Identifying Good Arms Fast and With Confidence: Strategies and Empirical Insights

Alicia Curth (*University of Cambridge)* amc253@cam.ac.uk

Alihan Hüyük (*University of Cambridge)* ah2075@cam.ac.uk

Mihaela van der Schaar (*University of Cambridge, UCLA, The ATI)* mv472@cam.ac.uk

Abstract

We investigate a *good* arm identification problem in which the goal is to discover as many arms above a threshold as fast as possible, subject to constraints on *both* confidence in each individual discovery *and* total budget. This problem resembles existing fixed confidence and fixed budget settings (which are usually considered *separately*) and aim for discovery of *all* good arms, but requires different considerations when designing learning algorithms because the combined constraint necessitates an early focus on *the most promising* arms (as not *all* arms need to be classified in our setting). We consider two types of possible solutions: successive discovery strategies for identifying *individual good arms*, and successive elimination strategies for identifying bundles of arms that are good *on average.* We empirically investigate their performance and provide insight into the (dis)advantages of these strategies as well as different components that can be used in their implementation.

1. Introduction

Consider a setting in which a player, given a set of options, aims to adaptively identify as many *good* options as *fast* as possible – subject to both a constraint on *confidence* in every discovery as well as a total *budget* limit. Here, good options are those that exceed a given quality threshold; that is, chosen options do not necessarily need to be among the *best* as long as they are "good enough". Our interest in this setting was originally motivated by the goal of designing adaptive clinical trials to discover subgroups in which a drug *works* – in this context, any group with *any* effect is of interest[1], yet any discovery needs to be with confidence to control Type I error *and* budget (in terms of patients that can be recruited) is very limited. Nonetheless, we believe that this setting could be of general interest for many more applications, e.g. recruitment or portfolio management, where the goal is to either confidently identify a *single* good arm, candidate or other type of option as fast as possible, or to confidently identify as many of them as possible given a pre-specified budget – and therefore study this problem in generality here[1]. This problem fits perfectly within the literatures on thresholding bandits [3–6] or good arm identification [7–9], but differs with respect to previous work which usually focusses on discovering *all* good arms with *either*

[1] In a companion paper [2] (also presented at this workshop), we study the original clinical trial problem and considerations arising in this context in more detail. The learning strategies under investigation in both papers are similar, however, the present paper differs in that it (i) takes a more general 'good arm identification' perspective, (ii) focusses on understanding different implementation choices within the considered learning strategies (i.e. the sampling and elimination rules discussed in Sec. 3) while the previous paper fixes a single choice and (iii) contains a new and extensive empirical investigation into the effects of those choices.

fixed confidence *or* fixed budget. We provide a more extensive comparison with related work and settings considered therein in Appendix A.

We hold that our problem – identifying good arms under a mixed fixed budget/fixed confidence constraint – is sufficiently different and interesting to be considered separately as it necessitates different considerations when designing learning algorithms: In the standard case where the ultimate goal is to identify *all* good options, performance is usually dominated by the last discoveries, i.e. those that are *hardest to make*, and the timing of the earliest discoveries is therefore not of primary interest. This changes in our setting: when budget is very limited or the goal is to make any single discovery as fast as possible, it instead becomes important to focus on the most promising options, i.e. those that can be proven to be good using the fewest samples possible. We therefore believe that investigating different learning strategies and their performance in this context is an interesting and novel addition to the literature on good arm identification.

Outlook. To tackle our problem, we consider adapting two types of strategies that have been used in the literature to solve related problem. On the one hand, we consider strategies that we will refer to as *successive discovery* strategies: these strategies successively sample individual arms according to some exploration rule, and after each sample make individual discoveries by testing whether it is possible to declare any individual arm as good/bad. When using upper confidence bound-based exploration (sampling) rules, this is similar to the good arm identification algorithms proposed in [7–9]. We consider alternative rules and provide insights into the strengths of different types of exploration rules in our experiments. On the other hand, we consider strategies that allow us to solve a slightly *relaxed* version of our problem: instead of demanding *individual* good arms to be identified, we now also accept bundles of arms (super-arms) that are *good on average*. This relaxation is inspired by clinical trials where a drug usually only has to be demonstrated to work across a population *on average*, but applies to other settings e.g. ensuring average quality of a workforce or average performance of a portfolio. In this case, it can be interesting to rely on *successive elimination* strategies, which sample *all* active arms (those that are still under consideration) at each time step and can subsequently either accept the complete active set of arms as good or remove individual arms that appear bad. In our experiments, we demonstrate when and how such elimination strategies can speed up identification as they allow to share statistical strength across arms.

2. Problem setting

We consider a setting with K arms (options) and a player that at each time t chooses an arm $J_t \in \mathcal{K}$ to observe reward $X_t \sim \nu_{J_t}$ from, where all X_t are subgaussian random variables with mean $\mu_{J_t} = \mathbb{E}_{\nu_{J_t}}[X_t]$. An arm j is considered good if $\mu_j > \mu_0$, where μ_0 is some prespecified threshold. Given prespecified weights for arms $\pi_j \in [0,1]$ with $\sum_{j=1}^{K} \pi_j = 1$, we also consider super-arms (sets/bundles of arms) $\mathcal{S} \subseteq \mathcal{K}$, which are considered good if $\mu_{\mathcal{S}} = \sum_{j\in\mathcal{S}} \frac{\pi_j \mu_j}{\sum_{j\in\mathcal{S}} \pi_j} > \mu_0$. Throughout, we assume $\pi_j = \frac{1}{K}$ so that $\mu_{\mathcal{S}} = |\mathcal{S}|^{-1} \sum_{j\in\mathcal{S}} \mu_j$.

The ultimate goal of the player is to identify the set $\mathcal{S}_{good} = \{j : \mu_j > \mu_0\}$ with high confidence, but budget is limited to B rounds of sampling. In the strict version of the problem, we therefore aim to discover the largest set of arms that can be individually declared good with high confidence, i.e. using at most B samples, find a set $\hat{S}^{strict} = \{j :$

$\mathbb{P}(\mu_j \leq \mu_0) \leq \alpha\}$. In the relaxed version of the problem, we instead aim to find the largest good super-arm, which is good *on average* with high confidence, i.e. a set $\hat{S}^{relax} \subseteq \mathcal{K}$ s.t. $\mathbb{P}(\mu_{\mathcal{S}} \leq \mu_0) \leq \alpha$. It is well-known that the number of samples needed to distinguish μ_j from μ_0 scales as $(\mu_j - \mu_0)^{-2}$, which can get very large as μ_j approaches the threshold μ_0 [7], making it disproportionally difficult to classify arms with very small distance to the threshold. Inspired by common practice in clinical trials [10], we therefore introduce a *minimum relevance threshold* $\mu_{min} > \mu_0$ (to be set depending on the application under consideration) and allow arms to be declared *bad* when they lie below μ_{min} while ensuring that power to detect larger effects is preserved, i.e. $\mathbb{P}(\text{j is declared bad}|\mu_j = \mu_{min}) \leq \beta$.

Throughout, we denote by $N_j(t) = \sum_{t' \leq t} \mathbb{1}\{J_{t'} = j\}$ and $N_{\mathcal{S}}(t) = \sum_{t' \leq t} \mathbb{1}\{J_{t'} \in \mathcal{S}\}$ the number of times an arm and super-arm, respectively, has been played by time t, and estimate means of individual arms by $\hat{\mu}_{j,N_j(t)} = N_j(t)^{-1} \sum_{t' \leq t} \mathbb{1}\{J_{t'} = j\} X_{t'}$ and, if all arms have been sampled equally often, of super-arms by $\hat{\mu}_{\mathcal{S},N_{\mathcal{S}}(t)} = N_{\mathcal{S}}(t)^{-1} \sum_{t' \leq t} \mathbb{1}\{J_{t'} \in \mathcal{S}\} X_{t'}$, respectively. As [7], we assume access to confidence intervals $\phi(t, \delta)$ for which it holds that $\mathbb{P}(\cap_{t=1}^{\infty} \{|\hat{\mu}_{\mathcal{S},t} - \mu_{\mathcal{S}}| \leq \phi(t, \delta)\}) \geq 1 - \delta$ for $\delta \in (0, 1)$ and instantiate it using Thm. 8 of [11].

3. Learning strategies

In this section, we introduce and discuss two types of learning strategies that can be used to discover (i) sets of individually good arms and (ii) good super-arms, respectively. Note that solutions to (i) are also valid solutions to (ii), but the reverse is not necessarily true.

3.1 Successive discovery strategies

We first consider successive discovery strategies, versions of which have been used for good arm identification (GAI) with fixed confidence [7–9], originally inspired by upper confidence bound (UCB)-style algorithms for *best* arm identification [12, 13]. We consider the following general procedure a successive discovery strategy: at each time step t before budget B is exhausted, the algorithm (i) chooses an arm J_t from the unclassified active set $\mathcal{A}$ to sample from using an exploration rule $\mathcal{E}$, (ii) checks whether any arm i can be identified as good and removed from $\mathcal{A}$ because $\hat{\mu}_{i,N_i(t)} - \phi(N_i(t), \alpha) > \mu_0$ and (iii) checks whether any arm i can be discarded[2] and removed from $\mathcal{A}$ because $\hat{\mu}_{i,N_i(t)} + \phi(N_i(t), \beta) < \mu_{min}$.

Clearly, the exploration rule $\mathcal{E}$ used in such a successive discovery strategy will have a large impact on how fast the first arms will be discovered. In our experiments, we therefore investigate the use of different rules:

- $\mathcal{E}_{unif}$: The simplest sampling rule would be to uniformly sample any one of the active arms, ignoring any information that has accumulated.
- $\mathcal{E}_{UCB}$: The standard sampling rule in the GAI literature [7–9] seems to use an optimistic UCB approach, i.e. sample $\arg\max_{j \in \mathcal{A}} \hat{\mu}_{j,N_j(t-1)} + \phi(N_j(t-1), \alpha)$. This rule selects the arm that currently appears *best* (i.e. likely to have the highest mean), however, it does not necessarily *exploit* accumulated knowledge by repeatedly sampling an arm who is close to be identified as good; in fact, as $\phi(t, \delta)$ shrinks with increasing t, we suspect that $\mathcal{E}_{UCB}$ may encourage frequent switching between similar arms which

[2] The formulation presented here is more general than existing GAI algorithms, which either do not remove arms [7, 9] or use the same threshold & confidence for removal and discovery [8].

may lead to no identifications when budget is very limited. Therefore, we explore the use of two new sampling strategies for this problem.

- $\mathcal{E}_{LCB}$: As our identification criterion relies on a lower confidence bound (LCB), sampling the arm with the best LCB would correspond to selecting arms that appear most promising for early identification. Thus, we also consider using $\mathcal{E}_{LCB}$ which chooses $\arg\max_{j\in\mathcal{A}} \hat{\mu}_{j,N_j(t-1)} - \phi(N_j(t-1), \alpha)$. Again, as $\phi(t, \delta)$ shrinks in t, this strategy may conversely risk *getting stuck* on an arm that only appeared good early on.
- $\mathcal{E}_{LUCB}$: We therefore consider a final strategy $\mathcal{E}_{LUCB} = \mathcal{E}_{UCB} \cup \mathcal{E}_{LCB}$ which will take two consecutive samples whenever sampling according to UCB and LCB disagree.

3.2 Successive elimination strategies

Second, we consider successive elimination strategies, versions of which have been used for best arm identification, e.g. [12, 14], where they have empirically been shown to be very wasteful [13]. In our *relaxed* problem formulation that aims to only find a good *super-arm*, on the other hand, we believe that successive elimination strategies can be *more efficient* than successive discovery when allowing to share statistical strength (sample size) across all sampled arms. We consider the following general successive elimination strategy: while budget is not depleted, (i) sample each arm in the active set (super-arm) $\mathcal{A}$ once, (ii) test whether the active super-arm can be identified as good as $\hat{\mu}_{\mathcal{A},N_{\mathcal{A}}(t)} - \phi(N_{\mathcal{A}}(t), \alpha) > \mu_0$ and (iii) remove bad individual arms from the active set using a removal (elimination) rule $\mathcal{R}$.

Here, the removal (elimination) rule $\mathcal{R}$ will determine how the algorithm behaves. We consider two possibilities:

- $\mathcal{R}_{arm}$: As in the successive discovery strategy, we use an arm-based removal rule that checks for any arm that can be removed individually through $\hat{\mu}_{i,N_i(t)} + \phi(N_i(t), \beta) < \mu_{min}$. Note that successive elimination with $\mathcal{R}_{arm}$ is almost identical to successive discovery with $\mathcal{E}_{unif}$ and differs only in the identification rule used (arm-based vs super-arm-based).
- $\mathcal{R}_{super}$: In addition, we exploit that we could also make use of super-arm information also when removing arms. The event $\mathbb{1}\{\hat{\mu}_{\mathcal{A},N_{\mathcal{A}}(t)} + \phi(N_{\mathcal{A}}(t), \beta) < \mu_{min}\}$ provides evidence that *at least* one arm does not meet the minimum quality μ_{min}, so when this event occurs, we remove the empirically worst arm $\arg\min_{j\in\mathcal{A}} \hat{\mu}_{j,N_j(t)} - \phi(N_j(t), \alpha)$

4. Empirical Investigation: Understanding the (dis)advantages of different strategies

We consider a stylized simulation setup to gain insight into the (dis)advantages of the two strategies (Successive Discovery – SDisc – and Successive Elimination – SElim) and their different sampling and removal rules. We consider $K = 10$ arms and assume that we observe $X_t \sim \mathcal{N}(\mu_{J_t}, 1)$. As [7], we use $\phi(t, \delta) = \sqrt{2^{\frac{\log(1/\delta)+3\log\log(1/\delta)+(3/2)\log\log(et/2))}{t}}}$. We let $\mu_0 = 0$, $\mu_{min} = 0.5$, $\alpha = 0.05$ and $\beta = 0.1$. In the main results presented in Fig. 1, we let $\mu_j \in \{\mu_b, \mu_g\}$, where $\mu_b = 0$ and $\mu_g = 0.5$ unless stated otherwise, and vary $n_g = |\{j : \mu_j \geq 0\}|$. Throughout, we do not restrict budget and report t_{stop}, the stopping time of the algorithm (i.e. the time when *all* arms are classified as good or not), as well as $t_g^{id,j}$ and $t_b^{id,j}$, the time taken to identify the j^{th} good arm and to discard the j^{th} bad arm,

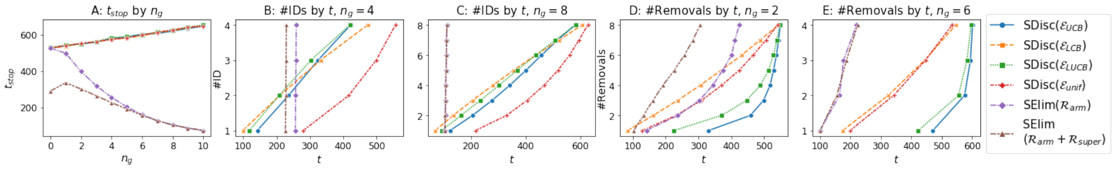


Figure 1: Results describing time until (A) termination, (B&C) identification of good arms and (D&E) removal of bad arms, avg. across 1000 replications. (A): Time to termination t_{stop} by number of good arms n_g. (B&C): Number of good arm identifications over time, for $n_g = 4$ (B) and $n_g = 8$ (C). (D&E): Number of removals of bad arms over time, for $n_g = 2$ (D) and $n_g = 6$ (E).

respectively; doing so allows us to understand what the algorithm would have identified given *any* budget. We discuss insights in turn below; additional results are in Appendix B.

Natural stopping times. In Fig. 1A, we investigate *how long* it would take the different algorithms to select/discard *all* arms for different n_g. First, we observe that the sampling strategy of SDisc has no impact on the stopping time; this is expected as identification of the final/worst arm determines t_{stop}. Second, the total time to termination *increases* as n_g increases for SDisc because the identification criterion is *stricter* than the removal criterion (i.e. $\beta > \alpha$). Third, SElim($\mathcal{R}_{arm}$), which is identical to SDisc($\mathcal{E}_{unif}$) except for the super-arm-based identification criterion, performs identically to SDisc when $n_g = 0$ but begins to terminate earlier when n_g increases as sample size can be shared across good arms. Finally, SElim($\mathcal{R}_{arm} + \mathcal{R}_{super}$) terminates fastest throughout, as it shares statistical strength across arms *both* when discarding and accepting arms; thus, the more homogeneous the arms (n_g close to 0 or 10) the faster it terminates.

Time to identify the j^{th} good arm. In Fig. 1B&C, we investigate *when* the different algorithms make *good* arm discoveries, for $n_g = 4, 8$. When comparing algorithms, we find that SDisc generally makes the *first* discovery before SElim, as SElim makes *all* discoveries at the same time (yet this often happens before SDisc even makes its second discovery). When comparing sampling strategies within SDisc, major differences become visible. (Non-adaptive) uniform sampling now clearly appears suboptimal; the first discovery happens much later than for adaptive sampling and subsequent discoveries happen much quicker after each other. Within the adaptive strategies, $\mathcal{E}_{LCB}$ indeed makes the first few discoveries faster than $\mathcal{E}_{UCB}$ in this setting, as the latter will unnecessarily switch between good arms as upper bounds cross (because the underlying good means are identical); as expected, $\mathcal{E}_{LUCB}$ lies inbetween.

If the good arms were to exhibit quantitatively very different effects, the arm with the largest μ_j should need least samples to be discovered – thus we would expect UCB-type strategies that haven proven successful in *best arm* identification [13] to be advantageous in this context. In Fig 2, we therefore further investigate the relative performance of sampling strategies when altering the underlying simulation: when the means in good arms are very different (Scen. 1: $\mu_1 = 0.5, \mu_2 = 1$; $\mu_j = 0, j > 2$) the relative performance indeed reverses. With more good arms and less spacing between means (Scen. 2: $\mu_j = 0.5 + \frac{0.5}{7}(j-1), j \leq 8$; $\mu_j = 0, j > 8$),

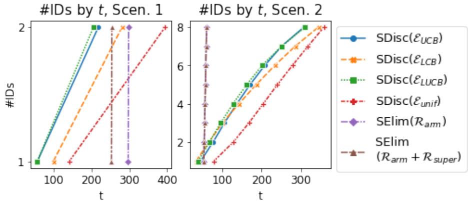


Figure 2: Good arm identifications over time for two additional scenarios.

this difference becomes less pronounced. In Appendix B, we additionally investigate how sampling strategies compare when outcome variance is known to differ across arms.

Time to discard the j^{th} bad arm. In Fig. 1D&E, we investigate when the different algorithms *discard* arms that do not appear good. First, we observe that, unsurprisingly, SElim – an algorithm operating by successive elimination – discards arms much faster than SDisc. Second, we observe that SElim($\mathcal{R}_{arm} + \mathcal{R}_{super}$) indeed benefits from the super-arm-based removal criterion as arms are discarded much faster especially when n_g small, which is when the super-arm-based removal criterion will be met earlier. Third, we note that uniform sampling leads to faster elimination than (L)UCB-based sampling, which is expected as the latter actively avoid sampling arms that appear bad. Perhaps more surprisingly, LCB sampling leads to similarly fast discarding of the first bad arms, which we attribute to LCB being more likely to continue sampling from a arm that has already been sampled often.

Incorrectly classified arms. Finally, we consider whether arms are (in)correctly classified as good. In Fig. 3A, we observe that good arms are seldomly missed by either algorithm (in fact, the rate lies far below the expected $\beta * n_g$, which we attribute to the used anytime confidence intervals being unnecessarily conservative as $t \ll \infty$ here); only SElim removes good arms slightly more often with the aggressive removal criterion $\mathcal{R}_{super}$. Further, in Fig. 3B, we observe interesting differences in arms below the threshold that are included in the selected set $\mathcal{S}$ (note: for SElim, this does *not* necessarily constitute an error as long as $\mu_{\mathcal{S}} > 0$). As SDisc identifies arms individually, $|\mathcal{S}| \approx n_g$ throughout, while SElim allows *free-riding* of arms with $\mu_j = \mu_b$ on the larger means of good arms, i.e. $|\mathcal{S}| > n_g$, especially when n_g is large, which leads to the super-arm mean being worse but still good (above μ_0). In Fig. 3C we set $\mu_b = -0.5$ instead of 0, and observe that this behavior decreases when arms contribute sufficiently large negative effects.

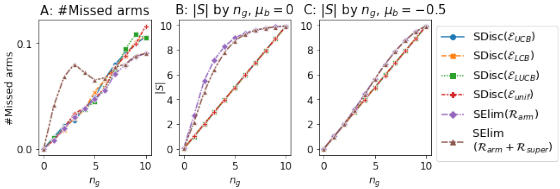


Figure 3: (A): Avg. number of missed arms by n_g. (B & C): Avg. $|\mathcal{S}|$ by n_g, for $\mu_b = 0, -0.5$.

5. Conclusion and Future Work

We investigated how to best identify many good arms fast in a setting where there are constraints on both confidence and budget, and presented empirical insights into (dis)advantages of successive discovery versus successive elimination strategies, as well as different exploration and removal rules that could be used in their implementation. We showed that the elimination-based strategy, which only discovers a good super-arm, generally terminates using fewer samples, but may include arms that are not truly good if other arms have a sufficiently high mean. Using successive discovery strategies, which discover individual arms, this can generally be avoided – if one is willing to use substantially more samples. Comparing sampling strategies for successive discovery, we found that there are scenarios where either UCB or LCB-sampling dominate, making LUCB-sampling a good intermediate choice. It would be an interesting next step to complement these empirical findings with theoretical ones, e.g. by theoretically investigating how the complexities of the strict and relaxed problem formulations compare or by theoretically characterizing the problem structures under which different exploration rules are expected to have an advantage.

References

[1] Christopher Jennison and Bruce W Turnbull. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of biopharmaceutical statistics*, 17(6):1135–1161, 2007.

[2] Alicia Curth, Alihan Hüyük, and Mihaela van der Schaar. Adaptively identifying good patient populations in clinical trials. *ICML workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022.

[3] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

[4] Jie Zhong, Yijun Huang, and Ji Liu. Asynchronous parallel empirical variance guided algorithms for the thresholding bandit problem. *arXiv preprint arXiv:1704.04567*, 2017.

[5] Chao Tao, Saúl Blanco, Jian Peng, and Yuan Zhou. Thresholding bandit with optimal aggregate regret. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Subhojyoti Mukherjee, Kolar Purushothama Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented ucb. *arXiv preprint arXiv:1704.02281*, 2017.

[7] Kevin G Jamieson and Lalit Jain. A bandit approach to sequential experimental design with false discovery control. *Advances in Neural Information Processing Systems*, 31, 2018.

[8] Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5):721–745, 2019.

[9] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR, 2020.

[10] Anne G Copay, Brian R Subach, Steven D Glassman, David W Polly Jr, and Thomas C Schuler. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal*, 7(5):541–546, 2007.

[11] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

[12] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

[13] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[14] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

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

[16] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[17] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[18] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

[19] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

[21] Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR, 2013.

[22] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[23] Oded Maron and Andrew Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. *Advances in neural information processing systems*, 6, 1993.

[24] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[25] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679, 2008.

[26] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: Theory and practice. In *ICML*, 2010.

[27] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[28] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[29] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[30] Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner, and Peter Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012. PMLR, 2016.

Appendix A. Extended literature review: Relationship to bandit literature

The prototypical goal in a bandit problem is to maximize the rewards of all arms that are played (e.g. [15]). Since the mean rewards are unknown initially, this requires striking a balance between *exploring* arms to gain information about their rewards and *exploiting* arms that appear to have high rewards. Instead, we focus here on a setting that is known as *pure exploration* in the bandit literature, where the rewards of played arms do not matter except for that of a singular arm identified at the end [16–19].

Different purely-exploratory objectives have been considered in the multi-armed bandit literature. Best arm identification (BAI) problems aim to identify the arm (or the top-K arms) with the largest mean reward (e.g. [12]). Here, the success can be measured via the reward gap between the identified arm and the true best arm. In the *fixed budget* setting, the goal is to maximize the probability of the identified arm indeed being the best given a fixed budget of samples [20–22], while in the *fixed confidence* setting, the goal is to minimize the number of samples necessary to guarantee a fixed level of confidence [20, 23–29]. Good arm identification (GAI) problems (sometimes called pure exploration in thresholding bandits) aim to identify arms with mean rewards that are higher than a pre-specified threshold. These problems too can be considered either in fixed budget [3, 6] or fixed confidence [8, 9] settings. In this literature, two dominant types of learning strategy seem to have emerged: the thresholding bandit solutions of e.g. [3, 6] focus on playing arms *close* to the threshold to increase confidence in classification of the hardest arms, while the literature titled good arm identification [8, 9] has relied on successive discovery strategies similar to what we describe in the main text. [9]'s successive discovery strategy additionally relies on a subsampling strategy (of the active set) to speed up discovery when there are many good arms; this improvement could be incorporated into the successive discovery strategy we consider and is orthogonal to our investigation of sampling rules.

Our strict problem formulation is thus essentially a type of GAI problem but it requires both the budget as well as the confidence in each identified arm being good to be fixed[3], and given those constraints, aims to identify as many good arms as possible. In existing formulations of GAI, the aim is usually to identify *all* good arms, which is only possible with the more relaxed constraint of either just the budget or the confidence being fixed (but not both at the same time). The relaxed version of our problem is similar in that it too requires both the budget and the confidence to be fixed but it only aims to identify a collection of arms that are good *on average* rather than arms that are all individually good.[4] Table 1 formally compares our strict and relaxed problems with existing pure exploration problems.

[3] In this paper, we consider a relatively weak requirement for error-control: we require *each* single discovery to be true at pre-specified confidence level α, which means that the total error can be as large as $K\alpha$ (strict problem) or $2^K\alpha$ (relaxed problem, due to combinatorial nature of super-arms). In our clinical trial focussed companion paper [2], we seek stricter control of the *Familywise Error Rate (FWER)* using Bonferroni-style adjustments.

[4] Our relaxed problem formulation could be seen as a generic *combinatorial bandit* problem [18, 30]; however, to the best of our knowledge no existing solutions exploit the idea of sharing statistical strength across arms by pooling samples and solutions derived from e.g. [18, 30] would therefore resemble standard GAI solutions.

Table 1: Comparison of pure exploration problems. The strict and relaxed verions of our problem uniquely require both the budget as well as the confidence to be fixed, and aim to identify as many suitable arms as possible within those constraints. In contrast, other problems aim to identify all suitable arms, which is only possible with the more relaxed constraint of either just the budget or just the confidence being fixed. FB and FC stand for fixed budget and fixed confidence respectively.

Problem	**Ref.**	**Type of arms identified**	**Number of arms identified**	**Budget**	**Confidence**	**Formulation**
BAI	[12]	Best arms $i^* = \text{argmax}_i\, \mu_i$	Top-K arms	Variable	Variable	minimize $\mu_{i^*} - \mu_{\hat{i}^*}$
BAI w/ FB	[22]			Fixed (T)	Maximized	maximize $\mathbb{P}(\hat{i}^*(T) = i^*)$
BAI w/ FC	[29]			Minimized	Fixed ($1-\delta$)	minimize T s.t. $\mathbb{P}(\hat{i}^*(T) \neq i^*) \leq \delta$
GAI w/ FB	[3]	Good arms $\mathcal{I} = \{i : \mu_i > \mu_0\}$	All good arms	Fixed (T)	Maximized	maximize $\mathbb{P}(\hat{\mathcal{I}}(T) = \mathcal{I})$
GAI w/ FC	[9]			Minimized	Fixed ($1-\delta$)	minimize T s.t. $\mathbb{P}(\hat{\mathcal{I}}(T) \neq \mathcal{I}) \leq \delta$
Strict	**(Ours)**	Good arms $\mathcal{I} = \{i : \mu_i > \mu_0\}$	Maximized	Fixed (T)	Fixed ($1-\delta$) w.r.t. single discovery	maximize $\lvert\hat{\mathcal{I}}(T)\rvert$ s.t. for $j \in \hat{\mathcal{I}}(T)$, $\mathbb{P}(\mu_j < \mu_0) \leq \delta$
Relaxed	**(Ours)**	Good super arms $\mathcal{I} : \frac{1}{\lvert\mathcal{I}\rvert} \sum_{i \in \mathcal{I}} \mu_i > \mu_0$	Maximized	Fixed (T)	Fixed ($1-\delta$)	maximize $\lvert\hat{\mathcal{I}}(T)\rvert$ s.t. $\mathbb{P}\Big(\frac{1}{\lvert\hat{\mathcal{I}}(T)\rvert} \sum_{i \in \hat{\mathcal{I}}(T)} \mu_i \ngtr \mu_0\Big) \leq \delta$

Appendix B. Additional Results

B.1 Identifications: Complete results

In Fig. 4, we present results capturing time until identification of each good arm for $n_g \in \{2, 4, 6, 8, 10\}$ (only $n_g = 4, 8$ are presented in the main text). In Fig. 5, we present results capturing time until removal of each bad arm for $n_g \in \{0, 2, 4, 6, 8\}$ (only $n_g = 2, 6$ are presented in the main text). These results reflect the same insights as those presented in the main text, both in terms of comparing algorithms and in terms of comparing sampling strategies.

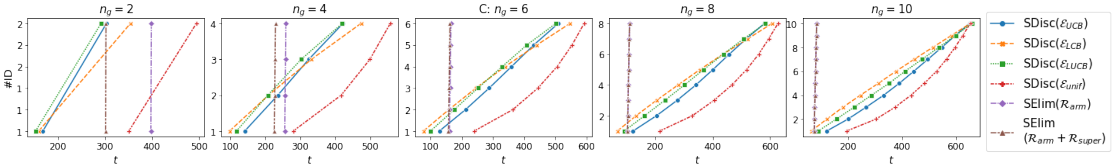


Figure 4: Results describing identification of good arms over time, for $n_g \in \{2, 4, 6, 8, 10\}$; avg. across 1000 replications.

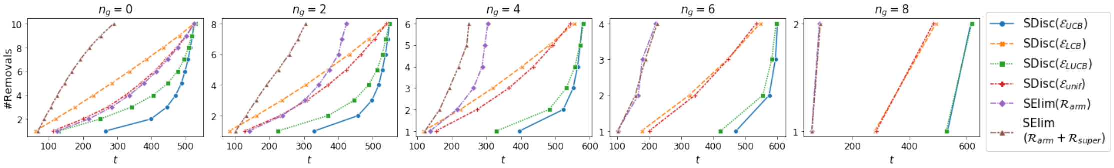


Figure 5: Results describing removal of bad arms over time, for $n_g \in \{0, 2, 4, 6, 8\}$; avg. across 1000 replications.

B.2 Additional simulation scenarios

Varying means We present additional results on the setting presented in Fig. 2 of the main text: for $n_g \in \{2, \ldots, 10\}$ we let $\mu_j = 0.5 + 0.5\frac{j-1}{n_g-1}$ for $j \leq n_g$ and $\mu_j = 0$ otherwise. As discussed in the main text, we observe that the relative performance of sampling strategies changes in this setting: $\mathcal{E}_{LCB}$ generally performs worse than $\mathcal{E}_{UCB}$ here; with increasing n_g and hence decreasing spacing between the good means, this effect reduces.

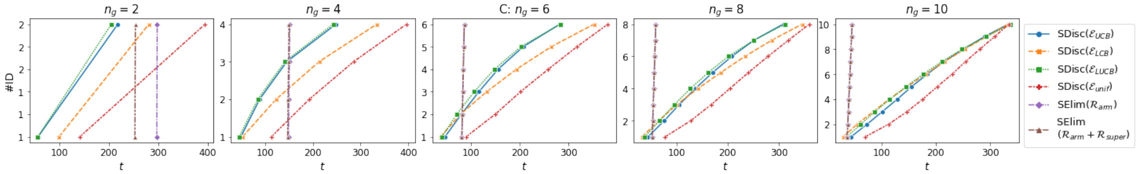
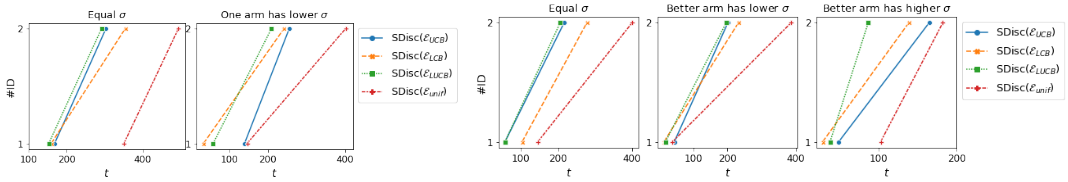


Figure 6: Results describing identification of good arms over time, for $n_g \in \{2, 4, 6, 8, 10\}$ for a setting with varying means; avg. across 1000 replications.

Different variances Finally, we consider how changing variance affects the performance of the different sampling algorithms[5]. In Fig. 7(a), we compare the original setting ($\mu_g = 0.5$) with $n_g = 2$ and $\sigma = 1$ for all arms to one where the means are the same but $\sigma_1 = 0.5$ and $\sigma_2 = 1$. In Fig. 7(b), we compare the setting with $n_g = 2$ and $\mu_1 = 0.5$ and $\mu_2 = 1$ with $\sigma = 1$ for all arms from the previous paragraph, to one where either the better arm ($j = 1$) or the worse arm ($j = 2$) has smaller variance $\sigma_j = 0.5$ than all other arms with $\sigma = 1$. We make a number of interesting observations: In the setting on the left hand side, where means are equal, we observe that $\mathcal{E}_{LCB}$ improves compared to $\mathcal{E}_{UCB}$ when variance differs, as it intrinsically makes use of the fact that arms with *lower* variance need less samples to be identified, while $\mathcal{E}_{UCB}$ may erroneously focus on arms with high variance (which have higher UCB). In the setting on the right hand side, where means differ, we observe that the advantage of $\mathcal{E}_{UCB}$ over $\mathcal{E}_{LCB}$ in discovering the first good arm observed in the constant variance setting essentially disappears as we let variance differ. This is expected when the better arm has a *lower* variance as the higher UCB of the worse arm may confuse $\mathcal{E}_{UCB}$ but $\mathcal{E}_{LCB}$ would sample the correct arm. Perhaps more surprisingly, it seems that even when the worse arm has the lower variance, $\mathcal{E}_{LCB}$ outperforms $\mathcal{E}_{UCB}$.



(a) Equal means

(b) Different means

Figure 7: Results describing identification of good arms over time for different variances, for $n_g = 2$ with equal means (left) and different means (right)

[5] We assume *known variance* and use $\phi(t, \delta) = \sigma\sqrt{2\frac{\log(1/\delta)+3\log\log(1/\delta)+(3/2)\log\log(et/2))}{t}}$ from [11]