Generalized Objectives in Adaptive Experiments: The Frontier between Regret and Speed

Chao Qin CQ2199@COLUMBIA.EDU
Division of Decision, Risk and Operations
Columbia Business School
New York, NY 10027, USA

Daniel Russo DJR2174@GSB.COLUMBIA.EDU
Division of Decision, Risk and Operations
Columbia Business School
New York, NY 10027, USA

The full paper will be posted on arXiv soon.

Abstract

This paper formulates a generalized model of multi-armed bandit experiments that accommodates both cumulative regret minimization and best-arm identification objectives. We identify the optimal instance-dependent scaling of the cumulative cost across experimentation and deployment, which is expressed in the familiar form uncovered by Lai and Robbins (1985). We show that the nature of asymptotically efficient algorithms is nearly independent of the cost functions, emphasizing a remarkable universality phenomenon. Balancing various cost considerations is reduced to an appropriate choice of exploitation rate. Additionally, we explore the Pareto frontier between the length of experiment and the cumulative regret across experimentation and deployment. A notable and universal feature is that even a slight reduction in the exploitation rate (from one to a slightly lower value) results in a substantial decrease in the experiment's length, accompanied by only a minimal increase in the cumulative regret.

Keywords: multi-armed bandits, regret minimization, best-arm identification, Thompson sampling

1. Introduction

The multi-armed bandit literature is divided into two distinct segments. One, focused on *best-arm identification* or *pure exploration*, aims to minimizing the expected number of measurements required to confidently identifying an optimal treatment arm. Another, focused on *cumulative regret minimization*, aims to interleave exploration and exploitation so as to maximize the cumulative reward earned.

The existence of these two widely studied problems reflects that practitioners conducting adaptive experiments have varied goals. Those conducting clinical trials or A/B tests may wish to end experimentation quickly and focus on deploying a specific treatment arm in the population. Formulating a best-arm identification problem appears natural in those cases, but it overlooks that the quality of treatment decisions during the experiment may be an important consideration (Berry, 2004). In simulation optimization problems (Hong

et al., 2021), the quality of decision-making during the experiment is unimportant, but the time required to simulate the performance may depend on the arm itself, and these distinct sampling costs should be reflected in the formulation.

We formulate a generalized model of a bandit experiments. In this model, there are K treatment arms with *unknown* quality $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and a population of T individuals is arriving sequentially at each period $t = 0, \ldots, T-1$. During the experimentation phase, a decision-maker pulls one treatment arm $I_t \in [K] \triangleq \{1, \ldots, K\}$ for the individual at period t and observes a noisy signal of arm I_t's quality. At any period $\tau \in \{0, 1, \ldots, T\}$, the decision-maker can choose to end the experimentation phase and deploy an arm $\hat{I}_\tau \in [K]$ to the remaining individuals throughout periods $\{\tau, \ldots, T-1\}$. In this work, we assume that noisy signals of treatment arm $i \in [K]$ are independently drawn from a canonical one-dimensional exponential family parameterized by mean parameter θ_i, and we focus on problem instances with a unique best arm. Formally, $\boldsymbol{\theta} \in \Theta$ where

$$\Theta \triangleq \left\{\boldsymbol{\theta} \in \mathbb{R}^K \,:\, I^*(\boldsymbol{\theta}) \text{ is unique}\right\} \quad \text{with} \quad I^*(\boldsymbol{\theta}) \triangleq \arg\max_{i \in [K]} \theta_i.$$

Distinct per-individual cost functions measure the cost of treatment decisions during the experimentation and deployment phase. For treatment arm $i \in [K]$, there is a per-individual cost $C_i(\boldsymbol{\theta})$ of employing it *within-experimentation* and a per-individual cost $\Delta_i(\boldsymbol{\theta})$ of deploying it *post-experimentation*. The total expected cost of the experiment is

$$\mathrm{Cost}_{\boldsymbol{\theta}}(T, \texttt{Alg}) \triangleq \mathbb{E}\left[\sum_{t=0}^{\tau-1} C_{I_t}(\boldsymbol{\theta}) + (T-\tau)\Delta_{\hat{I}_\tau}(\boldsymbol{\theta})\right],$$

where $\texttt{Alg}$ is an algorithm, i.e., an adaptive rule for selecting the treatment arms $(I_0, \ldots, I_{\tau-1})$, the stopping time τ and the deployed arm $\hat{I}_\tau$.

We place two assumptions on the cost functions:

Assumption 1 *For any $\boldsymbol{\theta} \in \Theta$, there is a strictly positive cost to continuing the experiment, i.e., $\min_{i \in [K]} C_i(\boldsymbol{\theta}) > 0$.*

Assumption 2 *For any $\boldsymbol{\theta} \in \Theta$, recommending the best arm $I^* = I^*(\boldsymbol{\theta})$ uniquely minimizes post-treatment costs, i.e., $\Delta_{I^*}(\boldsymbol{\theta}) = 0$ and $\Delta_j(\boldsymbol{\theta}) > 0$ for $j \neq I^*$.*

The next example offers some discussion of the generalized cost function and its relationship to past literature.

Example 1 (Cumulative regret minimization and best-arm identification) *For any arm $i \in [K]$, its per-individual post-experiment cost $\Delta_i(\boldsymbol{\theta}) = \theta_{I^*} - \theta_i$ and within-experiment cost $C_i(\boldsymbol{\theta}) = c + \Delta_i(\boldsymbol{\theta})$ where $c > 0$. Then the total cost*

$$\mathrm{Cost}_{\boldsymbol{\theta}}(T, \texttt{Alg}) = \mathbb{E}\left[c \cdot \tau + \sum_{t=0}^{\tau-1} (\theta_{I^*} - \theta_{I_t}) + (T-\tau)\left(\theta_{I^*} - \theta_{\hat{I}_\tau}\right)\right]$$

aggregates a term $\sum_{t=0}^{\tau-1} (\theta_{I^*} - \theta_{I_t})$ that captures within-experiment regret, a term $(T-\tau)\left(\theta_{I^*} - \theta_{\hat{I}_\tau}\right)$ that captures post-experiment regret of deploying the arm $\hat{I}_\tau$, and a term $c \cdot \tau$

that penalizes long experiments. The cost c might, for instance, reflect inherent operation costs involved with running a clinical trial.

For each given c, our analysis studies the limits of attainable total cost as the population size T grows. Varying c then allows us to unify insights from two threads of the bandit literature:

- *(Cumulative regret minimization) When c is nearly zero, our theory matches the classical theory of regret pioneered by Lai and Robbins (1985). This is natural, since when $c = 0$ this problem is equivalent to that of cumulative regret minimization in bandit problems*[1]*.*

- *(Best-arm identification) When c is very large, our theory recovers results on best-arm identification (Garivier and Kaufmann, 2016). In this regime, the term $\sum_{t=0}^{\tau-1}(\theta_{I^*} - \theta_{I_t})$ that captures the cost contribution of within-experiment regret is negligible compared to the other two sources of cost. Ignoring that cost, the decision-maker's goal is to stop rapidly while still gathering enough information to deploy an arm $\hat{I}_\tau$ with small expected regret.*

2. Contributions and informal presentation of the main results

Asymptotic performance limits. We study the nature of experimentation rules that minimize total costs $\mathrm{Cost}_{\boldsymbol{\theta}}(T, \mathtt{Alg})$ as the population size T tends to infinity. We show that asymptotically optimal algorithms incur costs that grow logarithmically with the population size. With this in mind, we define the normalized cost of an algorithm

$$\mathrm{NCost}_{\boldsymbol{\theta}}(\mathtt{Alg}) \triangleq \limsup_{T\to\infty} \frac{\mathrm{Cost}_{\boldsymbol{\theta}}(T, \mathtt{Alg})}{\log(T)}.$$

Then for every possible instance $\boldsymbol{\theta} \in \Theta$, its problem complexity is defined as

$$\mathrm{NC}^*_{\boldsymbol{\theta}} \triangleq \inf_{\mathtt{Alg}\in\mathcal{A}} \mathrm{NCost}_{\boldsymbol{\theta}}(\mathtt{Alg}), \tag{1}$$

where $\mathcal{A}$ is the class of *uniformly good rules* similar to that in the classical work of Lai and Robbins (1985).

The classic result of Lai and Robbins (1985) applies to Example 1 in the case of $c = 0$. In this case, $\mathrm{Cost}_{\boldsymbol{\theta}}(T, \mathtt{Alg})$ encodes an algorithm's expected regret, and they show the problem complexity $\mathrm{NC}^*_{\boldsymbol{\theta}} = \sum_{j\neq I^*} \frac{\theta_{I^*} - \theta_j}{\mathrm{KL}(\theta_j, \theta_{I^*})}$. The next theorem shows that for our generalized model, $\mathrm{NC}^*_{\boldsymbol{\theta}}$ takes on a related form.

Theorem 1 (Informal) *For any $\boldsymbol{\theta} \in \Theta$, there exist $\bar{\theta}_{I^*,j} \in (\theta_j, \theta_{I^*})$ for each $j \neq I^*$ such that*

$$\mathrm{NC}^*_{\boldsymbol{\theta}} = \sum_{j\neq I^*} \frac{C_j(\boldsymbol{\theta})}{\mathrm{KL}(\theta_j, \bar{\theta}_{I^*,j})}.$$

1. There is no disadvantage taking $\tau = T$ when $c = 0$. Stopping earlier is a feasible strategy which goes by the name 'explore-then-commit' in the bandit literature

The exact form of $\bar{\theta}_{I^*,j}$ is identified in our theory. If, in the limit, an efficient algorithm plays the best arm I^* much more often than any competing arm $j \neq I^*$, then $\bar{\theta}_{I^*,j} \approx \theta_{I^*}$ and the result of Lai and Robbins (1985) is essentially recovered. In settings where all within-experiment costs share the same value, i.e. $C_1(\boldsymbol{\theta}) = \cdots = C_K(\boldsymbol{\theta}) = \gamma > 0$, the term $\frac{\mathrm{NC}^*_{\boldsymbol{\theta}}}{\gamma}$ is equal to the problem complexity terms that have appeared previously in the best-arm identification literature (Garivier and Kaufmann, 2016; Russo, 2020). Indeed, the expressions in this literature appear different at first glance, and our new mathematical analysis is needed to put them into an expression that resembles Lai and Robbins (1985).

Universality of (asymptotically) optimal algorithms. We have defined what appears to be a substantial generalization of the typical mutli-armed bandit problem. One might expect these problems require substantially new bandit algorithms, which carefully calibrate their exploration on the basis of the within-experiment and post-experiment cost functions. Surprisingly, this not the case. Instead, a *universality* phenomenon emerges, where the nature of asymptotically optimal allocation rules is nearly independent of the details of the cost functions.

Let us illustrate the main idea informally by focusing on one concrete algorithm. We take Thompson sampling, one of the most widely used strategies for standard multi-armed bandit algorithms, and modify it as in Russo (2020) by introducing a single tuning parameter. Given observations gathered so far, Thompson sampling (TS) defines a randomized rule for sampling an arm to measure. Top-two Thompson sampling (TTTS) uses TS as a subroutine, sampling arms randomly until two distinct candidates are chosen and then flipping a biased coin to pick among those two. The bias of the coin, denoted by $\beta \in (0, 1)$, is an important tuning parameter. Russo (2020) explains that it governs the long-run proportion of measurement effort assigned to exploitation, and thus it is also called *exploitation rate*. When $\beta = 1$, TTTS is standard Thompson sampling and in the limit measures the true bast arm almost always. When $\beta < 1$, TTTS samples the true best arm β fraction of the time in the limit. The remaining $1 - \beta$ fraction of the time, the algorithm measures one of the $k - 1$ alternative arms and the allocation among these is determined automatically.

The next theorem states that, if the exploitation rate β is tuned appropriately to the problem instance, then TTTS with appropriate stopping and recommendation rules is asymptotically efficient. The surprising feature of this result is that TTTS is agnostic to the within-experiment and post-experiment cost functions. We show that the chosen stopping and recommendation rules are agnostic to these as well. The cost structure impacts the optimal exploitation rate but, fixing this, *has no impact on how remaining measurement effort is allocated among the other $k - 1$ arms.*

Theorem 2 (Informal) *For any $\boldsymbol{\theta} \in \Theta$, if top-two Thompson sampling is applied with optimally tuned β, generalized likelihood ratio stopping rule and plug-in recommendation rule, then*

$$\mathrm{NCost}_{\boldsymbol{\theta}}(\texttt{TTTS}) = \mathrm{NC}^*_{\boldsymbol{\theta}}.$$

Our focus on TTTS here is merely illustrative, so we will not discuss the algorithm's shortcomings or its strengths. For our theory, the only important feature of TTTS is its limiting allocation. Russo (2020) identified limiting proportions $\boldsymbol{p}^\beta(\boldsymbol{\theta}) = \left(p^\beta_1(\boldsymbol{\theta}), \ldots, p^\beta_K(\boldsymbol{\theta})\right)$,

which describe the fraction of measurements allocated to each arm in the limit under certain top-two sampling rules. When $\boldsymbol{\theta}$ is fixed, we write $\boldsymbol{p}^\beta = \boldsymbol{p}^\beta(\boldsymbol{\theta})$ and $\left(p_1^\beta, \ldots, p_K^\beta\right) = \left(p_1^\beta(\boldsymbol{\theta}), \ldots, p_K^\beta(\boldsymbol{\theta})\right)$ for simplicity. Figure 1 provides a visualizations of these limiting pro-

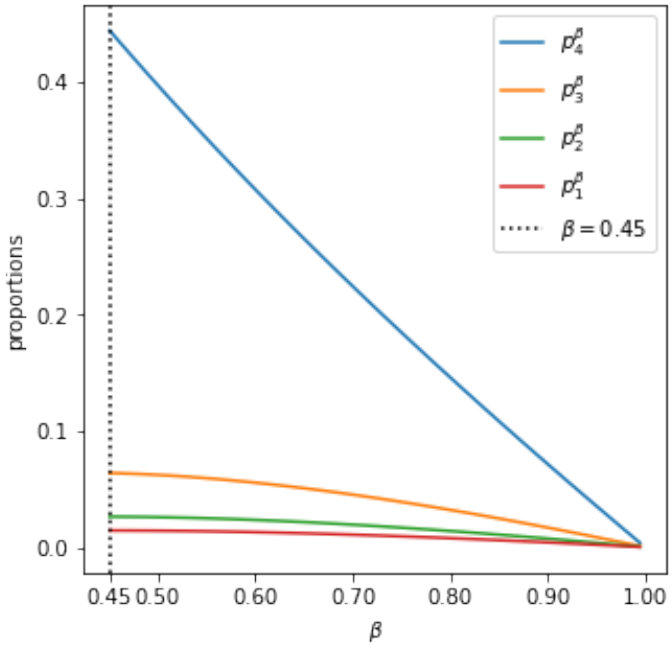

Figure 1: Proportions allocated to sub-optimal arms vs. exploitation rate

portions as β varies for the Gaussian instance with common known variance but unknown mean vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_5) = (1, 2, 3, 4, 5)$. Our theory applies, in an appropriate sense, to any algorithms with these limiting proportions.

Pareto frontier between regret and speed. We have identified, informally, that a one dimensional class of allocation proportions $(\boldsymbol{p}^\beta)_{\beta \in (0,1)}$ is optimal under a generic set of objective functions. By then plotting various performance measures as a function of the exploitation rate β, we can trace the Pareto frontier. We illustrate this in Figure 2 while focusing on two important performance measures: (1) experiment's length $\mathbb{E}_{\boldsymbol{\theta}}[\tau]$ and (2) total regret across experiment and deployment, $\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{\tau-1}(\theta_{I^*} - \theta_{I_t}) + (T - \tau)\left(\theta_{I^*} - \theta_{\hat{I}_\tau}\right)\right]$. One can think of this as varying the parameter c is Example 1.[2] One extreme point on this frontier, attained when $\beta \to 1$ (i.e. $c \to 0$), focuses solely on regret minimization. Other points prioritize committing quickly to a single decision. The best-arm identification problem is an extreme point here in which speed of confidently identifying I^* is prioritized; for this instance $\boldsymbol{\theta} = (1, 2, 3, 4, 5)$, the optimal $\beta = 0.45$. The extra regret incurred by such solutions can be understood as the price of early commitment. A striking and universal feature is that when decreasing β from 1 to a slightly lower value (e.g. 0.9 for this instance),

2. In Figure 2, we divide experiment's length and total regret by $\log(T)$ and let $T \to \infty$, which gives the normalized length and regret, respectively.

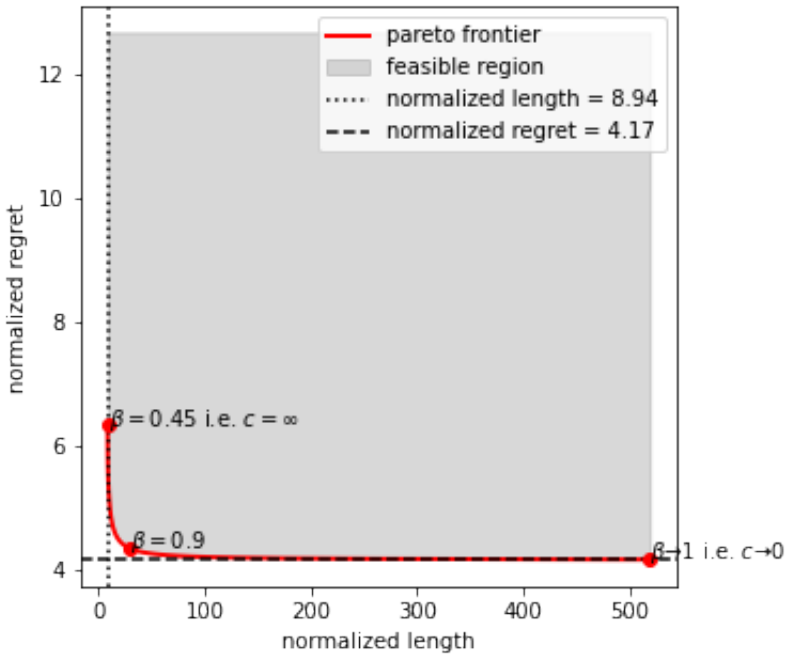

Figure 2: Pareto frontier between total regret and experiment's length

there is a significant decrease in experiment's length with only a minimal increase in total regret.

References

Donald A. Berry. Bayesian Statistics and the Efficiency and Ethics of Clinical Trials. *Statistical Science*, 19(1):175 – 187, 2004.

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

L Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 2020.