4) is achieved by CBT for rewards that need not be Bernoulli.

5.2 Optimality of CBT

We state the optimality of CBT in Theorem 1, after describing below the conditions on discrete rewards under (B1) and continuous rewards under (B2) for which the theorem holds. Let $M_\mu(\theta) = E_\mu e^{\theta X}$.

(B1) The rewards are integer-valued. For $0 < \delta \leq 1$, there exists $\theta_\delta > 0$ such that for $\mu > 0$ and $0 \leq \theta \leq \theta_\delta$,

$$M_\mu(\theta) \leq e^{(1+\delta)\theta\mu}, \tag{5.5}$$

$$M_\mu(-\theta) \leq e^{-(1-\delta)\theta\mu}. \tag{5.6}$$

In addition,

$$P_\mu(X > 0) \leq a_2\mu \text{ for some } a_2 > 0, \tag{5.7}$$

$$E_\mu X^4 = O(\mu) \text{ as } \mu \to 0. \tag{5.8}$$

(B2) The rewards are continuous random variables satisfying

$$\sup_{\mu>0} P_\mu(X \leq \gamma\mu) \rightarrow 0 \text{ as } \gamma \rightarrow 0. \tag{5.9}$$

Moreover (5.8) holds and for $0 < \delta \leq 1$, there exists $\tau_\delta > 0$ such that for $0 < \theta\mu \leq \tau_\delta$,

$$M_\mu(\theta) \leq e^{(1+\delta)\theta\mu}, \tag{5.10}$$

$$M_\mu(-\theta) \leq e^{-(1-\delta)\theta\mu}. \tag{5.11}$$

In addition for each $t \geq 1$, there exists $\xi_t > 0$ such that

$$\sup_{\mu \leq \xi_t} P_\mu(\widehat{\sigma}_t^2 \leq \gamma\mu^2) \rightarrow 0 \text{ as } \gamma \rightarrow 0, \tag{5.12}$$

where $\widehat{\sigma}_t^2 = t^{-1}\sum_{u=1}^t (X_u - \bar{X}_t)^2$ and $\bar{X}_t = t^{-1}\sum_{u=1}^t X_u$ for i.i.d. $X_u \stackrel{d}{\sim} F_\mu$.

Theorem 1. *Assume* (A1), (A2) *and either* (B1) *or* (B2). *For CBT with threshold* ζ_n *satisfying* (5.3) *and* b_n, c_n *satisfying* (4.1),

$$R_n \sim n\zeta_n \text{ as } n \rightarrow \infty. \tag{5.13}$$

Theorem 1 says that CBT is optimal as it attains the lower bound given in Lemma 2. In the examples below we show that the regularity conditions (A2), (B1) and (B2) are reasonable and checkable. The proof of Theorem 1 and the checking details in Examples 3–5 are in the supplementary document.

Example 2. If $X \stackrel{d}{\sim}$ Bernoulli(μ) under P_μ, then

$$M_\mu(\theta) = 1 - \mu + \mu e^\theta \leq \exp[\mu(e^\theta - 1)].$$

Hence (5.5) and (5.6) hold with $\theta_\delta > 0$ satisfying

$$e^{\theta_\delta} - 1 \leq \theta_\delta(1+\delta) \text{ and } e^{-\theta_\delta} - 1 \leq -\theta_\delta(1-\delta).$$

In addition (5.7) holds with $a_2 = 1$, and (5.8) holds because $E_\mu X^4 = \mu$. Condition (A2) holds with $a_1 = 1$.

Example 3. Let F_μ be a distribution with support on $0, \ldots, I$ for some positive integer $I > 1$ and having mean μ. Condition (A2) holds with $a_1 = I^{-1}$ and (B1) holds as well.

Example 4. Let F_μ be the Poisson distribution with mean μ. Condition (A2) holds with $a_1 = 1 - e^{-1}$ and (B1) holds as well.

Example 5. Let Z be a continuous non-negative random variable with mean 1, and with $Ee^{\tau_0 Z} < \infty$ for some $\tau_0 > 0$. Let F_μ be the distribution of μZ. Condition (A2) holds with $a_1 = 1$ and (B2) holds as well.

6 Methodology for unknown priors

The optimal implementation of CBT, in particular the computation of the optimal target ζ_n, assumes knowledge of how $g(\mu)$ behaves for μ near 0. For g

unknown we will rely on Theorem 1 to motivate an empirical implementation of CBT.

What is striking about (5.13) is that it relates the optimal target ζ_n to $\frac{R_n}{n}$, and moreover this relation does not depend on either the prior g or the reward distributions. We suggest therefore, in an empirical implementation of CBT, to apply targets

$$\zeta(m) := \frac{S'_m}{n}, \tag{6.1}$$

where S'_m is the sum of the losses X_{kt} over the first m trials.

In the beginning with m small, $\zeta(m)$ underestimates the optimal target, but this will only encourage exploration, which is the right strategy at the beginning. As m increases $\zeta(m)$ gets closer to the optimal target, and empirical CBT behaves like CBT when deciding whether to play an arm further. A key difference between CBT and empirical CBT is that empirical CBT decides from among all played arms which to play further whereas CBT plays the arms sequentially.

Empirical CBT

Notation: When there are m total rewards, let $n_k(m)$ denote the number of rewards from arm k and let K_m denote the number of arms played.

For $m = 0$, play arm 1. Hence $K_1 = 1$, $n_1(1) = 1$ and $n_k(1) = 0$ for $k > 1$.

For $m = 1, \dots, n-1$:

1. If $\min_{1 \le k \le K_m} L_{kn_k(m)} \le \zeta(m)$, then play the arm k minimizing $L_{kn_k(m)}$ at trial $m+1$.

2. If $\min_{1 \le k \le K_m} L_{kn_k(m)} > \zeta(m)$, then play a new arm $K_m + 1$ at trial $m+1$.

Empirical CBT, unlike CBT, does not achieve the smallest regret. This is because when a good arm (that is an arm with μ_k below optimal target) appears early, we are not sure whether this is due to good fortune or that the prior is disposed towards arms with small μ_k, so we explore more arms before we are certain and play the good arm for the remaining trials. Similarly when no good arm appears after many trials, we may conclude that the prior is disposed towards arms with large μ_k, and play an arm with μ_k above the optimal target for the remaining trials, even though it is advantageous to explore further.

As the analysis of the regret of empirical CBT is complicated, we consider an idealized version of empirical CBT in the supplementary document and derive its asymptotic regret there, to give us a sense of the additional regret when applying CBT empirically.

In the idealized version of empirical CBT, μ_k is revealed after the first positive loss of arm k is observed. The number of arms played is the smallest K satisfying

$$\min_{1 \leq k \leq K} \mu_k \leq \tfrac{K\lambda}{n},$$

and exploitation of the best arm begins after $\mu_1, \ldots, \mu_K$ have been revealed. The idealized empirical CBT is like the idealized algorithm described in the beginning of Section 5.1, but with a target $\zeta = \frac{k\lambda}{n}$, after k arms have been played. This is because λ is the mean of the first positive loss of each arm, so after k arms have been played the sum of losses has mean $k\lambda$. The idealized empirical CBT is a simplification of empirical CBT that captures the additional regret of empirical CBT over CBT when applying a target that does not depend on the prior.

Theorem 2. *The idealized empirical CBT has regret*

$$R'_n \sim I_\beta n \zeta_n, \tag{6.2}$$

where $I_\beta = (\frac{1}{\beta+1})^{\frac{1}{\beta+1}}(2 - \frac{1}{(\beta+1)^2})\Gamma(2 - \frac{\beta}{\beta+1})$ *and* $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

The constant I_β increases with β, with $I_0 = 1$ and $\lim_{\beta \to \infty} I_\beta = 2$. The increase is quite slow so that for reasonable values of β it is closer to 1 than

2. For example $I_1 = 1.10$, $I_3 = 1.24$ and $I_5 = 1.36$. Equation (6.2) says that empirical CBT should have regret not more than 36% over the baseline lower bound when $\beta \le 5$. This agrees with the simulation outcomes in Section 7.

7 Numerical studies

We study arms with Bernoulli rewards in Example 6 and arms with unspecified reward distributions in Example 7. In our simulations 10,000 datasets are generated for each entry in Tables 1–4, and standard errors are placed after the $\pm$ sign. In both CBT and empirical CBT, we select $b_n = c_n = \log \log n$. Aziz (2019) performed numerical studies involving various infinite arms bandit algorithms, including CBT and empirical CBT, with the objective of finding the arm with the best mean. There is also an application of infinite arms bandit there on an online dataset involving voting responses to 3795 proposed captions of a cartoon, on a New Yorker website.

Example 6. We consider Bernoulli rewards with uniform prior $g(\mu) = 1$, as well as the Beta priors $g(\mu) = 3\mu^2$ [i.e. Beta(3,1)], $g(\mu) = \frac{15}{16}\mu^2(1-\mu)^{-\frac{1}{2}}$ [i.e. Beta(3,$\frac{1}{2}$)], $g(\mu) = 5\mu^4$ [i.e. Beta(5,1)] and $g(\mu) = \frac{315}{256}\mu^4(1-\mu)^{-\frac{1}{2}}$ [i.e. Beta(5,$\frac{1}{2}$)].

We see from Tables 1–3 that the two-target algorithm does better with

		Regret			
		n =100	n =1000	n =10,000	n =100,000
CBT	$\zeta = \sqrt{2/n}$	14.6±0.1	51.5±0.3	162±1	504±3
	empirical	15.6±0.1	54.0±0.3	172±1	531±3
Berry et al.	1-failure	21.8±0.1	152.0±0.6	1123±4	8955±28
	$\sqrt{n}$-run	19.1±0.2	74.7±0.7	260±3	844±9
	$\sqrt{n}$-run (non-recall)	15.4±0.1	57.7±0.4	193±1	618±4
	$n^{\frac{1}{2}} \log n$-learning	18.7±0.1	84.4±0.6	311±3	1060±9
Two-target	$f = 3$	15.2±0.1	52.7±0.3	167±1	534±3
	$f = 6$	16.3±0.1	55.8±0.4	165±1	511±3
	$f = 9$	17.5±0.1	58.8±0.4	173±1	514±3
UCB-F	$K = \lfloor \sqrt{n/2} \rfloor$	39.2±0.1	206.4±0.4	1204±1	4432±5
Lower bound	$\sqrt{2n}$	14.1	44.7	141	447

Table 1: Regrets for Bernoulli rewards with uniform prior.

$f = 3$ at smaller n, and $f = 6$ or 9 at larger n. CBT is the best performer uniformly over sample size and prior, and empirical CBT is competitive against two-target with f fixed.

Even though CBT outperforms empirical CBT, its optimal target ζ depends on the prior. On the other hand when applying empirical CBT, the same algorithm is used for all priors here and on the URL dataset in Example 7 with unspecified prior. Hence though it seems that empirical CBT is

		Regret			
		Beta$(3,1)$	Beta$(3,\frac{1}{2})$	Beta$(5,1)$	Beta$(5,\frac{1}{2})$
CBT	$\zeta = Cn^{-\frac{1}{\beta+1}}$	284.2±0.9	363.0±1.0	474.0±1.0	554.3±1.1
	empirical	299.6±0.9	382.2±1.1	509.6±1.0	592.7±1.0
$n^{\frac{1}{\beta+1}}$-run	non-recall	346.3±1.3	445.7±1.7	546.5±1.4	658.8±1.6
Two-target	$f=3$	310.7±1.1	390.8±1.3	510.3±1.3	592.1±1.3
	$f=6$	301.2±1.2	385.9±1.4	520.9±1.5	619.5±1.6
	$f=9$	311.0±1.3	400.1±1.6	545.3±1.6	649.6±1.7
UCB-F		649.5±0.3	779.2±0.3	774.0±0.3	867.6±0.2
Lower bound	$Cn^{\frac{\beta}{\beta+1}}$	251.5	336.4	426.3	538.5

Table 2: Regrets for Bernoulli rewards with Beta priors at $n = 1000$.

numerically comparable to two-target and inferior to CBT, in applications where prior is unknown or incorrectly specified, it can perform much better.

For the uniform prior, the best performing among the algorithms in Berry et al. (1997) is the non-recalling $\sqrt{n}$-run algorithm. For UCB-F [cf. Wang et al. (2008)], the selection of $K = \lfloor (\frac{\beta}{\alpha})^{\frac{1}{\beta+1}} (\frac{n}{\beta+1})^{\frac{\beta}{\beta+1}} \rfloor$ $(\sim \frac{1}{p(\zeta_n)})$ and "exploration sequence" $\mathcal{E}_m = \sqrt{\log m}$ works well.

Example 7. We consider the URL dataset studied in Vermorel and Mohri (2005), where a POKER algorithm for dealing with large number of arms is proposed. We reproduce part of their Table 1 in our Table 4,

		Regret (×10)			
		Beta$(3,1)$	Beta$(3,\frac{1}{2})$	Beta$(5,1)$	Beta$(5,\frac{1}{2})$
CBT	$\zeta = Cn^{-\frac{1}{\beta+1}}$	866±3	1127±4	2122±5	2569±6
	empirical	1004±3	1318±4	2547±5	3149±6
$n^{\frac{1}{\beta+1}}$-run	non-recall	1476±7	1713±8	3874±13	4142±13
Two-target	$f=3$	1159±5	1501±6	2973±9	3559±11
	$f=6$	990±4	1308±5	2527±7	3060±9
	$f=9$	957±4	1257±5	2429±7	2992±9
UCB-F		3739±3	4522±4	6488±4	7499±5
Lower bound	$Cn^{\frac{\beta}{\beta+1}}$	795	1064	1979	2499

Table 3: Regrets (×10) for Bernoulli rewards with Beta priors at n=100,000.

		Regret	
Algorithm	ϵ	n =130	n =1300
emp. CBT		212±2	123.8±0.6
POKER		203	132
ϵ-greedy	0.05	733	431
ϵ-first	0.15	725	411
ϵ-decreasing	1.0	738	411

Table 4: Average regret R_n/n.

together with new simulations on empirical CBT. The dataset consists of the retrieval latency of 760 university home-pages, in milliseconds, with a sample size of more than 1300 for each home-page. The numbers in the dataset correspond to the non-negative losses X_{kt}. The dataset can be downloaded from "sourceforge.net/projects/bandit".

In our simulations the losses are randomly permuted within home-page in each run. At $n = 130$ POKER performs better than empirical CBT whereas at $n = 1300$ empirical CBT performs better. The other algorithms are uniformly worse than both POKER and empirical CBT.

The algorithm ϵ-first refers to exploring the first ϵn losses, with random selection of the arms to be played. This is followed by pure exploitation for the remaining $(1-\epsilon)n$ losses, on the "best" arm (with the smallest mean loss). The algorithm ϵ-greedy refers to selecting, in each trial, a random arm with probability ϵ, and the best arm with the remaining $1-\epsilon$ probability. The algorithm ϵ-decreasing is like ϵ-greedy except that in the mth trial, we select a random arm with probability $\min(1, \frac{\epsilon}{m})$, and the best arm otherwise. Both ϵ-greedy and ϵ-decreasing are disadvantaged by not making use of information on the total number of trials. Vermorel and Mohri also ran simulations on more complicated strategies like LeastTaken, SoftMax, Exp3, GaussMatch and IntEstim, with average regret ranging from 276–747 at $n = 130$ and 189–599 at $n = 1300$.

8 Conclusion

CBT optimizes the regret in the infinite arms bandit problem when it is possible for an arm to have reward mean arbitrarily close to the upper bound of the rewards. This optimality is over all bandit algorithms and does not require knowledge of the reward distribution for a given arm mean. It depends however on the correct selection of a target value that is computed from an assumed prior.

Empirical CBT is like CBT with the key difference that it computes the target value empirically. Though not optimal, it performs well in numerical studies and is more practical as it can be applied without assuming a prior.

We suggest here two extensions of CBT and empirical CBT for future work. The first is to handle the situation of sample size not known in advance. Bonald and Proutière (2013) have a version of two-target that they believe to be optimal for Bernoulli rewards, when sample size is not known in advance.

The second extension is to incorporate covariate information in the computation of confidence bounds, leading to recommended arms that are specific to subgroups of the population. Modern developments in the finite-arms bandit literature has centered on the handling of covariate information, see for example Goldenshluger and Zeevi (2013), Perchet and Rigollet (2013), Slivkins (2014), Wang, Kulkarni and Poor (2005) and Yang and Zhu (2002). When number of arms is comparable to or larger than sample size, an infinite-arms approach is more appropriate and will provide strategies that differ from a finite-arms framework.

9 Supplementary materials

The proofs of Lemma 2 and Theorems 1 and 2, as well as the verifications of (A2), (B1) and (B2) in Examples 3–5 are in the supplementary document.

References

[1] Agrawal, R. (1995a). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. in Appl. Probab.* **27** 1054–1078.

[2] Agrawal, R. (1995b). The continuum-armed bandit problem. *SIAM J. Control and Optimization* **33** 1926–1951.

[3] Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. *In The 25th Annual Conference on Learning Theory* **23** 39.1–39.26.

[4] Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.* **47** 235–256.

[5] Auer, P., Ortner, R. and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem. *In The 20th Annual Conference on Learning Theory* **18** 454–468.

[6] Aziz, M. (2019). *On Multi-Armed Bandits Theory and Applications.* Ph.D. thesis, Northeastern University.

[7] Baransi, A., Maillard, O.A. and Mannor, S. (2014). Sub-sampling for multi-armed bandits. *In Proc. of the European Conference on Mach. Learn.* 13.

[8] Berry, D., Chen, R., Zame, A., Heath, D. and Shepp, L. (1997). Bandit problems with infinitely many arms. *Ann. Statist.* **25** 2103–2116.

[9] Berry, D. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments.* CRC Press, London.

[10] Bonald, T. and Proutière, A. (2013). Two-target algorithms for infinite-armed bandits with Bernoulli rewards. *In NIPS* **26** 2184–2192.

[11] Bubeck, S. and Liu, C. Y. (2013). Prior-free and prior-dependent regret bounds for Thompson sampling. *In NIPS* **26** 638–646.

[12] Burnetas, A. and Katehakis, M. (1996). Optimal adaptive policies for sequential allocation problems. *Adv. in Appl. Math.* **17** 122–142.

[13] Cappé, O., Garivier, A., Maillard, O. A., Munos, R. and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* **41** 1516–1541.

[14] Carpentier, A. and Valko, M. (2015). Simple regret for infinitely many bandits. *In The 32th International Conference on Mach. Learn.* **37** 1133–1141.

[15] Chan, H. (2020). The multi-armed bandit problem: An efficient nonparametric solution. *Ann. Statist.* **48** 346–373.

[16] Cope, E. W. (2009). Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. on Autom. Control* **54** 1243–1253.

[17] Gittins, J. (1989). *Multi-armed Bandit Allocation Indices.* Wiley, New York.

[18] Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems* **3** 230–261.

[19] Hu, I. and Wei, C.Z. (1989). Irreversible adaptive allocation rules. *Ann. Statist.* **17** 801–822.

[20] Kaufmann, E., Cappé, O. and Garivier, A. (2012). On Bayesian upper confidence bounds for bandit problems. *In Proc. Mach. Learn. Res.* **22** 592–600.

[21] Kleinberg, R.D. (2004). Nearly tight bounds for the continuum-armed bandit problem. *In NIPS* **17** 697–704.

[22] Korda, N., Kaufmann, E. and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. *In NIPS* **26** 1448–1456.

[23] Lai, T.L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114.

[24] Lai, T.L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22.

[25] Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *Ann. Statist.* **41** 693–721.

[26] Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Math. Operations Res.* **39** 1221–1243.

[27] Slivkins, A. (2014). Contextual bandits with similarity information. *J. Mach. Learn. Res.* **15** 2533–2568.

[28] Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.

[29] Tyagi, H. and Gärtner, B. (2013). Continuum armed bandit problem of few variables in high dimensions. *In The 11th Workshop on Approximation and Online Algorithms* 108–119.

[30] Vermorel, J. and Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. *Machine Learning: ECML*, Springer, Berlin.

[31] Wang, Y., Audibert, J. and Munos, R. (2008). Algorithms for infinitely many-armed bandits. *In NIPS* **8** 1–8.

[32] Wang, C., Kulkarni, S. and Poor, H. (2005). Bandit problems with side information. *IEEE Trans. on Autom. Control* **50** 338–355.

[33] YANG, Y. and ZHU, D. (2002). Randomized allocation with nonparametric estimation for a multiarmed bandit problem with covariates. *Ann Statist.* **30** 100–121.