Fusing Reward and Dueling Feedback in Stochastic Bandits

Xuchuang Wang [1] **Qirun Zeng** [2] **Jinhang Zuo** [3] **Xutong Liu** [4] **Mohammad Hajiesmaili** [1] **John C.S. Lui** [5] **Adam Wierman** [6]

Abstract

This paper investigates the fusion of absolute (reward) and relative (dueling) feedback in stochastic bandits, where both feedback types are gathered in each decision round. We derive a regret lower bound, demonstrating that an efficient algorithm may incur only the smaller among the reward and dueling-based regret for each individual arm. We propose two fusion approaches: (1) a simple elimination fusion algorithm that leverages both feedback types to explore all arms and unifies collected information by sharing a common candidate arm set, and (2) a decomposition fusion algorithm that selects the more effective feedback to explore the corresponding arms and randomly assigns one feedback type for exploration and the other for exploitation in each round. The elimination fusion experiences a suboptimal multiplicative term of the number of arms in regret due to the intrinsic suboptimality of dueling elimination. In contrast, the decomposition fusion achieves regret matching the lower bound up to a constant under a common assumption. Extensive experiments confirm the efficacy of our algorithms and theoretical results.

1. Introduction

Relative feedback is a type of feedback that provides information about the relative quality of two or more items (i.e., which is better) rather than their *absolute* quality. The power of relative feedback has been widely recognized in various fields. For example, in the human alignment stage of the large language model (LLM) training (Ouyang et al., 2022; Rafailov et al., 2024), the human annotators are asked to compare the quality of two LLM-generated sentences, rather than scoring them using absolute values. Given LLMs' popularity, designing algorithms that can effectively leverage relative feedback has become a critical research topic in the online learning community, e.g., reinforcement learning with human feedback (RLHF) (Wang et al., 2023a; Xiong et al., 2024), bandits with human preferences (Ji et al., 2023), dueling bandits (Yan et al., 2022; Saha & Gaillard, 2022), etc. This line of work complements the traditional online learning algorithms that rely on absolute feedback, e.g., reward in bandits and reinforcement learning (Lai & Robbins, 1985; Szepesvári, 2022). We refer to Appendix A for a detailed discussion on related works.

Most of the prior literature has mainly focused on algorithms that leverage either relative or absolute feedback alone. In many real-world applications, however, these two types of feedback coexist. For example, in LLM training, the human annotators may provide both absolute scores and relative comparisons (Ouyang et al., 2022). Another example is in recommendation systems, where the user feedback can be both absolute ratings and pairwise comparisons (choose one among two recommendations) (Zhang et al., 2020). Despite its practical relevance, there is no prior literature where the underlying algorithms simultaneously utilize (dubbed as *fuse*) the absolute and relative feedback.

In this paper, we investigate how to fuse the absolute and relative feedback under the framework of the stochastic multi-armed bandit (MAB) (Lattimore & Szepesvári, 2020), which is a fundamental problem in online learning. To align with the terminology in bandit literature, we refer to the absolute feedback as *reward* and the relative feedback as *dueling* feedback. We consider a MAB with $K \in \mathbb{N}^+$ arms, where each arm k is associated with a reward distribution with an unknown mean, and each arm pair (k_1, k_2) is associated with an unknown dueling probability, representing the probability that arm k_1 is preferred over arm k_2 in a comparison. In each round t, the learner selects a tuple of arms $\{k_t, (k_{1,t}, k_{2,t})\}$, where the reward feedback is the observed reward drawn from the reward distribution of arm k_t, and the dueling feedback is the observed winning arm from the comparison between arms $k_{1,t}$ and $k_{2,t}$.

[1]University of Massachusetts, Amherst, MA [2]University of Science and Technology of China [3]City University of Hong Kong, Hong Kong [4]Carnegie Mellon University, Pittsburgh, PA [5]The Chinese University of Hong Kong, Hong Kong [6]California Institute of Technology, Pasadena, CA. Correspondence to: Xutong Liu <xutongl@andrew.cmu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1: Regret bounds for DR-MAB

Algorithm	Regret Bound
No Fusion†	$O\left(\sum_{k\neq 1}\frac{(\alpha\Delta_k^{(\mathrm{R})}+(1-\alpha)\Delta_k^{(\mathrm{D})})\log T}{\min\{(\Delta_k^{(\mathrm{R})})^2,(\Delta_k^{(\mathrm{D})})^2\}}\right)$
ElimFusion (Alg. 1)	$O\left(\sum_{k\neq 1}\frac{(\alpha\Delta_k^{(\mathrm{R})}+(1-\alpha)\Delta_k^{(\mathrm{D})})\log T}{\max\{(\Delta_k^{(\mathrm{R})})^2,(\Delta_k^{(\mathrm{D})})^2/K\}}\right)$
DecoFusion (Alg. 2)	$O\left(\sum_{k\neq 1}\frac{\log T}{\max\{\Delta_k^{(\mathrm{R})}/\alpha,\Delta_k^{(\mathrm{D})}/(1-\alpha)\}}\right)$
Simplified LB (Cor. 2.4)	$\Omega\left(\sum_{k\neq 1}\frac{\log T}{\max\{\Delta_k^{(\mathrm{R})}/\alpha,\Delta_k^{(\mathrm{D})}/(1-\alpha)\}}\right)$

† Two separate algorithms for reward and dueling feedback.

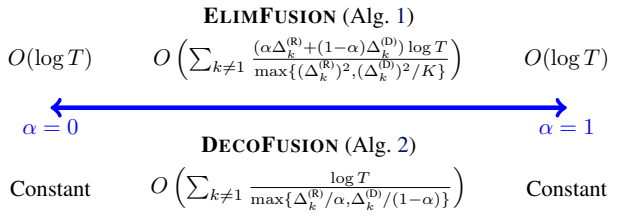


Figure 1: Impact of parameter α on regret R_T

The goal of the learner is to minimize the accumulative *regret* R_T over $T \in \mathbb{N}^+$ rounds. The regret attributes to reward and dueling feedback, called reward-based regret $R_T^{(\mathrm{R})}$ and dueling-based regret $R_T^{(\mathrm{D})}$. The $R_T^{(\mathrm{R})}$ (resp. $R_T^{(\mathrm{D})}$) measures the difference between the reward (resp., dueling probability) of the selected arms (resp., arm pairs) and that of the optimal arm (resp., optimal arm pair) in hindsight. One notable advantage of the dueling feedback is that querying relative feedback from suboptimal arm pairs is often more cost-efficient than that of the absolute feedback (Ouyang et al., 2022). Taking this cost difference into account, we introduce a weight parameter $\alpha \in [0,1]$ and define the final regret as $R_T := \alpha R_T^{(\mathrm{R})} + (1-\alpha)R_T^{(\mathrm{D})}$. We call this model *dueling-reward multi-armed bandit* (DR-MAB) and present its details in Section 2. In this paper, we aim to answer the following central questions:

How do we fuse reward and dueling feedback in DR-MAB *so as to reduce the accumulative regret? How does the weight factor α influence the fusion process and the regret?*

Technical challenges. One key challenge comes from the separated nature of reward and dueling feedback, each providing information on one aspect of the arms (i.e., reward means vs. dueling probabilities), both of which are incomparable. That is, the heterogeneity of both feedback types makes it difficult to fuse them. Furthermore, the relative costs of reward and dueling feedback (parameter α in the regret) further complicate the fusion process. For example, while in general, the learner needs to balance the utilization of two feedback types to minimize regret, there exist cases where one feedback is much cheaper than the other, making it more beneficial to rely solely on the cheaper one.

1.1. Contributions

This paper investigates the fusion of reward and dueling feedback in stochastic MAB setting. We summarize all our technical contributions in Table 1. Below are the details of our technical contributions.

We formulate the dueling-reward MAB problem and study its regret lower bound in Section 2. We first provide a general lower bound and then, with an additional assumption, derive a simplified version as $\Omega(\sum_k \min\{\alpha/\Delta_k^{(\mathrm{R})}, (1-\alpha)/\Delta_k^{(\mathrm{D})}\}\log T)$[1], where the $\Delta_k^{(\mathrm{R})}$ and $\Delta_k^{(\mathrm{D})}$ are the reward and dueling gaps of arm k (definition in Section 2). This bound highlights the benefit of fusing reward and dueling feedback: for each suboptimal arm k, among the two weighted regret costs induced from reward and dueling feedback, fusion makes it possible to only pay the smaller one.

We first propose a simple elimination fusion (ElimFusion) algorithm in Section 3. ElimFusion applies the elimination algorithms for stochastic and dueling bandits separately to both types of feedback and then fuses their information by sharing two algorithms' candidate arm sets (arms not identified as suboptimal yet). ElimFusion enjoys the benefit of fusion suggested by the lower bound and achieves a regret bound of $O(\sum_k(\alpha\Delta_k^{(\mathrm{R})} + (1-\alpha)\Delta_k^{(\mathrm{D})})\min\{1/(\Delta_k^{(\mathrm{R})})^2, K/(\Delta_k^{(\mathrm{D})})^2\}\log T)$, where the regret cost of dueling feedback is suboptimal in terms of the factor K, which is inherited from the intrinsic suboptimality of the dueling elimination algorithm.

We then devise a decomposition fusion (DecoFusion) algorithm in Section 4. Notice that the lower bound suggests a decomposition of the suboptimal arms into two subsets: one with the smaller regret cost induced by reward feedback and the other smaller by dueling feedback. DecoFusion leverages this insight by approximating this arm decomposition. However, without the knowledge of the reward and dueling gaps, the approximated decomposition may deviate from the ground-truth one, which we further address by proposing a novel randomized decision-making strategy that explicitly separates the exploration and exploitation. Putting all together, DecoFusion achieves a regret bound of $O(\sum_k \min\{\alpha/\Delta_k^{(\mathrm{R})}, (1-\alpha)/\Delta_k^{(\mathrm{D})}\}\log T)$, which matches the simplified lower bound (instead of the general form) and outperforms ElimFusion.

Furthermore, as Figure 1 shows, DecoFusion is able to fully utilize a free exploration property when either $\alpha = 0$ (free reward) or $\alpha = 1$ (free dueling), to achieve a constant

[1] Most big-O and big-Ω formulas in the introduction are informally tailored for readability. We refer the readers to corresponding theorems for their formal expressions.

(i.e., independent from T) regret, which is impossible for ELIMFUSION. Lastly, we conduct simulations to evaluate the performance of the proposed algorithms in Section 5.

2. Model Formulation

Consider a dueling-reward multi-armed bandit (DR-MAB) with $K \in \mathbb{N}^+$ arms. Each arm $k \in \mathcal{K} \coloneqq \{1, 2, \dots, K\}$ is associated with a Bernoulli[2] reward distribution with unknown mean $\mu_k \in (0, 1)$. Each arm pair $(k, \ell) \in \mathcal{K}^2$ is associated with a dueling probability $\nu_{k,\ell} \in (0, 1)$, representing the probability that arm k wins arm ℓ in a duel. Especially, $\nu_{k,\ell} > 0.5$ if $\mu_k > \mu_\ell$ and $\nu_{k,\ell} = 0.5$ if $\mu_k = \mu_\ell$. and arm pairs (k, ℓ) and (ℓ, k) are symmetric, i.e., $\nu_{k,\ell} = 1 - \nu_{\ell,k}$. Besides this ordering relation between reward means μ_k and dueling probabilities $\nu_{k,\ell}$, we do not assume or utilize any other specific relation between μ_k and $\nu_{k,\ell}$. For simplicity, we assume all reward means are distinct, and these arms are labeled in descending order regarding their reward means, i.e., $\mu_1 > \mu_2 > \cdots > \mu_K$. Therefore, arm 1 is the unique *optimal arm* and the *Condorcet winner* (i.e., $\nu_{1,k} > 0.5$ for any arm $k > 1$) (Urvoy et al., 2013).

Consider $T \in \mathbb{N}^+$ rounds in DR-MAB. In each round $t \in \{1, 2, \dots, T\}$, the learner picks a tuple $\{k_t, (k_{1,t}, k_{2,t})\}$ consisting of an arm k_t and a pair of arms $(k_{1,t}, k_{2,t})$. Then, the learner observes the reward realization $X_{k_t,t}$ of the chosen arm k_t and the winner of the pair duel $Y_{k_{1,t},k_{2,t},t}$, where $X_{k_t,t}$ is sampled from the Bernoulli distribution with mean μ_{k_t}, and $Y_{k_{1,t},k_{2,t},t}$ is determined by a sample from the Bernoulli distribution with mean $\nu_{k_{1,t},k_{2,t}}$: $Y_{k_{1,t},k_{2,t},t} = k_{1,t}$ if the sample is 1, and $Y_{k_{1,t},k_{2,t},t} = k_{2,t}$ if it is 0. Both $X_{k_t,t}$ and $Y_{k_{1,t},k_{2,t},t}$ are independent across rounds and arms (pairs). We call the $X_{k_t,t}$ as the *reward feedback* (absolute) and the $Y_{k_{1,t},k_{2,t},t}$ as the *dueling feedback* (relative), following the convention of bandit literature.

Regret objective. The learner aims to minimize the accumulative *regret* R_T over T rounds, composed by the reward-based regret $R_T^{(\mathrm{R})}$ and the dueling-based regret $R_T^{(\mathrm{D})}$. Denote $\Delta_k^{(\mathrm{R})} \coloneqq \mu_1 - \mu_k$ and $\Delta_k^{(\mathrm{D})} \coloneqq \nu_{1,k} - 0.5$ as the reward and dueling gaps between the optimal arm 1 and the suboptimal arm k, respectively. Then, $R_T^{(\mathrm{R})}$ is defined as the accumulation of the reward gaps $\Delta_{k_t}^{(\mathrm{R})}$ of all chosen arms k_t over the T rounds, while $R_T^{(\mathrm{D})}$ is defined as the accumulation of the average of the dueling gaps $(\Delta_{k_{1,t}}^{(\mathrm{D})} + \Delta_{k_{2,t}}^{(\mathrm{D})})/2$ of the picked arm pairs $(k_{1,t}, k_{2,t})$, i.e.,

$$R_T^{(\mathrm{R})} \coloneqq \sum_{t=1}^{T} \Delta_{k_t}^{(\mathrm{R})} = T \cdot \mu_1 - \sum_{t=1}^{T} \mu_{k_t},$$
$$R_T^{(\mathrm{D})} \coloneqq \sum_{t=1}^{T} \frac{\Delta_{k_{1,t}}^{(\mathrm{D})} + \Delta_{k_{2,t}}^{(\mathrm{D})}}{2} = \sum_{t=1}^{T} \frac{\nu_{1,k_{1,t}} + \nu_{1,k_{2,t}} - 1}{2}.$$

Lastly, we introduce a parameter $\alpha \in [0, 1]$ to balance the impact of the reward-based and the dueling-based regrets on the aggregated regret R_T, defined as follows,

$$R_T \coloneqq \alpha R_T^{(\mathrm{R})} + (1 - \alpha) R_T^{(\mathrm{D})}. \tag{1}$$

2.1. Lower Bound

This section provides the regret lower bound for any consistent algorithm (Definition 2.1) in DR-MAB. We denote $N_{k,t}$ as the number of times arm k is picked in the first t rounds for reward feedback, and $M_{k,\ell,t}$ as the number of times the pairs (k, ℓ) and (ℓ, k) (due to their symmetry) are picked in the first t rounds for dueling feedback.

Definition 2.1 (Consistent algorithm)**.** An algorithm is called *consistent* for DR-MAB if for any suboptimal arm $k \neq 1$ and parameter $\gamma > 0$, it fulfills $\mathbb{E}[N_{k,T}] = o(T^\gamma)$ and $\mathbb{E}[M_{k,\ell,T}] = o(T^\gamma)$ for any arm $\ell \neq k$.

The consistent definition covers all algorithms that achieve logarithmic regrets in DR-MAB, e.g., UCB and elimination (Auer, 2002; Auer & Ortner, 2010). This definition is a *"generalization"* of the consistent policy in stochastic bandits (Lai & Robbins, 1985). We first provide a lemma to bound the number of arm pulling and pair dueling.

Lemma 2.2. *For any suboptimal arm $k \neq 1$, under any consistent algorithm, we have*

$$N_{k,T}\,\mathrm{kl}(\mu_k, \mu_1) + \sum_{\ell<k} M_{k,\ell,T}\,\mathrm{kl}\left(\nu_{k,\ell}, 0.5\right) \geqslant (1 - o(1)) \log T, \tag{2}$$

where $\mathrm{kl}(p, q) \coloneqq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is KL-divergence between two Bernoulli distributions with means p and q.

The proof of Lemma 2.2 (formally in Appendix C) is based on the information-theoretic lower bounds for MAB (Lai & Robbins, 1985) and the dueling bandits (Komiyama et al., 2015). This proof needs a careful selection of pairs of the DR-MAB instances and constructions of the key information quantities, so that the attributions from reward and dueling feedback would both appear in the LHS of (2).

The LHS of (2) can be interpreted as the amount of "information" collected by the learner for arm k in the T rounds, and its two terms are "information" from the reward and

[2]The choice of Bernoulli distributions is mainly for simplicity. One can extend the assumption to distributions with bounded interval support, i.e., $[0, 1]$-bounded, via more sophisticated analysis.

dueling feedback. Their additive relation in the LHS implies that the "information" can be attributed to two parts: the reward-based (first term) and the dueling-based (second term). Especially, the second term for dueling-based information of suboptimal arm k is the sum over all pairs between the arm k and other better arms $\ell\,(<k)$, implying that the *effective* dueling-based information is collected by dueling with better arms.

Although Lemma 2.2 suggests that the necessary exploration on a suboptimal arm k may be fulfilled by aggregating information collected from the reward and dueling feedback, we prove the following theorem to show that the potentially optimal way to minimize regret due to each suboptimal arm k is to focus on either the reward-based or the dueling-based exploration, *exclusively*.

Theorem 2.3. *For any consistent algorithm and regret balanced by $\alpha \in [0,1]$, the following regret lower bound holds,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \geqslant \sum_{k\neq 1} \min\left\{ \frac{\alpha\Delta_k^{(\mathrm{R})}}{\mathrm{kl}(\mu_k,\mu_1)}, \min_{\ell<k} \frac{(1-\alpha)(\Delta_k^{(\mathrm{D})}+\Delta_\ell^{(\mathrm{D})})}{\mathrm{kl}\left(\nu_{k,\ell},\frac{1}{2}\right)} \right\}. \tag{3}$$

A full proof is given in Appendix C and is based on minimizing the regret decomposed in terms of the sampling times $N_{k,T}$ and $M_{k,\ell,T}$, given the constraints in Lemma 2.2. As both the regret minimization objective and the constraints are linear expressions, the minimization can be solved by linear programming, yielding the minimal attained in one of two end nodes of the constraint line segment in (2).

Theorem 2.3 provides a regret lower bound for any consistent algorithm in DR-MAB. The sum in the RHS of (3) is over all suboptimal arms $k \neq 1$. The two terms inside the outer minimization in the RHS correspond to the reward-based and dueling-based regrets for arm k. The outer minimization indicates that for arm k, an effective algorithm should explore it through either the reward or dueling feedback, depending on which yields a smaller regret.

The second term (inner minimization) in the RHS of (3) implies that for each suboptimal arm k, there exists a most effective arm (competitor) to duel with, denoted as $\ell_k^* \coloneqq \min_{\ell<k} (\Delta_k^{(\mathrm{D})} + \Delta_\ell^{(\mathrm{D})})/\mathrm{kl}\left(\nu_{k,\ell},\frac{1}{2}\right)$. However, as discussed in Komiyama et al. (2015, §3.2), the most effective arm ℓ_k^* is usually the optimal arm, i.e., $\ell_k^* = 1$, in many real world applications. Then, by assuming $\ell_k^* = 1$, the key term in the lower bound for any suboptimal arm k becomes $\min\{\alpha\Delta_k^{(\mathrm{R})}/\mathrm{kl}(\mu_k,\mu_1), (1-\alpha)\Delta_k^{(\mathrm{D})}/\mathrm{kl}\left(\nu_{k,1},\frac{1}{2}\right)\}$. Further noticing the $\mathrm{kl}(a,b) = \Theta((a-b)^2)$ property for the KL-divergence between two Bernoulli distributions (Lattimore & Szepesvári, 2020, §16), this term can be simplified as $\min\{\alpha/\Delta_k^{(\mathrm{R})}, (1-\alpha)/\Delta_k^{(\mathrm{D})}\} = 1/\max\{\Delta_k^{(\mathrm{R})}/\alpha, \Delta_k^{(\mathrm{D})}/(1-\alpha)\}$. This simplification is summarized in the following corollary.

Corollary 2.4. *If for each suboptimal arm $k \neq 1$, the most effective dueling arm ℓ_k^* is the optimal arm 1, the regret lower bound in Theorem 2.3 can be simplified as follows, for some universal positive constant C,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \geqslant C \sum_{k\neq 1} \frac{1}{\max\{\Delta_k^{(\mathrm{R})}/\alpha, \Delta_k^{(\mathrm{D})}/(1-\alpha)\}} \tag{4}$$

From the regret lower bounds in (3) and (4), one may get a counter-intuitive observation: when $\alpha = 0$ or 1, the regret lower bound would become sub-logarithmic for any consistent algorithm, which is unusual in the bandit literature. Later in this paper, we will show that a sub-logarithmic $o(\log T)$ regret is actually achievable (in fact, T-independent constant regret) in these two scenarios, revealing a novel phenomenon of DR-MAB.

3. ElimFusion: Fusing Reward and Dueling via Elimination

As the reward means and dueling probabilities only have a weak ordering relation—the dueling probability of arm k winning over arm ℓ is greater than 0.5 if and only if arm k has a higher reward mean than arm ℓ—and their observations are independently sampled from distributions with different parameters, the estimations of reward means μ_k and dueling probabilities $\nu_{k,\ell}$ are intrinsically separated. This separation makes it difficult to combine these two types of feedback in online learning directly.

To address this challenge, in this section, we introduce our first approach to fusing reward and dueling feedback in DR-MAB, named Elimination Fusion (ElimFusion). Although ElimFusion has a suboptimal regret in DR-MAB, it suggests a simple and effective way to fuse absolute and relative feedback. As the ultimate goal is to address the fusion in general online learning problems beyond bandits, the intuitive high-level idea of ElimFusion may have a broader application to other learning settings.

3.1. Algorithm Design of ElimFusion

Arm elimination is a common technique in multi-armed bandits (Auer & Ortner, 2010) and dueling bandits (Saha & Gaillard, 2022). The main idea is maintaining a candidate arm set $\mathcal{C}$, initialized as the full arm set $\mathcal{K}$, and as the learning proceeds, one gradually identifies and eliminates suboptimal arms from the set $\mathcal{C}$ based on the observed feedback, until only the optimal arm remains. ElimFusion leverages the maintained candidate arm set as a bridge to connect the two types of feedback. Specifically, ElimFusion maintains a single candidate arm set $\mathcal{C}$, and arms in

this set can be eliminated according to *either* reward *or* dueling feedback, whichever is first triggered.

ELIMFUSION in Algorithm 1 starts with a warm-up phase (detained in Algorithm 3 in Appendix B) to initialize the reward and dueling statistics, where each pair of arms is dueled once, and each arm is queried for rewards in a round-robin manner in the same time. In each time slots, ELIMFUSION picks the arm and arm pair with smallest number of pulls and dueling times, respectively, which uniformly explores arms (Lines 7–10). With the new reward and dueling observations, ELIMFUSION updates the reward and dueling statistics for each arm pair and arm, respectively (Line 11, detailed in Algorithm 4 in Appendix B). Then, ELIMFUSION eliminates arms based on the observed feedback (Line 14). To illustrate the arm elimination, we define the confidence radius $\mathrm{CR}^{(\mathrm{R})}_{k,t} := \sqrt{2\log(Kt/\delta)/N_{k,t}}$ and $\mathrm{CR}^{(\mathrm{D})}_{k,\ell,t} := \sqrt{2\log(Kt/\delta)/M_{k,\ell,t}}$ for the estimated reward mean and dueling probability, where $\delta > 0$ is the input confidence parameter of ELIMFUSION. Then, ELIMFUSION eliminates an arm k if *either* its upper confidence bound (UCB) of reward mean estimate $\hat{\mu}_{k,t+1} + \mathrm{CR}^{(\mathrm{R})}_{k,t+1}$ is less than the lower confidence bound (LCB) of highest mean estimate $\hat{\mu}_{\hat{k}^{(\mathrm{R})}_{t+1},t+1} - \mathrm{CR}^{(\mathrm{R})}_{\hat{k}^{(\mathrm{R})}_{t+1},t+1}$ (the $\hat{k}^{(\mathrm{R})}_{t+1}$ is the estimated optimal arm in Line 13) *or* there exists an arm ℓ in $\mathcal{C}$ such that the UCB of its dueling probability $\hat{\nu}_{k,\ell,t+1} + \mathrm{CR}^{(\mathrm{D})}_{k,\ell,t+1}$ is less than $1/2$ (i.e., arm ℓ outperforms arm k). ELIMFUSION keeps eliminating arms via the above loop, and when only the optimal arm remains in $\mathcal{C}$, i.e., $|\mathcal{C}| = 1$, it switches to exploitation (Lines 15–17).

Algorithm 1 Elimination Fusion (ELIMFUSION)

1: **Input:** Arm set $\mathcal{K}$, confidence parameter δ
2: **Initialize:** Candidate arm set $\mathcal{C} \leftarrow \mathcal{K}$, time slot $t \leftarrow 0$, reward pull counter $N_{k,t} \leftarrow 0$ and reward estimate $\hat{\mu}_{k,t} \leftarrow 0$ for all arms $k \in \mathcal{K}$, dueling pull counter $M_{k,\ell,t} \leftarrow 0$ and dueling estimate $\hat{\nu}_{k,\ell,t} \leftarrow 0$ for all pairs $(k,\ell) \in \mathcal{K}^2$
3: Warm-up (Algorithm 3)
4: **for** each time slot t **do**
5: `# decision making:`
6: **if** $|\mathcal{C}| > 1$ **then** # {uniform explore}
7: $k_t \leftarrow \arg\min_{k\in\mathcal{C}} N_{k,t}$
8: $(k_{t,1}, k_{t,2}) \leftarrow \arg\min_{(k,\ell)\in\mathcal{C}^2, k\neq\ell} M_{k,\ell,t}$
9: Pull arm k_t and observe arm reward $X_{k_t,t}$
10: Duel $(k_{t,1}, k_{t,2})$ and observe winner $Y_{k_{t,1},k_{t,2},t}$
11: Statistics update (Algorithm 4)
12: `# arm elimination:`
13: $\hat{k}^{(\mathrm{R})}_{t+1} \leftarrow \arg\max_{\ell\in\mathcal{C}} \hat{\mu}_{\ell,t+1}$
14: $\mathcal{C} \leftarrow \mathcal{C} \setminus \Big\{ k \in \mathcal{C} :$ either $\exists \ell \in \mathcal{C}\setminus\{k\}, \hat{\nu}_{k,\ell,t+1} + \mathrm{CR}^{(\mathrm{D})}_{k,\ell,t+1} < \frac{1}{2}$ or $\hat{\mu}_{k,t+1} + \mathrm{CR}^{(\mathrm{R})}_{k,t+1} \leqslant \hat{\mu}_{\hat{k}^{(\mathrm{R})}_{t+1},t+1} - \mathrm{CR}^{(\mathrm{R})}_{\hat{k}^{(\mathrm{R})}_{t+1},t+1} \Big\}$
15: **else** # {exploit when only the optimal arm left}
16: $k \leftarrow$ the only remaining arm in set $\mathcal{C}$
17: Pull arm k for reward and duel pair (k,k)

3.2. Regret Analysis of ELIMFUSION

Theorem 3.1 provides the regret upper bound of ELIMFUSION. Its proof is deferred to Appendix D.1.

Theorem 3.1. *Letting $\delta \leftarrow 1/K^2T^2$ and taking the expectation, the regret upper bound of* ELIMFUSION *can be upper bounded as follows,*

$$\mathbb{E}[R_T] \leqslant O\left(\sum_{k\neq 1} \frac{(\alpha\Delta^{(\mathrm{R})}_k + (1-\alpha)\Delta^{(\mathrm{D})}_k)\log T}{\max\{(\Delta^{(\mathrm{R})}_k)^2, (\Delta^{(\mathrm{D})}_k)^2/K\}}\right). \quad (5)$$

Applying elimination algorithms to reward and dueling feedback individually without sharing the candidate arm set yields the regret upper bound as follows,

$$\mathbb{E}[R_T] \leqslant O\left(\sum_{k\neq 1} \frac{(\alpha\Delta^{(\mathrm{R})}_k + (1-\alpha)\Delta^{(\mathrm{D})}_k)\log T}{\min\{(\Delta^{(\mathrm{R})}_k)^2, (\Delta^{(\mathrm{D})}_k)^2/K\}}\right), \quad (6)$$

where the denominator is the minimum of the two gap-dependent terms and can be much worse than the maximum in the denominator in (5) of ELIMFUSION. This indicates the improvement of fusing the reward and dueling feedback via sharing the same candidate arm set in ELIMFUSION. We note that the bound in (6) is not as tight as the one in Table 1 for the "No Fusion" algorithm.

While the fusion via the candidate arm set in the elimination algorithm is effective and easy to implement, the $1/K$ factor in the $(\Delta^{(\mathrm{D})}_k)^2/K$ term of (5) is *suboptimal*, which is inherited from the suboptimality of the elimination algorithm in the dueling bandits literature (Saha & Gaillard, 2022, Remark 1). Furthermore, the dependence on the parameter α in (5), compared to the lower bound in (4), is also suboptimal, reflecting the limitation of the elimination algorithm design for more sophisticated optimization in DR-MAB. In the next section, we introduce an advanced algorithm to address these limitations and achieve the optimal regret balance and dependence on the problem parameters.

4. DECOFUSION: Fusing Reward and Dueling via Suboptimal Arm Decomposition

This section presents a novel approach to fusing the reward and dueling feedback, called decomposition fusion (DECOFUSION). We first present the high-level ideas and technical challenges of DECOFUSION in Section 4.1, and then provide its detailed algorithm design in Section 4.2.

Finally, we present the theoretical regret upper bound of DECOFUSION in Section 4.3.

4.1. High-Level Ideas and Technical Challenges

Decomposed exploration for two set of suboptimal arms. We denote $\mathcal{K}^{(\mathrm{R})} := \{k \in \mathcal{K} : \alpha\Delta_k^{(\mathrm{R})}/\mathrm{kl}(\mu_k, \mu_1) < \min_{\ell<k}(1-\alpha)(\Delta_k^{(\mathrm{D})} + \Delta_\ell^{(\mathrm{D})})/\mathrm{kl}\left(\nu_{k,\ell}, \frac{1}{2}\right)\}$ as the subset of suboptimal arms incurring smaller regret from reward feedback, and $\mathcal{K}^{(\mathrm{D})} = \mathcal{K} \setminus \{\mathcal{K}^{(\mathrm{R})} \cup \{1\}\}$ as that of dueling feedback. Then, the RHS of regret lower bound in (3) can be decomposed as $\sum_{k\in\mathcal{K}^{(\mathrm{R})}} \alpha\Delta_k^{(\mathrm{R})}/\mathrm{kl}(\mu_k, \mu_1) + \sum_{k\in\mathcal{K}^{(\mathrm{D})}} \min_{\ell<k}(1-\alpha)(\Delta_k^{(\mathrm{D})} + \Delta_\ell^{(\mathrm{D})})/\mathrm{kl}\left(\nu_{k,\ell}, \frac{1}{2}\right)$. This lower bound decomposition implies that to minimize regret, arms in $\mathcal{K}^{(\mathrm{R})}$ and $\mathcal{K}^{(\mathrm{D})}$ should be explored by reward and dueling feedback, respectively. Inspired by this observation, an algorithm with this minimal regret also needs to decompose arms as above and explore them accordingly, which we call that the algorithm has a *decomposed exploration policy*. However, the above decomposition relies on the reward and dueling gaps $\Delta_k^{(\mathrm{R})}$ and $\Delta_k^{(\mathrm{D})}$, which are unknown a priori. This poses a significant challenge for designing a near-optimal algorithm.

Before delving into the details of our algorithm design, we introduce *"empirical log-likelihoods"* as the measures of the amount of information collected for distinguishing arm k up to time slot t from the reward and dueling feedback, $I_{k,t}^{(\mathrm{R})} := N_{k,t}\,\mathrm{kl}\left(\hat{\mu}_{k,t}, \max_{\ell\in\mathcal{K}}\hat{\mu}_{\ell,t}\right), I_{k,t}^{(\mathrm{D})} := \sum_{\ell\in\mathcal{K}:\hat{\nu}_{k,\ell,t}<\frac{1}{2}} M_{k,\ell,t}\,\mathrm{kl}\left(\hat{\nu}_{k,\ell,t}, \frac{1}{2}\right)$, introduced by Honda & Takemura (2010) and Komiyama et al. (2015) for devising optimal algorithms in stochastic and dueling bandits, respectively. Their algorithms use the conditions of $I_{k,t}^{(\mathrm{R})} \leqslant \log t + f(K)$ and $I_{k,t}^{(\mathrm{D})} - \min_{\ell\in\mathcal{K}} I_{\ell,t}^{(\mathrm{D})} \leqslant \log t + f(K)$ for choosing arms to explore, and the function $f(K)$ is independent of time horizon T and determined later.

Decomposition deadlock. The information measures $I_{k,t}^{(\mathrm{R})}$ and $I_{k,t}^{(\mathrm{D})}$ cannot work with the decomposed exploration policy. Because both definitions involve the full arm set $\mathcal{K}$ and thus need accuracy estimates of all arms. But under the decomposed arm exploration policy, arms in $\mathcal{K}^{(\mathrm{R})}$ only have good estimates of reward mean $\hat{\mu}_{k,t}$, while arms in $\mathcal{K}^{(\mathrm{D})}$ only have good estimates of dueling probability $\hat{\nu}_{k,\ell,t}$, as exploring these arms by feedback other than the corresponding one would incur redundant regrets. This is a "decomposition deadlock": applying the decomposed exploration policy makes the $I_{k,t}^{(\mathrm{R})}$ and $I_{k,t}^{(\mathrm{D})}$-based optimal bandit algorithms design invalid, and deploying the bandit algorithm design makes the decomposed exploration policy invalid. Similar "deadlock" challenges also exist when considering other bandit algorithms, e.g., KL-UCB (Cappé et al., 2013).

4.2. Algorithm Design of DECOFUSION

To address the above challenges, we propose the decomposition fusion (DECOFUSION) algorithm. DECOFUSION, presented in Algorithm 2, consists of three main components: (i) decomposition arm set construction (Lines 5–9), (ii) randomized decision making (Lines 10–22), and (iii) exploration arm set update (Lines 23–27). Specifically, for the challenge of unknown decomposition sets, we conservatively maintain two sets $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ in terms of the collected information instead of the incurred regret (details in Section 4.2.1). However, this set construction leads to a mismatch between the constructed and ground-truth decompositions, which our proposed randomized decision-making strategy in Section 4.2.2 can address. Finally, for the "deadlock" challenge of the conflict between the decomposed exploration policy and the calculation of information measures $I_{k,t}^{(\mathrm{R})}$ and $I_{k,\ell,t}^{(\mathrm{D})}$, in Section 4.2.3, we devise a novel approach to update the exploration arm set $\mathcal{E}$ based on the revised definitions of the information measures $\hat{I}_{k,t}^{(\mathrm{R})}$ and $\hat{I}_{k,t}^{(\mathrm{D})}$ and the approximated arm sets $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$.

4.2.1. DECOMPOSITION ARM SET CONSTRUCTION

After the warm-up (Line 3), in each time slot, DECOFUSION constructs two arm sets, $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$, as analogs (*not* estimates) of the sets $\mathcal{K}^{(\mathrm{R})}$ and $\mathcal{K}^{(\mathrm{D})}$. The reason that $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ are only analogs, not estimates, is because that the lack of knowledge of reward and dueling gaps $\Delta_k^{(\mathrm{R})}$ and $\Delta_k^{(\mathrm{D})}$ makes it hard to decompose arms according to the weighted reward and dueling-based regrets. Instead, we utilize the revised information measures $\hat{I}_{k,t}^{(\mathrm{R})}$ and $\hat{I}_{k,t}^{(\mathrm{D})}$ (defined in Section 4.2.3) to construct the arm sets. Because this information-based set construction does not consider the weight α, the constructed sets $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ mismatches the optimal decomposition sets $\mathcal{K}^{(\mathrm{R})}$ and $\mathcal{K}^{(\mathrm{D})}$, even when the information measures are accurate at the end of the time horizon; thus only analogs. We also note that while the optimal decomposition $\mathcal{K}^{(\mathrm{R})}$ and $\mathcal{K}^{(\mathrm{D})}$, together with the singleton of optimal arm $\{1\}$, forms an exclusive partition of the arm set $\mathcal{K}$, the conservatively constructed sets $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ often overlap, and their intersection $\hat{\mathcal{K}}_t^{(\mathrm{R})} \cap \hat{\mathcal{K}}_t^{(\mathrm{D})}$ contains the optimal arm 1 in almost all time slots.

Specifically, $\hat{\mathcal{K}}_t^{(\mathrm{D})} := \{k \in \mathcal{K} : \hat{I}_{k,t}^{(\mathrm{R})} \leqslant \log t + f(K)\}$ is defined in Line 6, where the condition $\hat{I}_{k,t}^{(\mathrm{R})} \leqslant \log t + f(K)$ implies the insufficient information on arm k from reward feedback. This is a conservative approximation of the dueling arm set $\mathcal{K}^{(\mathrm{D})}$: as long as there is insufficient information from reward feedback to exclude an arm from exploration, that arm is included in dueling feedback's corresponding set. The other set $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ is constructed similarly, as detailed in Line 8. Note that the construction of dueling arm set $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ does not rely on strictly exclusion from the reward arm, and vice versa. Because for the actual decomposition is

Algorithm 2 DECOFUSION: Decomposition Fusion

1: **Input:** Arm set $\mathcal{K}$, parameter α, function $f(K)$
2: **Initialize:** Exploration arm set $\mathcal{E} \leftarrow \mathcal{K}$, time slot $t \leftarrow 0$, decomposition arm set $\hat{\mathcal{K}}_t^{(\mathrm{R})}, \hat{\mathcal{K}}_t^{(\mathrm{D})} \leftarrow \mathcal{K}$, $\hat{I}_{k,t}^{(\mathrm{R})}, \hat{I}_{k,t}^{(\mathrm{D})} \leftarrow 0$ for all arms k, and other initializations in Algorithm 1
3: Warm-up (Algorithm 3)
4: **for** each time slot t **do**
5: # **decomposition arm set update:**
6: $\hat{\mathcal{K}}_t^{(\mathrm{D})} \leftarrow \{k \in \mathcal{K} : \hat{I}_{k,t}^{(\mathrm{R})} \leqslant \log t + f(K)\}$
7: $\hat{k}_t^{(\mathrm{D})} \leftarrow \arg\min_{k \in \hat{\mathcal{K}}_t^{(\mathrm{D})}} \hat{I}_{k,t}^{(\mathrm{D})}$
8: $\hat{\mathcal{K}}_t^{(\mathrm{R})} \leftarrow \left\{k \in \mathcal{K} : \hat{I}_{k,t}^{(\mathrm{D})} - \hat{I}_{\hat{k}_t^{(\mathrm{D})},t}^{(\mathrm{D})} \leqslant \log t + f(K)\right\}$
9: $\hat{k}_t^{(\mathrm{R})} \leftarrow \arg\max_{k \in \hat{\mathcal{K}}_t^{(\mathrm{R})}} \hat{\mu}_{k,t}$
10: # **randomized decision making:**
11: Pick an arm $k_t^{\exp}$ from $\mathcal{E}$ according to a fixed order
12: **if** Uniform$[0,1] > \frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$ **then**
13: # {reward explore, dueling exploit}
14: $k_t \leftarrow k_t^{\exp}$ and $(k_{1,t}, k_{2,t}) \leftarrow (\hat{k}_t^{(\mathrm{R})}, \hat{k}_t^{(\mathrm{R})})$
15: **else** # {dueling explore, reward exploit}
16: $\hat{\mathcal{O}}_{k_t^{\exp},t} \leftarrow \{k \in \hat{\mathcal{K}}_t^{(\mathrm{D})} \setminus \{k_t^{\exp}\} : \hat{\nu}_{k_t^{\exp},k,t} \leqslant \frac{1}{2}\}$
17: **if** $\hat{k}_t^{(\mathrm{D})} \in \hat{\mathcal{O}}_{k_t^{\exp},t}$ or $\hat{\mathcal{O}}_{k_t^{\exp},t} = \emptyset$ **then**
18: $k_t^{\mathrm{duel}} \leftarrow \hat{k}_t^{(\mathrm{D})}$
19: **else** $k_t^{\mathrm{duel}} \leftarrow \arg\min_{k \in \hat{\mathcal{K}}_t^{(\mathrm{D})} \setminus \{k_t^{\exp}\}} \hat{\nu}_{k_t^{\exp},k,t}$
20: $k_t \leftarrow \hat{k}_t^{(\mathrm{D})}$ and $(k_{1,t}, k_{2,t}) \leftarrow (k_t^{\exp}, k_t^{\mathrm{duel}})$
21: Pull arm k_t and observe arm reward $X_{k_t,t}$
22: Duel $(k_{t,1}, k_{t,2})$ and observe winner $Y_{k_{t,1},k_{t,2},t}$
23: # **exploration arm set update:**
24: $\mathcal{E} \leftarrow \mathcal{E} \setminus \{k_t^{\exp}\}$ # {remove the explored arm}
25: $\mathcal{B} \leftarrow \mathcal{B} \cup \{k \in \hat{\mathcal{K}}_t^{(\mathrm{R})} \setminus \mathcal{E} : \hat{I}_{k,t}^{(\mathrm{R})} \leqslant \log t + f(K)\} \cup \{k \in \hat{\mathcal{K}}_t^{(\mathrm{D})} \setminus \mathcal{E} : \hat{I}_{k,t}^{(\mathrm{D})} - \hat{I}_{\hat{k}_t^{(\mathrm{D})},t}^{(\mathrm{D})} \leqslant \log t + f(K)\}$
26: **if** $\mathcal{E}$ is empty **then** # {last $\mathcal{E}$ was traversed}
27: $\mathcal{E} \leftarrow \mathcal{B}$ and $\mathcal{B} \leftarrow \emptyset$ # {renew $\mathcal{E}$}
28: Statistics update (Algorithm 4)
29: $\hat{I}_{k,t+1}^{(\mathrm{R})} \leftarrow N_{k,t}\, \mathrm{kl}(\hat{\mu}_{k,t}, \max_{\ell \in \hat{\mathcal{K}}_t^{(\mathrm{R})}} \hat{\mu}_{\ell,t})$
30: $\hat{I}_{k,t+1}^{(\mathrm{D})} \leftarrow \sum_{\ell \in \hat{\mathcal{K}}_t^{(\mathrm{D})} : \hat{\nu}_{k,\ell,t} < \frac{1}{2}} M_{k,\ell,t}\, \mathrm{kl}(\hat{\nu}_{k,\ell,t}, \frac{1}{2})$

unknown a prior in DECOFUSION, one needs to be conservative (i.e., allow both sets $\hat{\mathcal{K}}_t^{(D)}$ and $\hat{\mathcal{K}}_t^{(R)}$ to have some overlap instead of taking exclusion) to ensure enough explorations from both feedback sides for uncertain arms so to learn the ground-truth decomposition at the end. Especially, we calculate the estimated optimal arms $\hat{k}_t^{(\mathrm{D})}$ and $\hat{k}_t^{(\mathrm{R})}$ among the approximated dueling and reward arm sets, respectively, in Lines 7 and 9. Because arms outside the corresponding sets may not have good estimate accuracy and thus are not considered.

4.2.2. RANDOMIZED DECISION MAKING

To tackle the mismatch between the constructed $(\hat{\mathcal{K}}_t^{(\mathrm{R})}, \hat{\mathcal{K}}_t^{(\mathrm{D})})$ and the ground truth $(\mathcal{K}^{(\mathrm{R})}, \mathcal{K}^{(\mathrm{D})})$ and realize the decomposed exploration policy, we devise a randomized decision-making strategy. In each time slot, DECOFUSION picks one arm $k_t^{\exp}$ from the exploration arm set $\mathcal{E}$ according to some fixed order in Line 11, e.g., the arm with the smallest index in the set. Then, DECOFUSION randomly chooses either the reward or dueling feedback to explore the arm $k_t^{\exp}$ in Line 12. Specifically, if the round's realization of the uniform distribution Uniform$[0,1]$ is greater than a threshold $\frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$, then the algorithm explores the arm $k_t^{\exp}$ by the reward feedback and exploits the pair of the reward estimated optimal arm $(\hat{k}_t^{(\mathrm{R})}, \hat{k}_t^{(\mathrm{R})})$ by the dueling feedback; otherwise, the algorithm explores the arm $k_t^{\exp}$ by the dueling feedback and exploits the dueling estimated optimal arm $\hat{k}_t^{(\mathrm{D})}$ by the reward feedback. When exploring $k_t^{\exp}$ via dueling feedback , we follow the RMED1 algorithm of Komiyama et al. (2015) for dueling bandits to pick the comparison arm k_t^{duel} (Lines 17–19). One key difference is that our comparison arm k_t^{duel} is selected from the set $\hat{\mathcal{K}}_t^{(\mathrm{D})}$ instead of the full arm set $\mathcal{K}$.

Finally, the threshold $\frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$ is chosen to compensate for the mismatch in the decomposition arm set construction. Although chosen by careful derivations, the threshold has an intuitive explanation: take individual squares on all terms in $\frac{\alpha}{\alpha+(1-\alpha)} = \alpha$. The square exponent can be regarded as the compensation for the mismatch between the optimal set $\mathcal{K}^{(\mathrm{R})}$ (omit $\mathcal{K}^{(\mathrm{D})}$) according to the regret with a linear dependence on $1/\Delta_k^{(\mathrm{R})}$ and the constructed set $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ via the collected information $\hat{I}_{k,t}^{(\mathrm{R})}$ with quadratic $(1/\Delta_k^{(\mathrm{R})})^2$ dependence.

4.2.3. EXPLORATION ARM SET CONSTRUCTION

Besides sets $\hat{\mathcal{K}}_t^{(\mathrm{R})}$ and $\hat{\mathcal{K}}_t^{(\mathrm{D})}$, DECOFUSION also maintains an exploration arm set $\mathcal{E}$ and an auxiliary arm set $\mathcal{B}$ for updating $\mathcal{E}$. In each time slot, the exploration arm set $\mathcal{E}$ outputs an arm to explore and removes it after exploration (Line 24), and the auxiliary arm set $\mathcal{B}$ adds arms that need further exploration regarding either the reward or dueling feedback, as detailed in Line 25, and each arm in the set $\mathcal{B}$ is unique and not duplicated. When all arms in $\mathcal{E}$ are explored,

i.e., $\mathcal{E} = \emptyset$, the algorithm renews the exploration arm set $\mathcal{E}$ by the auxiliary arm set $\mathcal{B}$ (Line 27).

At the end of a time slot, DECOFUSION updates the statistics via Algorithm 4 (Line 28) as well as the revised empirical log-likelihoods $\hat{I}^{(\mathrm{R})}_{k,t}$ and $\hat{I}^{(\mathrm{D})}_{k,t}$ in Lines 29 and 30. The key difference between the revised empirical log-likelihoods $\hat{I}^{(\mathrm{R})}_{k,t}$ and $\hat{I}^{(\mathrm{D})}_{k,t}$ and the original ones $I^{(\mathrm{R})}_{k,t}$ and $I^{(\mathrm{D})}_{k,t}$ is that the revised ones only involve the arms in the approximated sets $\hat{\mathcal{K}}^{(\mathrm{R})}_t$ and $\hat{\mathcal{K}}^{(\mathrm{D})}_t$ instead of the full arm set $\mathcal{K}$.

4.3. Regret Analysis of DECOFUSION

Theorem 4.1 (Regret upper bound of Algorithm 2)**.** *For sufficiently small $\xi > 0$, and a constant $c > 0$ depending on the bandit instance,* DECOFUSION *has the regret bound,*

$$\mathbb{E}[R_T] \leqslant O(K^2) + O(\xi^{-2}) + O(e^{cK - f(K)}) + \sum_{k \neq 1} \frac{(\Delta^{(\mathrm{R})}_k/\alpha + \Delta^{(\mathrm{D})}_k/(1-\alpha))((1+\xi)\log T + f(K))}{\max\left\{\mathrm{kl}(\mu_k, \mu_1)/\alpha^2, \mathrm{kl}(\nu_{k,1}, \frac{1}{2})/(1-\alpha)^2\right\}} \quad (7)$$

Simplifying the above bound by letting $T \to \infty$, $f(K) = cK^{1+\xi}$, and noticing $\mathrm{kl}(p, q) = \Theta((p-q)^2)$, we have

$$\mathbb{E}[R_T] \leqslant O\left(\sum_{k \neq 1} \frac{\log T}{\max\{\Delta^{(\mathrm{R})}_k/\alpha, \Delta^{(\mathrm{D})}_k/(1-\alpha)\}}\right). \quad (8)$$

The proof of Theorem 4.1 consists of two key claims (steps 2 and 3 of the proof in Appendix D.2): (i) in most time slots, the estimated optimal arms $\hat{k}^{(\mathrm{R})}_t$ and $\hat{k}^{(\mathrm{D})}_t$ from both feedback types are exactly the optimal arm 1, i.e., $\sum_{t=1}^{T} \mathbb{E}[\hat{k}^{(\mathrm{R})}_t = \hat{k}^{(\mathrm{D})}_t = 1] = T - O(1)$, and (ii) under the first claim, the regret of the randomized decision-making policy is upper bounded by the last regret term in (7). While the claim (i) is proved via similar techniques in Honda & Takemura (2010) and Komiyama et al. (2015), the proof of claim (ii) needs a novel analysis to handle the regret costs of randomized decision-making. It needs to construct a sampling threshold regarding both feedback and then bounds actual sampling times of reward feedback $N_{k,T}$ for each arm and dueling feedback for each pair $M_{k,\ell,T}$ when $\ell = 1$ and $\ell \neq 1$ case-by-case. The choice of the threshold $\frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$ is crucial for balancing the regret costs of both feedback types.

Near-optimal regret when $\ell^*_k = 1$. The simplified regret upper bound in (8) tightly matches the simplified regret lower bound in (4) in terms of all non-trivial factors (except for a universal constant). However, this simplified lower bound only holds when the most effective comparison arm ℓ^*_k is the optimal arm 1 for all arms k. Without this condition, the regret upper bounds in Theorem 4.1 are worse than the general lower bound in (3), where the main gap comes from that DECOFUSION often uses the dueling estimated optimal arm $\hat{k}^{(\mathrm{D})}_t$ to duel with the exploration arm k^{exp}_t, where the $\hat{k}^{(\mathrm{D})}_t$ may not be the most effective comparison arm ℓ^*_k. This issue also exists in the design of optimal dueling bandit algorithms (Komiyama et al., 2015, RMED1 and RMED2), which is only partially resolved by assuming the knowledge of time horizon T in dueling bandit literature.

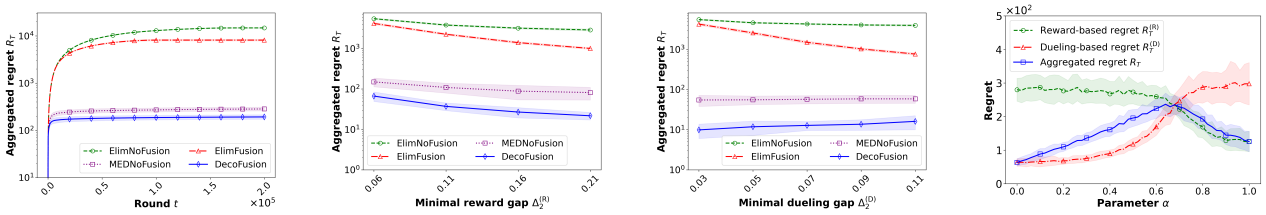
Physical meanings for regret when $\alpha = 0$ or $\alpha = 1$. When either $\alpha = 0$ or $\alpha = 1$, the leading term of the regret bound in (7) vanishes, and the regret becomes a T-independent constant. Because when $\alpha = 0$, the reward-based regret $R^{(\mathrm{R})}_T$ is not counted in the regret, i.e., reward feedback is *free*! In this case, the threshold $\frac{\alpha^2}{\alpha^2+(1-\alpha)^2} = 0$ in the randomized decision-making, and the algorithm always explores arms via reward feedback (i.e., *free exploration*) and exploit the estimated optimal arm $\hat{k}^{(\mathrm{R})}_t$ by the dueling feedback, which only incurs a constant regret. Similarly when $\alpha = 1$, the duel feedback is free, and the algorithm enjoys free exploration from dueling feedback and achieves a constant regret as well. As illustrated in Figure 1, such an advantage is only achieved by DECOFUSION, while ELIMFUSION always suffers $O(\log T)$ regret cost.

5. Experiments

This section reports the numerical experiments of the proposed fusion algorithms. We compare our algorithms with two baselines: 1) ELIMNOFUSION that maintains two separate sets for arm elimination, one based on reward feedback and the other on dueling; 2) MEDNOFUSION (Minimum Empirical Divergence) that deploys the optimal algorithms from Honda & Takemura (2010) and Komiyama et al. (2015) to choose arm k_t and dueling pair $(k_{1,t}, k_{2,t})$ separately.

Figures 2(a), 2(b), and 2(c) illustrate the trends of aggregated regret ($\alpha = 0.5$) in various settings. The setup details are in Appendix E. DECOFUSION outperforms all other algorithms across all settings. In Figure 2(a), the aggregated regret of DECOFUSION is around 41 times lower than that of ELIMFUSION. While ELIMFUSION outperforms ELIMNOFUSION by 81.3%, and DECOFUSION outperforms MEDNOFUSION by 47.5%, both elimination-based approaches are worse than the other two due to the suboptimality of the elimination mechanism. Figures 2(b) and 2(c) examine the impact of reward and dueling gaps on total regret. As either gap increases, distinguishing suboptimal arms becomes easier, thereby reducing the regret cost of the algorithms. The slight increase in the regret of DECOFUSION in Figure 2(c) is due to the regret incurred during the warm-up (Line 3), which scales with the dueling gap and constitutes a large portion of the relatively small total regret.

Figure 2(d) investigates the impact of the parameter α on DECOFUSION. The aggregated regret (blue) reaches its maximum when $\alpha = 0.5$ and decreases as α is close to either 0 or 1, which verifies its free exploration property and



(a) Aggregated regret R_T (b) Vary reward gap $\Delta_2^{(R)}$ (c) Vary dueling gap $\Delta_2^{(D)}$ (d) Vary α in DECOFUSION

Figure 2: Regret comparison in different settings

constant regrets when $\alpha = 0$ or 1. The curves of reward-based and dueling-based regrets (red and green) corroborate the effectiveness of the randomized decision-making policy: when $\alpha = 0$, the algorithm assigns all explorations to reward feedback, while the case of $\alpha = 1$ assigns all explorations to dueling feedback, and one can tune α to balance the regret cost between the two types of feedback.

6. Conclusion

This paper studies the fusion of reward and dueling feedback in stochastic multi-armed bandits, called DR-MAB. We derived regret lower bounds for DR-MAB, and proposed two algorithms, ELIMFUSION and DECOFUSION. ELIMFUSION provides a simple and efficient way to fusion both feedback types via sharing the same candidate arm set, but its regret is suboptimal regarding a multiplicative factor of the number of arms. DECOFUSION, on the other hand, is designed to achieve the optimal regret up to a constant factor, by decoupling the suboptimal arms into two sets for reward and dueling feedback. Both algorithms and the lower bound suggest that the advantage of fusing both feedback types is that it saves the higher regret costs among both feedback types for each suboptimal arm, achieving a better regret than solely relying on any one of the two.

The standard LLM training usually has two separate steps: (1) use the responses from human experts (absolute feedback) to conduct a supervised learning, called supervised fine-tuning (SFT), and (2) use the human preference (relative feedback) among multiple responses generated from the SFT model to conduct a preference directly training (either directed preference optimization (DPO) (Rafailov et al., 2024) or reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022)). Although not exactly the same, these two steps correspond to the reward (absolute) and dueling (relative) feedback in our bandits setting. Our algorithms suggest a training approach to conduct these two steps in parallel, which can potentially save the human labeling cost and improve the training efficiency (as our regret upper bounds improve over the existing ones). Investigating this potential approach in LLM training is an interesting direction.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The work of Jinhang Zuo was supported by CityUHK 9610706. The work of Mohammad Hajiesmaili was CAREER-2045641, CPS-2136199, and CNS-2325956. The work of John C.S. Lui was supported in part by the RGC SRFS2122-4S02. The work of Adam Wierman was supported by NSF grants CCF-2326609, CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648, as well as funding from the Resnick Sustainability Institute. Xutong Liu is the corresponding author.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Ailon, N., Karnin, Z., and Joachims, T. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pp. 856–864. PMLR, 2014.

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

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

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

Bhaskara, A., Gollapudi, S., Im, S., Kollias, K., and Munagala, K. Online learning and bandits with queried hints. *arXiv preprint arXiv:2211.02703*, 2022.

Bubeck, S. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pp. 1516–1541, 2013.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013.

Cutkosky, A., Dann, C., Das, A., and Zhang, Q. Leveraging initial hints for free in stochastic linear bandits. In *International Conference on Algorithmic Learning Theory*, pp. 282–318. PMLR, 2022.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Honda, J. and Takemura, A. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pp. 67–79. Citeseer, 2010.

Ji, X., Wang, H., Chen, M., Zhao, T., and Wang, M. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pp. 1141–1154. PMLR, 2015.

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

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Li, Z., Liu, M., and Lui, J. Fedconpe: Efficient federated conversational bandits with heterogeneous clients. *arXiv preprint arXiv:2405.02881*, 2024.

Lindståhl, S., Proutiere, A., and Johnsson, A. Predictive bandits. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 1170–1176. IEEE, 2020.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Saha, A. and Gaillard, P. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pp. 19011–19026. PMLR, 2022.

Saha, A., Koren, T., and Mansour, Y. Adversarial dueling bandits. In *International Conference on Machine Learning*, pp. 9235–9244. PMLR, 2021.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237. doi: 10.1561/2200000068. URL http://dx.doi.org/10.1561/2200000068.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. Advancements in dueling bandits. In *IJCAI*, pp. 5502–5510, 2018.

Sutton, R. S., Barto, A. G., et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Szepesvári, C. *Algorithms for reinforcement learning*. Springer nature, 2022.

Urvoy, T., Clerot, F., Féraud, R., and Naamane, S. Generic exploration and k-armed voting bandits. In *International conference on machine learning*, pp. 91–99. PMLR, 2013.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023a.

Wang, Z., Liu, X., Li, S., and Lui, J. C. Efficient explorative key-term selection strategies for conversational contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10288–10295, 2023b.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for

rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Yan, X., Luo, C., Clarke, C. L., Craswell, N., Voorhees, E. M., and Castells, P. Human preferences as dueling bandits. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 567–577, 2022.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yun, D., Proutiere, A., Ahn, S., Shin, J., and Yi, Y. Multi-armed bandit with additional observations. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–22, 2018.

Zhang, X., Xie, H., Li, H., and CS Lui, J. Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020*, pp. 662–672, 2020.

A. Related Works

Stochastic bandits with absolute or relative feedback. Stochastic multi-armed bandits (MAB) is a fundamental online learning model, introduced by Lai & Robbins (1985) and later comprehensively studied, e.g., by Bubeck et al. (2012); Slivkins (2019); Lattimore & Szepesvári (2020). Later on, various extensions of the stochastic bandits have been proposed, such as the linear bandits (Abbasi-Yadkori et al., 2011), contextual bandits (Li et al., 2010), and combinatorial bandits (Chen et al., 2013). The relative feedback in the stochastic bandits setting is known as dueling bandits, which is initialized by Yue et al. (2012) and studied by Ailon et al. (2014); Komiyama et al. (2015); Sui et al. (2018); Saha & Gaillard (2022), etc. All of the above models assume that the learner receives either the absolute feedback, i.e., the reward of the selected arm, or the relative feedback, i.e., the winning arm in a pair of arms, but not both, which are different from the study on fusion of absolute and relative feedback in this paper.

Bandits with additional information. While we are the first to study the fusion of absolute and relative feedback in the stochastic bandits setting, there are prior works considering the absolute (reward) feedback with some types of "augmentation" feedback. In the line of works of conversational bandits (Zhang et al., 2020; Wang et al., 2023b; Li et al., 2024), a learner interacts with a contextual linear bandit model, and besides receiving the reward feedback from puling arms, the learner can also occasionally query "key-terms" to collect some side information of the bandit model. Another line of works considers the bandits with "hints", where the "hint" can be additional reward observations (Yun et al., 2018; Lindståhl et al., 2020), initial reward mean guesses (Cutkosky et al., 2022), or the order of the reward realizations among two or more arms (Bhaskara et al., 2022), etc. However, these types of augmentation feedback provide information directly related to the reward mean parameters (i.e., used to estimate the reward means), which is different from the relative feedback depending on dueling probabilities in the dueling bandits setting, and therefore, they are different from our study on the fusion of absolute and relative feedback in this paper.

Beyond stochastic bandits. The stochastic bandit literature is only a small subset of the wide online learning research area (Orabona, 2019). In contrast to the stochastic setting, the adversarial bandits—including adversarial MAB (Auer et al., 2002) and adversarial dueling bandits (Saha et al., 2021)—assume that the environment is determined by an adversary and may adapt to the learner's strategy. Generalizing the bandits setting with state transition, reinforcement learning (RL) (Sutton et al., 1998) is another popular online learning model, where the learner interacts with the environment and receives reward feedback based on the state-action pairs. Noticeably, there is a reinforcement learning with human feedback (RLHF) model (Wang et al., 2023a; Ouyang et al., 2022) in RL literature, which can be regarded as a counterpart of the dueling bandits in bandits literature, where the learner receives pairwise comparison (relative) feedback. While there is vast online learning literature beyond the stochastic bandits, due ot the importance of the stochastic MAB, our study on the fusion of the absolute and relative feedback on bandits is a novel and unexplored topic.

B. Deferred Pseudo-code

This section presents the pseudo-code of initial warm-up phase (Algorithm 3) and the statistics update phase (Algorithm 4) of the proposed DR-MAB algorithms.

Algorithm 3 Warm-up (Initial phase)

1: Duel each pair of arms (k, ℓ) for $k \neq \ell$ once (in total $K(K-1)$ times), meanwhile, query each arm for rewards in a round-robin manner
2: $t \leftarrow K(K-1)$
3: Update $M_{k,\ell,t}$ and $\hat{\nu}_{k,\ell,t}$ for all pairs $(k, \ell) \in \mathcal{K} \times \mathcal{K}$ $(k \neq \ell)$, and $N_{k,t}$, $\hat{\mu}_{k,t}$ for all arms $k \in \mathcal{K}$

C. Proof of Lower Bound

Proof of Lemma 2.2. **Step 1. Instance and event construction.** Pick any suboptimal arm $k \neq 1$. We use the standard parameters defined in model section as the original instance $\mathcal{I}$ and then construct an alternative instances $\mathcal{I}'$ with parameters with a prime, e.g., μ'_k and $\nu'_{k,\ell}$, as follows,

- Reward means $\mu'_\ell = \begin{cases} \mu_\ell & \text{if } \ell \neq k \\ \mu_1 + \epsilon & \text{if } \ell = k \end{cases}$,

Algorithm 4 Statistics update

for all arm pairs $(k, \ell) \in \mathcal{K} \times \mathcal{K}$ **do**
  if $(k, \ell) = (k_{t,1}, k_{t,2})$ or $(k_{t,2}, k_{t,1})$ **then**
    $M_{k,\ell,t+1} \leftarrow M_{k,\ell,t} + 1$
    $\hat{\nu}_{k,\ell,t+1} \leftarrow \frac{\hat{\nu}_{k,\ell,t} M_{k,\ell,t} + \mathbb{1}\{Y_{k,\ell,t} = k\}}{M_{k,\ell,t+1}}$
  else
    $M_{k,\ell,t+1} \leftarrow M_{k,\ell,t}$ and $\hat{\nu}_{k,\ell,t+1} \leftarrow \hat{\nu}_{k,\ell,t}$
for all arms $k \in \mathcal{K}$ **do**
  if $k = k_t$ **then**
    $N_{k,t+1} \leftarrow N_{k,t} + 1$
    $\hat{\mu}_{k,t+1} \leftarrow \frac{\hat{\mu}_{k,t} N_{k,t} + X_{k_t,t}}{N_{k,t+1}}$
  else
    $N_{k,t+1} \leftarrow N_{k,t}$ and $\hat{\mu}_{k,t+1} \leftarrow \hat{\mu}_{k,t}$

- Dueling probabilities $\nu_{\ell_1,\ell_2} = \begin{cases} \frac{1}{2} + \epsilon_{\ell_2} & \text{if } \ell_1 = k \text{ and } \ell_2 < k \\ \frac{1}{2} - \epsilon_{\ell_1} & \text{if } \ell_1 < k \text{ and } \ell_2 = k \\ \nu_{\ell_1,\ell_2} & \text{otherwise} \end{cases}$,

where the $\epsilon > 0$ is a small constant to be determined later, and the ϵ_ℓ parameters are chosen such that $\mathrm{kl}(\nu_{k,\ell}, \frac{1}{2} + \epsilon_\ell) = \mathrm{kl}(\nu_{k,\ell}, \frac{1}{2}) + \epsilon$. Under this instance construction, the optimal arms are arm 1 and arm k in the original and alternative instances, respectively. All reward distributions are Bernoulli. We denote $\mathbb{E}, \mathbb{P}$ and $\mathbb{E}', \mathbb{P}'$ as the expectation and probability under the original and alternative instances, respectively.

To facilitate the rest of the analysis, we define the empirical KL-divergence as follows,

$$\begin{aligned}
\widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(n) &\coloneqq \sum_{s=1}^{n} \log\left(\frac{Y_{k,\ell,(s)}\nu_{k,\ell} + (1 - Y_{k,\ell,(s)})(1 - \nu_{k,\ell})}{Y_{k,\ell,(sn)}\nu'_{k,\ell} + (1 - Y_{k,\ell,(s)})(1 - \nu'_{k,\ell})}\right), \\
\widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n) &\coloneqq \sum_{s=1}^{n} \log\left(\frac{X_{k,(s)}\mu_k + (1 - X_{k,(s)})(1 - \mu_k)}{X_{k,(s)}\mu'_k + (1 - X_{k,(s)})(1 - \mu'_k)}\right), \\
\widehat{\mathrm{KL}}(\mathcal{I}, \mathcal{I}') &\coloneqq \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(N_{k,T}) + \sum_{\ell<k} \widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(M_{k,\ell,T}),
\end{aligned}$$

where the subscript (s) refers to the s^{th} observation of the corresponding random variable, which differs from the time index. With these empirical KL-divergences defined, for any event $\mathcal{E}$, we have the following relation holds (change of measure),

$$\mathbb{P}'(\mathcal{E}) = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\exp(-\widehat{\mathrm{KL}}(\mathcal{I}, \mathcal{I}'))\right].$$

In the end of the first step, we define two events as follows,

$$\begin{aligned}
\mathcal{D}_1 &\coloneqq \left\{\sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) M_{k,\ell,T} + \mathrm{kl}(\mu_k, \mu'_k) N_{k,T} < (1 - \epsilon)\log T\right\}, \\
\mathcal{D}_2 &\coloneqq \left\{\widehat{\mathrm{KL}}(\mathcal{I}, \mathcal{I}') \leqslant \left(1 - \frac{\epsilon}{2}\right)\log T\right\}.
\end{aligned}$$

In the remaining of this proof, we use two steps to prove that $\mathbb{P}(\mathcal{D}_1) = o(1)$, which then implies the lemma.

Step 2. Prove $\mathbb{P}(\mathcal{D}_1 \cap \mathcal{D}_2) = o(1)$. We first apply the change of measure argument to transfer the measure of the event $\mathcal{D}_1 \cap \mathcal{D}_2$ from instance $\mathcal{I}$ to instance $\mathcal{I}'$,

$$\mathbb{P}'(\mathcal{D}_1 \cap \mathcal{D}_2) = \mathbb{E}[\mathbb{1}\{\mathcal{D}_1 \cap \mathcal{D}_2\}\exp(-\widehat{\mathrm{KL}}(\mathcal{I}, \mathcal{I}'))] \geqslant \mathbb{E}[\mathbb{1}\{\mathcal{D}_1 \cap \mathcal{D}_2\}]T^{-(1-\frac{\epsilon}{2})},$$

which yields

$$\begin{aligned}
\mathbb{P}(\mathcal{D}_1 \cap \mathcal{D}_2) &\leqslant T^{1-\frac{\epsilon}{2}} \mathbb{P}'(\mathcal{D}_1 \cap \mathcal{D}_2) \\
&\leqslant T^{1-\frac{\epsilon}{2}} \mathbb{P}' \left(N_{k,T} < \frac{(1-\epsilon)\log T}{\mathrm{kl}(\mu_k, \mu'_k)} \right) \\
&= T^{1-\frac{\epsilon}{2}} \mathbb{P}' \left(T - N_{k,T} > T - \frac{(1-\epsilon)\log T}{\mathrm{kl}(\mu_k, \mu'_k)} \right) \\
&\leqslant T^{1-\frac{\epsilon}{2}} \frac{T - \mathbb{E}'[N_{k,T}]}{T - ((1-\epsilon)\log T / \mathrm{kl}(\mu_k, \mu'_k))} \qquad (9) \\
&= T^{1-\frac{\epsilon}{2}} \frac{\sum_{\ell \neq k} \mathbb{E}'[N_{\ell,T}]}{T - ((1-\epsilon)\log T / \mathrm{kl}(\mu_k, \mu'_k))} \\
&\leqslant o(T^{\gamma - \frac{\epsilon}{2}}) = o(1), \qquad (10)
\end{aligned}$$

where inequality (9) is due to the Markov inequality, and inequality (10) is due to the consistent policy assumption (Definition 2.1) that for any suboptimal arm $\ell \neq k$ under instance $\mathcal{I}'$, we have $\mathbb{E}'[N_{\ell,T}] = o(T^{\gamma})$ for any positive γ.

Step 3. Prove $\mathbb{P}(\mathcal{D}_1 \setminus \mathcal{D}_2) = o(1)$. To prove this claim, we first telescope the $\mathbb{P}(\mathcal{D}_1 \setminus \mathcal{D}_2)$ and focus on bounding of the empirical KL-divergence as follows,

$$\begin{aligned}
&\mathbb{P}(\mathcal{D}_1 \setminus \mathcal{D}_2) \\
&= \mathbb{P}\left(\sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) M_{k,\ell,T} + \mathrm{kl}(\mu_k, \mu'_k) N_{k,T} < (1-\epsilon)\log T,\ \widehat{\mathrm{KL}}(\mathcal{I}, \mathcal{I}') > \left(1 - \frac{\epsilon}{2}\right)\log T \right) \\
&\leqslant \mathbb{P}\left(\sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) M_{k,\ell,T} + \mathrm{kl}(\mu_k, \mu'_k) N_{k,T} < (1-\epsilon)\log T, \right. \\
&\qquad \left. \max_{\substack{n_k, m_\ell \in \mathbb{N}^+, \forall \ell<k: \sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) m_\ell \\ + \mathrm{kl}(\mu_k, \mu'_k) n_k < (1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k) + \sum_{\ell<k} \widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell) > \left(1 - \frac{\epsilon}{2}\right)\log T \right) \\
&\leqslant \mathbb{P}\left(\max_{\substack{n_k, m_\ell \in \mathbb{N}^+, \forall \ell<k: \sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) m_\ell \\ + \mathrm{kl}(\mu_k, \mu'_k) n_k < (1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k) + \sum_{\ell<k} \widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell) > \left(1 - \frac{\epsilon}{2}\right)\log T \right).
\end{aligned}$$

Next, we a variant of the maximal law of large numbers (LLN) (Bubeck, 2010, Lemma 10.5) to bound the empirical KL-divergence. To guarantee that all sample times $N_{k,T}$ and $M_{k,\ell,T}$ are large enough to apply LLN, we set the lower bound of $N_{k,T}$ and $M_{k,\ell,T}$ as $\delta \log T$ for some small constant $\delta > 0$ as follows,

$$\begin{aligned}
&\frac{\max_{\substack{n_k, m_\ell \in \mathbb{N}^+, \forall \ell<k: \sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) m_\ell \\ + \mathrm{kl}(\mu_k, \mu'_k) n_k < (1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k) + \sum_{\ell<k} \widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell)}{\log T} \\
&\leqslant \frac{\max_{\substack{n_k, m_\ell \in \mathbb{N}^+, n_k, m_\ell > \delta \log T, \forall \ell<k,: \\ \sum_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right) m_\ell + \mathrm{kl}(\mu_k, \mu'_k) n_k < (1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k) + \sum_{\ell<k} \widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell)}{\log T} \qquad (11) \\
&\qquad + \frac{\delta(k-1)}{\min_{\ell<k} \mathrm{kl}\left(\nu_{k,\ell}, \nu'_{k,\ell}\right)} + \frac{\delta}{\mathrm{kl}(\mu_k, \mu'_k)}.
\end{aligned}$$

Then, because the maximal LLM (Bubeck, 2010, Lemma 10.5) implies $\lim_{N\to\infty} \max_{1\leqslant n\leqslant N} \frac{\widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n)}{N} = \mathrm{kl}(\mu_k, \mu'_k)$ and $\lim_{M\to\infty} \max_{1\leqslant m\leqslant M} \frac{\widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m)}{M} = \mathrm{kl}(\nu_{k,\ell}, \nu'_{k,\ell})$ for any arm $\ell < k$, almost surely (a.s.), we bound the first term in the

right-hand side of (11) by the maximal LLN as follows,

$$\limsup_{T\to\infty} \frac{\max_{\substack{n_k,m_\ell\in\mathbb{N}^+, n_k,m_\ell>\delta\log T,\forall \ell<k,: \\ \sum_{\ell<k}\mathrm{kl}\left(\nu_{k,\ell},\nu'_{k,\ell}\right)m_\ell+\mathrm{kl}(\mu_k,\mu'_k)n_k<(1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k)+\sum_{\ell<k}\widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell)}{\log T} \leqslant 1-\epsilon. \tag{12}$$

Therefore, combining (11) and (12), we have

$$\frac{\max_{n_k,m_\ell\in\mathbb{N}^+,\forall\ell<k:\sum_{\ell<k}\mathrm{kl}\left(\nu_{k,\ell},\nu'_{k,\ell}\right)m_\ell \atop +\mathrm{kl}(\mu_k,\mu'_k)n_k<(1-\epsilon)\log T} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k)+\sum_{\ell<k}\widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell)}{\log T} \leqslant 1-\epsilon+\Theta(\delta),\ \text{a.s.}$$

Noticing $1-\epsilon<1-\frac{\epsilon}{2}$ and letting $\delta\to 0$, we proved that

$$\begin{aligned}\mathbb{P}(\mathcal{D}_1\setminus\mathcal{D}_2) &\leqslant \mathbb{P}\left(\max_{\substack{n_k,m_\ell\in\mathbb{N}^+,\forall\ell<k:\sum_{\ell<k}\mathrm{kl}\left(\nu_{k,\ell},\nu'_{k,\ell}\right)m_\ell \\ +\mathrm{kl}(\mu_k,\mu'_k)n_k<(1-\epsilon)\log T}} \widehat{\mathrm{KL}}_k^{(\mathrm{R})}(n_k)+\sum_{\ell<k}\widehat{\mathrm{KL}}_\ell^{(\mathrm{D})}(m_\ell) > \left(1-\frac{\epsilon}{2}\right)\log T\right) \\ &= o(1).\end{aligned}$$

Lastly, by letting ϵ be infinitesimally small, we have $\mathbb{P}(\mathcal{D}_1)=o(1)$, which implies the lemma. □

Proof of Theorem 2.3 (regret lower bound). We first present the regret decomposition in terms of the reward and dueling feedback as follows,

$$R_T=\sum_{k>1}\left(\alpha\Delta_k^{(\mathrm{R})}N_{k,T}+\sum_{\ell<k}(1-\alpha)\left(\Delta_k^{(\mathrm{D})}+\Delta_\ell^{(\mathrm{D})}\right)M_{k,\ell,T}\right). \tag{13}$$

Then, with Lemma 2.2, its corresponding regret of each suboptimal arm $k\neq 1$ can be lower bounded by the following optimizing problem,

$$\begin{aligned}\min\quad & \alpha\Delta_k^{(\mathrm{R})}N_{k,T}+\sum_{\ell<k}(1-\alpha)\left(\Delta_k^{(\mathrm{D})}+\Delta_\ell^{(\mathrm{D})}\right)M_{k,\ell,T} \\ \text{s.t.}\quad & \sum_{\ell<k}\mathrm{kl}\left(\nu_{k,\ell},\frac{1}{2}\right)M_{k,\ell,T}+\mathrm{kl}(\mu_k,\mu_1)N_{k,T}\geqslant(1-o(1))\log T,\end{aligned}$$

whose asymptotical solution is as follows,

$$\liminf_{T\to\infty}\frac{\mathbb{E}[\Delta_k^{(\mathrm{R})}N_{k,T}+\sum_{\ell<k}(1-\alpha)\left(\Delta_k^{(\mathrm{D})}+\Delta_\ell^{(\mathrm{D})}\right)M_{k,\ell,T}]}{\log T}\geqslant\min\left\{\frac{\alpha\Delta_k}{\mathrm{kl}(\mu_k,\mu_1)},\min_{\ell<k}\frac{(1-\alpha)(\Delta_k^{(\mathrm{D})}+\Delta_\ell^{(\mathrm{D})})}{\mathrm{kl}\left(\nu_{k,\ell},\frac{1}{2}\right)}\right\}.$$

Lastly, we substitute the above lower bound for each terms in (13) and obtain the regret lower bound. □

D. Proof of Upper Bounds

D.1. Regret Upper Bound for ELIMFUSION (Algorithm 1, Theorem 3.1)

Proof of Theorem 3.1. **Step 1. Bound small probability event.** We first show the concentration of the reward and dueling probabilities as follows,

$$\begin{aligned}\mathbb{P}\left(|\hat{\mu}_{k,t}-\mu_k| \geqslant \sqrt{\frac{2\log(Kt/\delta)}{N_{k,t}}}\right) &= \sum_{n=1}^{t}\mathbb{P}\left(|\hat{\mu}_{k,t}-\mu_k| \geqslant \sqrt{\frac{2\log(Kt/\delta)}{n}}\,\middle|\, N_{k,t}=n\right)\mathbb{P}\left(N_{k,t}=n\right) \qquad (14)\\ &\leqslant \sum_{n=1}^{t}\mathbb{P}\left(|\hat{\mu}_{k,t}-\mu_k| \geqslant \sqrt{\frac{2\log(Kt/\delta)}{n}}\right)\\ &\leqslant \sum_{n=1}^{t} 2\exp\left(-4\log(Kt/\delta)\right) \qquad (15)\\ &\leqslant \frac{2\delta^4}{K^4t^3}\end{aligned}$$

where inequality (14) follows from the formula of total probability, and inequality (15) follows from Hoeffding's inequality.

Therefore, we have

$$\mathbb{P}\left(\forall t\in\mathcal{T}, k\in\mathcal{K}, |\hat{\mu}_{k,t}-\mu_k| \geqslant \sqrt{\frac{2\log(Kt/\delta)}{N_{k,t}}}\right) \overset{(a)}{\leqslant} \sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\delta^4}{K^4t^3} \leqslant \frac{\delta^4}{K^3}\sum_{t=1}^{T}\frac{1}{t^3} \leqslant \frac{3\delta^4}{2K^3}, \qquad (16)$$

where inequality (a) follows from the union bound and the above concentration.

With a similar argument, we can show that the dueling probabilities also concentrate as follows,

$$\mathbb{P}\left(\forall t\in\mathcal{T}, k,\ell\in\mathcal{K}, |\hat{\nu}_{k,\ell,t}-\nu_{k,\ell}| \geqslant \sqrt{\frac{2\log(Kt/\delta)}{M_{k,\ell,t}}}\right) \leqslant \frac{3\delta^4}{4K^2},$$

where the RHS's dominator different from (16) is because the number of arm pairs is $K(K-1)/2$.

Step 2. Bound the maximal sample times for both types of feedback.

Elimination with reward feedback. One suboptimal arm k is eliminated by reward feedback if the following event happens,

$$\hat{\mu}_{1,t}-\hat{\mu}_{k,t} \geqslant \mu_1-\mu_k-2\sqrt{\frac{2\log(Kt/\delta)}{N_{k,t}}} > 2\sqrt{\frac{2\log(Kt/\delta)}{N_{k,t}}},$$

which would happen on or before the rearranged expression as follows,

$$N_{k,t} > \frac{32\log(Kt/\delta)}{(\Delta_k^{(\mathrm{R})})^2}.$$

In other words, from the reward feedback and its corresponding elimination, the sample times of arm k are upper bounded by $\frac{32\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2}$.

Elimination with dueling feedback. The sample times for an arm k from the dueling feedback is given by noticing that for any arm $\ell<k$, when

$$\hat{\nu}_{k,\ell,t}+\sqrt{\frac{2\log(Kt/\delta)}{M_{k,\ell,t}}} \leqslant \nu_{k,\ell}+2\sqrt{\frac{2\log(Kt/\delta)}{M_{k,\ell,t}}} < \frac{1}{2},$$

that is,

$$M_{k,\ell,t} \geqslant \frac{8\log(Kt/\delta)}{(\Delta_k^{(\mathrm{D})})^2},$$

for all arms ℓ, then the arm k must have been eliminated from the candidate arm set $\mathcal{C}$ by Line 14 of Algorithm 1. Therefore, the sampling times of arm pair (k, ℓ) is upper bounded by $\frac{8\log(KT/\delta)}{(\Delta_k^{(\mathrm{D})})^2}$.

Step 3. Bound the regret. Each suboptimal arm k may be eliminated by either reward or dueling feedback. So, the regret incurred by the reward querying of arm k is upper bounded as follows,

$$R_{k,T}^{(\mathrm{R})} \leqslant \Delta_k^{(\mathrm{R})} \cdot \min\left\{\frac{32\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2}, \frac{K-1}{2}\cdot\frac{8\log(KT/\delta)}{(\Delta_k^{(\mathrm{D})})^2}\right\} = \frac{4\Delta_k^{(\mathrm{R})}\log(KT/\delta)}{\max\{(\Delta_k^{(\mathrm{R})})^2/8, (\Delta_k^{(\mathrm{D})})^2/4(K-1)\}},$$

where the factor $\frac{K-1}{2}$ is an upper bound of the ratio of sampling collection rate between reward and dueling feedback because the number of arms in the candidate arm set $\mathcal{C}$ is at most $\frac{K-1}{2}$ times smaller than the number of arm pairs in the set.

Next, we bound the regret incurred by the dueling querying of arm k. If the arm k is eliminated by the dueling feedback, then the dueling regret due to arm k can be upper bounded as follows,

$$\sum_{\ell<k}\frac{(\Delta_\ell^{(\mathrm{D})}+\Delta_k^{(\mathrm{D})}-1)}{2}\cdot M_{\ell,k,T} \leqslant \sum_{\ell<k}\Delta_k^{(\mathrm{D})}\frac{8\log(KT/\delta)}{(\Delta_k^{(\mathrm{D})})^2} \leqslant \frac{8(k-1)\log(KT/\delta)}{\Delta_k^{(\mathrm{D})}}.$$

If the arm k is eliminated by the reward feedback, implying the algorithm has queried the reward of arm k for at most $\frac{32\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2}$ times. During this period, dueling regret cost of any arm pair (k, ℓ) from some arm $\ell < k$ is upper bounded as follows,

$$\frac{(\Delta_\ell^{(\mathrm{D})}+\Delta_k^{(\mathrm{D})}-1)}{2}\times 2\times\frac{32\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2} = \frac{64\Delta_k^{(\mathrm{D})}\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2},$$

where the factor 2 every reward query of arm k is accompanied by at most two dueling comparison of arm pairs (k, ℓ) from some arm $\ell < k$.

Taking the minimal among the above two cases, we have the dueling regret upper bound as follows,

$$R_{k,T}^{(\mathrm{D})} \leqslant \Delta_k^{(\mathrm{D})} \cdot \min\left\{\frac{64\log(KT/\delta)}{(\Delta_k^{(\mathrm{R})})^2}, \frac{8(k-1)\log(KT/\delta)}{(\Delta_k^{(\mathrm{D})})^2}\right\} = \frac{8\Delta_k^{(\mathrm{D})}\log(KT/\delta)}{\max\{(\Delta_k^{(\mathrm{R})})^2/8, (\Delta_k^{(\mathrm{D})})^2/(k-1)\}}.$$

Lastly, we bound the total regret as follows,

$$\begin{aligned}
R_T &\leqslant \sum_{k\neq 1}\alpha R_{k,T}^{(\mathrm{R})} + (1-\alpha)R_{k,T}^{(\mathrm{D})} \\
&\leqslant \sum_{k\neq 1}\frac{4\alpha\Delta_k^{(\mathrm{R})}\log(KT/\delta)}{\max\{(\Delta_k^{(\mathrm{R})})^2/8, (\Delta_k^{(\mathrm{D})})^2/(K-1)\}} + \frac{8(1-\alpha)\Delta_k^{(\mathrm{D})}\log(KT/\delta)}{\max\{(\Delta_k^{(\mathrm{R})})^2/8, (\Delta_k^{(\mathrm{D})})^2/(k-1)\}} \\
&\leqslant \sum_{k\neq 1}\frac{8(\alpha\Delta_k^{(\mathrm{R})}+(1-\alpha)\Delta_k^{(\mathrm{D})})\log(KT/\delta)}{\max\{(\Delta_k^{(\mathrm{R})})^2/8, (\Delta_k^{(\mathrm{D})})^2/(K-1)\}}.
\end{aligned}$$

□

D.2. Regret Upper Bound for DECOFUSION (Algorithm 2, Theorem 4.1)

Proof of Regret Upper Bound of Algorithm 2. **Step 1. Event definition.** We start the proof by defining the following events,

$$\begin{aligned}
\mathcal{A}_t &:= \left\{ \hat{\mu}_{1,t} \geqslant \hat{\mu}_{k,t}, \forall k \in \hat{\mathcal{K}}_t^{(\mathrm{R})}, \hat{\mu}_{1,t} \geqslant \mu_1 - \delta, \text{ and } \hat{\nu}_{1,k,t} \geqslant \frac{1}{2}, \forall k \in \hat{\mathcal{K}}_t^{(\mathrm{D})} \right\}, \\
\mathcal{G}_t &:= \left\{ \hat{\mu}_{k_t^{(\mathrm{R})},t} \leqslant \max_{k \in \hat{\mathcal{K}}_t^{(\mathrm{R})} \setminus \{1\}} \mu_k + \delta \right\}, \\
\mathcal{H}_t &:= \bigcup_{k \in \hat{\mathcal{K}}_t^{(\mathrm{R})}} \left\{ \hat{\mu}_{k_t^{(\mathrm{R})},t} = \hat{\mu}_{k,t} \text{ and } |\hat{\mu}_{k,t} - \mu_k| \geqslant \delta \right\}, \\
\mathcal{U}_t &:= \bigcup_{\mathcal{S} \in 2^{\hat{\mathcal{K}}_t^{(\mathrm{D})}} \setminus \{\emptyset\}} \left\{ \hat{\nu}_{1,k,t} \geqslant \frac{1}{2}, \forall k \in \mathcal{S} \text{ and } \hat{\nu}_{1,k,t} < \frac{1}{2}, \forall k \in \hat{\mathcal{K}}_t^{(\mathrm{D})} \setminus \mathcal{S} \right\},
\end{aligned}$$

where the parameter δ is a small positive constant that satisfies $0 < \delta < \Delta_2^{(\mathrm{R})}$. The $\mathcal{A}_t$ refers to a good event, implying that the algorithm has correctly estimated the optimal arm at time slot t, i.e., $\hat{k}_t^{(\mathrm{R})} = \hat{k}_t^{(\mathrm{D})} = 1$, and the other three refer to the bad events that the algorithm may not correctly estimate the optimal arm, either due to bad reward estimates or bad dueling probability estimates.

Notice that (1) if both events $\mathcal{G}_t$ and $\mathcal{H}_t$ do not happen, i.e., $\mathcal{G}_t^{\mathrm{C}} \cap \mathcal{H}_t^{\mathrm{C}}$, then the first condition of event $\mathcal{A}_t$ holds, and (2) if event $\mathcal{U}_t$ does not happen, then the second condition of event $\mathcal{A}_t$ holds. Putting these together, we have the following relation between these events,

$$\mathcal{G}_t^{\mathrm{C}} \cap \mathcal{H}_t^{\mathrm{C}} \cap \mathcal{U}_t^{\mathrm{C}} \subseteq \mathcal{A}_t, \text{ that is, } \mathcal{A}_t^{\mathrm{C}} \subseteq \mathcal{G}_t \cup \mathcal{H}_t \cup \mathcal{U}_t.$$

In the rest of this proof, we show two claims: (1) the number of times that any of these three bad events $\mathcal{G}_t$, $\mathcal{H}_t$, and $\mathcal{U}_t$ happens is bounded by $O(1)$ (a term independent of time horizon T), and (2) for the time slots that the good event $\mathcal{A}_t$ happens, the regret of DECOFUSION is upper bounded as (7) shows.

Step 2. Bound the number of happening times of the bad events. Following Honda & Takemura (2010, Lemmas 16 and 17), we bound $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{G}_t\}\right] \leqslant O(1)$, and $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{H}_t\}\right] \leqslant O(1)$. Following Komiyama et al. (2015, Lemma 5), we bound $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}_t\}\right] \leqslant O(e^{K-f(K)})$. Putting these results together, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_t^C\}\right] &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{G}_t\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{H}_t\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}_t\}\right] \\
&\leqslant O(1) + O(1) + O(e^{cK-f(K)}) \leqslant O(e^{cK-f(K)}),
\end{aligned}$$

for some constant $c > 0$.

Step 3. Bound regret. Denote $\beta := \frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$ as the threshold in Line 12 of Algorithm 2. We first define the following three quantities as the sufficient number of times for exploring arm k from different types of feedback, for sufficiently small $\epsilon > 0$,

$$\begin{aligned}
N_k^{(\mathrm{Suff})} &:= \frac{(1+\epsilon)\log T + f(K)}{\mathrm{kl}(\mu_k, \mu_1)}, \\
M_{k,1}^{(\mathrm{Suff})} &:= \frac{(1+\epsilon)\log T + f(K)}{\mathrm{kl}(\nu_{k,1}, \frac{1}{2})}, \\
L_k^{(\mathrm{Suff})} &:= \min\left\{ N_{k,t}^{(\mathrm{Suff})}/(1-\beta), M_{k,1,t}^{(\mathrm{Suff})}/\beta \right\} = \frac{(1+\epsilon)\log T + f(K)}{\max\left\{(1-\beta)\,\mathrm{kl}(\mu_k, \mu_1), \beta\,\mathrm{kl}(\nu_{k,1}, \frac{1}{2})\right\}}.
\end{aligned}$$

Next, we bound the expected number of times of exploring arm k for different types of feedback as follows,

$$\mathbb{E}[N_{k,T}] \leqslant K/2 + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_t^C\}\right] + (1-\beta)\cdot\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{J}_{k,t} \text{ and } \mathcal{A}_t\}\right], \tag{17}$$

$$\mathbb{E}[M_{k,1,T}] \leqslant 1 + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_t^C\}\right] + \beta\cdot\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{J}_{k,t} \text{ and } \mathcal{A}_t\}\right], \tag{18}$$

$$\mathbb{E}[M_{k,\ell,T}] \leqslant 1 + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_t^C\}\right], \quad \forall 1 < \ell \leqslant k, \tag{19}$$

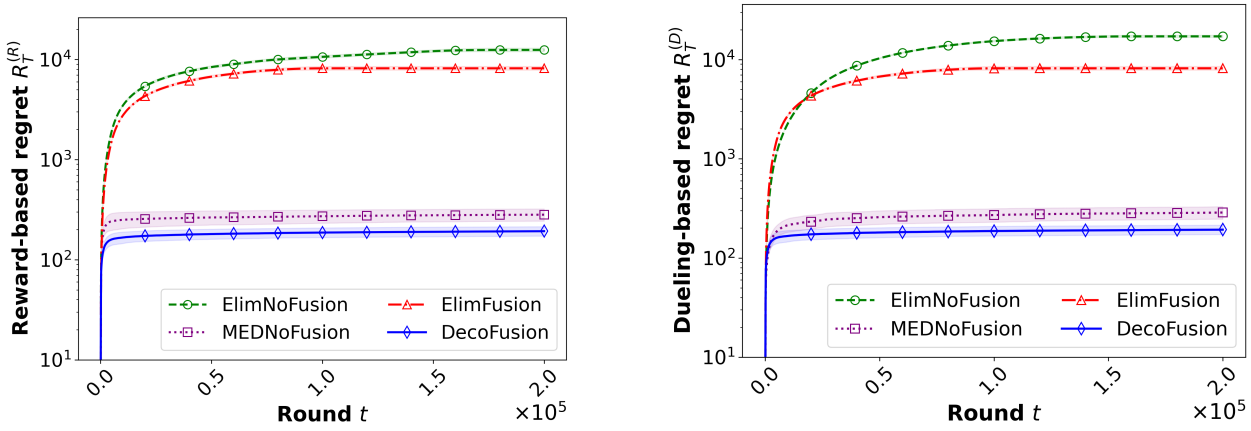
where the first constant term comes from the initialization of the algorithm (Algorithm 3), the last terms of (17) and (18) is due to the imbalanced exploration in Line 12 of Algorithm 2, and the missing third term of (19) is because when event $\mathcal{A}_t$ happens, the algorithm does not explore arm pair (k,ℓ) for any suboptimal arm $\ell \neq 1$.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{J}_{k,t} \text{ and } \mathcal{A}_t\}\right] \\
&\leqslant L_k^{(\text{Suff})} + \sum_{t=1}^{T} \sum_{n,m>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\{\mathcal{J}_{k,t} \text{ and } \mathcal{A}_t \text{ and } N_{k,t}=n, M_{k,1,t}=m\}\right] \\
&\leqslant L_k^{(\text{Suff})} + \sum_{t=1}^{T} \sum_{n,m>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\left\{\max\left\{I_{k,t}^{(\text{R})}, I_{k,t}^{(\text{D})} - I_{\hat{k}_t^{(\text{D})},t}^{(\text{D})}\right\} \leqslant \log t + f(K) \text{ and } \mathcal{A}_t \text{ and } N_{k,t}=n, M_{k,1,t}=m\right\}\right] \\
&\leqslant L_k^{(\text{Suff})} + \sum_{n,m>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\left\{L_k^{(\text{Suff})} \max\left\{\mathrm{kl}(\hat{\mu}_k(n), \mu_1-\delta), \mathrm{kl}(\hat{\nu}_{k,1}(m), \frac{1}{2})\right\} \leqslant \log T + f(K)\right\}\right] \qquad (20) \\
&\leqslant L_k^{(\text{Suff})} + \sum_{n,m>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\left\{\max\left\{\mathrm{kl}(\hat{\mu}_{k,t}(n), \mu_1-\delta), \mathrm{kl}(\hat{\nu}_{k,1,t}(m), \frac{1}{2})\right\} \leqslant \frac{\max\left\{\mathrm{kl}(\mu_k,\mu_1), \mathrm{kl}(\nu_{k,1}, \frac{1}{2})\right\}}{1+\epsilon}\right\}\right] \qquad (21) \\
&\leqslant L_k^{(\text{Suff})} + \sum_{n>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\left\{\mathrm{kl}(\hat{\mu}_{k,t}(n), \mu_1-\epsilon) \leqslant \frac{\mathrm{kl}(\mu_k,\mu_1)}{1+\epsilon}\right\}\right] + \sum_{m>L_k^{(\text{Suff})}} \mathbb{E}\left[\mathbb{1}\left\{\mathrm{kl}(\hat{\nu}_{k,1,t}(m), \frac{1}{2}) \leqslant \frac{\mathrm{kl}(\nu_{k,1}, \frac{1}{2})}{1+\epsilon}\right\}\right] \\
&\leqslant L_k^{(\text{Suff})} + O(\epsilon^{-2}), \qquad (22)
\end{aligned}$$

where inequality (20) applies the $\hat{\mu}_k(n)$ to denote the reward mean estimate with n samples, and $\hat{\nu}_{k,1}(m)$ to denote the dueling probability estimate with m samples, inequality (21) is by substituting $L_k^{(\text{Suff})}$ the definition of $L_k^{(\text{Suff})}$, inequality (22) is for that the last two terms are bounded as $O(\epsilon^{-2})$ in Honda & Takemura (2010, Lemma 15) and Komiyama et al. (2015, Lemma 6), respectively.

Therefore, the final regret upper bound is given as follows,

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}[R_T|\mathcal{A}_t] + \mathbb{E}[R_T|\mathcal{A}_t^{\mathsf{C}}] \\
&\leqslant \sum_{k\neq 1} \frac{(\Delta_k^{(\text{R})}/\alpha + \Delta_k^{(\text{D})}/(1-\alpha))((1+\epsilon)\log T + f(K))}{\max\left\{\mathrm{kl}(\mu_k,\mu_1)/\alpha^2, \mathrm{kl}(\nu_{k,1}, \frac{1}{2})/(1-\alpha)^2\right\}} + O(K^2) + O(\epsilon^{-2}) + O(e^{cK-f(K)}).
\end{aligned}$$



(a) Reward regret $R_T^{(\mathrm{R})}$ (b) Dueling regret $R_T^{(\mathrm{D})}$

Figure 3: Regret comparison in different settings (continue)

where we individual bound the two terms as follows,

$$\begin{aligned}
\mathbb{E}[R_T|\mathcal{A}_t] &\leqslant \sum_{k\neq 1} \alpha\Delta_k^{(\mathrm{R})}\mathbb{E}[N_{k,T}] + (1-\alpha)\Delta_k^{(\mathrm{D})}\mathbb{E}[M_{k,1,T}] \\
&\leqslant \sum_{k\neq 1} \frac{(\alpha(1-\beta)\Delta_k^{(\mathrm{R})} + (1-\alpha)\beta\Delta_k^{(\mathrm{D})})((1+\epsilon)\log T + f(K))}{\max\left\{(1-\beta)\,\mathrm{kl}(\mu_k,\mu_1), \beta\,\mathrm{kl}(\nu_{k,1},\frac{1}{2})\right\}} + O(\epsilon^{-2}) \\
&\leqslant \sum_{k\neq 1} \frac{(\Delta_k^{(\mathrm{R})}/\alpha + \Delta_k^{(\mathrm{D})}/(1-\alpha))((1+\epsilon)\log T + f(K))}{\max\left\{\mathrm{kl}(\mu_k,\mu_1)/\alpha^2, \mathrm{kl}(\nu_{k,1},\frac{1}{2})/(1-\alpha)^2\right\}} + O(\epsilon^{-2}),
\end{aligned}$$

where the last inequality is by substituting the definition of parameter β, and

$$\begin{aligned}
\mathbb{E}[R_T|\mathcal{A}_t^{\mathrm{C}}] &\leqslant \left(\alpha\Delta_2^{(\mathrm{R})} + (1-\alpha)\max_{k\in\mathcal{K}}\Delta_k^{(\mathrm{D})}\right)\left(K(K-1) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}_t^C\}\right]\right) \\
&\leqslant K^2 + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{A}_t^C\}\right] \leqslant O(K^2) + O(e^{cK-f(K)}),
\end{aligned}$$

where the $K(K-1)$ comes from the initialization of the algorithm (Algorithm 3). □

E. Experimental Setup and Additional Experiments

E.1. Additional Experiments

This section further reports other experiments in Figure 3. Figures 3(a) and 3(b) plots the decomposed reward- and dueling-based regrets of the aggregated regret in Figure 2(a) in the main paper. They illustrate that the two fusion algorithms outperform the no-fusion algorithms regarding reward-based and dueling-based regrets, respectively, highlighting the effectiveness of the fusion algorithms in mitigating regret costs to the other feedback.

E.2. Experiment Setup

The experiments of Figures 2(a), 2(d), 3(a), 3(b) are conducted with $K = 16$ arms, where their Bernoulli reward distributions are with means $\boldsymbol{\mu} = \{0.86, 0.80, 0.75, 0.70, 0.65, 0.60, 0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15, 0.10\}$. A dueling

probability matrix $\boldsymbol{\nu}$ determines the dueling feedback as follows,

$$\begin{bmatrix}
0.50 & 0.54 & 0.57 & 0.60 & 0.63 & 0.65 & 0.69 & 0.71 & 0.73 & 0.76 & 0.78 & 0.82 & 0.86 & 0.91 & 0.95 & 0.98 \\
0.46 & 0.50 & 0.54 & 0.58 & 0.61 & 0.64 & 0.67 & 0.70 & 0.74 & 0.76 & 0.79 & 0.81 & 0.84 & 0.87 & 0.89 & 0.92 \\
0.43 & 0.46 & 0.50 & 0.54 & 0.58 & 0.60 & 0.63 & 0.66 & 0.69 & 0.72 & 0.76 & 0.79 & 0.83 & 0.85 & 0.88 & 0.91 \\
0.40 & 0.42 & 0.46 & 0.50 & 0.54 & 0.58 & 0.61 & 0.64 & 0.66 & 0.69 & 0.72 & 0.76 & 0.79 & 0.82 & 0.85 & 0.88 \\
0.37 & 0.39 & 0.42 & 0.46 & 0.50 & 0.54 & 0.56 & 0.59 & 0.63 & 0.66 & 0.69 & 0.72 & 0.76 & 0.78 & 0.82 & 0.86 \\
0.35 & 0.36 & 0.40 & 0.42 & 0.46 & 0.50 & 0.54 & 0.57 & 0.59 & 0.63 & 0.67 & 0.70 & 0.73 & 0.76 & 0.79 & 0.82 \\
0.31 & 0.33 & 0.37 & 0.39 & 0.44 & 0.46 & 0.50 & 0.54 & 0.58 & 0.61 & 0.64 & 0.68 & 0.71 & 0.72 & 0.75 & 0.79 \\
0.29 & 0.30 & 0.34 & 0.36 & 0.41 & 0.43 & 0.46 & 0.50 & 0.54 & 0.57 & 0.59 & 0.62 & 0.65 & 0.68 & 0.72 & 0.76 \\
0.27 & 0.26 & 0.31 & 0.34 & 0.37 & 0.41 & 0.42 & 0.46 & 0.50 & 0.54 & 0.58 & 0.61 & 0.63 & 0.66 & 0.69 & 0.73 \\
0.24 & 0.24 & 0.28 & 0.31 & 0.34 & 0.37 & 0.39 & 0.43 & 0.46 & 0.50 & 0.54 & 0.56 & 0.59 & 0.62 & 0.66 & 0.69 \\
0.22 & 0.21 & 0.24 & 0.28 & 0.31 & 0.33 & 0.36 & 0.41 & 0.42 & 0.46 & 0.50 & 0.54 & 0.56 & 0.58 & 0.61 & 0.65 \\
0.18 & 0.19 & 0.21 & 0.24 & 0.28 & 0.30 & 0.32 & 0.38 & 0.39 & 0.44 & 0.46 & 0.50 & 0.54 & 0.57 & 0.58 & 0.62 \\
0.14 & 0.16 & 0.17 & 0.21 & 0.24 & 0.27 & 0.29 & 0.35 & 0.37 & 0.41 & 0.44 & 0.46 & 0.50 & 0.54 & 0.56 & 0.59 \\
0.09 & 0.13 & 0.15 & 0.18 & 0.22 & 0.24 & 0.28 & 0.32 & 0.34 & 0.38 & 0.42 & 0.43 & 0.46 & 0.50 & 0.54 & 0.56 \\
0.05 & 0.11 & 0.12 & 0.15 & 0.18 & 0.21 & 0.25 & 0.28 & 0.31 & 0.34 & 0.39 & 0.42 & 0.44 & 0.46 & 0.50 & 0.54 \\
0.02 & 0.08 & 0.09 & 0.12 & 0.14 & 0.18 & 0.21 & 0.24 & 0.27 & 0.31 & 0.35 & 0.38 & 0.41 & 0.44 & 0.46 & 0.50
\end{bmatrix}$$

where the value of $\nu_{i,j}$ between arm pairs (i, j) is in row i column j. The algorithms are run for $T = 200,000$ rounds with the following parameters for DECOFUSION and ELIMFUSION: $\alpha = 0.5$, $\delta = 1/T$, and $f(K) = 0.05K^{1.01}$. Each experiment is repeated 100 times, and we report the average regret and the standard deviation of all runs.

Then, Figures 2(b) and 2(c) report the final aggregated regrets under the following two experiments.

Fixing ν, varying μ: Fixing the dueling probability as in the matrix:

$$\boldsymbol{\nu} = \begin{bmatrix}
0.50 & 0.53 & 0.56 & 0.59 & 0.62 \\
0.47 & 0.50 & 0.53 & 0.56 & 0.59 \\
0.44 & 0.47 & 0.50 & 0.53 & 0.56 \\
0.41 & 0.44 & 0.47 & 0.50 & 0.53 \\
0.38 & 0.41 & 0.44 & 0.47 & 0.50
\end{bmatrix}$$

Vary $\boldsymbol{\mu} = \{0.9, 0.9 - \Delta, 0.9 - 2 \times \Delta, 0.9 - 3 \times \Delta, 0.9 - 4 \times \Delta\}$, where $\Delta \in \{0.06, 0.11, 0.16, 0.21\}$.

Fixing μ, varying ν: Fixing $\boldsymbol{\mu} = \{0.9, 0.84, 0.78, 0.72, 0.66\}$, we consider vary preference matrix in:

$$\boldsymbol{\nu} = \begin{bmatrix}
0.5 & 0.5 + 1 \times \Delta & 0.5 + 2 \times \Delta & 0.5 + 3 \times \Delta & 0.5 + 4 \times \Delta \\
0.5 - 1 \times \Delta & 0.5 & 0.5 + 1 \times \Delta & 0.5 + 2 \times \Delta & 0.5 + 3 \times \Delta \\
0.5 - 2 \times \Delta & 0.5 - 1 \times \Delta & 0.5 & 0.5 + 1 \times \Delta & 0.5 + 2 \times \Delta \\
0.5 - 3 \times \Delta & 0.5 - 2 \times \Delta & 0.5 - 1 \times \Delta & 0.5 & 0.5 + 1 \times \Delta \\
0.5 - 4 \times \Delta & 0.5 - 3 \times \Delta & 0.5 - 2 \times \Delta & 0.5 - 1 \times \Delta & 0.5
\end{bmatrix}$$

where $\Delta \in \{0.03, 0.05, 0.07, 0.09, 0.11\}$. All other settings of the two experiments are the same as above.

F. Extended Discussions on Regret Definition

If the preference probability $\nu_{k,\ell}$ is generated according to the reward means μ_k, μ_ℓ for any arm pair (k, ℓ). Assuming a relation between the preference probability and reward mean, like the Bradley-Terry model $\nu_{1,2} = \frac{\exp(\mu_1)}{\exp(\mu_1)+\exp(\mu_2)} = \frac{\exp(\mu_1-\mu_2)}{\exp(\mu_1-\mu_2)+1}$ or utility dueling bandits $\nu_{1,2} = \frac{1+(\mu_1-\mu_2)}{2}$ (Ailon et al., 2014), is stronger than the one in our paper and will lead to a better regret result.

Specifically, with this parametric relation, one only needs to focus on one set of the parameters, either the reward means or the dueling probabilities. Let us consider the case that the reward mean $(\mu_1, \mu_2, \ldots, \mu_K)$ as the basic parameters to estimate, and all observations from dueling feedback can be translated into reward feedback via the parametric relation. Under this point of view, the reward feedback provides direct observations from the reward mean parameters, and the dueling (preference) feedback between arms k and ℓ can be considered as a "parametric reward feedback" depending on the reward

mean parameters. For example, with the Bradley-Terry relation, the parametric form is a logistic function, which is similar to the logistic bandits (Faury et al., 2020), while with the utility dueling relation, the parametric form is a linear function, which is similar to the linear bandits (Abbasi-Yadkori et al., 2011).

With this interpretation, we believe that the final regret bound would depend on reward gaps $\Delta_k^{(R)}$ and have no explicit dependence on the dueling gaps $\Delta_k^{(D)}$ (or the other way round if we consider the dueling probabilities as the basic parameters), and the regret improvement may be in the actual dependence of the reward gaps $\Delta_k^{(R)}$ (e.g., the prefactor or the exponential order of this gap would be strictly less than than the best one without the dueling feedback option). To rigorously investigate the improvement in regret bounds under these stronger assumptions (out of the scope of current paper) is an interesting research direction.

Motivation of current regret definition in Eq. (1) The current linear combination definition—especially, the parameter $\alpha \in [0, 1]$—is motivated by the cost difference between querying a reward feedback, as a dueling feedback, as the dueling (relative) feedback is usually cost-efficient (Ouyang et al., 2022). Furthermore, this flexible definition covers many interesting scenarios, e.g., the case of $\alpha = 1$ is the regret from reward feedback only, and the case of $\alpha = 0$ is the regret from dueling feedback only, and in the case of $\alpha = \frac{1}{2}$, the regret is the simple sum of the two types of feedback.

Below, we provide three perspectives on changing the definition of regret in our paper. First, if one only considers the regret from reward feedback ("define the regret as the loss in the expected reward"), and treat the dueling feedback as a side free observations, then this regret reduces to the case of setting $\alpha = 1$ in our regret definition. In this scenario, our DECOFUSION algorithm achieves constant regret, as discussed in Lines 409–423 (left column).

Second, if one also considers the regret cost due to dueling feedback but count this part of the regret in terms of the expected regret instead of the dueling (preference) probability in the current definition, then the new regret cost in each decision round would be the sum of the reward gaps of the pulled three arms (one for reward, a pair for dueling). For this new regret definition, our algorithm and analysis still works, and the only modification is in the regret upper bound results, where one needs to change the dueling gap $\Delta_k^{(D)}$ in the nominator of Eq. (5) in Theorem 3.1 and Eq. (8) in Theorem 4.1 to the reward gap $2\Delta_k^{(R)}$.

Third, if one further assumes the Bradley-Terry model relation between reward means and dueling probabilities upon the new regret definition, this would lead to a problem that is similar to logistic bandits (Faury et al., 2020), as discussed in the previous response, which is an interesting research direction.