Adaptive Combinatorial Experimental Design: Pareto Optimality for Decision-Making and Inference

Hongrui Xie
University of Science and Technology of China

Junyu Cao
The University of Texas at Austin

Kan Xu
Arizona State University

Abstract

In this paper, we provide the first investigation into adaptive combinatorial experimental design, focusing on the trade-off between regret minimization and statistical power in combinatorial multi-armed bandits (CMAB). While minimizing regret requires repeated exploitation of high-reward arms, accurate inference on reward gaps requires sufficient exploration of suboptimal actions. We formalize this trade-off through the concept of Pareto optimality and establish equivalent conditions for Pareto-efficient learning in CMAB. We consider two relevant cases under different information structures, i.e., full-bandit feedback and semi-bandit feedback, and propose two algorithms `MixCombKL` and `MixCombUCB` respectively for these two cases. We provide theoretical guarantees showing that both algorithms are Pareto optimal, achieving finite-time guarantees on both regret and estimation error of arm gaps. Our results further reveal that richer feedback significantly tightens the attainable Pareto frontier, with the primary gains arising from improved estimation accuracy under our proposed methods. Taken together, these findings establish a principled framework for adaptive combinatorial experimentation in multi-objective decision-making.

1 INTRODUCTION

Combinatorial multi-armed bandits (CMAB) generalize the classical multi-armed bandit (MAB) framework to settings where the learner selects a structured combination of basic actions, referred to as a "super arm," at each round. This setup models many real-world problems, such as online advertising, sensor selection, network routing, and recommendation systems (Cesa-Bianchi and Lugosi, 2012; Chen et al., 2013), where multiple actions are taken jointly and rewards depend on the combination, and many of them require accurate estimation of reward gaps between combinatorial arms as well as regret minimization. For example, empirical evidence from a major video-sharing platform shows that joint treatment effects can be strongly non-additive (Ye et al., 2023). This setting is inherently combinatorial: at each round the platform selects a set of interventions (a super-arm) and observes only session-level outcomes (e.g., total watch time), aligning with the regret–inference trade-offs studied in our combinatorial bandit models. In such combinatorial problems, this tension is exacerbated by the large action space and dependencies between arms. The combinatorial nature of the action space increases the complexity of both exploration and optimization. We investigate the problem of deriving Pareto optimal policies for combinatorial bandits, where the learner must jointly minimize regret and accurately estimate reward gaps between combinatorial arms. A policy is Pareto optimal if no alternative policy can be better off in both regret and estimation error. Our goal is to identify a class of Pareto optimal algorithms that strike proper balance between regret control and inference accuracy.

To the best of our knowledge, our work provides the first systematic study of Pareto optimality in the context of combinatorial bandit feedback. We summarize our main contributions as follows:

1. **Pareto-Optimal Algorithms for Combinatorial Bandits.** We develop two Pareto optimal algorithms for two different combinatorial bandit settings – i.e., `MixCombKL` for full-bandit feedback and `MixCombUCB` for semi-bandit feedback. Our algorithms dynamically calibrate exploration to

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

preserve both estimation accuracy and regret performance to maintain Pareto optimality, regardless of the complexity of combinatorial action spaces.

2. **Theoretical Guarantees under Different Bandit Feedback Models.** We derive finite-sample regret bounds and estimation error guarantees for both bandit settings, and establish their Pareto optimality. Our analysis shows that semi-bandit feedback yields a sharper Pareto frontier than full-bandit feedback, with the improvement arising from reduced estimation error. Meanwhile, the regret under the two feedback regimes remains of the same order when using proposed algorithms.

2 RELATED LITERATURE

Our paper is related to the stochastic multi-armed bandit (MAB) literature of regret minimization and best-arm identification (BAI). Regret minimization seeks to reduce cumulative loss (Slivkins, 2019; Kuleshov and Precup, 2014), while BAI targets high-confidence identification of the optimal arm under fixed-budget or fixed-confidence settings (Audibert and Bubeck, 2010; Garivier and Kaufmann, 2016). The combinatorial multi-armed bandit (CMAB) problem generalizes the classical bandit setting by allowing the learner to select a subset of base arms at each round. The literature distinguishes between *semi-bandit* feedback, revealing individual rewards for basic actions (Kveton et al., 2015; Wang and Chen, 2018), and *full-bandit* feedback, revealing only total reward for super arms (Cesa-Bianchi and Lugosi, 2012; Combes et al., 2015). CMAB variants include contextual (Qin et al., 2014; Chen et al., 2018; Li et al., 2016), knapsack (Sankararaman and Slivkins, 2018), fairness (Li et al., 2019), relative-feedback settings (Saha and Gopalan, 2019) and pure exploration (Chen et al., 2014).

The goals of regret minimization and arms inference are inherently misaligned: better inference requires more exploration, and thus may increase regret. Prior work focuses primarily on classical K-arm bandits. Chen et al. (2017); Kaufmann et al. (2016); Degenne et al. (2019) study this trade-off, showing no single algorithm achieves optimal rates for both simultaneously. Similarily, this tradeoff is further studied in Faruk et al. (2025) for linear contextual bandit settings. Pareto frontiers for MAB problem were first developed in Zhong et al. (2023). Simchi-Levi and Wang (2025) then prove a sufficient and necessary condition of Pareto optimality and propose a Pareto optimal algorithm. Zuo and Qin (2025) extend this notion to multinomial logistic bandit problems and Zhang and Wang (2025) extended this framework to network inference problems. However, to the best of our knowledge, the trade-off between regret minimization and arms inference remains unexplored for CMAB problems.

Inspired from the Pareto optimality on classical MAB framework of Simchi-Levi and Wang (2025), our work characterized the regret-inference tradeoff through analyzing the Pareto optimality of CMAB problems. Applying the existing methods is computationally infeasible due to the exponential super-arm space in CMAB settings. Therefore, we design two different Pareto-optimal algorithms for bandit problems with different information structures, i.e., full-bandit and semi-bandit settings. Additionally, we show how feedback richness governs trade-offs, comparing the two regimes. We further overcome the challenge of CMAB problems on the concentration of information in super arm subspace, completing a unified Pareto-optimal framework.

3 PROBLEM FORMULATION

3.1 Models

We consider the stochastic CMAB problem and define a bandit instance as a tuple $(\mathcal{A}, \mathcal{M}, \nu)$. In particular, $\mathcal{A} = \{1, 2, \ldots, d\}$ is the set of base arms, $\mathcal{M} \subseteq 2^{\mathcal{A}}$ is the set of super arms that contains the feasible subsets of $\mathcal{A}$, and ν is the distribution of the reward vector w_t where $\mathbb{E}_{w_t \sim \nu}[w_t] = \boldsymbol{\mu} = [\mu(1), \ldots, \mu(d)]^\top \in [0,1]^d$. $\mathcal{V}_0$ is the set of all admissible CMAB instances ν. The basic actions in $\mathcal{A}$ are associated with a random reward vector w_t at time t, drawn i.i.d. from a distribution ν over $[0,1]^d$, i.e., $w_t \sim \nu$. For each base arm $e \in \mathcal{A}$, its reward is the e-th coordinate denoted $w_t(e) \in [0,1]$. Define the mapping $f : 2^{\mathcal{A}} \times \mathbb{R}^d \to \mathbb{R}, f(G, \varpi) = \sum_{e \in G} \varpi(e)$ as the total weight of the elements in an arbitrary set $G \subseteq \mathcal{A}$ under d-dimensional vector ϖ with entries $\varpi(e)$; then, $f(M(t), w_t) = \sum_{e \in M(t)} w_t(e)$ denotes the reward at time t for playing super arm $M(t)$.

The time horizon is set to be n. At each round $t \leq n$, the decision maker observes the history $\mathcal{H}_t = (M(1); w_1, \ldots, M(t); w_t)$, and selects a super arm $M(t) \subseteq \mathcal{M}$. An admissible policy $\pi = \{\pi_t\}_{t \geq 1}$ maps the history $\mathcal{H}_{t-1}$ to a super arm $M^\pi(t) \in \mathcal{M}$, where $\pi_t(M) = \mathbb{P}(M(t) = M \mid \mathcal{H}_{t-1})$ denotes the probability of selecting M at time t. We define $M^* := \arg\max_{M \in \mathcal{M}} \sum_{e \in M} \mu(e)$ to be the arm with the maximum true reward, and the performance of a policy π is evaluated via cumulative *regret*, which measures the expected loss relative to the optimal super arm M^*: $\mathcal{R}_\nu(n, \pi) = \sum_{t=1}^n [\mathbb{E}[f(M^*, w_t)] - \mathbb{E}^\pi [f(M^\pi(t), w_t)]]$.

We also want to quantify estimation errors for both base arms and super arms for any algorithm. We define the gap between any two super arms $M(\tau_i)$ and $M(\tau_j)$ as $\Delta_M^{(i,j)} := f(M(\tau_i), \boldsymbol{\mu}) - f(M(\tau_j), \boldsymbol{\mu}), \forall i \neq j \in [|\mathcal{M}|]$; here τ_i, τ_j are indexes of super arms. In particular,

the gap to the optimal super arm M^* for every M is $\Delta_M = f(M^*, \boldsymbol{\mu}) - f(M, \boldsymbol{\mu})$. We also define the gap between base arms as $\Delta_\mu^{(i,j)} := \mu(i) - \mu(j)$ for any $i \neq j \in [d]$. In addition, we assume each super arm satisfies the constraint $|M| = m$ for all $M \in \mathcal{M}$. For any algorithm that outputs estimation for bandit features, an algorithm-specific admissible adaptive estimator $\hat{\Delta}^{(i,j)} = \{\hat{\Delta}_t^{(i,j)}\}_{t\geq 1}$ maps $\mathcal{H}_t$ to an estimate of the pairwise gap $\Delta^{(i,j)}$ at time t in our stochastic CMAB setting, where $\Delta^{(i,j)}$ can be either $\Delta_M^{(i,j)}$ or $\Delta_\mu^{(i,j)}$. The estimation quality is quantified by the expected distance of $\Delta^{(i,j)}$ and $\hat{\Delta}^{(i,j)}$ (i.e., $\mathcal{E}(t, \hat{\Delta}^{(i,j)}) = \mathbb{E}[|\Delta^{(i,j)} - \hat{\Delta}_t^{(i,j)}|]$ for $\Delta_t^{(i,j)} = \Delta_{M,t}^{(i,j)}$ or $\Delta_{\mu,t}^{(i,j)}$), which is treated as the *estimation error*.

The richness of the bandit feedback influences the regret and estimation error and thus the Pareto frontier. To that end, we consider the two bandit feedback regimes upon selecting super arm $M(t)$ at time t: (i) *full-bandit feedback*, where only the aggregate reward $f(M(t), w_t)$ is revealed with no information about the contributions of individual basic actions, and (ii) *semi-bandit feedback*, where the individual component reward $w_t(e)$ is observed for $e \in M$ (when the respective basic action is chosen) and provides rich item-level information.

3.2 Pareto Optimality

To balance regret and estimation accuracy in combinatorial bandits, we begin by formalizing the notion of Pareto optimality. Because combinatorial feedback generally prevents estimation of all basic actions in $\mathcal{A}$ (see Appendix G for an example of such a restricted bandit structure), we define $\mathcal{A}_{ad}$ as the subset of actions in $\mathcal{A}$ that can be reliably estimated.

Consider a policy $(\pi, \widehat{\Delta})$, where π denotes the learning algorithm and $\widehat{\Delta}$ is an estimator of the reward gap between combinatorial super arms. The policy is said to be *Pareto optimal* if there is no alternative admissible policy $(\pi', \widehat{\Delta}')$ that performs at least as well in both cumulative regret and estimation error, and strictly better in at least one of them.

Definition 3.1 (Pareto Optimality)**.** $(\pi, \widehat{\Delta})$ *is Pareto optimal if there does not exist* $(\pi', \widehat{\Delta}')$ *such that:*

$$\mathcal{R}_\nu(n, \pi') \preceq \mathcal{R}_\nu(n, \pi), \max_{i\leq j} \mathcal{E}\left(\widehat{\Delta}'^{(i,j)}\right) \preceq \max_{i\leq j} \mathcal{E}\left(\widehat{\Delta}^{(i,j)}\right)$$

with at least one inequality [1] *being strict.*

[1] In our paper, for any two positive functions $f(n)$ and $g(n)$, we write $f(n) \preceq g(n)$ if $\frac{f(n)}{g(n)}$ is bounded by a positive constant for all n; specifically, the inequality holds true if and only if there exist constants $C_1, C_2 > 0$ such that $C_1 \leq \frac{f(n)}{g(n)} \leq C_2$ for all n.

Here we denote $\mathcal{R}_\nu(n, \pi)$ as the cumulative regret under the reward distribution ν. The estimation error Δ in the combinatorial bandit setting will refer either to a base-arm gap, defined as $\Delta_\mu^{(i,j)} = \mu(i) - \mu(j); \forall i \neq j \in [|\mathcal{A}_{ad}|]$, or to a super-arm gap, defined as $\Delta_M^{(i,j)} := f\big(M(\tau_i), \boldsymbol{\mu}\big) - f\big(M(\tau_j), \boldsymbol{\mu}\big), \forall i \neq j \in [|\mathcal{M}|]$.

The choice of ignoring constant or logarithmic factors is standard in bandit asymptotics as bandit asymptotics is about identifying the fundamental scaling of regret with problem parameters, and proof-dependent artifacts would not change the underlying rate of optimality. Similar factor choice is also seen in Simchi-Levi and Wang (2025).

We further define the Pareto frontier $\mathcal{P}_f$ to comprise all policies that are not strictly dominated in both regret and estimation accuracy:

Definition 3.2 (Pareto Frontier)**.** *The Pareto Frontier* $\mathcal{P}_f$ *is defined as*

$$\mathcal{P}_f = \Bigg\{ (\pi, \widehat{\Delta}) \;\Bigg|\; \nexists (\pi', \widehat{\Delta}') \text{ s.t. } \mathcal{R}_\nu(n, \pi') \preceq \mathcal{R}_\nu(n, \pi), \max_{i\leq j} \mathcal{E}\Big(\widehat{\Delta}'^{(i,j)}\Big) \preceq \max_{i\leq j} \mathcal{E}\Big(\widehat{\Delta}^{(i,j)}\Big) \Bigg\}.$$

Policies in $\mathcal{P}_f$ represent efficient trade-offs — any policy outside this set can be improved upon in at least one dimension without sacrificing the other. The search for Pareto optimal solutions can be cast as a multi-objective optimization problem:

$$\min_{(\pi, \widehat{\Delta})} \max_{\nu \in \mathcal{V}_0} \left(\mathcal{R}_\nu(n, \pi), \max_{i\leq j} \mathcal{E}\left(\widehat{\Delta}^{(i,j)}\right) \right).$$

This formulation captures the dual learning objective of minimizing regret while maintaining accurate reward gap estimations under worst-case scenarios.

4 ALGORITHMS AND THEORY

Because of the differences in information structure, our algorithmic design is tailored to the feedback model. In the full-bandit setting, the exponential size of the super-arm space renders classical UCB-style confidence construction impractical. Traditional UCB methods rely on per-arm play counts to form confidence intervals, but in this setting the rewards of individual base arms are not directly observed. Constructing valid confidence bounds requires projecting super-arm rewards into a high-dimensional linear space, which is statistically costly. KL-divergence-based methods bypass this issue by working on the probability simplex over super arms and using divergence constraints to guide exploration, enabling regret analysis without enumerating exponentially many actions. In contrast,

the semi-bandit feedback setting reveals feedback for each base arm, which makes per-action confidence intervals easy to compute and super-arm bounds derivable with only a single call to the optimization oracle. Here, applying KL-divergence methods adds computational overhead without improving statistical performance, while UCB-style algorithms already provide distribution-independent guarantees. Hence, our algorithm design adopts KL-divergence-based methods for full-bandit feedback and UCB-based approaches for semi-bandit, aligning algorithm design with feedback richness and computational feasibility. We explicitly analyze the computational efficiency of our algorithms in Appendix B.

4.1 Full-Bandit Feedback

Before introducing our algorithm, we first define some additional notation. We vectorize the super arm $M \in \mathcal{M}$ by setting $\boldsymbol{\theta}_M = [\mathbb{I}\{1 \in M\}, ..., \mathbb{I}\{d \in M\}]^\top$, where $|M| = m$ implies that $\|\boldsymbol{\theta}_M\|_1 = m$. The vectorization trick helps us construct matrix-based super-arm features for computation. Notice that $\max_{M\in\mathcal{M}} \boldsymbol{\theta}_M^\top X = \max_{\rho\in \mathrm{Co}(\boldsymbol{\theta})} \rho^\top X$, where $\mathrm{Co}(\boldsymbol{\theta})$ denotes the convex hull of $\{\boldsymbol{\theta}_M : M \in \mathcal{M}\}$. Next, by dividing each vector in the vectorized set $\mathcal{M}$ by m, we embed it into the d-dimensional simplex and define $\mathcal{Q}$ as the corresponding scaled convex set. Here and throughout, the notation $\tilde{\mathcal{O}}(\cdot)$ hides factors polylogarithmic in the time horizon T and the dimension parameters m, d.

Building on the Online Stochastic Mirror Descent (OSMD) (Audibert et al., 2014) framework, we propose the `MixCombKL` algorithm (Algorithm 1), which uses the Kullback–Leibler (KL) divergence as the Bregman divergence for projection onto $\mathcal{Q}$. The KL divergence between two distributions $p, q \in \mathcal{Q}$ is defined as $\mathrm{KL}(p, q) = \sum_{i=1}^d p(i) \log \frac{p(i)}{q(i)}$. The projection of q onto a closed convex set Ξ of distributions is given by $p^* = \arg\min_{p\in\Xi} \mathrm{KL}(p, q)$.

We set parameters $C = \lambda_{\min} m^{-\frac{3}{2}}, \gamma = [\sqrt{m \log \rho_{\min}^{-1}} + \sqrt{C\big(Cm^2 d + m\big)n}]^{-1} \sqrt{m \log \rho_{\min}^{-1}}, \eta = \gamma C$, and we define the problem-dependent constants $\lambda_{\min}$ and $\rho_{\min}$ as follows: Let $\lambda_{\min}$ be the smallest nonzero eigenvalue of $\mathbb{E}[\boldsymbol{\theta}_M \boldsymbol{\theta}_M^\top]$, where M is uniformly distributed over $\mathcal{M}$. We define the exploration-inducing distribution $\rho^0 \in \mathcal{P} : \rho_e^0 = \frac{1}{m|\mathcal{M}|} \sum_{M\in\mathcal{M}} \mathbb{I}(e \in M)$, $\forall e \in \mathcal{A}$, and let $\rho_{\min} = \min_{e\in\mathcal{A}} m\rho_e^0$; ρ^0 is the distribution over basic actions $\mathcal{A}$ induced by the uniform distribution over $\mathcal{M}$. We view the *estimable basic actions* under *full-combinatorial feedback* as special super arms M with $|M| = 1$, so we can define the set of basic actions covered by the span of all $\boldsymbol{\theta}_M$ (written as $\mathrm{span}(\boldsymbol{\theta})$) as $\mathcal{M}_{KL} = \{M \in \mathcal{A} : \boldsymbol{\theta}_M \in \mathrm{span}(\boldsymbol{\theta})\}$.

The `MixCombKL` algorithm proceeds as follows. First, at each round t, compute a probability distribution p_{t-1} over the super arms by decomposing it onto the predetermined distribution q'_{t-1}, mixed through $q'_{t-1} = (1-\gamma)q_{t-1} + \gamma\rho^0$.

Algorithm 1: Mixture-Based Combinatorial KL-divergence Algorithm (`MixCombKL`)

Input: $\alpha \in [0, \frac{1}{2}], \mathcal{M}, n$;
Initialization: Set $q_0 = \rho^0$, $R_0(M(\tau_k)) = 0, \forall k \in [|\mathcal{M}|], R_0(M(l)) = 0, \forall l \in [|\mathcal{M}_{KL}|]$;
for $t = 1, ..., n$ **do**
  Let $q'_{t-1} = (1-\gamma)q_{t-1} + \gamma\rho^0$;
  Decompose distribution p_{t-1} over $\mathcal{M}$ so that $\sum_{M\in\mathcal{M}} p_{t-1}(M)\boldsymbol{\theta}_M = mq'_{t-1}$;
  Set random variable U_t where $\mathbb{P}(U_t = 0) = 1 - \frac{1}{2t^\alpha}$ and $\mathbb{P}(U_t = 1) = \frac{1}{2t^\alpha}$;
  Select arm $M(t)$ with distribution $\forall M \in \mathcal{M}$, $\pi_t(M) = p_{t-1}(M)\mathbb{I}\{U_t = 0\} + \mathbb{I}\{U_t = 1\}/|\mathcal{M}|$;
  Observe reward $f(M(t), w_t)$;
  Let $\Sigma_{t-1} = \mathbb{E}[\boldsymbol{\theta}_M \boldsymbol{\theta}_M^\top]$, where M has law p_{t-1};
  Set $\tilde{w}_t(e) = f(M(t), w_t)\Sigma_{t-1}^+ \boldsymbol{\theta}_M$, where Σ_{t-1}^+ is the pseudo-inverse of matrix Σ_{t-1};
  Set $\tilde{q}_t(e) = \frac{q_{t-1}(i)\exp(\eta\tilde{w}_t(e))}{\sum_{j=1}^d q_{t-1}(j)\exp(\eta\tilde{w}_t(e))}$, $\forall e \in \mathcal{A}$;
  Compute $q_t = \mathbb{I}\{U_t = 1\}q_{t-1} + \mathbb{I}\{U_t = 0\} \arg\min_{p\in\mathcal{Q}} \sum_{e\in\mathcal{A}} p(e) \log \frac{p(e)}{\tilde{q}_t(e)}$;
  Set $R_t(M(\tau_k)) = 2t^\alpha \mathbb{I}\{U_t = 1\}\boldsymbol{\theta}_{M(\tau_k)}^\top \tilde{w}_t + R_{t-1}(M(\tau_k)), \forall k \in [|\mathcal{M}|]$;
  Set $R_t(M(l)) = 2t^\alpha \mathbb{I}\{U_t = 1\}\boldsymbol{\theta}_{M(l)}^\top \tilde{w}_t + R_{t-1}(M(l)), \forall k \in [|\mathcal{M}_{KL}|]$;

Output: $\hat{\Delta}_{M,n}^{(i,j)} = \frac{1}{n}(R_n(M(\tau_i)) - R_n(M(\tau_j))$, for $i, j \in [|\mathcal{M}|]$ and $i \neq j$;
$\hat{\Delta}_{\mu,n}^{(i,j)} = \frac{1}{n}(R_n(M(i)) - R_n(M(j))$, for $i, j \in [|\mathcal{M}_{KL}|]$ and $i \neq j$;

The KL projection in `MixCombKL` ensures that $mq_{t-1} \in \mathrm{Co}(\boldsymbol{\theta})$, and there exists ψ, a distribution over $\mathcal{M}$, such that $mq_{t-1} = \sum_M \psi(M)\boldsymbol{\theta}_M$. This guarantees that the system of linear equations in the decomposition step is consistent. The algorithm of Sherali (1987) also demonstrates that the decomposition step can be efficiently implemented.

Second, to ensure the balance between exploration for inference and regret minimization, we set a random variable U_t independently of previous history and the reward distribution, with probability $\mathbb{P}(U_t = 0) = 1 - \frac{1}{2t^\alpha}$ and $\mathbb{P}(U_t = 1) = \frac{1}{2t^\alpha}$. The pre-set parameter $\alpha \in [0, \frac{1}{2}]$ quantifies the decay of exploration. We select

a distribution: for all $M \in \mathcal{M}$,

$$\pi_t(M) = p_{t-1}(M)\mathbb{I}\{U_t = 0\} + |\mathcal{M}|^{-1}\mathbb{I}\{U_t = 1\},$$

and sample a super arm according to p_{t-1}. When choosing $\alpha = 0$, the algorithm has equal probability at each time of choosing a uniform distribution or the KL-divergence-based distribution, and as α grows the algorithm tends to focus more on regret. Thus, we can observe the full combinatorial feedback for the chosen super arm. Third, we update the empirical estimates of the reward parameters based on the observed feedback and compute an offline approximation oracle

$$q_t = \mathbb{I}_{\{U_t=1\}} q_{t-1} + \mathbb{I}_{\{U_t=0\}} \arg\min_{p \in \mathcal{Q}} \sum_{e \in \alpha} p(e) \log \frac{p(e)}{\tilde{q}_t(e)}$$

in order to find our pre-determined distribution q_t. We also compute R_t for future estimation of arms. Finally, we repeat the above steps for n rounds, dynamically refining the estimates and adjusting the distribution p_t to approach the Pareto-optimal trade-off between regret and estimation accuracy, and the estimation of basic actions and super arms is presented at the end.

Our algorithm departs from Combes et al. (2015) in that it improves inference accuracy by using a mixture of distributions as the super-arm sampling distribution, whereas prior work selects super arms solely based on distributions derived from KL-divergence properties.

Now, the following theorem provides an upper bound for the estimation error of super-arm gaps of `MixCombKL` through constructing a series of martingales.

Theorem 4.1. *With probability at least $1-\delta$, it holds*

$$|\hat{\Delta}_{M,n}^{(i,j)} - \Delta_M^{(i,j)}| \leq \frac{6}{\lambda_{\min}} \sqrt{\frac{m^3 d}{n^{1-\alpha}}} \ln(\frac{2d}{\delta})$$

where $\Delta_M^{(i,j)} = \mathbb{E}[\hat{\Delta}_{M,n}^{(i,j)}]$ for any $i, j \in [|\mathcal{M}|]$ and $i \neq j$.

Note that, when taking $\delta = \frac{1}{n^2}$, we have $\max_{i<j\leq|\mathcal{M}|} \mathcal{E}(n, \hat{\Delta}_{M,n}^{(i,j)}) = \tilde{\mathcal{O}}(\sqrt{n^{\alpha-1}})$. Similarly, we can derive an upper bound for the inference of $\Delta_\mu^{(i,j)}$.

Corollary 4.2. *With probability at least $1-\delta$, it holds*

$$|\hat{\Delta}_{\mu,n}^{(i,j)} - \Delta_\mu^{(i,j)}| \leq \frac{6m}{\lambda_{\min}} \sqrt{\frac{d}{n^{1-\alpha}}} \ln(\frac{2d}{\delta}),$$

where $\Delta_\mu^{(i,j)} = \mathbb{E}[\hat{\Delta}_{\mu,n}^{(i,j)}]$ for any $i, j \in [|\mathcal{M}_{KL}|]$ and $i \neq j$.

Similarly, we have $\max_{i<j\leq|\mathcal{M}_{KL}|} \mathcal{E}(n, \hat{\Delta}_{\mu,n}^{(i,j)}) = \tilde{\mathcal{O}}(\sqrt{n^{\alpha-1}})$ by taking $\delta = \frac{1}{n^2}$.

We now turn to establishing the regret upper bound. Our algorithm introduces a sequence of independent Bernoulli random variables U_t, with $\mathbb{P}(U_t = 1) = \frac{1}{2t^\alpha}$. When $U_t = 1$, the algorithm performs uniform exploration over all super arms, ensuring equal sampling. This uniform exploration step contributes an additional $\mathcal{O}(n^{1-\alpha})$ term to the regret bound originally derived in Combes et al. (2015).

For the case when $U_t = 0$, we establish a lemma of "triangle inequality" type, analogous to Lemma 4 in Combes et al. (2015), which relates the KL divergence to the difference in expected reward between the optimal super arm and the arm selected by `MixCombKL`. This lemma allows us to effectively bound the regret and leads to the final regret guarantee for `MixCombKL`.

Theorem 4.3. *The regret satisfies*

$$\mathcal{R}_\nu(n, \pi) \leq 2\sqrt{m^3 n \left(d + \frac{m^{1/2}}{\lambda_{\min}}\right) \log \rho_{\min}^{-1}} + \frac{mn^{1-\alpha}}{2(1-\alpha)} + \frac{m^{5/2}}{\lambda_{\min}} \log \rho_{\min}^{-1}.$$

For most classes of $\mathcal{M}$, we can observe that $\rho_{\min}^{-1} = \mathcal{O}(\text{poly}(d))$ and $m(d\lambda_{\min})^{-1} = \mathcal{O}(1)$ (Csiszar and Shields, 2004). In these cases, we can derive that our `MixCombKL` algorithm has a regret bound of $\mathcal{O}(\sqrt{m^3 dn \log(d/m)} + mn^{1-\alpha})$. This result indicates the effectiveness of our algorithm in tackling the intrinsic combinatorial nature of the full-bandit problem.

4.2 Semi-Bandit Feedback

For the semi-bandit setting, we relax the constraint on M to $|M| \leq m$. The `MixCombUCB` algorithm (Algorithm 3) starts with the initialization procedure `InitUCB` (Algorithm 2), which returns two variables. The first is a weight vector $\hat{w} \in [0,1]^d$, where $\hat{w}(e)$ is a single observation from the e-th marginal distribution ν. The second variable is the initialization step count plus one. `InitUCB` also outputs a set E of observed basic actions and m_0 pairs of corresponding super arms for every $e \in E$ as (e, M_e).

To compute $\hat{w}$, `InitUCB` repeatedly calls the optimization oracle $M(t) = \arg\max_{M\in\mathcal{M}} \sum_{e\in M} u(e)$ on an auxiliary binary weight vector $u \in \{0,1\}^d$ whose entries are initially set to all ones. When an item e is observed, $\hat{w}(e)$ is set to the observed reward from arm e, while the respective $u(e)$ is set to zero, and a corresponding super arm M_e is recorded for future observations. The procedure terminates once all entries of u are zero. Since at least one coordinate of u switches from one to zero in each iteration, `InitUCB` terminates in at most d steps. The number of initialization steps, denoted by m_0, corresponds to the number of basic arms included in $\mathcal{M}$, which we denote by $\mathcal{M}_{UCB}$—the *estimable basic actions* under the *semi-bandit feedback* regime.

At each time step t, `MixCombUCB` proceeds in four phases. First, it computes the UCB for the expected weight of each item e: $U_t(e) = \hat{w}_{T_{t-1}(e)}(e) + c_{t-1,T_{t-1}(e)}$, where $\hat{w}_s(e)$ is the average of s observed weights for item e, $T_t(e)$ is the number of times item e has been observed up to time t, and $c_{t,s} = \sqrt{\frac{2\log t}{s}}$ is the confidence radius such that $\mu(e) \in [\hat{w}_s(e) - c_{t,s}, \hat{w}_s(e) + c_{t,s}]$ holds with high probability.

Algorithm 2: `InitUCB`: `MixCombUCB` Initialization

Input: $\mathcal{M}$;
$\hat{\omega}(e) = 0, u(e) = 1, M_e = \emptyset,\ \forall e \in \mathcal{A}; E = \emptyset, t = 1$;
while $\exists e \in \mathcal{A} : u(e) = 1$ **do**
  $M(t) = \arg\max_{M\in\mathcal{M}} \sum_{e\in M} u(e)$;
  Observe $\{(e, w_t(e)) : e \in M(t)\}$ where $w_t \sim \nu$;
  forall $e \in M(t)$ **do**
    $\hat{\omega}(e) = \omega_t(e), u(e) = 0, E = E \cup \{e\}$;
    $M_e = M(t), t = t + 1$;
Output: $E, \hat{\omega}, t, \forall e \in E, (e, M_e)$;

Second, `MixCombUCB` queries the optimization oracle to solve the combinatorial maximization problem over the UCB super arm $\tilde{M}(t) = \arg\max_{M\in\mathcal{M}} f(M, U_t)$.

Third, we set $\alpha_t = \frac{1}{m_0 t^\alpha}$, then construct a probability distribution over super arms: for all $M \in \mathcal{M}$,

$$\pi_t(M) = (1 - m_0\alpha_t)\mathbb{I}\{M = \tilde{M}(t)\} + \sum_{e=1}^{m_0} \alpha_t \mathbb{I}\{M = M_e\}.$$

Then the algorithm selects M_t according to $\pi_t(M)$. Unlike the fully combinatorial case, the range of α is selected with respect to different classes of available action spaces $\mathcal{M}$, and we explain the α selection range based on different basic action gap properties below.

For each suboptimal basic action $e \in \tilde{\mathcal{A}}$, where $\tilde{\mathcal{A}} = \mathcal{A} \setminus \mathcal{M}^*$ denotes the set of items not included in any optimal super arm, we define the *minimum gap* $\Delta_{e,\min}$:

$$\Delta_{e,\min} = f(M^*, \boldsymbol{\mu}) - \max_{M\in\mathcal{M}: e\in M,\ \Delta_M > 0} f(M, \boldsymbol{\mu}). \quad (1)$$

We say the *large-gap property* holds when $\Delta_{e,\min} = \Theta(1)$ for all $e \in \mathcal{A}$. This means that no suboptimal action is "nearly optimal": the performance gap between any suboptimal action and the optimal choice is bounded away from zero by a constant, independent of the time horizon n or the problem size. In our algorithm, α can be chosen between $[0, 1]$ when the large-gap property holds, and $\alpha \in [0, \frac{1}{2}]$ when the large-gap property does not hold.

Fourth, `MixCombUCB` observes the weights of all selected items and then updates the estimates $\hat{w}(e)$ and arms estimator R_t accordingly. The estimation of basic actions and super arms is finally given after running for n time steps. The pseudocode for `MixCombUCB` is presented in Algorithm 3.

Our algorithmic framework differs from Kveton et al. (2015), where the UCB-optimal super arm is always selected. By introducing randomness in the selection process, our approach achieves a better balance between minimizing inference error and regret. We begin by presenting a theorem on estimation quality, derived through a martingale construction.

Algorithm 3: Mixture-Based Combinatorial UCB Algorithm (`MixCombUCB`)

Input: $\alpha, \mathcal{M}, n$;
$\{E_1, \hat{\omega}_1, m_0, (1, M_1), ..., (m_0, M_{m_0})\}$=`InitUCB`$(\mathcal{M})$, $T_{m_0-1}(e) = 1, R_{m_0-1}(e) = 0, \forall e \in \mathcal{A}$;
for $t = m_0, ..., n$ **do**
  Set $\alpha_t = \frac{1}{m_0 t^\alpha}$;
  $U_t(e) = \hat{w}_{T_{t-1}(e)}(e) + c_{t-1,T_{t-1}(e)}, \forall e \in \mathcal{A}$;
  Solve optimization problem $\tilde{M}(t) = \arg\max_{M\in\mathcal{M}} f(M, U_t)$;
  Set $\pi_t(M) = (1 - m_0\alpha_t)\mathbb{I}\{M = \tilde{M}(t)\} + \sum_{e=1}^{m_0} \alpha_t \mathbb{I}\{M = M_e\}, \forall M \in \mathcal{M}$;
  Select $M(t)$ according to π_t, Observe $\{(e, w_t(e)) : e \in M(t)\}$;
  $R_t(e) = R_{t-1}(e) + w_t(e)\frac{\mathbb{I}\{e\in M(t)\}}{\mathbb{P}(e\in M(t))}, \forall e \in \mathcal{A}$;
  Set $T_t(e) = T_{t-1}(e), \forall e \in \mathcal{A}, T_t(e) = T_{t-1}(e) + 1, \forall e \in M(t)$;
  $\hat{w}_{T_t(e)}(e) = \frac{T_{t-1}(e)\hat{w}_{T_{t-1}(e)}(e) + w_t(e)}{T_t(e)}, \forall e \in M(t)$;
Output: $\hat{\Delta}_{M,n}^{(i,j)} = \frac{1}{n}(\sum_{e\in M(\tau_i)} R_t(e) - \sum_{e\in M(\tau_j)} R_t(e))$ for $i, j \in [|\mathcal{M}|]$ and $i \neq j$; $\hat{\Delta}_{\mu,n}^{(i,j)} = \frac{1}{n}(\sum_{e\in M(i)} R_t(e) - \sum_{e\in M(j)} R_t(e))$ for $i, j \in [|\mathcal{M}_{UCB}|]$ and $i \neq j$;

Theorem 4.4. *With probability at least $1 - \delta$, it holds*

$$|\hat{\Delta}_{M,n}^{(i,j)} - \Delta_M^{(i,j)}| \leq 6md\sqrt{n^{\alpha-1}}\ln(\frac{2d}{\delta}),$$

where $\mathbb{E}[\hat{\Delta}_{M,n}^{(i,j)}] = \Delta_M^{(i,j)}$ for any $i, j \in [|\mathcal{M}|]$ and $i \neq j$.

We have $\max_{i<j\leq|\mathcal{M}|} \mathcal{E}(n, \hat{\Delta}_{M,n}^{(i,j)}) = \tilde{\mathcal{O}}(\sqrt{n^{\alpha-1}})$ after taking $\delta = \frac{1}{n^2}$. The corresponding error bound on the inference of $\Delta_\mu^{(i,j)}$ is characterized in the following Corollary.

Corollary 4.5. *With probability at least $1 - \delta$, it holds*

$$|\hat{\Delta}_{\mu,n}^{(i,j)} - \Delta_\mu^{(i,j)}| \leq 6d\sqrt{n^{\alpha-1}}\ln(\frac{2d}{\delta}),$$

where $\mathbb{E}[\hat{\Delta}_{\mu,n}^{(i,j)}] = \Delta_{\mu}^{(i,j)}$ *,* $\forall i,j \in [|\mathcal{M}_{UCB}|]$ *and* $i \neq j$*.*

It implies that when taking $\delta = \frac{1}{n^2}$, we have $\max_{i<j\leq|\mathcal{M}_{UCB}|} \mathcal{E}(t, \hat{\Delta}_{\mu,n}^{(i,j)}) = \tilde{\mathcal{O}}(\sqrt{n^{\alpha-1}})$.

Now we proceed with the regret analysis. Given the large-gap property result, we can derive a sharper regret bound, since a non-vanishing gap allows the algorithm to distinguish suboptimal actions more quickly. We formalize this in the following theorem:

Theorem 4.6. *Given the knowledge of gaps between basic actions, the regret of our algorithm is bounded by*

$$\mathcal{R}_\nu(n,\pi) \leq \sum_{e\in\tilde{\mathcal{A}}} \frac{712m}{\Delta_{e,\min}} \log n + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)},$$

where $\Delta_{e,\min}$ *is the minimum gap of suboptimal super arms that contain item* e*, which is defined in* (1)*.*

It shows that if for all $e \in \mathcal{A}$, we have $\Delta_{e,\min} = \Theta(1)$, then the regret of `MixCombUCB` is bounded by $\mathcal{O}(md\log n + mn^{1-\alpha})$. There's also a gap-free bound:

Proposition 4.7. *For all classes of* $\mathcal{M}$*, the regret of our algorithm is bounded as:*

$$\mathcal{R}_\nu(n,\pi) \leq 54\sqrt{mdn\log n} + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)}.$$

This result demostrates that even without the large-gap property, an $\tilde{\mathcal{O}}(\sqrt{n} + n^{1-\alpha})$ regret is still achievable, and it guarantees the exploration of basic action distributions even without assumptions on arms behaviors.

5 PARETO OPTIMAL CONDITION

We now present the conditions for Pareto optimality and verify that our algorithm achieves it. Our results in this section do not differentiate full- and semi-bandit settings, as our proofs are agnostic to the feedback model. We present the sufficient and necessary conditions for Pareto optimality for any CMAB algorithm. We first establish the equivalent condition of Pareto optimality based on super arm M. This setting is comparable to the classic K-arm MAB problem (Simchi-Levi and Wang, 2025); here each super arm M has reward $\sum_{e\in M} \boldsymbol{\mu}(e)$.

Theorem 5.1. *The necessary and sufficient condition for an admissible pair* $(\pi, \hat{\Delta}_M)$ *to be Pareto optimal is*

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

We also extend the trade-off to inference for base arms, where we consider the estimable basic actions set $\mathcal{A}_{ad}$ as aforementioned. Note that in full-bandit setting, $\mathcal{A}_{ad} = \mathcal{M}_{KL}$ and in semi-bandit setting, $\mathcal{A}_{ad} = \mathcal{M}_{UCB}$. Formally, the correponding sufficient and necessary condition can be characterized as follows.

Theorem 5.2. *The necessary and sufficient condition for an admissible pair* $(\pi, \hat{\Delta}_\mu)$ *to be Pareto optimal is*

$$\max_{\nu}\left(\max_{i<j\leq|\mathcal{A}_{\rm ad}|} \mathcal{E}(n, \widehat{\Delta}_{\mu}^{(i,j)})\right)\sqrt{\mathcal{R}_\nu(n,\pi)} = \widetilde{\mathcal{O}}(1).$$

We provide the proof sketch for our Theorem 5.2. We start with the simple case of $|\mathcal{A}_{ad}| = 2$ by proving the information-theoretic bound (in Lemma 5.3), then extend to general case (in Lemma 5.5).

Lemma 5.3. *When* $|\mathcal{A}_{ad}| = 2$*, the lower bound for all admissible pair* $(\pi, \hat{\Delta}_\mu)$ *has*

$$\inf_{(\pi,\widehat{\Delta}_\mu)} \max_{\nu\in\mathcal{V}_0}\left[\mathcal{E}(n, \widehat{\Delta}_\mu)\sqrt{\mathcal{R}_\nu(n,\pi)}\right] = \Omega(1), \qquad (2)$$

for any selected algorithm.

Above lemma demostrates that $\mathcal{E}(n, \widehat{\Delta}_\mu)\sqrt{\mathcal{R}_\nu(n,\pi)}$ will not perform better than a constant order in the worse case, no matter how the solution is chosen. Particularily, we are interested in how the upper bound of $\mathcal{E}(n, \widehat{\Delta}_\mu)\sqrt{\mathcal{R}_\nu(n,\pi)}$ influence our Pareto optimality condition. The following lemma shows that the sufficient condition for Pareto optimality holds when $\mathcal{E}(n, \widehat{\Delta}_\mu)\sqrt{\mathcal{R}_\nu(n,\pi)}$ achieves a constant order.

Lemma 5.4. *When* $|\mathcal{A}_{ad}| = 2$*, the sufficient condition for an admissible pair* $(\pi, \hat{\Delta}_\mu)$ *to be Pareto optimal is*

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

Now we extend our results to the general setting. We have $\max_{i<j\leq|\mathcal{A}_{\rm ad}|} \mathcal{E}(n, \Delta_{\mu}^{(i,j)}) = \tilde{\mathcal{O}}\left(\sqrt{n^{\alpha-1}}\right)$ from 4.2 and 4.5. Combining this with Theorem 4.3 and 4.7, it follows that for any combinatorial bandit instance ν, $\left(\max_{i<j\leq|\mathcal{A}_{\rm ad}|} \mathcal{E}(n, \widehat{\Delta}_{\mu}^{(i,j)})\right) \cdot \sqrt{\mathcal{R}_\nu(n,\pi)} = \widetilde{\mathcal{O}}(1)$. This observation allows us to generalize Lemma 5.4, yielding a *sufficient condition* for the general case: $\max_\nu \left(\max_{i<j\leq|\mathcal{A}_{\rm ad}|} \mathcal{E}(n, \widehat{\Delta}_{\mu}^{(i,j)})\right)\sqrt{\mathcal{R}_\nu(n,\pi)} = \widetilde{\mathcal{O}}(1)$. Moreover, combining the sufficient condition with the definition of Pareto optimality, we can proceed to the following lemma that demostrates the *necessary condition* for Pareto optimality by contradiction:

Lemma 5.5. *The necessary condition for an admissible pair* $(\pi, \hat{\Delta}_\mu)$ *to be Pareto optimal is equivalent to*

$$\max_{\nu}\left(\max_{i<j\leq|\mathcal{A}_{\rm ad}|} \mathcal{E}(n, \widehat{\Delta}_{\mu}^{(i,j)})\right)\sqrt{\mathcal{R}_\nu(n,\pi)} = \widetilde{\mathcal{O}}(1). \quad (3)$$

Having Lemma 5.4 and 5.5, the proof for Theorem 5.2 follows directly. Finally, given Theorem 5.2 and 5.2, we

Table 1: Pareto Frontier restrictions for fully and semi-bandit feedback models. We consider the estimation for both super arms and basic actions and we set $\lambda_{\min}$ for most instances of $\mathcal{M}$ so that $\lambda_{\min}^{-1} = \mathcal{O}(dm^{-1})$ (Cesa-Bianchi and Lugosi, 2012). The parameter α is chosen within the admissible range for each algorithm.

Feedback model	Full-Bandit		Semi-Bandit	
Admissible pair	$(\pi, \hat{\Delta}_{M,n})$	$(\pi, \hat{\Delta}_{\mu,n})$	$(\pi, \hat{\Delta}_{M,n})$	$(\pi, \hat{\Delta}_{\mu,n})$
Inference error $e(n, \hat{\Delta})$	$\tilde{\mathcal{O}}(\sqrt{md^3n^{\alpha-1}})$	$\tilde{\mathcal{O}}(\sqrt{d^3n^{\alpha-1}})$	$\tilde{\mathcal{O}}(md\sqrt{n^{\alpha-1}})$	$\tilde{\mathcal{O}}(d\sqrt{n^{\alpha-1}})$
Regret bound $\mathcal{R}_\nu(\pi, n)$	$\tilde{\mathcal{O}}(mn^{1-\alpha})$	$\tilde{\mathcal{O}}(mn^{1-\alpha})$	$\tilde{\mathcal{O}}(mn^{1-\alpha})$	$\tilde{\mathcal{O}}(mn^{1-\alpha})$
Pareto Frontier S_{PF}	$\tilde{\mathcal{O}}(m\sqrt{d^3})$	$\tilde{\mathcal{O}}(\sqrt{md^3})$	$\tilde{\mathcal{O}}(d\sqrt{m^3})$	$\tilde{\mathcal{O}}(d\sqrt{m})$

are now able to state that our algorithms `MixCombKL` and `MixCombUCB` are Pareto optimal:

Theorem 5.6. *`MixCombKL` and `MixCombUCB` are Pareto optimal when choosing $\alpha \in [0, \frac{1}{2}]$ in full- and semi-bandit settings. Specifically, `MixCombUCB` is Pareto optimal for all $\alpha \in [0, 1]$ if the large gap property $\Delta_{e,\min} = \Theta(1), \forall e \in \mathcal{A}$ holds.*

It is clear that the decision-maker has no incentive to select an α exceeding the specified upper bound, as doing so increases estimation error while offering negligible statistical benefit in terms of the regret upper bound. Thus, we successfully prove that our algorithms are Pareto optimal. Moreover, while the proofs of Pareto optimality in the full- and semi-bandit settings share a high-level structure, the differences in feedback lead to notable variations in parameter constraints. The reason is that the range of the exploration parameter α depends on the amount of information available from feedback. In the full-bandit case, only the total reward of super arms is observed. Over-exploring (choosing a large α) can significantly increase estimation error relative to any gains in regret reduction, limiting α to the range $[0, \frac{1}{2}]$. In contrast, in the semi-bandit case, the reward of each basic action is observed directly, providing richer and more precise feedback. This allows α to safely increase up to 1 when the large-gap property holds, enhancing regret performance, while still maintaining a conservative range of $\alpha \in [0, \frac{1}{2}]$ if the large-gap property does not hold.

6 PARETO FRONTIER AND FEEDBACK STRUCTURES

We consider the Pareto frontiers achieved by `MixCombKL` and `MixCombUCB` in the two bandit feedback settings. From Theorems 4.1, 4.3, 4.4, and 4.6, for any admissible pair $(\pi, \hat{\Delta}_M), (\pi, \hat{\Delta}_\mu)$ and reward distribution $\nu \in \mathcal{V}_0$, `MixCombKL` and `MixCombUCB` can achieve Pareto optimality. We can generally write their respective Pareto frontier $\mathcal{P}_f(\text{Alg}, \pi, \hat{\Delta})$ with respective algorithm and admissible pair $(\pi, \hat{\Delta}_n)$ in the following form:

$$\left\{ (\pi, \hat{\Delta}_n) \;\middle|\; \left(\max_{i<j} e_\nu(n, \hat{\Delta}_n^{(i,j)}) \right) \sqrt{\mathcal{R}_\nu(n, \pi)} = S_{PF} \right\},$$

where S_{PF} is the rate of the Pareto frontier for each algorithm and admissible pair.

In Table 1, we compare the rate differences of Pareto frontiers of our algorithms under feedback models with different information structures. We consider the Pareto frontiers for both super arms and base arms. For fair comparison, we set $\lambda_{\min}^{-1} = \mathcal{O}(dm^{-1})$ (Cesa-Bianchi and Lugosi, 2012). These results indicate that the Pareto frontier of the pair $(\pi, \hat{\Delta}_M)$ achieved by `MixCombUCB` is $\tilde{\mathcal{O}}(\sqrt{d/m})$ tighter than that of `MixCombKL`.

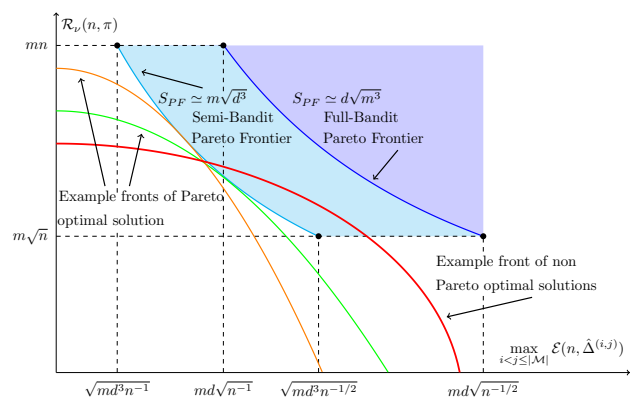

Figure 1: Summary of $\hat{\Delta}_{M,n}$ Pareto Frontier Results with Feedback Difference.

Figure 1 shows the Pareto frontier results under different feedback models. This difference stems from the richness of the bandit feedback. In the semi-bandit case, access to individual arm rewards provides more precise information, yielding a more favorable trade-off between regret and estimation error. In contrast, in the full-bandit case, limited feedback drives a worse Pareto frontier and constrains achievable trade-offs. Thus, semi-bandit algorithms attain more accurate estimations due to improved access to arm information. Yet, the regret in both cases is largely driven by random exploration under appropriate α, and in expectation this step occurs $\mathcal{O}(n^{1-\alpha})$ times under both algorithms, overshadowing the regret from arm estimation and keeping both cases at the $\mathcal{O}(mn^{1-\alpha})$ level.

7 EXPERIEMENT

We evaluate the empirical performance of `MixCombKL` and `MixCombUCB` under synthetic full-bandit and semi-bandit feedback settings with time horizon n. We use the following metrics to evaluate the algorithms:

- $R(n) = \sum_{t=1}^{n}[\mathbb{E}[f(M^*, w_t)] - \mathbb{E}^{\pi}[f(M^{\pi}(t), w_t)]]$
- $MSE_{\mu} = \frac{2}{d(d-1)} \sum_{1 \le i < j \le d} (\hat{\Delta}_{\mu,n}^{(i,j)} - \Delta_{\mu}^{(i,j)})^2$
- $MSE_M = \frac{2}{|\mathcal{M}|^2 - |\mathcal{M}|} \sum_{1 \le i < j \le |\mathcal{M}|} (\hat{\Delta}_{M,n}^{(i,j)} - \Delta_{M}^{(i,j)})^2$

Here $R(n), MSE_{\mu}$ and MSE_M represent the cumulative regret, mean square error of base arms and super arms. For each trial, we draw the expected reward $\mu(e) \sim \mathcal{U}(0.1, 0.9)$ for all $e \in [d]$ independently and allow any non-empty super-arm size at m. We repeat over 20 independent sampling trials for regret and gap estimation error analysis.

Result of MixCombKL Experiment Our method applies forced sampling with decay $\alpha_t = \frac{1}{2t^{\alpha}}$, $\alpha \in \{0, 0.25, 0.5, 1\}$. As the KL-projection based algorithm is harder to converge, we set $d = 8, m = 3$ and have in total $\binom{8}{3} = 56$ super arms. We run experiments for $n = 5000$ steps.

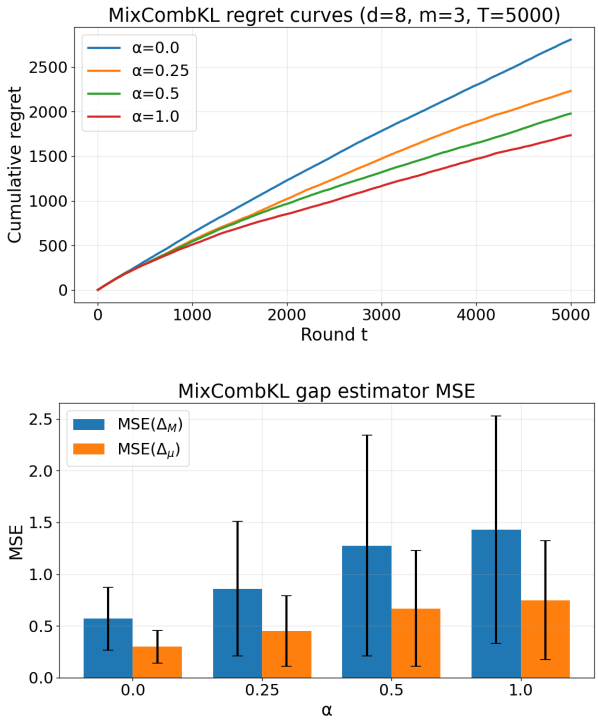

Figure 2: The regret and MSE plot of `MixCombKL`.

Result of MixCombUCB Experiment Similarily, our method applies forced sampling with decay $\alpha_t = \frac{1}{m_0 t^{\alpha}}$, $\alpha \in \{0, 0.25, 0.5, 1\}$. Here we set $d = 9, m = 4$ and we have in total $\binom{9}{4} = 126$ super arms. We run experiments for $n = 2000$ steps.

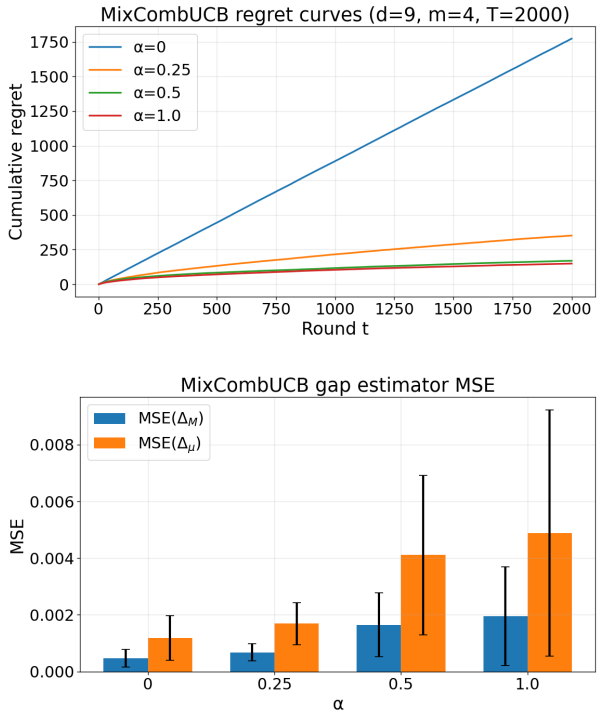

Figure 3: The regret and MSE plot of `MixCombUCB`.

8 CONCLUSION

In this paper, we investigate the concept of Pareto optimality for combinatorial bandits, framing it as the fundamental trade-off between minimizing cumulative regret and reducing average estimation errors of both base and super arm reward gaps. We characterize the sufficient and necessary conditions for achieving Pareto optimality in both full- and semi-bandit feedback settings. We propose two novel algorithms `MixCombKL` for the full-bandit setting and `MixCombUCB` for the semi-bandit case, and show that both algorithms are provably Pareto optimal. Looking ahead, future work could extend our Pareto-optimal framework to dynamic combinatorial settings and investigate Pareto optimality under alternative performance metrics, such as regret variants or average treatment effect (ATE), that better capture practical considerations in complex environments. Another valuable direction is to incorporate constraints (e.g., budgets or fairness) into the trade-off analysis, broadening the framework's applicability to real-world decision systems.

References

Audibert, J.-Y. and Bubeck, S. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT 2010 - Proceedings*, Haifa, Israel.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422. JCSS Special Issue: Cloud Computing 2011.

Chen, L., Li, J., and Qiao, M. (2017). Towards instance optimal bounds for best arm identification. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 535–592. PMLR.

Chen, L., Xu, J., and Lu, Z. (2018). Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA. PMLR.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., and lelarge, m. (2015). Combinatorial bandits revisited. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Csiszar, I. and Shields, P. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1.

Degenne, R., Nedelec, T., Calauzenes, C., and Perchet, V. (2019). Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1988–1996. PMLR.

Faruk, A. S., Shahverdikondori, M., and Zheleva, E. (2025). Learning peer influence probabilities with linear contextual bandits.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42.

Kuleshov, V. and Precup, D. (2014). Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

Kveton, B., Wen, Z., Ashkan, A., Eydgahi, H., and Eriksson, B. (2014). Matroid bandits: fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 420–429, Arlington, Virginia, USA. AUAI Press.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA. PMLR.

Li, F., Liu, J., and Ji, B. (2019). Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813.

Li, S., Wang, B., Zhang, S., and Chen, W. (2016). Contextual combinatorial cascading bandits. In *International conference on machine learning*, pages 1245–1253. PMLR.

Qin, L., Chen, S., and Zhu, X. (2014). Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM.

Saha, A. and Gopalan, A. (2019). Combinatorial bandits with relative feedback. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F.,

Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sankararaman, K. A. and Slivkins, A. (2018). Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pages 1760–1770. PMLR.

Sherali, H. D. (1987). A constructive proof of the representation theorem for polyhedral sets based on fundamental definitions. *American Journal of Mathematical and Management Sciences*, 7(3–4):253–270.

Simchi-Levi, D. and Wang, C. (2025). Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Management Science*, 71(6):4828–4846.

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

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5114–5122. PMLR.

Ye, Z., Zhang, Z., Zhang, D., Zhang, H., and Zhang, R. (2023). Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *SSRN Electronic Journal*.

Zhang, Z. and Wang, Z. (2025). Online experimental design with estimation-regret trade-off under network interference.

Zhong, Z., Cheung, W. C., and Tan, V. Y. F. (2023). Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits.

Zuo, J. and Qin, H. (2025). On pareto optimality for the multinomial logistic bandit. *arXiv preprint arXiv:2501.19277*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Table of Notations

Notation	Description
$[i]$	The set$\{1, ..., i\}$, given a positive integer i
n	Time horizon length
$\mathbb{E}[X]$	Expectation of random variable X
d	Total number of base arms
e	The base arms
m	Numbers of based arms in a super arm M
$M(t)$	The super arm chosen at time t
$\mathcal{A}$	The set of base arms
$\mathcal{M}$	The set of super arms
ω	Feedback vector of the model
$\boldsymbol{\mu}$	Expectation of feedback vector
ν	Distribution of Feedback
$f(G, \varpi)$	Mapping $f : 2^{\mathcal{A}} \times \mathbb{R}^d \to \mathbb{R}$ such that $f(G, \varpi) = \sum_{e \in G} \varpi(e)$
$\pi_t(M)$	Policy probability at time t choosing super arm M
$\mathcal{H}_t$	History $(M(1); w_1, \ldots, M(t); w_t)$ at time t
M^*	The super arm with maximun expected reward $\arg\max_{M \in \mathcal{M}} \sum_{e \in M} \mu(e)$
$\mathcal{R}_\nu(n, \pi)$	The expected regret with time horizon n and policy π
$\Delta_M^{(i,j)}$	Super arm gap $f(M(\tau_i), \boldsymbol{\mu}) - f(M(\tau_j), \boldsymbol{\mu}), \forall i \neq j \in [\lvert\mathcal{M}\rvert]$
$\Delta_\mu^{(i,j)}$	Base arm gap $\mu(i) - \mu(j), \forall i \neq j \in [d]$
Δ_M	The gap to best super arm $f(M^*, \boldsymbol{\mu}) - f(M, \boldsymbol{\mu})$
$\hat{\Delta}_{\mu,t}^{(i,j)}$	The adaptive base arm gap at time t
$\hat{\Delta}_{M,t}^{(i,j)}$	The adaptive super arm gap at time t
$\mathcal{E}(t, \hat{\Delta}_t^{(i,j)})$	The estimation quality $\mathbb{E}[\lvert\Delta^{(i,j)} - \hat{\Delta}_t^{(i,j)}\rvert]$ of estimator $\hat{\Delta}_t^{(i,j)}$ at time t
$\mathcal{P}_f$	The Pareto Frontier defined in 3.2
$f(n) \preceq g(n)$	$f(n)/g(n)$ bounded above and below by constant independent of n
$\boldsymbol{\theta}_M$	Vectorized M as $[\mathbf{I}\{1 \in M\}, ..., \mathbf{I}\{d \in M\}]^\intercal$
$\mathrm{KL}(p, q)$	KL-divergence of distribution p, q
$\lambda_{\min}$	Smallest nonzero eigenvalue of $\mathbb{E}[\boldsymbol{\theta}_M \boldsymbol{\theta}_M^\top]$ with M uniformly distributed
ρ_0	Distribution on $\mathcal{A}$ induced by uniform $\mathcal{M}$
$\rho_{\min}$	$\min_{e \in \mathcal{A}} m\rho_e^0$
C	Acuteness of super arms set $C = \lambda_{\min} m^{-\frac{3}{2}}$
γ	Mixing parameter $\gamma = [\sqrt{m \log \rho_{\min}^{-1}} + \sqrt{C(Cm^2 d + m)n}]^{-1} \sqrt{m \log \rho_{\min}^{-1}}$
η	Exponential control parameter $\eta = \gamma C$

Table 2: Summary of notations used in this paper.

B Computational Results

We define an algorithm as computationally efficient if it can be implemented efficiently whenever the offline version of the problem has an efficient solution (i.e., implementable up to polynomial factors). We now consider the computational problem of our algorithms `MixCombKL` and `MixCombUCB`. In this section, we demonstrate that our algorithms can achieve computational feasibility while maintaining a decent regret upper bound, thus retaining Pareto optimality even with limited computational resources.

B.1 Discussion on Full-Bandit Feedback

We first present Proposition B.1, which shows the computational efficiency of offline `MixCombKL`:

Proposition B.1. *`MixCombKL` is computationally efficient for any stochastic combinatorial bandits where the offline optimization oracle* $\arg\min_{p \in \mathcal{Q}} \mathrm{KL}(p, q)$ *can be implemented efficiently for any probability distribution* $q : q(i) \in \mathbb{R}^+, \forall i \in [d], \sum_{j=1}^d q(j) = 1$.

Proof. At each iteration t, the vector mq'_{t-1} can be expressed as a convex combination of at most $d+1$ arms. By Carathéodory's theorem, such a representation always exists and can be found within $\mathcal{O}(d^4)$ time. Consequently, the probability vector p_{t-1} has at most $d+1$ non-zero entries. Constructing the matrix Σ_{t-1} then requires $\mathcal{O}(d^2)$ time, since it is based on the weighted sum of at most $d+1$ rank-1 matrices $\boldsymbol{\theta}_M \boldsymbol{\theta}_M^\top$, each computed in $\mathcal{O}(d)$ time. Computing the pseudo-inverse of Σ_{t-1} incurs a computational cost of $\mathcal{O}(d^3)$. Aside from the offline optimization oracle, all other operations in the `MixCombKL` algorithm run in polynomial time with respect to m and d, establishing that `MixCombKL` is computationally efficient. □

The above proposition demonstrates that the offline version of `MixCombKL` can be implemented up to polynomial factors. For the online optimization step, one might argue that it may not be possible to compute the projection step exactly, as we are solving an optimization problem on a continuous space. Thus, we are interested in the case where the projection step can be solved up to accuracy ϵ_t in round t, namely, we find q_t such that $\mathrm{KL}(q_t, \tilde{q}_t) - \min_{p \in \Xi} \mathrm{KL}(p, \tilde{q}_t) \leq \epsilon_t$. Proposition B.2 shows that for $\epsilon_t = \mathcal{O}(t^{-2} \log^{-3}(t))$, the approximate projection gives the same regret as when the projection is computed exactly, and its computational complexity will be revealed later.

Proposition B.2. *If the projection step of `MixCombKL` is solved up to accuracy $\epsilon_t = \mathcal{O}(t^{-2} \log^{-3}(t))$, the regret has*

$$\mathcal{R}_\nu(n, \pi) \leq 2\sqrt{2m^3 n \left(d + \frac{m^{1/2}}{\lambda_{\min}}\right) \log \rho_{\min}^{-1}} + \frac{mn^{1-\alpha}}{2(1-\alpha)} + \frac{2m^{5/2}}{\lambda_{\min}} \log \rho_{\min}^{-1}.$$

The proof of the above proposition is shown in Section C.3. Our focus is to verify the computational efficiency of our algorithm when solving the KL-divergence projection step up to a certain accuracy. Suppose the convex set $\mathrm{Co}(\boldsymbol{\theta})$ is characterized by c linear equalities and r linear inequalities. The projection step corresponds to solving a convex optimization problem to an accuracy of $\epsilon_t = \mathcal{O}(t^{-2} \log^{-3}(t))$. We employ the Interior-Point Method (specifically, the barrier method) to solve this.

The total number of Newton iterations required to achieve accuracy ϵ_t is $\mathcal{O}(\sqrt{r} \log(r/\epsilon_t))$, and each iteration incurs a computational cost of $\mathcal{O}((d+c)^3)$ (Boyd and Vandenberghe, 2004, Ch. 10 and 11). Therefore, the overall complexity of the projection step at iteration t is $\mathcal{O}(\sqrt{r}(c+d)^3 \log(r/\epsilon_t))$. By substituting $\epsilon_t = \mathcal{O}(t^{-2} \log^{-3}(t))$, we see that the cumulative cost across n iterations becomes $\mathcal{O}(\sqrt{r}(c+d)^3 n \log n)$. Hence, when $\mathrm{Co}(\boldsymbol{\theta})$ is described by polynomially many linear constraints, the projection step can be implemented within polynomial factors.

Combining the results above shows that `MixCombKL` can be implemented within polynomial factors if the KL divergence is solved up to accuracy $\epsilon_t = \mathcal{O}(t^{-2} \log^{-3}(t))$. Proposition B.2 further shows that the regret remains the same as when the projection is computed exactly, and since the accuracy of arm estimation is unaffected by KL projection errors, it follows from the analysis in Section 5 that `MixCombKL` attains Pareto optimality even under limited computational resources.

B.2 Discussion on Semi-Bandit Feedback

Similarly, we have the following result for computational efficiency in the semi-bandit context:

Proposition B.3. *`MixCombUCB` is computationally efficient in any stochastic semi-bandit where the offline optimization oracle* $\arg\max_{M \in \mathcal{M}} f(M, w)$ *can be implemented efficiently for any* $w \in (\mathbb{R}^+)^d$.

Proof. In each step t, `MixCombUCB` calls the oracle once, and all of its remaining operations are polynomial in m and d. Therefore, `MixCombUCB` is guaranteed to be computationally efficient when the oracle is computationally efficient. □

We now see that the offline version of `MixCombUCB` can be implemented up to polynomial factors. Unlike the full-bandit feedback case in Section B.1, we can assume that the offline optimization oracle is exactly solved, as $\arg\max_{M \in \mathcal{M}} f(M, w)$ operates on a discrete space $\mathcal{M}$, thus keeping the regret bound and estimation error unchanged. Following the analysis of Section 5, `MixCombUCB` can achieve Pareto optimality under polynomial computational resources.

B.3 Computational Example on Problem-Dependent Constants

In this subsection, we discuss about the computational results of certain problem dependent constants with specific super arms set structure.

For example, when the super-arms set $\mathcal{M}$ consist of all subsets up to size m, we can compute that $\rho_{\min} = \frac{m}{d}, \lambda_{\min} = \frac{m(d-m)}{d(d-1)}$, and when $\mathcal{M}$ is the set of perfect matchings, where $d = m^2$ and $|\mathcal{M}| = m!$, we can compute that $\rho_{\min} = \frac{1}{m}, \lambda_{\min} = \frac{1}{m-1}$.

These concrete examples echos with the observation in Csiszar and Shields (2004) that for most classes of $\mathcal{M}$, there is $\rho_{\min}^{-1} = \mathcal{O}(\text{poly}(d))$ and $m(d\lambda_{\min})^{-1} = \mathcal{O}(1)$. Thus, the `MixCombKL` algorithm have a regret of $\mathcal{O}(\sqrt{m^3 dn \log(d/m)} + mn^{1-\alpha})$ on these problem classes.

C Analysis of Algorithm 1

C.1 Proof of Theorem 4.1.

We first prove a simple result:

Lemma C.1. *For all $x \in \mathbb{R}^d$, we have $\Sigma_{t-1}^{+}\Sigma_{t-1}x = \overline{x}$, where $\overline{x}$ is the orthogonal projection of x onto* $\text{span}(\boldsymbol{\theta})$*, the linear space spanned by $\boldsymbol{\theta}_M$.*

Proof. Note that for all $y \in \mathbb{R}^d$, if $\Sigma_{t-1}y = 0$, then we have

$$y^\top \Sigma_{t-1} y = \mathbb{E}\left[y^\top \boldsymbol{\theta}_{M(t)} \boldsymbol{\theta}_{M(t)}^\top y\right] = \mathbb{E}\left[(y^\top M(t))^2\right] = 0, \tag{4}$$

(i) If $\boldsymbol{\theta}_{M(t)}$ has law p_{t-1} such that $\sum_{M\in\mathcal{M}} \boldsymbol{\theta}_{M(t)}(i) p_{t-1}(M) = q'_{t-1}(i)$, $\forall i \in [d]$ and $q'_{t-1} = (1-\gamma)q_{t-1} + \gamma\mu^0$. By definition of μ^0, each $M(t) \in \mathcal{M}$ has a positive probability. (ii) If $M(t)$ is uniformly distributed, of course each $M(t) \in \mathcal{M}$ has a positive probability. Thus we conclude that $\forall t \in [n]$, each $M(t) \in \mathcal{M}$ has a positive probability. Hence, by 4, $y^\top \boldsymbol{\theta}_M = 0$ for all $M \in \mathcal{M}$. In particular, we see that the linear application Σ_{t-1} restricted to $\text{span}(\boldsymbol{\theta})$ is invertible and is zero on $\text{span}(\boldsymbol{\theta})^\perp$, hence we have $\Sigma_{t-1}^{+}\Sigma_{t-1}x = \overline{x}$. □

We consider fillteration $\mathcal{F}_n = \sigma(w_1, w_2, ..., w_n)$, we define the follow martingales

$$r_i(\tau) = \sum_{t=1}^{\tau}\sum_{j=1}^{d} 2\tau^\alpha \mathbb{I}\{U_t = 1\}\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\} w_t(j) - \tau \sum_{t=1}^{\tau}\sum_{j=1}^{d} \mathbb{E}[\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\}]\mathbb{E}[w_t(j)],\ \forall i \in [d], \tau \in [n].$$

It's easy to see that

$$\begin{aligned}
|r_i(t) - r_i(t-1)| &\le |\sum_{j=1}^{d} 2t^\alpha \mathbb{I}\{U_t = 1\}\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\} w_t(j)| + |\sum_{j=1}^{d} \mathbb{E}[\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\}]\mathbb{E}[w_t(j)]| \\
&\le \sum_{j=1}^{d} 2t^\alpha |\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\} w_t(j)| + \sum_{j=1}^{d} \mathbb{E}|\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\}|\mathbb{E}|w_t(j)| \\
&\le \sum_{j=1}^{d} 2t^\alpha |\mathbb{I}\{j \in M\}| + \sum_{i=1}^{d} \mathbb{E}|\mathbb{I}\{j \in M\}| \le 3mt^\alpha \le 3mt^{\frac{\alpha+1}{2}}
\end{aligned}$$

the third equality holds as $|\mathbb{I}\{e \in M\}|, |w_t(e)| \le 1$, $\forall e \in \mathcal{A}$. we denote $s_i = \sum_{j=1}^{d} \mathbb{E}[\mathbb{I}\{i \in M\}\mathbb{I}\{j \in M\}]\mathbb{E}[w_t(j)]$, for all $i \in [d]$, We write $V_{i,t} = \sum_{t=1}^{\tau} \mathbb{E}[(r_i(t) - r_i(t-1))^2 | \mathcal{F}_{t-1}]$ and we have:

$$\begin{aligned}\sum_{t=1}^{\tau}\mathbb{E}[(r_i(t)-r_i(t-1))^2|\mathcal{F}_{t-1}] &= \sum_{t=1}^{\tau}\mathbb{E}[(2t^{\alpha}\mathbb{I}\{U_t=1\}\sum_{j=1}^{d}\mathbb{I}\{i\in M\}\mathbb{I}\{j\in M\}w_t(j))^2|\mathcal{F}_{t-1}]-\tau s_i^2\\ &\le \sum_{t=1}^{\tau}\mathbb{E}[(2t^{\alpha}\mathbb{I}\{U_t=1\}\sum_{j=1}^{d}|\mathbb{I}\{i\in M\}\mathbb{I}\{j\in M\}w_t(j)|)^2|\mathcal{F}_{t-1}]\\ &\le \sum_{t=1}^{\tau}2t^{\alpha}\mathbb{E}[(\sum_{j=1}^{d}|\mathbb{I}\{j\in M\}|)^2|\mathcal{F}_{t-1}]\le\frac{2m^2(\tau+1)^{\alpha+1}}{\alpha+1}\end{aligned}$$

Now with $\delta\in(0,1]$, we have determistic upper bound $\hat{V}_t=\frac{4m^2(\tau+1)^{2\alpha+1}}{2\alpha+1}\ge V_{i,t},\forall i\in[d]$ such that

$$\hat{V}_t\le\frac{2m^2(\tau+1)^{\alpha+1}}{\alpha+1}\vee\frac{9m^2\tau^{\alpha+1}\ln(2/\delta)}{e-2}\le\frac{9m^2\tau^{\alpha+1}\ln(2/\delta)}{e-2}.$$

by Bernstein's Inequality, we know that with possibility more than $1-\delta$, we have

$$|r_i(t)|\le\sqrt{(e-2)\frac{9m^2t^{\alpha+1}(\ln(2/\delta))^2}{e-2}}=3m\ln(\frac{2}{\delta})t^{\frac{\alpha+1}{2}},\ \forall t\in[n].$$

Applying union bound, we can see with probability more than $1-\delta$, there is $|r_i(t)|\le 3m\ln(\frac{2}{\delta})t^{\frac{\alpha+1}{2}},\forall t\in[n],i\in[d]$. we write $\boldsymbol{r}(t)=(r_1(t),...,r_d(t))^\top$ and we can find that

$$\sum_{t=1}^{\tau}(\tilde{w}_t-\mathbb{E}[\tilde{w}_t])=\Sigma_{t-1}^{+}\boldsymbol{r}(\tau),\forall\tau\in[n].$$

By our update rule we know that all Σ_{t-1}^{+} in the above equality is the pseudo-inverse of $\Sigma_{t-1}=\mathbb{E}\left[\boldsymbol{\theta}_M\boldsymbol{\theta}_M^\top\right]$ where M has uniform distribution on $\mathcal{M}$. Now we can find everytime we update $R_t(M(\tau_k))$, there is

$$||\sum_{t=1}^{\tau}(\tilde{w}_t-\mathbb{E}[\tilde{w}_t])||_2\le||\Sigma_{t-1}^{+}||_{op}||\boldsymbol{r}(t)||_2\le\frac{1}{\lambda_{\min}}||\boldsymbol{r}(t)||_2\le\frac{3m\sqrt{dt^{\alpha+1}}}{\lambda_{\min}}\ln(\frac{2d}{\delta}).$$

We use $E_t(Z)=E(Z|w_t)$ for any random variable Z to denote conditional probability on w_t and we have

$$\mathbb{E}_t\left[\tilde{w}_t\right]=\mathbb{E}_t\left[f(M(t),w_t)\Sigma_{t-1}\boldsymbol{\theta}_{M(t)}\right]=\mathbb{E}_t\left[\Sigma_{t-1}^{+}\boldsymbol{\theta}_{M(t)}\boldsymbol{\theta}_{M(t)}^\top w_t\right]=\Sigma_{t-1}^{+}\Sigma_{t-1}w_t=\overline{w_t},\tag{5}$$

where the last equality follows from Lemma C.1 and $\overline{w_t}$ is the orthogonal projection of w_t onto span($\boldsymbol{\theta}$), therefore, we have $\mathbb{E}(\overline{w_t})\in\text{span}(\boldsymbol{\theta}),\forall t\in[n]$. We can derive

$$\boldsymbol{\theta}_{M(\tau_k)}^\top\mathbb{E}(\overline{w_t})=\mathbb{E}(\boldsymbol{\theta}_{M(\tau_k)}^\top\overline{w_t})=\mathbb{E}(\boldsymbol{\theta}_{M(\tau_k)}^\top w_t)=\boldsymbol{\theta}_{M(\tau_k)}^\top\boldsymbol{\mu}=f(M(\tau_k),\boldsymbol{\mu}),\forall k\in[|\mathcal{M}|]$$

In particular, for any $mq'\in\text{Co}(\boldsymbol{\theta})$,we have $\mathbb{E}_t\left[mq'^\top\tilde{w}_t\right]=mq'^\top\overline{w_t}=mq'^\top w_t$.

We now can find that

$$\begin{aligned}R_t(M(\tau_k))-tf(M(\tau_k),\boldsymbol{\mu})&=\boldsymbol{\theta}_{M(\tau_k)}^\top\sum_{t=1}^{\tau}(\tilde{w}_t-\mathbb{E}[\tilde{w}_t])=\hat{R}_t(M(\tau_k))\\ &\le||\boldsymbol{\theta}_{M(\tau_k)}||_2||\sum_{t=1}^{\tau}(\tilde{w}_t-\mathbb{E}[\tilde{w}_t])||_2\le\frac{3\sqrt{m^3dt^{\alpha+1}}}{\lambda_{\min}}\ln\left(\frac{2d}{\delta}\right).\end{aligned}\tag{6}$$

6 shows that

$$\begin{aligned}|\hat{\Delta}_n^{(i,j)}-\Delta^{(i,j)}|&=\mathbb{E}[|(\boldsymbol{\theta}_{M(\tau_i)}^\top\hat{\boldsymbol{\mu}}-\boldsymbol{\theta}_{M(\tau_i)}^\top\boldsymbol{\mu})-(\boldsymbol{\theta}_{M(\tau_j)})^\top\hat{\boldsymbol{\mu}}-\boldsymbol{\theta}_{M(\tau_j)}^\top\boldsymbol{\mu})|]\\ &\le\mathbb{E}[\frac{|\hat{R}_t(M(\tau_i))|+|\hat{R}_t(M(\tau_j))|}{n}]\le\frac{6}{\lambda_{\min}}\sqrt{\frac{m^3d}{n^{1-\alpha}}}\ln\left(\frac{2d}{\delta}\right).\end{aligned}$$

Taking $\delta=\frac{1}{n^2}$ and we can achieve that $\mathcal{E}(n,\hat{\Delta}_n^{(i,j)})=\mathcal{O}(\sqrt{n^{\alpha-1}})$.

C.2 Proof of Theorem 4.3.

For simplicity, we consider two class of time period denoted as $V_n = \{t : U_t = 0, t = 1, ..., n\}$ and $W_n = \{t : U_t = 1, t = 1, ..., n\}$, assume we have $mq^* = M^*$ as the optimal arm, i.e. $q^*(i) = \frac{1}{m}$ if $M_i^* = 1$, we can write our regret as follows:

$$\begin{aligned}\mathcal{R}_\nu(n,\pi) &= \mathbb{E}\Big[\sum_{t=1}^{n} mq^{\star\top}\tilde{w}_t - \sum_{t=1}^{n} mq'^{\top}_{t-1}\tilde{w}_t\Big] \\ &= \mathbb{E}\Big[\sum_{t\in V_n} mq^{\star\top}\tilde{w}_t - \sum_{t\in V_n} mq'^{\top}_{t-1}\tilde{w}_t\Big] + \mathbb{E}\Big[\sum_{t\in W_n} mq^{\star\top}\tilde{w}_t - \sum_{t\in W_n} mq'^{\top}_{t-1}\tilde{w}_t\Big] \\ &\le \mathbb{E}\Big[\sum_{t\in V_n} mq^{\star\top}\tilde{w}_t - \sum_{t\in V_n} mq'^{\top}_{t-1}\tilde{w}_t\Big] + m\mathbb{E}|W_n| \le \mathcal{R}_\nu^{\mathrm{Comb}}(n,\pi) + \frac{mn^{1-\alpha}}{2(1-\alpha)}.\end{aligned} \tag{7}$$

the first equality holds because of 5 and we use $\mathcal{R}_\nu^{\mathrm{Comb}}(n,\pi)$ to denote the regret caused by sample throught p_{t-1}. We first proceed with the following lemma:

Lemma C.2. *We have for any $\eta \le \frac{\gamma\lambda_{\min}}{m^{3/2}}$ and any $q \in \mathcal{P}$,*

$$\sum_{t\in V_n} q^\top\tilde{w}_t - \sum_{t\in V_n} q_{t-1}^\top\tilde{w}_t \le \eta\sum_{t\in V_n} q_{t-1}^\top\tilde{w}_t^2 + \frac{\mathrm{KL}(q,q_0)}{\eta},$$

where $\tilde{w}_t^2$ is the vector that is the coordinate-wise square of $\tilde{w}_t$.

We have:

$$\begin{aligned}\mathbb{E}_t\left[q_{t-1}^\top\tilde{w}_t^2\right] &= \sum_{i\in[d]} q_{t-1}(i)\mathbb{E}_t\left[\tilde{w}_t^2(i)\right] = \sum_{i\in[d]} \frac{q'_{t-1}(i) - \gamma\mu^0(i)}{1-\gamma}\mathbb{E}_t\left[\tilde{w}_t^2(i)\right] \\ &\le \frac{1}{m(1-\gamma)}\sum_{i\in[d]} mq'_{t-1}(i)\mathbb{E}_t\left[\tilde{w}_t^2(i)\right] = \frac{1}{m(1-\gamma)}\mathbb{E}_t\Big[\sum_{i\in[d]}\boldsymbol{\theta}_{\tilde{M}(t)}(i)\tilde{w}_t^2(i)\Big],\end{aligned}$$

where $\tilde{M}(t)$ is a random arm with the same law as $M(t)$ and independent of $M(t)$. Note that $\boldsymbol{\theta}^2_{\tilde{M}(t)} = \boldsymbol{\theta}_{\tilde{M}(t)}$, so that we have

$$\begin{aligned}\mathbb{E}_t\Big[\sum_{i\in[d]}\boldsymbol{\theta}_{\tilde{M}(t)}(i)\tilde{w}_t^2(i)\Big] &= \mathbb{E}_t\Big[w_t^\top\boldsymbol{\theta}_{M(t)}\boldsymbol{\theta}_{M(t)}^\top\Sigma_{t-1}^+\boldsymbol{\theta}_{\tilde{M}(t)}\boldsymbol{\theta}_{\tilde{M}(t)}^\top\Sigma_{t-1}^+\boldsymbol{\theta}_{M(t)}\boldsymbol{\theta}_{M(t)}^\top w_t\Big] \\ &\le m^2\mathbb{E}_t[\boldsymbol{\theta}_{M(t)}^\top\Sigma_{t-1}^+\boldsymbol{\theta}_{M(t)}],\end{aligned}$$

where we used the bound $\boldsymbol{\theta}_{M(t)}^\top w_t \le m$. By the Lemma 15 of Cesa-Bianchi and Lugosi (2012), $\mathbb{E}_t[\boldsymbol{\theta}_{M(t)}^\top\Sigma_{t-1}^+\boldsymbol{\theta}_{M(t)}] \le d$, so that we have:

$$\mathbb{E}_t\left[q_{t-1}^\top\tilde{w}_t^2\right] \le \frac{md}{1-\gamma}.$$

Observe that

$$\begin{aligned}\mathbb{E}_t\left[q^{\star\top}\tilde{w}_t - q'^{\top}_{t-1}\tilde{w}_t\right] &= \mathbb{E}_t\left[q^{\star\top}\tilde{w}_t - (1-\gamma)q_{t-1}^\top\tilde{w}_t - \gamma\mu^{0\top}\tilde{w}_t\right] \\ &= \mathbb{E}_t\left[q^{\star\top}\tilde{w}_t - q_{t-1}^\top\tilde{w}_t\right] + \gamma q_{t-1}^\top X(t) - \gamma\mu^{0\top}X(t) \\ &\le \mathbb{E}_t\left[q^{\star\top}\tilde{w}_t - q_{t-1}^\top\tilde{w}_t\right] + \gamma q_{t-1}^\top X(t) \\ &\le \mathbb{E}_t\left[q^{\star\top}\tilde{w}_t - q_{t-1}^\top\tilde{w}_t\right] + \gamma.\end{aligned}$$

Using Lemma and the above bounds, we get with $mq^\star$ the optimal arm, i.e. $q^\star(i) = \frac{1}{m}$ iff $M_i^\star = 1$,

$$\begin{aligned}\mathcal{R}_\nu^{\mathrm{Comb}}(n,\pi) &= \mathbb{E}\Big[\sum_{t\in V_n} mq^{\star\top}\tilde{w}_t - \sum_{t\in V_n} mq'^{\top}_{t-1}\tilde{w}_t\Big] \\ &\le \mathbb{E}\Big[\sum_{t\in V_n} mq^{\star\top}\tilde{w}_t - \sum_{t\in V_n} mq_{t-1}^{\top}\tilde{w}_t\Big] + mn\gamma \\ &\le \frac{\eta m^2 dn}{1-\gamma} + \frac{m\log\rho_{\min}^{-1}}{\eta} + m\gamma n,\end{aligned}$$

Since

$$\mathrm{KL}(q^*, q_0) = -\frac{1}{m}\sum_{i\in M^*} \log m\rho_i^0 \le \log \rho_{\min}^{-1}.$$

Choosing $\eta = \gamma C$ with $C = \frac{\lambda_{\min}}{m^{3/2}}$ gives

$$\begin{aligned}\mathcal{R}_\nu^{\mathrm{Comb}}(n,\pi) &\le \frac{\gamma C m^2 dn}{1-\gamma} + \frac{m\log\rho_{\min}^{-1}}{\gamma C} + m\gamma n \\ &= \frac{Cm^2 d + m - m\gamma}{1-\gamma}\gamma n + \frac{m\log\rho_{\min}^{-1}}{\gamma C} \\ &\le \frac{(Cm^2 d + m)\gamma n}{1-\gamma} + \frac{m\log\rho_{\min}^{-1}}{\gamma C}.\end{aligned}$$

Set $\gamma = \frac{\sqrt{m\log\rho_{\min}^{-1}}}{\sqrt{m\log\rho_{\min}^{-1}}+\sqrt{C(Cm^2d+m)n}}$ and by 7, we can see that

$$\forall n \ge 1 : \mathcal{R}_\nu(n,\pi) \le 2\sqrt{m^3 n\left(d + \frac{m^{1/2}}{\lambda_{\min}}\right)\log\rho_{\min}^{-1}} + \frac{mn^{1-\alpha}}{2(1-\alpha)} + \frac{m^{5/2}}{\lambda_{\min}}\log\rho_{\min}^{-1}.$$

C.3 Proof of Proposition B.2

We first introduce a simple result from Lemma 5 of Combes et al. (2015):

Lemma C.3. *The KL-divergence $z \mapsto KL(z,q)$ is 1-strongly convex with respect to the $\|\cdot\|_1$ norm.*

Recall that $u_t = \mathbb{I}\{U_t = 1\}q_{t-1} + \mathbb{I}\{U_t = 0\}\arg\min_{p\in\mathcal{Q}} \mathrm{KL}(p, \tilde{q}_t)$ and that q_t is an ϵ_t-optimal solution for the projection step, that is

$$\mathrm{KL}(u_t, \tilde{q}_t) \ge \mathrm{KL}(q_t, \tilde{q}_t) - \epsilon_t.$$

By Lemma C.3 we have

$$\mathrm{KL}(q_t, \bar{q}_t) - \mathrm{KL}(u_t, \bar{q}_t) \ge (q_t - u_t)^\top \nabla \mathrm{KL}(u_t, \bar{q}_t) + \frac{1}{2}\|q_t - u_t\|_1^2 \ge \frac{1}{2}\|q_t - u_t\|_1^2,$$

where we used $(q_t - u_t)^\top \nabla \mathrm{KL}(u_t, \tilde{q}_t) \ge 0$ due to first- order optimality condition for u_t. Hence, $\mathrm{KL}(q_t, \tilde{q}_t) - \mathrm{KL}(u_t, \tilde{q}_t) \le \epsilon_t$ implies that $\|q_t - u_t\|_\infty \le \|q_t - u_t\|_1 \le \sqrt{2\epsilon_t}$. Consider $q^\star$,the distribution over $\mathcal{P}$ for the optimal arm, i.e. $q^\star(i) = \frac{1}{m}$ iff $M_i^\star = 1$. Recall that from proof of Lemma C.2 for $q = q^\star$ we have

$$\mathrm{KL}(q^\star, \bar{q}_t) - \mathrm{KL}(q^\star, q_{t-1}) \le \eta q_{t-1}^\top \overline{w}_t - \eta q^{\star\top}\overline{w}_t + \eta^2 q_{t-1}^\top \overline{w}_t^2. \tag{8}$$

Generalized Pythagorean inequality (see Theorem 3.1 in Csiszar and Shields (2004)) basically says that if $U_t = 0$ then

$$\mathrm{KL}(q^\star, \tilde{q}_t) \ge \mathrm{KL}(q^\star, u_t) + \mathrm{KL}(u_t, \tilde{q}_t). \tag{9}$$

Let $\underline{q}_t = \min_{i\in[d]} q_t(i)$. Observe that for all $t \ge 1$:

$$\begin{aligned}\mathrm{KL}(q^\star, u_t) &= \sum_{i\in[d]} q^\star(i)\log\frac{q^\star(i)}{u_t(i)} = -\frac{1}{m}\sum_{i\in M^\star}\log m u_t(i) \\ &\ge -\frac{1}{m}\sum_{i\in M^\star}\log m(q_t(i) + \sqrt{2\epsilon_t}) \ge -\frac{1}{m}\sum_{i\in M^\star}\left(\log m q_t(i) + \frac{\sqrt{2\epsilon_t}}{\underline{q}_t}\right) \\ &\ge -\frac{\sqrt{2\epsilon_t}}{q_t} - \frac{1}{m}\sum_{i\in M^\star}\log m q_t(i) = -\frac{\sqrt{2\epsilon_t}}{\underline{q}_t} + \mathrm{KL}(q^\star, q_t),\end{aligned}$$

Plugging this into 9, we get that for all t that satisfy $U_t = 0$, there is:

$$\mathrm{KL}(q^\star, \tilde{q}_t) \geq \mathrm{KL}(q^\star, q_t) - \frac{\sqrt{2\epsilon_t}}{\underline{q}_t} + \mathrm{KL}(u_t, \tilde{q}_t) \geq \mathrm{KL}(q^\star, q_t) - \frac{\sqrt{2\epsilon_t}}{\underline{q}_t}.$$

Putting this together with 8 yields that for all t that satisfy $U_t = 0$, there is:

$$\mathrm{KL}(q^\star, q_t) - \mathrm{KL}(q^\star, q_{t-1}) \leq \eta q_{t-1}^\top \overline{w}_t - \eta q^{\star\top} \overline{w}_t + \eta^2 q_{t-1}^\top \overline{w}_t^2 + \frac{\sqrt{2\epsilon_t}}{\underline{q}_t}.$$

Finally, notice that $q_t = q_{t-1}$ for all t that have $U_t = 0$ and summing over t gives:

$$\sum_{t=1}^{n} \left(q^{\star\top} \tilde{w}_t - q_{t-1}^\top \tilde{w}_t\right) \leq \eta \sum_{t=1}^{n} q_{t-1}^\top \tilde{w}_t^2 + \frac{\mathrm{KL}(q^\star, q_0)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \frac{\sqrt{2\epsilon_t}}{\underline{q}_t}.$$

Defining

$$\epsilon_t = \frac{\left(\underline{q}_t \log \rho_{\min}^{-1}\right)^2}{32 t^2 \log^3(t)}, \ \forall t \geq 1,$$

and recalling that $\mathrm{KL}(q^\star, q_0) \leq \log \rho_{\min}^{-1}$,we get

$$\begin{aligned}\sum_{t=1}^{n} \left(q^{\star\top} \tilde{w}_t - q_{t-1}^\top \tilde{w}_t\right) &\leq \eta \sum_{t=1}^{n} q_{t-1}^\top \tilde{w}_t^2 + \frac{\log \rho_{\min}^{-1}}{\eta} + \frac{\log \rho_{\min}^{-1}}{\eta} \sum_{t=1}^{T} \sqrt{\frac{2}{32 n^2 \log^3(t+1)}} \\ &\leq \eta \sum_{t=1}^{n} q_{t-1}^\top \tilde{w}_t^2 + \frac{2 \log \rho_{\min}^{-1}}{\eta},\end{aligned}$$

where we used the fact $\sum_{t \geq 1} t^{-1} (\log(t+1))^{-3/2} \leq 4$. We remark that by the properties of KL-divergence and since $q'_{t-1} \geq \gamma \rho^0 > 0$, we have $\underline{q}_t > 0$ at every round t, so that $\epsilon_t > 0$ at every round t.

Using the above result and following the same lines as in the proof of Theorem 4.3, we have

$$\mathcal{R}_\nu^{\mathrm{Comb}}(n, \pi) \leq \frac{\eta m^2 d n}{1-\gamma} + \frac{2m \log \rho_{\min}^{-1}}{\eta} + m \gamma n$$

Choosing $\eta = \gamma C$ with $C = \frac{\lambda_{\min}}{m^{3/2}}$ gives

$$\mathcal{R}_\nu^{\mathrm{Comb}}(n, \pi) \leq \frac{(C m^2 d + m) \gamma n}{1-\gamma} + \frac{2m \log \rho_{\min}^{-1}}{\gamma C}.$$

The proof is completed by setting $\gamma = \frac{\sqrt{2m \log \rho_{\min}^{-1}}}{\sqrt{2m \log \rho_{\min}^{-1}} + \sqrt{C(Cm^2 d + m) n}}$.

D Analysis of Algorithm 2 and 3

D.1 Proof of Theorem 4.4.

Now we can establish a bound on the estimation error. We define the following martingales: $H_t(e) = R_t(e) - t\mu(e), \forall e \in \mathcal{A}, t \geq 1$ with fillteration $\mathcal{F}_n = \sigma(w_1, w_2, ..., w_n)$, we can verify that

$$\mathbb{E}[H_t(e)|\mathcal{F}_{t-1}] = H_{t-1}(e) + \mathbb{E}[w_t(e) \frac{\mathbb{I}\{e \in M(t)\}}{\mathbb{P}(e \in M(t))} |\mathcal{F}_{t-1}] - t\mu(e) = H_{t-1}(e) + \mathbb{E}[\frac{w_t(e)}{\mathbb{P}(e \in M(t))} | e \in M(t)] - t\mu(e)$$

and we can conclude that $\mathbb{E}[H_t(e)|\mathcal{F}_{t-1}] = H_{t-1}(e)$ and $(H_t(e))_{t\geq 1}$ is martingale for all $e \in \mathcal{A}$.
As $\mathbb{P}(e \in M(t)) \geq \mathbb{P}(M(t) = M_e) = \frac{1}{2m_0 t^\alpha} \geq \frac{1}{2dt^\alpha}$, we can see that for all $t \geq 1$ and $e \in \mathcal{A}$:

$$|H_t(e) - H_{t-1}(e)| \leq |w_t(e)\frac{\mathbb{I}\{e \in M(t)\}}{\mathbb{P}(e \in M(t))}| + \mu(e) \leq |w_t(e)||\frac{1}{\mathbb{P}(e \in M(t))}| + \mu(e) \leq 2dt^\alpha + 1 \leq 3dt^\alpha.$$

For all $e \in \mathcal{A}$, we write $W_t(e) = \sum_{t=1}^{\tau} \mathbb{E}[(H_t(e) - H_{t-1}(e))^2|\mathcal{F}_{t-1}]$ and we have

$$\begin{aligned}\sum_{t=1}^{\tau} \mathbb{E}[(H_t(e) - H_{t-1}(e))^2|\mathcal{F}_{t-1}] &\leq \sum_{t=1}^{\tau} \mathbb{P}(e \in M(t))^{-2}\mathbb{E}[(w_t(e)\mathbb{I}\{e \in M(t)\})^2|\mathcal{F}_{t-1}] - \tau\mu(e)^2 \\ &\leq \sum_{t=1}^{\tau} \mathbb{P}(e \in M(t))^{-2}\mathbb{E}[\mathbb{I}\{e \in M(t)\})^2|\mathcal{F}_{t-1}] \leq \sum_{t=1}^{\tau} \mathbb{P}(e \in M(t))^{-1}\end{aligned}$$

where the second inequality holds due to $|w_t(e)| \leq 1$. now we have $W_\tau(e) \leq \sum_{t=1}^{\tau} \mathbb{P}(e \in M(t))^{-1} \leq \sum_{t=1}^{\tau} 2dt^\alpha$. We let $\hat{W}_\tau = \frac{2d(\tau+1)^{\alpha+1}}{\alpha+1} \geq W_\tau(e), \forall e \in \mathcal{A}, \tau \geq 1$ denote the determistic upper bound and we have for all $\delta \in (0, 1]$:

$$\hat{W}_t \leq \frac{2d(\tau+1)^{\alpha+1}}{\alpha+1} \vee \frac{9d^2\tau^{\alpha+1}\ln(2/\delta)}{e-2} \leq \frac{9d^2\tau^{\alpha+1}\ln(2/\delta)}{e-2}.$$

by Bernstein's Inequality, we know that with possibility more than $1-\delta$, we have

$$|H_t(e)| \leq \sqrt{(e-2)\frac{9d^2t^{\alpha+1}(\ln(2/\delta))^2}{e-2}} = 3d\ln(\frac{2}{\delta})t^{\frac{\alpha+1}{2}}, \forall t \in [n].$$

Applying union bound, we can se with probability more than $1-\delta$, there is $|H_t(e)| \leq 3d\ln(\frac{2d}{\delta})t^{\frac{\alpha+1}{2}}, \forall t \in [n], e \in \mathcal{A}$. Now we can easily derive that

$$\begin{aligned}|\hat{\Delta}_t^{(i,j)} - \Delta^{(i,j)}| &= |\frac{1}{t}(R_t(M(\tau_i)) - f(M(\tau_i), \boldsymbol{\mu})) - \frac{1}{t}(R_t(M(\tau_j) - f(M(\tau_j), \boldsymbol{\mu}))| \\ &\leq \frac{1}{t}(\sum_{e_1 \in M(\tau_i)} |H_t(e_1)| + \sum_{e_2 \in M(\tau_j)} |H_t(e_2)|) \leq 6md\ln(\frac{2d}{\delta})\sqrt{t^{\alpha-1}}\end{aligned}$$

take $\delta = \frac{1}{n^2}$ and we finished our proof.

D.2 Proof of Theorem 4.6.

Now we introduce our first lemma used in our proof.

Lemma D.1. *Let:*

$$\mathcal{F}_t = \left\{\Delta_{M(t)} \leq 2\sum_{e \in \hat{M}(t)} c_{n,T_{t-1}(e)},\ \Delta_{M(t)} > 0\right\} \tag{10}$$

be the event that suboptimal solution $M(t)$ is "hard to distinguish" from M^* at time t, where $\hat{M}(t) = M(t) \setminus M^*$. Let $\mathcal{K}_t = \{M(t) = \tilde{M}(t)\}$ to denote the event that UCB best arm is chosen by algorithm. Then the regret of `MixCombUCB` *is bounded as:*

$$\mathcal{R}_\nu(n, \pi) \leq \mathbb{E}[\hat{\mathcal{R}}_\nu(n, \pi)] + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)}$$

where:

$$\hat{\mathcal{R}}_\nu(n, \pi) = \sum_{t=m_0}^{n} \Delta_{M(t)}\mathbb{I}\{\mathcal{F}_t\}\mathbb{I}\{\mathcal{K}_t\}. \tag{11}$$

The key step in our analysis is that we define a cascade of infinitely-many mutually-exclusive events and then bound the number of times that these events happen when a sub-optimal solution is chosen. The events are parametrized by two decreasing sequences of constants:

$$1 = \beta_0 > \beta_1 > \beta_2 > \ldots > \beta_k > \ldots \tag{12}$$

$$\alpha_1 > \alpha_2 > \ldots > \alpha_k > \ldots \tag{13}$$

such that $\lim_{i\to\infty} \alpha_i = \lim_{i\to\infty} \beta_i = 0$. We define:

$$m_{i,t} = \frac{\alpha_i m^2}{\Delta^2_{M(t)}} \log n$$

and assume that $m_{i,t} = \infty$ when $\Delta_{M(t)} = 0$. The events at time t are defined as in Kveton et al. (2015):

$$G_{1,t} = \{\text{at least } \beta_1 m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{1,t} \text{ times}\}$$

$$\begin{aligned} G_{2,t} = \{&\text{less than } \beta_1 m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{1,t} \text{ times,} \\ &\text{at least } \beta_2 m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{2,t} \text{ times}\} \end{aligned}$$

$$\ldots$$

$$\begin{aligned} G_{i,t} = \{&\text{less than } \beta_1 m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{1,t} \text{ times,} \\ &\ldots, \\ &\text{less than } \beta_{i-1} m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{i-1,t} \text{ times,} \\ &\text{at least } \beta_i m \text{ items in } \hat{M}(t) \text{ were observed at most } m_{i,t} \text{ times}\} \end{aligned}$$

$$\ldots$$

The following lemma establishes a sufficient condition under which events $G_{i,t}$ are exhaustive. This is the key step in the proofs in this section. We firstly introduce an lemma from Lemma 3 of Kveton et al. (2015):

Lemma D.2. *Let (α_i) and (β_i) be defined as in 12 and 13, respectively; and let:*

$$2\sqrt{2} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \leq 1. \tag{14}$$

Let event $\mathcal{F}_t$ happen. Then event $G_{i,t}$ happens for some i.

Now we continue our proof. Let $\mathcal{F}_t$ be the event in 10. By Lemma D.1 and D.2, it remains to bound

$$\hat{\mathcal{R}}_\nu(n, \pi) = \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t\} \mathbb{I}\{\mathcal{K}_t\} = \sum_{i=1}^{\infty} \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{G_{i,t}, \Delta_{M(t)} > 0\} \mathbb{I}\{\mathcal{K}_t\} \,.$$

In the next step, we define item-specific variants of events $G_{i,t}$ and associate the regret at time t with these events. In particular, let:

$$G_{e,i,t} = G_{i,t} \cap \left\{ e \in \hat{M}(t), T_{t-1}(e) \leq m_{i,t} \right\}$$

be the event that item e is not observed "sufficiently often" under event $G_{i,t}$. Then it follows that:

$$\mathbb{I}\{G_{i,t}, \Delta_{M(t)} > 0\} \leq \frac{1}{\beta_i m} \sum_{e \in \bar{\mathcal{A}}} \mathbb{I}\{G_{e,i,t}, \Delta_{M(t)} > 0\} \,,$$

because at least $\beta_i m$ items are not observed "sufficiently often" under event $G_{i,t}$. Therefore, we can bound $\hat{\mathcal{R}}_\nu(n, \pi)$ as:

$$\hat{\mathcal{R}}_\nu(n,\pi) \leq \sum_{e\in\tilde{\mathcal{A}}}\sum_{i=1}^{\infty}\sum_{t=m_0}^{n} \mathbb{I}\{G_{e,i,t}, \Delta_{M(t)} > 0\} \frac{\Delta_{M(t)}}{\beta_i m} .$$

Let each item e be contained in N_e suboptimal super arms and $\Delta_{e,1} \geq \ldots \geq \Delta_{e,N_e}$ be the gaps of these super arms, ordered from the largest gap to the smallest one, we denote those super arms as $M_{e,1},\ldots,M_{e,N_e}$. Then $\hat{\mathcal{R}}_\nu(n,\pi)$ can be further bounded as:

$$\begin{aligned}
\hat{\mathcal{R}}_\nu(n,\pi) &\leq \sum_{e\in\tilde{\mathcal{A}}}\sum_{i=1}^{\infty}\sum_{t=m_0}^{n}\sum_{k=1}^{N_e} \mathbb{I}\{G_{e,i,t}, M(t)=M_{e,k}\}\,\mathbb{I}\{\mathcal{K}_t\}\frac{\Delta_{e,k}}{\beta_i m} \\
&\overset{(a)}{\leq} \sum_{e\in\tilde{\mathcal{A}}}\sum_{i=1}^{\infty}\sum_{t=m_0}^{n}\sum_{k=1}^{N_e} \mathbb{I}\left\{e\in\hat{M}(t), T_{i-1}(e) \leq \alpha_i \frac{m^2}{\Delta_{e,k}^2}\log n\right\}\mathbb{I}\{M(t)=M_{e,k},\mathcal{K}_t\}\frac{\Delta_{e,k}}{\beta_i m} \\
&\overset{(b)}{\leq} \sum_{e\in\tilde{\mathcal{A}}}\sum_{i=1}^{\infty}\frac{\alpha_i m\log n}{\beta_i}\left[\Delta_{e,1}\frac{1}{\Delta_{e,1}^2} + \sum_{k=2}^{N_e}\Delta_{e,k}\left(\frac{1}{\Delta_{e,k}^2} - \frac{1}{\Delta_{e,k-1}^2}\right)\right]
\end{aligned}$$

where inequality (a) is by the definition of event $G_{e,i,t}$, inequality (b) follows from the solution to:

$$\max_{M(m_0),\ldots,M(n)} \sum_{t=m_0}^{n}\sum_{k=1}^{N_e} \mathbb{I}\left\{e\in\hat{M}(t), T_{i-1}(e) \leq \alpha_i \frac{m^2}{\Delta_{e,k}^2}\log n\right\}\mathbb{I}\{M(t)=M_{e,k},\mathcal{K}_t\}\frac{\Delta_{e,k}}{\beta_i m} , \tag{15}$$

The solution of 15 is based on three observations: First, the gaps have order $\Delta_{e,1} \geq \ldots \geq \Delta_{e,N_e}$. Second, by the design of our algorithm, the counter $T_e(t)$ increases when the event $\mathbb{I}\{M(t)=M_{e,k},\mathcal{K}_t\}$ happens. Finally, it's easy to see that $\sum_{k=1}^{N_e}\mathbb{I}\{M(t)=M_{e,k},\mathcal{K}_t\} \leq 1$ for any given e and t. Based on this facts, we can preceed with inequality (b). From Lemma 3 of Kveton et al. (2014), we have that:

$$\hat{\mathcal{R}}_\nu(n,\pi) \leq \sum_{e\in\tilde{\mathcal{A}}}\sum_{i=1}^{\infty}\frac{\alpha_i m\log n}{\beta_i}\frac{2}{\Delta_{e,\min}} = \sum_{e\in\tilde{\mathcal{A}}}\frac{2m}{\Delta_{e,\min}}\left[\sum_{i=1}^{\infty}\frac{\alpha_i}{\beta_i}\right]\log n \tag{16}$$

It remain to choose (α_i) and (β_i) such that:

- $\lim_{i\to\infty}\alpha_i = \lim_{i\to\infty}\beta_i = 0$.
- Monotonicity conditions in Equations 13 and 12 hold.
- Inequality 14 holds: $2\sqrt{2}\sum_{i=1}^{\infty}\frac{\beta_{i-1}-\beta_i}{\sqrt{\alpha_i}} \leq 1$.
- $\sum_{i=1}^{\infty}\frac{\alpha_i}{\beta_i}$ is minimized.

Similar to the proof of Lemma 3 in Kveton et al. (2015), we choose (α_i) and (β_i) to be geometric sequences, $\beta_i = b^i$ and $\alpha_i = la^i$ for $0 < a, b < 1$ and $l > 0$. For this setting, $\alpha_i \to 0$ and $\beta_i \to 0$, and the monotonicity conditions are also satisfied. Moreover, if $b \leq \sqrt{a}$, we have:

$$2\sqrt{2}\sum_{i=1}^{\infty}\frac{\beta_{i-1}-\beta_i}{\sqrt{\alpha_i}} = 2\sqrt{2}\sum_{i=1}^{\infty}\frac{b^{i-1}-b^i}{\sqrt{la^i}} = 2\sqrt{\frac{2}{l}}\cdot\frac{1-b}{\sqrt{a}-b} \leq 1$$

provided that $l \geq 8\left(\frac{1-b}{\sqrt{a}-b}\right)^2$. Furthermore, if $a < b$, we have:

$$\sum_{i=1}^{\infty}\frac{\alpha_i}{\beta_i} = \sum_{i=1}^{\infty}\frac{la^i}{b^i} = \frac{la}{b-a}.$$

Given the above, the best choice of l is $8\left(\frac{1-b}{\sqrt{a}-b}\right)^2$, and the problem of minimizing the constant in our regret bound can be written as:

$$\begin{aligned} \inf_{a,b} \quad & 8\left(\frac{1-b}{\sqrt{a}-b}\right)^2 \cdot \frac{a}{b-a} \\ \text{s.t.} \quad & 0 < a < b < \sqrt{a} < 1. \end{aligned}$$

We find the solution to the above problem numerically, and determine it to be $a = 0.1459$ and $b = 0.2360$. For these values of a and b, $8\left(\frac{1-b}{\sqrt{a}-b}\right)^2 \cdot \frac{a}{b-a} < 356$. We apply this upper bound to Equation 11 and 16, and it follows that the regret is bounded as:

$$\mathcal{R}_\nu(n,\pi) \le \mathbb{E}[\hat{\mathcal{R}}_\nu(n,\pi)] + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)} \le \sum_{e\in\tilde{\mathcal{A}}} \frac{712m}{\Delta_{e,\min}} \log n + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)}$$

D.3 Proof of Proposition 4.7.

The key idea is to decompose the regret into two parts, where the gaps are larger than ϵ and at most ϵ. We analyze each part separately and then set ϵ to get the desired result.

By D.1 ,it remains to bound $\hat{\mathcal{R}}_\nu(n,\pi) = \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t\}\mathbb{I}\{\mathcal{K}_t\}$, where the event $\mathcal{F}_t$ is defined in 10. We partition $\hat{\mathcal{R}}_\nu(n,\pi)$ as:

$$\begin{aligned} \hat{\mathcal{R}}_\nu(n,\pi) &= \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t, \Delta_{M(t)} < \epsilon\}\mathbb{I}\{\mathcal{K}_t\} + \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t, \Delta_{M(t)} \ge \epsilon\}\mathbb{I}\{\mathcal{K}_t\} \\ &\le \epsilon n + \sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t, \Delta_{M(t)} \ge \epsilon\}\mathbb{I}\{\mathcal{K}_t\} \end{aligned}$$

and bound the first term trivially. The second term is bounded in the same way as $\hat{\mathcal{R}}_\nu(n,\pi)$ in the proof of Theorem 4.6, except that we only consider the gaps $\Delta_{e,k} \ge \epsilon$. Therefore, $\Delta_{e,\min} \ge \epsilon$ and we get:

$$\sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\{\mathcal{F}_t, \Delta_{M(t)} \ge \epsilon\}\mathbb{I}\{\mathcal{K}_t\} \le \sum_{e\in\tilde{\mathcal{A}}} \frac{712m}{\epsilon} \log n \le \frac{712}{\epsilon} md \log n\,.$$

Based on the above inequalities:

$$\mathcal{R}_\nu(n,\pi) \le \frac{712md}{\epsilon} \log n + \epsilon n + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)}\,.$$

Finally, we choose $\epsilon = \sqrt{\frac{712md\log n}{n}}$ and get:

$$\mathcal{R}_\nu(n,\pi) \le 2\sqrt{712mdn\log n} + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)} < 54\sqrt{mdn\log n} + 5md + \frac{mn^{1-\alpha}}{2(1-\alpha)}\,,$$

which concludes our proof.

E Analysis of Pareto Optimality

E.1 Proof of Theorem 5.1

First, let us consider the Pareto optimality defined in classic multi-armed bandit with K arms introduced in Simchi-Levi and Wang (2025) which proposes the necessary and sufficient condition for Pareto optimality in MAB with general K arms and time horizon T. Specifically, an admissible pair $(\pi^*, \widehat{\Delta}^*)$ is Pareto optimal if and only if

$$\max_{\nu\in\mathcal{V}_0}\left[\left(\max_{i<j\le K}\mathcal{E}\big(T,\Delta^{*(i,j)}\big)\right)\sqrt{\mathcal{R}_\nu\big(T,\pi^*\big)}\right]=\tilde{\mathcal{O}}(1).$$

In the Combinatorial Bandit setting, we can see each super arm as an arm, thus we have $|\mathcal{M}|$ arms and each arm has its own reward distribution. Speciafically, arm M_τ has a reward distribution with mean $\mu_\tau=\sum_{e\in M_\tau}\mu(e)$. Therefore, it follows that there also exists Pareto optimality in the context of Combinatorial Bandit and the necessary and sufficient condition is

$$\max_{\nu\in\mathcal{V}_0}\left[\left(\max_{i<j\le|\mathcal{M}|}\mathcal{E}\big(n,\Delta_M^{(i,j)}\big)\right)\sqrt{\mathcal{R}_\nu\big(n,\pi\big)}\right]=\bar{\mathcal{O}}(1).$$

where $\mathcal{E}(n,\Delta_M^{(i,j)})$ is the estimation error of ATE between M_{τ_i} and M_{τ_j} with $\mathcal{R}_\nu\big(n,\pi\big)$ being the cumulative regret within n time steps under policy π.

E.2 Proof of Lemma 5.3

We first come up with a lemma:

Lemma E.1. *When $|\mathcal{A}_{ad}|=2$, for any given online decision-making policy π, the error of any estimator of parameter difference can be lower bounded as follows, for any function $f:n\to[0,\frac{1}{8}]$ and any $u\in\mathcal{E}$.*

$$\inf_{\widehat{\Delta}_\mu}\max_{\nu\in\mathcal{V}_0}\mathbb{P}_\nu\left(|\widehat{\Delta}_\mu-\Delta_\nu|\ge g(t)\right)\ge\frac{1}{2}\left[1-\sqrt{\frac{16g(t)^2\mathcal{R}_u\big(n,\pi\big)}{3|\Delta_u|}}\right].$$

Now we can proceed on our proof. Based on Lemma E.1, given policy π, and $\hat{\Delta}_n$, if $g(n)\le\sqrt{\frac{3|\Delta_u|}{64\mathcal{R}_u\left(n,\pi\right)}}$ for some $u\in\mathcal{V}_0$, there is

$$\begin{aligned}\max_{\nu\in\mathcal{V}_0}\mathbb{E}\left[|\widehat{\Delta}_\mu-\Delta_\nu|\right]&\ge g(n)\max_{\nu\in\mathcal{V}_0}\mathbb{P}_\nu\left(|\widehat{\Delta}_\mu-\Delta_\nu|_2\ge g(n)\right)\\&\ge\frac{g(n)}{2}\left[1-\sqrt{\frac{16g(n)^2}{3\Delta_u}\mathcal{R}_u\big(n,\pi\big)}\right]\ge\frac{g(n)}{4},\end{aligned}$$

where the second inequality holds because of Lemma E.1. We use $\nu_{\pi,\hat{\Delta}_\mu}$ to denote $\arg\max_{\nu\in\mathcal{V}_0}\mathbb{E}\left[|\widehat{\Delta}_\mu-\Delta_\nu|\right]$ given policy π and $\widehat{\Delta}_\mu$, and thus we have $e_{\nu_{\pi,\hat{\Delta}_\mu}}(n,\widehat{\Delta}_\mu)\ge\frac{g(n)}{4}$. After choosing $g(n)=\sqrt{\frac{3|\Delta_{\nu_{\pi,\hat{\Delta}_\mu}}|}{64\mathcal{R}_{\nu_{\pi,\hat{\Delta}_v}}(n,\pi)}}$,we retrieve for any given policy π and $\widehat{\Delta}_v$,

$$\begin{aligned}\max_{\nu\in\mathcal{V}_0}\left[e_\nu(n,\widehat{\Delta}_\mu)\sqrt{\mathcal{R}_\nu(n,\pi)}\right]&\ge e_{\nu_{\pi,\bar{\Delta}_\mu}}(n,\widehat{\Delta}_\mu)\sqrt{\mathcal{R}_{\nu_{\pi,\bar{\Delta}_\mu}}(n,\pi)}\\&\ge\frac{g(n)}{4}\sqrt{\mathcal{R}_{\nu_{\pi,\bar{\Delta}_\mu}}(n,\pi)}=\Theta(1),\end{aligned}$$

where the last equation holds because we consider respective $g(n)$ and $\Delta_\nu=\Theta(1)$ for $\nu\in\mathcal{V}_0$. Since the above inequalities hold true for any policy π and $\hat{\Delta}_\mu$, we finish the proof.

E.3 Proof of Lemma 5.4

We proceed by contradiction. Suppose that $(\pi_0,\widehat{\Delta}_0)$ satisfies the above condition, but is not Pareto optimal. Then, there exists a pair $(\pi_1,\widehat{\Delta}_1)$ that Pareto dominates $(\pi_0,\widehat{\Delta}_0)$. According to the lower bound established in Theorem 5.3, there must exist a point on the Pareto front of $(\pi_1,\widehat{\Delta}_1)$, denoted as

$$(\mathcal{E}_{\nu_1}(n,\widehat{\Delta}_1),\mathcal{R}_{\nu_1}(n,\pi_1)),$$

such that

$$\mathcal{E}_{\nu_1}(n, \widehat{\Delta}_1)\sqrt{\mathcal{R}_{\nu_1}(n, \pi_1)} = \Omega(1).$$

By the definition of Pareto dominance, there must exist a point

$$(\mathcal{E}_{\nu_2}(n, \widehat{\Delta}_0), \mathcal{R}_{\nu_2}(n, \pi_0)) \in \mathcal{F}(\pi_0, \widehat{\Delta}_0)$$

such that

$$\mathcal{E}_{\nu_2}(n, \widehat{\Delta}_0)\sqrt{\mathcal{R}_{\nu_2}(n, \pi_0)} > \mathcal{E}_{\nu_1}(n, \widehat{\Delta}_1)\sqrt{\mathcal{R}_{\nu_1}(n, \pi_1)} = \Omega(1).$$

Note that the inequality above is strict with respect to the dependence on n. This implies that

$$\mathcal{E}_{\nu_2}(n, \widehat{\Delta}_0) \cdot \sqrt{\mathcal{R}_{\nu_2}(n, \pi_0)} = \Omega(n^p)$$

for some constant $p > 0$, which contradicts our original assumption.

F Proof of Technical Lemmas

F.1 Proof of Lemma C.2

We have

$$\mathrm{KL}(q, \tilde{q}_t) - \mathrm{KL}(q, q_{t-1}) = \sum_{i\in[d]} q(i) \log \frac{q_{t-1}(i)}{\tilde{q}_t(i)} = -\eta \sum_{i\in[d]} q(i)\tilde{w}_t(i) + \log Z_t,$$

with

$$\begin{aligned} \log Z_t &= \log \sum_{i\in[d]} q_{t-1}(i) \exp\left(\eta \tilde{w}_t(i)\right) \\ &\le \log \sum_{i\in[d]} q_{t-1}(i) \left(1 + \eta\tilde{w}_t(i) + \eta^2 \tilde{w}_t^2(i)\right) \\ &\le \eta q_{t-1}^\top \tilde{w}_t + \eta^2 q_{t-1}^\top \tilde{w}_t^2, \end{aligned} \tag{17}$$

where we used $\exp(z) \le 1 + z + z^2$ for all $|z| \le 1$ in the first inequality and $\log(1 + z) \le z$ for all $z > -1$ in the second inequality. Later we verify the condition for the former inequality.

Hence we have

$$\mathrm{KL}(q, \tilde{q}_t) - \mathrm{KL}(q, q_{t-1}) \le \eta q_{t-1}^\top \tilde{w}_t - \eta q^\top \tilde{w}_t + \eta^2 q_{t-1}^\top \tilde{w}_t^2.$$

Generalized Pythagorean inequality (see Theorem 3.1 in Csiszar and Shields (2004)) basically says that if $U_t = 0$ and $q_t = \arg\min_{p\in\mathcal{Q}} \sum_{i\in[d]} p(i) \log \frac{p(i)}{\tilde{q}_t(i)}$, there is

$$\mathrm{KL}(q, q_t) + \mathrm{KL}(q_t, \tilde{q}_t) \le \mathrm{KL}(q, \tilde{q}_t).$$

Since $\mathrm{KL}(q_t, \tilde{q}_t) \ge 0$, we get if $U_t = 0$,

$$\mathrm{KL}(q, q_t) - \mathrm{KL}(q, q_{t-1}) \le \eta q_{t-1}^\top \overline{w}_t - \eta q^\top \overline{w}_t + \eta^2 q_{t-1}^\top \overline{w}_t^2.$$

Obviously if $U_t = 1$, $q_t = q_{t-1}$, therefore, summing over t gives

$$\sum_{t\in V_n} \left(q^\top \tilde{w}_t - q_{t-1}^\top \tilde{w}_t\right) \le \eta \sum_{t\in V_n} q_{t-1}^\top \tilde{w}_t^2 + \frac{\mathrm{KL}(q, q_0)}{\eta}.$$

To satisfy the inequality condition in 17, i.e., $\eta|\tilde{w}_t(i)| \le 1$, $\forall i \in [d]$, we find the upper bound for $\max_{i\in[d]} |\tilde{w}_t(i)|$ as follows:

$$\begin{aligned} \max_{i\in[d]} |\tilde{w}_t(i)| &\le \|\tilde{w}_t\|_2 = \|\Sigma_{t-1}^+ \boldsymbol{\theta}_{M(t)} f(M(t), w_t)\|_2 \le m\|\Sigma_{t-1}^+ \boldsymbol{\theta}_{M(t)}\|_2 \\ &\le m\sqrt{\boldsymbol{\theta}_{M(t)}^\top \Sigma_{t-1}^+ \Sigma_{t-1}^+ \boldsymbol{\theta}_{M(t)}} \le m\|\boldsymbol{\theta}_{M(t)}\|_2 \sqrt{\lambda_{\max}\left(\Sigma_{t-1}^+ \Sigma_{t-1}^+\right)} \\ &= m^{3/2}\sqrt{\lambda_{\max}\left(\Sigma_{t-1}^+ \Sigma_{t-1}^+\right)} = m^{3/2}\, \lambda_{\max}\left(\Sigma_{t-1}^+\right) = \frac{m^{3/2}}{\lambda_{\min}\left(\Sigma_{t-1}\right)}, \end{aligned}$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively denote the maximum and the minimum nonzero eigenvalue of matrix A. Note that ρ^0 induces uniform distribution over $\mathcal{M}$. Thus by $q'_{t-1} = (1-\gamma)q_{t-1} + \gamma\rho^0$ we see that p_{t-1} is a mixture of uniform distribution and the distribution induced by q_{t-1}. Note that, we have:

$$x^\top \Sigma_{t-1} x = \mathbb{E}\left[x^\top \boldsymbol{\theta}_{M(t)} \boldsymbol{\theta}_{M(t)}^\top x\right] = \mathbb{E}\left[\left(\boldsymbol{\theta}_{M(t)}^\top x\right)^2\right] \geq \gamma \mathbb{E}\left[\left(\boldsymbol{\theta}_M^\top x\right)^2\right],$$

where in the last inequality M has law ρ^0. By definition, we have for any $x \in \operatorname{span}(\boldsymbol{\theta})$ with $\|x\|_2 = 1$, there is $\mathbb{E}\left[(\boldsymbol{\theta}_M^\top x)^2\right] \geq \lambda_{\min}$, so that in the end, we get $\lambda_{\min}(\Sigma_{n-1}) \geq \gamma\lambda_{\min}$, and hence $\eta|\tilde{w}_t(i)| \leq \frac{\eta m^{3/2}}{\gamma\lambda_{\min}}, \forall i \in [d]$. Finally, we choose $\eta \leq \frac{\gamma\lambda_{\min}}{m^{3/2}}$ to satisfy the condition for the inequality we used in 17.

F.2 Proof of Lemma D.1.

Let $R_t = R(M(t), w_t)$ be the stochastic regret of `MixCombUCB` at time t, where $M(t)$ and w_t are the super arm and the weights of the items at time t, respectively. Furthermore, let $\mathcal{E}_t = \{\exists e \in \mathcal{A} : \left|\mu(e) - \hat{w}_{T_{t-1}(e)}(e)\right| \geq c_{t-1,T_{t-1}(e)}\}$ be the event that $\mu(e)$ is outside of the high-probability confidence interval around $\hat{w}_{T_{t-1}(e)}(e)$ for some item e at time t; $\mathcal{K}_t = \{M(t) = \tilde{M}(t)\}$ to denote the event that UCB best arm is chosen by algorithm; and let $\bar{\mathcal{E}}_t$ be the complement of $\mathcal{E}_t$, $\mu(e)$ is in the high-probability confidence interval around $\hat{w}_{T_{t-1}(e)}(e)$ for all e at time t. Then we can decompose the regret of CombUCB1 as:

$$\mathcal{R}_\nu(n,\pi) = \mathbb{E}\left[\sum_{t=1}^{m_0-1} R_t\right] + \mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\left\{\mathcal{E}_t\right\} \mathbb{I}\{\mathcal{K}_t\} R_t\right] + \mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\left\{\bar{\mathcal{E}}_t\right\} \mathbb{I}\left\{\mathcal{K}_t\right\} R_t\right] + \mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\{\bar{\mathcal{K}}_t\} R_t\right].$$

Now we bound each term in our regret decomposition.

The regret of the initialization, $\mathbb{E}\left[\sum_{t=1}^{m_0-1} R_t\right]$, is bounded by md because Algorithm 2 terminates in at most d steps, and $R_t \leq m$ for any $M(t)$ and w_t. The second term in our regret decomposition, $\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\{\mathcal{E}_t\}\mathbb{I}\{\mathcal{K}_t\} R_t\right]$, is small because all of our confidence intervals hold with high probability. In particular, for any e, s, and t:

$$P\left(|\mu(e) - \hat{w}_s(e)| \geq c_{t,s}\right) \leq 2\exp[-3\log t],$$

and therefore:

$$\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\{\mathcal{E}_t\}\mathbb{I}\{\mathcal{K}_t\}\right] \leq \sum_{e\in E}\sum_{t=1}^{n}\sum_{s=1}^{t} P\left(|\mu(e) - \hat{w}_s(e)| \geq c_{t,s}\right) \leq 2\sum_{e\in E}\sum_{t=1}^{n}\sum_{s=1}^{t} \exp[-3\log t] \leq 2\sum_{e\in E}\sum_{t=1}^{n} t^{-2}.$$

Since $R_t \leq m$ for any $M(t)$ and w_t, $\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\left\{\mathcal{E}_t\right\} \mathbb{I}\{\mathcal{K}_t\} R_t\right] \leq 4md$.

Finally, we rewrite the last term in our regret decomposition as:

$$\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\left\{\overline{\mathcal{E}}_t\right\} \mathbb{I}\{\mathcal{K}_t\} R_t\right] \overset{(a)}{=} \sum_{t=m_0}^{n} \mathbb{E}\left[\mathbb{I}\left\{\overline{\mathcal{E}}_t\right\} \mathbb{I}\{\mathcal{K}_t\} \mathbb{E}\left[R_t \mid M(t)\right]\right] \overset{(b)}{=} \mathbb{E}\left[\sum_{t=m_0}^{n} \Delta_{M(t)} \mathbb{I}\left\{\overline{\mathcal{E}}_t, \Delta_{M(t)} > 0\right\} \mathbb{I}\{\mathcal{K}_t\}\right].$$

In equality (a), the outer expectation is over the history of the agent up to time t,which in turn determines $M(t)$ and $\overline{\mathcal{E}}_t$; and $\mathbb{E}\left[R_t \mid M(t)\right]$is the expected regret at time t conditioned on super arm $M(t)$. Equality (b) follows from $\Delta_{M(t)} = \mathbb{E}\left[R_t \mid M(t)\right]$. Now we bound $\Delta_{M(t)} 1\{\overline{\mathcal{E}}_t, \Delta_{M(t)} > 0\}$ for any suboptimal $M(t)$. The bound is derived based on two facts. First, when `MixCombUCB` chooses $M(t)$, $f(M(t), U_t) \geq f(M^*, U_t)$. This further implies that $\sum_{e\in M(t)\setminus M^*} U_t(e) \geq \sum_{e\in M^*\setminus M(t)} U_t(e)$. Second, when event $\overline{\mathcal{E}}_t$ happens, $\left|\mu(e) - \hat{w}_{T_{t-1}(e)}(e)\right| < c_{t-1,T_{t-1}(e)}$ for all items e. Therefore:

$$\sum_{e\in M(t)\setminus M^*} \mu(e) + 2\sum_{e\in M(t)\setminus M^*} c_{t-1,T_{t-1}(e)} \geq \sum_{e\in M(t)\setminus M^*} U_t(e) \geq \sum_{e\in M^*\setminus M(t)} U_t(e) \geq \sum_{e\in M^*\setminus M(t)} \mu(e),$$

and $2\sum_{e\in M(t)\setminus M^*} c_{t-1,T_{t-1}}(e) \geq \Delta_{M(t)}$ follows from the observation that $\Delta_{M(t)} = \sum_{e\in M^*\setminus M(t)} \mu(e) - \sum_{e\in M(t)\setminus M^*} \mu(e)$. Now note that $c_{n,T_{t-1}}(e) \geq c_{t-1,T_{t-1}(e)}$ for any time $t \leq n$. Therefore, the event $\mathcal{F}_t$ must happen and:

$$\mathbb{E}\left[\sum_{t=m_0}^{n} \Delta_{M(t)}\mathbb{I}\left\{\overline{\mathcal{E}}_t, \Delta_{M(t)} > 0\right\}\mathbb{I}\{\mathcal{K}_t\}\right] \leq \mathbb{E}\left[\sum_{t=m_0}^{n} \Delta_{M(t)}\mathbb{I}\left\{\mathcal{F}_t\right\}\mathbb{I}\{\mathcal{K}_t\}\right].$$

Noticed that

$$\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\{\bar{\mathcal{K}}_t\}R_t\right] \leq m\mathbb{E}\left[\sum_{t=m_0}^{n} \mathbb{I}\{\bar{\mathcal{K}}_t\}\right] = \frac{m}{2}\sum_{t=m_0}^{n}\frac{1}{t^\alpha} \leq \frac{mn^{1-\alpha}}{2(1-\alpha)}$$

This conclude our proof.

F.3 Proof of Lemma D.2.

We follow the proof framework of Kveton et al. (2015) and we firstly fix t such that $\Delta_{M(t)} > 0$. Because t is fixed, we use shorthands $G_i = G_{i,t}$ and $m_i = m_{i,t}$. Letr

$$S_i = \left\{e \in \hat{M}(t) : T_{t-1}(e) \leq m_i\right\}$$

be the set of items in $\hat{M}(t)$ that are not observed "sufficiently often" under event G_i. Then event G_i can be written as:

$$G_i = \left(\bigcap_{j=1}^{i-1}\{|S_j| < \beta_j m\}\right) \cap \{|S_i| \geq \beta_i m\}.$$

Now we prove that event G_i happens for some i by showing that the event that none of our events happen cannot happen. Note that this event can be written as:

$$\bar{G} = \overline{\bigcup_{i=1}^{\infty} G_i} = \bigcap_{i=1}^{\infty}\left[\left(\bigcup_{j=1}^{i-1}\{|S_j| \geq \beta_j m\}\right) \cup \{|S_i| < \beta_i m\}\right] = \bigcap_{i=1}^{\infty}\{|S_i| < \beta_i m\}.$$

Let $\bar{S}_i = \hat{M}(t) \setminus S_i$ and $S_0 = \hat{M}(t)$. Then by the definitions of $\bar{S}_i$ and $S_i, \bar{\bar{S}}_{i-1} \subseteq \bar{S}_i$ for all $i > 0$. Furthermore, note that $\lim_{i\to\infty} m_i = 0$. So there must exist an integer j such that $\bar{S}_i = M(t)$ for all $i > j$, and $\hat{M}(t) = \bigcup_{i=1}^{\infty}(\bar{S}_i \setminus \bar{\bar{S}}_{i-1})$. Finally, by the definition of $\bar{S}_i, T_{t-1}(e) > m_i$ for all $e \in \bar{S}_i$. Now suppose that event $\bar{G}$ happens. Then:

$$\sum_{e\in\hat{M}(t)} \frac{1}{\sqrt{T_{t-1}(e)}} < \sum_{i=1}^{\infty}\sum_{e\in\bar{S}_i\setminus\bar{S}_{i-1}} \frac{1}{\sqrt{m_i}} = \sum_{i=1}^{\infty}\frac{|\bar{S}_i \setminus \bar{S}_{i-1}|}{\sqrt{m_i}},$$

We apply the basic properties of $S_i, \bar{S}_i$ and m_i we can know:

$$\begin{aligned}
\sum_{i=1}^{\infty}|\bar{S}_i \setminus \bar{S}_{i-1}|\frac{1}{\sqrt{m_i}} &= \sum_{i=1}^{\infty}(|S_{i-1} \setminus S_i|)\frac{1}{\sqrt{m_i}} = \sum_{i=1}^{\infty}(|S_{i-1}| - |S_i|)\frac{1}{\sqrt{m_i}} \\
&= \frac{|S_0|}{\sqrt{m_1}} + \sum_{i=1}^{\infty}|S_i|\left(\frac{1}{\sqrt{m_{i+1}}} - \frac{1}{\sqrt{m_i}}\right) \\
&< \frac{\beta_0 K}{\sqrt{m_1}} + \sum_{i=1}^{\infty}\beta_i m\left(\frac{1}{\sqrt{m_{i+1}}} - \frac{1}{\sqrt{m_i}}\right) = \sum_{i=1}^{\infty}(\beta_{i-1} - \beta_i)m\frac{1}{\sqrt{m_i}}.
\end{aligned}$$

The first two equalities follow from the definitions of $\bar{S}_i$ and S_i. The inequality follows from the facts that $|S_i| < \beta_i m$ for all $i > 0$ and $|S_0| \leq \beta_0 m$. In addition, let event $\mathcal{F}_t$ happen. Then:

$$\Delta_{M(t)} \leq 2 \sum_{e \in \hat{M}(t)} \sqrt{\frac{2\log n}{T_{t-1}(e)}} < 2\sqrt{2\log n} \sum_{i=1}^{\infty} \frac{(\beta_{i-1} - \beta_i)m}{\sqrt{m_i}} = 2\sqrt{2}\Delta_{M(t)} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \leq \Delta_{M(t)}$$

where the last inequality is due to our assumption in 14 The above is clearly a contradiction. As a result, $\bar{G}$ cannot happen, and event G_i must happen for some i.

F.4 Proof of Lemma E.1

First, we define distribution $\mathcal{B}$ as if $X \sim B(p)$ then $X = 1$ with probability p and $X = 0$ with probability $1 - p$ Then we construct combinatorial model instance $\nu = (\nu_1, \nu_2)$ and two combinatorial bandit instance $\nu_1 = (B(\frac{1}{2} - \zeta), B(\frac{1}{2}))$ and $\nu_2 = (B(\frac{1}{2} - \zeta), B(\frac{1}{2} + 2g(t))$.(we can ignore numbers of basic action d as we can construct $(X_1, X_2) = \nu_i$ with $i = 1, 2$ and $(X_3, ..., X_d) = \vec{0}$ therefore for every super arm $M \in \mathcal{M}$ we can only observe feedback from X_1, X_2)

Without loss of generality we can assume $\zeta \in [0, 1)$ and $g(t) \in [0, \frac{1}{8}]$. Then we have the estimation error Δ_μ being $\Delta_{\nu_1} = \zeta$ and $\Delta_{\nu_2} = \zeta + 2g(t)$ in two cases. We define the minimum distance test $\psi(\widehat{\Delta}_\mu)$ that is associated to $\widehat{\Delta}_\mu$ by

$$\psi(\widehat{\Delta}_\mu) = \arg\min_{i=1,2} |\widehat{\Delta}_\mu - \Delta_{\nu_i}|.$$

If $\psi(\widehat{\Delta}_\mu) = 1$, there is obviously $|\widehat{\Delta}_\mu - \Delta_{\nu_1}| \leq |\widehat{\Delta}_\mu - \Delta_{\nu_2}|$. By the triangle inequality, we know that, if $\psi(\hat{\Delta}_\mu) = 1$,

$$|\widehat{\Delta}_\mu - \Delta_{\nu_2}| \geq |\Delta_{\nu_1} - \Delta_{\nu_2}| - |\widehat{\Delta}_\mu - \Delta_{\nu_1}| \geq |\Delta_{\nu_1} - \Delta_{\nu_2}| - |\widehat{\Delta}_\mu - \Delta_{\nu_2}|,$$

which directly implies that

$$|\widehat{\Delta}_\mu - \Delta_{\nu_2}| \geq \frac{1}{2}|\Delta_{\nu_1} - \Delta_{\nu_2}| = g(t).$$

Symmetrically, if $\psi(\widehat{\Delta}_\mu) = 2$, we wiil see

$$|\widehat{\Delta}_\mu - \Delta_{\nu_1}| \geq \frac{1}{2}|\Delta_{\nu_1} - \Delta_{\nu_2}| = g(t).$$

Therefore, we can apply the above to show

$$\begin{aligned}
\inf_{\hat{\Delta}_\mu} \max_{\nu \in \mathcal{V}_0} \mathbb{P}_\nu \left(|\widehat{\Delta}_\mu - \Delta_\nu| \geq g(t) \right) &\geq \inf_{\hat{\Delta}_\mu} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} \left(|\widehat{\Delta}_\mu - \Delta_{\nu_i}| \geq g(t) \right) \\
&\geq \inf_{\hat{\Delta}_\mu} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} \left(\psi(\widehat{\Delta}_\mu) \neq i \right) \\
&\geq \inf_{\psi} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} \left(\psi \neq i \right).
\end{aligned}$$

where the last infimum is taken over all tests ψ based on $\mathcal{H}_t$ that take values in $\{1, 2\}$.

$$\begin{aligned}
\inf_{\hat{\Delta}_\mu} \max_{\nu \in \mathcal{V}_0} \mathbb{P}_\nu \left(|\widehat{\Delta}_\mu - \Delta_\nu| \geq g(t) \right) &\geq \frac{1}{2} \inf_{\psi} \left(\mathbb{P}_{\nu_1}(\psi = 2) + \mathbb{P}_{\nu_2}(\psi = 1) \right) = \frac{1}{2} \left[1 - \mathrm{TV}(\mathbb{P}_{\nu_1}, \mathbb{P}_{\nu_2}) \right] \\
&\geq \frac{1}{2} \left[1 - \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_{\nu_1}, \mathbb{P}_{\nu_2})} \right] \geq \frac{1}{2} \left[1 - \sqrt{\frac{16g(t)^2}{3\Delta_{\nu_1}} \mathcal{R}_{\nu_1}(n, \pi)} \right].
\end{aligned}$$

where the equality holds due to Neyman-Pearson lemma and the second inequality holds due to Pinsker's inequality, and the third inequality holds due to the following:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\nu_1},\mathbb{P}_{\nu_2}) &= \sum_{t=1}^{n}\mathbb{E}_{\nu_1}[\mathrm{KL}(P_{1,A_t},P_{2,A_t})] = \sum_{i=1}^{2}\mathbb{E}_{\nu_1}[T_i(n)]\mathrm{KL}(P_{1,i},P_{2,i})\\ &= \mathrm{KL}(B(\frac{1}{2}),B(\frac{1}{2}+2g(t)))\,(\mathbb{E}_{\nu_1}[T_2(n)]) \le \frac{32(g(t))^2}{3\Delta_{\nu_1}}\mathcal{R}_{\nu_1}\big(n,\pi\big).\end{aligned}$$

where we use

$$\begin{aligned}\mathrm{KL}(B(\frac{1}{2}),B(\frac{1}{2}+2g(t))) &= \frac{1}{2}\cdot\log\frac{\frac{1}{2}}{\frac{1}{2}+2g(t)}+\frac{1}{2}\cdot\log\frac{\frac{1}{2}}{\frac{1}{2}-2g(t)} = \frac{1}{2}\cdot\log\left(\frac{1}{1-16g(t)^2}\right)\\ &\le \frac{1}{2}(\frac{1}{1-16g(t)^2}-1) \le \frac{8g(t)^2}{1-16g(t)^2} \le \frac{32}{3}g(t)^2.\end{aligned}$$

As we know the history $\mathcal{H}_t$ is generated by selection rule π and $\Delta_{\nu_1}\mathbb{E}_{\nu_1}[T_2(n)]$ is basically the expected regret of ν_1, which is just the definition of regret. It shows that the last inequality holds and thus we finish our proof.

G Construction of Restricted Bandit Structure

We consider the $2d_0$-th ($d_0 \in \mathbb{N}^*$) armed fully combinatorial bandit with super-arm set as

$$\mathcal{M} = \{\{1\},\{2\},\{3,4\},\{5,6\},....,\{2d_0-1,2d_0\}\}$$

and we can see that for every basic action $e \in \{3,4,...,2d_0\} \subset \mathcal{A}$, we can see that $\boldsymbol{\theta}_e \notin \mathrm{span}\{\boldsymbol{\theta}_M : M \in \mathcal{M}\}$, that means we cannot express the reward of basic action $\boldsymbol{\theta}_e w_t$ within the linear space drawn by observable reward $\{\boldsymbol{\theta}_M w_t : M \in \mathcal{M}\}$. This shows that only the reward of basic action $\{1\},\{2\}$ is estimable.

H Existing Techincal Results

In this section we state some well-known theorems used in our paper:

Theorem H.1. *(Bernstein's Inequality). Let $X_1, X_2, \cdots$ be a martingale difference sequence, such that $|X_t| \le \alpha_t$ for a non-decreasing deterministic sequence $\alpha_1, \alpha_2, \cdots$ with probability 1. Let $M_t := \sum_{\tau=1}^{t} X_\tau$ be martingale. Let $\bar{V}_1, \bar{V}_2, \cdots$ be a deterministic upper bounds on the variance $V_t := \sum_{\tau=1}^{t}\mathbb{E}[X_\tau^2|X_1,\cdots,X_{\tau-1}]$ of the martingale M_t, such that $\bar{V}_t - s$ satisfy $\sqrt{\frac{\ln(\frac{2}{\delta})}{(e-2)V_t}} \le \frac{1}{\alpha_t}$. Then, with probability greater than $1-\delta$, for all t:*

$$|M_t| \le 2\sqrt{(e-2)\bar{V}_t \ln\frac{2}{\delta}}.$$

Theorem H.2. *(Neyman-Pearson Lemma). Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Then for any test ψ, it holds*

$$\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0) \ge \int \min(p_0,p_1).$$

Moreover, the equality holds for the Likelihood Ratio test $\psi^\star = \mathbb{I}(p_1 \ge p_0)$.

An direct corollary is that:

Corollary H.3.

$$\inf_{\psi}[\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)] = 1 - TV(\mathbb{P}_0,\mathbb{P}_1).$$

Proof. Denote that $\mathbb{P}_0$ and $\mathbb{P}_1$ are defined on the probability space $(\mathcal{X},\mathcal{A})$. By the definition of the total variation

distance, we have

$$\begin{aligned} TV(\mathbb{P}_0, \mathbb{P}_1) &= \sup_{R \in \mathcal{A}} |\mathbb{P}_0(R) - \mathbb{P}_1(R)| \\ &= \sup_{R \in \mathcal{A}} \left| \int_R p_0 - p_1 \right| \\ &= \frac{1}{2} \int |p_0 - p_1| \\ &= 1 - \int \min(p_0, p_1) \\ &= 1 - \inf_{\psi} \left[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \right]. \end{aligned}$$

where the last equality applies the Neyman-Pearson Lemma, and the fourth equality holds due to the fact that

$$\begin{aligned} \int |p_0 - p_1| &= \int_{p_1 \geq p_0} (p_1 - p_0) + \int_{p_1 < p_0} (p_0 - p_1) \\ &= \int_{p_1 \geq p_0} p_1 + \int_{p_1 < p_0} p_0 - \int \min(p_0, p_1) \\ &= 1 - \int_{p_1 < p_0} p_1 + 1 - \int_{p_1 \geq p_0} p_0 - \int \min(p_0, p_1) \\ &= 2 - 2 \int \min(p_0, p_1). \end{aligned}$$

And we finished our proof. □

Theorem H.4. *(Pinsker's inequality). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures such that $\mathbb{P}_1 \ll \mathbb{P}_2$. Then,*

$$TV(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\frac{1}{2} KL(\mathbb{P}_1, \mathbb{P}_2)}.$$