Short-Term Pain for Long-Term Gain: Adaptive Experiment with Post-Commitment Reward Shift

Puping Jiang* Wei Tang†

Abstract

Decision-makers in learning environments face a dilemma when their short-term optimal actions may not favor their long-term benefits the most. To understand the fundamental tradeoff behind the dilemma, we consider an adaptive experiment with post-commitment reward shift problem, which consists of two phases: (1) an experiment phase, where the decision-maker is free to experiment with different options, and (2) a commitment phase, where they must commit to a single option for the remainder of the phase. More importantly, each option's reward may change after the commitment. Our main result is a simple yet effective algorithm to solve the experimenter's problem. Our algorithm follows an arm-elimination framework but with special treatments dedicated to dealing with the post-commitment reward shift. We analyze the regret upper bounds of our algorithm across all parameter regimes, and we provide matching lower bounds. We further extend our analysis and results to the setting (1) when the experimenter has prior structural knowledge about the reward shift; (2) when the experimenter has concave rewards in the commitment phase.

*Antai College of Economics and Management, Shanghai Jiao Tong University. Email: jiangpuping@sjtu.edu.cn
†The Chinese University of Hong Kong, Email: weitang@cuhk.edu.hk

1 Introduction

Balancing short-term benefits against long-term goals, especially in an uncertain decision-making environment, has long challenged decision-makers. The dilemma is especially acute when the decision that maximizes near-term rewards conflicts with what is most desirable in the long run. It is often the case that the decision-maker can experiment with multiple options to identify which might be most suitable for the current decision task, yet they must eventually commit to an option for long-run goals.

Practitioners and researchers alike have debated how to reconcile these competing objectives in applications ranging from online platform experimentation to technology development and product manufacturing. For example, online platforms have experienced significant regulatory environment policy shifts with the EU's General Data Protection Regulation (GDPR) – adopted in 2016 and applicable from May 25, 2018 (Johnson et al., 2023; Aridor et al., 2023). During the two-year implementation window before GDPR took effect, a platform needs to test out new strategies, preparing for a platform-wide transition after the environment shift. Examples of pre-GDPR testing and staged rollouts include Criteo's pre-GDPR consent-interface tests (Tiku, 2018), IAB Europe's launch of the Transparency & Consent Framework in April 2018 (IAB Europe), and Google's March 2018 GDPR policy updates (Ads, 2018). Anticipated policy shifts – ranging from environmental regulation (e.g., Clean Air Act; EU ETS) to trade and geopolitical tensions – have also repeatedly altered firms' supply chain, production, and investment decisions. For example, Shapiro and Walker (2018) show U.S. environmental regulation (Clean Air Act) drove large changes in manufacturing emissions – clear evidence that actual policy shifts reshape firm manufacturing decisions; Colmer et al. (2025) provide firm-level evidence that carbon pricing under the EU ETS changes production choices and investments. Firms have to experiment with substitute production plans in order to achieve a resilient transition (Tomlin, 2009; Ang et al., 2017).

Figure 1 depicts the high-level operational problem the firms encounter in the environment discussed in this paper. As illustrated in Figure 1, before the anticipated environment shift, firms

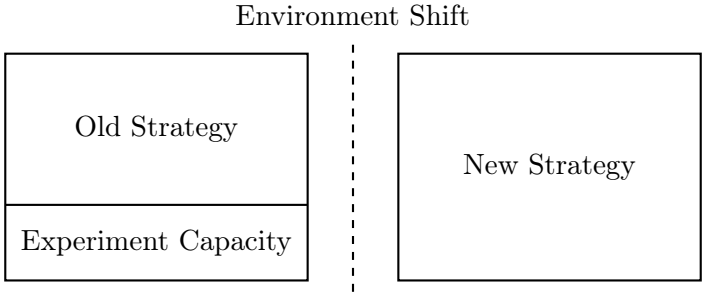

Figure 1: Illustration of Experiment and Commitment Transition

would typically reserve a relatively small portion of their capacity (for online platforms, this represents a portion of users and computing resources; for manufacturing firms, this represents a portion of production capacity and material resources) for experiments. The experiment capacity can be used not only to test strategies that might attain promising performance after the environment shift, but also to explore and roll out strategies that could deliver high payoffs under the current operational environment. While the experiment capacity is flexible for strategy switches, the strategy run on the remaining capacity is generally restricted to be stable due to cost and other practical concerns. However, a whole-capacity-wide strategy transition is inevitable after the environmental shift, since the old strategy could become highly unfavorable or even infeasible in the new envi-

ronment. The key decisions firms have to make are the experiment design within the experiment capacity and the new strategy to commit to after the shift. The fundamental question we aim to address is:

What is the rule of thumb for the optimal experiment design when there is an anticipated environment shift?

Such a decision dilemma is becoming more and more relevant as many industrial sectors are facing more and more anticipated economic and political environment shifts, and aim to build more resilient and sustainable supply chains (de Araujo and Robbins, 2019). Earlier academic work derives qualitative observations on firms' optimal exploration effort under changing environments (see, e.g., Posen and Levinthal, 2012). Our paper advances the understanding of this problem from an adaptive learning perspective via a parsimonious model. Before we briefly mention the high-level framework of our model, we elaborate more about the two aforementioned examples here.

Example 1.1 (Platform regulation shift: GDPR example)**.** *The EU announced GDPR on April 27th, 2016, which was scheduled to be enacted on May 25th, 2018.GDPR led to a significant impact on leading tech companies. Anticipating the policy environment change, an online platform could experiment with various operational strategies in preparation for the new rule, including strategies of data collection, advertisement, pricing, etc. Consider a platform that aims to transition globally to an interface design more compliant with the new regulations following the enactment of GDPR (empirical evidence shows significant industry-wide strategy switches after the enactment of GDPR, see Peukert et al. (2022), Aridor et al. (2023)). Typically, before the platform-wide transition, the platform would experiment within a relatively small group of users, i.e., beta users. Crucially, a strategy that will perform best after the rollout of new regulations may not be optimal for current operations (e.g., failing to comply with GDPR may lead to fines up to 4% of annual global revenue (see Cutbill (2019))). Routing a share of beta traffic to any candidate design yields a random outcome – (*CONVERSION-RATE, PRIVACY-LEVEL*), which follows a distribution tied to the given design. The platform's payoff can be simplified as* PROFIT−LEGAL-COST*, where* PROFIT *and* LEGAL-COST *are functions of* CONVERSION-RATE *and* PRIVACY-LEVEL*, respectively. At the enforcement date, the* LEGAL-COST *function shifts, so the same* PRIVACY-LEVEL *may imply a different cost than before. Finding the optimal strategies for the long run while balancing the short-term gains from the beta users poses a major challenge for the platform's experiment effort allocation.*

Example 1.2 (Manufacturing supply chain examples)**.** *In manufacturing contexts, state-level regulations limiting gas vehicle development, scheduled enactment of carbon tariffs, and other environmental regulations put automobile manufacturers in a similar position (see Blenkinsop (2025)). The profitability of different car models can be impacted quite differently depending on their carbon emission volumes. In some industrial contexts, such as the semiconductor industry, export restrictions led by geopolitical tensions force firms in key supply chains to explore substitute solutions before they lose access to critical components or technology (see The Economist (2024)). Different supply chains can be impacted differently depending on their resilience to those export restrictions. As another example, consider manufacturing firms that would experience per-unit cost reduction after mass production launches. The mass production facilities often start construction way before the experiments and trial production end. Some production plans may enjoy more cost reduction than others during mass production. For example, products with cost structures weighted more towards manufacturing expenses can benefit more from automation, learning curve effects, and other factors. Importantly, firms in manufacturing supply chains typically reserve a portion of their operating capacity for experiments during the grace period, after which mass production of substitutes or transitions of business structures across all capacities must roll out.*

There are two key features of these motivating examples: (i). *Two-phase horizon.* The time horizon consists of two phases, an *experiment phase* and a *commitment phase*. The commitment date is known in advance and is typically marked by an exogenous environment shift. Before the commitment date, experiments can be conducted due to the flexible nature of the experiment capacity. A long-term commitment must be made after the commitment date due to the inflexible nature of the global strategy transition, the construction of mass production capacity, etc. (ii). *Predictable reward shift.* The reward shift after the policy environment shift is largely predictable. For example, the legal costs associated with a given interface design are quantifiable in advance under GDPR, as long as the regulations are specified clearly. Similarly, the taxation, tariffs, or even fines imposed on producing products with high carbon emissions are also quantifiable in advance once the laws are announced. Export restrictions on key industrial components would turn production plans that heavily rely on those components infeasible, i.e., the rewards simply go to zero. Such export restrictions are often either announced in advance to give supply chains a grace period or are predictable due to geopolitical games.

As we can see from the above examples, decision-makers (i.e., the online platforms, the manufacturing firms) often face the dilemma of short-term pain versus long-term gain, and it is challenging to determine how much resource should be allocated to exploration in preparation for a future shift. Our work contributes to resolving the above decision-maker's problem from a quantitative perspective, with the goal of maximizing the following objective:

$$\text{Short-term earning} + \text{Long-term earning}.$$

Clearly, the relative size of the long-term earnings compared to the short-term earnings plays a key role in the decision-maker's experiment design. One possible interpretation of the size factor could be the relative length of the commitment. For online platforms, new policy tests can take weeks to months, while the global deployment could last for months to years. For manufacturing sectors, like the semiconductor and automobile industries, experiments for a new generation of products can take a much longer time, like years, while the commercial lives also typically last for years. Following this interpretation, we consider an adaptive experiment problem with post-commitment reward shift: The decision-maker faces a sequential decision-making problem over a total of T periods and must choose from K candidate treatments. The whole process consists of two phases: In the first phase (the first N rounds referred to as the *experiment phase*), she can experiment with any different treatments. The decision-maker initially has no prior knowledge of the mean reward of each treatment but can adaptively choose one of the treatments to observe a corresponding random outcome generated from this treatment and collect the corresponding reward. Importantly, this experimentation budget N is exogenously determined by the environment. In the second phase (the remaining $T - N$ rounds, referred to as the *commitment phase*), the decision-maker must commit to a single treatment and implement it throughout this phase. However, due to anticipated environment changes after N periods, a reward shift may occur for each treatment in the commitment phase. As a result, the treatment that appeared to be optimal during the experiment phase may become suboptimal in the commitment phase.

Notice that our objective has another interpretation that can characterize other important problems. We can think of each treatment (or arm) as having two attributes: the mean reward and an attribute that the decision-maker cares about only after the commitment. The decision-maker has to select one treatment after the experiment ends. The decision-maker faces the problem of maximizing the cumulative reward while ensuring that the selected treatment performs well in the second attribute (or, conversely, maximizing the selected treatment's performance in the second attribute while guaranteeing a decent cumulative reward). If we write the problem in Lagrangian form, the formulation returns to the aforementioned model, with the relative length in the previous

interpretation becoming the Lagrangian multiplier, which should now be interpreted as the relative importance the decision-maker attaches to the second attribute.

Under this framework, our main research question can now be stated as:

> *What is the optimal way to address the exploration-exploitation dilemma in the presence of such a post-commitment reward shift?*

We seek to answer the above question in two aspects: (1) whether an efficient online learning algorithm can achieve provable performance, and (2) what the optimal performance guarantee is for any online algorithm.

We measure the performance of online learning algorithms by the notion of *Regret*, which is the opportunity cost of our algorithm compared against the optimal clairvoyant strategy – one that knows the best treatments for both the experiment and commitment phases. Formally, it is defined as the difference between the total reward accumulated by an optimal clairvoyant strategy and the reward achieved by our online learning algorithm.

1.1 Main Results

Our main results. Our first main result introduces a simple yet effective online learning algorithm, termed as *Reserved Arm Eliminations for Commitment* algorithm (RAEC), which achieves provable performance across all parameter regimes of experimentation budget N. At a high level, RAEC operates as an elimination-based algorithm with phased learning. During the experiment phase, it adaptively eliminates suboptimal arms using increasingly stringent hypothesis tests. In particular, in Algorithm RAEC, a predetermined portion of the experiment phase is reserved for identifying the optimal arm for the commitment phase, while the remaining rounds will focus on standard regret minimization. A key feature of Algorithm RAEC is that the confidence intervals used for eliminating suboptimal arms are constructed to account for the commitment constraint after N rounds. RAEC's simplicity enables us to derive a full spectrum of regret upper bounds across different parameter settings. Throughout this work, we use $\widetilde{O}$ and $\widetilde{\Omega}$ to denote upper and lower bounds on the growth rate (up to logarithmic factors), and $\widetilde{\Theta}$ to characterize matching rates (up to logarithmic factors).

- *Short experiment scenario* (i.e., $N \leq (T-N)^{2/3}(K\log(T-N))^{1/3}$): The regret bound is $\widetilde{O}\left(\sqrt{KT^2/N}\right)$;
- *Balanced scenario* (i.e., $N \geq (T-N)^{2/3}(K\log(T-N))^{1/3}$ and $(T-N) \geq K^{1/4}T^{3/4}/(\log(T-N))^{1/2}$): The regret bound is $\widetilde{O}\left(K^{1/3}(T-N)^{2/3}\right)$;
- *Short commitment scenario* (i.e., $(T-N) < K^{1/4}T^{3/4}/(\log(T-N))^{1/2}$): The regret bound is $\widetilde{O}\left(\sqrt{KT}\right)$.

Beyond the algorithmic results, we establish an information-theoretic lower bound that matches the upper bound achieved by RAEC, which proves that the regret bound cannot be further improved by other algorithms (in a certain sense). We define the instance-independent regret lower bound (i.e., minimax lower bound) across different parameter regimes as follows:

- *Short experiment scenario*: The experiment phase is relatively short, and the optimal strategy is to focus on pure exploration. The minimax lower bound is $\Omega\left(\sqrt{\frac{KT^2}{N}}\right)$;

Experiment budget N	$N \lesssim K^{1/3}T^{2/3}$	$K^{1/3}T^{2/3} \lesssim N \lesssim T - K^{1/4}T^{3/4}$	$N \gtrsim T - K^{1/4}T^{3/4}$
Optimal regret $\mathbf{REG}[N,T]$	$\widetilde{\Theta}\left(\sqrt{\frac{KT^2}{N}}\right)$	$\widetilde{\Theta}(K^{1/3}\left(T-N\right)^{2/3})$	$\widetilde{\Theta}(\sqrt{KT})$

Table 1: The optimal regret for different parameter regimes. Here, we use $A \lesssim B$, $A \gtrsim B$ to denote $A = \widetilde{O}(B)$, $A = \widetilde{\Omega}(B)$, respectively.

- *Balanced scenario* The experiment phase is sufficiently long, but so is the commitment phase, which makes this case particularly more interesting. Here, the minimax lower bound is $\Omega\left(K^{1/3}(T-N)^{2/3}\right)$;
- *Short commitment scenario*: The commitment phase is short, making the problem resemble a standard K-armed bandit setting. The minimax lower bound is the well-known $\Omega\left(\sqrt{KT}\right)$.

Our results highlight that while asymptotic optimality (sublinear regret) remains achievable across different parameter regimes of N, the problem is inherently more challenging due to the reward shift between the experiment and commitment phases. Interestingly, the regret lower bound grows more slowly in K compared to traditional MAB problems. This is because the presence of a reward shift encourages the optimal strategy to explore more arms in the experiment phase. Consequently, the regret lower bound becomes less sensitive to increases in K. In summary, our work provides a tight characterization of the regret landscape in adaptive experimentation with post-commitment reward shift. Our results provide a quantitative framework for decision-makers balancing the short-term vs. long-term trade-off while offering guidance on how much effort should be invested in exploring optimal solutions for long-term gains and what costs can be expected in the process. Another interesting takeaway is that our algorithm predetermines the effort allocated to explore good arms of the commitment phase, and it is shown to be good enough. More sophisticated adaptive methods while learning the instance structure will not fundamentally improve further.

Finally, we advance our discussions to two extensions. The first extension is that, with a more nuanced ex-ante understanding of the reward shift structure, one can improve regret performance. In practical applications, ex-ante knowledge of reward shifts can arise from decomposing the various factors influencing rewards. For example, macroeconomic conditions may lead to a universal shift in rewards across all arms, while industry-specific factors, such as export restrictions on key components, may affect different arms disproportionately. These factors may not only cause significant shifts in reward values across different arms but may also alter the relative ranking of arms' mean rewards. Our results clarify that, in terms of regret minimization, correctly identifying the ranking-changing effect in reward shifts is far more critical than specifying the absolute magnitude of these shifts.

The second extension is that, instead of committing to a single arm, the decision-maker can commit to a portfolio of arms in the commitment phase, and the total reward in the commitment phase is a concave function of the sum of the observed outcomes. For example, the short-term payoff of a firm is the total profit, while the long-run payoff is linear in profit minus the convex increasing penalty of environmental impact (i.e., the marginal penalty increases as the environmental impact grows). In this setting, the optimal decision in the commitment phase corresponds to selecting an optimal distribution of arms rather than a single arm. This model strictly generalizes our baseline model. We propose a new algorithm, *Reserved Online Stochastic Convex Optimization for*

Commitment (ROSCOC), which, like RAEC, allocates a predetermined portion of the experiment phase to learning the optimal commitment decision. However, instead of using an arm-elimination technique, ROSCOC uses an online stochastic convex optimization approach to optimize the arm portfolio selection. Unlike the arm-elimination technique that will specify the optimal arm at the end, the online stochastic convex optimization only generates an execution history. We propose the idea of using the execution history as a proxy for the portfolio that will be committed. Our analysis shows that ROSCOC achieves a similar regret upper bound as in Table 1. Since this model is a strict generalization of the baseline model, our derived upper bound is also tight. Although this generalized problem appears to be much more complicated, we demonstrate that the effectiveness of the predetermined exploration effort allocation for the commitment phase still applies, and the problem is not fundamentally harder than the baseline problem.

1.2 Related Work

Multi-armed bandits (MAB) is a classical framework to study how to balance exploitation and exploration in sequential decision-making problems. Two archetypal objectives – regret minimization and best-arm identification – have been extensively studied in the MAB literature. Seminal works on regret minimization include the class of Upper Confidence Bound (UCB) type algorithms (Auer et al., 2002), and the class of Thompson sampling (TS) algorithms (Agrawal and Goyal, 2012; Russo and Van Roy, 2014; Agrawal and Goyal, 2017). While for any online algorithms, Lai and Robbins (1985) characterize asymptotic lower bounds on the expected regret for the MAB problems.

Algorithms designed for regret minimization typically explore different options without committing to a single arm. However, in practice, many applications may prefer a commitment to action instead of continuous exploration. In light of this motivation, many works have been devoted to studying the best-arm identification. In this line of work, two settings are considered: (i) a fixed-budget setting – given a time budget T, the decision-maker aims to maximize the probability of finding the optimal arm in at most T steps (Audibert et al., 2010; Karnin et al., 2013); (ii) a fixed-confidence setting – given a confidence level $\delta > 0$, the decision-maker aims to find the optimal item with the probability of at least $1 - \delta$ in the smallest number of time steps (Bubeck et al., 2013; Kaufmann and Kalyanakrishnan, 2013). Kaufmann et al. (2016) examines the lower bound complexity of both settings. Unlike the algorithms for regret minimization, the best-arm identification algorithms do not account for the regret incurred during exploration.

To mitigate the drawbacks of regret minimization and best-arm identification, more recently, many works (see, e.g., Bui et al., 2011; Degenne et al., 2019; Kim et al., 2023; Zhang and Ying, 2023; Zhong et al., 2023; Simchi-Levi and Wang, 2023; Qin and Russo, 2024; Yang et al., 2024) have started to explore designing algorithms that can achieve best-of-both-worlds guarantees in various unified frameworks. However, most of these works either consider that the experimenter is allowed to choose when to stop the experimentation adaptively, given the collected information so far, or assume that the post-commitment reward remains the same as the reward in the experiment phase. Among these works, a particularly relevant work is by Bui et al. (2011), where the authors also consider that the experimenter needs to stop the experimentation within an exogenously given budget, and the reward earned after the commitment is scaled up by a constant factor. In other words, there exists a linear reward shift in the commitment phase. Our work considers a more general reward shift structure, and we provide a more complete characterization of regret bounds with different parameter regimes of the experimentation budget. It is worth mentioning that Qin and Russo (2024) also consider different before- and post-commitment costs, but their main focus is a setting where the experimenter can adaptively choose when to stop. Although our model shares some conceptual similarities with Qin and Russo (2024), due to the different objective structures,

there is no straightforward reduction from our setting to match theirs.

Conceptually, our work also relates to recent growing literature that uses regret minimization to study the adaptive experimentation in online platforms (to name a few, see, e.g., Bibaut et al., 2021; Athey and Wager, 2021; Farias et al., 2022; Hu et al., 2024). Our work differs from these works as we consider an adaptive experimentation problem with commitment and potential reward shift after the commitment. Notably, some earlier work studies firms' exploration effort under changing environments also derives insights under the bandit framework, e.g., Posen and Levinthal (2012) uses simulation methods to obtain qualitative observations on this issue.

Should cite Simchi-Levi, Wang, Zheng 2026 MS.

2 Preliminaries

We consider a sequential experiment with a commitment problem. There are T different experimental units participating in the experiment sequentially, where T is fixed and known to the experimenter. Upon the arrival of each t-th subject, the experimenter assigns a treatment (arm), among K independent options. The whole process consists of two phases.

The first phase is the *experiment phase*, which includes the first N ($N < T$) periods, where N is fixed and known to the experimenter. In each period $t \in [N]$, the experimenter can pick one treatment $I_t \in [K]$ ($K < N$) to assign it to the subject. The environment generates a random outcome $o_{t,I_t} \in O$ which is independently and identically drawn from a fixed unknown outcome distribution ν_{I_t} over a common outcome space O. The outcome space could be a vector space, in which case a random outcome could be interpreted as the random feature vector of a treatment. Each ν_i belongs to a general distribution space $\mathcal{V}$. We do not make any assumptions on the distribution of the observed outcomes. As will soon become clear, what matters is the corresponding reward distribution. The experimenter will observe the realized outcome o_{t,I_t} and also collect a reward $f(o_{t,I_t})$ where function $f(\cdot) : O \to [0,1]$ is a known reward function in the experiment phase.

The second phase is the *commitment phase* which includes the remaining periods, namely, from $N+1$ to T. In each period $t > N$ of the commitment phase, the experimenter commits to assign the same treatment, denoted as $J_N \in [K]$ and collects a realized reward $g(o_{t,J_N})$ where $g(\cdot) : O \to [0,1]$ is the known reward function in the commitment phase. This reward function is not necessarily the same as the reward function $f(\cdot)$ and is known to the experimenter. Here, the outcome o_{t,J_N} is also independently and identically distributed according to the same distribution ν_{J_N} as in the experiment phase.

The experimenter initially knows the experimentation budget N and both reward functions f, g. Her goal is to design a policy π, which executes over the two decision phases, i.e., $\pi = (\pi_{\mathsf{exp}}, \pi_{\mathsf{com}})$, a sequential experiment policy π_{exp} and a commitment policy π_{com}, both possibly randomized, to maximize the total expected rewards. More formally, the sequential experiment policy π_{exp} is a function that maps the total time horizon T, the experiment length N, the historical observations $(I_s, o_{s,I_s})_{s\in[t-1]}$, and also a random variable U, which encodes any additional sources of randomization, to an assignment $I_t = \pi_{\mathsf{exp}}(N, T, (I_s, o_{s,I_s})_{s\in[t-1]}, U) \in [K]$. Similarly, the commitment policy π_{com} is a function that maps all historical observations up to the end of N-th period, together with the knowledge of T, N and an additional random variable U to an assignment $J_N = \pi_{\mathsf{com}}(N, T, (I_s, o_{s,I_s})_{s\in[N]}, U) \in [K]$. Specifically, the experimenter aims to maximize the following cumulative rewards:

$$\mathbf{REW}_{\pi}[N,T] = \sum_{t=1}^{N} \mathbb{E}[f(o_{t,I_t})] + \sum_{t=N+1}^{T} \mathbb{E}[g(o_{t,J_N})] \ , \tag{1}$$

where the expectations are over the randomness of the policy π, and the randomness from the outcome distributions $(\nu_i)_{i\in[K]}$. Directly optimizing (1) is not tractable given that the parameters $(\nu_i)_{i\in[K]}$ are not known to the experimenter a priori. Thus, it is also more convenient to focus on the regret. The objective then is to design the policy π that minimizes the regret defined as:

$$\mathbf{REG}_\pi[N,T] = \sum_{t=1}^{N}\left(\mathbb{E}\Big[f(o_{t,I_f^*})\Big] - \mathbb{E}[f(o_{t,I_t})]\right) + \sum_{t=N+1}^{T}\left(\mathbb{E}\Big[g(o_{t,I_g^*})\Big] - \mathbb{E}[g(o_{t,J_N})]\right) \ , \quad (2)$$

where $I_f^* \triangleq \arg\max_{i\in[K]} \mathbb{E}_{o\sim\nu_i}[f(o)]$ denotes the optimal treatment w.r.t. the reward function f, and $I_g^* \triangleq \arg\max_{i\in[K]} \mathbb{E}_{o\sim\nu_i}[g(o)]$ denotes the optimal treatment w.r.t. the reward function g. Throughout the paper, we assume $N \geq \Omega\left(\mathsf{poly}(T)\right)$, i.e., N grows at least polynomially in T. This assumption is used to rule out the less interesting scenario when the length of the experiment phase, N, is extremely small compared to the whole time horizon, T.

We now instantiate our setting using the previously mentioned GDPR and manufacturing example.

Example 2.1 (The GDPR example (continued))**.** *One treatment $I_t \in [K]$ could represent one interface design. The length of the experiment phase N is the length of the grace period before the enactment of GDPR. The length of the commitment phase $T-N$ is the length of the platform-wide deployment of the committed treatment before a new design rolls out. The commitment date is marked by the enactment of GDPR. Picking treatment I_t generates a treatment-associated random outcome o_{t,I_t} which consists of two attributes, $(\textsc{revenue}_{t,I_t}, \textsc{privacy-level}_{t,I_t})$. $\textsc{revenue}_{t,I_t}$ is a random variable, while $\textsc{privacy-level}_{t,I_t} = \textsc{privacy-level}_{I_t}$ is likely a constant for a given treatment I_t, and the larger the number is, the more privacy the customers enjoy. Before the enactment of GDPR, the reward function has the form $f(o_{t,I_t}) = \textsc{revenue}_{t,I_t}$. After the enactment of GDPR, the reward function has the form $g(o_{t,I_t}) = \textsc{revenue}_{t,I_t} + \textsc{privacy-level}_{I_t}$, which captures the legal penalty on privacy.*

Example 2.2 (The manufacturing supply chain examples (continued))**.** *One treatment $I_t \in [K]$ could represent one new production plan. The length of the experiment phase N is the length of the grace period before the enactment of carbon tariffs, export restrictions, or some other policy environment shifts. The length of the commitment phase $T-N$ is the market life of one generation of product. The commitment date is marked by the enactment of the carbon tariffs, export restrictions, full production capacity transition, and so on. The random outcome o_{t,I_t} consists of two attributes, $(\textsc{revenue}_{t,I_t}, \textsc{carbon-emission}_{t,I_t})$ or $(\textsc{revenue}_{t,I_t}, \textsc{risky-supplier}_{t,I_t})$. $\textsc{revenue}_{t,I_t}$ is a random variable, while $\textsc{carbon-emission}_{t,I_t} = \textsc{carbon-emission}_{I_t}$ recording carbon emission volume and $\textsc{risky-supplier}_{t,I_t} = \textsc{risky-supplier}_{I_t} \in \{0,1\}$ taking binary values with indicating whether the production involves risky suppliers, are constants for a given treatment I_t. Before the commitment, the reward function is again simply $f(o_{t,I_t}) = \textsc{revenue}_{t,I_t}$. After the enactment of the carbon tariffs, the reward function can have the form $g(o_{t,I_t}) = \textsc{revenue}_{t,I_t} - \textsc{carbon-emission}_{I_t}$. Or, after the enactment of export restrictions, the reward function can have the form $g(o_{t,I_t}) = (1 - \textsc{risky-supplier}_{I_t}) \cdot \textsc{revenue}_{t,I_t}$.*

More Notations. We define the following notations that will be useful for our analysis. For each treatment $i \in [K]$, we define its corresponding reward distributions for the *experiment phase* and the *commitment phase* as $V_{f,i}(x) = \mathbb{P}_{o\sim\nu_i}[f(o) = x]$ and $V_{g,i}(x) = \mathbb{P}_{o\sim\nu_i}[g(o) = x]$, respectively. We would like to note that although the outcome distributions $(\nu_i)_{i\in[K]}$ may be general, the induced reward distributions have bounded support between $[0,1]$ by the mapping of the reward functions f, g. And, because the distributions $(\nu_i)_{i\in[K]}$ are unknown, the induced distributions $(V_{f,i})_{i\in[K]}$ and $(V_{g,i})_{i\in[K]}$ are also unknown. We use $\mathcal{V}_{f,i}$, $\mathcal{V}_{g,i}$ to denote the space of the reward distributions for function f, g for the treatment i, respectively. That is, $V_{f,i} \in \mathcal{V}_{f,i}$ and $V_{g,i} \in \mathcal{V}_{g,i}$.

2.1 Additional Discussions

We conclude this section with some additional discussions.

The exogenously given experimentation budget N. In our framework, the experimentation budget N is exogenously given by the environment. This modeling choice is primarily motivated by the practical observations in which, for example, a regulatory policy may allow a duration of time before it takes effect, or the decision-maker may face an internal deadline limiting the total experimentation time. By contrast, a soft commitment model – where the decision maker can endogenously choose when to commit – has been studied in earlier work (see, e.g., Bui et al., 2011; Qin and Russo, 2024). Indeed, when the reward functions satisfy $f = g$, the standard explore-then-commit strategies (see, e.g., Robbins, 1952; Rusmevichientong and Tsitsiklis, 2010; Garivier et al., 2016) in the multi-armed bandit literature can capture this soft-commitment scenario as well.

Known reward functions f and g. In our paper, we assume both reward functions f and g are known. What is unknown is each arm's outcome distribution. An arm's outcome represents a random signal of the intrinsic nature of the arm itself, while the reward functions represent the environment's evaluation of the outcome. Recalling the motivations for our problem, the decision-maker faces an anticipated policy environment change. Such policy environment changes are typically announced much earlier than their enactment, and their impacts on evaluating a given outcome are measurable. This corresponds to the situation where f and g are known in advance. An alternative modeling is to consider a model where the realized rewards in the experiment phase and commitment phase are $f_i + \varepsilon$ and $g_i + \varepsilon$, respectively, where f_i, g_i are the arm's unknown true expected rewards. We note that this model is indeed a special case of ours if the learner can observe two noisy reward signals $f_i + \varepsilon$ and $g_i + \varepsilon$ simultaneously upon pulling arm i. However, if pulling an arm in experimentation only reveals $f_i + \varepsilon$, then the data collected in the experimentation phase tells us nothing about post-shift rewards, which would lead to linear regret unless extra structure (e.g., a parametric relation) is imposed on f_i, g_i.

Reducing to regret minimization and best-arm identification. We observe that $g(o) \equiv 0$ for all $o \in O$, our sequential experiment with the commitment problem reduces to regret minimization in the classic multi-armed bandit (MAB) problem with time horizon N (Auer et al., 2002; Lai and Robbins, 1985). Another variant of the MAB problem is the "best-arm identification problem" (often referred to as the pure exploration problem), where the goal is to identify the best treatment at the end of the experiment. In particular, when $f(o) \equiv 0$ for all $o \in O$, our sequential experiment with the commitment problem reduces to the fixed-budget pure exploration problem (Audibert et al., 2010; Bubeck et al., 2011; Kaufmann et al., 2016; Carpentier and Locatelli, 2016). Note that when $f(o) \equiv 0$, our regret definition (2) is essentially minimizing the *simple regret* $(T - N) \cdot \mathbb{E}\Big[g(o_{I_g^*}) - g(o_{J_N})\Big]$.

3 Our Algorithm and Results

In this section, we introduce the Reserved Arm Eliminations for Commitment (RAEC) algorithm for our sequential experiment and commitment problem and discuss its regret performance.

3.1 Our Algorithm

We begin with elaborating on the details of Algorithm RAEC. Our algorithm operates as an elimination-type algorithm (see, e.g., Even-Dar et al., 2006; Auer and Ortner, 2010) that acts in phases (epochs) and eliminates arms using increasingly sensitive hypothesis tests.

At a high level, there are two stages of the algorithm in the experiment phase, followed by a straightforward committing stage. In Stage I, the algorithm focuses on collecting arm information for the commitment-phase reward function g until the experimenter gathers sufficient information for the reward function g to make the commitment decision. The algorithm moves into Stage II, in which the algorithm uses the remaining rounds in the experiment phase to conduct the regret minimization over arms w.r.t. reward function f. Throughout the experiment phase, the algorithm will maintain two active arm sets $\mathcal{A}_{f,\ell}, \mathcal{A}_{g,\ell}$ for each epoch ℓ and for each reward function f, g, respectively. The active arm sets are both initialized as the whole arm space $[K]$.

Stage I: Reserved arm elimination for reward function g. In Stage I, which includes the first L epochs, the algorithm explores arms in action arm $\mathcal{A}_{g,\ell}$ and eliminates arms that have bad performances based on their empirical average rewards. Designing the number L is one key of our algorithm. The choice of this value represents the amount of effort the decision-maker should exert during the experiment phase to learn the optimal arm for the commitment phase. In particular, in each epoch $\ell \in [L]$, if there is more than one arm in the active arm set $\mathcal{A}_{g,\ell}$ remained, the algorithm will select each arm in the arm set $\mathcal{A}_{g,\ell}$ until the total number of times it has been selected achieves $m_{g,\ell}$. At the end of each epoch ℓ, the algorithm computes the empirical average rewards $\hat{\mu}_{g,j,\ell}$ for each arm $i \in \mathcal{A}_{g,\ell}$, then updates the arm set $\mathcal{A}_{g,\ell+1}$ by eliminating arms that deviate too much from the arm that has empirically best reward $\max_{j\in\mathcal{A}_{g,\ell}} \hat{\mu}_{g,j,\ell}$. Specifically, arms are eliminated from $\mathcal{A}_{g,\ell}$ if the gap between their empirical reward and the best empirical reward exceeds a threshold $\widetilde{\Delta}_\ell$, which is initialized to $1/2$. The threshold $\widetilde{\Delta}_\ell$ will then be halved for the next epoch. On the other hand, when there is only one arm remaining in the arm set $\mathcal{A}_{g,\ell}$, the algorithm would stop the arm elimination for the reward function g.

Stage II: Continuing arm elimination for reward function f. Stage II includes the remaining rounds in the experiment phase. When entering Stage II, no matter how many arms remain in the active arm set $\mathcal{A}_{g,L}$, the algorithm will not sample arms in $\mathcal{A}_{g,L}$. Instead, the algorithm will keep eliminating the arms $\mathcal{A}_{f,L}$ using a similar procedure in Stage I. Especially, the elimination criteria also remain the same as before.

We carefully design the total number of selections $m_{f,\ell}, m_{g,\ell}$ of the arms in each epoch as follows,

$$m_{f,\ell} = \frac{4\log(N)}{\widetilde{\Delta}_\ell^2}, \quad m_{g,\ell} = \frac{4\log(T-N)}{\widetilde{\Delta}_\ell^2} \ . \tag{3}$$

Here, notice that the number of pulls for arms in $\mathcal{A}_{f,\ell}$ and $\mathcal{A}_{g,\ell}$ differ in the numerator, where $m_{f,\ell}$ adjusts for the length of the experiment phase, N, while $m_{g,\ell}$ adjusts for the length of the commitment phase, $T-N$.

Commitment Stage. Note that at the beginning of the $(N+1)$-th round, at least one arm remains in the active arm set $\mathcal{A}_{g,N}$. The algorithm then uniformly at random commits to an arm in this set for the commitment phase.

We name our algorithm as *reserved arm eliminations for commitment* because we predetermine L epochs reserved for exploring the optimal arm in the commitment phase. Instead of choosing adaptively while learning the problem structure, a predetermined L is not only easy to implement in industrial environments but also explicitly indicates how much effort the decision-maker should exert for long-term benefits.

We highlight that during the algorithm execution Stage II, we sample each arm in $\mathcal{A}_{f,\ell}$ until it has been pulled $m_{f,\ell}$ times, but the $m_{f,\ell}$ number of pulls may not all occur during Stage II. This is because the arm is totally possible to have been pulled many times during Stage I. Because we observe the outcome realization o_{t,I_t} upon pulling each arm I_t and function forms f and g are known,

Algorithm 1 Reserved Arm Eliminations for Commitment (RAEC)

Input: A set of arms $\{1, 2, \ldots, K\}$, N, T, parameter ε;
Initialization: Set $\widetilde{\Delta}_1 = 1/2$, $\mathcal{A}_{f,1} = \mathcal{A}_{g,1} = [K]$, $L = \lceil \log_2 \frac{1}{\varepsilon} \rceil$.
Whenever N rounds are exhausted in Stage I or II, the algorithm enters the Commitment Stage.
/* Below $m_{f,\ell}, m_{g,\ell}$ are defined as in (3). */
for $\ell = 1, \ldots, L$ **do** /* Stage I: Reserved arm elimination for reward function g */
  if $|\mathcal{A}_{g,\ell}| > 1$ **then**
    Sample each arm in $\mathcal{A}_{g,\ell}$ until the total number of times it has been chosen is $m_{g,\ell}$ times.
    At the end of epoch ℓ, compute the empirical average reward $\hat{\mu}_{g,i,\ell}$ for reward function g for each $i \in [K]$.
    Update $\mathcal{A}_{g,\ell+1} \leftarrow \left\{ i \in [K] : \max_{j \in \mathcal{A}_{g,\ell}} \hat{\mu}_{g,j,\ell} - \hat{\mu}_{g,i,\ell} \le \widetilde{\Delta}_\ell \right\}$.
  else
    $\mathcal{A}_{g,\ell+1} \leftarrow \mathcal{A}_{g,\ell}$.
  end if
  Set $\widetilde{\Delta}_{\ell+1} \leftarrow \widetilde{\Delta}_\ell / 2$.
end for
/* Note that below ℓ restarts from 1. */
for $\ell = 1, 2, \ldots$ **do** /* Stage II: Arm eliminations for reward function f */
  Sample each arm in $\mathcal{A}_{f,\ell}$ until the total number of times it has been chosen is $m_{f,\ell}$ times.
  At the end of epoch ℓ, compute the empirical average reward $\hat{\mu}_{f,i,\ell}$ for reward function f for each $i \in [K]$.
  Update $\mathcal{A}_{f,\ell+1} \leftarrow \left\{ i \in [K] : \max_{j \in \mathcal{A}_{f,\ell}} \hat{\mu}_{f,j,\ell} - \hat{\mu}_{f,i,\ell} \le \widetilde{\Delta}_\ell \right\}$.
  Set $\widetilde{\Delta}_{\ell+1} \leftarrow \widetilde{\Delta}_\ell / 2$.
end for
/* Denote by $\mathcal{A}_{g,N}$ the current active arm set for reward function g. */
Uniformly at random committing to an arm in $\mathcal{A}_{g,l}$. /* Commitment Stage */

we can calculate both $f(o_{t,I_t})$ and $g(o_{t,I_t})$. When executing the algorithm, we actually record the history of outcome realization for each arm pull, and that is used to calculate the empirical means $\hat{\mu}_{f,i,\ell}$ and $\hat{\mu}_{g,i,\ell}$.

3.2 The Regret Upper Bound of Algorithm RAEC

In this section, we provide the regret upper bound of Algorithm RAEC. The main result of this section is summarized as follows:

Theorem 3.1 (Regret upper bound)**.** *With the parameter*

$$\varepsilon = \max\left\{ \sqrt{\frac{K \log(T-N)}{N}}, \left(\frac{K \cdot \log(T-N)}{T-N} \right)^{1/3}, \sqrt{\frac{KT \log(T)}{(T-N)^2}} \right\},$$

we have the following regret upper bound for Algorithm RAEC:

$$\mathbf{REG}[N,T] \le \widetilde{O}\left(\sqrt{\frac{K(T-N)^{2/3} \cdot \max\left\{ (T-N)^{4/3}, K^{1/3} N \right\}}{\min\left\{ N, K^{1/3}(T-N)^{2/3} \right\}}} \right).$$

We offer more detailed discussions and explanations about the above regret upper bound. Our presented regret bound shows a phase transition of the attainable regret bounds from the Algorithm RAEC. In particular,

1. Short experiment scenario (Region I) – when $N \leq O\left(T^{2/3}(K\log(T))^{1/3}\right)$, by choosing parameter $\varepsilon = \sqrt{\frac{K\log(T)}{N}}$, the regret bound $\mathbf{REG}[N,T] = \widetilde{O}\left(\sqrt{\frac{KT^2}{N}}\right)$. This is the scenario when the experiment phase is rather short, such that all effort should be invested in exploring the optimal arm for the commitment phase.

2. Balanced scenario (Region II) – when $N \geq \Omega\left(T^{2/3}(K\log(T))^{1/3}\right)$ and $(T-N) \geq \Omega\left(K^{1/4}N^{3/4}/(\log(T-N))^{1/2}\right)$, by choosing parameter $\varepsilon = \left(\frac{K\log(T-N)}{T-N}\right)^{1/3}$, the regret bound $\mathbf{REG}[N,T] = \widetilde{O}\left(K^{1/3}(T-N)^{2/3}\right)$. This is the most interesting scenario where balancing the experiment phase and the commitment phase becomes a challenge.

3. Short commitment scenario (Region III) – when $(T-N) < O\left(\frac{K^{1/4}T^{3/4}}{(\log(T-N))^{1/2}}\right)$, by choosing parameter $\varepsilon = \sqrt{\frac{KT\log(T)}{(T-N)^2}}$, the regret bound $\mathbf{REG}[N,T] = \widetilde{O}\left(\sqrt{KT}\right)$. This is the scenario when the commitment phase is rather short that the algorithm starts tuning down the accuracy ε (i.e., larger ε) at a faster speed than in the balanced scenario as N grows. But notice that it is only until $(T-N) < \widetilde{O}\left(\sqrt{KT}\right)$ (i.e., $\varepsilon \geq 1$) does the algorithm totally give up exploring good arms for the commitment phase and minimize regret only for the experiment phase.

As a sanity check, consider the case $g \equiv 0$. In this scenario, our problem reduces to the classic K-armed bandit setting. Without loss of generality, we may set $T = N+1$ (to avoid potential indefinite calculations caused by $T = N$), in which case ε could be considered as 1 – implying there is no exploration during the commitment phase. Consequently, our regret upper bound reduces to $\widetilde{O}(\sqrt{KN})$ which matches the standard result (see, e.g., Auer et al., 2002). On the other hand, when $f = 0$, the problem becomes a pure exploration setting focused on minimizing simple regret. In this case, the total regret is just the simple regret at the commitment period $N+1$, multiplied by the commitment phase length $T-N$. To fully recover known results of pure exploration, we let $N = o(T)$, i.e., the experiment phase is entirely devoted to exploration before the commitment phase. Under this regime, $\varepsilon = \sqrt{K\log(T)/T}$ and our regret is bounded by $\widetilde{O}(\sqrt{KT^2/N})$. Dividing the regret by the length of the commitment phase, we get simple regret of $\widetilde{O}(\sqrt{K/N})$, which again aligns with standard bounds (see, e.g., Chapter 33 in Lattimore and Szepesvári, 2020). Finally, we observe that over a wide range of regimes (i.e., in Regions I and II), our regret upper bound is significantly larger than the typical $\widetilde{O}(\sqrt{KT})$ found in standard K-armed bandit problems. However, as becomes clearer in Section 6.1, when there is no reward shift (i.e., $f = g$), the regret can be substantially improved, which mirrors the effect of reward shift ($f \neq g$). We will discuss more on this in Section 5.

We summarize the phase transition of the regret upper bound with respect to the different parameter regimes of N in Figure 2. We highlight that in our algorithm RAEC, designing ε is equivalent to determining the effort allocation between exploring the optimal arms in the experiment phase and the commitment phase. Our algorithm RAEC determines ε in advance instead of adaptively via learning the instance structure. This simple yet effective design makes RAEC easy to implement in practice while also providing transparency in how much effort is devoted to exploring good arms for the commitment phase. However, a key question is whether our simple

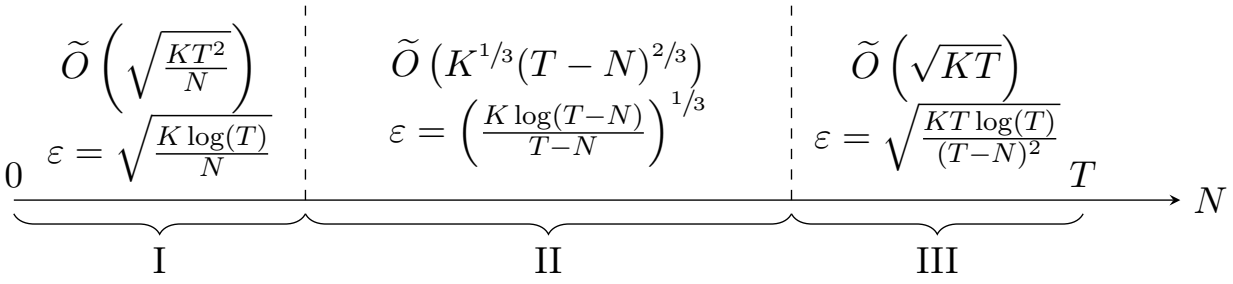

Figure 2: Phase Transition of The Regret Upper Bound of Algorithm RAEC

design can achieve the best possible performance for any online learning algorithm. We answer this affirmatively by establishing a matching lower bound (see Theorem 5.2).

4 The Regret Upper Bounds Analysis

In this section, we provide the analysis for the regret upper bound of Algorithm RAEC, i.e., we prove Theorem 3.1. Instead of directly proving the instance-independent upper bound presented in Theorem 3.1, we first prove the following stronger result (see Proposition 4.1) – an instance-dependent regret bound for Algorithm RAEC. We then discuss how to obtain the upper bound in Theorem 3.1 from Proposition 4.1.

Before discussing Algorithm RAEC's instance-dependent regret upper bound, we first define a problem instance. We define the instance space as $\mathcal{E} = \mathcal{V}_{f,1} \times \mathcal{V}_{f,2} \times \cdots \times \mathcal{V}_{f,K} \times \mathcal{V}_{g,1} \times \mathcal{V}_{g,2} \times \cdots \times \mathcal{V}_{g,K}$. An instance $\mathcal{I} \in \mathcal{E}$ is then defined as $\mathcal{I} = (V_{f,1}, \ldots, V_{f,K}, V_{g,1}, \ldots, V_{g,K})$. Let μ be the function that maps a distribution P to its mean. Define $\mu_{f,i} = \mu(V_{f,i})$ and $\mu_{g,i} = \mu(V_{g,i})$. Let $\mu_f^* = \max_{i\in[K]} \mu_{f,i}$ and $\mu_g^* = \max_{i\in[K]} \mu_{g,i}$ be the mean of the optimal arm w.r.t. reward function f, g respectively. Then, let $\Delta_{f,i} = \mu_f^* - \mu_{f,i}$ and $\Delta_{g,i} = \mu_g^* - \mu_{g,i}$ be the suboptimality gap. With these definitions, we are ready to present the instance-dependent regret upper bound of Algorithm RAEC.

Proposition 4.1. *Given an instance $\mathcal{I} \in \mathcal{E}$, the expected regret of Algorithm RAEC with parameter ε can be upper bounded as follows:*

$$\mathbf{REG}[N, T \mid \mathcal{I}] = O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\max\{\Delta_{g,i}, \varepsilon\}^2}, \frac{\log(N)}{\Delta_{f,i}^2}\right\} + (T-N) \cdot \max_{i:\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}\right). \tag{4}$$

The first term in regret bound (4) is the upper bound of the instance-dependent regret incurred in the experiment phase, and the second term is the upper bound of the regret in the commitment phase.

Proof Sketch of Proposition 4.1. We now describe a proof sketch of Proposition 4.1, which consists of three main steps. **Step 1** – We first upper bound the expected regret, denoted by $\mathbf{REG}_1[N]$, of the arm elimination process for the reward function f (namely, eliminating arms from the active arm sets $\mathcal{A}_{f,\ell}$) from epoch 1 to $\widetilde{L}$. See below Lemma 4.2.

Lemma 4.2. $\mathbf{REG}_1[N] \le O\left(\sum_{i\in[K]} \frac{\log(N)}{\Delta_{f,i}}\right)$.

Step 2 – We next upper bound the expected regret, $\mathbf{REG}_2[N]$, of arm elimination process for the reward function g (namely, eliminating arms from the active arm sets $\mathcal{A}_{g,\ell}$). We note that

to bound the regret incurred by pulling arms in the active arm set $\mathcal{A}_{g,\ell}$, it essentially reduces to upper-bounding the expected total number of rounds used to pull the arms in the set $\mathcal{A}_{g,\ell}$. See below Lemma 4.3.

Lemma 4.3. *Given target commitment precision ε,*

$$\mathbf{REG}_2[N] \leq O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\max\{\Delta_{g,i},\varepsilon\}^2}, \frac{\log(N)}{\Delta_{f,i}^2}\right\}\right).$$

Step 3 – Finally, we upper bound the expected regret, denoted by $\mathbf{REG}_3[N,T]$, incurred in the commitment phase. See Lemma 4.4.

Lemma 4.4. *Given target commitment precision ε, $\mathbf{REG}_3[N,T] \leq O\left((T-N)\max_{i:\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}\right)$.*

Notice that in the design of the Algorithm RAEC, we have the implicit constraint that we can complete exploring arms in $\mathcal{A}_{g,L}$. Therefore, when we design ε, we set ε such that $K \cdot m_{g,L} \leq N$, or equivalently, $\varepsilon \geq \sqrt{(4\beta \cdot K \cdot \log(T-N))/N}$. Without changing the asymptotic rate of the results, we simply write the implicit constraint as $\varepsilon \geq \sqrt{(K \cdot \log(T-N))/N}$.

As specified in Theorem 3.1, we choose $\varepsilon = \max\left\{\sqrt{\frac{K\log(T-N)}{N}}, \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}, \sqrt{\frac{KT\log(T)}{(T-N)^2}}\right\}$. As both N and T grow, since $\log(T-N)/N$ diminishes (due to $N \geq \Omega(\mathsf{poly}(T))$), the instance-dependent upper bound could be further written as

$$\mathbf{REG}[N,T \mid \mathcal{I}] \leq O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\Delta_{g,i}^2}, \frac{\log(N)}{\Delta_{f,i}^2}\right\}\right). \tag{5}$$

□

5 The Regret Lower Bound

In this section, we complement the regret upper bound of Algorithm RAEC with a matching lower bound. We present an instance-dependent lower bound, followed by an instance-independent lower bound (i.e., minimax lower bound).

We first introduce some definitions that will be useful for our analysis. Let $\mathcal{F}$ be a space of distributions with finite means. For any $\mu \in \mathbb{R}$ and a distribution $P \in \mathcal{F}$ such that its mean satisfies $\mu(P) < \mu$, we define the following

$$d_{\inf}(P,\mu,\mathcal{F}) = \inf_{P'\in\mathcal{F}}\{D(P,P') : \mu(P') > \mu\} \ ,$$

where $D(\cdot,\cdot)$ is the Kullback-Leibler divergence. Given an instance $\mathcal{I} \in \mathcal{E}$, we define

$$d_{f,i} = d_{\inf}(V_{f,i}, \mu_f^*, \mathcal{V}_{f,i}); \quad d_{g,i} = d_{\inf}(V_{g,i}, \mu_g^*, \mathcal{V}_{g,i}) \ .$$

For common distributions like Gaussian, $d_{f,i}$ ($d_{g,i}$) is of the same order as $\Delta_{f,i}$ ($\Delta_{g,i}$). Then, to derive a nontrivial instance-dependent regret lower bound, we need to specify the family of policies that we investigate, conventionally referred to as *consistent policies*, which rules out policies like guessing policies that may perform extremely well in certain instances. Following the spirit of *consistent policies* defined in the standard K-armed bandit setting, we extend the definition to our problem with modifications.

Definition 5.1 (Consistent policy)**.** *A policy π is called consistent over a class of bandits $\mathcal{E}$ if for all $\mathcal{I} \in \mathcal{E}$, arm permutation σ, and $p > 0$, it holds that*

$$\lim_{T-N\to\infty} \frac{\mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}]}{N^p} = 0 \ , \ \lim_{T-N\to\infty} \frac{\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}]}{(T-N)^p} = 0 \ ;$$

and

$$\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] = \mathbf{REG}_{\mathsf{com}}[T-N \mid \sigma(\mathcal{I})] \ ,$$

where $\mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}]$ and $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}]$ represent the expected accumulative regret of the experiment phase and the commitment phase, respectively.

Recall that we have assumed $N \geq \Omega(\mathsf{poly}(T))$, so the regime where $T-N$ goes to infinity implies both N and T grow to infinity. The last condition $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] = \mathbf{REG}_{\mathsf{com}}[T-N \mid \sigma(\mathcal{I})]$ says that the algorithm should result in the same expected regret for the commitment phase if we simply permutate the arms' identities, i.e., the policy is symmetric under arm permutation.

Theorem 5.1 (Instance-dependent lower bound)**.** *Given instance $\mathcal{I} \in \mathcal{E}$, there exists an instance-dependent regret lower bound for all consistent policies as follows,*

$$\mathbf{REG}[N, T \mid \mathcal{I}] = \Omega\left(\sum\nolimits_{i\in[K]} \max\left\{\frac{\log(N)}{d_{f,i}}, \frac{\log(T-N)}{d_{g,i}}\right\} \cdot \Delta_{f,i}\right).$$

Notice that for common distributions, including Gaussian-type distributions, we have that $d_{f,i} \approx \Delta_{f,i}$ and $d_{g,i} \approx \Delta_{g,i}$. Recalling the asymptotic instance-dependent regret upper bound of Algorithm RAEC, (5), we find that the above lower bound approximately matches the upper bound.

Furthermore, using different instance constructions and analysis, we derive the following instance-independent regret lower bound (i.e., minimax lower bound).

Theorem 5.2 (Instance-independent lower bound)**.** *There exists an instance-independent regret lower bound for all policies as follows,*

$$\mathbf{REG}[N, T] = \Omega\left(\sqrt{\frac{K(T-N)^{2/3} \cdot \max\left\{(T-N)^{4/3}, K^{1/3}N\right\}}{\min\left\{N, K^{1/3}(T-N)^{2/3}\right\}}}\right).$$

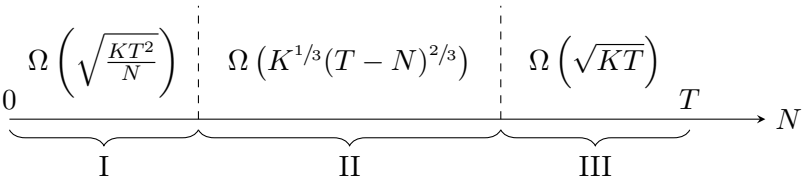

Figure 3: Phase Transition of Minimax Lower Bound

Figure 3 depicts the phase transition of the instance-independent regret lower bound with respect to the scale of N. Regions I, II, and III are defined as follows: (i). Region I: $N < K^{1/3}(T-N)^{2/3}$. (ii). Region II: $N \geq K^{1/3}(T-N)^{2/3}$ and $(T-N) \geq K^{1/4}N^{3/4}$. (iii). Region III: $(T-N) < K^{1/4}T^{3/4}$.

We observe that the instance-independent lower bound also matches the upper bound of RAEC presented in Theorem 3.1. This is especially noteworthy because our Algorithm RAEC allocates exploration effort for the commitment phase in a simple, predetermined way rather than adaptively.

This leads to a useful managerial implication for practitioners: Reserving effort in advance to explore good arms for the long-term commitment is both easy to implement and sufficient. More sophisticated and adaptive methods of allocating effort with the help of learning the instance structure will not help much. Meanwhile, we notice that in a wide range of regimes (Region I and II), our tight regret bound is significantly larger than a typical $\Omega\left(\sqrt{KT}\right)$ result. This implies that the presence of the commitment phase and the reward shift makes the problem fundamentally more challenging than standard bandit problems. As clarified in Section 6.1, such difficulty mainly comes from the reward shift.

We provide a proof sketch below for Theorem 5.2. At a high level, to prove the minimax lower bound, given any policy, we construct two instances such that this policy will suffer a large regret in at least one of these two instances (see, e.g., Slivkins et al., 2019; Lattimore and Szepesvári, 2020). We first construct the following instance $\mathcal{I}$: for some positive constants $\alpha, \delta_f, \delta_g$ (their values are specified in the proof),

$$\mu_{f,i} = \begin{cases} \alpha + \delta_f, & \text{if } i = 1; \\ \alpha, & \text{if } i \in [2, K/2]; \\ 0, & \text{if } i \in [K/2 + 1, K]; \end{cases} \qquad \mu_{g,i} = \begin{cases} \delta_g, & \text{if } i = K/2 + 1; \\ 0, & \text{otherwise} \end{cases}$$

We group the first $K/2$ arms as arm group A and the other half of arms as arm group B. Given a policy π, let $\mathbb{E}_{\mathcal{I}}[T_A]$ and $\mathbb{E}_{\mathcal{I}}[T_B]$ be the expected number of pulls distributed to group A and B, respectively. Clearly, $\mathbb{E}_{\mathcal{I}}[T_A] + \mathbb{E}_{\mathcal{I}}[T_B] = N$. Let $\mathbb{E}_{\mathcal{I}}[T_l]$ be policy π's expected number of pulls of arm l. Then, we define $i^\dagger \triangleq \arg\max_{l \in A \setminus \{1\}} \mathbb{E}_{\mathcal{I}}[T_l]$ and $j^\dagger = \arg\max_{l \in B \setminus \{K/2+1\}} \mathbb{E}_{\mathcal{I}}[T_l]$. Clearly, $\mathbb{E}_{\mathcal{I}}[T_{i^\dagger}] \leq \frac{\mathbb{E}_{\mathcal{I}}[T_A]}{K/2-1}$ and $\mathbb{E}_{\mathcal{I}}[T_{j^\dagger}] \leq \frac{\mathbb{E}_{\mathcal{I}}[T_B]}{K/2-1}$. With these definitions, we define the instance $\mathcal{I}^\dagger$ as follows:

$$\mu^\dagger_{f,i} = \begin{cases} \alpha + \delta_f, & \text{if } i = 1; \\ \alpha + 2\delta_f, & \text{if } i = i^\dagger; \\ \alpha, & \text{if } i \in [2, K/2] \setminus \{i^\dagger\}; \\ 0, & \text{if } i \in [K/2 + 1, K]; \end{cases} \qquad \mu^\dagger_{g,i} = \begin{cases} \delta_g, & \text{if } i = K/2 + 1; \\ 2\delta_g, & \text{if } i = j^\dagger; \\ 0, & \text{otherwise} \end{cases}$$

We show that the policy π would suffer a large regret either in instance $\mathcal{I}$ or in instance $\mathcal{I}^\dagger$.

6 Extensions

In this section, we discuss two extensions of our basic model: (1) In the first extension, we consider the situation when the decision-maker has access to more nuanced prior knowledge about the reward shift structure (i.e., the functional relationship between reward functions f and g). One main insight from the previous basic model is that with the information from learning the instance structure, adaptively determining the effort allocation between exploring the optimal arms in the experiment phase and the commitment phase will not help much. However, we argue that prior knowledge of the reward shift structure can sometimes help. The key is to identify the reward shift's ranking-changing effect instead of focusing on its absolute value. (2) In the second extension, we consider a generalized version of the basic model in which the decision-maker is allowed to commit to a portfolio of arms. The reward in the commitment phase is a Lipschitz concave function of the portfolio's combined outcome. We argue that the effectiveness of the idea of predetermined effort allocation persists in this general setting.

6.1 Improved Regret with Prior Knowledge on Reward Shift

In the basic model, we do not pose any assumptions on the relation between reward functions $f(\cdot)$ and $g(\cdot)$ except that both are bounded within $[0, 1]$. Our algorithm RAEC predetermines the accuracy level ε that we aim to achieve for the commitment phase. And, as we argued, adaptively determining the target accuracy level ε via learning the problem structure won't help improve the asymptotic regret performance. However, as we demonstrate below, prior knowledge of the problem structure helps.

In this extension, we consider a more nuanced situation where the decision-maker has some prior knowledge about the relation between the reward functions $f(\cdot)$ and $g(\cdot)$.

Definition 6.1 (Perturbed functional relationship)**.** *Reward functions f and g exhibit a perturbed functional relationship if there exists a constant $M \geq 0$ and a noise function $\delta(\cdot) : O \to [-D/2, D/2]$ for some $D \geq 0$:*

$$g(o) = M \cdot f(o) + \delta(o), \quad o \in O \ . \tag{6}$$

Intuitively, $M \cdot f(\cdot)$ is a ranking-preserving mapping, i.e., the arms' mean reward ranking remains the same as under $f(\cdot)$ after times M. On the other hand, δ could be ranking-changing. Relating to the application environments, if functions $f(\cdot)$ and $g(\cdot)$ indicate the profits of products, M could represent the impact of a common tariff, δ could represent product-specific policy restrictions like export controls on certain products or components.

Both M and δ can significantly shift the value of a treatment, but we argue that the size of D is significantly more important than the size of M. For details, please refer to Section C. The managerial takeaway is that specifying the ranking-changing effect is much more important than focusing on the absolute value of the reward shifts.

6.2 Generalized Concave Commitment Reward

Recall our motivation setting where a firm whose short-term payoff is the accumulated profit, while due to policy environment changes, the long-term goal is to find a business solution that balances the profitability and other factors like average quality, aggregated environmental impact across all product lines, average supply chain resilience over time, etc. In this extension, the firm is allowed to commit to a portfolio of candidate solutions for the commitment phase. A portfolio of candidate solutions means an action plan for deploying each candidate solution in the commitment phase. This means the action for each period in the commitment phase could be different, but is determined at the beginning of the commitment phase. We consider the case where the portfolio's commitment phase payoff is Lipschitz concave, defined in the outcome space. For example, the long-run performance is linear in profit minus the convex increasing penalty of environmental impact (i.e., the marginal penalty increases as the environmental impact grows). Formally, we formulate the decision-maker's problem as follows,

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{N} f(o_{t,I_t}) + (T-N) \cdot g\left(\frac{\sum_{t=N+1}^{T} o_{t,I_t}}{T-N}\right)\right], \tag{7}$$

where I_{N+1} to I_T are committed at the beginning of the commitment phase. We require O to be a compact space in this extension. For simplicity, we assume $O = [0, 1]^d$. And, for any $o_1,\ o_2 \in O$, $|g(o_1) - g(o_2)| \leq L \cdot \|o_1 - o_2\|$.

First, we argue that our basic problem (1) is a special case of problem (7). To see this, let the input spaces of f and g be scalar spaces, then with a slight abuse of notations, substitute o_{t,I_t} by

$h(o_{t,I_t})$ where $h(\cdot): O \mapsto \mathbb{R}$. Then, (7) can be written as

$$\max_{\pi} \mathbb{E}\left[\sum\nolimits_{t=1}^{N} f(h(o_{t,I_t})) + (T-N)\cdot g\left(\frac{\sum_{t=N+1}^{T} h(o_{t,I_t})}{T-N}\right)\right] .$$

If we regard the function composition $f \circ h$ as function f in (1), function h as function g in (1), and let function g in the above as linear function $g(x) = x$, we return to the formulation of (1) except that in the basic problem (1), the decision-maker can only commit to one arm for the commitment phase. However, notice that if we allow the decision-maker to commit to a portfolio of arms in (1), the optimal decision of the benchmark is still committing to the single best arm in the commitment phase. What might make the optimal decision to commit to a nontrivial portfolio is the nonlinearity of g in (7). Notice that we have demonstrated the formulation (7) is an extension of the basic model formulation, but the analysis later relies on the mild assumption that O is compact.

According to Agrawal and Devanur (2014), the total reward of the commitment phase can be upper bounded by $\mathsf{OPT}_g = \max_{\boldsymbol{p}\in\Delta([K])} g\left(\sum_{i=1}^{K} p_i \cdot \mathbb{E}_{o\sim\nu_i}[o]\right)$ due to concavity. Therefore, we consider the following definition of regret,

$$\mathbf{REG}_{\pi}[N,T] = \sum\nolimits_{t=1}^{N}\left(\mathbb{E}\left[f(o_{t,I_f^*})\right] - \mathbb{E}[f(o_{t,I_t})]\right) + (T-N)\cdot\left(\mathsf{OPT}_g - \mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T} o_{t,I_t}}{T-N}\right)\right]\right). \tag{8}$$

Following the spirit of RAEC algorithm, we propose the *reserved online stochastic convex optimization for commitment* algorithm (ROSCOC), where instead of applying the arm elimination technique during the reserved periods, we apply the well-established online stochastic convex optimization algorithm. Unlike the arm elimination method, which identifies the optimal arm for the commitment phase, online stochastic convex optimization only generates an execution path in the experiment phase. The challenge is how to relate this execution path to the portfolio of arms to commit. The novel idea is to use the execution path to approximate the ideal portfolio of arms. Following this idea, we generate the action plan for the commitment phase by uniformly sampling from the execution path. And we show that this actually works. For details, please refer to the pseudo-code of the Algorithm ROSCOC detailed in Algorithm 2 in the appendix (Algorithm ROSCOC uses a subroutine – the *online stochastic convex optimization* algorithm adopted from Agrawal and Devanur (2014), we include this subroutine in Algorithm 3). The main result is summarized as follows:

Theorem 6.1. *The regret of Algorithm ROSCOC has the form*

$$\mathbf{REG}_{\pi}[N,T] = \widetilde{O}\left(T\cdot\left(\sqrt{\frac{Kd}{N}} + L\cdot\sqrt{\frac{d}{N}}\right) + \sqrt{KN} + K^{1/3}d^{1/3}(T-N)^{2/3}\right), \tag{9}$$

where ROSCOC reserves $\tau = \min\left\{N, K^{1/3}d^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right\}$ *periods to exploration for commitment.*

When L and d are constants, the regret bound stated in Theorem 6.1 essentially reduces to the regret bound derived in Theorem 3.1: $\widetilde{O}\left(\sqrt{\frac{K(T-N)^{2/3}\cdot\max\left\{(T-N)^{4/3}, K^{1/3}N\right\}}{\min\left\{N, K^{1/3}(T-N)^{2/3}\right\}}}\right)$. As we have demonstrated, our basic model is a special case of the model in this section. Therefore, the minimax regret lower bound derived in Theorem 5.2 still holds, i.e., the above upper bound is basically tight.

In summary, although the general model appears to be much more challenging, we find that the effectiveness of predetermined effort allocation of exploration for the commitment phase still applies in the more general setting. The execution path of the online stochastic convex optimization algorithm is a good enough proxy for the optimal portfolio in the commitment phase. With the help of these algorithmic design techniques, we show that the general problem is not fundamentally harder than the basic model.

7 Conclusion

Motivated by the increasingly prominent operational challenge of balancing short-term performance with long-term objectives—where actions that maximize immediate rewards may conflict with long-run optimal outcomes due to shifts in the operational environment, such as state-level environmental policies and geopolitical disruptions in critical supply chains—we study an adaptive experimentation problem with post-commitment reward shifts. We first introduce a simple yet effective online learning algorithm – RAEC – which achieves provable performance across all parameter regimes of experimentation budget N. We establish an information-theoretic lower bound that matches the upper bound achieved by RAEC, which proves that the regret bound cannot be further improved by other algorithms. Finally, we advance our discussions to two extensions. (1) With prior knowledge of the reward shift structure, one can improve regret performance. Our results clarify that in terms of regret minimization, correctly identifying the ranking-changing effect in reward shifts is far more critical than specifying the absolute magnitude of these shifts. (2) Instead of committing to a single arm, the decision-maker can commit to a portfolio of arms in the commitment phase, and the total reward in the commitment phase is a concave function of the sum of the observed outcomes. We propose a new algorithm following the spirit of RAEC that allocates a predetermined portion of the experiment phase to learning the optimal commitment decision. Our analysis shows that the proposed new algorithm achieves a similar tight regret bound as RAEC across the full spectrum of parameter regimes.

Interesting future research questions include algorithm design when the commitment time is exogenously given but remains unknown (i.e., N is unknown) to the decision-maker or when reward functions f and g admit more fine-grained structural relationships, etc.

References

Google Ads. Changes to our ad policies to comply with the gdpr, 2018. URL https://blog.google/products/ads/changes-to-our-ad-policies-to-comply-with-the-gdpr/.

Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1405–1424. SIAM, 2014.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.

Erjie Ang, Dan A Iancu, and Robert Swinney. Disruption risk and optimal sourcing in multitier supply networks. *Management Science*, 63(8):2397–2419, 2017.

Guy Aridor, Yeon-Koo Che, and Tobias Salz. The effect of privacy regulation on the data industry: empirical evidence from gdpr. *RAND Journal of Economics (Wiley-Blackwell)*, 54(4), 2023.

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

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van Der Laan. Post-contextual-bandit inference. *Advances in neural information processing systems*, 34:28548–28559, 2021.

Philip Blenkinsop. Eu gives automakers 'breathing space' on co2 emission targets. *Reuters*, March 2025. URL https://www.reuters.com/business/autos-transportation/eu-propose-giving-automakers-three-years-meet-co2-emission-targets-2025-03-03/.

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

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

Loc Bui, Ramesh Johari, and Shie Mannor. Committing bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

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.

Jonathan Colmer, Ralf Martin, Mirabelle Muûls, and Ulrich J Wagner. Does pricing carbon mitigate climate change? firm-level evidence from the european union emissions trading system. *Review of Economic Studies*, 92(3):1625–1660, 2025.

David Cutbill. Gdpr, ccpa reshape customer privacy strategies. *The Wall Street Journal: Risk & Compliance Journal*, 2019. URL https://deloitte.wsj.com/riskandcompliance/gdpr-ccpa-reshape-customer-privacy-strategies-01561683726.

Beatriz Pessoa de Araujo and Adam Robbins. The modern dilemma: Balancing short- and long-term business pressures. *https://corpgov.law.harvard.edu/2019/06/20/the-modern-dilemma-balancing-short-and-long-term-business-pressures/*, 2019.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019.

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.

Vivek Farias, Ciamac Moallemi, Tianyi Peng, and Andrew Zheng. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.

Yuchen Hu, Henry Zhu, Emma Brunskil, and Stefan Wager. Minimax-regret sample selection in randomized experiments. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 1209–1235, 2024.

IAB Europe. Transparency & consent framework (tcf). URL https://iabeurope.eu/transparency-consent-framework/. Notes the launch of TCF v1.1 on 25 April 2018.

Garrett A Johnson, Scott K Shriver, and Samuel G Goldberg. Privacy and market concentration: Intended and unintended consequences of the gdpr. *Management Science*, 69(10):5695–5721, 2023.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251. PMLR, 2013.

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.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Learning the pareto front using bootstrapped observation samples. *arXiv preprint arXiv:2306.00096*, 2023.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Christian Peukert, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. Regulatory spillovers and data governance: Evidence from the gdpr. *Marketing Science*, 41(4):746–768, 2022.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Hart E Posen and Daniel A Levinthal. Chasing a moving target: Exploitation and exploration in dynamic environments. *Management science*, 58(3):587–601, 2012.

Chao Qin and Daniel Russo. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Joseph S Shapiro and Reed Walker. Why is pollution from us manufacturing declining? the roles of environmental regulation, productivity, and trade. *American economic review*, 108(12):3814–3854, 2018.

David Simchi-Levi and Chonghuan Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*, pages 3086–3097. PMLR, 2023.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

The Economist. China is quietly reducing its reliance on foreign chip technology. The Economist, February 2024. URL https://www.economist.com/business/2024/02/13/china-is-quietly-reducing-its-reliance-on-foreign-chip-technology. Accessed 2026-01-31.

Nitasha Tiku. Europe's new privacy law will change the web, and more, March 2018. URL https://www.wired.com/story/europes-new-privacy-law-will-change-the-web-and-more/.

Brian Tomlin. Impact of supply learning when suppliers are unreliable. *Manufacturing & Service Operations Management*, 11(2):192–209, 2009.

Junwen Yang, Vincent YF Tan, and Tianyuan Jin. Best arm identification with minimal regret. *arXiv preprint arXiv:2409.18909*, 2024.

Qining Zhang and Lei Ying. Fast and regret optimal best arm identification: fundamental limits and low-complexity algorithms. *Advances in Neural Information Processing Systems*, 36, 2023.

Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *Transactions on Machine Learning Research*, 2023.

Online Supplement

A Missing Proofs in Section 4

Proof for Lemma 4.2. Let $\widetilde{L} = \min\{\ell : m_{f,\ell} \geq N\}$ be the smallest epoch index that there are already N rounds exhausted. We recall that $I_f^* \in [K]$ and $I_g^* \in [K]$ denote the true optimal arm of the *experiment phase* and the *commitment phase*, respectively.

We decompose the total expected regret into the summation of regret incurred by each suboptimal arm $i \neq I_f^*$. To facilitate the analysis, we define the following events :

- $\text{Event}_1\,(i, \ell+1)$: The optimal arm I_f^* remains uneliminated till epoch ℓ, and arm i is eliminated after epoch ℓ, i.e., $\left\{i \notin \mathcal{A}_{f,\ell+1}, i \in \mathcal{A}_{f,\ell}, I_f^* \in \mathcal{A}_{f,\ell}\right\}$.
- $\text{Event}_1\left(i, \ell'+1, I_f^*, \ell+1\right)$: The optimal arm I_f^* eliminated after epoch ℓ, and arm i is eliminated after epoch ℓ', i.e., $\left\{i \notin \mathcal{A}_{f,\ell'+1}, i \in \mathcal{A}_{f,\ell'}, I_f^* \notin \mathcal{A}_{f,\ell+1}, I_f^* \in \mathcal{A}_{f,\ell}\right\}$.

Then,

$$\begin{aligned}
&\mathbf{REG}_1[N] \\
\leq\ & \sum\nolimits_{i\in[K]} \left(\Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{\widetilde{L}} m_{f,\ell} \cdot \mathbb{P}\left(\text{Event}_1\,(i,\ell+1)\right) + \right. \\
& \left. \Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{\widetilde{L}} \sum\nolimits_{\ell'=\ell+1}^{\widetilde{L}} m_{f,\ell'} \cdot \mathbb{P}\left(\text{Event}_1\left(i,\ell'+1,I_f^*,\ell+1\right)\right) \right) \\
\overset{(a)}{\leq}\ & \sum\nolimits_{i\in[K]} \left(\Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{\widetilde{L}} m_{f,\ell} \cdot \mathbb{P}\left(\text{Event}_1\,(i,\ell+1)\right) + \Delta_{f,i} \cdot m_{f,\widetilde{L}} \cdot \mathbb{P}\left(\exists\ \ell \in [\widetilde{L}] : I_f^* \notin \mathcal{A}_{f,\ell}\right) \right),
\end{aligned} \tag{10}$$

where inequality (a) is due to the facts that $m_{f,\ell}$ is increasing in ℓ and

$$\sum\nolimits_{\ell=1}^{\widetilde{L}} \sum\nolimits_{\ell'=\ell+1}^{\widetilde{L}} \mathbb{P}\left(\text{Event}_1\left(i,\ell'+1,I_f^*,\ell+1\right)\right) \leq \mathbb{P}\left(\exists\ \ell \in [\widetilde{L}] : I_f^* \notin \mathcal{A}_{f,\ell}\right).$$

Then, we bound the terms in (10), separately. In the first summation, since $\text{Event}_1\,(i,\ell)$ are mutually exclusive events across ℓ, we have

$$\sum\nolimits_{\ell=1}^{\widetilde{L}} m_{f,\ell} \cdot \mathbb{P}\left(\text{Event}_1\,(i,\ell+1)\right) \leq \Delta_{f,i} \cdot m_{f,L_i^f} + \Delta_{f,i} \cdot \sum\nolimits_{\ell=L_i^f+1}^{\widetilde{L}} m_{f,\ell} \cdot \mathbb{P}\left(\text{Event}_1\,(i,\ell+1)\right).$$

here $L_i^f = \min\left\{\ell : \widetilde{\Delta}_\ell \leq \Delta_{f,i}/2\right\}$. Since $\text{Event}_1\,(i,\ell+1)$ implies that arm i is not eliminated after epoch $\ell-1$ given the presence of I_f^*, due to Azuma's inequality, for $\ell \geq L_i^f + 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\text{Event}_1\,(i,\ell+1)\right) &\leq \mathbb{P}\left(\hat{\mu}_{f,I_f^*,\ell-1} - \hat{\mu}_{f,i,\ell-1} \leq \widetilde{\Delta}_{\ell-1}\right) \\
&\leq \mathbb{P}\left(\left(\hat{\mu}_{f,i,\ell-1} - \hat{\mu}_{f,I_f^*,\ell-1}\right) + \Delta_{f,i} \geq \Delta_{f,i} - \widetilde{\Delta}_{\ell-1}\right) \\
&\leq \exp\left(-\frac{m_{f,\ell} \cdot (\Delta_{f,i}/2)^2}{2}\right).
\end{aligned}$$

On the other hand, in the second summation of (10), we have

$$\begin{aligned}
\mathbb{P}\left(\exists\, \ell \in [\widetilde{L}] : I_f^* \notin \mathcal{A}_{f,\ell}\right) &= \sum\nolimits_{\ell=1}^{\widetilde{L}} \mathbb{P}\left(I_f^* \notin \mathcal{A}_{f,\ell+1}, I_f^* \in \mathcal{A}_{f,\ell}\right) \\
&\le \sum\nolimits_{\ell=1}^{\widetilde{L}} \mathbb{P}\left(\bigcup_{i\in[K]} \left(\hat{\mu}_{f,I_f^*,\ell} + \widetilde{\Delta}_\ell < \hat{\mu}_{f,i,\ell}, I_f^* \in \mathcal{A}_{f,\ell}, i \in \mathcal{A}_{f,\ell}\right)\right) \\
&\le \sum\nolimits_{\ell=1}^{\widetilde{L}} \sum\nolimits_{i\in[K]} \mathbb{P}\left(\widetilde{\Delta}_\ell < \hat{\mu}_{f,i,\ell} - \hat{\mu}_{f,I_f^*,\ell}\right) \\
&\le \sum\nolimits_{\ell=1}^{\widetilde{L}} \sum\nolimits_{i\in[K]} \mathbb{P}\left(\widetilde{\Delta}_\ell < \left(\hat{\mu}_{f,i,\ell} - \hat{\mu}_{f,I_f^*,\ell}\right) + \Delta_{f,i}\right) \\
&\le K \cdot \sum\nolimits_{\ell=1}^{\widetilde{L}} \exp\left(-\frac{m_{f,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right),
\end{aligned}$$

Recall that $m_{f,\ell} = \frac{\alpha\cdot\log(N)}{(\widetilde{\Delta}_\ell)^2}$. Then,

$$\begin{aligned}
(10) &\le \sum\nolimits_{i\in[K]} \left(\Delta_{f,i}\cdot m_{f,L_i^f} + \Delta_{f,i}\cdot \sum\nolimits_{\ell=L_i^f+1}^{\widetilde{L}} m_{f,\ell}\cdot\exp\left(-\frac{m_{f,\ell}\cdot(\Delta_{f,i}/2)^2}{2}\right) + \right. \\
&\qquad \left. \Delta_{f,i}\cdot N\cdot K\cdot \sum\nolimits_{\ell=1}^{\widetilde{L}} \exp\left(-\frac{m_{f,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)\right) \\
&\le \sum\nolimits_{i\in[K]} \left(\Delta_{f,i}\cdot m_{f,L_i^f} + \Delta_{f,i}\cdot\frac{8}{\Delta_{f,i}^2}\cdot\int_{(\alpha/2)\log(N)}^{\infty} xe^{-x}\,\mathrm{d}x + \Delta_{f,i}\cdot N\cdot K\cdot\frac{\widetilde{L}}{N^{\alpha/2}}\right) \\
&\le \sum\nolimits_{i\in[K]} \left(\Delta_{f,i}\cdot\frac{\alpha\cdot\log(N)}{(\Delta_{f,i}/2)^2} + \Delta_{f,i}\cdot\frac{8}{\Delta_{f,i}^2}\cdot\int_{(\alpha/2)\log(N)}^{\infty} xe^{-x}\,\mathrm{d}x + \Delta_{f,i}\cdot N\cdot K\cdot\frac{\widetilde{L}}{N^{\alpha/2}}\right) \\
&\le \sum\nolimits_{i\in[K]} \left(\frac{4\alpha\cdot\log(N)}{\Delta_{f,i}} + \frac{8}{\Delta_{f,i}}\cdot\frac{(\alpha/2)\log(N)+1}{N^{\alpha/2}} + \Delta_{f,i}\cdot N\cdot K\cdot\frac{\lceil\log\left(N/(\alpha\cdot\log(N))\right)\rceil}{N^{\alpha/2}}\right) \\
&\le O\left(\sum\nolimits_{i\in[K]}\frac{\log(N)}{\Delta_{f,i}}\right),
\end{aligned}$$

where the last inequality holds for $\alpha = 4$ and $K/N \le O(1)$. □

Proof for Lemma 4.3. We decompose the regret as the summation of the regret incurred by each suboptimal arm. Then, the high-level idea of the following upper bound is that for each suboptimal arm, we further decompose the regret into three scenarios:

(i) I_f^* and I_g^* are not eliminated before epoch $L+1$;

(ii) I_f^* is eliminated before epoch $L+1$;

(iii) I_g^* is eliminated before epoch $L+1$.

The probabilities of the occurrence of (*ii*) and (*iii*) should be small. Then, we take a closer look at the case of (*i*). According to the definitions of L_i^f and L_i^g, it should be of high probability that arm i is eliminated from $\mathcal{A}_{f,\ell}$ and $\mathcal{A}_{g,\ell}$ around epoch L_i^f and L_i^g, respectively. Our calculation of the additional number of pulls of arm i after it is eliminated from $\mathcal{A}_{f,\ell}$ and before it is eliminated from $\mathcal{A}_{g,\ell}$ also falls into four cases. Roughly speaking,

(a) Arm i is eliminated from $\mathcal{A}_{f,\ell}$ and $\mathcal{A}_{g,\ell}$ in epoch L_i^f and $L_i^g - 1$, respectively;

(b) Arm i is eliminated from $\mathcal{A}_{g,\ell}$ after epoch L_i^g;

(c) Arm i is eliminated from $\mathcal{A}_{g,\ell}$ no later than epoch L_i^g and eliminated from $\mathcal{A}_{f,i}$ before epoch $L_i^f - 1$;

(d) Arm i is eliminated from $\mathcal{A}_{g,\ell}$ no later than epoch $L_i^g - 1$ and eliminated from $\mathcal{A}_{f,i}$ no earlier than $L_i^f - 1$.

Case (a) occurs with high probability and leads to $(m_{g,L_i^g} - m_{f,L_i^f-1})^+$ number of additional pulls. (b), (c) and (d) occur with low probability.

Similar to before, in order to facilitate the analysis, we define the following events :

- $\text{Event}_2\left(i, n+1; i, \ell+1\right)$: Arms I_f^* and I_g^* remain uneliminated till epoch n and ℓ, respectively. And, arm i is eliminated from $\mathcal{A}_{f,n}$ and $\mathcal{A}_{g,\ell}$ after epoch n and ℓ, respectively, i.e., $\left\{i \notin \mathcal{A}_{f,n+1}, i \in \mathcal{A}_{f,n}, i \notin \mathcal{A}_{g,\ell+1}, i \in \mathcal{A}_{g,\ell}, I_f^* \in \mathcal{A}_{f,n}, I_g^* \in \mathcal{A}_{g,\ell}\right\}$.
- $\text{Event}_2\left(i, n+1; i, \ell'+1; I_f^*, n+1\right)$: Arm i is eliminated from $\mathcal{A}_{f,n'}$ and $\mathcal{A}_{g,\ell}$ after epoch n' and ℓ, respectively. And, arm I_f^* is eliminated after epoch n, respectively, i.e., $\left\{i \notin \mathcal{A}_{g,\ell+1}, i \in \mathcal{A}_{g,\ell}, i \notin \mathcal{A}_{f,n'+1}, i \in \mathcal{A}_{f,n'}, I_f^* \notin \mathcal{A}_{f,n+1}, I_f^* \in \mathcal{A}_{f,n}\right\}$.
- $\text{Event}_2\left(i, n+1; i, \ell'+1; I_g^*, \ell+1\right)$: Arm i is eliminated from $\mathcal{A}_{f,n}$ and $\mathcal{A}_{g,\ell'}$ after epoch n and ℓ', respectively. And, arm I_g^* is eliminated after epoch ℓ, respectively, i.e., $\left\{i \notin \mathcal{A}_{f,n+1}, i \in \mathcal{A}_{f,n}, i \notin \mathcal{A}_{g,\ell'+1}, i \in \mathcal{A}_{g,\ell'}, I_g^* \notin \mathcal{A}_{g,\ell+1}, I_g^* \in \mathcal{A}_{g,\ell}\right\}$.

Then,

$$
\begin{aligned}
\mathbf{REG}_2[N] \leq \sum\nolimits_{i\in[K]} \Bigg(& \Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{n=1}^{L} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
& + \Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{n=1}^{L} \sum\nolimits_{n'=n+1}^{L} (m_{g,\ell} - m_{f,n'})^+ \times \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell'+1; I_f^*, n+1\right)\right) \\
& + \Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{\ell'=\ell+1}^{L} \sum\nolimits_{n=1}^{L} (m_{g,\ell'} - m_{f,n})^+ \times \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell'+1; I_g^*, \ell+1\right)\right) \Bigg) \\
\leq \sum\nolimits_{i\in[K]} \Bigg(& \Delta_{f,i} \cdot \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{n=1}^{L} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
& + \Delta_{f,i} \cdot m_{g,L} \cdot \left(\mathbb{P}\left(\exists\ \ell \in [L] : I_f^* \notin \mathcal{A}_{f,\ell}\right) + \mathbb{P}\left(\exists\ \ell \in [L] : I_g^* \notin \mathcal{A}_{g,\ell}\right)\right) \Bigg), \qquad (11)
\end{aligned}
$$

For the first summation in (11), we decompose it into four parts,

$$
\begin{aligned}
& \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
& = \sum\nolimits_{\ell=L_i^g+1}^{L} \sum\nolimits_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) + \sum\nolimits_{\ell=1}^{L_i^g} \sum\nolimits_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
& = \sum\nolimits_{\ell=L_i^g+1}^{L} \sum\nolimits_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) + \sum\nolimits_{\ell=1}^{L_i^g} \sum\nolimits_{n=1}^{L_i^f-2} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
& \quad + \sum\nolimits_{\ell=1}^{L_i^g} \sum\nolimits_{n=L_i^f-1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{n=L_i^f-1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, L_i^g\right)\right) + \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, L_i^g+1\right)\right) \\
&\quad + \sum_{\ell=L_i^g+1}^{L} \sum_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g} \sum_{n=1}^{L_i^f-2} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g-2} \sum_{n=L_i^f-1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right).
\end{aligned}
$$

Notice that all the events in the above summations are mutually exclusive. Then,

$$
\begin{aligned}
&\sum_{\ell=1}^{L} \sum_{n=1}^{L} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&= \sum_{n=L_i^f-1}^{L} \left((m_{g,L_i^g-1} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, L_i^g\right)\right) + (m_{g,L_i^g} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, L_i^g+1\right)\right) \right) \\
&\quad + \sum_{\ell=L_i^g+1}^{L} \sum_{n=1}^{L} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g} \sum_{n=1}^{L_i^f-2} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g-2} \sum_{n=L_i^f-1}^{L} (m_{g,\ell} - m_{f,n})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\le \left(m_{g,L_i^g} - m_{f,L_i^f-1}\right)^+ + \sum_{\ell=L_i^g+1}^{L} m_{g,\ell} \cdot \sum_{n=1}^{L} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g} m_{g,L_i^g} \cdot \sum_{n=1}^{L_i^f-2} \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right) \\
&\quad + \sum_{\ell=1}^{L_i^g-2} \sum_{n=L_i^f-1}^{L} (m_{g,\ell} - m_{f,L_i^f-1})^+ \cdot \mathbb{P}\left(\text{Event}_2\left(i, n+1; i, \ell+1\right)\right)
\end{aligned}
$$

On the other hand, for the second term in (11), according to the proof for Lemma 4.2, we know

$$
\begin{aligned}
\mathbb{P}\left(\exists\ \ell \in [\widetilde{L}] : I_f^* \notin \mathcal{A}_{f,\ell}\right) &\le K \cdot \sum_{\ell=1}^{\widetilde{L}} \exp\left(-\frac{m_{f,\ell} \cdot \widetilde{\Delta}_\ell^2}{2}\right), \\
\mathbb{P}\left(\exists\ \ell \in [\widetilde{L}] : I_g^* \notin \mathcal{A}_{g,\ell}\right) &\le K \cdot \sum_{\ell=1}^{\widetilde{L}} \exp\left(-\frac{m_{g,\ell} \cdot \widetilde{\Delta}_\ell^2}{2}\right).
\end{aligned}
$$

Combining the above and remembering that L_i^g cannot exceed L, we have

$$
\begin{aligned}
&\text{(11)} \\
&\le \sum_{i\in[K]} \left(\Delta_{f,i} \cdot \left(\min\left\{ m_{g,L_i^g}, m_{g,L} \right\} - m_{f,L_i^f-1} \right)^+ + \Delta_{f,i} \cdot \sum_{\ell=L_i^g+1}^{L} m_{g,\ell} \cdot \mathbb{P}\left(i \notin \mathcal{A}_{g,\ell+1}, i \in \mathcal{A}_{g,\ell}, I_g^* \in \mathcal{A}_{g,\ell}\right) \right. \\
&\quad + \Delta_{f,i} \cdot m_{g,L_i^g} \cdot \sum_{n=1}^{L_i^f-2} \mathbb{P}\left(i \notin \mathcal{A}_{f,n+1}, i \in \mathcal{A}_{f,n}, I_f^* \in \mathcal{A}_{f,n}\right) \\
&\quad + \Delta_{f,i} \cdot \sum_{\ell=1}^{L_i^g-2} (m_{g,\ell} - m_{f,L_i^f-1})^+ \cdot \mathbb{P}\left(i \notin \mathcal{A}_{g,\ell+1}, i \in \mathcal{A}_{g,\ell}, I_g^* \in \mathcal{A}_{g,\ell}\right) \\
&\quad \left. + \Delta_{f,i} \cdot m_{g,L} \cdot \left(\mathbb{P}\left(\exists\ \ell \in [L] : I_f^* \notin \mathcal{A}_{f,\ell}\right) + \mathbb{P}\left(\exists\ \ell \in [L] : I_g^* \notin \mathcal{A}_{g,\ell}\right) \right) \right) \\
&\le \sum_{i\in[K]} \left(\Delta_{f,i} \cdot \left(\min\left\{ m_{g,L_i^g}, m_{g,L} \right\} - m_{f,L_i^f-1} \right)^+ + \Delta_{f,i} \cdot \sum_{\ell=L_i^g+1}^{L} m_{g,\ell} \cdot \exp\left(-\frac{m_{g,\ell} \cdot (\Delta_{g,i}/2)^2}{2}\right) \right.
\end{aligned}
$$

$$
\begin{aligned}
&+\Delta_{f,i}\cdot m_{g,L_i^g}\cdot\sum\nolimits_{\ell=1}^{L_i^f-2}\exp\left(-\frac{m_{f,\ell}\cdot\left(\widetilde{\Delta}_\ell-\Delta_{f,i}\right)^2}{2}\right)\\
&+\Delta_{f,i}\cdot\sum\nolimits_{\ell=1}^{L_i^g-2}\left(m_{g,\ell}-m_{f,L_i^f-1}\right)^+\cdot\exp\left(-\frac{m_{g,\ell}\cdot\left(\widetilde{\Delta}_\ell-\Delta_{g,i}\right)^2}{2}\right)\\
&+\Delta_{f,i}\cdot m_{g,L}\cdot K\cdot\sum\nolimits_{\ell=1}^{L}\left(\exp\left(-\frac{m_{f,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)+\exp\left(-\frac{m_{g,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)\right)\Bigg).
\end{aligned}
$$

Here $L_i^f=\min\left\{\ell:\widetilde{\Delta}_\ell\le\Delta_{f,i}/2\right\}$ and $L_i^g=\min\left\{\ell:\widetilde{\Delta}_\ell\le\Delta_{g,i}/2\right\}$. Let $m_{g,\ell}=\frac{\beta\cdot\log(T-N)}{(\widetilde{\Delta}_\ell)^2}$, then

$$
\begin{aligned}
&\le\sum\nolimits_{i\in[K]}\Delta_{f,i}\cdot\Bigg(\Delta_{f,i}\cdot\left(\min\left\{m_{g,L_i^g},m_{g,L}\right\}-m_{f,L_i^f-1}\right)^+
+\Delta_{f,i}\cdot\sum\nolimits_{\ell=L_i^g+1}^{L}m_{g,\ell}\cdot\exp\left(-\frac{m_{g,\ell}\cdot\left(\Delta_{g,i}/2\right)^2}{2}\right)\\
&\quad+\Delta_{f,i}\cdot\frac{\beta\log(T-N)}{\left(\Delta_{g,i}/2\right)^2}\cdot\sum\nolimits_{\ell=1}^{L_i^f-2}\exp\left(-\frac{\frac{\beta\log(T-N)}{(\widetilde{\Delta}_\ell)^2}\cdot\left(\widetilde{\Delta}_\ell/2\right)^2}{2}\right)\\
&\quad+\Delta_{f,i}\cdot\sum\nolimits_{\ell=1}^{L_i^g-2}\left(m_{g,\ell}-\frac{\alpha\log(N)}{\Delta_{f,i}^2}\right)^+\cdot\exp\left(-\frac{m_{g,\ell}\cdot\left(\widetilde{\Delta}_\ell/2\right)^2}{2}\right)\\
&\quad+\Delta_{f,i}\cdot\min\left\{m_{g,L},N\right\}\cdot K\cdot\sum\nolimits_{\ell=1}^{L}\left(\exp\left(-\frac{m_{f,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)+\exp\left(-\frac{m_{g,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)\right)\Bigg)\\
&\le\sum\nolimits_{i\in[K]}\Bigg(\Delta_{f,i}\cdot\Bigg(\min\left\{\frac{\beta\cdot\log(T-N)}{\left(\Delta_{g,i}/2\right)^2},\frac{\beta\cdot\log(T-N)}{(\varepsilon/2)^2}\right\}-\frac{\alpha\cdot\log(N)}{\left(2\Delta_{f,i}\right)^2}\Bigg)^+\\
&\quad+\Delta_{f,i}\cdot\frac{8}{\Delta_{g,i}^2}\cdot\int_{(\beta/2)\log(T-N)}^{\infty}xe^{-x}\,\mathrm{d}x\\
&\quad+\Delta_{f,i}\cdot\frac{\beta\log(T-N)}{\left(\Delta_{g,i}/2\right)^2}\cdot\frac{L_i^f}{(T-N)^{\beta/8}}+\Delta_{f,i}\cdot\frac{8}{\Delta_{g,i}^2}\cdot\left(\int_{\frac{(\alpha/8)\log(N)}{(\Delta_{f,i}/\Delta_{g,i})^2}}^{(\beta/2)\log(T-N)}xe^{-x}\,\mathrm{d}x\right)^+\\
&\quad+\Delta_{f,i}\cdot\min\left\{m_{g,L},N\right\}\cdot K\cdot L\cdot\left(\frac{1}{N^{\alpha/2}}+\frac{1}{(T-N)^{\beta/2}}\right)\Bigg)\\
&\le\sum\nolimits_{i\in[K]}\Bigg(\Delta_{f,i}\cdot\Bigg(\min\left\{\frac{\beta\cdot\log(T-N)}{\left(\Delta_{g,i}/2\right)^2},\frac{\beta\cdot\log(T-N)}{(\varepsilon/2)^2}\right\}-\frac{\alpha\cdot\log(N)}{\left(2\Delta_{f,i}\right)^2}\Bigg)^+\\
&\quad+\Delta_{f,i}\cdot\frac{8}{\Delta_{g,i}^2}\cdot\frac{(\beta/2)\log(T-N)+1}{(T-N)^{\beta/2}}\\
&\quad+\Delta_{f,i}\cdot\frac{\beta\log(T-N)}{\left(\Delta_{g,i}/2\right)^2}\cdot\frac{\lceil\log\left(1/\Delta_{f,i}\right)\rceil}{(T-N)^{\beta/8}}+\Delta_{f,i}\cdot\frac{8}{\Delta_{g,i}^2}\cdot\left(\frac{\frac{(\alpha/8)\log(N)}{(\Delta_{f,i}/\Delta_{g,i})^2}+1}{N^{\frac{(\alpha/8)}{(\Delta_{f,i}/\Delta_{g,i})^2}}}-\frac{(\beta/2)\log(T-N)+1}{(T-N)^{\beta/2}}\right)^+\\
&\quad+\Delta_{f,i}\cdot\min\left\{\frac{\beta\cdot\log(T-N)}{\left(\varepsilon/2\right)^2},N\right\}\cdot K\cdot\lceil\log(1/\varepsilon)\rceil\cdot\left(\frac{1}{N^{\alpha/2}}+\frac{1}{(T-N)^{\beta/2}}\right)\Bigg)
\end{aligned}
$$

$$\leq O\left(\sum\nolimits_{i\in[K]}\Delta_{f,i}\cdot\max\left\{\frac{\log(T-N)}{\max\{\Delta_{g,i},\varepsilon\}^2},\frac{\log(N)}{\Delta_{f,i}^2}\right\}\right),$$

where the last inequality holds with $\alpha=4$, $\beta=4$, $\varepsilon=o(1)$ since $N>K$ and $N<T$. □

Proof for Lemma 4.4. In the commitment phase, we consider two possible scenarios:

(i). all arms remained in the active arm set $\mathcal{A}_{g,L+1}$ satisfy $\max_{i\in\mathcal{A}_{g,L+1}}\Delta_{g,i}<2\varepsilon$;

(ii). at least one arm in $\mathcal{A}_{g,L+1}$ has $\Delta_{g,i}\geq 2\varepsilon$.

Intuitively, scenario (i) should occur with high probability, and once it occurs, the regret incurred in the commitment phase is upper bounded by $(T-N)\cdot\max_{\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}$. On the other hand, scenario (ii) should occur with low probability, and once it occurs, since we commit to an arbitrary arm in $\mathcal{A}_{g,L+1}$, the regret incurred by committing to arm $i\in\mathcal{A}_{g,L+1}$ is bounded by $(T-N)\cdot\Delta_{g,i}$. Thus, we need to upper bound the probability of scenario (ii). We decompose it into the probability summation of the following two events:

- $\text{Event}_3(i)$: Both arms i and I_g^* remain uneliminated from $\mathcal{A}_{g,L}$ after epoch L, i.e., $\left\{i\in\mathcal{A}_{g,L+1},I_g^*\in\mathcal{A}_{g,L+1}\right\}$.
- $\text{Event}_3\left(i,I_g^*\right)$: Arm i remains uneliminated from $\mathcal{A}_{g,L}$ after epoch L but arm I_g^* is eliminated earlier, i.e., $\left\{i\in\mathcal{A}_{g,L+1},I_g^*\notin\mathcal{A}_{g,L+1}\right\}$.

We use Azuma's inequality to upper bound the probabilities of those two events. In particular, we have

$$\begin{aligned}
\mathbb{P}\left(\text{Event}_3\left(i\right)\right)&\leq\exp\left(-\frac{m_{g,L}\cdot\left((\Delta_{g,i}-\widetilde{\Delta}_L)\vee 0\right)^2}{2}\right)\leq\exp\left(-\frac{m_{g,L}\cdot(\Delta_{g,i}/2)^2}{2}\right)\\
&\leq\exp\left(-\frac{\beta\cdot\log(T-N)}{\varepsilon^2}\cdot\frac{\Delta_{g,i}^2}{8}\right),\\
\mathbb{P}\left(\text{Event}_3\left(i,I_g^*\right)\right)&\leq K\cdot\sum\nolimits_{\ell=1}^{L}\exp\left(-\frac{m_{g,\ell}\cdot\widetilde{\Delta}_\ell^2}{2}\right)=\frac{K\cdot L}{(T-N)^{\beta/2}}=\frac{K\cdot\lceil\log(1/\varepsilon)\rceil}{(T-N)^{\beta/2}}.
\end{aligned}$$

where we have used the fact that $m_{g,\ell}=\frac{\beta\cdot\log(T-N)}{(\widetilde{\Delta}_\ell)^2}$. Now we can upper bound the regret $\mathbf{REG}_3[N,T]$ as follows:

$$\begin{aligned}
&\mathbf{REG}_3[N,T]\\
&\leq(T-N)\max_{\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}+(T-N)\max_{i:\Delta_{g,i}\geq 2\varepsilon}\left\{\Delta_{g,i}\mathbb{P}\left(i\in\mathcal{A}_{g,L+1}\right)\right\}\\
&\leq(T-N)\max_{\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}+(T-N)\max_{i:\Delta_{g,i}\geq 2\varepsilon}\left\{\Delta_{g,i}\left(\mathbb{P}\left(\text{Event}_3\left(i\right)\right)+\mathbb{P}\left(\text{Event}_3\left(i,I_g^*\right)\right)\right)\right\}\\
&\leq(T-N)\max_{\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}+(T-N)\max_{i:\Delta_{g,i}\geq 2\varepsilon}\left\{\Delta_{g,i}\left(\exp\left(-\frac{\beta\log(T-N)}{\varepsilon^2}\frac{\Delta_{g,i}^2}{8}\right)+\frac{K\lceil\log(1/\varepsilon)\rceil}{(T-N)^{\beta/2}}\right)\right\}\\
&\leq O\left((T-N)\max_{i:\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}\right),
\end{aligned}$$

where the last inequality holds for $\beta=4$, $K/T\leq O(1)$. □

With the above three lemmas, we are ready to prove Proposition 4.1.

Proof of Proposition 4.1. We simply take the summation of the regret upper bounds established in Lemmas 4.2 to 4.4,

$$\begin{aligned}\mathbf{REG}[N,T \mid \mathcal{I}] &= \mathbf{REG}_1[N] + \mathbf{REG}_2[N] + \mathbf{REG}_3[N,T] \\ &\leq O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\max\{\Delta_{g,i},\varepsilon\}^2}, \frac{\log(N)}{\Delta_{f,i}^2}\right\} + (T-N)\cdot \max_{i:\Delta_{g,i}<2\varepsilon}\{\Delta_{g,i}\}\right),\end{aligned}$$

which finishes the proof. □

Theorem 3.1 (Regret upper bound)**.** *With the parameter*

$$\varepsilon = \max\left\{\sqrt{\frac{K\log(T-N)}{N}}, \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}, \sqrt{\frac{KT\log(T)}{(T-N)^2}}\right\},$$

we have the following regret upper bound for Algorithm RAEC:

$$\mathbf{REG}[N,T] \leq \widetilde{O}\left(\sqrt{\frac{K(T-N)^{2/3}\cdot\max\left\{(T-N)^{4/3}, K^{1/3}N\right\}}{\min\left\{N, K^{1/3}(T-N)^{2/3}\right\}}}\right).$$

Proof for Theorem 3.1. Under the choice of ε in Theorem 3.1, we have three region discussions as N grows.

Region I. When $N < K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}$, we have $\varepsilon = \sqrt{\frac{K\log(T-N)}{N}}$, then

$$(4) \leq O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i}\cdot\max\left\{\frac{N}{K}, \frac{\log(N)}{\Delta_{f,i}^2}\right\} + (T-N)\cdot\sqrt{\frac{K\log(T-N)}{N}}\right)$$

We divide arms $[K]$ into two groups:

- Group 1 includes all arms in $[K]$ satisfying $\Delta_{f,i} < \sqrt{\frac{K\log(N)}{N}}$;
- Group 2 includes all arms in $[K]$ satisfying $\Delta_{f,i} \geq \sqrt{\frac{K\log(N)}{N}}$.

Notice that $\sum_{i\in\text{Group 1}} \mathbb{E}[n_i(N)] \leq N$, then the above,

$$\begin{aligned}&\leq O\left(\sum\nolimits_{i\in\text{Group 1}} \Delta_{f,i}\cdot\mathbb{E}[n_i(N)] + \sum\nolimits_{i\in\text{Group 2}} \Delta_{f,i}\cdot\frac{N}{K} + (T-N)\cdot\sqrt{\frac{K\log(T-N)}{N}}\right) \\ &= O\left(T\cdot\sqrt{\frac{K\log(T)}{N}} + N\right) = O\left(T\cdot\sqrt{\frac{K\log(T)}{N}}\right), \qquad (12)\end{aligned}$$

where the last equality holds due to the scenario condition that $N < K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}$.

Region II. When $N \geq K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}$ and $(T-N) \geq K^{1/4}N^{3/4}\log(N)^{3/4}/\log(T-N)^{1/2}$, we have $\varepsilon = \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}$, then

$$(4) \leq O\left(\sum\nolimits_{i\in[K]} \Delta_{f,i}\max\left\{\frac{(T-N)^{2/3}\log(T-N)^{1/3}}{K^{2/3}}, \frac{\log(N)}{\Delta_{f,i}^2}\right\} + (T-N)\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}\right)$$

Similarly, we divide arms $[K]$ into two groups:

- Group 1 includes all arms in $[K]$ satisfying $\Delta_{f,i} < \left(\frac{K^{2/3}\cdot\log(N)}{(T-N)^{2/3}\cdot\log(T-N)^{1/3}}\right)^{1/2}$. We further divide all arms in Group 1 as following two subgroups:
  - Group 1a includes all arms in Group 1 satisfying $\Delta_{f,i} < \frac{K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}}{N}$
  - Group 1b includes all arms in Group 1 satisfying $\frac{K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}}{N} \leq \Delta_{f,i} < \left(\frac{K^{2/3}\cdot\log(N)}{(T-N)^{2/3}\cdot\log(T-N)^{1/3}}\right)^{1/2}$
- Group 2 includes all arms in $[K]$ satisfying $\Delta_{f,i} \geq \left(\frac{K^{2/3}\cdot\log(N)}{(T-N)^{2/3}\cdot\log(T-N)^{1/3}}\right)^{1/2}$

Notice that $\sum_{i\in\text{Group 1}}\mathbb{E}[n_i(N)] \leq N$, then the above,

$$
\begin{aligned}
&\leq O\left(\sum\nolimits_{i\in\text{Group 1a}}\Delta_{f,i}\mathbb{E}[n_i(N)] + \sum\nolimits_{i\in\text{Group 1b}}\frac{\log(N)}{\Delta_{f,i}} + \sum\nolimits_{i\in\text{Group 2}}\Delta_{f,i}\frac{(T-N)^{2/3}\log(T-N)^{1/3}}{K^{2/3}}\right.\\
&\quad\left.+(T-N)\left(\frac{K\log(T-N)}{T-N}\right)^{1/3}\right)\\
&\overset{(a)}{\leq} O\left(K\frac{(T-N)^{4/3}\log(T-N)^{2/3}/K^{1/3}}{K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}} + (T-N)\left(\frac{K\log(T-N)}{T-N}\right)^{1/3}\right)\\
&\leq O\left(K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right), \qquad (13)
\end{aligned}
$$

where inequality (a) utilizes the fact that $(T-N) \geq K^{1/4}N^{3/4}\log(N)^{3/4}/\log(T-N)^{1/2}$ which is equivalent to $N\log(N) \leq (T-N)^{4/3}\log(T-N)^{2/3}/K^{1/3}$.

Region III. When $\sqrt{KN\log(N)} \leq (T-N) < K^{1/4}N^{3/4}\log(N)^{3/4}/\log(T-N)^{1/2}$ (which implies $N = T - o(T)$), the condition can be equivalently written as $\sqrt{KT\log(T)} \leq (T-N) < K^{1/4}T^{3/4}\log(T)^{3/4}/\log(T-N)^{1/2}$. We have $\varepsilon = \sqrt{\frac{KT\log(T)}{(T-N)^2}}$, then

$$
(4) \leq O\left(\sum\nolimits_{i\in[K]}\Delta_{f,i}\cdot\max\left\{\frac{(T-N)^2\log(T-N)}{KT\log(T)}, \frac{\log(T)}{\Delta_{f,i}^2}\right\} + (T-N)\cdot\sqrt{\frac{KT\log(T)}{(T-N)^2}}\right)
$$

Again, we divide arms $[K]$ into two groups:

- Group 1 includes all arms in $[K]$ satisfying $\Delta_{f,i} < \left(\frac{KT(\log T)^2}{(T-N)^2\cdot\log(T-N)}\right)^{1/2}$. We further divide all arms in Group 1 as following two subgroups:
  - Group 1a includes all arms in Group 1 satisfying in $\Delta_{f,i} < \sqrt{\frac{K\log(T)}{T}}$;
  - Group 1b includes all arms in Group 1 satisfying $\sqrt{\frac{K\log(T)}{T}} \leq \Delta_{f,i} < \left(\frac{KT(\log T)^2}{(T-N)^2\cdot\log(T-N)}\right)^{1/2}$
- Group 2 includes all arms in $[K]$ satisfying $\Delta_{f,i} \geq \left(\frac{KT(\log T)^2}{(T-N)^2\cdot\log(T-N)}\right)^{1/2}$

Notice that $\sum_{i\in\text{Group 1a}}\mathbb{E}[n_i(N)]\le N$, then the above,

$$\begin{aligned}
&\le O\left(\sum\nolimits_{i\in\text{Group 1a}}\Delta_{f,i}\cdot\mathbb{E}[n_i(N)]+\sum\nolimits_{i\in\text{Group 1b}}\Delta_{f,i}\cdot\frac{\log(N)}{\Delta_{f,i}^2}+\right.\\
&\quad\left.\sum\nolimits_{i\in\text{Group 2}}\Delta_{f,i}\cdot\frac{(T-N)^2\cdot\log(T-N)}{KT\log(T)}+\sqrt{KT\log(T)}\right)\\
&\le O\left(\sqrt{\frac{K\log(T)}{T}}N+K\sqrt{\frac{T\log(T)}{K}}+\frac{(T-N)^2\log(T-N)}{T\log(T)}+\sqrt{KT\log(T)}\right)\le O\left(\sqrt{KT\log(T)}\right).
\end{aligned}\tag{14}$$

When $T-N<\sqrt{KT\log(T)}$, we have $\varepsilon\ge 1$, that is, the algorithm does not explore for the optimal arm in the commitment phase at all. It is easy to check that

$$(4)\le O\left(\sqrt{KT\log(T)}\right).\tag{15}$$

Combining (12), (13), (14), and (15), the instance-independent upper bound has form

$$O\left(\sqrt{\frac{K(T-N)^{2/3}\log(T-N)\cdot\max\left\{(T-N)^{4/3},K^{1/3}N\log(N)/\log(T-N)^{2/3}\right\}}{\min\left\{N,K^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right\}}}\right)\tag{16}$$

Drop the logarithmic terms in (16) and write it in the notation $\widetilde{O}$ would give us the desired bound. We completed the proof. □

B Missing Proofs in Section 5

Theorem 5.1 (Instance-dependent lower bound)**.** *Given instance $\mathcal{I}\in\mathcal{E}$, there exists an instance-dependent regret lower bound for all consistent policies as follows,*

$$\mathbf{REG}[N,T\mid\mathcal{I}]=\Omega\left(\sum\nolimits_{i\in[K]}\max\left\{\frac{\log(N)}{d_{f,i}},\frac{\log(T-N)}{d_{g,i}}\right\}\cdot\Delta_{f,i}\right).$$

Proof for Theorem 5.1. For convenience, we define the experiment phase instance space as $\mathcal{E}_{\mathsf{exp}}=\mathcal{V}_{f,1}\times\mathcal{V}_{f,2}\times\cdots\times\mathcal{V}_{f,K}$ and the commitment phase instance space as $\mathcal{E}_{\mathsf{com}}=\mathcal{V}_{g,1}\times\mathcal{V}_{g,2}\times\cdots\times\mathcal{V}_{g,K}$.

Consider an instance $\mathcal{I}=(P_j)_{j=1}^K\in\mathcal{E}_{\mathsf{exp}}$. Let μ_i be the mean of the ith arm in $\mathcal{I}$ and $d_{f,i}=d_{\inf}(P_i,\mu^*,\mathcal{V}_{f,i})$ with $\mu^*=\max_{i\in[K]}\mu_i$. Fix a suboptimal arm i, and let $\varepsilon>0$ be arbitrary and $\mathcal{I}'=(P'_j)_{j=1}^K\in\mathcal{E}_{\mathsf{exp}}$ be a bandit with $P'_j=P_j$ for $j\ne i$ and $P'_i\in\mathcal{V}_{f,i}$ such that $D(P_i,P'_i)\le d_{f,i}+\varepsilon$ and $\mu(P'_i)>\mu^*$, which exists by the definition of $d_{f,i}$. By *Divergence Decomposition Theorem*, we have $D(P_{\mathcal{I}},P_{\mathcal{I}'})\le\mathbb{E}_{\mathcal{I}}[T_i(N)](d_{f,i}+\varepsilon)$, where $T_i(N)$ is the realized number of pulls of arm i throughout the experiment phase. And by *Bretagnolle–Huber inequality*, for any event A,

$$P_{\mathcal{I}}(A)+P_{\mathcal{I}'}(A^c)\ge\frac{1}{2}\exp(-D(P_{\mathcal{I}},P_{\mathcal{I}'}))\ge\frac{1}{2}\exp(-\mathbb{E}_{\mathcal{I}}[T_i(N)](d_{f,i}+\varepsilon)).$$

Now choose $A=\{T_i(N)\ge N/2\}$, and let $\mathbf{REG}_{\mathsf{exp}}[N\mid\mathcal{I}]$ be the expected regret of executing policy π under instance $\mathcal{I}$ in the experiment phase. And similarly, we define $\mathbf{REG}_{\mathsf{exp}}[N\mid\mathcal{I}']$. Then,

$$\begin{aligned}
\mathbf{REG}_{\mathsf{exp}}[N\mid\mathcal{I}]+\mathbf{REG}_{\mathsf{exp}}\left[N\mid\mathcal{I}'\right]&\ge\mathbb{E}_{\mathcal{I}}[T_i(N)]\cdot\Delta_i+(N-\mathbb{E}_{\mathcal{I}'}[T_i(N)])\cdot(\mu'_i-\mu^*)\\
&\ge\frac{N}{2}(\Delta_iP_{\mathcal{I}}(A)+P_{\mathcal{I}'}(A^c)(\mu'_i-\mu^*))
\end{aligned}$$

$$\geq \frac{N}{2}\min\{\Delta_i, \mu_i' - \mu^*\}(P_{\mathcal{I}}(A) + P_{\mathcal{I}'}(A^c))$$
$$\geq \frac{N}{4}\min\{\Delta_i, \mu_i' - \mu^*\}\exp(-\mathbb{E}_{\mathcal{I}}[T_i(N)](d_{f,i}+\varepsilon)),$$

where the first inequality uses the regret decomposition formula.

Rearranging and taking the limit inferior leads to

$$\liminf_{T-N\to\infty}\frac{\mathbb{E}_{\mathcal{I}}[T_i(N)]}{\log(N)} \geq \frac{1}{d_{f,i}+\varepsilon}\cdot\liminf_{N\to\infty}\frac{\log\left(\frac{N\min\{\Delta_i,\mu_i'-\mu^*\}}{4\left(R_N+R_N'\right)}\right)}{\log(N)}$$
$$= \frac{1}{d_{f,i}+\varepsilon}\cdot\left(1-\limsup_{N\to\infty}\frac{\log\left(R_N+R_N'\right)}{\log(N)}\right) = \frac{1}{d_{f,i}+\varepsilon},$$

where the last equality follows from the definition of consistency, which says that for any $p>0$, there exists a constant C_p such that for sufficiently large N, $\mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}'] \leq C_p N^p$, which implies that

$$\limsup_{N\to\infty}\frac{\log(\mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{exp}}[N \mid \mathcal{I}'])}{\log(N)} \leq \limsup_{N\to\infty}\frac{p\log(N)+\log(C_p)}{\log(N)} = p,$$

which gives the result since $p>0$ was arbitrary and by taking the limit as ε tends to zero.

Analogously, we let $\mathcal{I} = (P_i)_{i=1}^K \in \mathcal{E}_{\mathsf{com}}$, $d_{g,i} = d_{\inf}(P_i, \mu^*, \mathcal{V}_{g,i})$. We also define $\mathcal{I}' = (P_i')_{i=1}^K \in \mathcal{E}_{\mathsf{com}}$ similar as before. We choose event $A = \{J_N = i\}$ and define the expected regrets in the committing phase under instance $\mathcal{I}$ and $\mathcal{I}'$ as $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}]$ and $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}']$, respectively. Notice that the regret in the committing phase is the simple regret multiplied by the committing phase length. Then,

$$\begin{aligned}
&\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{com}}\left[T-N \mid \mathcal{I}'\right] \\
&\geq (T-N)\cdot(\Delta_i P_{\mathcal{I}}(A) + P_{\mathcal{I}'}(A^c)(\mu_i' - \mu^*)) \\
&\geq (T-N)\cdot\min\{\Delta_i, \mu_i' - \mu^*\}(P_{\mathcal{I}}(A) + P_{\mathcal{I}'}(A^c)) \\
&\geq \frac{(T-N)}{2}\cdot\min\{\Delta_i, \mu_i' - \mu^*\}\exp(-\mathbb{E}_{\mathcal{I}}[T_i(N)](d_{g,i}+\varepsilon)).
\end{aligned}$$

Rearranging and taking the limit inferior leads to

$$\begin{aligned}
\liminf_{T-N\to\infty}\frac{\mathbb{E}_{\mathcal{I}}[T_i(N)]}{\log(T-N)} &\geq \frac{1}{d_{g,i}+\varepsilon}\cdot\liminf_{T-N\to\infty}\frac{\log\left(\frac{(T-N)\min\{\Delta_i,\mu_i'-\mu^*\}}{4(\mathbf{REG}_{\mathsf{com}}[T-N|\mathcal{I}]+\mathbf{REG}_{\mathsf{com}}[T-N|\mathcal{I}'])}\right)}{\log(T-N)} \\
&= \frac{1}{d_{g,i}+\varepsilon}\cdot\left(1-\limsup_{T-N\to\infty}\frac{\log\left(\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}']\right)}{\log(T-N)}\right) \\
&= \frac{1}{d_{g,i}+\varepsilon},
\end{aligned}$$

where the last equality follows from the definition of consistency, which says that for any $p>0$, there exists a constant C_p such that for sufficiently large $T-N$, $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}'] \leq C_p(T-N)^p$, which implies that

$$\limsup_{T-N\to\infty}\frac{\log(\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}'])}{\log(T-N)} \leq \limsup_{T-N\to\infty}\frac{p\log(T-N)+\log(C_p)}{\log(T-N)} = p,$$

which gives the result since $p > 0$ was arbitrary and by taking the limit as ε tends to zero.

In terms of the regret lower bound for the committing phase, again, we let $\mathcal{I} = (P_i)_{i=1}^K \in \mathcal{E}_{\mathsf{com}}$, $d_{g,i} = d_{\inf}(P_i, \mu^*, \mathcal{V}_{g,i})$. Assume the optimal arm is i^*, fix a suboptimal arm i, and let $\mathcal{I}' = (P'_j)_{j=1}^K \in \mathcal{E}_{\mathsf{com}}$ be a bandit with $P'_j = P_j$ for $j \neq i, i^*$ and $P'_i = P_{i^*}$, $P'_{i^*} = P_i$, i.e., we switch the position of arm i and i^* in $\mathcal{I}'$. Recall that *consistent* policies are symmetric under arm permutations. That is, the policies will generate the same expected regret for the instances that differ by only arm permutations. Therefore, for the above two instances $\mathcal{I}$ and $\mathcal{I}'$, $\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] = \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}']$. We still let $A = \{J_N = i\}$. Then, we have

$$
\begin{aligned}
&2\mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] = \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] + \mathbf{REG}_{\mathsf{com}}\left[T-N \mid \mathcal{I}'\right] \\
&\geq (T-N) \cdot (\Delta_i P_{\mathcal{I}}(A) + \Delta P_{\mathcal{I}'}(A^c)) \\
&\geq (T-N) \cdot \Delta (P_{\mathcal{I}}(A) + P_{\mathcal{I}'}(A^c)) \\
&\geq \frac{(T-N)}{2} \cdot \Delta \exp(-(\mathbb{E}_{\mathcal{I}}[T_i(N)] D(P_i, P_{i^*}) + \mathbb{E}_{\mathcal{I}}[T_{i^*}(N)] D(P_{i^*}, P_i))) \\
&\geq \frac{(T-N)}{2} \cdot \Delta \exp(-N \cdot D_{\max}),
\end{aligned}
$$

where Δ is the mean difference between the optimal arm and the second-best arm in $\mathcal{I}$ and $D_{\max} = D(P_i, P_{i^*}) \vee D(P_{i^*}, P_i)$. In summary, the instance-dependent lower bound for the total regret is then given by

$$
\begin{aligned}
\mathbf{REG}[N, T \mid \mathcal{I}] &\geq \sum\nolimits_{i \in [K]} \mathbb{E}_{\mathcal{I}}[T_i(N)] \cdot \Delta_i + \mathbf{REG}_{\mathsf{com}}[T-N \mid \mathcal{I}] \\
&\geq \sum\nolimits_{i \in [K]} \max\left\{ \frac{\log(N)}{d_{f,i}}, \frac{\log(T-N)}{d_{g,i}} \right\} \cdot \Delta_i + \frac{T-N}{4} \cdot \Delta \cdot \exp\left(-N \cdot D_{max}\right)
\end{aligned}
$$

Recall that $N \geq \Omega(\mathsf{poly}(T))$, the lower bound can be asmptotically written as $\Omega\left(\sum_{i \in [K]} \max\left\{ \frac{\log(N)}{d_{f,i}}, \frac{\log(T-N)}{d_{g,i}} \right\} \cdot \Delta_i \right)$. □

Theorem 5.2 (Instance-independent lower bound)**.** *There exists an instance-independent regret lower bound for all policies as follows,*

$$
\mathbf{REG}[N, T] = \Omega\left(\sqrt{ \frac{K(T-N)^{2/3} \cdot \max\left\{(T-N)^{4/3}, K^{1/3} N\right\}}{\min\left\{N, K^{1/3}(T-N)^{2/3}\right\}} } \right).
$$

Proof for Theorem 5.2. We consider an instance where in the *experiment phase*, the mean reward vector has form $\mu_f = (\alpha + \delta_f, \alpha, \ldots, \alpha, 0, \ldots, 0)$ where the last $K/2$ arms have zero mean, and in the *commitment phase*, the mean reward vector has form $\mu_g = (0, \ldots, 0, \delta_g, 0, \ldots, 0)$ where δ_g is on the kth arm's position and $k \equiv K/2 + 1$, α is some positive constant. We group the first $K/2$ arms as arm group A and the other half of arms as arm group B. Given policy π, let $\mathbb{E}_{\mathcal{I}}[T_A]$ and $\mathbb{E}_{\mathcal{I}}[T_B]$ be the expected number of pulls distributed to group A and B, respectively. Clearly, $\mathbb{E}_{\mathcal{I}}[T_A] + \mathbb{E}_{\mathcal{I}}[T_B] = N$. Let $\mathbb{E}_{\mathcal{I}}[T_l]$ be policy π's expected number of pulls of arm l. Then, we define $i^\dagger = \arg\min_{l \in A \setminus \{1\}} \mathbb{E}_{\mathcal{I}}[T_l]$ and $j^\dagger = \arg\min_{l \in B \setminus \{k\}} \mathbb{E}_{\mathcal{I}}[T_l]$. Clearly, $\mathbb{E}_{\mathcal{I}}[T_{i^\dagger}] \leq \frac{\mathbb{E}_{\mathcal{I}}[T_A]}{\frac{K}{2}-1}$ and $\mathbb{E}_{\mathcal{I}}\left[T_{j^\dagger}\right] \leq \frac{\mathbb{E}_{\mathcal{I}}[T_B]}{\frac{K}{2}-1}$. Now, we construct an alternative instance with $\mu_f^\dagger = (\alpha + \delta_f, \alpha, \ldots, \alpha + 2\delta_f, \ldots, \alpha, 0, \ldots, 0)$ where arm $i^\dagger$'s position has value $\alpha + 2\delta_f$, and $\mu_g^\dagger = (0, \ldots, 0, \delta_g, 0, \ldots, 2\delta_g, \ldots, 0)$ where arm $j^\dagger$'s position has value $2\delta_g$. Parameters α, δ_f and δ_g will be specified shortly.

$$
\mathbf{REG}_\pi[N, T \mid \mathcal{I}] + \mathbf{REG}_\pi\left[N, T \mid \mathcal{I}^\dagger\right]
$$

$$
\begin{aligned}
&\geq \mathbb{E}_{\mathcal{I}}[T_B]\cdot\alpha + \mathbb{E}_{\mathcal{I}^\dagger}[T_B]\cdot\alpha + \mathbb{P}_{\mathcal{I}}\left[T_1 \leq \frac{N}{2}\right]\cdot\frac{N\cdot\delta_f}{2} + \mathbb{P}_{\mathcal{I}^\dagger}\left[T_1 > \frac{N}{2}\right]\cdot\frac{N\cdot\delta_f}{2} \\
&\quad + \mathbb{P}_{\mathcal{I}}[I_{N+1}\neq k]\cdot(T-N)\cdot\delta_g + \mathbb{P}_{\mathcal{I}^\dagger}[I_{N+1}=k]\cdot(T-N)\cdot\delta_g \\
&\geq \mathbb{E}_{\mathcal{I}}[T_B]\cdot\alpha + \frac{N\cdot\delta_f}{2}\cdot\frac{1}{2}\exp\left(-D\left(\mathbb{P}_{\mathcal{I}},\mathbb{P}_{\mathcal{I}^\dagger}\right)\right) + (T-N)\cdot\delta_g\cdot\frac{1}{2}\exp\left(-D\left(\mathbb{P}_{\mathcal{I}},\mathbb{P}_{\mathcal{I}^\dagger}\right)\right) \\
&= \mathbb{E}_{\mathcal{I}}[T_B]\cdot\alpha + \left(\frac{N\cdot\delta_f}{2} + (T-N)\cdot\delta_g\right)\cdot\frac{1}{2}\exp\left(-\mathbb{E}_{\mathcal{I}}[T_{i^\dagger}]\cdot D(\alpha,\alpha+2\delta_f) - \mathbb{E}_{\mathcal{I}}\left[T_{j^\dagger}\right]\cdot D(0,2\delta_g)\right) \\
&\sim \Omega\left(\mathbb{E}_{\mathcal{I}}[T_B] + (N\cdot\delta_f + (T-N)\cdot\delta_g)\cdot\exp\left(-\frac{N-\mathbb{E}_{\mathcal{I}}[T_B]}{K}\cdot(\delta_f)^2 - \frac{\mathbb{E}_{\mathcal{I}}[T_B]}{K}\cdot(\delta_g)^2\right)\right), \qquad (17)
\end{aligned}
$$

where the second inequality again uses the *Bretagnolle-Huber inequality*, the last approximation utilizes the fact that for many common distributions like Gaussian, $D(a,b)\approx(a-b)^2$.

If $N = o\left(K^{1/3}T^{2/3}\right)$, the regret-minimizing policy should have $\mathbb{E}_{\mathcal{I}}[T_B] = N$, then

$$
(17) = \Omega\left(N + (N\cdot\delta_f + (T-N)\cdot\delta_g)\cdot\exp\left(-\frac{N}{K}\cdot(\delta_g)^2\right)\right).
$$

Let $\delta_g = \sqrt{\frac{K}{N}}$, we get the minimax lower bound of $\Omega\left(\sqrt{\frac{K\cdot(T-N)^2}{N}}\right)$. If $N \geq \Omega\left(K^{1/3}(T-N)^{2/3}\right)$, the regret-minimizing policy should have $\mathbb{E}_{\mathcal{I}}[T_B] = \Omega\left(K^{1/3}(T-N)^{2/3}\right)$, then

$$
(17) = \Omega\left(K^{1/3}(T-N)^{2/3} + (N\cdot\delta_f + (T-N)\cdot\delta_g)\cdot\exp\left(-\frac{N}{K}\cdot(\delta_f)^2 - \left(\frac{T-N}{K}\right)^{2/3}\cdot(\delta_g)^2\right)\right).
$$

Let $\delta_f = \sqrt{\frac{K}{N}}$, $\delta_g = \left(\frac{K}{T-N}\right)^{1/3}$, we get the minimax lower bound of $\Omega\left(\max\left\{K^{1/2}N^{1/2}, K^{1/3}(T-N)^{2/3}\right\}\right)$. When $(T-N) \geq \Omega(K^{1/4}N^{3/4})$, the bound can be further written as $\Omega\left(K^{1/3}(T-N)^{2/3}\right)$; otherwise, the lower bound is $\Omega(K^{1/2}N^{1/2})$.

In summary, the minimax lower bound has form $\Omega\left(\sqrt{\frac{K(T-N)^{2/3}\cdot\max\left\{(T-N)^{4/3},K^{1/3}N\right\}}{\min\left\{N,K^{1/3}(T-N)^{2/3}\right\}}}\right)$. □

C Missing Proofs in Section 6

C.1 Missing Proof in Section 6.1

Proposition C.1. *When the reward shift satisfies the structure specified in Definition 6.1 and the parameters M, D satisfy*

$$
M\sqrt{K/(T-N)} + D \lesssim \left(\frac{K}{T-N}\right)^{1/3} \qquad (18)
$$

then with $\varepsilon = M\cdot\left(\frac{K}{T-N}\right)^{1/2} + D$, the expected regret of Algorithm RAEC in the balanced scenario (i.e., $\sqrt{\frac{K}{N}} \vee \frac{\sqrt{KN}}{T-N} \lesssim \left(\frac{K}{T-N}\right)^{1/3}$) can be improved as

$$
\mathbf{REG}[N,T] = \widetilde{O}\left(\frac{K\cdot\left(\sqrt{K/(T-N)} + 2\cdot D/M\right)}{\left(M\cdot\sqrt{K/(T-N)} + D\right)^2} + T\cdot\left(M\cdot\sqrt{\frac{K}{T-N}} + D\right)\right). \qquad (19)
$$

Proof for Proposition C.1. Notice that under the reward shift structure specified in Definition 6.1,

$$\Delta_{g,i} = g_j - g_i = M \cdot (f_j - f_i) + (\delta_j - \delta_i) \leq M \cdot (f_j - f_i) + D \leq M \cdot (f_k - f_i) + D = M \cdot \Delta_{f,i} + D,$$

where $f_i \equiv \mathbb{E}[f(o_i)]$, $g_i \equiv \mathbb{E}[g(o_i)]$, $k = \arg\max_{i\in[K]} f_i$, $j = \arg\max_{i\in[K]} g_i$. Similarly, we have $\Delta_{g,i} \geq M \cdot \Delta_{f,i} - D$. In summary, we have

$$M \cdot \Delta_{f,i} - D \leq \Delta_{g,i} \leq M \cdot \Delta_{f,i} + D. \tag{20}$$

When $\max\left\{\sqrt{\frac{K\log(T-N)}{N}}, \sqrt{\frac{KN\log(N)}{(T-N)^2}}\right\} \leq \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}$, we set the stopping criteria as $\varepsilon = \min\left\{M \cdot \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D, \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}\right\}$. If $M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D \geq \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}$, we go back to the analysis in the basic model. So, we focus on the scenario where $M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D < \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/3}$. That is, $\varepsilon = M \cdot \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D$. Then, we divide arms $[K]$ into two groups: Group 1 with $\Delta_{g,i} < M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D$ and Group 2 with $\Delta_{g,i} \geq M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D$. Then

$$\begin{aligned}
(4) &\leq O\left(\sum\nolimits_{i\in Group1} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\varepsilon^2}, \frac{\log(N)}{(\Delta_{f,i})^2}\right\}\right. \\
&\quad \left. + \sum\nolimits_{i\in Group2} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{(\Delta_{g,i})^2}, \frac{\log(N)}{(\Delta_{f,i})^2}\right\} + (T-N)\cdot\varepsilon\right) \\
&= O\left(\sum\nolimits_{i\in Group1} \Delta_{f,i} \cdot \max\left\{\frac{\log(T-N)}{\left(M \cdot \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right)^2}, \frac{\log(N)}{(\Delta_{f,i})^2}\right\} + \sum\nolimits_{i\in Group2} \frac{\log(T-N)}{\Delta_{f,i}}\right) \\
&\quad + O\left((T-N)\cdot\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right)\right)
\end{aligned}$$

Due to (20), for arms in Group 1, we have $\Delta_{f,i} < \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + \frac{2D}{M}$. And, for arms in Group 2, we know that $\Delta_{f,i} \geq \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}$. We further divide arms Group 1 into two groups: Group 1a with $\Delta_{f,i} < \left(M \cdot \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right) \cdot \sqrt{\frac{\log(N)}{\log(T-N)}}$ and Group 1b with $\left(M \cdot \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right) \cdot \sqrt{\frac{\log(N)}{\log(T-N)}} \leq \Delta_{f,i} < \left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + \frac{2D}{M}$. Notice that $\sum_{i\in Group1a} \mathbb{E}[n_i(N)] \leq N$. Then the above,

$$\begin{aligned}
&\leq O\left(\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right)\cdot\sqrt{\frac{\log(N)}{\log(T-N)}}\cdot N\right. \\
&\quad \left. + K\cdot\left(\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + 2^{D}/_{M}\right)\cdot\frac{\log(T-N)}{\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2} + D\right)^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&+K\cdot\frac{\log(T-N)}{\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}}+(T-N)\cdot\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D\right)\Bigg)\\
\leq O&\left(K\cdot\left(\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+2^{D}/M\right)\cdot\frac{\log(T-N)}{\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D\right)^{2}}\right.\\
&\left.+K\cdot\frac{\log(T-N)}{\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}}+\left(T+\left(\sqrt{\frac{\log(N)}{\log(T-N)}}-1\right)N\right)\cdot\left(M\cdot\left(\frac{K\cdot\log(T-N)}{T-N}\right)^{1/2}+D\right)\right)\\
\leq\widetilde{O}&\left(\frac{K\cdot\left(\sqrt{\frac{K}{T-N}}+2^{D}/M\right)}{\left(M\cdot\sqrt{\frac{K}{T-N}}+D\right)^{2}}+T\cdot\left(M\cdot\sqrt{\frac{K}{T-N}}+D\right)\right). \qquad (21)
\end{aligned}
$$

We proved the proposition. □

C.2 Missing Algorithms and Proofs in Section 6.2

Algorithm 2 Reserved Online Stochastic Convex Optimization for Commitment (ROSCOC)

Input: A set of arms $\{1,2,\ldots,K\}$, N, T, τ;
Initialization: Set $\mathcal{H}_0=\emptyset$, $\widetilde{\Delta}_1=1/2$, $\mathcal{A}_{f,1}=[K]$.
Whenever N rounds are exhausted in Stage I or II, the algorithm enters the Commitment Stage.
for $t=1,\ldots,\tau$ **do** `/* Stage I: Reserved online stochastic convex optimization for g */`
    Implement *Online Stochastic Convex Optimization* algorithm for function g.
    Record the execution path $\mathcal{H}_t=\mathcal{H}_{t-1}\cup\{I_t\}$.
end for
for $\ell=1,2,\ldots$ **do** `/* Stage II: Arm eliminations for reward function f */`
    Sample each arm in $\mathcal{A}_{f,\ell}$ until the total number of times it has been chosen is $m_{f,\ell}$ times.
    At the end of epoch ℓ, for each arm $i\in[K]$, compute the empirical average reward $\hat{\mu}_{f,i,\ell}$ for reward function f.
    Update $\mathcal{A}_{f,\ell+1}\leftarrow\left\{i\in[K]:\max_{j\in\mathcal{A}_{f,\ell}}\hat{\mu}_{f,j,\ell}-\hat{\mu}_{f,i,\ell}\leq\widetilde{\Delta}_\ell\right\}$.
    Set $\widetilde{\Delta}_{\ell+1}\leftarrow\widetilde{\Delta}_\ell/2$.
end for
Uniformly sample $T-N$ elements from $\mathcal{H}_\tau$ with replacement. Denote the sample set as sequence $S=[I_1,\ldots,I_{T-N}]$. `/* Commitment Stage: Execution-history-induced portfolio */`
for $t=N+1,\ldots,T$ **do**
    Pull the $(t-N)-$th arm in S.
end for

In Stage I of the above algorithm, we utilize the algorithm *Online Stochastic Convex Optimization*, which was proposed by Agrawal and Devanur (2014), and we present it here.

Algorithm 3 Online Stochastic Convex Optimization

for $t = 1, \ldots, \tau$ **do**
  $I_t = \arg\max_{i\in[K]} g^*(x_t) - x_t^T \tilde{o}_{t,i}$. Here, $g^*(x) \equiv \max_y x^T y + g(y)$, is the Fenchel dual of g.
  Pull arm I_t and observe outcome o_{t,I_t}.
  /* Below $sign(x_t)$ is the vector that indicates the sign of each dimension of x_t, $n_{t,i}$ is the total number of pulls of arm i till time t. */
  Update $\tilde{o}_{t+1,i} = \bar{o}_{t,i} - sign(x_t) \cdot \sqrt{\frac{\ln(\tau)}{n_{t,i}}}$.
  Observe $h_t(x) = g^*(x) - x^T o_{t,I_t}$ and we update x_{t+1} by applying *Online Gradient Descent* (OGD).
end for

Theorem 6.1. *The regret of Algorithm ROSCOC has the form*

$$\mathbf{REG}_\pi[N,T] = \widetilde{O}\left(T \cdot \left(\sqrt{\frac{Kd}{N}} + L \cdot \sqrt{\frac{d}{N}}\right) + \sqrt{KN} + K^{1/3} d^{1/3} (T-N)^{2/3}\right), \tag{9}$$

where ROSCOC reserves $\tau = \min\left\{N, K^{1/3} d^{1/3} (T-N)^{2/3} \log(T-N)^{1/3}\right\}$ *periods to exploration for commitment.*

Proof for Theorem 6.1. Directly borrow the results from Agrawal and Devanur (2014), we have

$$\tau \cdot \left(\mathsf{OPT}_g - \mathbb{E}\left[g\left(\frac{\sum_{t=1}^{\tau} o_{t,I_t}}{\tau}\right)\right]\right) \le O\left(L\sqrt{d\tau} + \sqrt{Kd\tau \log(\tau)}\right). \tag{22}$$

On the other hand, due to Pinelis' inequality (an extension of Azuma's inequality in high dimensional spaces, see Pinelis (1994) and Alistarh et al. (2018)) (define $S_\tau = \sum_{t=1}^{\tau}(o_{t,I_t} - \mathbb{E}[o_{t,I_t}])$, then S_τ is a martingale with bounded differences, $\|o_{t,I_t} - \mathbb{E}[o_{t,I_t}]\| \le \sqrt{d}$), we have the following key observation:

$$\mathbb{P}\left(\left\|\frac{\sum_{t=1}^{\tau} o_{t,I_t}}{\tau} - \frac{\sum_{t=1}^{\tau} \mathbb{E}[o_{t,I_t}]}{\tau}\right\| \ge \varepsilon\right) \le 2\exp\left(\frac{-\tau \cdot \varepsilon^2}{2d}\right).$$

This, together with g's Lipschitz condition, will help bind the regret in the committing phase. Choosing $\varepsilon = \sqrt{\frac{d\log(\tau d)}{\tau}}$, then

$$\begin{aligned}\mathbb{E}\left[g\left(\frac{\sum_{t=1}^{\tau} o_{t,I_t}}{\tau}\right)\right] - g\left(\frac{\sum_{t=1}^{\tau} \mathbb{E}[o_{t,I_t}]}{\tau}\right) &\le L \cdot \left(\sqrt{\frac{d\log(\tau d)}{\tau}} + \sqrt{d} \cdot 2\exp\left(-\frac{\log(\tau d)}{2}\right)\right) \\ &= O\left(L \cdot \sqrt{\frac{d\log(\tau d)}{\tau}}\right).\end{aligned} \tag{23}$$

In the committing phase, the policy is nonadaptive. Due to the same argument, we have

$$g\left(\frac{\sum_{t=N+1}^{T} \mathbb{E}[o_{t,I_t}]}{T-N}\right) - \mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T} o_{t,I_t}}{T-N}\right)\right] \le O\left(L \cdot \sqrt{\frac{d\log((T-N)d)}{T-N}}\right). \tag{24}$$

Now, we are ready to bound the regret. In the experiment phase, Step 1's gradient descent would incur at most τ amount of regret. Step 2's arm elimination would incur the conventional regret of $\sqrt{K(N-\tau)\log(N-\tau)}$. And finally, plus the regret of the committing phase. Then,

$\mathbf{REG}[N,T]$

$$
\begin{aligned}
&\leq O\left(\tau+\sqrt{K(N-\tau)\log(N-\tau)}\right)+(T-N)\cdot\left(\mathsf{OPT}_g-\mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T}o_{t,I_t}}{T-N}\right)\right]\right)\\
&=O\left(\tau+\sqrt{K(N-\tau)\log(N-\tau)}\right)\\
&\quad+(T-N)\cdot\left(\mathsf{OPT}_g-\mathbb{E}\left[g\left(\frac{\sum_{t=1}^{\tau}o_{t,I_t}}{\tau}\right)\right]+\mathbb{E}\left[g\left(\frac{\sum_{t=1}^{\tau}o_{t,I_t}}{\tau}\right)\right]-g\left(\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}\right)\right.\\
&\quad\left.+g\left(\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}\right)-\mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T}\mathbb{E}[o_{t,I_t}]}{T-N}\right)\right]+\mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T}\mathbb{E}[o_{t,I_t}]}{T-N}\right)-\mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T}o_{t,I_t}}{T-N}\right)\right]\right]\right)\\
&\leq O\left(\tau+\sqrt{K(N-\tau)\log(N-\tau)}\right)\\
&\quad+(T-N)\cdot O\left(L\sqrt{\frac{d}{\tau}}+\sqrt{\frac{Kd\log(\tau)}{\tau}}+L\cdot\sqrt{\frac{d\log(\tau d)}{\tau}}+L\cdot\sqrt{\frac{d\log((T-N)d)}{T-N}}\right), \qquad (25)
\end{aligned}
$$

where the last inequality utilizes (22), (22), (24), and the fact that

$$
\begin{aligned}
g\left(\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}\right)-\mathbb{E}\left[g\left(\frac{\sum_{t=N+1}^{T}\mathbb{E}[o_{t,I_t}]}{T-N}\right)\right]&=\mathbb{E}\left[g\left(\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}\right)-g\left(\frac{\sum_{t=N+1}^{T}\mathbb{E}[o_{t,I_t}]}{T-N}\right)\right]\\
&\leq L\cdot\mathbb{E}\left[\left\|\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}-\frac{\sum_{t=N+1}^{T}\mathbb{E}[o_{t,I_t}]}{T-N}\right\|\right]\\
&\overset{(a)}{\leq} L\cdot\mathbb{E}\left[\sqrt{\frac{d\log((T-N)d)}{T-N}}\right],
\end{aligned}
$$

where inequality (a) is due to the fact that $\{\mathbb{E}[o_{t,I_t}]\}_{t=N+1,\ldots,T}$ are iid random variables due to the uniform sampling step in the algorithm. And, each of them has the same mean value $\frac{\sum_{t=1}^{\tau}\mathbb{E}[o_{t,I_t}]}{\tau}$. Applying Pinelis's inequality again, we get the above result.

Setting $\tau=\min\left\{N,K^{1/3}d^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right\}$, when $N=o\left(K^{1/3}d^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right)$, we have

$$
(25)\leq\widetilde{O}\left(N+T\cdot\left(\sqrt{\frac{Kd}{N}}+L\cdot\sqrt{\frac{d}{N}}\right)\right)=\widetilde{O}\left(T\cdot\left(\sqrt{\frac{Kd}{N}}+L\cdot\sqrt{\frac{d}{N}}\right)\right).
$$

When $N\geq\Omega\left(K^{1/3}d^{1/3}(T-N)^{2/3}\log(T-N)^{1/3}\right)$, we have

$$
(25)\leq\widetilde{O}\left(\sqrt{KN}+K^{1/3}d^{1/3}(T-N)^{2/3}\right).
$$

In combination, we can write the minimax regret as follows,

$$
\widetilde{O}\left(T\cdot\left(\sqrt{\frac{Kd}{N}}+L\cdot\sqrt{\frac{d}{N}}\right)+\sqrt{KN}+K^{1/3}d^{1/3}(T-N)^{2/3}\right)
$$

When L and d are constants, the regret upper bound has form

$$
\widetilde{O}\left(\sqrt{\frac{K(T-N)^{2/3}\cdot\max\left\{(T-N)^{4/3},K^{1/3}N\right\}}{\min\left\{N,K^{1/3}(T-N)^{2/3}\right\}}}\right),
$$

which is the same as the minimax bound in the basic model. Recall that our basic model is a special case of (7), so the above regret result is tight. □