Greedy Confidence Pursuit: A Pragmatic Approach to Multi-bandit Optimization

Philip Bachman and Doina Precup

McGill University, School of Computer Science
phil.bachman@gmail.com, dprecup@cs.mcgill.ca

Abstract. We address the practical problem of maximizing the number of high-confidence results produced among multiple experiments sharing an exhaustible pool of resources. We formalize this problem in the framework of bandit optimization as follows: given a set of multiple multi-armed bandits and a budget on the total number of trials allocated among them, select the top-m arms (with high confidence) for as many of the bandits as possible. To solve this problem, which we call *greedy confidence pursuit*, we develop a method based on *posterior sampling*. We show empirically that our method outperforms existing methods for top-m selection in single bandits, which has been studied previously, and improves on baseline methods for the full greedy confidence pursuit problem, which has not been studied previously.

1 Introduction

Clinical and scientific teams often pursue multiple research objectives on a fixed budget. To obtain as many significant results as possible, they must intelligently allocate their limited resources among one or more concurrent experiments. The machine learning community has developed ways to formulate and address variations on this problem. For example, budgeted learning [12] and subsequent work considers the problem of active learning when a fixed budget is given for probing which model among a collection of models is best for a given task.

In this paper, we adopt the framework provided by bandit problems [3] to address resource allocation among multiple concurrent tasks. Bandits offer a simple way of formalizing many decision problems, e.g. deciding which among a set of drugs most effectively treats a particular disease. In the standard formulation a bandit has multiple arms with unknown expected payoffs and one must probingly pull the arms in order to find the best one. Most bandit optimization problems focus on regret minimization, i.e. minimizing some measure of loss incurred over the course of an experiment. The goal in practical experimental settings, e.g. clinical trials, is often different: one typically has a fixed budget for acquiring patients to be treated, and the goal is to identify the best treatment option *at the end* of the experiment. Hence, payoffs during the experiment are not counted, in contrast to regret minimization, and the objective is solely to maximize the (statistical) confidence with which the best action can be selected

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part I, LNAI 8188, pp. 241–256, 2013.

after the experiment is over. In a recent series of papers, this idea has been developed under the label "pure exploration" in multi-armed bandits [4, 2, 8].

The problem of selecting the best arm with high confidence using a minimum number of trials has also been tackled by [13] and [7]. In [9], the authors extended the approach of [7] to the case in which one wants to select not just the best arm, but the m arms with highest payoffs. In recent work [10], the same authors provided an alternative algorithm with stronger PAC guarantees. Note that best-arm selection is the special case of top-m selection where $m = 1$.

In the clinical trial setting, significant interest is currently directed towards personalized medicine based on treatments which only work for specific sub-populations. For example, it is understood that diseases like cancer may evolve differently based on certain genetic mutations, and thus any treatment for such a disease may only benefit certain types of patients. In such cases, no budgeted clinical trial can hope to show that a treatment is universally effective; instead, one should try to identify sub-populations within which the treatment works with high confidence. One can naturally describe this problem using multiple bandits (i.e. sub-populations) each comprising multiple arms (i.e. available treatments). Typically, a fixed total number of patients can be enrolled (corresponding to a fixed total number of trials). Hence, patients should be recruited and allocated among the sub-populations and treatments to maximize the number of sub-populations for which an effective treatment is confidently identified.

We formalize this problem as multi-bandit top-m selection: given a set of n multi-armed bandits, a trial budget T, and a target confidence τ, maximize the number of top-m groups identified with confidence $\rho > \tau$ after performing T trials. We refer to this general problem as *greedy confidence pursuit*, as it preferentially directs resources (i.e. trials) towards experiments (i.e. bandits) in which confident results are easiest to achieve. Work in [8] addresses a related problem which focuses, roughly speaking, on minimizing the probability of incorrectly identifying any top-m group. We will discuss the relation between [8] and our own work in detail. Similarly, work in [6] considers a multi-bandit objective which focuses on minimizing the maximum uncertainty among the estimated per-arm returns. In contrast, we propose the more pragmatic objective of maximizing the number of confident results achieved on a fixed budget[1].

In Section 2 of this paper, we define greedy confidence pursuit and contrast it with objectives previously considered in the multi-bandit setting. In Section 3 we develop an algorithm for intra-bandit top-m selection in bandits with Bernoulli-distributed returns. In Section 5 we develop an algorithm for inter-bandit trial allocation which completes our approach to greedy confidence pursuit. In Sections 4 and 6, we compare the performance of our algorithms with existing algorithms across a range of problems, illustrating the power of our approach and highlighting the differences between greedy confidence pursuit and other

[1] Our objective is pragmatic as many practical scenarios (e.g. scientific publication) require surpassing some confidence threshold for capturing *any* value, with extra confidence beyond the threshold providing rapidly diminishing additional value.

objectives previously considered in the multi-bandit setting. We conclude the paper and discuss future work in Section 7.

2 Motivating and Formulating Our Objective

Consider a pharmaceutical developer evaluating a new drug for potential use in multiple sub-populations of patients. Given a fixed budget for processing trial patients, the developer may seek to maximize the number of sub-populations for which their proposed drug is identified as significantly better than existing treatments[2]. We can formalize this problem as follows:

- Each sub-population is represented by a bandit b_i.
- Each bandit has a set of n arms $A_i = \{a_{i1}, ..., a_{in}\}$, with arm a_{i1} representing the new drug and the rest representing existing treatments.
- The random variables $\mathbf{R}_i = \{\mathbf{r}_{i1}, ..., \mathbf{r}_{in}\}$ give the per-trial outcomes for b_i.
- The objective is to maximize the number of bandits b_i for which we find $\mathbb{E}[\mathbf{r}_{i1}] > \max_{j \neq 1} \mathbb{E}[\mathbf{r}_{ij}]$ with confidence $\rho_i > \tau$.

In the above scenario, only confident results involving a particular target arm (i.e. the pharmaceutical developer's proposed drug) are considered worth pursuing. This represents a variant of the general greedy confidence pursuit problem, in which confident results involving any best arm are pursued equally.

We approach greedy confidence pursuit by decomposing trial allocation into three stages: bandit selection, arm selection, and belief updates based on the trial outcome. Methods for these stages can be combined "a la carte", which facilitates algorithm development and eases comparison with existing work.

In previous work [8], given a set of N bandits $B = \{b_1, ...b_N\}$, Gabillon et. al proposed the following objective for multi-bandit subset selection:

$$\text{maximize } \mathbb{E}_H[\min_i \rho_i] \,, \tag{1}$$

where ρ_i measures the confidence that the top-m group selected for bandit b_i is correct. In contrast, our objective can be written as follows:

$$\text{maximize } \mathbb{E}_H[\sum_i \mathbb{I}\{\rho_i > \tau\}] \,, \tag{2}$$

where ρ_i is as above, τ is a confidence threshold and $\mathbb{I}$ is the indicator function. The expectations are over *histories* (i.e. sequences of observed trial outcomes). Intuitively, (1) maximizes a lower bound on the per-bandit confidences and (2) maximizes the number of bandits for which the top-m group can be selected with high confidence. The precise confidence measure we use is given in (3).

For both objectives (1) and (2), the trials allocated to bandit b_i should be distributed among its arms to maximize ρ_i. Hence, good arm selection for (1) will

[2] While human drug trials are slow to adopt novel experimental designs, one could analogously consider trials of a new consumer product across multiple potential target demographics, or exploratory drug trials in non-human model systems.

also be good for (2). However, methods for optimizing these objectives will select bandits quite differently. Intuitively, methods optimizing (1) will tend to allocate trials to bandits with relatively low confidence, while methods optimizing (2) will tend to allocate trials to bandits with relatively low expected *completion cost* (out of the bandits b_i for which $\rho_i \leq \tau$). The practical differences between (1) and (2) are most striking when some bandit b_i is effectively intractable with respect to the operative confidence measure and trial budget; an algorithm optimizing (1) will still sink its budget into (hopelessly) pursuing improvements in ρ_i, while algorithms optimizing (2) will ignore b_i in favor of lower-hanging fruit.

The completion cost is a critical concept when working with (2), which we define as follows: a bandit b_i has completion cost c_i if efficiently allocating c_i trials among the arms of b_i is expected to push ρ_i above τ. Note that if each c_i were deterministic and known a priori, an optimal trial allocation policy for (2) would be to sort the bandits such that $c_1 \leq c_2 \leq ... \leq c_N$, then allocate c_1 trials to b_1, c_2 trials to b_2 etc., until budget exhaustion. This greedy policy maximizes the number of tasks completed on a fixed budget when each task has a known cost. The difficulty in our case is that each c_i is neither known a priori nor deterministic. Thus, a balance between exploring (to better estimate each c_i) and exploiting (to push each ρ_i past τ) must be struck.

In Section 5 we describe an estimator for the completion costs c_i and discuss how to use these estimates for inter-bandit trial allocation during greedy confidence pursuit. Next, we present our method for intra-bandit top-m selection.

3 Bayesian Top-m Selection

Our intra-bandit subset selection algorithm uses Bayesian estimates of the per-arm returns and follows a general approach called posterior sampling, of which Thompson sampling [16] is perhaps the best-known example. The notation introduced in this section will be reused throughout the remainder of this paper.

3.1 Definitions and Notation

For a set B of N bandits, where each $b_i \in B$ has a set A_i of n arms with Bernoulli-distributed returns, our algorithm maintains its beliefs about the return of each arm $a_{ij} \in A_i$ using a beta distribution $\mathcal{B}_{ij} = \mathcal{B}(\alpha_{ij}, \beta_{ij})$, where α_{ij} and β_{ij} count the observed successes and failures for arm a_{ij}, respectively. We set priors over the returns by initializing all parameters α_{ij} and β_{ij} to a common value (e.g. we set them to 1 in all of our tests). The belief for arm a_{ij} is updated by incrementing α_{ij} or β_{ij} following each trial allocated to a_{ij}. The MAP estimate of the return of a_{ij} is given by $\alpha_{ij}(\alpha_{ij} + \beta_{ij})^{-1}$.

For a bandit b_i with current MAP return estimates $\bar{R}_i = \{\bar{r}_{i1}, ..., \bar{r}_{in}\}$, we define its current MAP gap location $\bar{\gamma}_i$ as $\frac{1}{2}(\bar{r}_{im} + \bar{r}_{i(m+1)})$, in which $\bar{r}_{im}$ and $\bar{r}_{i(m+1)}$ refer to the m^{th} and $(m+1)^{th}$ largest MAP return estimates respectively. Given $\bar{\gamma}_i$, we define the current MAP per-arm gaps $\bar{\Gamma}_i = \{\bar{\gamma}_{i1}, ...\bar{\gamma}_{in}\}$ such that $\bar{\gamma}_{ij} = |\bar{r}_{ij} - \bar{\gamma}_i|$. We also refer to a bandit's true returns and gaps (R_i, Γ_i) as its parameters $\theta \in \Theta$, where Θ spans all bandits permitted by the prior.

We associate each bandit b_i with a confidence ρ_i, which should give the probability that its current top-m group (based on the MAP return estimates) is correct. Since this is impractical to compute exactly, we use a lower bound[3]. For a bandit b_i with current MAP return estimates $\bar{R}_i = \{\bar{r}_{i1}, ..., \bar{r}_{in}\}$, we compute this bound as follows:

$$\bar{\rho}_i = 1 - \sum_{j=1}^{n} 1 - \Phi\left(\frac{\sqrt{t_{ij}}\,|\bar{r}_{ij} - \bar{\gamma}_i|}{\bar{\sigma}_{ij}}\right) , \tag{3}$$

where $\bar{\sigma}_{ij} = \sqrt{\bar{r}_{ij}(1-\bar{r}_{ij})}$ is the current MAP estimate of the standard deviation of the return for arm a_{ij}, t_{ij} is the number of trials previously allocated to arm a_{ij}, $\bar{\gamma}_i$ is the MAP gap location derived from $\bar{R}_i$, and Φ is the CDF for a standard normal distribution. This bound uses a normal approximation to the posterior distribution of the return estimate for each arm and computes a union bound on the probability that all MAP return estimates are on the same side of $\bar{\gamma}_i$ as their true values. When (3) is negative, we define $\bar{\rho}_i = 0$.

The algorithms presented in this paper all sample from the current posterior over a bandit's returns and gaps (i.e. its parameters $\theta \in \Theta$) as follows: sample a return for each arm from its current Beta distribution, compute the gap location implied by the sampled returns, and compute the per-arm gaps using the sampled returns and the computed gap location.

3.2 Posterior Sampling and Its Merits

Posterior sampling, or randomized probability matching, is a flexible approach to sequential optimization problems drawing increasing interest from the theoretical and applied sides of machine learning [1, 11, 15, 5]. For bandit problems, posterior sampling policies π^p select arms as follows:

$$\pi^p(a_{ij}|H) \propto p\left(a_{ij} = \arg\max_{a_{kl}} f_\theta^H(a_{kl}) \,\middle|\, H\right) , \tag{4}$$

in which $\pi^p(a_{ij}|H)$ is the probability of π^p selecting a_{ij} given H, the trial history H records the outcomes of all previous trials, $\theta \in \Theta$ is an unobserved parameter specifying the distribution of the bandit's returns, and f_θ^H is any deterministic function with bounded range. The remaining component of any posterior sampling policy π^p is the conditional distribution $p(\theta|H)$, which describes the posterior over $\theta \in \Theta$ after observing the trials recorded in H. Alg. (1) gives the general form followed by posterior sampling algorithms.

While the f_θ^H used in (4) must be deterministic given particular values for θ and H, its use in posterior sampling induces a stochastic policy by virtue of our imperfect knowledge of θ, which we observe only through the trials recorded in

[3] The true (Bayesian) confidence for a bandit can be computed to arbitrary precision by repeatedly sampling from the joint posterior over its per-arm returns and observing the frequency with which its MAP top-m group appears as the top-m group among the sampled sets of returns.

H. Thus, while f_θ^H must be deterministic, its value for a particular arm a_{ij} given a particular history H is stochastic, with stochasticity provided by entropy in the posterior $p(\theta|H)$.

Algorithm 1. PostSample(f_θ^H, $p(\theta|H)$, H, T)

1: **for** $1 \leq t \leq T$:
3: Sample $\hat{\theta} \in \Theta$ from the posterior given by $p(\hat{\theta}|H)$
4: Let $\hat{a}_{ij}^* = \arg\max_{a_{kl}} f_{\hat{\theta}}^H(a_{kl})$
5: Pull arm $\hat{a}_{ij}^*$ and update H based on the outcome
6: **end for**

The performance of a posterior sampling policy π^p is most naturally measured by its *Bayes risk* with respect to f_θ^H, which can be written as follows:

$$\mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - f_\theta^H(\pi_t^p) \right] \ , \tag{5}$$

in which π_t^p indicates an arm selected according to the probabilities given by $\pi^p(a_{ij}|H)$ and a_{ij}^* is an arm which maximizes f_θ^H. The Bayes risk describes the sub-optimality of π^p with respect to an optimal policy π^* that always knows a_{ij}^*, with respect to a prior over Θ chosen a priori. Based on work in [14], we decompose the Bayes risk for posterior sampling policies as follows:

$$\begin{aligned}
(5) &= \mathbb{E}_H \mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - f_\theta^H(\pi_t^p) \right] \\
&= \mathbb{E}_H \mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - U_t^H(\pi_t^p) + U_t^H(\pi_t^p) - f_\theta^H(\pi_t^p) \right] \\
&= \mathbb{E}_H \mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - U_t^H(a_{ij}^*) + U_t^H(\pi_t^p) - f_\theta^H(\pi_t^p) \right] \\
&= \mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - U_t^H(a_{ij}^*) \right] + \mathbb{E}_\theta \sum_{t=1}^{T} \left[U_t^H(\pi_t^p) - f_\theta^H(\pi_t^p) \right]
\end{aligned}$$

in which U_t^H is any function that is deterministic and bounded given H. The key step in this decomposition relies on the property that $\mathbb{E}_{\theta|H}[U_t^H(a_{ij}^*)] = \mathbb{E}_{\theta|H}[U_t^H(\pi_t^p)]$, which results from the posterior sampling construction of π^p according to (4), which makes the distributions $\pi^p(a_{ij}|H)$ and $p(a_{ij} = a_{ij}^*|H)$ identical. We emphasize that this decomposition is valid for any π^p based on posterior sampling for any f_θ^H and U_t^H meeting the stated contraints.

Analyses of the Bayes risk for UCB policies follow a decomposition parallel to that for posterior sampling, with a final step that results in:

$$\mathbb{E}_\theta \sum_{t=1}^{T} \left[f_\theta^H(a_{ij}^*) - U_t^H(a_{ij}^*) \right] + \mathbb{E}_\theta \sum_{t=1}^{T} \left[U_t^H(\pi_t^u) - f_\theta^H(\pi_t^u) \right] \ ,$$

in which U_t^H meets the same constraints as for posterior sampling and π_t^u is the arm selected by a UCB policy π^u based on U_t^H, i.e. one where $\pi_t^u = \arg\max_{a_{ij}} U_t^H(a_{ij})$. The key step in the Bayes risk decomposition for UCB policies relies on the fact that $U_t^H(\pi_t^u) \geq U_t^H(a_{ij}^*)$ for all t, due to the UCB construction of π^u.

The parallel decompositions of the Bayes risks for posterior sampling and UCB algorithms show that, if for some f_θ^H there exists an upper bound U_t^H which produces a UCB policy π^u with provably good Bayes risk, then substituting that U_t^H into the decomposed Bayes risk for the policy π^p which performs posterior sampling with respect to f_θ^H proves an equivalent Bayes risk for π^p. Thus, the Bayes risk of posterior sampling with respect to any f_θ^H is upper-bounded by the lowest Bayes risk upper bound for any π^u constructed from any upper bound U_t^H on f_θ^H. For detailed coverage of this result and its implications, see [14].

3.3 Top-m Selection via Posterior Sampling

Motivated by the preceding result, we derive a function f_θ^H for which good Bayes risk ensures good subset selection performance. We begin by restating an efficient static allocation policy π^s for subset selection described in detail by [8]:

$$\pi_\theta^s(a_{ij}) = \frac{Tb^2}{\gamma_{ij}^2 \sum_{kl} \frac{b^2}{\gamma_{kl}^2}} \ , \tag{6}$$

in which $\pi_\theta^s(a_{ij})$ gives the number of trials to allocate to a_{ij} assuming the gaps γ_{ij} are known a priori (the gaps are determined by the bandit parameters θ), T gives the total number of trials to allocate, and b is a bound on the range of the returns (e.g., $b = 1$ for Bernoulli bandits). The policy induced by (6) is optimal with respect to a lower bound on selection confidence analogous to that in (3). Next, for any trial history H, define $H(a_{ij})$ as the number of trials recorded for a_{ij} in H[4]. Finally, for history H and bandit parameters θ, define the log misallocation ratio as:

$$f_\theta^H(a_{ij}) = \log\left(\frac{\pi_\theta^s(a_{ij})}{H(a_{ij})}\right) \ . \tag{7}$$

Note that this f_θ^H implicitly depends on the desired subset size m through the definition of the per-arm gaps used in computing $\pi_\theta^s(a_{ij})$ for each arm and that it is bounded by $\pm\log(T)$. This f_θ^H provides a particularly interesting target for posterior sampling because we only ever observe it indirectly, through the information recorded in H over the course of an experiment.

Intuitively, posterior sampling with respect to (7) will select arms in proportion to their posterior probability of being most under-sampled relative to their sample density in the optimal static policy π_θ^s. Any policy whose Bayes risk with respect to (7) grows sublinearly in T has performance asymptotically equivalent

[4] Without loss of generality, we assume all arms have at least one trial in H.

to that of π^s_θ for the true θ as $T \to \infty$. And, from the earlier result, the existence of any UCB policy with good Bayes risk with respect to (7) suggests good Bayes risk for posterior sampling with respect to (7).

We perform intra-bandit top-m selection by posterior sampling with respect to the value in (7). Alg. (2) describes how our algorithm allocates trials at each round. While our full approach to greedy confidence pursuit calls Alg. (2) one round at a time, it can also be iterated following the form of Alg. (1) for application to single bandit subset selection problems.

Algorithm 2. SelectArm(bandit b_i, trial history H)

1: Sample $\hat{\theta}_i = (\hat{R}_i, \hat{\Gamma}_i)$ according to $p(\hat{\theta}_i|H)$.
2: Compute $\pi^s_{\hat{\theta}_i}(a_{ij})$ for each $a_{ij} \in A_i$ according to (6).
3: Compute $\hat{a}^*_{ij} = \arg\max_{a_{ij}} f^H_{\hat{\theta}_i}(a_{ij})$, with $f^H_{\hat{\theta}_i}$ as in (7).
4: Return $\hat{a}^*_{ij}$.

As further justification for our algorithm, consider the relation:

$$\arg\max_{a_{ij}} \log\left(\frac{\pi^s_\theta(a_{ij})}{H(a_{ij})}\right) = \arg\min_{a_{ij}} \frac{\sqrt{H(a_{ij})}\gamma_{ij}}{b} \ , \tag{8}$$

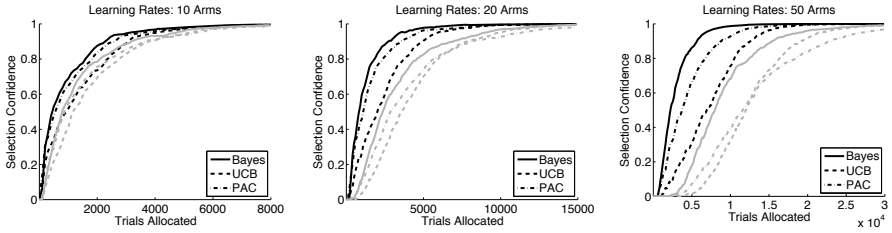
which follows from a straightforward derivation. If one were to model all arms using the same bound b on their standard deviation, then the values in the argmin above are equivalent to the values passed to Φ in (3) when computing the contribution of each arm to a bandit's confidence ρ_i. Thus, by posterior sampling with respect to (7), our algorithm selects arms according to their posterior probability of having the lowest confidence in (3). This can be interpreted as stochastic greedy maximization of the following lower bound on ρ_i:

$$\rho_i \geq 1 - n\left(1 - \min_j \left[\Phi\left(\frac{\sqrt{t_{ij}}\,|\bar{r}_{ij} - \bar{\gamma}_i|}{\bar{\sigma}_{ij}}\right)\right]\right) \ . \tag{9}$$

4 Testing Top-m Selection

This section empirically compares our subset selection algorithm with two existing methods. The first one [10] offers a standard PAC guarantee on sample complexity and success probability that matches a theoretical lower bound on the optimal samples/accuracy tradeoff (up to constant factors). The second method is based on the optimally efficient (up to constant factors) method for best arm selection presented in [8], which we adapt for use in subset selection. We refer to our arm selection method as Bayes and the respective baseline methods as PAC and UCB. We now describe the PAC and UCB methods as used in our tests.

Using the notation from the previous section, both the PAC and UCB methods rely primarily on the current MAP estimates of the gaps (i.e. $\{\bar{\gamma}_{i1}, ..., \bar{\gamma}_{in}\}$) for



(a) 10 arms (b) 20 arms (c) 50 arms

Fig. 1. These plots show average confidence lower bounds as a function of trials allocated for three different arm selection methods and two subset sizes at each of three arm counts. To generate each line, confidence lower bounds were averaged over 100 tests using bandits generated as described in the main text. Methods are indicated by line style. In each subfigure, the darker lines correspond to selecting the best arm and the lighter lines correspond to selecting the top half of the arms.

each of the arms in bandit b_i. Both methods allocate the next trial to an arm a_{ij} such that $-\bar{\gamma}_{ij} + \beta_{ij} = \max_k[-\bar{\gamma}_{ik} + \beta_{ik}]$, in which the negative gap $-\bar{\gamma}_{ij}$ encourages a focus on arms near the boundary and the term β_{ij} encourages exploration to improve the per-arm gap estimates. The PAC and UCB methods differ only in their computation of the β_{ij} term.

The PAC method, referred to in [10] as LUCB1, computes β_{ij} as follows:

$$\beta_{ij} = \sqrt{\frac{1}{2t_{ij}} \ln\left(\frac{5nt^4}{4\delta}\right)}\,, \tag{10}$$

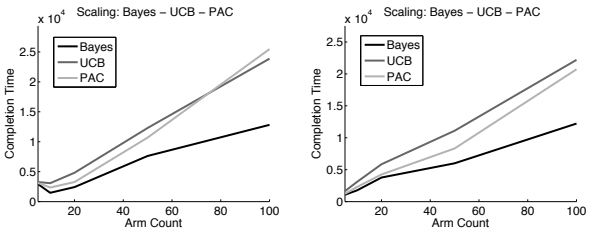
in which t_{ij} gives the number of trials previously allocated to a_{ij}, t gives the total number of trials previously allocated, n is the number of bandit arms, and $(1-\delta)$ is the desired probability of correct subset selection (we set $\delta = 0.05$ in our tests). UCB computes β_{ij} as follows:

$$\beta_{ij} = \sqrt{\frac{2\kappa_i \bar{\sigma}_{ij}^2}{t_{ij}}} + \frac{7\kappa_i \nu_i}{3(t_{ij}-1)}\,, \tag{11}$$

in which $\bar{\sigma}_{ij}^2$ is the current empirical (i.e. MAP) estimate of the variance of the return of a_{ij}, t_{ij} is as in (10), and κ_i/ν_i are constants computed from continuously updated empirical estimates of the complexity of bandit b_i. A full description of the κ_i/ν_i computations is beyond the scope of this paper and appears in [8][5].

All tests underlying Figures 1 and 2 used bandits with return distributions generated by the same process. Four parameters determined the return distribution of each bandit used in these tests: the minimum allowed gap γ_{min}, the

[5] For those familiar with the source material, we have implemented AGapE-V with the per-arm gaps Δ_{mk} redefined to permit top-m selection. This redefinition of the gaps permits simpler notation, while effecting only a constant shift in all gap values, thus leaving the selection process unchanged when $m = 1$.



(a) Instant (b) Delayed

Fig. 2. This plot compares best arm selection performance of the Bayes, PAC, and UCB algorithms. The lines show the median completion times achieved by each method over 100 tests at each arm count in $\{5, 10, 20, 50, 100\}$, with bandits generated as described in the text. Tests were considered complete when a confidence ≥ 0.98 was maintained for at least 100 rounds. Feedback in (a) was instant, while feedback in (b) was delayed 100 trials.

maximum allowed gap γ_{max}, the number of arms n, and the number of top arms to select m. Without loss of generality, we assume that the arms are sorted in order of descending returns. We generated a random set of returns meeting the constraints imposed by these parameters by generating sets of n returns uniformly distributed over $[0.1...0.9]$ until the gap between the m^{th} and $(m+1)^{th}$ largest returns was in the range $[\gamma_{min}...\gamma_{max}]$. For the tests in this section, and those in the remaining sections, for a given set of per-arm returns (i.e. a bandit), we presented each algorithm with matching sequences of trial outcomes. This allowed us to expose all methods tested to problems of equivalent difficulty. For tests in this section we set $\gamma_{min} = 0.05$ and $\gamma_{max} = 0.15$.

In Figure 1 we show the results of running the Bayes, PAC, and UCB methods on bandits with various arm counts when selecting either the best arm or the top half of the arms. We plot the average learning curves over 100 bandits for each arm count/subset size pair. The confidence values plotted in these curves were computed according to (3). Confidence curves for all other tests in this paper were computed similarly. The tests in Figure 1 show our method consistently outperforming existing methods over all arm counts and subset sizes.

Figure 2 compares Bayes, PAC, and UCB methods across a larger range of arm counts, in the context of best arm selection. For these tests, we compute *completion time* as the first round at which the confidence bound $\bar{\rho}_i$ was at least 0.98 for the previous 100 trials. Our method clearly has a large advantage as the number of arms increases. While the absolute advantage at arm counts ≤ 10 is smaller, it still represents a $10\% - 20\%$ reduction in completion time.

5 Bayesian Greedy Confidence Pursuit

Recall that, if the completion cost c_i for each bandit b_i were deterministic and known a priori, an optimal policy for greedy confidence pursuit would be to complete bandits in order of increasing completion costs, until budget exhaustion.

To compensate for the uncertain completion costs encountered in practical scenarios, we address greedy confidence pursuit by posterior sampling with respect to an approximate per-bandit completion cost.

For use in greedy confidence pursuit, an approximate completion cost need only predict the relative ranking of a set of bandits in terms of their true completion costs, as this permits mimicking the optimal greedy policy for known completion costs, which depends only on the cost-induced bandit order. For a bandit b_i with true returns R_i and gaps Γ_i, we use the following cost estimate:

$$\hat{c}_i = \sum_{j=1}^{n} \frac{\left(\sigma_{ij} + \sqrt{\sigma_{ij}^2 + (16/3)\gamma_{ij}}\right)^2}{\gamma_{ij}^2} \,, \tag{12}$$

in which σ_{ij} is the standard deviation associated with the return r_{ij}. The value in (12) comes from a bandit complexity measure described in [8]. Figure 3 supports the predictiveness of (12) with respect to relative empirical costs.

Note that, if one assumes the same target confidence τ for all bandits $b_i \in B$, then accounting for τ in $\hat{c}_i$ would not affect the ordering of bandits according to $\hat{c}_i$, as an "easier" bandit according to (12) would also have a smaller expected completion cost for any value of τ. By using (12), we also ignore the effort previously expended on a given bandit. While considering the number of trials already spent on a bandit could improve on the performance of (12), it would require steps to avoid the "sunk-cost" fallacy of economics, as manifested by premature commitment to bandits wrongly identified as "easy".

We perform bandit selection for greedy confidence pursuit by (minimum) posterior sampling with respect to $f_\theta^H(b_i) = \hat{c}_i$. The resulting algorithm is given in Alg. (3). Note that (12) captures dependence on the subset size m through its use of Γ_i and that $\hat{c}_i$ becomes stochastic when sampled with respect to the

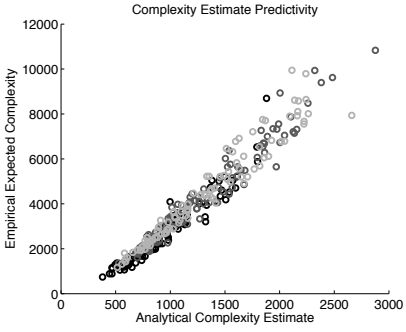


Fig. 3. This figure examines the predictive power of the completion cost in (12). From darkest to lightest the points represent selecting the top 1, 3, and 5 arms of a 10-armed bandit. Points correspond to particular bandits for which 20 runs of our Bayesian subset selection were performed, using independently generated trial outcomes for each run. The x coordinate of each point is the value of (12) for the true returns and gaps underlying its runs, while the y coordinate is the mean completion time for its runs.

per-bandit posteriors over returns and gaps. A theoretical analysis of our allocation process is beyond the scope of this paper, but the properties of posterior sampling described in Section 3 suggest it will efficiently direct trials towards the bandits with minimal $\hat{c}_i$. Section 6 empirically supports the design of this approach.

Algorithm 3. SelectBandit(bandit set B, trial history H)

1: **for each** $b_i \in B$:
2: Sample $\hat{\theta}_i = (\hat{R}_i, \hat{\Gamma}_i)$ according to $p(\hat{\theta}_i|H)$.
3: Compute $\hat{c}_i$ according to (12) using $\hat{R}_i$ and $\hat{\Gamma}_i$.
4: **end for**
5: Let $b_i^* = \arg\min_{b_i:\, \rho_i < \tau} \hat{c}_i$.
6: Return b_{i^*}.

5.1 Greedy Confidence Pursuit for "Targeted" Tasks

Now, consider the following problem:

- Given a finite trial budget T and N bandits b_i with returns $R_i = \{r_{i1}, ...r_{in}\}$
- Maximize the number of bandits b_i for which we can say with confidence ρ_i greater than τ that (without loss of generality) $r_{i1} > \max_{j \neq 1} r_{ij}$,

which reformulates the example scenario from Section 2. The twist in this scenario is that we only care about bandits for which a specific arm is best.

We address this problem by extending our algorithm for bandit selection in general greedy confidence pursuit. Intuitively, we sample bandits in proportion to their probability of having the lowest completion cost among bandits in which the targeted arm is best. Alg. (4) describes our extension of Alg. (3).

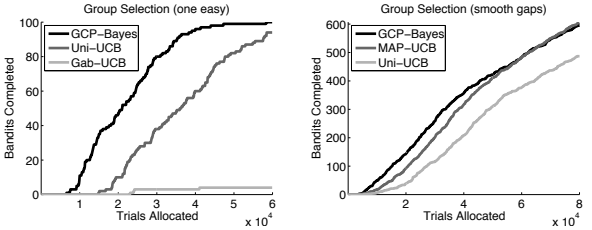
Algorithm 4. TargetedBanditSelection(bandit set B, trial history H)

01: $\forall i$, set $\hat{c}_i = \infty$.
02: **while** $(\min_i \hat{c}_i == \infty)$
03: **for each** $b_i \in B$:
04: Sample $\hat{\theta}_i = (\hat{R}_i, \hat{\Gamma}_i)$ according to $p(\hat{\theta}_i|H)$.
05: Compute $\hat{c}_i$ according to (12) using $\hat{R}_i$ and $\hat{\Gamma}_i$.
06: If $\hat{r}_{i1} < \max_{j \neq 1} \hat{r}_{ij}$, set $\hat{c}_i = \infty$
08: **end for**
09: **end while**
10: Let $b_i^* = \arg\min_{b_i:\, \rho_i < \tau} \hat{c}_i$.
11: Return b_{i^*}.

In the next section, we empirically support the value of Alg. (4) in situations where one is focused on maximizing "positive" results involving specific arms. For practical reasons, we upper bound the number of runs through the "resampling" loop of lines $02 - 09$. If, prior to reaching the upper bound, no bandit has been found for which $\hat{r}_{i1} > \max_{j \neq 1} \hat{r}_{ij}$, we select a bandit according to Alg. (3).

6 Testing Greedy Confidence Pursuit

We begin our empirical examination of greedy confidence pursuit with tests supporting (12) as an approximate completion cost. These tests were based on selecting the top 1, 3, and 5 arms of 10-armed bandits, with returns distributed as in Section 4. For each test, we generated a bandit, computed its cost according to (12) using the true returns, and then ran our Bayes arm selection 20 times on the bandit, using independently simulated trials for each run. Each point in Figure 3 corresponds to the analytically computed cost and the empirical expected cost for a particular bandit, with empirical completion costs measured as for Figure 2. These tests show that the cost estimates given by (12) are highly predictive with respect to the behavior of our algorithm.



(a) One easy group (b) Graded complexities

Fig. 4. This figure examines the performance of our method for greedy confidence pursuit, when selecting best arms. Curves in (a) were computed over 100 tests, each of which used 20 10-armed bandits, with the gap for one bandit set to 0.1 and the remaining gaps set to 0.01. Curves in (b) were computed over 100 tests, each of which used 15 10-armed bandits, with the gaps for the bandits evenly spaced on a log scale from 0.01 to 0.1. Bandit generation for the tests in (a) and (b) is described in the text. The curves in (a) and (b) show the number of bandits confidently completed prior to a given trial, aggregated across the relevant tests, with completion defined as for Fig. 2.

For the tests underlying Figures 4 and 5, the per-bandit objective was best arm selection. These tests compared our method for greedy confidence pursuit (tag: GCP-Bayes), comprising the bandit selection described in Section 5 and the arm selection described in Section 3, to three baseline methods. The first baseline was Uni-UCB, which uniformly selected bandits and then applied the UCB arm selection described in Section 4. The second baseline was Gab-UCB, which used UCB arm selection applied jointly over the bandits as described for GapE-V in [8][6]. The final baseline was provided by MAP-UCB, which selected bandits stochastically in inverse proportion to estimates of their completion costs computed by plugging MAP estimates of the relevant values into (12), and then used UCB for intra-bandit arm selection.

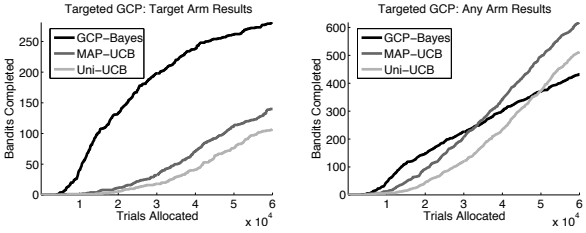
[6] Note that Gab-UCB is designed to optimize (1) rather than (2). By selecting jointly over all arms/bandits, our Bayesian approach to top-m selection can also be applied towards (1).

Figure 4 examines whether our approach to greedy confidence pursuit can improve the rate at which confident results are achieved. In each test underlying (a), 20 bandits were generated such that one had gap 0.1 and the rest had gap 0.01. For each test underlying (b), 15 bandits were generated to have gaps evenly spaced on a logarithmic scale over [0.01...0.10]. Given the desired gap size γ for each bandit, the best arm was set to return $0.5 + \gamma$, and the remaining returns were set uniformly at random in [0.0...0.5] and then uniformly shifted such that the second best arm had return 0.5. The curves in Figure 4 show the cumulative confident results achieved by each method prior to a given trial, computed based on 100 independently generated sets of test bandits for both (a) and (b).

Overall, Figure 4 shows that, in comparison to Uni-UCB and Gab-UCB, our method significantly accelerates the achievement of confident results. The tests in (a) show that Gab-UCB, which optimizes the objective described in (1), performs poorly with respect to the rate at which confident results are achieved when the bandits under consideration span a wide range of costs. The tests in (b) show that GCP-Bayes and MAP-UCB both maintain a large performance advantage over Uni-UCB even when the difference between easy and hard bandits is less pronounced than for the tests in (a). Note that MAP-UCB is a novel algorithm which we have introduced to provide non-trivial competition for GCP-Bayes.

6.1 Testing "Targeted" Greedy Confidence Pursuit

Figure 5 examines the performance of our approach to group selection for greedy confidence pursuit in the context of the targeted scenario from Section 2. In each test underlying the plots, 20 10-armed bandits were generated with gaps distributed uniformly at random over [0.01...0.10]. In each test, 5 bandits had their target arm best and the other 15 bandits had some other arm best. Given the gap size and best arm index for each bandit, the per-arm returns were set as for the tests underlying Figure 4.



(a) First arm groups (b) Any arm groups

Fig. 5. This figure gives two views of the cumulative number of best arms confidently selected prior to a given trial, similar to Figure 4. Each of the 100 tests on which these plots are based used 20 10-armed bandits, of which 5 had the target arm best while the remaining 15 had some other arm best. Gaps for all bandits were set uniformly at random in [0.01...0.10]. Bandit generation for these tests is described in the text. Plot (a) shows the cumulative number of bandits completed among those whose target arm was best, while (b) shows cumulative completions among all bandits.

The curves in Figure 5 show the rate at which each considered method achieved confident results, as measured by the number of bandits confidently completed prior to a given round, aggregated over 100 independently generated sets of bandits. For (a), only completed bandits among those with their target arm best were considered when computing the plotted curves. For (b), all completed bandits were considered when computing the plotted curves.

The curves in (a) show that, in comparison to both Uni-UCB and MAP-UCB, the targeted version of GCP-Bayes from Section 5.1 dramatically increases the rate at which confident results are achieved among bandits with their target arm best. The curves in (b) show that the increased focus of this version of GCP-Bayes on a particular subset of the bandits also modestly increases the initial rate at which confident results are achieved among all bandits, but that this early advantage fades as easy target-arm-best results are exhausted. After completing the easiest target-arm-best results, GCP-Bayes falls behind MAP-UCB, which greedily and impartially pursues all easy results.

7 Conclusion and Future Work

We presented a new multi-bandit optimization objective, called greedy confidence pursuit, which captures the general problem of maximizing the number of significant results achieved among a set of experiments sharing a finite pool of fungible resources. We derived algorithms for optimizing this objective in the context of top-m arm identification, both for single and multiple bandits. Our methods compare favorably to existing UCB-style algorithms in terms of empirical performance. In particular, for subset selection, our method scales much better with increasing arm counts than existing algorithms, which suggests its applicability in domains frequently involving numerous actions, such as online advertising and Monte-Carlo tree search for games with high branching factors.

While we used Bernoulli bandits in this paper, our methods directly extend to other return types e.g. normally-distributed continuous returns, through a simple change of priors. Structured priors, e.g. Gaussian processes, can also be used to capture both inter-bandit and intra-bandit relationships between returns. We used bandits with homogenous arm counts, but our methods handle heterogenous arm counts with no changes. With minor modifications, our methods can be used with bandits that share arms and for tasks other than subset selection, e.g. estimating quantiles or rank-ordering all returns. For practical applications, it may also be useful to account for variability in the value of completing each bandit. Such extensions are beyond the scope of the current paper, but provide rich material for future work. We gave one brief illustration of the flexibility granted by our use of posterior sampling by transforming Alg. (3) into Alg. (4), for application to problems in which only specific confident results are pursued.

References

[1] Agrawal, S., Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. In: COLT (2012)
[2] Audibert, J.-Y., Bubeck, S., Munos, R.: Best arm identification in multi-armed bandits. In: COLT (2010)
[3] Berry, D.A., Fristedt, B.: Bandit Problems. Chapman and Hall Ltd. (1985)
[4] Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in multi-armed bandits problems. In: Gavaldà, R., Lugosi, G., Zeugmann, T., Zilles, S. (eds.) ALT 2009. LNCS, vol. 5809, pp. 23–37. Springer, Heidelberg (2009)
[5] Chappelle, O., Li, L.: An empirical evaluation of thompson sampling. In: Advances in Neural Information Processing Systems (2011)
[6] Deng, K., Pineau, J., Murphy, S.: Active learning for personalizing treatment. In: IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (2011)
[7] Even-Dar, E., Mannor, S., Mansour, Y.: Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of Machine Learning Research 7, 1079–1105 (2006)
[8] Gabillon, V., Ghavamzadeh, M., Lazaric, A., Bubeck, S.: Multi-bandit best arm identification. In: Advances in Neural Information Processing Systems (2011)
[9] Kalyanakrishnan, S., Stone, P.: Efficient selection of multiple bandit arms: Theory and practice. In: International Conference on Machine Learning (2010)
[10] Kalyanakrishnan, S., Tewari, A., Auer, P., Stone, P.: Pac subset selection in stochastic multi-armed bandits. In: International Conference on Machine Learning (2012)
[11] Li, L., Chappelle, O.: Open problem: Regret bounds for thompson sampling. In: COLT (2012)
[12] Madani, O., Lizotte, D.J., Greiner, R.: The budgeted multi-armed bandit problem. In: COLT (2004)
[13] Mannor, S., Tsitsiklis, J.N.: The sample complexity of exploration in the multi-armed bandit problem. Journal of Machine Learning Research 5, 623–648 (2004)
[14] Russo, D., Van Roy, B.: Learning to optimize via posterior sampling. arXiv:1301.2609v1 [cs.LG] (2013)
[15] Scott, S.L.: A modern bayesian look at the multi-armed bandit. Applied Stochastic Models in Business and Industry 26, 639–658 (2010)
[16] Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25(3-4), 285–294 (1933)