Multi-armed Bandit Experimental Design: Online Decision-making and Adaptive Inference

David Simchi-Levi, Chonghuan Wang
Laboratory for Information and Decision Systems, MIT

Abstract

Multi-armed bandit has been well-known for its efficiency in online decision-making in terms of minimizing the loss of the participants' welfare during experiments (i.e., the regret). In clinical trials and many other scenarios, the statistical power of inferring the treatment effects (i.e., the gaps between the mean outcomes of different arms) is also crucial. Nevertheless, minimizing the regret entails harming the statistical power of estimating the treatment effect, since the observations from some arms can be limited. In this paper, we investigate the trade-off between efficiency and statistical power by casting the multi-armed bandit experimental design into a *minimax multi-objective optimization problem*. We introduce the concept of *Pareto optimality* to mathematically characterize the situation in which neither the statistical power nor the efficiency can be improved without degrading the other. We derive a useful *sufficient and necessary condition* for the Pareto optimal solutions. Additionally, we design an effective Pareto optimal multi-armed bandit experiment that can be tailored to different levels of the trade-off between the two objectives.

1 Introduction

Multi-armed bandit (MAB), one of the most effective frameworks for sequential decision making, is renowned for its adaptability as more evidence becomes available. The adaptive allocation has been demonstrated to be more efficient than some traditional random experiments, such as classical random control trials (RCTs), in Lai et al. (1985). The current theoretical literature has studied MAB extensively, mainly focusing on understanding the best achievable efficiency in the hope of minimizing the loss of the participants' welfare during experiments (i.e., the regret). The minimax optimal regret of stochastic MAB has been well understood to be $\widetilde{\Theta}(\log n)$, which can be achieved by the famous upper confidence bound (UCB) based algorithms (see, Lai et al. 1985) and Thompson sampling (TS) based algorithms (see, Thompson 1933).

However, for many real-world problems, regret is not the only metric that matters when conducting experiments. Consider the following hypothetical scenario as a motivating example where a drug company is using the MAB to evaluate the efficacy of a new drug. Regret measures the overall loss of patients' welfare, so we want to keep it to a minimum. Especially, for rare or fatal diseases, it is expected to treat the patients within the trial as effectively as possible. Additionally, it is always crucial to report what kind of difference the new treatment can make compared with the control, i.e., to infer the average treatment effect (ATE) (see, e.g., Angrist and Imbens 1995). It is also of great interest to understand the difference between any two drugs if more than one is being tested simultaneously. Such kind of inference can be quite instructive in that it may be used to determine which alternative should be used when the best drug is unavailable due to shortage, regulation or some other factors. This illustrates *the necessity of adaptive statistical inference of ATE while making online assignment decisions* in MAB experiments.

Online decision-making and statistical inference of ATE have been extensively investigated separately, but when they are jointly considered, several new and crucial challenges arise. First, as MAB algorithms collect data adaptively and consequently, it is not appropriate to consider the collected data as independent and identically distributed (i.i.d.). This imposes extra challenges to conducting inference, and may harm the statistical power. For example, using the sample average to estimate ATE (as practitioners usually do in RCTs) in some MAB algorithms including UCB and TS creates a non-negligible bias, as pointed out by much work in literature (see, e.g., Xu et al. 2013, Luedtke and Van Der Laan 2016, Nie et al. 2018). A recent stream of literature develops several offline, post-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

experiment analysis methods based on the adaptively collected data (see, e.g., Kato et al. 2020, Hadad et al. 2021, Chen et al. 2022). They all have specific constraints or assumptions on the data collection process, which may usually reduce the efficiency of MAB algorithms. In other words, MAB algorithms are effective in learning the optimal online decision-making policy, but lack statistical power. In contrast, some existing well-established ATE estimation methods are recognized for their strong statistical power, but will incur significant regrets. Specifically, since a predetermined percentage of the population will be assigned to the suboptimal arm, the regret of a traditional RCT will increase linearly with the number of samples. Another commonly adopted assumption for adaptive ATE estimators is "overlap", which states that the probability of playing all arms is lower bounded away from zero by a universal constant over time (see, e.g., Khan and Ugander 2021, Hahn et al. 2011). Then, the accumulative regret is also linear because of the linear growth of the expected times the suboptimal arms are played. One of our goals in this paper is *to design MAB experiments with both online decision-making policies and adaptive inference for ATE with finite sample guarantees* by harnessing the efficiency and statistical power from each side.

Moreover, the relationship between the two tasks, adaptive inference and online decision-making, may change throughout the experiment. On the one hand, these two tasks can complement one another, particularly in the initial phase of the experiment. An accurate estimation of ATE is undoubtedly informative for learning the best decision-making policy. An efficient online decision-making algorithm typically explores all arms actively including the suboptimal ones at the beginning explicitly or implicitly, in order to gather enough information. This is advantageous for inference as well. On the other hand, the success of gauging ATE may be hindered by online decision-making, which is only concerned about being able to identify the best arm without considering how much better it is. To be more precise, minimizing the regret requires the algorithm to stop performing the suboptimal action as soon as it gains enough confidence about the suboptimality. Due to the potentially relatively few observations, it will make an accurate estimation of the suboptimal arm virtually impossible. However, since ATE is the difference, an accurate inference heavily relies on the estimators of both arms, and thus the statistical power of estimating ATE may be limited by efficient online decision-making. Therefore, there exists an important trade-off between these two tasks. In order to describe such a trade-off, we introduce the term "Pareto optimality" to characterize the circumstance where neither regret nor estimating error of ATE can be made better off without making the other worse off. *How to statistically quantify and practically achieve the Pareto optimality in MAB experiments* remains an open question.

1.1 Preliminaries

In stochastic MAB experiments, there is a finite set $\mathcal{A}$ of arms (i.e., treatments or actions) $a \in \mathcal{A}$ with $|\mathcal{A}| = K$. Without loss of generality, the control can also be seen as an action. n is the total number of experimental units (or the time horizon). At each time $t \leq n$, the environment generates a reward $r_t(a)$ for every arm $a \in \mathcal{A}$. After choosing arm a_t, only the reward of the chosen arm $r_t := r_t(a_t)$ can be observed. The expectation $\mathbb{E}[r_t(a)] = \mu_a \in [-1, 1]$ where μ_a is the unknown true reward of arm a which is disturbed by an i.i.d. noise to generate $r_t(a) \in [-1, 1]$. A stochastic MAB instance can be denoted by $\nu = (P_1, \cdots, P_K)$, where P_i is the distribution of the rewards of arm i. The optimal arm is the arm with the maximum mean reward denoted by $a^* := \arg\max_{a \in \mathcal{A}} \mu_a$, and thus is also unknown. We define the gap between arm i and arm j as $\Delta^{(i,j)} := \mu_i - \mu_j$, for any $i \neq j \in [K]$. Among all $\Delta^{(i,j)}$, we additionally define the difference between μ_{a^*} and μ_a as the suboptimality gap, i.e., $\Delta(a) := \mu_{a^*} - \mu_a$ for $a \in \mathcal{A} \backslash \{a^*\}$. Specifically, when $K = 2$, we can define arm 1 to be the treatment of interest and arm 2 to be a control, and thus $\Delta^{(1,2)}$ is the ATE. From now on, we only need to focus on $\Delta^{(i,j)}$ as the ATE follows naturally by the definition of each arm when designing the experiments. In this paper, we will elaborate on $|\Delta^{(i,j)}| = \Theta(1)$ for all $i \neq j \in [K]$, which is arguably the most fundamental case. Denote all stochastic MAB instances satisfying the mentioned assumptions to constitute a feasible set $\mathcal{E}_0$. We will discuss the case where $|\Delta^{(i,j)}|$ is extremely small i.e., $|\Delta^{(i,j)}| = \mathcal{O}(n^{-p})$ for some $p > 0$ in Sec. 4.

At every time t, the decision maker observes the history $\mathcal{H}_t = (a_1, r_1, \cdots, a_t, r_t)$. An admissible policy $\pi = \{\pi_t\}_{t \geq 1}$ maps the history $\mathcal{H}_{t-1}$ to an action a_t. Denote the probability with which arm a is chosen at time t as $\pi_t(a) = \mathbb{P}(a_t = a \mid \mathcal{H}_{t-1})$ under policy π. To measure the efficiency of online learning, we use the accumulative *regret*, defined as the expected difference between the reward under the optimal policy and the policy π, i.e., $\mathcal{R}(n, \pi) = \mathbb{E}^{\pi}[n\mu_{a^*} - \sum_{i=1}^{n} r_i(a_i)]$. In addition, an admissible adaptive estimator $\hat{\Delta}^{(i,j)} = \{\hat{\Delta}_t^{(i,j)}\}_{t \geq 1}$ maps the history $\mathcal{H}_t$ to an estimation of $\Delta^{(i,j)}$ at each time t. We use the *error* defined as the expected distance of $\Delta^{(i,j)}$ and $\hat{\Delta}_t^{(i,j)}$, (i.e., $e(t, \hat{\Delta}^{(i,j)}) = \mathbb{E}[|\Delta^{(i,j)} - \hat{\Delta}_t^{(i,j)}|]$) to measure the quality of the estimation. We define $\hat{\Delta} := \{\hat{\Delta}(i,j)\}_{i<j \leq K}$ to represent all the estimators on the gap between any two arms. A design of an MAB experiment can then be represented by an admissible pair $(\pi, \hat{\Delta})$. The optimal design of MAB experiments in this paper is solving the following minimax multi-objective optimization problem:

$$\min_{(\pi, \hat{\Delta})} \max_{\nu \in \mathcal{E}_0} \left(\mathcal{R}_\nu(n, \pi), \max_{i<j \leq K} e_\nu(n, \hat{\Delta}^{(i,j)}) \right), \quad (1)$$

where we use the subscript ν to denote the MAB instance. Eq. (1) mathematically describes the two goals: minimiz-

ing the regret and the inference error under the worst case. For traditional MAB problems, $\min_\pi \max_{\nu\in\mathcal{E}_0} \mathcal{R}(n,\pi)$ is usually the only objective. The asymptotic behavior of $\hat{\Delta}$ is one of the central focuses of the existing ATE literature. Note that π and $\hat{\Delta}$ are complicatedly correlated through the history $\mathcal{H}$, and thus the optimization problem (1) has some implicit constraints on π and $\hat{\Delta}$. Such kinds of constraints may lead to a complicated feasible region and thus impose great challenges on solving the optimization problem.

In order to define the optimality and understand the structure of solutions, we start from the inner maximization over all $\nu \in \mathcal{E}_0$ problem in Eq. (1). Intuitively, under a given $(\pi, \hat{\Delta})$, the pairs $(\mathcal{R}_\nu, \max_{i<j\le K} e_\nu)$ for all $\nu \in \mathcal{E}_0$ can constitute an accessible region, and the *front* of the accessible region can be defined to be the optimal values of the inner maximum problem (see, Figure 1). Formally, we define the front of a given pair $(\pi, \hat{\Delta})$ as follows.

Definition 1 (Front)**.** *The front for $(\pi, \hat{\Delta})$, denoted by $\mathcal{F}(\pi, \hat{\Delta})$, consists of all pairs of (R, e) satisfying: (i) $\exists \nu \in \mathcal{E}_0, (\mathcal{R}_\nu(n,\pi), \max_{i<j\le K} e_\nu(n, \hat{\Delta}^{(i,j)})) = (R, e)$; (ii) $\nexists \nu \in \mathcal{E}_0$, $\max_{i<j\le K} e_\nu(n, \hat{\Delta}^{(i,j)}) > e$ and $\mathcal{R}_\nu(n,\pi) \ge R$; (iii) $\nexists \nu \in \mathcal{E}_0$, $\max_{i<j\le K} e_\nu(n, \hat{\Delta}^{(i,j)}) \ge e$ and $\mathcal{R}_\nu(n,\pi) > R$.*

The first condition ensures the achievability of the front. The second and the third conditions are describing that there does not exist any instance that will incur no fewer values on both objectives and a strictly larger value on at least one. From now on, including in Definition 1, when it comes to comparing regrets or errors, we only focus on the order of n ignoring the universal constant and the logarithm terms, since n is usually relatively large. In Figure 1, the yellow region and its boundary are an example of the accessible region and front, respectively. We also present the traditional RCTs in Figure 1. Since $|\Delta^{(i,j)}| = \Theta(1)$ for the instance class $\mathcal{E}_0$, the RCTs will usually incur linear regrets, and thus the accessible region is a line. By the theoretical results for RCTs (see, e.g., Wainwright 2019), the best achievable accuracy is $\Theta(n^{-\frac{1}{2}})$, and thus the front for RCTs is at the point $(n, n^{-\frac{1}{2}})$ ignoring the constant independent of n and the logarithm terms. Additionally, the well-known results in MAB tells that the worst-case regret of any policy is no smaller than $\log(n)$, and thus any accessible region will inevitably have some parts on or above the $\log(n)$ line shown in Figure 1.

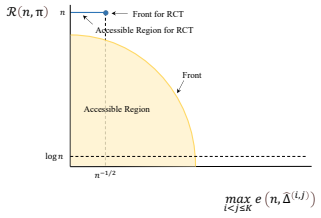

Figure 1: Examples of accessible regions and fronts.

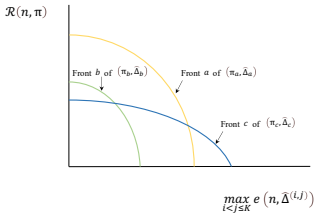

Figure 2: Examples of Pareto dominance.

In order to define the optimality for our minimax objective (1), we first define the *Pareto dominace* between two feasible solutions based on the definition of the front as follows.

Definition 2 (Pareto dominance)**.** *A feasible solution $(\pi_1, \hat{\Delta}_1)$ Pareto dominates another solution $(\pi_2, \hat{\Delta}_2)$ if $\forall (R_1, e_1) \in \mathcal{F}(\pi_1, \hat{\Delta}_1)$, $\exists (R_2, e_2) \in \mathcal{F}(\pi_2, \hat{\Delta}_2)$, such that at least one of the following two conditions holds: (i) $R_1 \le R_2$ and $e_1 < e_2$ or (ii) $R_1 < R_2$ and $e_1 \le e_2$.*

The definition formally describes that $(\pi_1, \hat{\Delta}_1)$ is Pareto better than $(\pi_2, \hat{\Delta}_2)$ if for any point (R_1, e_1) on the front of $(\pi_1, \hat{\Delta}_1)$, there exists some point (R_2, e_2) on the front of $(\pi_2, \hat{\Delta}_2)$ such that (R_1, e_1) is no larger than (R_2, e_2) on both coordinates and is strictly better on at least one coordinate. In Figure 2, we present toy examples of the fronts of three different solutions. By the definition, $(\pi_b, \hat{\Delta}_b)$ Pareto dominates $(\pi_a, \hat{\Delta}_a)$, and $(\pi_c, \hat{\Delta}_c)$ can neither Pareto dominate nor be Pareto dominated by $(\pi_a, \hat{\Delta}_a)$ or $(\pi_b, \hat{\Delta}_b)$. Based on such Pareto dominance, we can have the definition of the crucial concept in the paper *Pareto optimality*.

Definition 3 (Pareto Optimality)**.** *An admissible pair of $(\pi^*, \hat{\Delta}^*)$ is Pareto optimal in terms of the dependence on n, if it is not Pareto dominated by any other solution. Pareto frontier denoted as $\mathcal{P}$ is the envelop of the fronts of all the Pareto optimal solutions.*

That a pair $(\pi^*, \hat{\Delta}^*)$ is Pareto optimal does not mean it can Pareto dominate every other policy, and thus there may exist a group of admissible pairs that are all Pareto optimal. The reason why we emphasize that our Pareto optimality is in terms of the dependence on n is that in Definitions 1 and 2 we focus on the dependence of n ignoring the constants and the logarithm terms.

In general, designing Pareto optimal MAB experiments is in the center of this work. The first natural research question is *how to solve the minimax multi-objective optimization problem to get the Pareto optimal solutions.* The first challenge stems from the multiple objectives, which are arguably more challenging to tackle than solving the single objective optimization problem like the traditional MAB problems. Moreover, π and $\hat{\Delta}$ have the measurability constraints and are highly correlated through the history $\mathcal{H}_t$, which are hard to explicitly integrate into the optimization problem and endow the feasible region with complicated structures. Furthermore, finding only one Pareto optimal solution is always not enough. It is important to design experiments flexibly under different requirements for the trade-off between these two objectives. This is indeed asking *how to obtain the optimal Pareto optimal solutions given different levels of trade-off in these two objectives*.

1.2 Contributions and Main Results

The main contribution of this paper is the Pareto optimal design of MAB experiments, especially the statisti-

cal understanding of the trade-off between online decision-making and adaptive inference in MAB experiments. To the best of our knowledge, this work is the first to jointly consider the efficiency and statistical power in the experimental design literature. We are also the first to introduce the minimax multi-objective optimization framework to experimental designs. It also greatly generalizes the existing minimax optimization framework in MAB literature. We next summarize our main results from two folds.

First, we find a *sufficient and necessary condition* for the Pareto optimal solutions of the minimax multi-objective optimization problem (1). Specifically, an admissible pair $(\pi^*, \hat{\Delta}^*)$ is Pareto optimal if and only if

$$\max_{\nu\in\mathcal{E}_0}\left[\left(\max_{i<j\leq K} e_\nu(n,\hat{\Delta}^{*(i,j)})\right)\sqrt{\mathcal{R}_\nu(n,\pi^*)}\right]=\widetilde{\mathcal{O}}(1).$$

Compared with the definition of Pareto optimality, this condition seems more straightforwards to interpret and easier to verify. Technically, we establish an information-theoretical minimax lower bound to portray the trade-off between these two objectives. Specifically,

$$\inf_{(\pi,\hat{\Delta})}\max_{\nu\in\mathcal{E}_0}\left[\left(\max_{i<j\leq K} e_\nu(n,\hat{\Delta}^{(i,j)})\right)\sqrt{\mathcal{R}_\nu(n,\pi)}\right]=\Omega(1).$$

This lower bound tells that no solution can do better on $(\max_{i<j\leq K} e_\nu(n,\hat{\Delta}^{(i,j)}))\sqrt{\mathcal{R}_\nu(n,\pi)}$ than a constant order in the worst case. Particularly, since UCB/TS algorithms have the regret upper bound of $\mathcal{O}(\log n)$, then any ATE estimator based on the UCB/TS algorithm cannot avoid an error of $\Omega(\frac{1}{\sqrt{\log n}})$ in the worst case. This explicitly shows the lack of statistical power of UCB/TS algorithms on ATE inference. On the other hand, we construct a series of Pareto optimal policies that can Pareto dominate any policy violating the condition to show its necessity. We also show that the Pareto frontier is the curve that satisfies $(\max_{i<j\leq K} e_\nu(n,\hat{\Delta}^{(i,j)}))\sqrt{\mathcal{R}_\nu(n,\pi)}=\widetilde{\Theta}(1)$.

Second, we propose an efficient Pareto optimal algorithm for stochastic MAB experiments which can adapt to different levels of trade-off between the two objectives. Specifically, we combine the well-known EXP3 algorithm (see, e.g., Auer et al. 2002a, Seldin et al. 2013) and the idea of extra forced exploration which means that the algorithm purposely plays the other arms after it identifies the best one. For any given $\alpha\in[0,1]$ as input, we prove that the regret of our algorithm is $\widetilde{\mathcal{O}}(n^{1-\alpha})$ and the estimation error of ATE is $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{t^{1-\alpha}}})$ for all $t\leq n$, showing its Pareto optimality. Note that α balances the two objectives. If α is large, the practitioners emphasize the control of regret. Small α will lead to a more accurate ATE estimation. Technically, a common practice when using the EXP3 algorithm is to estimate the expected rewards of each arm with inverse propensity weighted (IPW) estimators, which imposes the challenge on careful variance control.

Notably, most of the existing results on adaptive estimators in the inference literature are typically in the style of central limit theorem bounds which are asymptotic in nature, whereas all our bounds are finite in nature.

1.3 Related Work

Learning Efficiency in MABs. A main body of literature in MABs has focused on the online learning efficiency, i.e., minimizing regret (see, e.g., Lai et al. 1985, Sutton and Barto 2018, Lattimore and Szepesvári 2020). Two representative classes of algorithms that can provide optimal regret bounds, i.e., $\Theta(\log n)$ regret bounds, are UCB-based algorithms (see, e.g., Lai et al. 1985, Agrawal 1995, Auer 2002, Auer et al. 2002a, Garivier and Moulines 2011, Garivier and Cappé 2011, Carpentier et al. 2011) and TS-based algorithms (see, e.g., Thompson 1933, Chapelle and Li 2011, Kaufmann et al. 2012, Russo and Van Roy 2016). Both UCB/TS algorithms have been extended to the setting where contextual information of actions exists (see, e.g., Filippi et al. 2010, Chu et al. 2011, Russo and Van Roy 2014, Russo and Van Roy 2016, Li et al. 2017). Fan and Glynn (2021) and Simchi-Levi et al. (2022) reveal that efficiency-optimized bandit algorithms may suffer from serious heavy-tailed risk. In this paper, our design is based on the idea of EXP3, which was initially designed for adversarial MABs Auer et al. (2002b). Recently, it has gradually gained its own popularity in the stochastic setting (Seldin et al. 2011, Seldin et al. 2012, Seldin et al. 2013) and the mixed stochastic-adversarial setting (see, Bubeck and Slivkins 2012). The version of Bernstein's inequality we used is inspired by Seldin et al. (2013). These mentioned works only focus on minimizing the regret. Another growing body of MAB literature is aiming at identifying the best arm (see, e.g., Jennison et al. 1982, Mannor and Tsitsiklis 2004, Chan and Lai 2006, Gabillon et al. 2012, Garivier and Kaufmann 2016, Agrawal et al. 2021, Kato and Ariu 2021). Zhong et al. (2021) carefully study the trade-off between regret minimization and best-arm identification, which is different from our objective. An emerging field is the multitasking bandit, where minimizing regret is not the only objective (see, e.g., Yang et al. 2017, Yao et al. 2021, Deshmukh et al. 2017). Erraqabi et al. (2017) also want to balance the trade-off between regret and estimation error. They redefine a new reward function based on the observed rewards and the error bounds. By such a new reward to guide online decision-making, they formulate the problem into a single objective optimization, integrating the two objectives into one. In this way, they do not explicitly capture the trade-off as we do, and thus cannot describe the optimality of their design.

Adaptive experimental design. Experimental design is becoming more and more popular in operations research, econometrics, and statistics (see, e.g., Johari et al. 2015, Eckles et al. 2017, Aronow and Samii 2017, Athey et al.

2018, Xiong et al. 2019, Bojinov et al. 2020, Wager and Xu 2021, Bojinov et al. 2021, Johari et al. 2022, Farias et al. 2022b). Adaptive experimental design is the area that is most pertinent to this work (Hahn et al. 2011, Atan et al. 2019, Offer-Westort et al. 2021, Kasy and Sautmann 2021, Bhat et al. 2020). MAB itself can also be seen as a type of adaptive experimental design, but here we focus on the designs different from traditional MAB. Kato et al. (2020) investigate adaptive experiments for ATE when contexts can be observed. Glynn et al. (2020) propose a theoretical model to study optimal experimental design when temporal interference exists by transforming it into to a Markov decision problem. Adusumilli (2021) investigates the asymptotic Bayes and minimax risk for bandit experiments. Farias et al. (2022a) combine synthetic control and MAB to study the settings where experimental units are coarse due to interference or other concerns. Different from our stationary treatment effect, Qin and Russo (2022) investigate bandit experiments where a potentially nonstationary sequence of contexts influences arms' performance.

Inference in MABs. The work from Villar et al. (2015) is a pioneer work in revealing that MAB algorithms offer significant advantages in assigning more patients to better treatments, and severe limitations on resulting statistical power from an empirical perspective. We statistically describe and quantify such a issue. There is a substantial literature on post-experiment inference from logged adaptively collected data (see, e.g., Zhang et al. 2020, Zhang et al. 2021, Bibaut et al. 2021). One of the central tasks along this line is the evaluation of a new policy given historic/observational data which cannot be seen as i.i.d. samples (see, e.g., Dudík et al. 2011, Dudík et al. 2014, Swaminathan and Joachims 2015, Li et al. 2015, Wang et al. 2017, Kallus and Zhou 2018, Farajtabar et al. 2018, Athey and Wager 2021, Zhan et al. 2021, Zhou et al. 2022, Hadad et al. 2021, Chen et al. 2022). Bareinboim et al. (2015) study the issue of unobserved confounding in MAB, and consider how the observational data can be used to empower TS algorithms. Dimakopoulou et al. (2021) focus on conducting inference on the true mean of each arm based on data collected by stochastic MAB so far at each step. They incorporate the adaptively weighted doubly robust estimator into TS algorithms, which is proved to achieve the optimal regret and has outstanding empirical performances. Dimakopoulou et al. (2017) and Dimakopoulou et al. (2019) consider the case where context exists and estimate the conditional expectation of each action's reward under different contexts.

Finally, we remark that the full version of this paper (containing additional theoretical results, computational experiments, and missing proofs) is available at `https://ssrn.com/abstract=4224969`.

2 MAB Experimental Design for $K = 2$

In this section, we focus on $K = 2$ to illustrate our ideas. We first establish the crucial lower bound and the sufficient condition for the Pareto optimality. Then, we propose a series of Pareto optimal designs and show the necessity of the condition based on the constructed Pareto optimal solutions. For brevity, we adopt the Δ instead of $\Delta^{(1,2)}$, since there is no ambiguity when $K = 2$.

2.1 A Lower Bound and A Sufficient Condition

In this subsection, we start with establishing an lower bound for $(e_\nu(n, \hat{\Delta}))\sqrt{\mathcal{R}_\nu(n, \pi)}$. In the following theorem, we establish an important minimax lower bound.

Theorem 1. *For any admissible pair $(\pi, \hat{\Delta}_n)$, there always exists a hard instance $\nu \in \mathcal{E}_0$ that $e_\nu(n, \hat{\Delta}_n)\sqrt{\mathcal{R}_\nu(n, \pi)}$ is no less than a constant order, i.e.,*

$$\inf_{(\pi, \hat{\Delta}_n)} \max_{\nu \in \mathcal{E}_0} \left[e_\nu(n, \hat{\Delta}_n)\sqrt{\mathcal{R}_\nu(n, \pi)} \right] = \Omega(1). \quad (2)$$

Theorem 1 states that for any admissible pair $(\pi, \hat{\Delta}_n)$, there usually exists a challenging instance $\nu \in \mathcal{E}$ such that the product of estimation error and regret is lower-bounded by n^p for some positive value of p. This mathematically highlights the trade-off between the two objectives. A small regret will inevitably have a large error on the ATE estimation. Roughly speaking, the expected error is almost lower bounded by the inverse of the square root of the regret in the worst case, i.e., $e_\nu(n, \hat{\Delta}_n) = \Omega(\frac{1}{\sqrt{\mathcal{R}_\nu(n,\pi)}})$. In particular, since $\mathcal{R}_\nu(n, \pi) = \mathcal{O}(\log(n))$ for UCB and TS algorithms, no estimators can not achieve smaller error than the order $\Omega(\frac{1}{\sqrt{\log(n)}})$ consistently over all the possible instances. Although $\log(n)$ increases with n, the speed is rather slow which explicitly shows the limitation of regret-optimal policies in terms of statistical power for estimating the ATE.

In Theorem 1, we have shown that no solution can perform better than a constant order in terms of $e_\nu(n, \hat{\Delta}_n)\sqrt{\mathcal{R}_\nu(n, \pi)}$ in the worst case. The following theorem states one policy is Pareto optimal if it can achieve the constant order on $e_\nu(n, \hat{\Delta}_n)\sqrt{\mathcal{R}_\nu(n, \pi)}$ in terms of the dependence on n.

Theorem 2. *An admissible pair $(\pi, \hat{\Delta})$ is Pareto optimal if*

$$\max_{\nu \in \mathcal{E}_0} \left[e_\nu(n, \hat{\Delta})\sqrt{\mathcal{R}_\nu(n, \pi)} \right] = \widetilde{\mathcal{O}}(1). \quad (3)$$

Together with Theorem 1, if Eq. (3) is satisfied, we can directly draw the conclusion that $(\pi, \hat{\Delta})$ is optimal in terms of the metric $\max_{\nu \in \mathcal{E}_0}[e_\nu(n, \hat{\Delta})\sqrt{\mathcal{R}_\nu(n, \pi)}]$. However, whether the optimality on such a metric can guarantee the Pareto optimality is what we want to answer in Theorem 2. Comparing with the definition of Pareto optimality, Eq.

(3) seems relatively more straightforwards and practical to verify. For example, consider the traditional RCTs where a half of experimental units are treated and controlled, respectively. For any $\nu \in \mathcal{E}_0$, $e_\nu(n, \hat{\Delta}_{\text{RCT}}) = \widetilde{O}(1/\sqrt{n})$ and $\mathcal{R}_\nu(n, \pi_{\text{RCT}}) = \Theta(n)$, and thus they are Pareto optimal.

2.2 An Algorithm and A Necessary Condition

Although the RCTs are Pareto optimal, they can not be easily adapted to different levels of trade-off between the two objectives. In this subsection, we propose a flexible algorithm satisfying Eq. (3) for $K = 2$ with the analysis of the regret upper bound and the error bound for inference. Then, we will prove that the condition (3) is necessary.

2.2.1 Algorithm and regret upper bound

We adopt the idea of the famous EXP3 algorithm for adversarial MAB (see, e.g., Seldin et al. 2013, Auer et al. 2002a), together with the idea to force the algorithm to actively explore the suboptimal arm, to design our EXP3 with exploration (EXP3E) algorithm shown in Algorithm 1.

We first define a set of random variables $\hat{R}_t(a)$ for $a \in \{1, 2\}$ based on inverse propensity score weight (IPW) as: $\hat{R}_t(a) = \hat{R}_{t-1}(a) + \frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t}$, which can provide an unbiased estimation of μ_a after being divided by t, i.e., $\mathbb{E}[\hat{R}_t(a)] = \mu_a t$. We also define $\hat{R}_t^{\max}$ as $\max_{a\in\{1,2\}} \hat{R}_t(a)$. One may think a more straightforward way to estimate μ_a is the simple sample average $\frac{\sum_{s=1}^t \mathbb{I}_{a=A_t} R_t}{\sum_{s=1}^t \mathbb{I}_{a=A_t}}$. However, such an estimator is neither unbiased nor asymptotically normal because whether we take action a at time t is highly correlated with the past history as is pointed by the recent works (see, e.g., Xu et al. 2013, Luedtke and Van Der Laan 2016, Hadad et al. 2021, Nie et al. 2018, Zhang et al. 2020). Thus, the ATE based on the simple sample average will inevitably be biased. The first phase of our algorithm is aiming at identifying the best arm with well-controlled regret. In this phase, the algorithm is adaptively polishing its decision policy to gain confidence about which arm is the optimal one, according to the estimated reward $\hat{R}_t(a)$. There are many ways to map $\hat{R}_t(a)$ into probabilities, among which a popular choice is exponential weighting as $\pi_t(a) = \frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{\sum_{a\in\mathcal{A}} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}$. Note that the decision maker knows the exactly $\pi_t(a)$, different from the classical offline ATE inference. If at time t there exists an arm a such that $\hat{R}_t(a)$ is larger than the other by at least $\Omega(\sqrt{t})$, the algorithm believes a is the optimal arm and eliminates the other arm. Formally, our elimination rule is $\mathcal{A}_{t+1} = \mathcal{A}_t \backslash \{a \in \mathcal{A}_t : \hat{R}_t^{\max} - \hat{R}_t(a) > 2\sqrt{Ct}\}$, where C is a constant defined in Algorithm 1. Note that when the first phase ends is a stopping time with respect to the history $\mathcal{H}_t$. We define two stopping times as $\tau(a) = \max\{t : a \in \mathcal{A}_t\}$ for $a \in \{1, 2\}$, and then the first phase ends after $\min_{a\in\{1,2\}} \tau(a)$ periods. By a careful analysis, the length of the first phase can be shown in the order $1/\Delta^2$.

Algorithm 1: EXP3 with exploration for $K = 2$ (EXP3E)

Input: α and δ
Initialization: $\mathcal{A}_1 = \{1, 2\}$, $\hat{R}_0(a) = 0$ for $a \in \{1, 2\}$, $\varepsilon_0 = 0$, $C = 4(e^2 + 2)^2(\log(\frac{2}{\delta}))^2$
for $t = 1, 2, \cdots, n$ **do**
  $\varepsilon_t = \frac{1}{\sqrt{Ct}}$, $\alpha_t = \frac{1}{2t^\alpha}$;
  if $|\mathcal{A}_t| = 2$: // Phase 1
    $\pi_t(a) = \frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{e^{\varepsilon_{t-1}\hat{R}_{t-1}(1)} + e^{\varepsilon_t \hat{R}_{t-1}(2)}}$ for $a \in \{1, 2\}$;
  else: // Phase 2
    $\pi_t(a) = 1 - \alpha_t$ if $a \in \mathcal{A}_t$; otherwise $\pi_t(a) = \alpha_t$;
  Select A_t according to π_t;
  Observe reward R_t;
  $\forall a \in \{1, 2\}$: $\hat{R}_t(a) = \hat{R}_{t-1}(a) + \frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t}$;
  $\mathcal{A}_{t+1} = \mathcal{A}_t \backslash \{a \in \mathcal{A}_t : \hat{R}_t^{\max} - \hat{R}_t(a) > 2\sqrt{Ct}\}$;
  Output: $\hat{\Delta}_t = \frac{1}{t}(\hat{R}_t(1) - \hat{R}_t(2))$;
end for

After eliminating the suboptimal arm, EXP3E operates into the second phase. The algorithm is forced to play the arm which was identified as the suboptimal one in the first phase with a carefully controlled probability $\alpha_t = \frac{1}{2t^\alpha}$. α is an important input parameter that balances our two tasks. In the following, we will see soon that a small α can help the algorithm to have a more accurate estimator of Δ, while sacrificing the regret.

Theorem 3. *Let Algorithm 1 runs with any given $\alpha \in [0, 1]$ and $\delta = \frac{1}{2n^2}$. The regret is*

$$\mathcal{O}\left(\frac{\log(n)}{\Delta} + n^{1-\alpha}\Delta\log(n)\right). \tag{4}$$

The regret bound in Theorem 3 decreases with α, which is consistent with our intuition that a large α restricts the probability to play the suboptimal arm in the second phase. When $\alpha = 1$, the regret bound in Theorem 3 becomes $\mathcal{O}(\frac{\log(n)}{\Delta} + \Delta\log(n))$, which matches with the optimal regret bound of MAB in current literature (see, e.g., Lattimore and Szepesvári 2020) up to logarithmic factors. This means that if minimizing the accumulative regret is the only objective (i.e., ignoring the inference task), by setting $\alpha = 1$, the performance of our EXP3E is unimprovable in terms of the dependency on the learning horizon n. Another extreme case is when $\alpha = 0$, the regret upper bound grows linearly with the learning horizon T. When α is set to be 0, in the second phase, the exploration probability remains to be $\frac{1}{2}$. This indicates that in the second phase EXP3E is doing random control trials. Moreover, if $|\Delta| = \Theta(1)$, Theorem 3 has an immediate corollary.

Corollary 1. *With any given $\alpha \in [0, 1]$, $\delta = \frac{1}{2n^2}$ and $|\Delta| = \Theta(1)$, the regret of Algorithm 1 is $\widetilde{\mathcal{O}}(n^{1-\alpha})$.*

2.2.2 Inference for ATE

Now, we are going to focus on the problem of inference of Δ. Since we have shown $\mathbb{E}[\hat{R}_t(a)] = \mu_a t$, we can define a set of martingales as $M_t^a = \hat{R}_t(a) - \mu_a t$ for $a \in \mathcal{A}$, and $M_t^{(1,2)} := M_t^1 - M_t^2 = t\hat{\Delta}_t - t\Delta$. We can directly have,

Theorem 4. *For $t \in [n]$, $\hat{\Delta}_t$ is unbiased, i.e., $\mathbb{E}[\hat{\Delta}_t] = \Delta$.*

Then, how well our $\hat{\Delta}_t$ can estimate Δ becomes the central problem that we need to solve. For any $t \in [n]$, the martingale difference of $M_t^{(1,2)}$ can be bounded by $|M_t^{(1,2)} - M_{t-1}^{(1,2)}| = 2 + 1/\pi_t(1) + 1/\pi_t(2)$. Moreover, the variance of the martingale $M_t^{(1,2)}$ can be bounded as $\sum_{t=1}^n \mathbb{E}[(\frac{R_t}{\pi_t(1)}\mathbb{I}_{A_t=1} - \frac{R_t}{\pi_t(2)}\mathbb{I}_{A_t=2} - \Delta)^2 \mid \mathcal{H}_{t-1}] \leq \sum_{t=1}^n \frac{1}{\pi_t(1)} + \frac{1}{\pi_t(2)}$. A common concern about IPW-based method is its large variance, especially when $\pi_t(a)$ is small (e.g., Dimakopoulou et al. 2021). The blessing of online learning is that we can control the propensity score. However, it also imposes an important challenge. Assigning a small probability to the suboptimal arm can bring us a small regret but a large variance of the estimators which may harm the inference. In the second phase of our EXP3E, $\frac{1}{\pi_t(1)} + \frac{1}{\pi_t(2)} \leq 2t^\alpha$ by the design of our algorithm. As for the first phase, from the proof of Theorem 3, our elimination rules contribute to securing $\frac{1}{\pi_t(a)} \leq 1 + e^2$ for $a \in \mathcal{A}$. In this way, we control the variance of $M_t^{(1,2)}$. By Bernstein's inequality (Freedman 1975), we can have the following theorem.

Theorem 5. *If Algorithm 1 runs with $\alpha \in [0,1]$ and $\delta < 2/e$, with probability at least $1-\delta$, for all $t \in [n]$,*

$$\left|\hat{\Delta}_t - \Delta\right| \leq \frac{(8e^2+16)\log\frac{2}{\delta}}{\sqrt{t^{1-\alpha}}}. \tag{5}$$

Furthermore, since we can take $\delta = \frac{1}{2n^2}$, $e(n,\hat{\Delta}) = \mathbb{E}[|\hat{\Delta}_n - \Delta|] = \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n^{1-\alpha}}})$.

Different from most existing results on adaptive estimators in the inference literature that are asymptotic in nature, Theorem 5 is a finite sample properties which has its own advantages when the number or samples are relatively small. Another important difference with offline inference lies in that Theorem 5 is an any-time bound since it holds for all $t \in [n]$. In addition, the RHS of Eq. (5) increases with the value of α. Intuitively, when α is small, EXP3E is more likely to explore during the second phase of our algorithm, which will help to estimate of the reward of the suboptimal arm, and thus improve our inference. However, there is no free lunch. In contrast, the regret upper bound in Theorem 3 will be large. Such observations illustrate the role that α plays in balancing online learning and inference.

Together with Corollary 1 and Theorem 5, we can safely draw the following statement.

Corollary 2. *For any instance $\nu \in \mathcal{E}_0$ and $\alpha \in [0,1]$, Algorithm 1 can guarantee $e_\nu(n,\hat{\Delta})\sqrt{\mathcal{R}_\nu(n,\pi)} = \widetilde{\mathcal{O}}(1)$. Furthermore, by Theorem 1, Algorithm 1 achieves the Pareto optimality for any $\alpha \in [0,1]$.*

For simplicity, we denote the online decision-making policy and ATE estimator with input parameter α in Algorithm 1 as $(\pi_\alpha, \hat{\Delta}_\alpha)$. By Theorem 1, the front of $(\pi_\alpha, \hat{\Delta}_\alpha)$ is $\mathcal{F}(\pi_\alpha, \hat{\Delta}_\alpha) = \{(n^{1-\alpha}, \frac{1}{\sqrt{n^{1-\alpha}}})\}$. With different input of $\alpha \in [0,1]$, the front of our $(\pi_\alpha, \hat{\Delta}_\alpha)$ will cover the line $e(n,\hat{\Delta})\sqrt{\mathcal{R}_\nu(n,\pi)} = 1$, which is the Pareto frontier. The front for $\alpha = 0$ coincides with that of RCTs.

2.2.3 The sufficient and necessary condition

Based on the series of $(\pi_\alpha, \hat{\Delta}_\alpha)$, the following theorem illustrates the necessity of the condition (3).The main idea is that any policy violating condition (3) will be Pareto dominated by $(\pi_\alpha, \hat{\Delta}_\alpha)$ for some α, and thus can not be Pareto optimal.

Theorem 6. *Any Pareto optimal $(\pi^*, \hat{\Delta}^*)$ satisfies $\max_{\nu\in\mathcal{E}_0}\left[e_\nu(n,\hat{\Delta}^*)\sqrt{\mathcal{R}_\nu(n,\pi^*)}\right] = \widetilde{\mathcal{O}}(1)$.*

Together with Theorems 2 and 6, we formally confirm the condition (3) is sufficient and necessary for the Pareto optimal solutions for $K = 2$.

Corollary 3. *The admissible pair $(\pi^*, \hat{\Delta}^*)$ is Pareto optimal to the minimax multi-objective optimization (1) if and only if $\max_{\nu\in\mathcal{E}_0}\left[e_\nu(n,\hat{\Delta}^*)\sqrt{\mathcal{R}_\nu(n,\pi^*)}\right] = \widetilde{\mathcal{O}}(1)$.*

In Figure 3, we present several examples of the possible fronts of the admissible policies. First, the lower bound in Eq. (2) tells that the front of any admissible policy has intersection of the region $\mathcal{R}(n,\pi) \gtrsim 1/(e(n,\hat{\Delta}))^2$ (the blue region). The sufficient and necessary condition indicates that the Pareto optimal solutions will only intersect with the blue region on the boundary $\mathcal{R}(n,\pi) \simeq 1/(e(n,\hat{\Delta}))^2$. In turn, any policy that intersects with the region on the boundary are Pareto optimal. The red curve is non Pareto optimal, since it partly falls into the interior of the region.

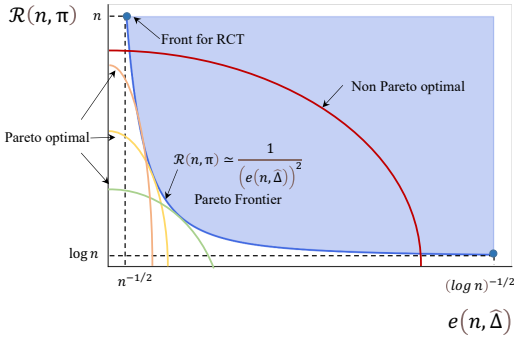

Figure 3: Examples of Pareto (non)optimal solutions.

3 Extension to General K

In this section, we extend our model, algorithm and analysis to a general $K \geq 2$. The main ideas of EXP3EG in Algorithm 2 follow from those of EXP3E. However, since the suboptimal arms are usually unlikely to be eliminated at the same time, EXP3EG can not be divided into two phases explicitly anymore. Following the same notatition as before, we assign $a \notin \mathcal{A}_t$ a time-varying but fixed probability $\alpha_t = \frac{1}{Kt^\alpha}$ to be played. For $a \in \mathcal{A}_t$, we assign the probability $(1 - |\mathcal{A}_t^c|\alpha_t)\frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{\sum_{a' \in \mathcal{A}_t} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}$. The elimination rule is still $\mathcal{A}_{t+1} = \mathcal{A}_t \backslash \{a : \hat{R}_t^{\max} - \hat{R}_t(a) > 2\sqrt{Ct}\}$. Moreover, notably, EXP3EG is powerful enough to output $\hat{\Delta}_{i,j}$ for all $i \neq j \in [K]$ at the same time, which means EXP3EG does not need to know which $\Delta_{i,j}$ is of interest in advance. We can have the following theorem on regret.

Theorem 7. *Let Algorithm 2 run with $\alpha \in [0, 1]$ and $\delta = \frac{1}{2n^2}$. The regret is $\mathcal{O}(\sum_{a \in [K]/\{a^*\}} \frac{\log(n)}{\Delta(a)} + \Delta(a) n^{1-\alpha} \log(n))$.*

When $\alpha = 1$, the regret upper bound in Theorem 7 matches with the minimax lower bounds for MAB problems up to a logarithmic factor. Also, since we mainly care about $|\Delta(a)| = \Theta(1)$, the regret bounds becomes $\widetilde{\mathcal{O}}(n^{1-\alpha})$. In Theorem 7, the regret upper bound is only dependent on the gaps with the optimal arm $\Delta(a)$ instead of all $|\Delta^{(i,j)}|$. Intuitively, $\Delta(a)$ has a more important role than $|\Delta^{(i,j)}|$, since the regret is defined to compete with the optimal arm.

For inference, following the notation defined in Section 2.2, we introduce a series of martingales $M_t^{(i,j)} := M_t^i - M_t^j$ for any $i \neq j \in [K]$, where recall that $M_t^i = \hat{R}_t(i) - \mu_i t$. An immediate result is unbiasedness, i.e., $\mathbb{E}[\hat{\Delta}_t^{(i,j)}] = \Delta^{(i,j)}$ for $t \in [n]$. We extend the result in Theorem 5 as following to a fixed pair of i, j.

Algorithm 2: EXP3E for general K (EXP3EG)

Input: α and δ
Initialization: $\mathcal{A}_1 = \{1, 2, \cdots, K\}$, $\hat{R}_0(a) = 0$ for $a \in \{1, \cdots, K\}$, $C = (4K^2(e^2+1)+2)^2(\log(\frac{2}{\delta}))^2$
for $t = 1, 2, \cdots, n$ **do**
  $\varepsilon_t = \frac{1}{\sqrt{Ct}}$, $\alpha_t = \frac{1}{Kt^\alpha}$;
  $\forall a \in \mathcal{A}_t$: $\pi_t(a) = (1 - |\mathcal{A}_t^c|\alpha_t)\frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{\sum_{a' \in \mathcal{A}_t} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}$;
  $\forall a \notin \mathcal{A}_t$: $\pi_t(a) = \alpha_t$;
  Select A_t according to π_t;
  Observe reward R_t;
  $\forall a \in \{1, \cdots, K\}$: $\hat{R}_t(a) = \hat{R}_{t-1}(a) + \frac{R_t}{\pi_t(a)}\mathbb{I}_{a=A_t}$;
  $\mathcal{A}_{t+1} = \mathcal{A}_t \backslash \{a : \hat{R}_t^{\max} - \hat{R}_t(a) > 2\sqrt{Ct}\}$;
  Output: For $i \neq j \in [K]$: $\hat{\Delta}_t^{(i,j)} = \frac{1}{t}(\hat{R}_t(i) - \hat{R}_t(j))$;
end for

Theorem 8. *If Algorithm 2 runs with $\alpha \in [0, 1]$, for any fixed $i, j \in [k]$, $i \neq j$, and $\delta < 2/e$, with probability at least $1 - \delta$, for all $t \in [n]$,*

$$|\hat{\Delta}_t^{(i,j)} - \Delta^{(i,j)}| \leq \frac{4(2 + 2K^2(1+e^2))\log\frac{2}{\delta}}{\sqrt{t^{1-\alpha}}}. \quad (6)$$

Particularly, after taking $\delta = \frac{1}{2n^2}$, we can derive $\max_{i<j\leq K} e(n, \hat{\Delta}_n^{(i,j)}) = \mathcal{O}(\frac{1}{\sqrt{n^{1-\alpha}}})$.

Then, together with Theorem 7, $(\max_{i<j\leq K} e_\nu(n, \hat{\Delta}_n^{(i,j)}))\sqrt{\mathcal{R}_\nu(n, \pi)} = \widetilde{O}(1)$ holds for any $\nu \in \mathcal{E}_0$. Taking i^* and j^* to be the best and the second best arm respectively, we always have $\max_{i<j\leq K} e(n, \hat{\Delta}_n^{(i,j)}) \geq e(n, \hat{\Delta}_n^{(i^*,j^*)})$. By such a fact, we can reduce the problem with $K > 2$ to $K = 2$. Theorem 2 can be easily generalized, and thus the sufficient condition for Pareto optimality can be $\max_{\nu\in\mathcal{E}_0}(\max_{i<j\leq K} e_\nu(n, \hat{\Delta}_n^{(i,j)}))\sqrt{\mathcal{R}_\nu(n, \pi)} = \widetilde{O}(1)$. Hence, Algorithm 2 is Pareto optimal for all $\alpha \in [0, 1]$. Then the necessity of the condition follows similarly from Theorem 6. We can naturally extend Corollary 3 as follows.

Theorem 9. *The admissible pair $(\pi^*, \hat{\Delta}^*)$ is Pareto optimal to the optimization problem (1) if and only if $\max_{\nu\in\mathcal{E}_0}[(\max_{i<j\leq K} e_\nu(n, \hat{\Delta}_n^{*(i,j)}))\sqrt{\mathcal{R}_\nu(n, \pi^*)}] = \widetilde{\mathcal{O}}(1)$.*

4 Discussion

In the previous sections, we restrict ourselves to the instance class $\mathcal{E}_0$, where $\Delta^{(i,j)} = \Theta(1)$ for all $i, j \in [K]$, which is usually referred to as the "well-separated" instance class (see, e.g., Kalvit and Zeevi 2021). Such an instance class allows us to see the magnitude of $\Delta^{(i,j)}$ as a universal constant independent of n and ignore its influence when deriving the necessary and sufficient condition for Pareto optimality. In this section, we first discuss about the case where $\Delta^{(i,j)}$ is extremely small comparing with the time horizon n or can even shrink with n (i.e., $\Delta^{(i,j)} = \mathcal{O}(n^{-p})$ for some strictly positive $p > 0$). For simplicity, we will focus on $K = 2$ since $K > 2$ naturally follows.

Case 1: $\Delta = \mathcal{O}(n^{-1/2})$. As known in current literature (see, e.g., Lattimore and Szepesvári 2020 and Kalvit and Zeevi 2021), approximately $\frac{1}{\Delta^2}$ samples are unavoidable to distinguish between two distributions with means separated by Δ. This indicates that if $\Delta = \mathcal{O}(n^{-1/2})$, even the most efficient adaptive algorithm which only cares about the regret rate will spend $\Theta(n)$ samples on the suboptimal arm and thus the regret is doomed to be roughly $n\Delta$. No one can expect to increase the statistical power by sacrificing the online decision-making efficiency, which itself has no room to be sacrificed. Therefore, the main question becomes what level the estimation error can be controlled. By slight modifications of the proof of Theorem 5, we can show our estimator $\hat{\Delta}$ can achieve $e(n, \hat{\Delta}) = \widetilde{O}(n^{-1/2})$ when $\Delta = \mathcal{O}(n^{-1/2})$, which can not be further improved

either. In this case, we have the strongest statistical power and an unavoidably large regret.

Case 2: $\Delta = \Omega(n^{-1/2})$ and $\Delta = \mathcal{O}(n^{-(1-\alpha)/2})$. Recall that α is the input of EXP3E controlling the trade-off between the two objectives. In this case, the regret upper bound in Theorem 3 reduces to $\widetilde{\mathcal{O}}(\frac{\log(n)}{\Delta})$, which is not influenced by α and matches with the regret lower bound in the worst case (see, e.g., Lattimore and Szepesvári 2020). This means that when $n^{-1/2} \lesssim \Delta \lesssim n^{-(1-\alpha)/2}$, our EXP3E always has the optimal efficiency in online decision-making. Note that Theorem 5 still holds here, i.e., $e(n, \hat{\Delta}) = \widetilde{O}(n^{-(1-\alpha)/2})$. We want to point out that by a simple modification of Lemma **??**, in this case one feasible lower error bound is $\Omega(\sqrt{\frac{\Delta}{\log(n)}})$, which we do not match with. Such kind of mismatch may be caused by the large variance of IPW-based estimator, especially during the second phase of the algorithm. It is also possible that $\Omega(\sqrt{\frac{\Delta}{\log(n)}})$ underestimates the difficulty of the problem with $n^{-1/2} \lesssim \Delta \lesssim n^{-(1-\alpha)/2}$. We leave this issue to our future research.

Case 3: $\Delta = \Omega(n^{-(1-\alpha)/2})$ and $\Delta < \Theta(1)$. In this case, Theorem 3 offers a regret upper bound of $\widetilde{\mathcal{O}}(n^{1-\alpha}\Delta)$, whose proof inplies that the algorithm plays the suboptimal arm exact $\Theta(n^{1-\alpha})$ times. And thus, by an easy extension of Lemma **??**, the $\widetilde{O}(n^{-(1-\alpha)/2})$ error bound offered by Theorem 5 is rate optimal. The problem here is that since we do not know the order of the magnitude of Δ in advance, we can hardly control the regret to the level that we want in the order of n. That is the price for the strong statistical power when $\Delta = \Omega(n^{-(1-\alpha)/2})$.

Up till now, we have discussed that EXP3E is somewhat optimal in some other senses under different orders of the small Δ. However, what is the sufficient and necessary condition for the desired Pareto optimal for extremely small Δ is still unknown and we leave it to our future work.

The second aspect to consider is the neglect of constants and logarithm terms in defining Pareto optimality. While this is a common practice in MAB, ignoring these terms can be significant, especially in cases where sample collection is costly or the number of samples is limited. Admittedly, such a simplification is another limitation of our work. It is very challenging to get optimal rates on the dependence of both constant and n. Even for the traditional MAB problem without the added complexity of inference, there is a lack of research addressing the best achievable dependence on constants. Another important consideration is the choice of the parameter α, which plays a crucial role in our design. Choosing the appropriate α can be challenging in some cases, and there may be scenarios where a dynamic α would be more appropriate, especially in experiments that have a long duration. Understanding how people make decisions about choosing α is also important, but it is outside the scope of this work. Finally, a crucial next step is extending our work to continuous arm bandit problems, as our design and analysis are currently limited to discrete arms.

5 Concluding Remarks

In this paper, we statistically investigate the trade-off between efficiency in decision-making and statistical power of ATE in MAB experiments. We novelly introduce the general minimax multi-objective optimization framework and Pareto optimality to formally describe and theoretically analyze such a trade-off. Moreover, we derive a useful sufficient and necessary condition for Pareto optimal designs, i.e., $(\max_{i<j\leq K} e_\nu(n, \hat{\Delta}^{*(i,j)}))\sqrt{\mathcal{R}_\nu(n, \pi^*)} = \widetilde{\mathcal{O}}(1)$ for any instance $\nu \in \mathcal{E}_0$. Additionally, we propose an efficient Pareto optimal design with $max_{i<j\leq K} e_\nu(n, \hat{\Delta}^{(i,j)}) = \mathcal{O}(n^{-(1-\alpha)/2})$ and $\mathcal{R}_\nu(n, \pi) = \mathcal{O}(n^{1-\alpha})$ for any give $\alpha \in [0, 1]$ controlling the desired level of trade-off.

References

Adusumilli, K. (2021). Risk and optimal policies in bandit experiments. *arXiv preprint arXiv:2112.06363*.

Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, S., Koolen, W. M., and Juneja, S. (2021). Optimal best-arm identification methods for tail-risk measures. *Advances in Neural Information Processing Systems*, 34:25578–25590.

Angrist, J. and Imbens, G. (1995). Identification and estimation of local average treatment effects.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

Atan, O., Zame, W. R., and Schaar, M. (2019). Sequential patient recruitment and allocation for adaptive clinical trials. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1891–1900. PMLR.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28.

Bhat, N., Farias, V. F., Moallemi, C. C., and Sinha, D. (2020). Near-optimal ab testing. *Management Science*, 66(10):4477–4495.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., and van Der Laan, M. (2021). Post-contextual-bandit inference. *Advances in neural information processing systems*, 34:28548–28559.

Bojinov, I., Rambachan, A., and Shephard, N. (2021). Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196.

Bojinov, I., Simchi-Levi, D., and Zhao, J. (2020). Design and analysis of switchback experiments. *arXiv preprint arXiv:2009.00148*.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings.

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R., and Auer, P. (2011). Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 189–203. Springer.

Chan, H. P. and Lai, T. L. (2006). Sequential generalized likelihood ratios and adaptive treatment allocation for optimal sequential selection. *Sequential Analysis*, 25(2):179–201.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24.

Chen, N., Gao, X., and Xiong, Y. (2022). Debiasing samples from online learning using bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 8514–8533. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Deshmukh, A. A., Dogan, U., and Scott, C. (2017). Multi-task learning for contextual bandits. *Advances in neural information processing systems*, 30.

Dimakopoulou, M., Ren, Z., and Zhou, Z. (2021). Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2019). Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453.

Dudík, M., Erhan, D., Langford, J., and Li, L. (2014). Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Eckles, D., Karrer, B., and Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1).

Erraqabi, A., Lazaric, A., Valko, M., Brunskill, E., and Liu, Y.-E. (2017). Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR.

Fan, L. and Glynn, P. W. (2021). The fragility of optimized bandit algorithms. *arXiv preprint arXiv:2109.13595*.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR.

Farias, V., Moallemi, C., Peng, T., and Zheng, A. (2022a). Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*.

Farias, V. F., Li, A. A., Peng, T., and Zheng, A. T. (2022b). Markovian interference in experiments. *arXiv preprint arXiv:2206.02371*.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.

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

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer.

Glynn, P. W., Johari, R., and Rasouli, M. (2020). Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15).

Hahn, J., Hirano, K., and Karlan, D. (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108.

Jennison, C., Johnstone, I. M., and Turnbull, B. W. (1982). Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical decision theory and related topics III*, pages 55–86. Elsevier.

Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. (2022). Experimental design in two-sided platforms: An analysis of bias. *Management Science*.

Johari, R., Pekelis, L., and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. *Advances in neural information processing systems*, 31.

Kalvit, A. and Zeevi, A. (2021). A closer look at the worst-case behavior of multi-armed bandit algorithms. *Advances in Neural Information Processing Systems*, 34:8807–8819.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

Kato, M. and Ariu, K. (2021). The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*.

Kato, M., Ishihara, T., Honda, J., Narita, Y., et al. (2020). Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

Khan, S. and Ugander, J. (2021). Adaptive normalization for ipw estimation. *arXiv preprint arXiv:2106.07695*.

Lai, T. L., Robbins, H., et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org.

Li, L., Munos, R., and Szepesvári, C. (2015). Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR.

Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2018). Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR.

Offer-Westort, M., Coppock, A., and Green, D. P. (2021). Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4):826–844.

Qin, C. and Russo, D. (2022). Adaptivity and confounding in multi-armed bandit experiments. *arXiv preprint arXiv:2202.09036*.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Seldin, Y., Auer, P., Shawe-taylor, J., Ortner, R., and Laviolette, F. (2011). Pac-bayesian analysis of contextual bandits. *Advances in neural information processing systems*, 24.

Seldin, Y., Cesa-Bianchi, N., Auer, P., Laviolette, F., and Shawe-Taylor, J. (2012). Pac-bayes-bernstein inequality for martingales and its application to multiarmed bandits. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 98–111. JMLR Workshop and Conference Proceedings.

Seldin, Y., Szepesvári, C., Auer, P., and Abbasi-Yadkori, Y. (2013). Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. In *European Workshop on Reinforcement Learning*, pages 103–116. PMLR.

Simchi-Levi, D., Zheng, Z., and Zhu, F. (2022). A simple and optimal policy design with safety against

heavy-tailed risk for multi-armed bandits. *arXiv preprint arXiv:2206.02969*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Swaminathan, A. and Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.

Wager, S. and Xu, K. (2021). Experimenting in equilibrium. *Management Science*, 67(11):6694–6715.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, Y.-X., Agarwal, A., and Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR.

Xiong, R., Athey, S., Bayati, M., and Imbens, G. (2019). Optimal experimental design for staggered rollouts. *arXiv preprint arXiv:1911.03764*.

Xu, M., Qin, T., and Liu, T.-Y. (2013). Estimation bias in multi-armed bandit algorithms for search advertising. *Advances in Neural Information Processing Systems*, 26.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. (2017). A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Advances in Neural Information Processing Systems*, 30.

Yao, J., Brunskill, E., Pan, W., Murphy, S., and Doshi-Velez, F. (2021). Power constrained bandits. In *Machine Learning for Healthcare Conference*, pages 209–259. PMLR.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135.

Zhang, K., Janson, L., and Murphy, S. (2020). Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829.

Zhang, K., Janson, L., and Murphy, S. (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in neural information processing systems*, 34:7460–7471.

Zhong, Z., Cheung, W. C., and Tan, V. Y. (2021). On the pareto frontier of regret minimization and best arm identification in stochastic bandits. *arXiv preprint arXiv:2110.08627*.

Zhou, Z., Athey, S., and Wager, S. (2022). Offline multi-action policy learning: Generalization and optimization. *Operations Research*.