FINDING EPSILON-BEST ARMS IN STRUCTURED BANDITS VIA FUSED FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

To model real-world scenarios where a learner receives diverse, potentially biased feedback signals, we introduce a structured multi-armed bandit framework with unaligned fused feedback. In this setting, at each round, the learner can either select a single arm to observe its standard reward feedback or select a pair of arms to observe the corresponding dueling feedback. Crucially, each feedback modality has its own latent parameter: the expected feedback associated with an arm is the inner product of its feature vector and the corresponding modality-specific parameter. We propose WIT-BOTELIM, an algorithm tailored for ε-best arm identification under fused feedback, which leverages a novel fused-elimination rule to efficiently aggregate disparate feedback modalities. By proving a generic sample complexity lower bound, we demonstrate that WIT-BOTELIM achieves near-optimal performance in key practical regimes. Finally, numerical experiments indicate the superiority of WIT-BOTELIM over shared-parameter methods under latent-parameter mismatch, which is attributed to the delicately-designed fused-elimination.

1 INTRODUCTION

Modern machine learning systems increasingly rely on heterogeneous forms of supervision to improve model performance (Christiano et al., 2017; Ouyang et al., 2022; Wang et al., 2025b). Consider, for example, a post-training LLM system that generates a finite set of candidate responses and must select one for deployment or further training. The system may request an absolute evaluation of an individual response, such as a human rating, task-success indicator, or verifier score, which we term *reward feedback*; alternatively, it may ask a human or AI judge to compare a pair of responses, which we term *dueling feedback*. When only one query can be issued at a time, the learner must decide both *which feedback channel to query* and *which arm or pair of arms to evaluate*. Despite the growing availability of such heterogeneous feedback, how to make full use of *fused feedback*, that is, to acquire and integrate information from multiple modalities efficiently, remains insufficiently understood.

Bandit models provide a natural abstraction for studying this problem. In a structured bandit setting, a learner may query an individual arm for reward feedback or a pair of arms for dueling feedback, with observations within each modality linked through a modality-specific latent parameter. Existing approaches to reward–dueling fusion typically impose an *alignment assumption*, requiring the two modalities to induce a consistent arm ranking and share the same optimal arm (Wang et al., 2025a; Zeng et al., 2026). In contrast, we allow reward and dueling feedback to be governed by distinct modality-specific latent parameters. This consequently allows different arm rankings and optimal arms. This formulation is both more general and better suited to practical settings in which the two modalities capture different aspects of candidate quality. For instance, absolute ratings and pairwise preferences are often elicited through distinct procedures and may therefore convey complementary or discrepant information about the same candidate.

Beyond the post-training example, such feedback unalignment raises a fundamental question in structured bandits: how can a learner efficiently utilize unaligned reward and dueling feedback to identify the best arm? Existing algorithms developed under alignment assumption may not fit such unaligned setting well, as they pool observations across modalities into a shared estimator (Wang et al., 2025a; Zeng et al., 2026). To bridge this gap, we propose an efficient algorithm,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the Witness-Bottleneck Elimination (WIT-BOTELIM) algorithm, for best arm identification (BAI) under unaligned fusion and establish the generic lower bound for any BAI algorithm. Our study demonstrates that effective utilization of unaligned feedback of different modalities remains possible. In specific, WIT-BOTELIM evaluates arms separately within each modality while integrating information from both modalities for decision making.

Main contributions. We study fixed-confidence BAI in structured bandits with unaligned fused feedback. First, we formulate this unaligned setting, where reward and dueling feedback are governed by distinct latent parameters and may therefore induce different rankings and optimal arms. This more general setting can model practical scenarios in which the two modalities capture different aspects of arm quality, but such unalignment prevents a learner from simply pooling information across modalities for efficient BAI, thereby posing a fundamental challenge.

Secondly, due to the misalignment between the modality-specific parameters, a globally optimal arm may not exist. Nevertheless, it remains valuable to identify an ε-optimal arm whose gaps under both modalities are less than ε. Accordingly, we characterize the problem difficulty through a lower bound shared by all algorithms.

Furthermore, we design WIT-BOTELIM to identify all ε-optimal arms, when the efficacy of WIT-BOTELIM stems from an embedded fused-elimination rule. WIT-BOTELIM maintains separate estimators for different modalities while utilizing observations from either modality individually or from both modalities jointly to eliminate arms from the common candidate set. We prove that WIT-BOTELIM can succeed in the BAI task with high probability efficiently. The analysis of WIT-BOTELIM, along with our lower bound result, implies the near-optimality of WIT-BOTELIM in some practical regimes. Empirical results further demonstrate that WIT-BOTELIM outperforms shared-parameter methods under latent-parameter mismatch.

1.1 LITERATURE REVIEW

Standard best arm identification. Best-arm identification (BAI) has been studied extensively under both fixed-budget and fixed-confidence settings. In the fixed-budget setting, Audibert et al. (2010) proposed the Successive Rejects algorithm, while Karnin et al. (2013) designed the Sequential Halving algorithm; the latter also introduced the Exponential-Gap Elimination algorithm for fixed-confidence BAI. Besides, Garivier & Kaufmann (2016) characterized the generic lower bound on time complexity of fixed-confidence BAI algorithms and proposed the Track-and-Stop algorithm accordingly, which adaptively tracks the optimal sampling proportions.

Structured BAI. Structured bandits, which assume arm rewards are coupled through some shared structure, include linear and dueling bandits, and, more recently, bandits with multiple or fused feedback modalities. In linear BAI, where the expected arm reward is modeled as the inner product of an arm feature vector and an unknown latent parameter, *experimental design* such as G-allocation and $\mathcal{XY}$-allocation can be used to allocate arm samples toward informative directions (Soare et al., 2014; Jedra & Proutiere, 2020). Closely related to our algorithmic design, Fiez et al. (2019) proposed the Randomized Adaptive Gap Elimination (RAGE) algorithm, which is equipped with both experimental design and phased elimination rule for efficient BAI in linear bandits. Generalized-linear BAI extends the framework of experimental design beyond the linear setting: Kazerouni & Wein (2021) studied BAI under generalized linear reward models, while Jun et al. (2021) developed sharper confidence bounds for the linear logistic model and designed the RAGE-GLM algorithm for BAI in transductive logistic bandits. The latter is particularly relevant to the statistical treatment of our dueling modality.

Fused bandits. Meanwhile, recent studies began to design efficient algorithms in fused bandits for both BAI and regret minimization (RM). The regret is defined as the gap between the maximum reward and a learner's reward. In specific, Wang et al. (2025a) introduced the fused bandits without linear assumption among arm rewards and proposed the ElimFusion and DecoFusion algorithm for RM. Moreover, Zeng et al. (2026) extended this fused setting to a generalized linear setting with aligned fused feedback and designed the HyTS-GLB algorithm, which adaptively chooses between reward and dueling queries with the objective of BAI. In contrast, our unaligned fused setting allows reward and dueling feedback to be governed by distinct latent parameters and to induce different

optimal arms. BAI in unaligned environments represents a key yet unexamined problem; this study directly addresses that gap.

Bandits with multiple criteria. Our problem is also related to bandit settings with heterogeneous measurements or multiple criteria. Learners in multi-fidelity bandits study how measurements with different costs and fidelity levels should be allocated (Poiani et al., 2024), but the different fidelities are ultimately used to estimate the same underlying arm value. Constrained Bandits allow arm performance and feasibility to be assessed separately once either criterion is sufficiently resolved (Cai & Kandasamy, 2026); here arm feasibility conditions are defined relative to fixed constraints. Moreover, multi-objective problems may require learner to make decisions utilizing multiple criteria (Kone et al., 2025; Chen et al., 2025). In representative multi-objective bandit formulations, pulling an arm reveals multiple objective components simultaneously, and the goal of a BAI learner is to identify the best arm associated with each objective (Chen et al., 2025). In contrast, an efficient learner in our setting with unaligned fused feedback must actively choose between two qualitatively different feedback modalities—a single-arm reward query and a pairwise dueling query—and seeks an arm whose gaps under both modalities are at most ε. Hence, although these problems seem to be similar, the involved dynamic and reward design differ fundamentally from our fused bandit setting.

2 Problem Formulation

We formulate a structured bandit setting with unaligned fused feedback. Consider K arms indexed by $[K] := \{1, \dots, K\}$ with each arm $i \in [K]$ associated with a known feature vector $x_i \in \mathbb{R}^d$. Let $\mathcal{P} := \{(j,k) \in [K]^2 : j < k\}$. The environment is characterized by two latent parameters, $\theta_R \in \mathbb{R}^d$ for reward feedback and $\theta_D \in \mathbb{R}^d$ for dueling feedback. In contrast to models that impose a common latent parameter or a common ordering across the two feedback modalities, we allow $\theta_R \neq \theta_D$. For brevity, we use R to denote the reward modality and D to denote the dueling modality. We use $\tilde{m}$ to denote the other modality, e.g., $\tilde{R} = D$ and $\tilde{D} = R$.

At each round t, the learner adaptively chooses one of two feedback modalities. For a reward query, it selects an arm $i_t \in [K]$ and observes $R_t = x_{i_t}^\top \theta_R + \iota_t$, where ι_t is conditionally zero-mean and 1-sub-Gaussian. For a dueling query, it selects a pair (j_t, k_t) from $\mathcal{P}$ and observes $Y_t \sim \text{Bernoulli}\Big(\mu(x_{j_t,k_t}^\top \theta_D)\Big)$, with $x_{j,k} := x_j - x_k$, $\mu(z) = (1+e^{-z})^{-1}$ is the logistic link, and $Y_t = 1$ indicates a preference for arm j_t over arm k_t.

For a feedback modality $m \in \{R, D\}$, let $i_m^\star \in \arg\max_{j \in [K]} x_j^\top \theta_m$ denote a modality-wise optimal arm. Accordingly, we define the modality-specific gap of arm i by $\Delta_{m,i} := \max_{j \in [K]} x_j^\top \theta_m - x_i^\top \theta_m = \max_{j \in [K]} (x_j - x_i)^\top \theta_m$ for all $m \in \{R, D\}$. To identify an arm that is simultaneously competitive under both modalities, we define the *fused gap* $\Delta_i := \max\{\Delta_{R,i}, \Delta_{D,i}\}$, and the set of jointly ε-optimal arms $\mathcal{S}_\varepsilon := \{i \in [K] : \Delta_i < \varepsilon\} = \{i \in [K] : \Delta_{R,i} < \varepsilon,\ \Delta_{D,i} < \varepsilon\}$. Furthermore, we consider two identification objectives. In the ε^{one}*-BAI* problem, the learner returns any arm in $\mathcal{S}_\varepsilon$; throughout this setting we assume $\mathcal{S}_\varepsilon \neq \emptyset$. In the ε^{all}*-BAI* problem, the learner aims to recover $\mathcal{S}_\varepsilon$ itself.

We now formalize the class of sequential algorithms. Let $\mathcal{A}$ denote the query space $\mathcal{A} := \big(\{R\} \times [K]\big) \cup \big(\{D\} \times \mathcal{P}\big)$. Let $A_t \in \mathcal{A}$ be the query selected at round t, and let O_t denote the resulting reward or dueling observation. The history filtration is $\mathcal{F}_t := \sigma(A_1, O_1, \dots, A_t, O_t)$. We let $\mathcal{F}_0 := \{\emptyset, \Omega\}$. A non-anticipatory algorithm is therefore specified by $\pi = \big((\pi_t)_{t \geq 1}, \tau^\pi, \rho^\pi\big)$, where the sampling rule π_t selects A_t using only $\mathcal{F}_{t-1}$, the stopping rule τ^π is an almost surely finite stopping time with respect to $(\mathcal{F}_t)_{t \geq 0}$, and the recommendation rule ρ^π is $\mathcal{F}_{\tau^\pi}$-measurable. For ε^{one}-BAI problem, we write its target output as $\widehat{i}^\pi \in [K]$; for the ε^{all}-BAI problem, we write the target output as $\widehat{\mathcal{S}}^\pi \subseteq [K]$.

Under the fixed-confidence setting, the target failure probability $\delta \in (0,1)$ is specified in advance. For the one ε-optimal problem, an algorithm π is (ε, δ)-PAC if, for every admissible instance, $\mathbb{P}^\pi_{\theta_R, \theta_D}\Big(\widehat{i}^\pi \in \mathcal{S}_\varepsilon\Big) \geq 1 - \delta$. For the ε^{all}-BAI problem, π is (ε, δ)-PAC if $\mathbb{P}^\pi_{\theta_R, \theta_D}\Big(\widehat{\mathcal{S}}^\pi = \mathcal{S}_\varepsilon\Big) \geq 1 - \delta$. The sample complexity of π is its stopping time τ^π. Our objective is to design an (ε, δ)-PAC algorithm with small τ^π in expectation and with high probability, while exploiting the information provided by both feedback modalities.

2.1 LOWER BOUND

This section establishes an information-theoretic lower bound on the expected number of queries required by any (ε, δ)-PAC algorithm. We focus on instances satisfying $\mathcal{S}_\varepsilon \neq \emptyset$ and distinguish between ε^{one}-BAI and ε^{all}-BAI problem.

We assume that $\text{span}\{x_i : i \in [K]\} = \mathbb{R}^d, \text{span}\{x_j - x_k : (j,k) \in \mathcal{P}\} = \mathbb{R}^d$. Let Δ_K and $\Delta_\mathcal{P}$ denote the probability simplices over $[K]$ and $\mathcal{P}$ respectively. For designs $\nu_R \in \Delta_K$ and $\nu_D \in \Delta_\mathcal{P}$, define $M_R(\nu_R) := \sum_{a\in[K]} \nu_{R,a} x_a x_a^\top$, $M_D(\nu_D) := \sum_{(j,k)\in\mathcal{P}} \nu_{D,j,k} x_{j,k} x_{j,k}^\top$. We restrict the design optimization to $\mathfrak{D}_R := \{\nu_R \in \Delta_K : M_R(\nu_R) \succ 0\}, \mathfrak{D}_D := \{\nu_D \in \Delta_\mathcal{P} : M_D(\nu_D) \succ 0\}$. Hence, all design matrices considered below are invertible.

For every $i \in \mathcal{S}_\varepsilon$, define

$$H_i^R := \inf_{\nu_R\in\mathfrak{D}_R} \max_{k\neq i} \frac{\|x_k - x_i\|^2_{M_R(\nu_R)^{-1}}}{\left(\varepsilon - (x_k - x_i)^\top \theta_R\right)^2}, \qquad H_i^D := \inf_{\nu_D\in\mathfrak{D}_D} \max_{k\neq i} \frac{\|x_k - x_i\|^2_{M_D(\nu_D)^{-1}}}{\left(\varepsilon - (x_k - x_i)^\top \theta_D\right)^2}.$$

These quantities characterize the *modality-specific hardness* of arm i identification under an (ε, δ)-PAC criterion and make explicit how ε enters the lower bound. Within each modality, the difficulty of arm i identification is dictated by the hardest arm to distinguish from arm i under the optimal allocation of information, jointly with the near-optimality design reflected by ε.

Theorem 1 (Lower bound for a nonempty joint answer)**.** *Suppose $\mathcal{S}_\varepsilon \neq \emptyset$ and let $s := |\mathcal{S}_\varepsilon|$.*

For ε^{all}-BAI, if $0 < \delta < 1/(s+1)$, every (ε, δ)-PAC algorithm satisfies

$$\mathbb{E}^\pi_{\theta_R,\theta_D}[\tau^\pi] \geq 2\,\text{kl}\left(\frac{1-\delta}{s}, \delta\right) \min_{i\in\mathcal{S}_\varepsilon} \left(H_i^R + 4H_i^D\right).$$

For ε^{all}-optimal identification, if $0 < \delta < 1/2$, every (ε, δ)-PAC algorithm satisfies

$$\mathbb{E}^\pi_{\theta_R,\theta_D}[\tau^\pi] \geq 2\,\text{kl}(1-\delta, \delta) \left[\max_{i\in\mathcal{S}_\varepsilon} H_i^R + 4 \max_{i\in\mathcal{S}_\varepsilon} H_i^D\right].$$

These lower bounds characterize the minimum expected number of time steps required by any (ε, δ)-PAC algorithm. The way the modality-specific hardness terms are integrated further reflects the distinction between the two identification targets. When the learner is required to return only one arm from $\mathcal{S}_\varepsilon$, it suffices to certify a single valid arm. In contrast, for ε^{all}-optimal identification, the returned set should be identical to $\mathcal{S}_\varepsilon$.

The quantities above align with those in the classical transductive linear-bandit lower bound (Fiez et al., 2019). Specifically, in the reward-only setting, taking i to be the unique best arm and setting $\varepsilon = 0$ recovers H_i^R to $\inf_{\nu_R\in\mathfrak{D}_R} \max_{k\neq i} \|x_i - x_k\|^2_{M_R(\nu_R)^{-1}} / ((x_i - x_k)^\top \theta_R)^2$, which is the corresponding characteristic experimental-design quantity in Fiez et al. (2019). Thus, the reward component retains the standard transductive geometry, while the unaligned fused setting introduces an ε-shifted decision boundary and an additional logistic feedback modality.

3 WIT-BOTELIM: FUSING REWARD AND DUELING VIA ELIMINATION

For ε-BAI in our proposed bandit setting, we designed the Witness-Bottleneck Elimination (WIT-BOTELIM) algorithm. When adapting the design of RAGE (Fiez et al., 2019), our WIT-BOTELIM algorithm apparently differ from RAGE as WIT-BOTELIM must accommodate two modalities governed by distinct parameters θ, which requires modality-specific estimation and a different elimination mechanism, as detailed in Section 3.1.

As presented in Algorithm 1, WIT-BOTELIM introduces a sharper fused-elimination rule with separate RAGE-style allocations for the two modalities. During initialization, the algorithm performs a burn-in phase to obtain initial estimates for the dueling modality. At each phase r, it constructs modality-specific candidate sets and uses them to determine the corresponding sampling allocations. It then collects rounded batches according to these allocations and performs estimation separately for the two modalities. The algorithm subsequently applies fused elimination using $\text{CR}_{m,r}(i,j)$ for

Algorithm 1 WIT-BOTELIM: Witness–Bottleneck Elimination

Require: $\{x_i\}_{i=1}^K$, ε, δ, κ_0, ζ, Target Mode $q \in \{\texttt{one}, \texttt{all}\}$, $\gamma \in (0,1)$

Initialize $\gamma_0 = \gamma, C_1 = W_{R,0} = W_{D,0} = [K]$, $K_{R,1} = K_{D,1} = \widehat{\mathcal{S}} = \emptyset$, and $\Phi = \{i, j \in [K] : i < j\}$

Duel-side Warm up: $\widehat{\theta}_{D,0} \leftarrow \textsc{BurnIn}(\Phi, \kappa_0, \delta/2, \zeta)$

for $r = 1, 2, \ldots$ **do**

  Initialize phase-wise parameters for $m \in \{R, D\}$: $\gamma_r = \gamma_{r-1}/2$, $\delta_r = 3\delta/(\pi^2 r^2)$, $\delta_{m,r} = \delta_r/2$, and $\delta_{m,r}^{\mathrm{gap}} = \delta_{m,r}^{\mathrm{wit}} = \delta_r/4$

  Construct candidate set and active target for $m \in \{R, D\}$, and computes allocation rule: $G_{m,r} \leftarrow C_r \setminus K_{\bar{m},r}$ and $\mathcal{Y}_{m,r} \leftarrow \{x_j - x_i : i \in G_{m,r},\ j \in W_{m,r-1},\ j \neq i\}$

  $(\nu_{m,r}, n_{m,r}) \leftarrow \textsc{Allocation}\left(m, \mathcal{Y}_{m,r}, \gamma_r, \delta_{m,r}^{\mathrm{gap}}, \delta_{m,r}^{\mathrm{wit}}, \widehat{\theta}_{D,r-1}, \zeta\right)$ for $m \in \{R, D\}$

  Collect the rounded batches and compute $\widehat{\theta}_{m,r} \leftarrow \textsc{Estimate}(m, \nu_{m,r}, n_{m,r})$ for $m \in \{R, D\}$

  Remove witnesses having a valid strict-domination certificate for $m \in \{R, D\}$: $W_{m,r} \leftarrow \left\{ j \in W_{m,r-1} : (x_k - x_j)^\top \widehat{\theta}_{m,r} \le \mathrm{CR}_{m,r}(k,j),\ \forall k \in W_{m,r-1} \setminus \{j\} \right\}$

  Construct $[L_{m,i}^{(r)}, U_{m,i}^{(r)}]$ for $m \in \{R, D\}$ and $i \in G_{m,r}$: $\widehat{\Delta}_{m,i}^{(r)} \leftarrow \max_{j \in W_{m,r}} (x_j - x_i)^\top \widehat{\theta}_{m,r}$, $L_{m,i}^{(r)} \leftarrow \max\{0, \widehat{\Delta}_{m,i}^{(r)} - \gamma_r\}$, $U_{m,i}^{(r)} \leftarrow \widehat{\Delta}_{m,i}^{(r)} + \gamma_r$

  Construct the fused interval of each $i \in C_r$ and the certified sets: $K_{m,r}^+ \leftarrow K_{m,r} \cup \left\{ i \in C_r \setminus (K_{R,r} \cup K_{D,r}) : U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)} \right\}$ for $m \in \{R, D\}$

  $$(L_i^{F,(r)}, U_i^{F,(r)}) \leftarrow \begin{cases} (L_{m,i}^{(r)}, U_{m,i}^{(r)}), & i \in K_{m,r}^+ \text{ for some } m, \\ (\max_m L_{m,i}^{(r)}, \max_m U_{m,i}^{(r)}), & \text{otherwise,} \end{cases} \quad \text{for } i \in C_r$$

  $\mathcal{I}_r \leftarrow \{i \in C_r : U_i^{F,(r)} < \varepsilon\}$, $\mathcal{O}_r \leftarrow \{i \in C_r : L_i^{F,(r)} \ge \varepsilon\}$, $C_{r+1} \leftarrow C_r \setminus (\mathcal{I}_r \cup \mathcal{O}_r)$

  if $q = \texttt{one}$ **then**

    if $\mathcal{I}_r \neq \emptyset$ **then return** any $\widehat{i} \in \mathcal{I}_r$

    else if $C_{r+1} = \emptyset$ **then return** $\emptyset$

    end if

  end if

  if $q = \texttt{all}$ **then**

    $\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \cup \mathcal{I}_r$

    if $C_{r+1} = \emptyset$ **then return** $\widehat{\mathcal{S}}$

    end if

  else $C_{r+1} \leftarrow C_r \setminus \mathcal{O}_r$

  end if

  $K_{m,r+1} \leftarrow K_{m,r}^+ \cap C_{r+1}$ for $m \in \{R, D\}$

end for

arms i, j in the witness set, while exploiting the fused gaps to determine which modality forms the bottleneck for each arm in the candidate set. Finally, it checks the stopping condition corresponding to either the $\varepsilon^{\mathrm{one}}$-BAI or $\varepsilon^{\mathrm{all}}$-BAI objective and returns the required output once the condition is satisfied. Details are attached in Appendix C.

3.1 FUSED ELIMINATION

The fused elimination mechanism used by WIT-BOTELIM consists of two complementary components: *witness elimination* and *bottleneck elimination*. We now elaborate them as follows.

We first describe the *witness elimination* rule. At the beginning of phase r, let $C_r \subseteq [K]$ denote the arms whose fused status remains unresolved. For each modality $m \in \{R, D\}$, we maintain a *witness set* $W_{m,r}$ containing arms that have not yet been certified as strictly dominated under modality m. A witness $j \in W_{m,r-1}$ is removed whenever there exists $k \in W_{m,r-1} \setminus \{j\}$ such that $(x_k - x_j)^\top \widehat{\theta}_{m,r} - \mathrm{CR}_{m,r}(k,j) > 0$. Here $\mathrm{CR}_{m,r}(k,j)$ is a simultaneous confidence radius for the direction $x_k - x_j$; its reward and dueling forms are given in Appendix C.4. The purpose of witness

pruning is to remove competitors that can no longer determine a modality-wise gap while retaining those that remain informative for eliminating unresolved candidates. In particular, a modality-wise optimal arm may remain an essential witness even when it is not jointly ε-optimal. For example, if $j = i_R^\star \notin \mathcal{S}_\varepsilon$, arm j must still be retained because it determines the reward-side gap of every candidate, whereas reward-suboptimal witnesses that are confidently dominated can be discarded. Hence witness elimination compresses the competitor side of the phase-wise design problem without removing the arms needed to evaluate the remaining modality-wise gaps.

Secondly, the *bottleneck* elimination rule exploits the fused gap $\Delta_i = \max\{\Delta_{R,i}, \Delta_{D,i}\}$. Even when an arm i remains unresolved, both modality-wise gaps need not be refined indefinitely: once one modality is known to determine their maximum, improving the estimate of the other modality can no longer change the fused gap. We therefore maintain *bottleneck sets* $K_{R,r}$ and $K_{D,r}$. Membership $i \in K_{m,r}$ indicates that modality m has already been certified to satisfy $\Delta_{m,i} > \Delta_{\bar{m},i}$ and hence controls the fused gap, $\Delta_i = \Delta_{m,i}$. Once such a certificate is obtained, the opposite modality no longer needs to treat arm i as an active target. Consequently, modality m only needs to resolve $G_{m,r} := C_r \setminus K_{\bar{m},r}$. Thus, while witness elimination reduces the set of relevant competitors, bottleneck elimination reduces the set of relevant target–modality pairs. Together, they remove comparisons that are already irrelevant to the fused decision and concentrate subsequent sampling on the directions that can still affect the final classification.

Combining these two elimination rules yields the phase-r comparison set

$$\mathcal{Y}_{m,r} := \{x_j - x_i : i \in G_{m,r},\ j \in W_{m,r-1},\ j \neq i\}, \qquad m \in \{R, D\}. \tag{1}$$

Thus, $G_{m,r}$ performs target-side bottleneck compression, whereas $W_{m,r-1}$ performs competitor-side witness compression. Importantly, the previous-phase witness set is used in equation 1, so the entire phase-r allocation is fixed before observing the fresh data collected in that phase.

After the fresh batch is collected, the witness set is updated using the strict-domination rule above. For every $i \in G_{m,r}$, define $\widehat{\Delta}_{m,i}^{(r)} := \max_{j \in W_{m,r}} (x_j - x_i)^\top \widehat{\theta}_{m,r}$, and construct $L_{m,i}^{(r)} := \max\{0, \widehat{\Delta}_{m,i}^{(r)} - \gamma_r\}, U_{m,i}^{(r)} := \widehat{\Delta}_{m,i}^{(r)} + \gamma_r$.

For an arm whose controlling modality has not yet been identified, modality m receives a new bottleneck certificate whenever $U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)}$. Accordingly, define $K_{m,r}^{+} := K_{m,r} \cup \left\{ i \in C_r \setminus (K_{R,r} \cup K_{D,r}) : U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)} \right\}$. Once obtained, a bottleneck certificate persists as long as the arm remains unresolved, so that $K_{m,r+1} = K_{m,r}^{+} \cap C_{r+1}$.

For the confidence allocation, let $\delta_r := \frac{3\delta}{\pi^2 r^2}, \delta_{m,r}^{\mathrm{gap}} = \delta_{m,r}^{\mathrm{wit}} = \frac{\delta_r}{4}$ for $m \in \{R, D\}$. The following lemma establishes the statistical guarantees for every phase, confirming the sound performance of our algorithm. The detailed proof is deferred to Appendix C.5.

Lemma 2 (Phase-wise fused-elimination guarantee)**.** *Fix any phase r. Conditioned on the history before phase r and on the validity of the logistic reference point inherited from the preceding phases, with probability at least $1 - \delta_r$, the following statements hold simultaneously for both $m \in \{R, D\}$:*

1. *for every $i \in G_{m,r}$, $\left|\widehat{\Delta}_{m,i}^{(r)} - \Delta_{m,i}\right| \le \gamma_r, \qquad \widehat{\Delta}_{m,i}^{(r)} := \max_{j \in W_{m,r}} (x_j - x_i)^\top \widehat{\theta}_{m,r}$;*
2. *every witness removed from $W_{m,r-1}$ is strictly suboptimal under modality m; in particular, $i_m^\star \in W_{m,r}$;*
3. *if $U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)}$, then $\Delta_{\bar{m},i} < \Delta_{m,i}$ and therefore $\Delta_i = \Delta_{m,i}$.*

Consequently, on the same event, $\Delta_{m,i} \in [L_{m,i}^{(r)}, U_{m,i}^{(r)}]$ for every active target, and every witness or bottleneck elimination made in phase r is consistent with the true modality gaps.

3.2 Upper Bound of Wit-BotElim

We next establish an upper bound on the sample complexity of Wit-BotElim. To state the bound, let $\Phi := \{x_i - x_j : i, j \in [K], i < j\}$ and set $\overline{G}_{R,1} = \overline{G}_{D,1} := [K]$. For each $r \ge 2$ and

$m \in \{R, D\}$, define

$$\overline{G}_{m,r} := \{i \in [K] : \Delta_{R,i} \le \varepsilon + 2\gamma_{r-1},\ \Delta_{D,i} \le \varepsilon + 2\gamma_{r-1},\ \Delta_{\bar{m},i} - \Delta_{m,i} \le 4\gamma_{r-1}\}$$

and $\overline{\mathcal{Y}}_{m,r} := \{x_j - x_i : i \in \overline{G}_{m,r},\ j \in [K],\ j \ne i\}$, when the latter is the corresponding deterministic direction envelope. These sets depend only on the underlying bandit instance and are independent of the realized history of the algorithm.

For every phase $r \ge 1$, define

$$\Gamma_r := 64\left(d\log 6 + \log\frac{8\pi^2 r^2 \max\{K, |\Phi|\}(2+|\Phi|)^2}{3\delta}\right).$$

When $\overline{\mathcal{Y}}_{R,r} \ne \emptyset$,
define $\overline{P}_{R,r} := \min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \overline{\mathcal{Y}}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}}$, and set $\overline{P}_{R,r} := 0$ when $\overline{\mathcal{Y}}_{R,r} = \emptyset$. Similarly, when $\overline{\mathcal{Y}}_{D,r} \ne \emptyset$, define

$$\overline{P}_{D,r} := \min_{\nu_D \in \mathfrak{D}_D} \max\left\{\Gamma_r \max_{\phi \in \Phi} \|\phi\|^2_{H_D(\nu_D,\theta_D)^{-1}},\ \frac{(2.4)^2\, 4^r}{\gamma^2} \max_{y \in \overline{\mathcal{Y}}_{D,r}} \|y\|^2_{H_D(\nu_D,\theta_D)^{-1}}\right\},$$

and set $\overline{P}_{D,r} := 0$ when $\overline{\mathcal{Y}}_{D,r} = \emptyset$. For ε^{one} identification, define

$$\rho_{\text{one}} := \begin{cases} \max_{i \in \mathcal{S}_\varepsilon}(\varepsilon - \Delta_i), & \mathcal{S}_\varepsilon \ne \emptyset, \\ \min_{i \in [K]}(\Delta_i - \varepsilon) & \mathcal{S}_\varepsilon = \emptyset \end{cases}.$$

When $\mathcal{S}_\varepsilon = \emptyset$, we assume $\rho_{\text{one}} > 0$ to obtain a finite-phase guarantee. For ε^{all} identification, define $\rho_{\text{all}} := \min_{i \in [K]} |\varepsilon - \Delta_i|$, and assume $\rho_{\text{all}} > 0$ for the corresponding finite-phase guarantee. For $q \in \{\text{one}, \text{all}\}$, define $R_q := \min\{r \ge 1 : 2\gamma_r < \rho_q\} = \max\left\{1, 1 + \left\lfloor \log_2 \frac{2\gamma}{\rho_q} \right\rfloor\right\}$.

Equipped with these quantities, our WIT-BOTELIM algorithm can efficiently complete the ε-BAI task. Its performance guarantee is established as below, with the proof deferred to Appendix C.

Theorem 3 (PAC guarantee and sample complexity)**.** *Let $\delta < 1/e$ and $\zeta > 0$, and suppose the standing full-rank design assumptions hold. Assume also that $\|\phi\|_2 \le 1$ for every $\phi \in \Phi$ and that $\dot{\mu}(\phi^\top \theta_D) \ge \kappa_0 > 0, \forall \phi \in \Phi$. For $q \in \{\text{one}, \text{all}\}$ satisfying the separation conditions above, with probability at least $1 - \delta$,* WIT-BOTELIM *returns the correct target with upper bound of* WIT-BOTELIM

$$\tau^q_{\text{WIT-BOTELIM}} = O\Bigg(R_q \frac{d^2}{\zeta} + (1+\zeta)\kappa_0^{-1} d^2 \log\frac{|\Phi|}{\delta} + (1+\zeta)\kappa_0^{-1} d \log^2\frac{|\Phi|}{\delta} + (1+\zeta)\sum_{r=1}^{R_q}\left[\frac{4^r \overline{P}_{R,r}}{\gamma^2}\log\frac{r^2(1+|\overline{\mathcal{Y}}_{R,r}|)}{\delta} + \overline{P}_{D,r}\log\frac{r^2\max\{K,|\Phi|\}}{\delta}\right]\Bigg).$$

For ε^{one} identification, correctness means that the returned arm belongs to $\mathcal{S}_\varepsilon$ whenever $\mathcal{S}_\varepsilon \ne \emptyset$, and that the algorithm returns $\emptyset$ when $\mathcal{S}_\varepsilon = \emptyset$. For ε^{all} identification, correctness means exact recovery of $\mathcal{S}_\varepsilon$.

Comparison with existing works. Compared with the phased transductive elimination mechanisms underlying RAGE (Fiez et al., 2019) and RAGE-GLM (Jun et al., 2021), the upper bound of WIT-BOTELIM shares a similar structure with those of RAGE and RAGE-GLM, corroborating the soundness of our analysis.

Comparison to the lower bound in Theorem 1. We next specialize the upper and lower bounds to a regime where the unique jointly ε-optimal arm is explicit. Without loss of generality, let $i^\star_R = 1$, $i^\star_D = 2$, and $\mathcal{S}_\varepsilon = \{3\}$, while allowing an arbitrary number of arms K. Define $\eta_R := \varepsilon - \Delta_{R,3} > 0$ and $\eta_D := \varepsilon - \Delta_{D,3} > 0$, and assume that $|\varepsilon - \Delta_{m,3}| \le |\varepsilon - \Delta_{m,i}|, \forall m \in \{R, D\},\ i \in [K]$. In this case, arm 3 is closest to the modality-wise decision boundary under both feedback types. Such

a regime can arise when coarse prior or pilot estimates of the modality-wise gaps are available and ε is chosen near the estimated gaps of an arm that appears robust under both modalities.

When $\mathcal{S}_\varepsilon = \{3\}$, Theorem 1 indicates, for either target mode, $\mathbb{E}^\pi_{\theta_R,\theta_D}[\tau^\pi] = \Omega\left((H_3^R + H_3^D)\log\frac{1}{\delta}\right)$ as $\delta \downarrow 0$. For comparison with the directional terms in the upper bound, define $J_R := \inf_{\nu_R \in \mathfrak{D}_R} \max_{k\neq 3} \frac{\|x_k - x_3\|^2_{M_R(\nu_R)^{-1}}}{\eta_R^2}$, $J_D := \inf_{\nu_D \in \mathfrak{D}_D} \max_{k\neq 3} \frac{\|x_k - x_3\|^2_{M_D(\nu_D)^{-1}}}{\eta_D^2}$. After irrelevant overheads, Appendix C.11 shows that the directional contribution of Theorem 3 is $\widetilde{O}\left((J_R + \kappa_0^{-1} J_D)\log\frac{1}{\delta}\right)$.

Moreover, $H_3^m \leq J_m$, while under the closest-challenger dominance condition of Appendix C.11, over instance families for which c_R, c_D, and κ_0^{-1} are uniformly bounded, the dominant directional upper and lower bounds have the same instance-dependent order up to logarithmic factors. The comparison concerns the directional terms only: the rounding, burn-in, and global Fisher-leverage costs remain additional overhead in the upper bound. Altogether, we observe the near-optimality of our WIT-BOTELIM in this specialized regime.

Remark 4. We note that a lower bound on $\mathbb{E}_{\theta_R,\theta_D}[\tau^\pi]$ should not be directly compared with a high-probability upper bound of the form $\Pr(\tau^\pi > T) \leq 1 - \delta$ without an additional comparability assumption. However, since prior work often established that these two types of sample complexity bounds are usually of the same order (Hou et al., 2023), we make the comparison under this assumption.

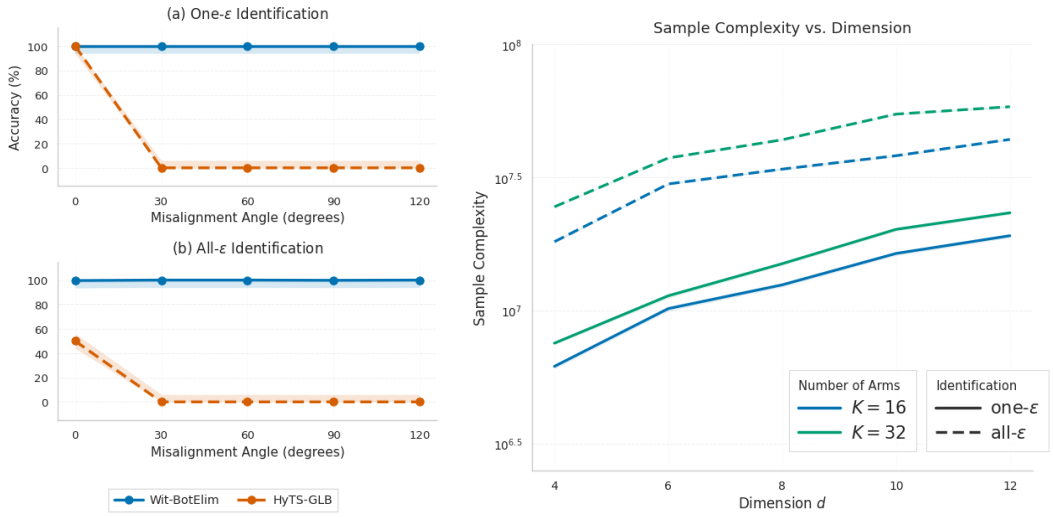
4 EXPERIMENTS

This section evaluates WIT-BOTELIM along two complementary dimensions: robustness to heterogeneous feedback, and scaling with problem size. To examine robustness to model heterogeneity, we compare against HYTS-GLB from Zeng et al. (2026, Algorithm 1), which models reward and dueling feedback through a shared latent parameter. This comparison serves as a model-misspecification stress test: when $\theta_R \neq \theta_D$, the shared-parameter assumption is violated, allowing us to examine whether an algorithm designed under aligned feedback remains reliable when the two modalities encode distinct preferences. We separately vary the feature dimension and number of arms to study how the sample complexity of WIT-BOTELIM scales with problem size. Detailed instance construction, parameter choices, additional diagnostics, and mechanism ablations are provided in Appendix D.

Figure 1a reports empirical identification accuracy as the angle between θ_R and θ_D increases. In the aligned case, the shared-parameter model is compatible with the data, and both methods succeed in the ε^{one} task. Once the reward and dueling parameters are misaligned by $30°$, $60°$, $90°$, or $120°$, however, the empirical success rate of HYTS-GLB drops to zero on our instances, whereas WIT-BOTELIM remains essentially perfectly accurate. It succeeds in all 500 trials at every tested angle for ε^{one} identification and achieves exact ε^{all} recovery in all trials except one at $90°$. These empirical success rates should be distinguished from the prescribed confidence level $1 - \delta = 0.95$: the latter is the target fixed-confidence guarantee, whereas the figure reports observed frequencies over repeated trials.

The realized query allocation provides additional insight into this behavior. On these instances, HYTS-GLB allocates its queries to reward feedback and issues no dueling queries. Under its shared-parameter model, information from one modality can in principle identify the common latent parameter; when θ_R and θ_D are distinct, however, reward observations alone need not certify performance under the dueling parameter. In contrast, WIT-BOTELIM explicitly certifies the relevant directions under both modalities and therefore does not require the two feedback sources to induce the same parameter or ranking. Additional fine-grained misalignment results and the corresponding geometric construction are given in Appendix D.

Figure 1b studies the scaling behavior of WIT-BOTELIM by varying $d \in \{4, 6, 8, 10, 12\}$ and $K \in \{16, 32\}$. The purpose of this study is to verify how its query complexity changes as the underlying structured identification problem becomes larger. For both values of K and for both identification objectives, mean sample complexity increases monotonically with the feature dimension. Increasing d from 4 to 12 increases the mean sample complexity by approximately a factor of 3.1 for ε^{one} identification and by approximately a factor of 2.4 for ε^{all} identification.



(a) Identification accuracy under increasing parameter misalignment.

(b) Sample complexity as the feature dimension and number of arms increase.

Figure 1: Robustness to heterogeneous feedback and scaling of WIT-BOTELIM with problem size. Shaded regions denote confidence intervals over repeated trials.

The number of arms has a consistent effect as well. At every tested dimension, the $K = 32$ instances require more samples than their $K = 16$ counterparts. Depending on d, doubling the number of arms increases mean sample complexity by approximately 12%–23% for ε^{one} identification and 25%–43% for ε^{all} identification. These results indicate that the empirical query cost responds systematically to both the geometric dimension and the number of candidate arms, rather than being dominated by a single fixed overhead.

Exact-set recovery is consistently more demanding than returning a single acceptable arm. Across the tested configurations, ε^{all} identification requires substantially more samples than ε^{one} identification because every potentially relevant arm must be classified under both feedback modalities, whereas the one-arm objective may terminate as soon as a single jointly acceptable arm is certified. The appendix further examines how the witness and bottleneck mechanisms reduce the number of active comparison directions during this certification process. In particular, witness pruning primarily reduces the typical within-phase workload, while bottleneck pruning becomes increasingly important for expensive late-phase trajectories in exact-set recovery.

Overall, the empirical results support three aspects of the proposed method. First, explicitly modeling heterogeneous reward and dueling parameters is important for correctness when the two feedback modalities disagree. Second, the sample complexity of WIT-BOTELIM increases systematically with both feature dimension and arm-set size, while remaining well behaved over the tested range. Furthermore, the mechanism-level analysis in Appendix D shows that witness and bottleneck pruning reduce the set of directions that must be resolved, explaining the observed savings in costly feedback queries.

5 CONCLUSION

This paper studies the fusion of reward and dueling feedback in structured bandits when the two feedback types are governed by different latent models. We establish a sample-complexity lower bound for this setting and propose WIT-BOTELIM, an efficient adaptive algorithm that combines the two modalities through witness-based elimination and a bottleneck mechanism that safely removes directions that no longer need to be resolved. Our theoretical and empirical results show the efficacy of our proposd WIT-BOTELIM algorithm, which also indicating the benefit of unaligned fused feedback that allows information obtained from one modality to simplify the decision problem faced by the other.

AI Use Statement. Generative AI tools, including ChatGPT, were used as interactive writing assistants during the preparation of this work. Their use included providing feedback on the formulation and presentation of the theoretical framework, assisting with the review and refinement of mathematical claims and proof arguments, supporting literature searches, and editing the manuscript. The authors independently reviewed and verified all statements and references included in the final manuscript. Any AI-generated or AI-suggested code was inspected and tested by the authors, and any suggested mathematical arguments were independently validated before inclusion. The authors take full responsibility for the content and correctness of this paper.

Reproducibility Statement. We provide the information needed to reproduce the theoretical and empirical results of this work. The main text specifies the problem setting, assumptions, and feedback models, and gives a complete description of WIT-BOTELIM. The appendix contains complete proofs of all theoretical results, including the upper and lower bounds and the supporting lemmas. Upon publication, we will release a code repository containing the implementation and scripts needed to reproduce all experiments.

REFERENCES

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 41–53, 2010.

Ting Cai and Kirthevasan Kandasamy. Constrained best arm identification with tests for feasibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19808–19816, 2026. doi: 10.1609/aaai.v40i24.39063.

Zhirui Chen, P. N. Karthik, Yeow Meng Chee, and Vincent Y. F. Tan. Optimal multi-objective best arm identification with fixed confidence. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1162–1170. PMLR, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10666–10676, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016.

Yunlong Hou, Vincent Y. F. Tan, and Zixin Zhong. Almost optimal variance-constrained best arm identification. *IEEE Transactions on Information Theory*, 69(4):2603–2634, 2023. doi: 10.1109/TIT.2022.3222231.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5148–5157. PMLR, 2021.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1238–1246. PMLR, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Abbas Kazerouni and Lawrence M. Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021. doi: 10.1016/j.orl.2021.03.011.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Constrained pareto set identification with bandit feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31342–31378. PMLR, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 27, pp. 828–836, 2014.

Xuchuang Wang, Qirun Zeng, Jinhang Zuo, Xutong Liu, Mohammad Hajiesmaili, John C. S. Lui, and Adam Wierman. Fusing reward and dueling feedback in stochastic bandits. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65346–65367. PMLR, 2025a.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *International Conference on Learning Representations*, 2025b.

Qirun Zeng, Xuchuang Wang, Jiayi Shen, Xutong Liu, Fang Kong, and Jinhang Zuo. Best arm identification in generalized linear bandits via hybrid feedback, 2026.

A Discussion on Further Work

Beyond the motivating post-training example, similar forms of heterogeneous feedback arise broadly in practice. For instance, in A/B testing systems such as those used for conversational AI services, companies may collect both absolute human ratings, corresponding to *reward* feedback, and pairwise preferences between two candidate responses, corresponding to *dueling* feedback, to support service-level decisions. A related interaction also appears in modern training pipelines. For example, LLM training often combines supervised fine-tuning (SFT), which is driven by demonstration-based supervision, with preference-based methods such as DPO (Rafailov et al., 2024), which rely on pairwise comparisons. Because these stages are governed by different supervision signals and optimization objectives, they need not induce the same notion of response quality or rank candidate behaviors consistently. Nevertheless, both sources of feedback contribute to the same downstream model-selection or improvement objective, providing a natural example of why effectively combining potentially unaligned feedback modalities can be useful. Extending this principle to richer sequential decision-making and post-training systems is a natural direction for future work.

B Proofs of Lower Bound

Throughout this appendix, we restrict the reward noise to $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. This is a valid subclass of the 1-sub-Gaussian reward model considered in Section 2. For an algorithm π, let N_a^R denote the number of reward queries of arm a, and let $N_{j,k}^D$ denote the number of dueling queries of pair $(j,k) \in \mathcal{P}$, both counted before τ^π. Since exactly one query is made at each round,

$$\tau^\pi = \sum_{a\in[K]} N_a^R + \sum_{(j,k)\in\mathcal{P}} N_{j,k}^D.$$

Recall that $M_R(\nu_R) = \sum_a \nu_{R,a} x_a x_a^\top$ and $M_D(\nu_D) = \sum_{(j,k)\in\mathcal{P}} \nu_{D,j,k} x_{j,k} x_{j,k}^\top$. All experimental-design optimizations below are restricted to $\mathfrak{D}_R$ and $\mathfrak{D}_\mathcal{P}$, respectively, so the corresponding design matrices are positive definite.

B.1 Change of Measure

The proof uses the standard transportation argument for adaptive bandit experiments (Kaufmann et al., 2016).

Lemma 5 (Stopped change of measure)**.** *For any alternative parameters $\theta'_R, \theta'_D \in \mathbb{R}^d$ and every $E \in \mathcal{F}_{\tau^\pi}$,*

$$\begin{aligned}
&\frac{1}{2}\sum_{a\in[K]} \mathbb{E}^\pi_{\theta_R,\theta_D}[N_a^R]\big(x_a^\top(\theta_R-\theta'_R)\big)^2 \\
&\quad + \sum_{(j,k)\in\mathcal{P}} \mathbb{E}^\pi_{\theta_R,\theta_D}[N_{j,k}^D]\,\mathrm{kl}\left(\mu(x_{j,k}^\top\theta_D), \mu(x_{j,k}^\top\theta'_D)\right) \\
&\geq \mathrm{kl}\left(\mathbb{P}^\pi_{\theta_R,\theta_D}(E), \mathbb{P}^\pi_{\theta'_R,\theta'_D}(E)\right).
\end{aligned}$$

Proof. Consider the complete interaction transcript stopped at τ^π. Since the sampling rule at round t is measurable with respect to $\mathcal{F}_{t-1}$, conditional on the same history the distribution used to select the next query is identical under the true and alternative instances. Hence only the conditional observation distributions contribute to the likelihood ratio.

For a reward query of arm a,

$$\mathrm{kl}\left(\mathcal{N}(x_a^\top\theta_R, 1), \mathcal{N}(x_a^\top\theta'_R, 1)\right) = \frac{1}{2}\big(x_a^\top(\theta_R-\theta'_R)\big)^2.$$

For a dueling query (j,k), the corresponding contribution is $\mathrm{kl}(\mu(x_{j,k}^\top\theta_D), \mu(x_{j,k}^\top\theta'_D))$. Applying the chain rule for KL divergence to the stopped transcript and then the data-processing inequality to $\mathbf{1}_E$ proves the claim. □

We next record the quadratic projection used to construct the alternative parameters.

Lemma 6 (Half-space projection)**.** *For $M \succ 0$, $y \in \mathbb{R}^d$, and $g > 0$,*

$$\inf_{h:\, y^\top h \geq g} h^\top M h = \frac{g^2}{\|y\|^2_{M^{-1}}}.$$

Proof. Cauchy–Schwarz in the M-inner product gives $(y^\top h)^2 \leq (h^\top M h)(y^\top M^{-1} y)$. Hence every feasible h satisfies $h^\top M h \geq g^2/(y^\top M^{-1} y)$. Equality is attained at

$$h^\star = \frac{g}{y^\top M^{-1} y} M^{-1} y.$$

□

Although the design complexities are optimized only over full-rank designs, the expected sampling proportions induced by an arbitrary algorithm need not themselves be full rank. The following regularization resolves this distinction.

Lemma 7 (Full-rank regularization)**.** *Let $\bar{\nu}$ be any probability distribution over the available measurements and let ν_0 be any full-rank design. For $\alpha \in (0, 1)$, define*

$$\nu_\alpha := (1 - \alpha)\bar{\nu} + \alpha \nu_0.$$

Then ν_α is full rank and

$$M(\bar{\nu}) \preceq \frac{1}{1 - \alpha} M(\nu_\alpha).$$

Moreover, for every nonnegative function L on the measurements,

$$\sum_q \bar{\nu}_q L(q) \leq \frac{1}{1 - \alpha} \sum_q \nu_{\alpha,q} L(q).$$

Proof. By construction, $M(\nu_\alpha) = (1 - \alpha) M(\bar{\nu}) + \alpha M(\nu_0)$, and $M(\nu_0) \succ 0$. Hence $M(\nu_\alpha) \succ 0$ and $M(\nu_\alpha) \succeq (1 - \alpha) M(\bar{\nu})$. The second statement follows identically from $\nu_{\alpha,q} \geq (1 - \alpha)\bar{\nu}_q$ and $L(q) \geq 0$. □

B.2 Nonempty Joint Answer Set

Suppose $\mathcal{S}_\varepsilon \neq \emptyset$. For every $i \in \mathcal{S}_\varepsilon$, recall

$$H_i^R := \inf_{\nu_R \in \mathfrak{D}_R} \max_{k \neq i} \frac{\|x_k - x_i\|^2_{M_R(\nu_R)^{-1}}}{\big(\varepsilon - (x_k - x_i)^\top \theta_R\big)^2}$$

and

$$H_i^D := \inf_{\nu_D \in \mathfrak{D}_\mathcal{P}} \max_{k \neq i} \frac{\|x_k - x_i\|^2_{M_D(\nu_D)^{-1}}}{\big(\varepsilon - (x_k - x_i)^\top \theta_D\big)^2}.$$

Because $i \in \mathcal{S}_\varepsilon$, $(x_k - x_i)^\top \theta_m < \varepsilon$ for every k and both modalities.

Lemma 8 (Boundary alternatives)**.** *Fix $i \in \mathcal{S}_\varepsilon$ and $k \neq i$. If $(x_k - x_i)^\top \theta'_R \geq \varepsilon$, then $i \notin \mathcal{S}_\varepsilon(\theta'_R, \theta_D)$. Likewise, if $(x_k - x_i)^\top \theta'_D \geq \varepsilon$, then $i \notin \mathcal{S}_\varepsilon(\theta_R, \theta'_D)$.*

Proof. For the reward alternative,

$$\max_{j \in [K]} (x_j - x_i)^\top \theta'_R \geq (x_k - x_i)^\top \theta'_R \geq \varepsilon.$$

Hence arm i violates the strict requirement $\Delta_{R,i} < \varepsilon$. The dueling case is identical. □

B.2.1 ε^{ALL}-OPTIMAL IDENTIFICATION

Proposition 9 (Finite-confidence lower bound)**.** *If $0 < \delta < 1/2$, every (ε, δ)-PAC algorithm for ε^{all}-optimal identification satisfies*

$$\mathbb{E}^{\pi}_{\theta_R,\theta_D}[\tau^{\pi}] \geq 2\,\mathrm{kl}(1-\delta,\delta)\left[\max_{i\in\mathcal{S}_\varepsilon} H_i^R + 4\max_{i\in\mathcal{S}_\varepsilon} H_i^D\right].$$

Proof. Let $E := \{\widehat{\mathcal{S}}^{\pi} = \mathcal{S}_\varepsilon\}$. PAC correctness gives $\mathbb{P}^{\pi}_{\theta}(E) \geq 1-\delta$. Every alternative in Lemma 8 changes the correct target set, and therefore $\mathbb{P}^{\pi}_{\theta'}(E) \leq \delta$.

We first consider reward feedback. Let $n_R := \sum_a \mathbb{E}[N_a^R]$. If $n_R = +\infty$, the reward-side claim is immediate. Otherwise define $\bar{\nu}_{R,a} := \mathbb{E}[N_a^R]/n_R$. Fix any $\nu_R^0 \in \mathfrak{D}_R$ and set $\nu_{R,\alpha} := (1-\alpha)\bar{\nu}_R + \alpha\nu_R^0$.

Fix $i \in \mathcal{S}_\varepsilon$ and $k \neq i$, and write $y := x_k - x_i$ and $g := \varepsilon - y^\top\theta_R > 0$. With $M_\alpha := M_R(\nu_{R,\alpha})$, define

$$h_\alpha := \frac{g}{y^\top M_\alpha^{-1} y} M_\alpha^{-1} y, \qquad \theta'_R := \theta_R + h_\alpha.$$

Then $y^\top\theta'_R = \varepsilon$, so Lemma 8 applies.

By Lemma 7,

$$\begin{aligned}
&\frac{1}{2}\sum_a \mathbb{E}[N_a^R]\big(x_a^\top h_\alpha\big)^2 \\
&= \frac{n_R}{2} h_\alpha^\top M_R(\bar{\nu}_R) h_\alpha \\
&\leq \frac{n_R}{2(1-\alpha)} h_\alpha^\top M_\alpha h_\alpha \\
&= \frac{n_R}{2(1-\alpha)} \frac{g^2}{y^\top M_\alpha^{-1} y}.
\end{aligned}$$

Combining this with Lemma 5 gives

$$n_R \geq 2(1-\alpha)\,\mathrm{kl}(1-\delta,\delta)\frac{\|x_k - x_i\|^2_{M_R(\nu_{R,\alpha})^{-1}}}{\big(\varepsilon - (x_k-x_i)^\top\theta_R\big)^2}.$$

This holds for every $k \neq i$. Since $\nu_{R,\alpha} \in \mathfrak{D}_R$,

$$n_R \geq 2(1-\alpha)\,\mathrm{kl}(1-\delta,\delta) H_i^R.$$

The same conclusion holds for every $i \in \mathcal{S}_\varepsilon$. Letting $\alpha \downarrow 0$ gives

$$n_R \geq 2\,\mathrm{kl}(1-\delta,\delta)\max_{i\in\mathcal{S}_\varepsilon} H_i^R.$$

We next consider dueling feedback. Let $n_D := \sum_{(j,\ell)\in\mathcal{P}} \mathbb{E}[N_{j,\ell}^D]$ and normalize the expected counts to $\bar{\nu}_D$. Fix any $\nu_D^0 \in \mathfrak{D}_\mathcal{P}$ and let $\nu_{D,\alpha} := (1-\alpha)\bar{\nu}_D + \alpha\nu_D^0$.

For fixed $i \in \mathcal{S}_\varepsilon$ and $k \neq i$, write $y := x_k - x_i$ and $g := \varepsilon - y^\top\theta_D > 0$. Let $M_\alpha := M_D(\nu_{D,\alpha})$ and take the quadratic projection

$$h_\alpha := \frac{g}{y^\top M_\alpha^{-1} y} M_\alpha^{-1} y, \qquad \theta'_D := \theta_D + h_\alpha.$$

Again $y^\top\theta'_D = \varepsilon$, so this is a valid boundary alternative.

For the logistic model, let $b(z) := \log(1+e^z)$. Since $b''(z) = \mu(z)(1-\mu(z)) \leq 1/4$, the Bernoulli KL divergence in natural parameters satisfies

$$\mathrm{kl}(\mu(u),\mu(v)) \leq \frac{1}{8}(u-v)^2.$$

Using Lemma 7 and the quadratic projection above, the dueling information under the true expected sampling proportions is at most

$$\frac{n_D}{8(1-\alpha)} \frac{g^2}{y^\top M_\alpha^{-1} y}.$$

Lemma 5 therefore yields

$$n_D \geq 8(1-\alpha)\,\mathrm{kl}(1-\delta,\delta) \frac{\|x_k - x_i\|^2_{M_D(\nu_{D,\alpha})^{-1}}}{\left(\varepsilon - (x_k - x_i)^\top \theta_D\right)^2}.$$

Taking the maximum over $k \neq i$, using $\nu_{D,\alpha} \in \mathfrak{D}_\mathcal{P}$, then maximizing over $i \in \mathcal{S}_\varepsilon$ and letting $\alpha \downarrow 0$ gives

$$n_D \geq 8\,\mathrm{kl}(1-\delta,\delta) \max_{i \in \mathcal{S}_\varepsilon} H_i^D.$$

Finally, $\mathbb{E}[\tau^\pi] = n_R + n_D$. Adding the reward and dueling requirements proves the proposition. □

B.2.2 ε^{ALL}-BAI PROBLEM

Proposition 10 (Finite-confidence lower bound)**.** *Let $s := |\mathcal{S}_\varepsilon| > 0$. If $0 < \delta < 1/(s+1)$, every (ε,δ)-PAC algorithm for ε^{all}-BAI problem satisfies*

$$\mathbb{E}^\pi_{\theta_R,\theta_D}[\tau^\pi] \geq 2\,\mathrm{kl}\left(\frac{1-\delta}{s},\delta\right) \min_{i \in \mathcal{S}_\varepsilon} \left(H_i^R + 4H_i^D\right).$$

Proof. For $i \in \mathcal{S}_\varepsilon$, let $p_i := \mathbb{P}^\pi_\theta(\widehat{i}^\pi = i)$. PAC correctness implies $\sum_{i \in \mathcal{S}_\varepsilon} p_i \geq 1-\delta$. Hence there exists $i_\star \in \mathcal{S}_\varepsilon$ such that

$$p_{i_\star} \geq \frac{1-\delta}{s}.$$

Since $\delta < 1/(s+1)$, $(1-\delta)/s > \delta$.

Let $E_\star := \{\widehat{i}^\pi = i_\star\}$. Under any boundary alternative from Lemma 8, $i_\star$ is no longer a valid recommendation. Therefore PAC correctness gives $\mathbb{P}^\pi_{\theta'}(E_\star) \leq \delta$. Monotonicity of binary KL over $p \geq (1-\delta)/s > \delta \geq q$ yields

$$\mathrm{kl}\left(\mathbb{P}^\pi_\theta(E_\star), \mathbb{P}^\pi_{\theta'}(E_\star)\right) \geq \mathrm{kl}\left(\frac{1-\delta}{s},\delta\right).$$

Repeating the regularized reward-side argument in the proof of Proposition 9 gives

$$n_R \geq 2\,\mathrm{kl}\left(\frac{1-\delta}{s},\delta\right) H^R_{i_\star}.$$

Likewise, the dueling-side argument gives

$$n_D \geq 8\,\mathrm{kl}\left(\frac{1-\delta}{s},\delta\right) H^D_{i_\star}.$$

Adding the two inequalities and using $i_\star \in \mathcal{S}_\varepsilon$ yields

$$\begin{aligned}\mathbb{E}[\tau^\pi] &\geq 2\,\mathrm{kl}\left(\frac{1-\delta}{s},\delta\right)\left(H^R_{i_\star} + 4H^D_{i_\star}\right) \\ &\geq 2\,\mathrm{kl}\left(\frac{1-\delta}{s},\delta\right) \min_{i \in \mathcal{S}_\varepsilon}\left(H_i^R + 4H_i^D\right).\end{aligned}$$

This proves the proposition and, together with Proposition 9, proves Theorem 1. □

B.3 Empty Joint Answer Set

We now consider $\mathcal{S}_\varepsilon = \emptyset$. Unlike the nonempty case, the true all-arm target is $\emptyset$, so an alternative must make at least one arm jointly ε-optimal.

For each modality $m \in \{R, D\}$, fix $i^\star_m \in \arg\max_{j\in[K]} x_j^\top \theta_m$, and choose

$$i^\circ \in \arg\min_{i\in[K]} \left[\max_{m\in\{R,D\}} (x_{i^\star_m} - x_i)^\top \theta_m - \varepsilon \right].$$

For every modality satisfying $(x_{i^\star_m} - x_{i^\circ})^\top \theta_m > \varepsilon$, write

$$y_m := x_{i^\star_m} - x_{i^\circ}, \qquad g_m := y_m^\top \theta_m - \varepsilon > 0.$$

For a full-rank design of modality m, the quadratic minimum-information projection onto $y_m^\top \vartheta = \varepsilon$ is

$$\bar{\theta}_m(\nu_m) := \theta_m - \frac{g_m}{y_m^\top M_m(\nu_m)^{-1} y_m} M_m(\nu_m)^{-1} y_m.$$

Although this projection minimizes the quadratic cost of reaching the boundary, it need not preserve the current modality-wise optimal arm. We therefore impose the following condition.

Definition 11 (Winner-preserving projection)**.** *The empty-set instance satisfies the winner-preserving projection condition for i° if, for every modality $m \in \{R, D\}$ satisfying $g_m > 0$, every full-rank design ν_m of that modality, and every $j \neq i^\star_m$,*

$$(x_{i^\star_m} - x_j)^\top \bar{\theta}_m(\nu_m) > 0.$$

To see the meaning of this condition, fix a modality and write $M := M_m(\nu_m)$ and $d_{m,j} := x_{i^\star_m} - x_j$. At the projected parameter,

$$d_{m,j}^\top \bar{\theta}_m = d_{m,j}^\top \theta_m - g_m \frac{d_{m,j}^\top M^{-1} y_m}{y_m^\top M^{-1} y_m}.$$

Hence the condition requires

$$d_{m,j}^\top \theta_m > g_m \frac{d_{m,j}^\top M^{-1} y_m}{y_m^\top M^{-1} y_m}, \qquad \forall j \neq i^\star_m.$$

Equivalently, along $\theta_m(s) = \theta_m - sM^{-1}y_m$, the target arm reaches its ε boundary at $s_\varepsilon = g_m/(y_m^\top M^{-1} y_m)$ before any competitor can tie $i^\star_m$.

The strict inequality allows the projection to continue slightly past the boundary. For sufficiently small $\rho > 0$,

$$\theta'_{m,\rho} := \theta_m - \frac{g_m + \rho}{y_m^\top M^{-1} y_m} M^{-1} y_m$$

still has $i^\star_m$ as a modality-wise optimal arm and satisfies $y_m^\top \theta'_{m,\rho} = \varepsilon - \rho < \varepsilon$. Consequently,

$$\max_j (x_j - x_{i^\circ})^\top \theta'_{m,\rho} < \varepsilon.$$

If the corresponding modality already satisfies the strict inequality, it is left unchanged. If it is exactly on the boundary, an arbitrarily small positive contraction of θ_m preserves the arm ordering and moves it into the strict interior.

Proposition 12 (Lower bound for an empty joint answer)**.** *Let $c_R := 2$ and $c_D := 8$. Suppose $\mathcal{S}_\varepsilon = \emptyset$, $0 < \delta < 1/2$, and Definition 11 holds. Then, if*

$$\max_{m\in\{R,D\}} (x_{i^\star_m} - x_{i^\circ})^\top \theta_m > \varepsilon,$$

every (ε, δ)-PAC algorithm for all ε-optimal identification satisfies

$$\mathbb{E}^\pi_{\theta_R,\theta_D}[\tau^\pi] \geq \mathrm{kl}(1-\delta, \delta) \min_{\substack{m\in\{R,D\}:\\ g_m>0}} \frac{c_m \inf_{\nu_m} \|y_m\|^2_{M_m(\nu_m)^{-1}}}{g_m^2},$$

where the infimum is over $\mathfrak{D}_R$ for $m = R$ and over $\mathfrak{D}_\mathcal{P}$ for $m = D$.

Proof. Let $E_\emptyset := \{\widehat{\mathcal{S}}^\pi = \emptyset\}$. Under the true instance, $\mathbb{P}_\theta^\pi(E_\emptyset) \geq 1 - \delta$.

Let n_R and n_D denote the expected numbers of reward and dueling queries, and let $\bar{\nu}_R, \bar{\nu}_D$ be their normalized expected sampling proportions. For each modality choose an arbitrary full-rank reference design ν_m^0 and define $\nu_{m,\alpha} = (1-\alpha)\bar{\nu}_m + \alpha\nu_m^0$. By Lemma 7, $\nu_{m,\alpha}$ is full rank.

For every modality with $g_m > 0$, construct the winner-preserving projection using $\nu_{m,\alpha}$ and continue it an arbitrarily small amount beyond the boundary. Modalities already satisfying the strict ε requirement are left unchanged, while a modality exactly on the boundary is moved into the interior by an arbitrarily small contraction. The resulting alternative satisfies $i^\circ \in \mathcal{S}_\varepsilon$, and therefore $\mathbb{P}_{\theta'}^\pi(E_\emptyset) \leq \delta$.

For reward feedback, the regularization and half-space projection give an information contribution, in the boundary limit, of at most

$$\frac{n_R}{2(1-\alpha)} \frac{g_R^2}{\|y_R\|^2_{M_R(\nu_{R,\alpha})^{-1}}}$$

whenever $g_R > 0$. For dueling feedback, the logistic inequality $\mathrm{kl}(\mu(u), \mu(v)) \leq (u-v)^2/8$ analogously gives at most

$$\frac{n_D}{8(1-\alpha)} \frac{g_D^2}{\|y_D\|^2_{M_D(\nu_{D,\alpha})^{-1}}}$$

whenever $g_D > 0$.

Applying Lemma 5 gives

$$\mathrm{kl}(1-\delta, \delta) \leq \frac{1}{1-\alpha} \sum_{\substack{m \in \{R,D\}: \\ g_m > 0}} \frac{n_m g_m^2}{c_m \|y_m\|^2_{M_m(\nu_{m,\alpha})^{-1}}}.$$

Since $n_R + n_D = \mathbb{E}[\tau^\pi]$, the right-hand side is bounded above by

$$\frac{\mathbb{E}[\tau^\pi]}{1-\alpha} \max_{\substack{m \in \{R,D\}: \\ g_m > 0}} \frac{g_m^2}{c_m \|y_m\|^2_{M_m(\nu_{m,\alpha})^{-1}}}.$$

Taking reciprocals, using that each $\nu_{m,\alpha}$ belongs to the corresponding full-rank design class, and letting $\alpha \downarrow 0$ proves the proposition. □

A weaker bound depending only on the fused distance follows immediately. Define

$$G^\circ := \max_{m \in \{R,D\}} (x_{i_m^\star} - x_{i^\circ})^\top \theta_m - \varepsilon.$$

Since every positive g_m satisfies $g_m \leq G^\circ$, Proposition 12 implies

$$\mathbb{E}[\tau^\pi] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{(G^\circ)^2} \min_{\substack{m \in \{R,D\}: \\ g_m > 0}} c_m \inf_{\nu_m} \|y_m\|^2_{M_m(\nu_m)^{-1}}.$$

If $G^\circ = 0$, the true instance lies on the boundary of the strict joint answer set. There exists a sequence of alternatives under which $i^\circ \in \mathcal{S}_\varepsilon$ and whose observation distributions converge to those of the true instance. The change-of-measure inequality then rules out finite expected stopping time for an algorithm that is uniformly PAC at such an instance.

B.4 Comparison with Transductive Linear Bandits

The reward component has the same experimental-design geometry as the information-theoretic lower bound for transductive linear best-arm identification in Fiez et al. (2019). Their characteristic quantity is

$$\inf_\nu \max_{z \neq z^\star} \frac{\|z^\star - z\|^2_{(\sum_x \nu_x x x^\top)^{-1}}}{\left((z^\star - z)^\top \theta^\star\right)^2}.$$

In the reward-only setting, taking i to be the unique best arm and letting $\varepsilon \downarrow 0$ gives

$$\varepsilon - (x_k - x_i)^\top \theta_R \longrightarrow (x_i - x_k)^\top \theta_R,$$

and therefore H_i^R reduces to the corresponding transductive characteristic quantity.

Thus the reward side of our lower bound retains the standard transductive linear-bandit geometry. The fused setting introduces two additional features: an ε-shifted decision boundary and a second, logistic observation channel. After the quadratic relaxation, the latter exhibits the same directional experimental-design structure, with the constant factor arising from the global curvature of the logistic link.

C Analysis of Wit-BotElim

This appendix develops the statistical and deterministic ingredients used in the proof of Theorem 3. We first restate the rounding and logistic fixed-design results inherited from RAGE (Fiez et al., 2019) and RAGE-GLM (Jun et al., 2021). We then specify the burn-in, allocation, and estimation routines used by Algorithm 1. Based on these primitives, we establish the phase-wise confidence guarantee underlying witness and bottleneck elimination, construct a simultaneous good event, and derive deterministic envelopes and phase-wise sample-cost bounds.

Throughout this appendix, let

$$\gamma_0 := \gamma, \qquad \gamma_r := \frac{\gamma}{2^r}, \qquad \delta_B := \frac{\delta}{2}, \qquad \delta_r := \frac{3\delta}{\pi^2 r^2}.$$

For each $m \in \{R, D\}$, split the phase risk as

$$\delta_{m,r}^{\mathrm{gap}} = \delta_{m,r}^{\mathrm{wit}} = \frac{\delta_r}{4}.$$

Thus,

$$\delta_B + \sum_{r \geq 1} \delta_r = \delta.$$

Recall

$$\Phi := \{x_{j,k} : (j,k) \in \mathcal{P}\}.$$

After a common rescaling, assume

$$\|\phi\|_2 \leq 1, \qquad \forall \phi \in \Phi,$$

and assume that a known constant $\kappa_0 > 0$ satisfies

$$\dot{\mu}(\phi^\top \theta_D) \geq \kappa_0, \qquad \forall \phi \in \Phi.$$

For phase r, let $\mathcal{H}_{r-1}$ denote the sigma-field generated by all observations and algorithmic decisions made before the fresh phase-r batches are drawn. Conditional on $\mathcal{H}_{r-1}$, the sets C_r, $W_{m,r-1}$, $K_{m,r}$, $\mathcal{Y}_{m,r}$, and the continuous phase-r designs are fixed.

C.1 Statistical Primitives

We first restate the external ingredients used by the algorithm.

Lemma 13 (Effective rounding)**.** *Let $\zeta > 0$ and*

$$r(\zeta) := \left\lceil \frac{d^2}{\zeta} \right\rceil.$$

For every full-rank continuous design and every $n \geq r(\zeta)$, the effective-rounding procedure used by RAGE (Fiez et al., 2019) returns a discrete allocation whose information matrix is at least $1/(1+\zeta)$ times the corresponding continuous information matrix.

In particular, for every reward design $\nu_R \in \mathfrak{D}_R$, the rounded batch satisfies

$$V_R := \sum_{s=1}^{n} x_{a_s} x_{a_s}^\top \succeq \frac{n}{1+\zeta} M_R(\nu_R).$$

For every dueling design $\nu_D \in \mathfrak{D}_D$ and every fixed reference parameter ϑ, the rounded batch satisfies

$$V_D(\vartheta) := \sum_{s=1}^{n} \dot{\mu}(\phi_s^\top \vartheta)\phi_s\phi_s^\top \succeq \frac{n}{1+\zeta} H_D(\nu_D, \vartheta).$$

Lemma 13 is the effective-rounding interface used by RAGE and RAGE-GLM (Fiez et al., 2019; Jun et al., 2021). We use it as a primitive and do not reproduce its proof.

Lemma 14 (G-optimal design)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a finite measurement set spanning $\mathbb{R}^d$, and define*

$$M(\nu) := \sum_{x\in\mathcal{X}} \nu_x x x^\top.$$

Then

$$\min_{\nu\in\Delta_{\mathcal{X}}} \max_{x\in\mathcal{X}} \|x\|^2_{M(\nu)^{-1}} = d.$$

This is the standard G-optimality identity used in linear experimental design and in the burn-in construction of RAGE-GLM; see Soare et al. (2014); Jun et al. (2021).

We next restate the fixed-design logistic confidence theorem used by the dueling component.

Lemma 15 (Fixed-design logistic confidence)**.** *Consider a fixed batch $\{\phi_s\}_{s=1}^n$ with $\|\phi_s\|_2 \le 1$ and observations*

$$Y_s \sim \text{Bernoulli}\left(\mu(\phi_s^\top \theta_D)\right).$$

Let $\widehat{\theta}$ be the logistic MLE and define

$$V_D(\theta_D) := \sum_{s=1}^{n} \dot{\mu}(\phi_s^\top \theta_D)\phi_s\phi_s^\top.$$

Let t_{eff} be the number of distinct measurement vectors appearing in the batch. For $\alpha \le e^{-1}$, define

$$\Gamma_{\log}(\alpha, t_{\text{eff}}) := 64\left(d\log 6 + \log\frac{2+t_{\text{eff}}}{\alpha}\right).$$

If

$$\max_{s\in[n]} \|\phi_s\|^2_{V_D(\theta_D)^{-1}} \le \frac{1}{\Gamma_{\log}(\alpha, t_{\text{eff}})},$$

then Theorem 1 of Jun et al. (2021) gives, on its variance-stability event,

$$|u^\top(\widehat{\theta} - \theta_D)| \le 2.4\|u\|_{V_D(\theta_D)^{-1}} \sqrt{\log\frac{2(2+t_{\text{eff}})}{\alpha}}$$

for every fixed direction u with $\|u\|_2 \le 1$.

Moreover, Remark 1 of Jun et al. (2021) gives the corresponding empirical-Fisher form

$$|u^\top(\widehat{\theta} - \theta_D)| \le 3.6\|u\|_{V_D(\widehat{\theta})^{-1}} \sqrt{\log\frac{2(2+t_{\text{eff}})}{\alpha}}.$$

The same conclusions hold simultaneously over a finite collection of test directions after allocating the failure probability by a union bound.

The fixed-design condition in Lemma 15 is the reason for using fresh batches in Algorithm 1. Conditional on $\mathcal{H}_{r-1}$, the phase-r measurement vectors are fixed before the fresh observations are generated.

Algorithm 2 BURNIN: Dueling logistic burn-in

Require: Φ, κ_0, δ_B, ζ

1: $\nu_{D,B} \leftarrow \arg\min_{\nu_D \in \mathfrak{D}_D} \max_{\phi \in \Phi} \|\phi\|^2_{M_D(\nu_D)^{-1}}$

2: $\Gamma_B \leftarrow 64\left(d\log 6 + \log \frac{2+|\Phi|}{\delta_B}\right)$

3: $n_B \leftarrow \max\left\{r(\zeta), \left\lceil 3(1+\zeta)\kappa_0^{-1} d\Gamma_B \log \frac{2|\Phi|(2+|\Phi|)}{\delta_B} \right\rceil\right\}$

4: $\phi_1, \ldots, \phi_{n_B} \leftarrow$ ROUND$(D, n_B, \zeta, \nu_{D,B})$

5: Query the associated duels and observe $Y_1, \ldots, Y_{n_B}$

6: $\widehat{\theta}_{D,0} \leftarrow$ LOGISTICMLE $(\{(\phi_s, Y_s)\}_{s=1}^{n_B})$

7: **return** $\widehat{\theta}_{D,0}$

C.2 DUELING BURN-IN

The dueling Fisher information depends on the unknown parameter θ_D. We therefore use the G-optimal burn-in of Jun et al. (2021, Algorithm 2), specialized to the pairwise measurement set Φ.

Define

$$\nu_{D,B} \in \arg\min_{\nu_D \in \mathfrak{D}_D} \max_{\phi \in \Phi} \|\phi\|^2_{M_D(\nu_D)^{-1}}.$$

By Lemma 14,

$$\max_{\phi \in \Phi} \|\phi\|^2_{M_D(\nu_{D,B})^{-1}} = d.$$

Define

$$\Gamma_B := 64\left(d\log 6 + \log \frac{2+|\Phi|}{\delta_B}\right)$$

and

$$n_B := \max\left\{r(\zeta), \left\lceil 3(1+\zeta)\kappa_0^{-1} d\Gamma_B \log \frac{2|\Phi|(2+|\Phi|)}{\delta_B} \right\rceil\right\}.$$

For $r \geq 0$, define the dueling localization event

$$\mathcal{L}_r := \left\{|\phi^\top(\widehat{\theta}_{D,r} - \theta_D)| \leq 1, \quad \forall \phi \in \Phi\right\}.$$

Lemma 16 (Burn-in guarantee)**.** *The burn-in procedure in Algorithm 2 satisfies*

$$\mathbb{P}(\mathcal{L}_0) \geq 1 - \delta_B.$$

Moreover,

$$n_B = O\left(\frac{d^2}{\zeta} + (1+\zeta)\kappa_0^{-1} d^2 \log \frac{|\Phi|}{\delta} + (1+\zeta)\kappa_0^{-1} d \log^2 \frac{|\Phi|}{\delta}\right).$$

Lemma 16 is the direct specialization of the burn-in analysis of Jun et al. (2021) to Φ; we do not repeat its proof.

C.3 PHASE-WISE ALLOCATION

We now make the routine ALLOCATION in Algorithm 1 explicit.

For the reward side, if $\mathcal{Y}_{R,r} \neq \emptyset$, define

$$P_{R,r} := \min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \mathcal{Y}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}}.$$

Let $\nu_{R,r}$ be a minimizer and define

$$n_{R,r} := \max\left\{r(\zeta), \left\lceil \frac{2(1+\zeta)P_{R,r}}{\gamma_r^2} \log \frac{2|\mathcal{Y}_{R,r}|}{\delta_{R,r}^{\mathrm{gap}}} \right\rceil\right\}.$$

If $\mathcal{Y}_{R,r} = \emptyset$, set $n_{R,r} := 0$.

For the dueling side, let

$$L_\Phi := \max\{K, |\Phi|\}$$

and

$$\eta_r := \frac{\delta_{D,r}^{\mathrm{gap}}}{2L_\Phi(2+|\Phi|)}.$$

Since

$$\delta_{D,r}^{\mathrm{gap}} = \frac{3\delta}{4\pi^2 r^2},$$

we equivalently have

$$\eta_r = \frac{3\delta}{8\pi^2 r^2 L_\Phi(2+|\Phi|)}.$$

Define

$$\Gamma_r := 64\left(d\log 6 + \log\frac{2+|\Phi|}{\eta_r}\right).$$

Equivalently,

$$\Gamma_r = 64\left(d\log 6 + \log\frac{8\pi^2 r^2 L_\Phi(2+|\Phi|)^2}{3\delta}\right).$$

If $\mathcal{Y}_{D,r} \neq \emptyset$, define

$$P_{D,r} := \min_{\nu_D \in \mathfrak{D}_D} \max\left\{\Gamma_r \max_{\phi\in\Phi} \|\phi\|^2_{H_D(\nu_D,\widehat{\theta}_{D,r-1})^{-1}}, \right.$$
$$\left. \frac{(2.4)^2}{\gamma_r^2} \max_{y\in\mathcal{Y}_{D,r}} \|y\|^2_{H_D(\nu_D,\widehat{\theta}_{D,r-1})^{-1}}\right\}.$$

Let $\nu_{D,r}$ be a minimizer and define

$$n_{D,r} := \max\left\{r(\zeta), \left\lceil 3(1+\zeta)P_{D,r}\log\frac{1}{\eta_r}\right\rceil\right\}.$$

If $\mathcal{Y}_{D,r} = \emptyset$, set $n_{D,r} := 0$.

The first branch of $P_{D,r}$ controls the global Fisher leverage required by the logistic fixed-design confidence theorem. The second branch controls the active directions in $\mathcal{Y}_{D,r}$. This two-objective design is adapted from RAGE-GLM (Jun et al., 2021); the reduced active direction set is specific to WIT-BOTELIM.

C.4 FRESH-BATCH ESTIMATION AND CONFIDENCE BOUNDS

For a nonempty reward batch, define

$$V_{R,r} := \sum_{s=1}^{n_{R,r}} x_{a_s} x_{a_s}^\top$$

and

$$\widehat{\theta}_{R,r} := V_{R,r}^{-1} \sum_{s=1}^{n_{R,r}} x_{a_s} R_s.$$

For a nonempty dueling batch $\{\phi_s\}_{s=1}^{n_{D,r}}$, let $\widehat{\theta}_{D,r}$ be its logistic MLE and define

$$V_{D,r}(\vartheta) := \sum_{s=1}^{n_{D,r}} \dot{\mu}(\phi_s^\top\vartheta)\phi_s\phi_s^\top,$$

and

$$\widehat{V}_{D,r} := V_{D,r}(\widehat{\theta}_{D,r}).$$

If $n_{D,r} = 0$, use the convention

$$\widehat{\theta}_{D,r} := \widehat{\theta}_{D,r-1}.$$

We first record the reward concentration result.

Algorithm 3 ALLOCATION: Phase-wise experimental design

Require: $m, \mathcal{Y}_{m,r}, \gamma_r, \delta^{\mathrm{gap}}_{m,r}, \delta^{\mathrm{wit}}_{m,r}, \widehat{\theta}_{D,r-1}, \zeta$

if $\mathcal{Y}_{m,r} = \emptyset$ **then**
  return $(\varnothing, 0)$
end if
if $m = R$ **then**
  $\nu_{R,r} \leftarrow \arg\min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \mathcal{Y}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}}$
  $P_{R,r} \leftarrow \max_{y \in \mathcal{Y}_{R,r}} \|y\|^2_{M_R(\nu_{R,r})^{-1}}$
  $n_{R,r} \leftarrow \max\left\{r(\zeta), \left\lceil \frac{2(1+\zeta)P_{R,r}}{\gamma_r^2} \log \frac{2|\mathcal{Y}_{R,r}|}{\delta^{\mathrm{gap}}_{R,r}} \right\rceil\right\}$
  return $(\nu_{R,r}, n_{R,r})$
else
  $L_\Phi \leftarrow \max\{K, |\Phi|\}$
  $\eta_r \leftarrow \delta^{\mathrm{gap}}_{D,r} / \left(2L_\Phi(2 + |\Phi|)\right)$
  $\Gamma_r \leftarrow 64\left(d \log 6 + \log \frac{2+|\Phi|}{\eta_r}\right)$
  $\nu_{D,r} \leftarrow \arg \min_{\nu_D \in \mathfrak{D}_D} \max\Big\{\Gamma_r \max_{\phi \in \Phi} \|\phi\|^2_{H_D(\nu_D, \widehat{\theta}_{D,r-1})^{-1}},$
    $\frac{(2.4)^2}{\gamma_r^2} \max_{y \in \mathcal{Y}_{D,r}} \|y\|^2_{H_D(\nu_D, \widehat{\theta}_{D,r-1})^{-1}}\Big\}$
  Let $P_{D,r}$ be the value of the preceding objective
  $n_{D,r} \leftarrow \max\left\{r(\zeta), \left\lceil 3(1+\zeta)P_{D,r} \log \frac{1}{\eta_r} \right\rceil\right\}$
  return $(\nu_{D,r}, n_{D,r})$
end if

Lemma 17 (Reward fixed-design confidence)**.** *Conditional on $\mathcal{H}_{r-1}$, let $\mathcal{U}$ be any finite set of directions. Then, for every $\alpha \in (0,1)$,*

$$\mathbb{P}\left(|u^\top(\widehat{\theta}_{R,r} - \theta_R)| \le \|u\|_{V_{R,r}^{-1}} \sqrt{2 \log \frac{2|\mathcal{U}|}{\alpha}}, \quad \forall u \in \mathcal{U} \;\middle|\; \mathcal{H}_{r-1} \right) \ge 1 - \alpha.$$

Proof. Conditional on $\mathcal{H}_{r-1}$, the reward design is fixed. For every fixed direction u, $u^\top(\widehat{\theta}_{R,r} - \theta_R)$ is $\|u\|_{V_{R,r}^{-1}}$-sub-Gaussian. Applying the corresponding tail bound and a union bound over $\mathcal{U}$ proves the claim. □

Define the reward witness radius

$$\mathrm{CR}_{R,r}(k,j) := \|x_k - x_j\|_{V_{R,r}^{-1}} \sqrt{2 \log \frac{2|\Phi|}{\delta^{\mathrm{wit}}_{R,r}}}.$$

Lemma 18 (Reward phase accuracy)**.** *Conditional on $\mathcal{H}_{r-1}$, with probability at least*

$$1 - \delta^{\mathrm{gap}}_{R,r} - \delta^{\mathrm{wit}}_{R,r},$$

both of the following statements hold:

$$|y^\top(\widehat{\theta}_{R,r} - \theta_R)| \le \gamma_r, \qquad \forall y \in \mathcal{Y}_{R,r},$$

and

$$|(x_k - x_j)^\top(\widehat{\theta}_{R,r} - \theta_R)| \le \mathrm{CR}_{R,r}(k,j), \qquad \forall k \ne j.$$

Proof. Lemma 13 gives

$$V_{R,r} \succeq \frac{n_{R,r}}{1+\zeta} M_R(\nu_{R,r}),$$

and therefore

$$\max_{y \in \mathcal{Y}_{R,r}} \|y\|^2_{V_{R,r}^{-1}} \le \frac{(1+\zeta)P_{R,r}}{n_{R,r}}.$$

The definition of $n_{R,r}$ and Lemma 17 applied to $\mathcal{Y}_{R,r}$ imply the active-direction statement with failure probability at most $\delta^{\mathrm{gap}}_{R,r}$.

Applying Lemma 17 to all pairwise directions gives the witness statement with failure probability at most $\delta^{\mathrm{wit}}_{R,r}$. A union bound completes the proof. □

We next record the deterministic comparison between nearby logistic Fisher matrices.

Lemma 19 (Logistic Fisher stability)**.** *Suppose $\vartheta \in \mathbb{R}^d$ satisfies*

$$|\phi^\top(\vartheta - \theta_D)| \le 1, \qquad \forall \phi \in \Phi.$$

Then, for every $\nu_D \in \mathfrak{D}_D$,

$$e^{-1}H_D(\nu_D, \theta_D) \preceq H_D(\nu_D, \vartheta) \preceq eH_D(\nu_D, \theta_D).$$

Consequently,

$$e^{-1}H_D(\nu_D, \theta_D)^{-1} \preceq H_D(\nu_D, \vartheta)^{-1} \preceq eH_D(\nu_D, \theta_D)^{-1}.$$

Proof. For the logistic link,

$$|(\log \dot\mu)'(z)| = |1 - 2\mu(z)| \le 1.$$

Hence

$$|\phi^\top(\vartheta - \theta_D)| \le 1$$

implies

$$e^{-1}\dot\mu(\phi^\top\theta_D) \le \dot\mu(\phi^\top\vartheta) \le e\dot\mu(\phi^\top\theta_D).$$

Summing over the design proves the first matrix inequality. Matrix inversion reverses the Loewner order and yields the second. □

We use the following consequence of the RAGE-GLM fixed-design analysis.

Lemma 20 (RAGE-GLM phase confidence)**.** *Fix phase $r \ge 1$ and condition on $\mathcal{H}_{r-1}$ and $\mathcal{L}_{r-1}$. Consider the design $\nu_{D,r}$ and sample size $n_{D,r}$ defined in Appendix C.3. Then, with failure probability at most $\delta^{\mathrm{gap}}_{D,r}$,*

$$\mathcal{L}_r$$

holds and

$$|y^\top(\widehat\theta_{D,r} - \theta_D)| \le \gamma_r, \qquad \forall y \in \mathcal{Y}_{D,r}.$$

Moreover, define

$$\mathrm{CR}_{D,r}(k,j) := 3.6\|x_k - x_j\|_{\widehat V_{D,r}^{-1}} \sqrt{\log \frac{2(2+t_{\mathrm{eff},r})|\Phi|}{\delta^{\mathrm{wit}}_{D,r}}},$$

where $t_{\mathrm{eff},r}$ is the number of distinct pairwise measurement vectors in the phase-r batch. Then, with failure probability at most $\delta^{\mathrm{wit}}_{D,r}$,

$$|(x_k - x_j)^\top(\widehat\theta_{D,r} - \theta_D)| \le \mathrm{CR}_{D,r}(k,j), \qquad \forall k \ne j.$$

If $n_{D,r} = 0$, no new dueling estimate is formed, $\widehat\theta_{D,r} = \widehat\theta_{D,r-1}$, and $\mathcal{L}_r = \mathcal{L}_{r-1}$.

Lemma 20 is the RAGE-GLM phase argument of Jun et al. (2021), specialized to the phase-dependent direction set $\mathcal{Y}_{D,r}$. The global branch in $P_{D,r}$ guarantees the leverage condition required by Lemma 15; the active-direction branch guarantees the target accuracy γ_r; and the fresh-batch construction allows the fixed-design theorem to be applied conditional on $\mathcal{H}_{r-1}$. We therefore use this result without reproducing the RAGE-GLM concentration proof.

C.5 Phase-Wise Validity of Fused Elimination

Define

$$\mathcal{E}_{R,r}^{\mathrm{gap}} := \left\{ |y^\top(\widehat{\theta}_{R,r} - \theta_R)| \le \gamma_r, \quad \forall y \in \mathcal{Y}_{R,r} \right\},$$

and

$$\mathcal{E}_{R,r}^{\mathrm{wit}} := \left\{ |(x_k - x_j)^\top(\widehat{\theta}_{R,r} - \theta_R)| \le \mathrm{CR}_{R,r}(k,j), \quad \forall k \ne j \right\}.$$

Similarly, define

$$\mathcal{E}_{D,r}^{\mathrm{gap}} := \mathcal{L}_r \cap \left\{ |y^\top(\widehat{\theta}_{D,r} - \theta_D)| \le \gamma_r, \quad \forall y \in \mathcal{Y}_{D,r} \right\},$$

and

$$\mathcal{E}_{D,r}^{\mathrm{wit}} := \left\{ |(x_k - x_j)^\top(\widehat{\theta}_{D,r} - \theta_D)| \le \mathrm{CR}_{D,r}(k,j), \quad \forall k \ne j \right\}.$$

Let

$$\mathcal{E}_r := \bigcap_{m \in \{R,D\}} \left(\mathcal{E}_{m,r}^{\mathrm{gap}} \cap \mathcal{E}_{m,r}^{\mathrm{wit}} \right).$$

Lemma 21 (Phase-wise fused-elimination guarantee)**.** *Conditional on $\mathcal{H}_{r-1}$ and $\mathcal{L}_{r-1}$,*

$$\mathbb{P}\left(\mathcal{E}_r \mid \mathcal{H}_{r-1}, \mathcal{L}_{r-1}\right) \ge 1 - \delta_r.$$

On $\mathcal{E}_r$, the following statements hold simultaneously for $m \in \{R, D\}$:

1. *every witness removed from $W_{m,r-1}$ is strictly suboptimal under modality m, and*
$$i_m^\star \in W_{m,r};$$
2. *for every $i \in G_{m,r}$,*
$$\left| \widehat{\Delta}_{m,i}^{(r)} - \Delta_{m,i} \right| \le \gamma_r;$$
3. *for every $i \in C_r \setminus (K_{R,r} \cup K_{D,r})$, if*
$$U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)},$$
then
$$\Delta_{\bar{m},i} < \Delta_{m,i}$$
and therefore
$$\Delta_i = \Delta_{m,i}.$$

Proof. By Lemma 18, the reward gap and witness events fail with total probability at most

$$\delta_{R,r}^{\mathrm{gap}} + \delta_{R,r}^{\mathrm{wit}}.$$

By Lemma 20, the dueling gap and witness events fail with total probability at most

$$\delta_{D,r}^{\mathrm{gap}} + \delta_{D,r}^{\mathrm{wit}}.$$

Since each term equals $\delta_r/4$, a union bound gives

$$\mathbb{P}\left(\mathcal{E}_r \mid \mathcal{H}_{r-1}, \mathcal{L}_{r-1}\right) \ge 1 - \delta_r.$$

Suppose witness j is removed under modality m. Then there exists $k \in W_{m,r-1} \setminus \{j\}$ such that

$$(x_k - x_j)^\top \widehat{\theta}_{m,r} - \mathrm{CR}_{m,r}(k,j) > 0.$$

On $\mathcal{E}_{m,r}^{\mathrm{wit}}$,

$$(x_k - x_j)^\top \theta_m \ge (x_k - x_j)^\top \widehat{\theta}_{m,r} - \mathrm{CR}_{m,r}(k,j) > 0.$$

Hence j is strictly suboptimal. In particular, if $j = i_m^\star$, then

$$(x_k - x_j)^\top \theta_m \le 0$$

for every k, so such a certificate is impossible. Therefore

$$i_m^\star \in W_{m,r}.$$

Now fix $i \in G_{m,r}$. Since

$$W_{m,r} \subseteq W_{m,r-1},$$

every nonzero direction appearing in

$$\widehat{\Delta}_{m,i}^{(r)} = \max_{j \in W_{m,r}} (x_j - x_i)^\top \widehat{\theta}_{m,r}$$

belongs to $\mathcal{Y}_{m,r}$. Moreover, $i_m^\star \in W_{m,r}$, so

$$\max_{j \in W_{m,r}} (x_j - x_i)^\top \theta_m = \Delta_{m,i}.$$

The active-direction event then gives

$$\left| \widehat{\Delta}_{m,i}^{(r)} - \Delta_{m,i} \right| \le \gamma_r.$$

Finally, suppose

$$U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)}.$$

The preceding gap guarantee implies that both modality-wise confidence intervals contain their corresponding true gaps. Hence

$$\Delta_{\bar{m},i} \le U_{\bar{m},i}^{(r)} < L_{m,i}^{(r)} \le \Delta_{m,i}.$$

Thus

$$\Delta_{\bar{m},i} < \Delta_{m,i}$$

and

$$\Delta_i = \max\{\Delta_{R,i}, \Delta_{D,i}\} = \Delta_{m,i}.$$

□

C.6 Simultaneous Good Event and Interval Validity

Let

$$\mathcal{E}_B := \mathcal{L}_0$$

and define

$$\mathcal{E} := \mathcal{E}_B \cap \bigcap_{r \ge 1} \mathcal{E}_r.$$

Lemma 22 (Simultaneous good event)**.** *The event $\mathcal{E}$ satisfies*

$$\mathbb{P}_{\theta_R, \theta_D}^{\pi^q_{\text{WIT-BOTELIM}}}(\mathcal{E}) \ge 1 - \delta.$$

Proof. Lemma 16 gives

$$\mathbb{P}(\mathcal{E}_B^c) \le \delta_B.$$

If the burn-in and the first $r-1$ phase events hold, then $\mathcal{E}_{D,r-1}^{\text{gap}}$ implies $\mathcal{L}_{r-1}$, so Lemma 21 applies at phase r. Therefore, by a first-failure union bound,

$$\mathbb{P}(\mathcal{E}^c) \le \delta_B + \sum_{r \ge 1} \delta_r.$$

Using

$$\delta_B = \frac{\delta}{2}$$

and

$$\sum_{r \ge 1} \delta_r = \frac{3\delta}{\pi^2} \sum_{r \ge 1} \frac{1}{r^2} = \frac{\delta}{2},$$

we obtain

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta.$$

□

Lemma 23 (Validity of modality-wise and fused intervals)**.** *On $\mathcal{E}$, for every phase r, modality m, and $i \in G_{m,r}$,*

$$\Delta_{m,i} \in [L_{m,i}^{(r)}, U_{m,i}^{(r)}].$$

Moreover,

$$L_{m,i}^{(r)} \geq \Delta_{m,i} - 2\gamma_r, \qquad U_{m,i}^{(r)} \leq \Delta_{m,i} + 2\gamma_r.$$

Every bottleneck certificate is correct and remains correct while its arm is unresolved. Consequently, for every unresolved arm i,

$$\Delta_i \in [L_i^{F,(r)}, U_i^{F,(r)}],$$

and

$$L_i^{F,(r)} \geq \Delta_i - 2\gamma_r, \qquad U_i^{F,(r)} \leq \Delta_i + 2\gamma_r.$$

In particular,

$$i \in \mathcal{I}_r \Longrightarrow \Delta_i < \varepsilon,$$

and

$$i \in \mathcal{O}_r \Longrightarrow \Delta_i \geq \varepsilon.$$

Proof. By Lemma 21,

$$\left|\widehat{\Delta}_{m,i}^{(r)} - \Delta_{m,i}\right| \leq \gamma_r.$$

Since

$$L_{m,i}^{(r)} = \max\left\{0, \widehat{\Delta}_{m,i}^{(r)} - \gamma_r\right\}$$

and

$$U_{m,i}^{(r)} = \widehat{\Delta}_{m,i}^{(r)} + \gamma_r,$$

the modality-wise interval is valid and satisfies the displayed endpoint bounds.

If no bottleneck certificate has been produced for arm i, the fused interval is obtained by taking the maximum of the two valid modality-wise intervals, so it contains

$$\Delta_i = \max\{\Delta_{R,i}, \Delta_{D,i}\}.$$

If modality m has been certified as controlling, then Lemma 21 gives

$$\Delta_i = \Delta_{m,i},$$

and therefore the modality-m interval is also a valid fused interval. The endpoint and classification statements follow directly. □

C.7 Deterministic Envelope

Recall

$$\overline{G}_{R,1} = \overline{G}_{D,1} = [K].$$

For $r \geq 2$, define

$$\overline{G}_{m,r} := \{i \in [K] : \Delta_{R,i} \leq \varepsilon + 2\gamma_{r-1},\ \Delta_{D,i} \leq \varepsilon + 2\gamma_{r-1},\ \Delta_{\bar{m},i} - \Delta_{m,i} \leq 4\gamma_{r-1}\}.$$

Define

$$\overline{\mathcal{Y}}_{m,r} := \left\{x_j - x_i : i \in \overline{G}_{m,r},\ j \in [K],\ j \neq i\right\}.$$

Lemma 24 (Deterministic envelope)**.** *On $\mathcal{E}$, for every phase r and modality m,*

$$G_{m,r} \subseteq \overline{G}_{m,r}, \qquad \mathcal{Y}_{m,r} \subseteq \overline{\mathcal{Y}}_{m,r}.$$

Proof. The claim is immediate for $r = 1$. Let $r \geq 2$.

First take $i \in C_r$. Suppose, for some modality m,

$$\Delta_{m,i} > \varepsilon + 2\gamma_{r-1}.$$

If modality m was active for i in phase $r-1$, then Lemma 23 gives

$$L_{m,i}^{(r-1)} \geq \Delta_{m,i} - 2\gamma_{r-1} > \varepsilon.$$

Hence the fused lower confidence endpoint also exceeds ε, and i would have been rejected before phase r, contradicting $i \in C_r$.

If modality m had already been certified as noncontrolling, the certified controlling modality has a strictly larger true gap. Its valid interval therefore yields the same rejection conclusion. Thus every $i \in C_r$ satisfies

$$\Delta_{R,i} \le \varepsilon + 2\gamma_{r-1}, \qquad \Delta_{D,i} \le \varepsilon + 2\gamma_{r-1}.$$

Now take $i \in G_{m,r}$ and suppose

$$\Delta_{\bar{m},i} - \Delta_{m,i} > 4\gamma_{r-1}.$$

If neither modality had previously been certified as controlling, then Lemma 23 gives

$$U_{m,i}^{(r-1)} \le \Delta_{m,i} + 2\gamma_{r-1} < \Delta_{\bar{m},i} - 2\gamma_{r-1} \le L_{\bar{m},i}^{(r-1)}.$$

Hence $\bar{m}$ would have been certified as controlling at the end of phase $r-1$. This would imply

$$i \in K_{\bar{m},r},$$

contradicting

$$i \in G_{m,r} = C_r \setminus K_{\bar{m},r}.$$

If a bottleneck certificate had already been produced, correctness of that certificate gives the same contradiction. Therefore

$$G_{m,r} \subseteq \overline{G}_{m,r}.$$

Finally,

$$W_{m,r-1} \subseteq [K]$$

implies

$$\mathcal{Y}_{m,r} \subseteq \overline{\mathcal{Y}}_{m,r}.$$

□

The deterministic envelope deliberately replaces the realized witness set $W_{m,r-1}$ by the full arm set $[K]$. Hence it explicitly captures target-side bottleneck compression, while the additional elimination produced by witness elimination remains adaptive and path dependent.

C.8 Domination of Phase-Wise Costs

We now compare the realized phase costs with the deterministic quantities appearing in Theorem 3. Whenever an active direction set is empty, its directional contribution is interpreted as zero.

Define

$$\overline{P}_{R,r} := \min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \overline{\mathcal{Y}}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}}.$$

For dueling feedback, define

$$\overline{P}_{D,r} := \min_{\nu_D \in \mathfrak{D}_D} \max \left\{ \Gamma_r \max_{\phi \in \Phi} \|\phi\|^2_{H_D(\nu_D,\theta_D)^{-1}}, \right.$$
$$\left. \frac{(2.4)^2}{\gamma_r^2} \max_{y \in \overline{\mathcal{Y}}_{D,r}} \|y\|^2_{H_D(\nu_D,\theta_D)^{-1}} \right\}.$$

Since

$$\gamma_r^{-2} = \frac{4^r}{\gamma^2},$$

this is the same deterministic dueling design quantity used in Theorem 3.

Lemma 25 (Domination of realized phase designs)**.** *On $\mathcal{E}$, for every phase r,*

$$P_{R,r} \le \overline{P}_{R,r}$$

and

$$P_{D,r} \le e\,\overline{P}_{D,r}.$$

Proof. For reward feedback, Lemma 24 gives

$$\mathcal{Y}_{R,r} \subseteq \overline{\mathcal{Y}}_{R,r}.$$

For every fixed design ν_R,

$$\max_{y \in \mathcal{Y}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}} \le \max_{y \in \overline{\mathcal{Y}}_{R,r}} \|y\|^2_{M_R(\nu_R)^{-1}}.$$

Minimizing over ν_R proves

$$P_{R,r} \le \overline{P}_{R,r}.$$

For dueling feedback, $\mathcal{E}$ implies $\mathcal{L}_{r-1}$. Hence Lemma 19 gives

$$H_D(\nu_D, \widehat{\theta}_{D,r-1})^{-1} \preceq e H_D(\nu_D, \theta_D)^{-1}$$

for every ν_D. Together with

$$\mathcal{Y}_{D,r} \subseteq \overline{\mathcal{Y}}_{D,r},$$

both branches of the realized dueling objective are bounded by e times their deterministic counterparts. Minimizing over ν_D proves

$$P_{D,r} \le e\, \overline{P}_{D,r}.$$

□

Lemma 26 (Phase sample-cost bounds)**.** *On $\mathcal{E}$, for every phase r,*

$$n_{R,r} = O\left(\frac{d^2}{\zeta} + (1+\zeta)\frac{4^r \overline{P}_{R,r}}{\gamma^2} \log \frac{r^2 |\overline{\mathcal{Y}}_{R,r}|}{\delta}\right),$$

and

$$n_{D,r} = O\left(\frac{d^2}{\zeta} + (1+\zeta)\overline{P}_{D,r} \log \frac{r^2 \max\{K, |\Phi|\}}{\delta}\right).$$

Proof. For reward feedback, use

$$\gamma_r^{-2} = \frac{4^r}{\gamma^2}$$

and

$$\delta_{R,r}^{\mathrm{gap}} = \frac{3\delta}{4\pi^2 r^2}.$$

By Lemma 25,

$$P_{R,r} \le \overline{P}_{R,r},$$

and by Lemma 24,

$$|\mathcal{Y}_{R,r}| \le |\overline{\mathcal{Y}}_{R,r}|.$$

Substitution into the definition of $n_{R,r}$ gives the reward bound.

For dueling feedback,

$$\eta_r = \frac{3\delta}{8\pi^2 r^2 \max\{K, |\Phi|\}(2 + |\Phi|)}.$$

Therefore

$$\log \frac{1}{\eta_r} = O\left(\log \frac{r^2 \max\{K, |\Phi|\}}{\delta}\right),$$

up to factors depending only logarithmically on $|\Phi|$. Using Lemma 25 and absorbing the universal factor e into $O(\cdot)$ gives the dueling bound. □

Finally, Lemma 16 gives

$$n_B = O\left(\frac{d^2}{\zeta} + (1+\zeta)\kappa_0^{-1} d^2 \log \frac{|\Phi|}{\delta} + (1+\zeta)\kappa_0^{-1} d \log^2 \frac{|\Phi|}{\delta}\right).$$

We have therefore established all ingredients needed for Theorem 3: the simultaneous event $\mathcal{E}$ occurs with probability at least $1-\delta$; all modality-wise and fused intervals are valid on $\mathcal{E}$; the random active sets are contained in their deterministic envelopes; and the realized reward and dueling sampling costs are dominated by the deterministic phase costs appearing in the theorem.

C.9 Stopping Horizon and Correctness

Lemma 27 (Stopping horizon and correctness)**.** *Work on the simultaneous event $\mathcal{E}$ of Lemma 22.*

For ε^{one} identification, suppose first that $\mathcal{S}_\varepsilon \neq \emptyset$. Then Wit-BotElim *returns an arm in $\mathcal{S}_\varepsilon$ no later than phase R_{one}.*

If $\mathcal{S}_\varepsilon = \emptyset$ and $\rho_{\mathrm{one}} > 0$, then Wit-BotElim *returns $\emptyset$ no later than phase R_{one}.*

For ε^{all} identification, if $\rho_{\mathrm{all}} > 0$, then Wit-BotElim *returns exactly $\mathcal{S}_\varepsilon$ no later than phase R_{all}.*

Proof. By Lemma 23, on $\mathcal{E}$ every accepted arm satisfies

$$i \in \mathcal{I}_r \Longrightarrow \Delta_i < \varepsilon,$$

whereas every rejected arm satisfies

$$i \in \mathcal{O}_r \Longrightarrow \Delta_i \geq \varepsilon.$$

Thus every classification made by the algorithm is correct.

Consider first $\varepsilon^{\mathrm{one}}$ identification with $\mathcal{S}_\varepsilon \neq \emptyset$. Choose

$$i^\star_\varepsilon \in \arg\max_{i \in \mathcal{S}_\varepsilon} (\varepsilon - \Delta_i),$$

so that

$$\varepsilon - \Delta_{i^\star_\varepsilon} = \rho_{\mathrm{one}}.$$

If the algorithm has not already stopped before phase R_{one}, then $i^\star_\varepsilon$ cannot have been rejected, because rejection is correct on $\mathcal{E}$. Hence it remains unresolved. Since

$$2\gamma_{R_{\mathrm{one}}} < \rho_{\mathrm{one}},$$

Lemma 23 gives

$$U^{F,(R_{\mathrm{one}})}_{i^\star_\varepsilon} \leq \Delta_{i^\star_\varepsilon} + 2\gamma_{R_{\mathrm{one}}} < \varepsilon.$$

Therefore $i^\star_\varepsilon \in \mathcal{I}_{R_{\mathrm{one}}}$ and the algorithm returns a valid arm.

Now suppose $\mathcal{S}_\varepsilon = \emptyset$ and $\rho_{\mathrm{one}} > 0$. Every arm satisfies

$$\Delta_i - \varepsilon \geq \rho_{\mathrm{one}}.$$

At phase R_{one}, every arm that remains unresolved satisfies

$$L^{F,(R_{\mathrm{one}})}_i \geq \Delta_i - 2\gamma_{R_{\mathrm{one}}} > \varepsilon.$$

Hence every remaining candidate is rejected and $C_{R_{\mathrm{one}}+1} = \emptyset$. The $\varepsilon^{\mathrm{one}}$ stopping rule therefore returns $\emptyset$.

Finally consider $\varepsilon^{\mathrm{all}}$ identification. If $i \in \mathcal{S}_\varepsilon$, then

$$\varepsilon - \Delta_i \geq \rho_{\mathrm{all}}.$$

Thus, if i has not already been accepted,

$$U^{F,(R_{\mathrm{all}})}_i \leq \Delta_i + 2\gamma_{R_{\mathrm{all}}} < \varepsilon,$$

so i is accepted by phase R_{all}.

If $i \notin \mathcal{S}_\varepsilon$, the assumption $\rho_{\mathrm{all}} > 0$ implies

$$\Delta_i - \varepsilon \geq \rho_{\mathrm{all}}.$$

Hence, if i has not already been rejected,

$$L^{F,(R_{\mathrm{all}})}_i \geq \Delta_i - 2\gamma_{R_{\mathrm{all}}} > \varepsilon,$$

so i is rejected by phase R_{all}.

Therefore every arm is correctly classified by R_{all}, and the accumulated accepted set is exactly $\mathcal{S}_\varepsilon$. □

C.10 Proof of Theorem 3

Proof. Work on $\mathcal{E}$, which occurs with probability at least $1-\delta$.

Suppose first that $\mathcal{S}_\varepsilon \neq \emptyset$ in ε^{one} mode. Choose

$$i^\star_\varepsilon \in \arg\max_{i\in\mathcal{S}_\varepsilon}(\varepsilon - \Delta_i),$$

so that

$$\varepsilon - \Delta_{i^\star_\varepsilon} = \rho_{\text{one}}.$$

Whenever

$$2\gamma_r < \rho_{\text{one}},$$

validity of the fused interval gives

$$U^{F,(r)}_{i^\star_\varepsilon} \leq \Delta_{i^\star_\varepsilon} + 2\gamma_r < \varepsilon.$$

Hence $i^\star_\varepsilon \in \mathcal{I}_r$ and the algorithm returns a valid jointly ε-optimal arm.

If ε^{one} identification is also defined for $\mathcal{S}_\varepsilon = \emptyset$, let

$$\rho_{\text{one}} = \min_{i\in[K]}(\Delta_i - \varepsilon).$$

Once $2\gamma_r < \rho_{\text{one}}$, every unresolved arm satisfies

$$L^{F,(r)}_i \geq \Delta_i - 2\gamma_r > \varepsilon,$$

so all remaining arms are rejected. With the corresponding empty-set stopping rule in Algorithm 1, the algorithm returns $\emptyset$.

For ε^{all} identification, assume $\rho_{\text{all}} > 0$. Once

$$2\gamma_r < \rho_{\text{all}},$$

every arm with $\Delta_i < \varepsilon$ has

$$U^{F,(r)}_i < \varepsilon,$$

while every arm with $\Delta_i > \varepsilon$ has

$$L^{F,(r)}_i \geq \varepsilon.$$

Thus all arms are correctly classified by phase R_{all}.

Since

$$R_q = \min\left\{r \geq 1 : \frac{2\gamma}{2^r} < \rho_q\right\},$$

either target mode stops by phase R_q on $\mathcal{E}$. Summing the phase-wise bounds gives

$$\sum_{r=1}^{R_q}(n_{R,r}+n_{D,r}) = O\Bigg(R_q\frac{d^2}{\zeta} + (1+\zeta)\sum_{r=1}^{R_q}\left[\frac{4^r\overline{P}_{R,r}}{\gamma^2}\log\frac{r^2|\overline{\mathcal{Y}}_{R,r}|}{\delta} + \overline{P}_{D,r}\log\frac{r^2\max\{K,|\Phi|\}}{\delta}\right]\Bigg).$$

Adding

$$n_B = O\left(\frac{d^2}{\zeta} + (1+\zeta)\kappa_0^{-1}d^2\log\frac{|\Phi|}{\delta} + (1+\zeta)\kappa_0^{-1}d\log^2\frac{|\Phi|}{\delta}\right)$$

and absorbing the burn-in rounding floor into $R_q d^2/\zeta$ yields

$$\tau^q_{\text{WIT-BOTELIM}} = O\Bigg(R_q\frac{d^2}{\zeta} + (1+\zeta)\kappa_0^{-1}d^2\log\frac{|\Phi|}{\delta} + (1+\zeta)\kappa_0^{-1}d\log^2\frac{|\Phi|}{\delta} + (1+\zeta)\sum_{r=1}^{R_q}\left[\frac{4^r\overline{P}_{R,r}}{\gamma^2}\log\frac{r^2|\overline{\mathcal{Y}}_{R,r}|}{\delta} + \overline{P}_{D,r}\log\frac{r^2\max\{K,|\Phi|\}}{\delta}\right]\Bigg),$$

which is the bound in Theorem 3. Correctness follows from the interval-validity argument on $\mathcal{E}$. □

C.11 Specialized Comparison of the Upper and Lower Bounds

We formalize the comparison discussed after Theorem 3. Throughout this subsection, suppose $i^\star_R = 1$, $i^\star_D = 2$, $\mathcal{S}_\varepsilon = 3$, and define $\eta_m := \varepsilon - \Delta_{m,3} > 0$ for $m \in R, D$. We further assume that arm 3 is closest to the modality-wise decision boundary, namely $|\varepsilon - \Delta_{m,3}| \le |\varepsilon - \Delta_{m,i}|$ for every $m \in R, D$ and $i \in [K]$.

Why the specialized regime is natural. The closest-boundary condition can arise when ε is selected using approximate gap information obtained from historical data or a pilot stage. The following observation formalizes the robustness of this regime to estimation error.

Lemma 28 (Robustness to approximate gap information)**.** *Suppose estimates $\widetilde{\Delta}_{m,i}$ satisfy $|\widetilde{\Delta}_{m,i} - \Delta_{m,i}| \le \beta_m$ for every $m \in R, D$ and $i \in [K]$. Assume that $\widetilde{\Delta}_{m,3} \le \varepsilon - \beta_m$ for both modalities, and that every $i \ne 3$ satisfies $\widetilde{\Delta}_{m,i} \ge \varepsilon + \beta_m$ for at least one modality m. If, in addition, $|\varepsilon - \widetilde{\Delta}_{m,i}| \ge |\varepsilon - \widetilde{\Delta}_{m,3}| + 2\beta_m$ for every m and $i \ne 3$, then $\mathcal{S}_\varepsilon = 3$ and arm 3 is closest to the true decision boundary in both modalities.*

Proof. The condition $\widetilde{\Delta}_{m,3} \le \varepsilon - \beta_m$ implies $\Delta_{m,3} \le \varepsilon$ for both modalities. For every $i \ne 3$, there exists a modality m such that $\widetilde{\Delta}_{m,i} \ge \varepsilon + \beta_m$, and therefore $\Delta_{m,i} \ge \varepsilon$. Hence $\mathcal{S}_\varepsilon = 3$ under the strict convention used in the main text.

Moreover, $|\varepsilon - \Delta_{m,i}| \ge |\varepsilon - \widetilde{\Delta}_{m,i}| - \beta_m$, whereas $|\varepsilon - \Delta_{m,3}| \le |\varepsilon - \widetilde{\Delta}_{m,3}| + \beta_m$. The assumed $2\beta_m$ separation therefore yields $|\varepsilon - \Delta_{m,i}| \ge |\varepsilon - \Delta_{m,3}|$ for every m and $i \ne 3$. □

Thus, exact prior knowledge of the gaps is not required: the specialized regime is preserved whenever the pilot estimation error is smaller than the separation between the estimated boundary distances.

Modality-specific active resolution. We next show that the deterministic envelope of modality m remains active only until the phase accuracy reaches the scale of its own boundary margin η_m.

Lemma 29 (Modality-specific envelope horizon)**.** *Under the specialized regime above, there exists a universal constant $C > 0$ such that, for either target mode and each $m \in R, D$,*

$$\sum_{\substack{r \le R_q: \\ \overline{\mathcal{Y}}_{m,r} \neq \emptyset}} \gamma_r^{-2} \le C \eta_m^{-2}.$$

Proof. Fix $m \in R, D$. Consider first an arm $i \neq 3$. If $\Delta_{m,i} > \varepsilon$, the closest-boundary assumption gives $\Delta_{m,i} \ge \varepsilon + \eta_m$. Hence i cannot belong to $\overline{G}_{m,r}$ once $2\gamma_{r-1} < \eta_m$, because the envelope requires $\Delta_{m,i} \le \varepsilon + 2\gamma_{r-1}$.

If instead $\Delta_{m,i} < \varepsilon$, the same assumption gives $\Delta_{m,i} \le \varepsilon - \eta_m$. Since $\mathcal{S}_\varepsilon = 3$, arm i must violate the ε condition in the opposite modality $\bar{m}$, so $\Delta_{\bar{m},i} > \varepsilon$ and therefore $\Delta_{\bar{m},i} \ge \varepsilon + \eta_{\bar{m}}$. Hence $\Delta_{\bar{m},i} - \Delta_{m,i} \ge \eta_m + \eta_{\bar{m}} \ge \eta_m$. Such an arm cannot belong to $\overline{G}_{m,r}$ once $4\gamma_{r-1} < \eta_m$, because the envelope additionally requires $\Delta_{\bar{m},i} - \Delta_{m,i} \le 4\gamma_{r-1}$.

It remains to consider arm 3. If modality m controls its fused gap, then $\eta_m = \min \eta_R, \eta_D$, and the global stopping condition reaches the scale $\Theta(\eta_m)$. If modality m does not control its fused gap, then $\eta_m > \eta_{\bar{m}}$ and $\Delta_{\bar{m},3} - \Delta_{m,3} = \eta_m - \eta_{\bar{m}}$. If $\eta_m - \eta_{\bar{m}} \ge \eta_m/2$, the cross-modal envelope condition removes modality m once $\gamma_{r-1} < \eta_m/8$. Otherwise $\eta_{\bar{m}} > \eta_m/2$, so the global stopping scale $\min \eta_R, \eta_D = \eta_{\bar{m}}$ is already within a constant factor of η_m.

Therefore, whenever $\overline{\mathcal{Y}}_{m,r}$ is nonempty, γ_r is bounded below by a universal constant multiple of η_m. Since γ_r decreases geometrically, summing γ_r^{-2} over the active phases gives the stated bound. □

elimination to directions centered at arm 3. For any positive definite matrix A and any $i, j \in [K]$, write $x_j - x_i = (x_j - x_3) - (x_i - x_3)$. Then

$$\|x_j - x_i\|_{A^{-1}}^2 \le 2\|x_j - x_3\|_{A^{-1}}^2 + 2\|x_i - x_3\|_{A^{-1}}^2 \le 4 \max_{k \neq 3} \|x_k - x_3\|_{A^{-1}}^2 .$$

Thus, up to a universal constant, the full pairwise geometry in the deterministic envelope is controlled by the star of directions centered at arm 3.

For reward feedback, define

$$A_R := \inf_{\nu_R \in \mathfrak{D}_R} \max_{k \neq 3} \|x_k - x_3\|_{M_R(\nu_R)^{-1}}^2 , \qquad J_R := \frac{A_R}{\eta_R^2}.$$

The preceding inequality implies $\overline{P}_{R,r} \le 4A_R$. Combining this with the modality-specific envelope horizon shows that the accumulated reward directional term in Theorem 3 is $\widetilde{O}(J_R \log(1/\delta))$, where the suppressed factors are the phase-dependent logarithms already displayed in the theorem.

For dueling feedback, define

$$A_D^F := \inf_{\nu_D \in \mathfrak{D}_D} \max_{k \neq 3} \|x_k - x_3\|_{H_D(\nu_D, \theta_D)^{-1}}^2 , \qquad J_D^F := \frac{A_D^F}{\eta_D^2}.$$

The active-resolution branch of the dueling design then contributes $\widetilde{O}(J_D^F \log(1/\delta))$. Since $H_D(\nu_D, \theta_D) \succeq \kappa_0 M_D(\nu_D)$, matrix inversion gives $H_D(\nu_D, \theta_D)^{-1} \preceq \kappa_0^{-1} M_D(\nu_D)^{-1}$, and therefore $J_D^F \le \kappa_0^{-1} J_D$, where

$$J_D := \inf_{\nu_D \in \mathfrak{D}_D} \max_{k \neq 3} \frac{\|x_k - x_3\|_{M_D(\nu_D)^{-1}}^2}{\eta_D^2}.$$

The global Fisher-leverage and active-resolution branches can also be separated up to a universal constant. Let ν_D^{glob} minimize the global-leverage objective and ν_D^{dir} minimize A_D^F. For their

equal mixture $\nu_D^{\mathrm{mix}} := (\nu_D^{\mathrm{glob}} + \nu_D^{\mathrm{dir}})/2$, linearity of the Fisher matrix gives $H_D(\nu_D^{\mathrm{mix}}, \theta_D) \succeq H_D(\nu_D^{\mathrm{glob}}, \theta_D)/2$ and $H_D(\nu_D^{\mathrm{mix}}, \theta_D) \succeq H_D(\nu_D^{\mathrm{dir}}, \theta_D)/2$. Hence each corresponding inverse-Fisher norm increases by at most a factor two. The full dueling design is therefore bounded, up to universal constants, by the sum of a global Fisher-leverage contribution and the directional contribution above.

Comparison with the lower-bound characteristic. For $m \in R, D$ and $k \neq 3$, define $d_k := x_k - x_3$ and $g_{m,k} := \varepsilon - d_k^\top \theta_m$. Since $i_m^\star$ maximizes $d_k^\top \theta_m$, we have $g_{m,i_m^\star} = \eta_m$ and $g_{m,k} \geq \eta_m$ for every $k \neq 3$. Recall

$$H_3^m = \inf_{\nu_m \in \mathfrak{D}_m} \max_{k \neq 3} \frac{\|d_k\|^2_{M_m(\nu_m)^{-1}}}{g_{m,k}^2}.$$

It follows immediately that $H_3^m \leq J_m$.

A reverse comparison requires an additional geometric condition describing which directions dominate an approximately optimal design.

Definition 30 (Closest-challenger dominance)**.** *For modality m, closest-challenger dominance holds with constant $c_m \geq 1$ if, for every $\xi > 0$, there exists $\nu_m^\xi \in \mathfrak{D}_m$ such that*

$$\max_{k \neq 3} \frac{\|d_k\|^2_{M_m(\nu_m^\xi)^{-1}}}{g_{m,k}^2} \leq H_3^m + \xi$$

and

$$\max_{k \neq 3} \|d_k\|^2_{M_m(\nu_m^\xi)^{-1}} \leq c_m \left\|d_{i_m^\star}\right\|^2_{M_m(\nu_m^\xi)^{-1}}.$$

The closest-challenger dominance condition is satisfied by broad classes of structured instances. For example, if the challenger directions are collinear with $x_{i_m^\star} - x_3$, so that $x_k - x_3 = a_k(x_{i_m^\star} - x_3)$ with uniformly bounded $|a_k|$, then $\|x_k - x_3\|^2_{M_m(\nu_m)^{-1}} = a_k^2 \|x_{i_m^\star} - x_3\|^2_{M_m(\nu_m)^{-1}}$ for every design ν_m, and the condition holds with a constant determined by the radial spread. More generally, if near-optimal designs for H_3^m have uniformly bounded condition number and the challenger directions have uniformly bounded relative Euclidean lengths, then there exists a uniform constant c_m such that $\max_{k \neq 3} \|x_k - x_3\|^2_{M_m(\nu_m)^{-1}} \leq c_m \|x_{i_m^\star} - x_3\|^2_{M_m(\nu_m)^{-1}}$. This includes approximately isotropic, symmetric, and tight-frame-type geometries, as well as bounded perturbations of such instances, showing that the condition holds for a broad family of structured arm sets.

Such condition requires the modality-wise optimal challenger, whose boundary slack is exactly η_m, to control the largest prediction variance under designs approaching the lower-bound optimum, up to a factor c_m.

Lemma 31 (Comparison of J_m and H_3^m)**.** *If closest-challenger dominance holds with constant c_m, then*

$$H_3^m \leq J_m \leq c_m H_3^m.$$

Proof. The first inequality follows from $g_{m,k} \geq \eta_m$. For the second, fix $\xi > 0$ and take ν_m^ξ from the definition. Since $g_{m,i_m^\star} = \eta_m$,

$$
\begin{aligned}
J_m &\le \max_{k \neq 3} \frac{\|d_k\|^2_{M_m(\nu_m^\xi)^{-1}}}{\eta_m^2} \\
&\le c_m \frac{\left\|d_{i_m^\star}\right\|^2_{M_m(\nu_m^\xi)^{-1}}}{\eta_m^2} \\
&= c_m \frac{\left\|d_{i_m^\star}\right\|^2_{M_m(\nu_m^\xi)^{-1}}}{g^2_{m,i_m^\star}} \le c_m(H_3^m + \xi).
\end{aligned}
$$

Letting $\xi \downarrow 0$ proves the claim. □

Consequently, over any family of specialized instances for which c_R, c_D, and κ_0^{-1} are uniformly bounded, the dominant directional part of Theorem 3 is $\widetilde{O}((H_3^R + H_3^D)\log(1/\delta))$. Since $\mathcal{S}_\varepsilon = 3$, Theorem 1 gives the corresponding lower order $\Omega((H_3^R + H_3^D)\log(1/\delta))$. This comparison concerns the instance-dependent directional terms: Theorem 3 is a high-probability upper bound, whereas Theorem 1 is an expected-sample lower bound, and the rounding, burn-in, and global Fisher-leverage terms remain additional overhead.

C.12 Effect of Witness and Bottleneck Compression

We finally formalize the comparison with an uncompressed phased design. For modality m, define $\mathcal{Y}_{m,r}^{\text{no-wit}} := x_j - x_i : i \in G_{m,r},, j \in [K],, j \neq i$ and $\mathcal{Y}_r^{\text{full}} := x_j - x_i : i \in C_r,, j \in [K],, j \neq i$. By construction, $\mathcal{Y}_{m,r} \subseteq \mathcal{Y}_{m,r}^{\text{no-wit}} \subseteq \mathcal{Y}_r^{\text{full}}$ pathwise.

For any direction sets $\mathcal{A} \subseteq \mathcal{B}$, monotonicity of the maximum gives

$$
\min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \mathcal{A}} \|y\|^2_{M_R(\nu_R)^{-1}} \le \min_{\nu_R \in \mathfrak{D}_R} \max_{y \in \mathcal{B}} \|y\|^2_{M_R(\nu_R)^{-1}} .
$$

The same conclusion holds for the active-direction branch of the dueling design:

$$
\min_{\nu_D \in \mathfrak{D}_D} \max_{y \in \mathcal{A}} \|y\|^2_{H_D(\nu_D,\vartheta)^{-1}} \le \min_{\nu_D \in \mathfrak{D}_D} \max_{y \in \mathcal{B}} \|y\|^2_{H_D(\nu_D,\vartheta)^{-1}} .
$$

Thus witness compression cannot increase the phase-wise directional design value, while bottleneck compression cannot increase it relative to the corresponding full target set. The global Fisher-leverage branch of the dueling design is unchanged because it depends on Φ rather than on the active direction set.

The deterministic theorem captures the bottleneck elimination explicitly. For $r \ge 2$, define $\overline{C}_r := i \in [K] : \Delta_{R,i} \le \varepsilon + 2\gamma_{r-1},, \Delta_{D,i} \le \varepsilon + 2\gamma_{r-1}$ and

$$
\overline{\mathcal{Y}}_r^{\text{full}} := \left\{x_j - x_i : i \in \overline{C}_r,\, j \in [K],\, j \neq i\right\}.
$$

Since $\overline{G}_{m,r} \subseteq \overline{C}_r$, we have $\overline{\mathcal{Y}}_{m,r} \subseteq \overline{\mathcal{Y}} * r^{\text{full}}$. Therefore the reward design value and the active-direction branch of the dueling design in Theorem 3 are no larger than their counterparts based on the full deterministic target envelope. Moreover, once $\overline{G} * m, r$ becomes empty, modality m incurs no further directional sampling cost.

By contrast, $\overline{\mathcal{Y}} * m, r$ deliberately retains every $j \in [K]$ as a possible competitor. Hence the deterministic upper bound does not capture the additional elimination due to the realized witness set $W * m, r - 1$. Witness elimination may further decrease the actual phase-wise design cost, but this gain is adaptive and path dependent rather than part of the deterministic worst-case envelope.

D Additional Experimental Details

This appendix provides the common experimental protocol, additional diagnostics for parameter misalignment, scaling results with respect to the number of arms and feature dimension, and ablations of the cross-modal elimination mechanisms.

D.1 Common Experimental Setup

Statistical parameters. Unless otherwise specified, all experiments use

$$\varepsilon = 0.20, \qquad \delta = 0.05, \qquad \sigma_R = 1.$$

Reward observations are generated with Gaussian noise of standard deviation σ_R. For WitBotElim, the initial phase resolution is set to

$$\gamma_0 = \varepsilon,$$

and is geometrically refined as

$$\gamma_r = \frac{\varepsilon}{2^r}.$$

We evaluate both ε^{one} identification, where the algorithm returns one arm in $\mathcal{S}_\varepsilon$, and ε^{all} identification, where it must recover the complete set $\mathcal{S}_\varepsilon$. All principal instances satisfy

$$|\mathcal{S}_\varepsilon| = 3.$$

Latent parameters. We control feedback heterogeneity through the angle between the reward and dueling parameters. We set

$$\theta_R = e_1, \qquad \theta_D = \cos(\alpha)e_1 + \sin(\alpha)e_2,$$

where α is the misalignment angle. Hence, $\alpha = 0^\circ$ corresponds to the aligned setting $\theta_R = \theta_D$, while larger α produces increasingly heterogeneous reward and dueling preferences.

Feature construction. The arms are constructed from prescribed modality-specific gaps rather than sampled independently at random. Let

$$g_{R,i} = \frac{\Delta_{R,i}}{\varepsilon}, \qquad g_{D,i} = \frac{\Delta_{D,i}}{\varepsilon}$$

denote the normalized reward and dueling gaps of arm i. Since only score differences affect the identification problem, we normalize the optimal score in each modality to zero. For $\alpha > 0$, writing the component of arm x_i in the signal plane as

$$x_i^{\parallel} = a_i e_1 + b_i e_2,$$

we choose

$$a_i = -\varepsilon g_{R,i}, \qquad b_i = \varepsilon \frac{g_{R,i}\cos\alpha - g_{D,i}}{\sin\alpha}.$$

This gives

$$x_i^\top \theta_R = -\varepsilon g_{R,i}, \qquad x_i^\top \theta_D = -\varepsilon g_{D,i},$$

and therefore realizes the desired reward and dueling gaps exactly. The aligned case is constructed directly from the common gap profile.

For dimensions larger than two, we add deterministic components orthogonal to $\operatorname{span}\{\theta_R, \theta_D\}$. These components do not alter the prescribed reward or dueling gaps, but make the feature geometry genuinely d-dimensional and full rank. All arms satisfy the imposed norm constraint. We use normalized gap levels such as

$$0.77, \qquad 0.88,$$

and additionally 0.96 in the larger scaling instances. These near-boundary arms ensure that the identification problem remains nontrivial across multiple phases.

Table 1 summarizes the experimental configurations. A seed count refers to independent observation seeds for each algorithm–instance–target-mode cell. When algorithms are compared on the same instance, we use matched observation seeds.

Table 1: Summary of the experimental configurations.

Study	(K, d)	Misalignment α	Seeds/cell
Misalignment	$(16, 4)$	$0, 30, 60, 90, 120^\circ$	500
Fine-grained supplement	$(16, 4)$	$15, 20, 25^\circ$	500
Scaling	$K \in \{16, 32\}$, $d \in \{4, 6, 8, 10, 12\}$	60°	1000
Mechanism ablation	$(32, 4)$	$0, 30, 60, 90, 120^\circ$	1000

D.2 Fine-Grained Misalignment

The main misalignment study uses a relatively coarse angular grid. We additionally examine smaller nonzero angles. For this diagnostic, however, merely choosing $\theta_R \neq \theta_D$ is insufficient: if either modality-specific optimum remains in $\mathcal{S}_\varepsilon$, an algorithm that effectively resolves only that modality may still succeed in the ε^{one} task. A more informative instance therefore makes the reward-optimal and dueling-optimal arms distinct and excludes both from the joint target set.

Let $x_R^\star$ and $x_D^\star$ denote the reward-optimal and dueling-optimal arms, respectively. To exclude each modal optimum because of the other modality, we require

$$\theta_R^\top (x_R^\star - x_D^\star) > \varepsilon$$

and

$$\theta_D^\top (x_D^\star - x_R^\star) > \varepsilon.$$

Adding the two inequalities yields

$$(\theta_R - \theta_D)^\top (x_R^\star - x_D^\star) > 2\varepsilon.$$

Under

$$\|\theta_R\| = \|\theta_D\| = 1 \qquad \text{and} \qquad \|x_i\| \leq 1,$$

we have

$$\|\theta_R - \theta_D\| = 2 \sin \frac{\alpha}{2}, \qquad \|x_R^\star - x_D^\star\| \leq 2.$$

Therefore,

$$(\theta_R - \theta_D)^\top (x_R^\star - x_D^\star) \leq 4 \sin \frac{\alpha}{2}.$$

A necessary condition for the desired split-winner construction is consequently

$$2 \sin \frac{\alpha}{2} > \varepsilon.$$

For $\varepsilon = 0.20$,

$$\alpha > 2 \arcsin(0.1) \approx 11.48^\circ.$$

This threshold motivates the supplemental angles

$$15^\circ, \qquad 20^\circ, \qquad 25^\circ.$$

They are substantially finer than the 30° spacing in the main study while remaining above the geometric threshold at which both single-modality optima can be excluded from $\mathcal{S}_\varepsilon$. Thus, success can no longer be achieved by identifying only the reward-optimal or only the dueling-optimal arm.

For all three supplemental instances, the reward and dueling optima are different and the joint target set contains three other arms. Table 2 reports the empirical success rates.

The behavior observed on the coarser angular grid is therefore already present at 15°, shortly above the smallest angle at which this type of conflicting joint-ε instance is geometrically feasible.

Table 2: Fine-grained identification accuracy under small parameter misalignment.

	ε^{one}		ε^{all}	
α	HyTS-GLB	WIT-BOTELIM	HyTS-GLB	WIT-BOTELIM
15°	0.0%	99.6%	0.0%	99.6%
20°	0.0%	100%	0.0%	100%
25°	0.0%	100%	0.0%	100%

Behavior of HyTS-GLB. The realized query allocation provides a direct explanation. In all three supplemental instances, HyTS-GLB allocates reward queries but no dueling queries. Its ε^{one} output therefore follows the reward-side estimate. By construction, the reward-optimal arm lies outside $\mathcal{S}_\varepsilon$, leading to zero empirical success in all three cells.

This observation also clarifies the role of the shared-parameter assumption. HyTS-GLB permits both feedback modalities, but under a shared latent representation, observations from one modality may be treated as sufficient for identifying the common parameter. On these heterogeneous instances, the realized allocation is entirely reward-side. In contrast, membership in our target set requires simultaneous certification under both θ_R and θ_D, and WIT-BOTELIM therefore maintains modality-specific certificates for the relevant directions.

D.3 Scaling with the Number of Arms and Feature Dimension

We next study how the sample complexity of WIT-BOTELIM changes with the size of the structured identification problem. We fix

$$\alpha = 60^\circ, \qquad K \in \{16, 32\}, \qquad d \in \{4, 6, 8, 10, 12\},$$

and use 1000 observation seeds per configuration.

The primary purpose of this study is to evaluate the dependence of the query cost on K and d, rather than to establish a separate algorithmic comparison. For both identification objectives and both values of K, the mean sample complexity increases as the feature dimension grows. Likewise, for a fixed dimension, the $K = 32$ instances consistently require more samples than their $K = 16$ counterparts.

Increasing d from 4 to 12 raises the mean sample complexity by approximately a factor of 3.1 for ε^{one} identification and by approximately a factor of 2.4 for ε^{all} identification. Doubling the number of arms from $K = 16$ to $K = 32$ increases the mean query cost by approximately 12%–23% for ε^{one} and 25%–43% for ε^{all}, depending on the dimension.

The increase in d reflects genuine geometric complexity rather than zero-padding. The additional coordinates are orthogonal to the reward–dueling signal plane and preserve the prescribed gaps, but they enlarge the experimental-design problem and produce full-rank d-dimensional feature sets. Exact-set recovery is consistently more expensive because every potentially relevant arm must be classified under both feedback modalities, whereas ε^{one} identification may terminate after a single jointly acceptable arm is certified.

D.4 Ablation of Cross-Modal Fusion

We use INDEPENDENTRAGESTYLE as the complete no-fusion baseline. It preserves the same modality-specific estimation, confidence, phased elimination, and experimental-design machinery as WIT-BOTELIM, but runs the reward and dueling identification procedures independently. In particular, the certification state of one modality is never used to modify the sampling targets or comparison directions of the other modality. The two sides interact only at the final decision stage through the intersection of their certified sets.

Hence, INDEPENDENTRAGESTYLE isolates the contribution of cross-modal fusion itself:

INDEPENDENTRAGESTYLE = complete modality-wise elimination without cross-modal fusion.

This baseline is conceptually different from the internal mechanism ablations below. The latter retain the fused framework and selectively remove witness or bottleneck compression.

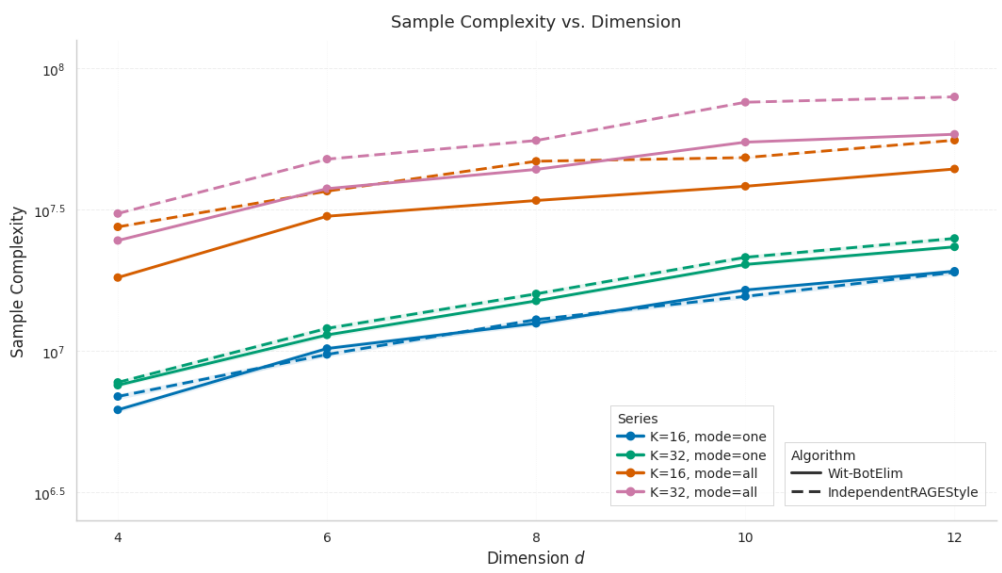


Figure 2: Sample complexity of WIT-BOTELIM and INDEPENDENTRAGESTYLE as the feature dimension increases. Colors distinguish the number of arms and identification objective, while solid and dashed curves distinguish the fused and no-fusion procedures. Shaded regions denote 95% confidence intervals over repeated trials.

No-fusion baseline. Figure 2 compares the complete fused method with INDEPENDENTRAGESTYLE across the scaling grid. Both methods use both feedback modalities; the difference is whether information certified on one side is allowed to simplify the experimental-design problem on the other.

For ε^{one} identification, the two procedures have similar query complexity, with an average elimination of approximately 2.7% for WIT-BOTELIM across the tested (K, d) configurations. The difference is much larger for exact-set recovery: for ε^{all} identification, WIT-BOTELIM uses fewer samples in every tested configuration, with an average elimination of approximately 23.8%. The larger gain is consistent with the need to classify all relevant arms under both modalities, which creates more opportunities for cross-modal pruning.

Witness and bottleneck ablations. After isolating the effect of removing cross-modal fusion entirely, we further decompose the two compression mechanisms within the fused framework. At phase r, the complete direction set has the form

$$\mathcal{Y}_{m,r} = \{x_j - x_i : i \in G_{m,r}, j \in W_{m,r-1}, i \neq j\}.$$

We independently ablate the target-side bottleneck set G and the comparator-side witness set W, yielding

$$\begin{aligned} \text{NOCOMPRESSION} &: C \times Z, \\ \text{WITNESSONLY} &: C \times W, \\ \text{BOTTLENECKONLY} &: G \times Z, \\ \text{FULL} &: G \times W. \end{aligned}$$

Here, FULL is exactly WIT-BOTELIM. Importantly, NOCOMPRESSION is not the same as INDEPENDENTRAGESTYLE: the former remains inside the fused framework and retains the shared cross-modal candidate state, whereas the latter maintains two completely independent modality-specific elimination processes until the final decision.

The four fused variants use the same estimators, confidence schedule, stopping conditions, and experimental-design machinery. Their correctness is essentially identical, allowing the sample-complexity differences to be interpreted as mechanism effects.

Table 3: Median sample-complexity elimination relative to NoCompression.

Mode	WitnessOnly	BottleneckOnly	Full
ε^{one}	28.3%	0.7%	28.5%
ε^{all}	31.0%	8.2%	42.3%

Table 4: Phase-wise comparison for the 60° ε^{all} instance. Direction counts are reported as reward/dueling counts.

Phase	NoComp. dirs.	Full dirs.	NoComp. samples	Full samples
1	1024/1024	992/992	0.50M	0.50M
2	1024/1024	61/31	0.64M	0.60M
3	812/812	27/28	2.27M	1.64M
4	160/162	6/7	7.67M	5.06M
5	68/64	3/3	30.06M	15.83M

Table 3 reports the median elimination relative to NoCompression, averaged over the nonzero-angle instances.

For ε^{one} identification, witness compression explains almost all of the median improvement. Replacing the broad comparator set Z with the safe witness set W sharply reduces the number of comparison directions, whereas target-side bottleneck compression provides little additional elimination in the typical trajectory.

For ε^{all} identification, the two mechanisms become complementary. Witness compression alone reduces median sample complexity by approximately 31%, while the complete $G \times W$ construction reduces it by approximately 42%. Bottleneck compression has an even stronger effect on difficult trajectories, where removing targets that no longer require resolution in one modality substantially reduces late-phase sampling.

Phase-wise behavior. Table 4 provides a phase-wise summary for the 60° ε^{all} instance. We report the median reward/dueling direction counts and median total samples requested within each phase.

The first phase is nearly identical across variants because little instance-specific information is available before the initial observations. The effect of witness pruning appears immediately afterward: by phase 2, the complete method has reduced the direction set from roughly one thousand comparisons per modality to only a few dozen.

The statistical value of compression becomes larger in later phases. Because the phase resolution decreases geometrically, resolving the remaining hard directions becomes progressively more expensive. At phase 5 in the representative ε^{all} cell, the median phase cost decreases from approximately 30.1M samples under NoCompression to 15.8M under Full.

The ε^{one} results help separate this effect from early stopping. Most trajectories terminate after approximately the same number of phases across the four fused variants. The roughly 28% elimination of Full is therefore primarily a within-phase effect: witness and bottleneck compression reduce the directions passed to the experimental-design problem rather than merely eliminating an entire phase. For ε^{all} identification, this compression becomes especially important on difficult late-phase trajectories.

Computational trade-off. The query savings require maintaining the cross-modal witness and bottleneck states and solving the corresponding structured experimental-design problems. The number of design-solver calls is approximately unchanged across the fused ablations, while the optimization effort increases only moderately. Thus, the empirical trade-off is between additional learner-side computation and fewer reward or preference queries, with the largest statistical benefit appearing in exact-set recovery.