Best Arm Identification in Generalized Linear Bandits via Hybrid Feedback

Qirun Zeng[1] **Xuchuang Wang**[2] **Jiayi Shen**[3]
Xutong Liu[4] **Fang Kong**[5] **Jinhang Zuo**[1*]

[1]City University of Hong Kong, Hong Kong SAR, China
[2]Hong Kong Baptist University, Hong Kong SAR, China
[3]University of Science and Technology of China, Hefei, China
[4]University of Washington Tacoma, Tacoma, WA, USA
[5]Southern University of Science and Technology, Shenzhen, China

qirun.zeng@my.cityu.edu.hk xuchuangw@hkbu.edu.hk
jiayishen@mail.ustc.edu.cn xutongl@uw.edu
kongf@sustech.edu.cn jinhang.zuo@cityu.edu.hk

Abstract

We study fixed-confidence best arm identification in generalized linear bandits under a hybrid feedback model: at each round, the learner may query either (i) absolute reward feedback from a single arm or (ii) relative (dueling) feedback from an arm pair, both governed by generalized linear models. We introduce a likelihood-ratio–based confidence sequence that unifies heterogeneous generalized linear observations and yields an explicit ellipsoidal confidence set under a self-concordance assumption. Building on this confidence set, we propose a hybrid Track-and-Stop algorithm that adaptively allocates queries by tracking a minimax-optimal design over a joint action space of arms and pairs. We establish δ-correctness and provide almost-sure asymptotic upper bounds on the stopping time. We further extend the framework to a cost-aware setting that accounts for heterogeneous acquisition costs across feedback modalities. Empirical experiments demonstrate that the proposed algorithms significantly improve sample efficiency over baseline methods.

1 Introduction

Large-scale interactive learning systems increasingly rely on heterogeneous user feedback. In modern conversational and recommendation platforms, especially LLM assistants trained with reinforcement learning from human feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022], supervision often comes from pairwise preferences, while absolute signals such as ratings, clicks, or task completion indicators may also be available. Although these feedback modalities may arise from a shared latent utility structure, they provide fundamentally different statistical information: absolute feedback offers noisy evaluations of individual actions, whereas relative feedback captures pairwise comparisons of their underlying utilities. Designing learning algorithms that can systematically exploit such heterogeneous yet complementary feedback remains a central challenge in interactive learning.

From a theoretical perspective, existing bandit models mostly treat different feedback modalities in isolation. Stochastic and linear bandits focus exclusively on reward-based observations, while dueling bandits and preference-based learning frameworks assume access only to pairwise comparisons [Yue and Joachims, 2011, Yue et al., 2012, Sui et al., 2018]. Recent work has started to bridge

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

this gap by studying hybrid feedback bandits [Wang et al., 2025]. However, existing approaches mainly combine reward and dueling feedback at the algorithmic level via joint elimination procedures. As a result, they do not fully exploit the shared latent structure underlying the two modalities within a unified statistical inference framework.

A natural way to move from algorithmic fusion to statistical integration is to model absolute and relative feedback as different observations of a shared latent utility. Generalized linear bandits (GLBs) provide such a framework: absolute observations provide noisy nonlinear feedback on individual arm utilities, while dueling observations provide noisy nonlinear feedback on utility differences. Thus, both modalities provide complementary information about the same latent parameter through distinct observation channels. This motivates our study of fixed-confidence best-arm identification (BAI) with hybrid feedback under generalized linear structure, where the goal is to identify the best arm with high confidence while adaptively choosing between absolute and relative queries.

Addressing this problem requires more than simply combining two feedback channels, and introduces several new challenges. First, reward and dueling feedback carry heterogeneous information: reward queries observe individual arm utilities, while dueling queries observe utility differences. Their distinct nonlinear likelihoods and curvature profiles make naive aggregation statistically inefficient. Second, fixed-confidence inference is harder under generalized linear structure, since uncertainty depends on both the unknown latent parameter and the queried arm or arm pair. This complicates the construction of time-uniform confidence regions for adaptive sampling and stopping. Third, exploration must be optimized jointly over actions and feedback modalities. The learner must decide which arms or arm pairs to query, when to use absolute or relative feedback, and how to account for potentially heterogeneous acquisition costs. Together, these challenges call for a unified inference, stopping, and allocation framework that estimates the shared latent utility, certifies the best arm, and adaptively balances absolute and relative feedback during pure exploration.

1.1 Our Contributions

We study fixed-confidence BAI in generalized linear bandits under a hybrid feedback model that allows both reward and dueling queries. Table 1 summarizes representative BAI results and highlights the gap addressed by this work. Our main contributions are summarized as follows.

- We formulate structured pure exploration with *both reward and dueling feedback* under a shared GLM parameterization. Unlike unstructured hybrid-bandit models [Wang et al., 2025], this setting transfers information through a latent parameter while allowing the two modalities to have different likelihoods and curvatures.
- We extend likelihood-ratio–based confidence sequences for self-concordant GLMs to hybrid observations. Despite the heterogeneous likelihoods induced by reward and dueling feedback, the resulting confidence region admits a single explicit ellipsoidal form, enabling unified inference and fixed-confidence stopping.
- Building on these confidence sets, we design HyTS-GLB, a Track-and-Stop–style algorithm that tracks a minimax design over the joint query space of arms and arm pairs. We prove δ-correctness and almost-sure asymptotic upper bounds on the stopping time as $\delta \to 0$.
- We extend HyTS-GLB to *heterogeneous acquisition costs* and provide an almost-sure asymptotic upper bound on total cost as $\delta \to 0$. This lets the learner trade off both information and cost across reward and dueling queries.
- We conduct experiments showing that hybrid feedback reduces sample complexity relative to single-modality baselines, while the cost-aware variant further adapts to asymmetric query costs across modalities.

Together, these contributions yield a unified statistical and algorithmic framework for fixed-confidence best-arm identification with hybrid feedback in generalized linear bandits. All detailed proofs are deferred to the appendix due to space limitations.

1.2 Related Work

Generalized Linear Bandits. GLMs [McCullagh, 2019] model exponential-family observations with natural parameter $\boldsymbol{x}^\top \boldsymbol{\theta}^\star$ and mean $\mu(\boldsymbol{x}^\top \boldsymbol{\theta}^\star)$. Filippi et al. [2010] introduced GLM-based structured

Table 1: Comparison of fixed-confidence BAI across feedback models and structural assumptions.

	Model	C_{reward}	C_{dueling} *	Sample Complexity†
Jedra and Proutiere [2020]	Linear	1	$\emptyset$	$\tilde{\mathcal{O}}\big(T^\star_{1,\emptyset}(\boldsymbol{\theta}^\star)\log\frac{1}{\delta}\big)$
Jun et al. [2021]	Logistic	1 ($\emptyset$)	$\emptyset$ (1)	$\tilde{\mathcal{O}}\left(T'_{1(\emptyset),\emptyset(1)}(\boldsymbol{\theta}^\star)\log\frac{1}{\delta}\right)$
Rivera and Tewari [2024]	Logistic	1	$\emptyset$	$\tilde{\mathcal{O}}\left(T^\star_{1,\emptyset}(\boldsymbol{\theta}^\star)\log^2\frac{1}{\delta}\right)$
Kazerouni and Wein [2021]	GLM	1	$\emptyset$	$\tilde{\mathcal{O}}\left(T''_{1,\emptyset}(\boldsymbol{\theta}^\star)\log^2\frac{1}{\delta}\right)$
HyTS (Sec. 4.1)	Hybrid GLM	1	1	$\tilde{\mathcal{O}}\big(T^\star_{1,1}(\boldsymbol{\theta}^\star)\log\frac{1}{\delta}\big)$
Cost-aware-HyTS (Sec. 4.2)	Hybrid GLM	a	b	$\tilde{\mathcal{O}}\big(T^\star_{a,b}(\boldsymbol{\theta}^\star)\log\frac{1}{\delta}\big)$

* C_{reward} and C_{dueling} denote the acquisition cost of reward and dueling feedback, respectively, with $\emptyset$ indicating unavailable.
† The sample complexity is expressed in terms of model-specific characteristic times or complexity measures. For our algorithms, $T^\star$ denotes the characteristic time induced by the minimax-optimal design under the corresponding feedback costs (Section 4.1). T' and T'' denote the finite-time complexity measures of Jun et al. [2021] and Kazerouni and Wein [2021], respectively.

bandits, followed by work on regret and inference under assumptions such as self-concordance [Russac et al., 2021]. Lee et al. [2024] developed likelihood-ratio–based confidence sequences for self-concordant GLMs, yielding convex and numerically tight confidence regions. We go beyond Lee et al. [2024] by developing confidence sequences for hybrid reward–dueling feedback, where adaptive queries may arise from different GLM channels. This requires a single confidence certificate that aggregates heterogeneous Fisher curvature and feature geometry across modalities, a setting absent in prior single-channel GLM analyses. Beyond that, Kirschner et al. [2025], Clerico et al. [2025] also construct similar confidence sequences for GLMs.

Best-Arm Identification and Track-and-Stop. Best-arm identification has been widely studied for stochastic and linear bandits [Bubeck et al., 2009, Audibert and Bubeck, 2010, Soare et al., 2014, Jourdan and Degenne, 2022, Shao and Fang, 2025]. Garivier and Kaufmann [2016] established a general fixed-confidence lower bound for multi-armed bandits and proposed Track-and-Stop, which asymptotically achieves it. For linear bandits, Soare [2015] derived an instance-dependent lower bound, and Jedra and Proutiere [2020] gave a matching algorithm. Degenne et al. [2020] showed that iterative saddle-point solvers can design asymptotically optimal BAI algorithms for structured bandits. For GLBs, Kazerouni and Wein [2021] studied (ϵ, δ)-PAC BAI and provided corresponding sample-complexity guarantees.

Relative and Hybrid Feedback. Relative (dueling) feedback has been studied extensively in bandit learning, particularly in dueling bandits and preference-based learning [Yue and Joachims, 2011, Yue et al., 2012, Sui et al., 2018]. It is also practically attractive because relative feedback can be easier to elicit than absolute rewards in conversational recommendation [Zhang et al., 2020, Yang et al., 2024, Zou et al., 2026]. Most closely related, Wang et al. [2025] study stochastic bandits with reward and dueling feedback and propose regret-minimization algorithms based on elimination and decomposition, showing that hybrid feedback helps even without structure. We instead study fixed-confidence BAI under shared GLM structure, which requires time-uniform confidence sequences, confidence-based stopping, and design-based allocation over arms and arm pairs. Related hybrid settings have also been explored by He et al. [2024] for multi-armed bandits, dueling bandits, and bandits with offline data.

2 Problem Formulation

Generalized Linear Model. We begin by introducing the generalized linear model [McCullagh, 2019] underlying both reward and dueling feedback. Let $r \in \mathbb{R}$ denote a generic scalar observation. In a GLM, the conditional distribution of r given a scalar parameter $\eta \in \mathbb{R}$ belongs to an exponential family and is specified by

$$p(r \mid \eta) = \exp\left(\frac{r\eta - b(\eta)}{\zeta(\varphi)} + c(r, \varphi)\right), \tag{1}$$

where η is the canonical (natural) parameter, φ is a dispersion parameter, $b(\cdot)$ is the log-partition function, and $c(r, \varphi)$ is the base measure. We use $p(r \mid \eta)$ to denote the conditional probability density (or mass) function of r given η. A fundamental property of GLMs is that the conditional mean of r is given by $\mathbb{E}[r \mid \eta] = b'(\eta) \triangleq \mu(\eta)$, where $\mu(\cdot)$ is the mean function (inverse link function).

Many common observation models, including Gaussian, Poisson, and Bernoulli distributions, fall within this framework.

Hybrid Generalized Linear Bandits. We consider a stochastic decision-making problem with K arms indexed by $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$. Each arm $i \in \mathcal{K}$ is associated with a known feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, and the environment is governed by an unknown parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$. We assume the set of arms $\mathcal{K}$ spans $\mathbb{R}^d$. We also assume that the dueling feature vectors span $\mathbb{R}^d$, i.e., $\operatorname{span}\{\boldsymbol{x}_i - \boldsymbol{x}_j : 1 \le i < j \le K\} = \mathbb{R}^d$. The learner interacts with the environment sequentially and may acquire information through two different feedback modalities.

At each round t, the learner chooses either a *reward query* or a *dueling query*. In the former, it selects $i_t \in \mathcal{K}$ and observes $R_t \sim p_{\mathrm{c}}(\cdot \mid \boldsymbol{x}_{i_t}^\top \boldsymbol{\theta}^\star)$. In the latter, it selects $(j_t, k_t) \in \mathcal{G} \triangleq \{(j,k) \in \mathcal{K}^2 : j < k\}$ and observes $Y_t \in \{0, 1\}$ drawn from $p_{\mathrm{d}}(\cdot \mid \boldsymbol{x}_{j_t,k_t}^\top \boldsymbol{\theta}^\star)$, where $\boldsymbol{x}_{j,k} \triangleq \boldsymbol{x}_j - \boldsymbol{x}_k$, and $Y_t = 1$ indicates preference for j_t.

Both feedback modalities are modeled as GLMs with a shared parameter $\boldsymbol{\theta}^\star$, but possibly different observation spaces and mean functions μ_{c} and μ_{d}. We use subscripts c and d to distinguish reward and dueling quantities. Define the joint action space $\mathcal{A} \triangleq \mathcal{K} \cup \mathcal{G}$ and, for $a = (j,k) \in \mathcal{G}$, write $\boldsymbol{x}_a = \boldsymbol{x}_{j,k}$. For each $a \in \mathcal{A}$, we define the feedback modality $m(a)$ to be c for $a \in \mathcal{K}$ and d for $a \in \mathcal{G}$. Selecting $a_t \in \mathcal{A}$ corresponds to querying a single arm ($a_t \in \mathcal{K}$) or a pair ($a_t \in \mathcal{G}$), with observations generated by the corresponding model $p_{m(a_t)}(\cdot \mid \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^\star)$.

Remark 2.1. The shared parameter $\boldsymbol{\theta}^\star$ represents a latent user- or evaluator-preference vector, which is natural in recommender systems and RLHF-style data collection (e.g., Yue et al. [2012], Negahban et al. [2012], Christiano et al. [2017], Yang et al. [2024]). An item or response may receive an absolute rating through $\mu_{\mathrm{c}}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star)$, while a pairwise comparison is generated from the utility difference through $\mu_{\mathrm{d}}((\boldsymbol{x}_j - \boldsymbol{x}_k)^\top \boldsymbol{\theta}^\star)$. Thus, the two modalities may have different links and noise levels while still providing complementary observations of the same underlying preference structure. This flexibility is not merely cosmetic: choosing different link functions for reward and dueling feedback can change the local curvature and information carried by each query, and may yield gains in regimes where one modality is saturated, noisier, or better aligned with the relevant decision boundary.

Best Arm Identification Objective. Under the hybrid generalized linear bandit problem described above, we study a fixed-confidence best arm identification (BAI) problem. The learner sequentially queries reward or dueling feedback and aims to identify the single arm with the highest expected reward using as few samples as possible. Formally, the optimal arm is defined as

$$i^\star \triangleq \arg\max_{i \in \mathcal{K}} \mu_{\mathrm{c}}\big(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star\big)\,,$$

and we assume it is unique. We consider adaptive strategies that select both the feedback modality and the queried arms over time. Let τ denote a stopping time and let $\hat{i}_\tau^\star$ be the arm recommended upon termination.

Definition 2.2. Given a confidence level $\delta \in (0,1)$, an exploration strategy is said to be δ-*correct* if, for every $\boldsymbol{\theta} \in \Theta$ with a unique best arm $i^\star(\boldsymbol{\theta})$,

$$\Pr_{\boldsymbol{\theta}}(\hat{i}_\tau^\star \neq i^\star(\boldsymbol{\theta})) \le \delta \quad \text{and} \quad \Pr_{\boldsymbol{\theta}}(\tau < +\infty) = 1.$$

The goal is to design a δ-correct exploration strategy with a small stopping time. We adopt the following standard assumptions from the generalized linear bandit literature [Russac et al., 2021, Lee et al., 2024].

Assumption 2.3 (Bounded parameter)**.** The unknown parameter satisfies $\boldsymbol{\theta}^\star \in \Theta \subseteq \mathcal{B}^d(S) \triangleq \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le S\}$, where Θ is compact and convex.

Assumption 2.4 (Bounded features)**.** The arm feature set satisfies $\mathcal{X} \triangleq \{\boldsymbol{x}_i : i \in \mathcal{K}\} \subseteq \mathcal{B}^d(1)$. This implies $\|\boldsymbol{x}_{j,k}\|_2 \le 2$ for all $(j,k) \in \mathcal{G}$.

Assumption 2.5 (Smoothness and convexity)**.** The log-partition functions $b_{\mathrm{c}}(\cdot)$ and $b_{\mathrm{d}}(\cdot)$ are three times continuously differentiable and convex, with $\mu'_{m(a)}(\cdot) = b''_{m(a)}(\cdot) \ge 0$ for every $a \in \mathcal{A}$.

Bounded parameters and features control the predictors and likelihood derivatives. Compactness of Θ ensures MLE existence in Section 3.1, while convexity makes its computation a convex problem. Monotonicity of μ_{c} and uniqueness of the best arm imply $i^\star = \arg\max_{i \in \mathcal{K}} \boldsymbol{x}_i^\top \boldsymbol{\theta}^\star$.

3 Confidence Sequences for Hybrid GLB

In this section, we develop time-uniform confidence sequences (CS) for $\boldsymbol{\theta}^\star$ under the hybrid feedback model. Our construction follows the likelihood-ratio framework of Lee et al. [2024], adapted to mixed feedback. We first build a likelihood-ratio–based sequence, then use self-concordance to derive an explicit ellipsoidal set for algorithmic design.

3.1 Maximum Likelihood Estimation

To distinguish between reward and dueling feedback, we define the following index sets according to the action a_s selected at round s: $\mathcal{C}_t \triangleq \{1 \le s < t : a_s \in \mathcal{K}\}$ and $\mathcal{D}_t \triangleq \{1 \le s < t : a_s \in \mathcal{G}\}$. Let Z_s denote the observation at round s, i.e., $Z_s = R_s$ if $a_s \in \mathcal{K}$ and $Z_s = Y_s$ if $a_s \in \mathcal{G}$.

Under the GLM in Equation (1), the cumulative negative log-likelihood at time t can be decomposed as

$$\mathcal{L}_t(\boldsymbol{\theta}) \triangleq \sum_{s=1}^{t-1} \frac{b_{m(a_s)}(\boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}) - Z_s \boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}}{\zeta_{m(a_s)}(\varphi_{m(a_s)})}.$$

We define the ℓ_2-constrained maximum likelihood estimator (MLE) as

$$\hat{\boldsymbol{\theta}}_t \triangleq \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_t(\boldsymbol{\theta}), \tag{2}$$

where $\Theta \subset \mathbb{R}^d$ is a compact, convex parameter set. Equivalently, whenever $\hat{\boldsymbol{\theta}}_t$ lies in the interior of Θ, it satisfies the first-order optimality condition $\nabla \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) = 0$.

3.2 Likelihood-Ratio Confidence Sequences

We first construct a likelihood-ratio confidence sequence based on the cumulative loss. Let L_t be any almost-sure upper bound on the Lipschitz modulus of $\mathcal{L}_t$ over Θ:

$$L_t \ge \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta} \frac{|\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}')|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2}.$$

Following Lee et al. [2024], we use an upper bound on the Lipschitz modulus that holds almost surely. In the hybrid model, L_t decomposes over reward and dueling observations and thus scales with their sample counts. Since $\|\boldsymbol{x}_{i,j}\|_2 \le 2$ while $\|\boldsymbol{x}_i\|_2 \le 1$, the dueling term incurs an additional factor of 2.

Lemma 3.1 (Unified likelihood-ratio CS for hybrid GLMs, Theorem 3.1 of Lee et al. [2024])**.** *Suppose Assumptions 2.3–2.5 hold. Then for any $\delta \in (0, 1)$,*

$$\Pr\left(\exists\, t \ge 1 : \mathcal{L}_t(\boldsymbol{\theta}^\star) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \ge \beta_t(\delta)\right) \le \delta,$$

with a time-uniform confidence radius:

$$\beta_t(\delta) \triangleq \log \tfrac{1}{\delta} + \inf_{c_t \in (0,1]} \left\{ d \log \tfrac{1}{c_t} + 2SL_t c_t \right\} \le \log \tfrac{1}{\delta} + d \log\left(\max\left\{e,\ \tfrac{2eSL_t}{d}\right\}\right). \tag{3}$$

The bound constructed in Appendix A.1 satisfies $L_t = O(t)$ almost surely, and hence $\beta_t(\delta) = O(\log(1/\delta) + d \log t)$ almost surely as $t \to \infty$.

3.3 Ellipsoidal Confidence Sequences under Self-Concordance

The likelihood-ratio CS controls the loss difference but is not directly helpful to algorithmic design. We therefore exploit self-concordance to convert it into an ellipsoidal confidence sequence.

Assumption 3.2 (Self-concordant GLMs [Russac et al., 2021])**.** The GLMs of reward and dueling feedback are self-concordant. That is, there exist finite constants M_{c} and M_{d} such that, for all $a \in \mathcal{A}$, $|\mu''_{m(a)}(\eta)| \le M_{m(a)} \mu'_{m(a)}(\eta), \forall\, \eta \in \{\boldsymbol{x}_a^\top \boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta\}$.

This assumption enables the ellipsoidal conversion in Theorem 3.3. Although $\hat{\boldsymbol{\theta}}_t$ minimizes the combined loss $\mathcal{L}_t$ rather than each component, self-concordance holds uniformly over Θ and does not require it to minimize $\mathcal{L}_t^{\mathrm{c}}$ or $\mathcal{L}_t^{\mathrm{d}}$. This is what lets the two feedback streams share one estimator

and one confidence certificate rather than producing two separate ellipsoids that would have to be reconciled afterward. To unify the two feedback modalities in inference and design, we define the aggregated information matrix

$$\boldsymbol{A}_t \triangleq \sum_{s=1}^{t-1} \frac{\mu'_{m(a_s)}(\boldsymbol{x}_{a_s}^\top \hat{\boldsymbol{\theta}}_t)}{2(1 + S\rho_{a_s} M_{m(a_s)})\zeta_{m(a_s)}(\varphi_{m(a_s)})} \boldsymbol{x}_{a_s} \boldsymbol{x}_{a_s}^\top,$$

where $\rho_a = 1$ for $a \in \mathcal{K}$ and $\rho_a = 2$ for $a \in \mathcal{G}$. Intuitively, the matrix $\boldsymbol{A}_t$ aggregates curvature information from both feedback types and serves as the effective information matrix in the ellipsoidal confidence set.

Identifiability. Throughout, we impose the identifiability condition (Kazerouni and Wein [2021]): there exists a (random but finite) time E such that the matrix $\boldsymbol{A}_E$ is positive definite. Consequently, $\boldsymbol{A}_t \succ 0$ for all $t \geq E$, and $\boldsymbol{A}_t^{-1}$ is well defined.

Despite heterogeneous curvature induced by reward and dueling observations, their contributions combine additively into a single information matrix, yielding a unified ellipsoidal confidence region.

Theorem 3.3 (CS for hybrid GLMs)**.** *Define* $\mathcal{E}_t(\delta) \triangleq \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{A}_t} \leq \beta_t(\delta)\}$. *We have*

$$\Pr\left(\exists\, t \geq 1 : \boldsymbol{\theta}^\star \notin \mathcal{E}_t(\delta)\right) \leq \delta.$$

This ellipsoidal form enables a simple stopping rule and a design-based allocation strategy.

Proof sketch. It suffices to show $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{A_t} \leq \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t)$. By Lee et al. [2024, Lemma D.1] and Assumption 3.2, each component loss admits the corresponding quadratic lower bound at $\hat{\boldsymbol{\theta}}_t$; summing yields $\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \geq \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{A_t}$, which with Lemma 3.1 concludes the proof. □

Relation to Lee et al. [2024]. Our construction adapts the likelihood-ratio analysis to reward and dueling channels with potentially different mean functions and dispersion scales. The matrix $\boldsymbol{A}_t$ combines their Hessian contributions according to the realized queries and modality-specific self-concordance factors, preserving each channel's curvature and feature geometry. The resulting ellipsoid provides a common confidence certificate for the stopping rule and guides query allocation across the two feedback channels.

4 Hybrid Track-and-Stop Algorithms

This section introduces a hybrid sampling framework that integrates reward and dueling feedback within a unified pure-exploration procedure. The central idea is to exploit the complementary statistical information provided by the two feedback modalities, while adaptively allocating samples across them. Our approach extends the Track-and-Stop paradigm [Garivier and Kaufmann, 2016, Jedra and Proutiere, 2020] to hybrid generalized linear bandits.

4.1 HyTS-GLB Algorithm

Minimax-Optimal Design. Following the Track-and-Stop principle, we define sampling proportions through a minimax optimal-design problem. The criterion allocates samples so as to minimize the largest confidence width among all comparisons between the current best arm and its competitors. It is closely related to classical G-optimal design [Kiefer and Wolfowitz, 1960, Pukelsheim, 2006], which has played a central role in best-arm identification for linear bandits [Soare et al., 2014].

To leverage the ellipsoidal confidence sets for efficient pure exploration, we now specify how samples should be allocated across actions. Let $\hat{\boldsymbol{\theta}}_t$ denote the maximum likelihood estimator at time t, and define the current empirical best arm, with ties broken arbitrarily, as $\hat{i}_t^\star = \arg\max_{i \in \mathcal{K}} \boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_t$.

Given a sampling weight vector $\boldsymbol{w} \in \Delta_{\mathcal{A}} \triangleq \{\boldsymbol{w} \mid w_a \geq 0 \wedge \sum_{a \in \mathcal{A}} w_a = 1\}$, we define the hybrid information matrix

$$\boldsymbol{A}(\boldsymbol{w}, \boldsymbol{\theta}) \triangleq \sum_{a \in \mathcal{A}} \boldsymbol{w}_a \frac{\mu'_{m(a)}(\boldsymbol{x}_a^\top \boldsymbol{\theta})}{2(1 + S\rho_a M_{m(a)})\zeta_{m(a)}(\varphi_{m(a)})} \boldsymbol{x}_a \boldsymbol{x}_a^\top.$$

The minimax-optimal sampling proportions $\boldsymbol{w}^\star(t)$ are any solution to

$$\boldsymbol{w}^\star(t) = \arg\min_{\boldsymbol{w}\in\Delta_{\mathcal{A}}} \phi(\hat{\boldsymbol{\theta}}_t, \boldsymbol{w}) \triangleq \begin{cases} \max_{i\neq \hat{i}_t^\star} \left\| \boldsymbol{x}_{\hat{i}_t^\star,i} \right\|^2_{\boldsymbol{A}(\boldsymbol{w},\hat{\boldsymbol{\theta}}_t)^{-1}}, & \text{if } \boldsymbol{A}(\boldsymbol{w},\hat{\boldsymbol{\theta}}_t) \succ 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Since they specify asymptotic sampling frequencies rather than an explicit sequential policy, we adopt a tracking rule to realize these proportions online.

Tracking Rule. The learner tracks the target proportions by selecting the most under-sampled action relative to its cumulative target mass, while incorporating a vanishing forced-exploration mechanism to guarantee persistent excitation. Let $\mathcal{A}_0 \subset \mathcal{A}$ be a fixed subset of arms such that $\lambda_{\min}\big(\sum_{a\in\mathcal{A}_0} \boldsymbol{x}_a \boldsymbol{x}_a^\top\big) > 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue, and let π_0 denote a probability distribution supported on $\mathcal{A}_0$ with $\pi_0(a) > 0$ for every $a \in \mathcal{A}_0$ (e.g., the uniform distribution). Let $\mathcal{T}_t$ be the set of tracking rounds up to time t, and let $N_a^{\mathcal{T}}(t) = \sum_{s\in\mathcal{T}_t} \mathbf{1}\{a_s = a\}$ and $W_a^{\mathcal{T}}(t) = \sum_{s\in\mathcal{T}_t} w_a^\star(s-1)$ denote the corresponding pull counts and cumulative target mass. At each round,

$$a_t \in \begin{cases} \arg\min_{a\in\mathcal{A}} \big(N_a^{\mathcal{T}}(t-1) - W_a^{\mathcal{T}}(t-1)\big), & \text{w.p. } 1-\epsilon_t, \\ a \sim \pi_0, & \text{w.p. } \epsilon_t, \end{cases} \tag{4}$$

where $\epsilon_t = t^{-\alpha}$ for some $\alpha \in (0,1)$. The forced-exploration component guarantees the non-degeneracy of the information matrix at only an asymptotically lower-order cost, while the tracking component ensures that the empirical proportions approach the optimal-design set. Moreover, we assume the curvature is bounded away from zero on the relevant domain (Assumption 4.1), a standard condition in best arm identification that ensures each exploration action provides non-vanishing curvature [Jun et al., 2021, Kazerouni and Wein, 2021]; in these works, the sample complexity also depends on κ^{-1}.

Assumption 4.1. There exists $\kappa > 0$ such that $\inf_{a\in\mathcal{A},\boldsymbol{\theta}\in\Theta} \mu'_{m(a)}(\boldsymbol{x}_a^\top\boldsymbol{\theta}) \geq \kappa$.

Remark 4.2. With forced exploration, Assumption 4.1 ensures information growth for consistency, almost-sure stopping, and asymptotic complexity. It is not needed for confidence coverage or error control at finite stopping times. If minimax-optimal tracking alone ensures persistent excitation, i.e., the induced information matrix is almost surely full rank with $\lambda_{\min}(A_t) \gtrsim t^\alpha$ for some $\alpha > 0$, then forced exploration is unnecessary and one may set $\epsilon_t \equiv 0$.

Stopping Rule. The algorithm stops at the first time t such that the current empirically best arm $\hat{i}_t^\star$ is certified optimal against all competitors with high confidence:

$$\forall j \neq \hat{i}_t^\star : \quad \inf_{\boldsymbol{\theta}\in\mathcal{E}_t(\delta)} \boldsymbol{x}_{\hat{i}_t^\star,j}^\top \boldsymbol{\theta} > 0. \tag{5}$$

Theorem 4.3 (Correctness of the stopping rule)**.** *The stopping rule* (5) *guarantees that, upon stopping, the returned arm satisfies* $\hat{i}_\tau^\star = i^\star$ *with probability at least* $1-\delta$.

Proof. Equation (5) implies that if $\boldsymbol{\theta}^\star \in \mathcal{E}_t(\delta)$, $\boldsymbol{x}_{\hat{i}_t^\star,j}^\top \boldsymbol{\theta}^\star > 0, \forall j \neq \hat{i}_t^\star$, which yields $\hat{i}_t^\star = \arg\max_i \boldsymbol{x}_i^\top \boldsymbol{\theta}^\star$. The time-uniform coverage of $\mathcal{E}_t(\delta)$ ensures this event holds with probability at least $1-\delta$. □

We now describe our proposed algorithm, Hybrid Track-and-Stop for GLBs, in Algorithm 1. For warm-up, querying d actions with linearly independent feature vectors once each is enough; Assumption 4.1 then ensures $\boldsymbol{A}_{d+1} \succ 0$, so $E = d$ suffices and is independent of δ. The constant κ is an analytical property of the links and parameter domain, not an input or a tuning parameter of the algorithm.

Definition 4.4 (Minimax-Optimal characteristic time)**.** For a given instance $\boldsymbol{\theta}^\star$, define the minimax characteristic time

$$T^\star(\boldsymbol{\theta}^\star) \triangleq \inf_{\boldsymbol{w}\in\Delta_{\mathcal{A}}} \frac{\max_{i\neq i^\star} \|\boldsymbol{x}_{i^\star,i}\|^2_{\boldsymbol{A}(\boldsymbol{w},\boldsymbol{\theta}^\star)^{-1}}}{\Delta_{\min}^2},$$

where $\Delta_{\min} \triangleq \min_{i\neq i^\star} \boldsymbol{x}_{i^\star,i}^\top \boldsymbol{\theta}^\star$ denotes the smallest optimality gap.

Algorithm 1: HyTS-GLB

Input: Confidence level δ, warm-up rounds E

Warm-up: select actions for E rounds to ensure $\boldsymbol{A}_{E+1}$ is positive definite;
for $t = E+1, E+2, \ldots$ **do**
  Compute the MLE $\hat{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_t(\boldsymbol{\theta})$;
  if *the stopping condition* (5) *holds* **then**
    return $\hat{i}_t^\star = \arg\max_{i\in\mathcal{K}} \boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_t$;
  end
  Compute the target proportions $\boldsymbol{w}^\star(t)$ and then select action a_t according to Equation (4);
  Execute action a_t and observe feedback Z_t, then record the information;
end

This design-dependent quantity measures the cost of statistically separating the optimal arm from sub-optimal arms under the worst comparison induced by the confidence-width stopping rule. Appendix B discusses an information-theoretic lower bound for hybrid GLM BAI and compares it directly with $T^\star(\boldsymbol{\theta}^\star)$ in the linear–Gaussian benchmark. This supports interpreting HyTS-GLB as tracking the minimax allocation induced by our confidence-width criterion, while stopping-time guarantees still incur conservative constants and logarithmic factors.

Theorem 4.5 (Almost sure sample complexity upper bound)**.** *Assume Assumptions 2.3 to 2.5, 3.2 and 4.1 hold. Then HyTS-GLB's sample complexity τ verifies*

$$\Pr\left(\limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \lesssim T^\star(\boldsymbol{\theta}^\star)\right) = 1.$$

Proof sketch. Ignoring the negligible forced exploration phase, we focus on tracking. As the empirical proportions approach the design set, the largest comparison variance is asymptotically controlled by $\Delta_{\min}^2 T^\star(\boldsymbol{\theta}^\star)/t$. Stopping is ensured once the confidence radius is below $\Delta_{\min}/2$. Hence $t \gtrsim T^\star(\boldsymbol{\theta}^\star)\beta_t(\delta)$ suffices, and the claim follows by a standard fixed-point argument. □

Theorem 4.5 shows that HyTS-GLB achieves instance-dependent sample complexity governed by the minimax characteristic time $T^\star(\boldsymbol{\theta}^\star)$, which captures the joint information from reward and dueling feedback under the adopted confidence-width criterion. Unlike single-modality strategies, the minimax design underlying $T^\star(\boldsymbol{\theta}^\star)$ balances absolute and relative information, enabling faster separation along informative directions. Consequently, HyTS-GLB can strictly outperform reward-only or dueling-only methods when neither modality alone is sufficient.

Relation to Jun et al. [2021]. They study transductive pure exploration under a common logistic observation model, using variance-dependent fixed-design confidence bounds and the phased elimination algorithm RAGE-GLM. Taking their measurement set to contain both arm features and pairwise differences also covers the special case in which our two channels share the same logistic model. Our framework allows the channels to have different GLM likelihoods and combines a time-uniform likelihood-ratio confidence sequence with Track-and-Stop allocation over the joint query space.

4.2 Cost-Aware HyTS-GLB

In many applications, different feedback modalities incur heterogeneous acquisition costs. To capture this asymmetry, we extend HyTS-GLB to a cost-aware setting, where the objective is to minimize the *total acquisition cost* required to identify the best arm with high confidence. All statistical modeling, confidence sequences, and stopping rules remain unchanged from the time-minimization setting.

Cost Model. Each action $a \in \mathcal{A}$ incurs a known acquisition cost $c_a > 0$. For example, if all reward queries have cost C_{reward} and all dueling queries have cost C_{dueling}, then $c_a = C_{\text{reward}}$ for $a \in \mathcal{K}$ and $c_a = C_{\text{dueling}}$ for $a \in \mathcal{G}$. The cumulative cost up to time t is

$$B_t \triangleq \sum_{s=1}^{t} c_{a_s}.$$

The goal is to design a δ-correct strategy that minimizes B_τ.

Cost-normalized Design and Tracking. Let $\boldsymbol{c} \in \mathbb{R}^{|\mathcal{A}|}$ denote the vector of action costs. We introduce a cost-normalized design vector $\boldsymbol{p} \in \mathbb{R}^{\mathcal{A}}_{\geq 0}$ satisfying $\langle \boldsymbol{c}, \boldsymbol{p} \rangle = 1$, where p_a represents the sampling intensity per unit cost. The induced information matrix is $\boldsymbol{A}(\boldsymbol{p}, \boldsymbol{\theta})$. Given the current estimate $\hat{\boldsymbol{\theta}}_t$, the cost-aware minimax-optimal design is defined as

$$\boldsymbol{p}^\star(t) = \operatorname*{arg\,min}_{\boldsymbol{p} \geq 0:\, \langle \boldsymbol{c}, \boldsymbol{p} \rangle = 1} \max_{i \neq \hat{i}^\star_t} \|\boldsymbol{x}_{\hat{i}^\star_t, i}\|^2_{\boldsymbol{A}(\boldsymbol{p}, \hat{\boldsymbol{\theta}}_t)^{-1}}. \tag{6}$$

To implement the design in a round-based procedure, we normalize $\bar{\boldsymbol{p}}^\star(t) = \frac{\boldsymbol{p}^\star(t)}{\sum_{a \in \mathcal{A}} p^\star_a(t)}$ and apply the same tracking rule as in HyTS-GLB.

Definition 4.6 (Cost-aware characteristic time)**.** The cost-aware characteristic time is

$$T^\star_{\text{cost}}(\boldsymbol{\theta}^\star) \triangleq \inf_{\boldsymbol{p} \geq 0:\, \langle \boldsymbol{c}, \boldsymbol{p} \rangle = 1} \frac{\max_{i \neq i^\star} \|\boldsymbol{x}_{i^\star, i}\|^2_{\boldsymbol{A}(\boldsymbol{p}, \boldsymbol{\theta}^\star)^{-1}}}{\Delta^2_{\min}}.$$

Theorem 4.7. *Under the same assumptions as in Theorem 4.5. Then Cost-Aware HyTS-GLB's cost B_τ verifies*

$$\Pr\left(\limsup_{\delta \to 0} \frac{B_\tau}{\log(1/\delta)} \lesssim T^\star_{\text{cost}}(\boldsymbol{\theta}^\star) \right) = 1.$$

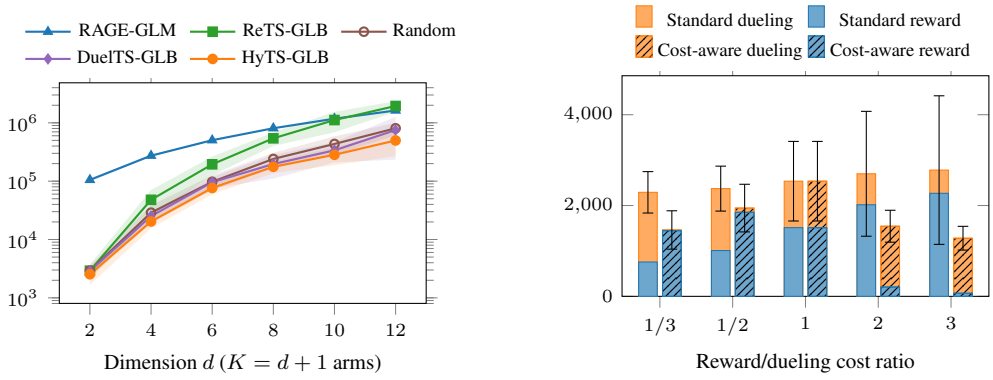
Compared with standard HyTS-GLB, which implicitly assumes uniform query costs, the cost-aware variant optimizes sampling in a cost-normalized design space. By minimizing the worst-case uncertainty per unit cost, the algorithm adaptively favors the more cost-efficient feedback modality while preserving the same confidence sequences and stopping rule. When acquisition costs differ, this can reduce total cost, while recovering standard HyTS-GLB as a special case under uniform costs.

5 Experiments

Our experiments examine two questions: whether hybrid feedback improves sample efficiency over single-modality baselines, and whether the cost-aware design adapts to heterogeneous acquisition costs. Due to the high cost of exact optimal allocation, we follow Jedra and Proutiere [2020] and compute designs with Frank–Wolfe [Jaggi, 2013]. All methods are evaluated on the same generated instances under the same logistic observation model and confidence level. We set $K = d + 1$ with $d \in \{2, 4, 6, 8, 10, 12\}$, $\delta = 0.05$, and $\boldsymbol{\theta}^\star = \frac{S-1}{\sqrt{d}} \cdot \mathbf{1}$ with $S = 5$. Arms are constructed so that $\boldsymbol{x}_1^\top \boldsymbol{\theta}^\star = 0.9\|\boldsymbol{\theta}^\star\|_2$, while for $i \geq 2$, $\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star = 0.8 \cdot \frac{K-i}{K-1}\|\boldsymbol{\theta}^\star\|_2$, with remaining coordinates random. We compare HyTS-GLB with RageGLM [Jun et al., 2021], ReTS-GLB, DuelTS-GLB, and a random hybrid baseline. We also tested GLGapE [Kazerouni and Wein, 2021], but it did not terminate within the simulation budget, consistent with observations in Jun et al. [2021]. Results report means over 30 runs on shared instances. As shown in Figure 1(a), HyTS-GLB consistently outperforms RageGLM, ReTS-GLB and DuelTS-GLB across all dimensions. This suggests that hybrid allocation is most useful when reward and dueling feedback reduce uncertainty along complementary directions, especially as the dimension grows. We further evaluate cost-aware allocation with $C_{\text{reward}} + C_{\text{dueling}} = 2$ and $\frac{C_{\text{reward}}}{C_{\text{dueling}}} \in \{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}$. Figure 1(b) reports the reward and dueling cost decomposition for $K = 3, d = 2$. The cost-aware variant uniformly reduces total cost relative to standard HyTS-GLB by shifting samples toward the cheaper modality; when $\frac{C_{\text{reward}}}{C_{\text{dueling}}} = 1$, the two coincide, matching our results in Section 4.2.

Beyond these main experiments, we provide additional evidence for the role of hybrid allocation in the appendix. We also give two concrete complexity comparisons: one where dueling feedback is more informative and one where reward feedback is more informative. These examples reinforce the main message that HyTS-GLB does not commit to a single feedback modality, but adapts its sampling allocation to the instance geometry and feedback utility. Code and raw results are available at GitHub[1].

[1] https://github.com/qirunzeng/HyTS-GLB



(a) Stopping queries versus dimension

(b) Acquisition cost by modality

Figure 1: Scalability and acquisition cost of hybrid-feedback learning. Results are averaged over 30 independent runs, with shaded regions and error bars indicating one standard deviation.

6 Concluding Remarks

We study fixed-confidence best-arm identification in generalized linear bandits with hybrid reward and dueling feedback. Our unified likelihood-based confidence sequence supports a geometry-aware Track-and-Stop algorithm with δ-correctness and instance-dependent almost-sure asymptotic stopping-time bounds as $\delta \to 0$. The framework allocates queries according to the information provided by each channel and extends to heterogeneous acquisition costs.

Several limitations remain. The two channels share a latent preference parameter; if they reflect different preferences, pooling their likelihoods may invalidate our guarantees. Computationally, repeated MLE and design updates over the $O(K^2)$ pairwise action space can make each query expensive. Pair screening, less frequent updates, and approximate design oracles may improve scalability, but their effects on stopping-time guarantees remain to be established. Our empirical evaluation is limited to synthetic logistic instances, and validation on real feedback data is still needed.

Future work includes closing the gap to fully instance-optimal guarantees for general GLMs, developing robust models with partially shared parameters, and extending the framework to time-varying costs, contextual or non-stationary environments, and richer preference signals.

Acknowledgement

The work of Jinhang Zuo is supported by the Early Career Scheme (No. CityU 21212025) of the University Grants Committee (UGC) and CityU 9610706. The work of Xuchuang Wang is supported in part by the RGC JRFS2526-4S06.

References

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

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Eugenio Clerico, Hamish Flynn, Gergely Neu, et al. Confidence sequences for generalized linear models via regret analysis. *arXiv preprint arXiv:2504.16555*, 2025.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

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

Qijia He, Minghan Wang, Xutong Liu, Zhiyong Wang, and Fang Kong. Learning across the gap: Hybrid multi-armed bandits with heterogeneous offline and online data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2024.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Marc Jourdan and Rémy Degenne. Choosing answers in epsilon-best-answer identification for linear bandits. In *International Conference on Machine Learning*, pages 10384–10430. PMLR, 2022.

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

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

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

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Johannes Kirschner, Andreas Krause, Michele Meziu, and Mojmir Mutny. Confidence estimation via sequential likelihood mixing. *arXiv preprint arXiv:2502.14689*, 2025.

Jungyhun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Eduardo Ochoa Rivera and Ambuj Tewari. Near optimal pure exploration in logistic bandits. *arXiv preprint arXiv:2410.20640*, 2024.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2021.

Yuming Shao and Zhixuan Fang. Linear streaming bandit: Regret minimization and fixed-budget epsilon-best arm identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20354–20361, 2025.

Marta Soare. *Sequential resource allocation in linear stochastic bandits*. PhD thesis, Université Lille 1-Sciences et Technologies, 2015.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27, 2014.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *IJCAI*, pages 5502–5510, 2018.

Xuchuang Wang, Qirun Zeng, Jinhang Zuo, Xutong Liu, Mohammad Hajiesmaili, John Lui, and Adam Wierman. Fusing reward and dueling feedback in stochastic bandits. *arXiv preprint arXiv:2504.15812*, 2025.

Shuhua Yang, Hui Yuan, Xiaoying Zhang, Mengdi Wang, Hong Zhang, and Huazheng Wang. Conversational dueling bandits in generalized linear models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3806–3817, 2024.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 241–248, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Xiaoying Zhang, Hong Xie, Hang Li, and John CS Lui. Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020*, pages 662–672, 2020.

Yurui Zou, Qirun Zeng, and Jinhang Zuo. Conversational generalized linear bandits with hybrid feedback. *arXiv preprint arXiv:2610.35262*, 2026.

Appendix

Contents

1 Introduction **1**

1.1 Our Contributions . . . 2

1.2 Related Work . . . 2

2 Problem Formulation **3**

3 Confidence Sequences for Hybrid GLB **5**

3.1 Maximum Likelihood Estimation . . . 5

3.2 Likelihood-Ratio Confidence Sequences . . . 5

3.3 Ellipsoidal Confidence Sequences under Self-Concordance . . . 5

4 Hybrid Track-and-Stop Algorithms **6**

4.1 HyTS-GLB Algorithm . . . 6

4.2 Cost-Aware HyTS-GLB . . . 8

5 Experiments **9**

6 Concluding Remarks **10**

A Proofs **14**

A.1 Proof of Lemma 3.1 . . . 14

A.2 Proof of Theorem 3.3 . . . 15

A.3 Proof of Theorem 4.5 . . . 16

A.4 Proof of Theorem 4.7 . . . 17

B KL Lower Bound and Gap-Ratio Comparison **17**

C Additional Experiments **18**

D Concrete Complexity Comparisons **19**

A Proofs

A.1 Proof of Lemma 3.1

Let $\mathcal{F}_t$ denote the history after observing Z_t. At round t, the learner selects a_t using an $\mathcal{F}_{t-1}$-measurable randomized policy, before observing Z_t. Write $m_t \triangleq m(a_t)$ and $\boldsymbol{x}_t \triangleq \boldsymbol{x}_{a_t}$. For each modality $m \in \{\mathrm{c}, \mathrm{d}\}$, let ν_m be a dominating measure for its conditional distributions. For any fixed $\boldsymbol{\theta} \in \Theta$, define

$$M_t(\boldsymbol{\theta}) \triangleq \exp\{\mathcal{L}_t(\boldsymbol{\theta}^\star) - \mathcal{L}_t(\boldsymbol{\theta})\} = \prod_{s=1}^{t-1} \frac{p_{m_s}(Z_s \mid \boldsymbol{x}_s^\top \boldsymbol{\theta})}{p_{m_s}(Z_s \mid \boldsymbol{x}_s^\top \boldsymbol{\theta}^\star)}.$$

The base-measure terms in Equation (1) cancel in the likelihood ratio. Although the density and even the observation space may depend on m_s, conditioning on the selected modality gives

$$\mathbb{E}_{\boldsymbol{\theta}^\star}[M_{t+1}(\boldsymbol{\theta}) \mid \mathcal{F}_{t-1}, a_t] = M_t(\boldsymbol{\theta}) \int \frac{p_{m_t}(z \mid \boldsymbol{x}_t^\top \boldsymbol{\theta})}{p_{m_t}(z \mid \boldsymbol{x}_t^\top \boldsymbol{\theta}^\star)} p_{m_t}(z \mid \boldsymbol{x}_t^\top \boldsymbol{\theta}^\star)\, \mathrm{d}\nu_{m_t}(z) = M_t(\boldsymbol{\theta}).$$

Taking conditional expectation once more over the learner's action randomization gives $\mathbb{E}[M_{t+1}(\boldsymbol{\theta}) \mid \mathcal{F}_{t-1}] = M_t(\boldsymbol{\theta})$. Hence $\{M_t(\boldsymbol{\theta})\}_{t\geq 1}$ is a nonnegative martingale with respect to $\{\mathcal{F}_{t-1}\}_{t\geq 1}$, with $M_1(\boldsymbol{\theta}) = 1$. This also shows directly why adaptive switching between the two modalities does not require a union bound.

We next apply the time-uniform PAC–Bayes argument. As in Lee et al. [2024], let $Q = \mathrm{Unif}(\Theta)$, where the uniform measure is taken on the affine hull of Θ. The volume scaling argument below then gives a KL term no larger than $d \log \frac{1}{c}$. By Tonelli's theorem, $\overline{M}_t \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim Q}[M_t(\boldsymbol{\theta})]$ is also a nonnegative martingale with $\overline{M}_1 = 1$. Ville's inequality therefore gives an event $\mathcal{V}_\delta$ with $\Pr(\mathcal{V}_\delta) \geq 1 - \delta$ on which, simultaneously for all $t \geq 1$,

$$\log \mathbb{E}_{\boldsymbol{\theta}\sim Q}[M_t(\boldsymbol{\theta})] \leq \log \tfrac{1}{\delta}.$$

For any $\mathcal{F}_{t-1}$-measurable posterior $P_t \ll Q$, the Donsker–Varadhan variational inequality and the preceding display imply

$$\mathcal{L}_t(\boldsymbol{\theta}^\star) - \mathbb{E}_{\boldsymbol{\theta}\sim P_t}[\mathcal{L}_t(\boldsymbol{\theta})] \leq \log \tfrac{1}{\delta} + D_{\mathrm{KL}}(P_t \| Q), \quad \text{for all } t$$

on $\mathcal{V}_\delta$.

Fix $c \in (0, 1]$ and choose

$$P_{t,c} = \mathrm{Unif}\Big((1-c)\hat{\boldsymbol{\theta}}_t + c\Theta\Big).$$

Convexity of Θ ensures that this posterior is supported on Θ, while scaling of Lebesgue volume yields

$$D_{\mathrm{KL}}(P_{t,c} \| Q) \leq d \log \tfrac{1}{c}.$$

Every $\boldsymbol{\theta}$ in its support can be written as $\boldsymbol{\theta} = (1-c)\hat{\boldsymbol{\theta}}_t + c\widetilde{\boldsymbol{\theta}}$ for some $\widetilde{\boldsymbol{\theta}} \in \Theta$, and consequently $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|_2 \leq 2Sc$. If L_t is any pathwise upper bound on the Lipschitz modulus of $\mathcal{L}_t$ over Θ, then

$$\mathbb{E}_{\boldsymbol{\theta}\sim P_{t,c}}[\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t)] \leq 2SL_t c.$$

Combining these inequalities, and then minimizing over $c \in (0, 1]$, gives

$$\mathcal{L}_t(\boldsymbol{\theta}^\star) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \leq \log \tfrac{1}{\delta} + \inf_{c\in(0,1]} \left\{ d \log \tfrac{1}{c} + 2SL_t c \right\}$$

simultaneously for all t on $\mathcal{V}_\delta$. Choosing $c = \min\{1, \frac{d}{2SL_t}\}$ gives the final upper bound in Equation (3) and proves Lemma 3.1.

Following the Lipschitz-bound argument of Lee et al. [2024], convexity of Θ and the mean-value and triangle inequalities give, on every sample path,

$$\sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta \\ \boldsymbol{\theta}\neq\boldsymbol{\theta}'}} \frac{|\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\boldsymbol{\theta}')|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2} \leq \sum_{s=1}^{t-1} \frac{\rho_{a_s}}{\zeta_{m(a_s)}(\varphi_{m(a_s)})} \left(|Z_s| + \sup_{\boldsymbol{\vartheta}\in\Theta} \left|\mu_{m(a_s)}(\boldsymbol{x}_{a_s}^\top \boldsymbol{\vartheta})\right| \right).$$

The inner supremum is finite by compactness and continuity, so the right-hand side is an admissible choice of L_t almost surely. Consistent with the linear-growth bounds in Lee et al. [2024, Table 1 and Appendix C], this choice also satisfies $L_t = O(t)$ almost surely: under our assumptions, the observations have uniformly bounded conditional second moments, and the martingale strong law gives $\sum_{s=1}^{t-1} |Z_s| = O(t)$ almost surely. The feature-norm bounds 1 for reward feedback and 2 for dueling feedback affect only constants.

A.2 Proof of Theorem 3.3

We follow the proof of Lee et al. [2024, Theorem 3.2].

Define

$$\tilde{G}_t^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \triangleq \frac{1}{\zeta_c(\varphi_c)} \sum_{s \in \mathcal{C}_t} \tilde{\alpha}_s^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \boldsymbol{x}_{i_s} \boldsymbol{x}_{i_s}^\top,$$
$$\tilde{G}_t^{\mathrm{d}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \triangleq \frac{1}{\zeta_d(\varphi_d)} \sum_{s \in \mathcal{D}_t} \tilde{\alpha}_s^{\mathrm{d}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \boldsymbol{x}_{a_s} \boldsymbol{x}_{a_s}^\top,$$

where

$$\tilde{\alpha}_s^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \triangleq \int_0^1 (1-v) \mu_{\mathrm{c}}'(\langle \boldsymbol{x}_{a_s}, \boldsymbol{\theta} + v(\boldsymbol{\nu} - \boldsymbol{\theta}) \rangle) \mathrm{d}v,$$
$$\tilde{\alpha}_s^{\mathrm{d}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \triangleq \int_0^1 (1-v) \mu_{\mathrm{d}}'(\langle \boldsymbol{x}_{a_s}, \boldsymbol{\theta} + v(\boldsymbol{\nu} - \boldsymbol{\theta}) \rangle) \mathrm{d}v.$$

Lemma A.1 (Lee et al., 2024, Lemma D.1)**.** *Let $\mu_{\mathrm{c}}, \mu_{\mathrm{d}}$ be increasing ($\mu_{\mathrm{c}}', \mu_{\mathrm{d}}' \geq 0$) and self-concordant with constants $M_{\mathrm{c}}, M_{\mathrm{d}}$. Let $\mathcal{Z}^{\mathrm{c}}, \mathcal{Z}^{\mathrm{d}} \subset \mathbb{R}$ be bounded. Then, for any $z_1^{\mathrm{c}}, z_2^{\mathrm{c}} \in \mathcal{Z}^{\mathrm{c}}$ and $z_1^{\mathrm{d}}, z_2^{\mathrm{d}} \in \mathcal{Z}^{\mathrm{d}}$,*

$$\int_0^1 (1-v) \mu_{\mathrm{c}}'(z_1^{\mathrm{c}} + v(z_2^{\mathrm{c}} - z_1^{\mathrm{c}})) \mathrm{d}v \geq \frac{\mu_{\mathrm{c}}'(z_1^{\mathrm{c}})}{2 + M_{\mathrm{c}} |z_1^{\mathrm{c}} - z_2^{\mathrm{c}}|},$$
$$\int_0^1 (1-v) \mu_{\mathrm{d}}'(z_1^{\mathrm{d}} + v(z_2^{\mathrm{d}} - z_1^{\mathrm{d}})) \mathrm{d}v \geq \frac{\mu_{\mathrm{d}}'(z_1^{\mathrm{d}})}{2 + M_{\mathrm{d}} |z_1^{\mathrm{d}} - z_2^{\mathrm{d}}|}.$$

Consequently, since $|z_1 - z_2| \leq \|\boldsymbol{x}_i^\top\|_2 \|\boldsymbol{\nu} - \boldsymbol{\theta}\|_2 \leq 2S$ for reward observations and $|z_1 - z_2| \leq \|\boldsymbol{x}_{j,k}^\top\|_2 \|\boldsymbol{\nu} - \boldsymbol{\theta}\|_2 \leq 4S$ for dueling observations,

$$\tilde{G}_t^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \succeq \frac{\nabla^2 \mathcal{L}_t^{\mathrm{c}}(\boldsymbol{\theta})}{2(1 + S M_{\mathrm{c}})},$$
$$\tilde{G}_t^{\mathrm{d}}(\boldsymbol{\theta}, \boldsymbol{\nu}) \succeq \frac{\nabla^2 \mathcal{L}_t^{\mathrm{d}}(\boldsymbol{\theta})}{2(1 + 2S M_{\mathrm{d}})}.$$

Let

$$\mathcal{L}_t^{\mathrm{c}} \triangleq \sum_{s \in \mathcal{C}_t} (b_{\mathrm{c}}(\boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}) - R_s \boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}) / \zeta_{\mathrm{c}}(\varphi_{\mathrm{c}}),$$
$$\mathcal{L}_t^{\mathrm{d}} \triangleq \sum_{s \in \mathcal{D}_t} (b_{\mathrm{d}}(\boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}) - Y_s \boldsymbol{x}_{a_s}^\top \boldsymbol{\theta}) / \zeta_{\mathrm{d}}(\varphi_{\mathrm{d}})$$

denote the cumulative loss incurred by reward observation and dueling observation separately. By Taylor's theorem with integral remainder,

$$\mathcal{L}_t^{\mathrm{c}}(\boldsymbol{\theta}) - \mathcal{L}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t) = \langle \nabla \mathcal{L}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t \rangle + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})},$$
$$\mathcal{L}_t^{\mathrm{d}}(\boldsymbol{\theta}) - \mathcal{L}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t) = \langle \nabla \mathcal{L}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t \rangle + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})}.$$

Summing the two equalities yields

$$\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) = \langle \nabla \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t \rangle + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})} + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})}.$$

Since $\hat{\boldsymbol{\theta}}_t \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_t(\boldsymbol{\theta})$ and Θ is convex, the first-order optimality condition implies that for any $\boldsymbol{\theta} \in \Theta$,

$$\langle \nabla \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t \rangle \geq 0.$$

Therefore,

$$\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \geq \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})} + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\tilde{G}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})}.$$

Lemma A.2. *For $t \geq 1$, consider a hybrid cumulative loss*

$$\mathcal{L}_t(\boldsymbol{\theta}) = \mathcal{L}_t^{\mathrm{c}}(\boldsymbol{\theta}) + \mathcal{L}_t^{\mathrm{d}}(\boldsymbol{\theta}),$$

with $\boldsymbol{H}_t^{\mathrm{c}}(\boldsymbol{\theta}) \triangleq \nabla^2 \mathcal{L}_t^{\mathrm{c}}(\boldsymbol{\theta})$ and $\boldsymbol{H}_t^{\mathrm{d}}(\boldsymbol{\theta}) \triangleq \nabla^2 \mathcal{L}_t^{\mathrm{d}}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_t \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_t(\boldsymbol{\theta})$. Then for all $\boldsymbol{\theta} \in \Theta$,

$$\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \geq \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{A}_t}.$$

Proof of Lemma A.2. Combining the above inequalities yields

$$\begin{aligned}
\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) &\geq \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{H}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t)}}{2(1 + SM_{\mathrm{c}})} + \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{H}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t)}}{2(1 + 2SM_{\mathrm{d}})} \\
&= (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t)^\top \left(\frac{\boldsymbol{H}_t^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_t)}{2(1 + SM_{\mathrm{c}})} + \frac{\boldsymbol{H}_t^{\mathrm{d}}(\hat{\boldsymbol{\theta}}_t)}{2(1 + 2SM_{\mathrm{d}})} \right) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t) \\
&= \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{A}_t},
\end{aligned}$$

as claimed. □

Proof of Theorem 3.3. By Lemma A.2,

$$\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}_t\|^2_{\boldsymbol{A}_t} \leq \mathcal{L}_t(\boldsymbol{\theta}^\star) - \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \leq \beta_t(\delta).$$

□

A.3 Proof of Theorem 4.5

Write $n_t = |\mathcal{T}_t|$ for the number of tracking rounds and consider a common infinite sampling trajectory, with the sampling rule and tie-breaking independent of δ. For a spanning exploration distribution π_0 (e.g., uniform on $\mathcal{A}_0$), the strong law for exploration counts and Assumption 4.1 give $\frac{n_t}{t} \to 1$ and $\lambda_{\min}(\boldsymbol{A}_t) \gtrsim t^{1-\alpha}$ a.s. Applying Theorem 3.3 with $\delta_t = t^{-2}$ and using Borel–Cantelli and $L_t = O(t)$ a.s. yields $\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\|_2^2 = O(\frac{\log t}{t^{1-\alpha}}) \to 0$ a.s.

Lemma A.3 (Adapted from Lemmas 1 and 2 of Jedra and Proutiere [2020])**.** *Under the assumptions of Theorem 4.5, ϕ is jointly continuous wherever the best arm is unchanged and $\boldsymbol{A}(\boldsymbol{w}, \boldsymbol{\theta}) \succ 0$. Define*

$$\phi^\star(\boldsymbol{\theta}) \triangleq \min_{\boldsymbol{w}\in\Delta_{\mathcal{A}}} \phi(\boldsymbol{\theta}, \boldsymbol{w}), \qquad \mathcal{W}^\star(\boldsymbol{\theta}) \triangleq \arg\min_{\boldsymbol{w}\in\Delta_{\mathcal{A}}} \phi(\boldsymbol{\theta}, \boldsymbol{w}).$$

Then $\phi^\star$ is continuous at $\boldsymbol{\theta}^\star$, and $\mathcal{W}^\star(\boldsymbol{\theta}^\star)$ is nonempty, compact, and convex, with every optimal design positive definite. The optimal-design correspondence is upper hemicontinuous at $\boldsymbol{\theta}^\star$. The same properties hold on the cost-normalized domain $\{\boldsymbol{p} \geq 0 : \langle \boldsymbol{c}, \boldsymbol{p} \rangle = 1\}$.

Proof. Following the compactness and maximum-theorem argument of Jedra and Proutiere [2020, Lemmas 1 and 2], it suffices to exclude singular limits of near-optimal designs. The dueling spanning assumption implies $\boldsymbol{G} \triangleq \sum_{i\neq i^\star} \boldsymbol{x}_{i^\star,i} \boldsymbol{x}_{i^\star,i}^\top \succ 0$, and, for $\boldsymbol{A} \succ 0$,

$$\max_{i\neq i^\star} \boldsymbol{x}_{i^\star,i}^\top \boldsymbol{A}^{-1} \boldsymbol{x}_{i^\star,i} \geq \frac{\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{G})}{K-1} \geq \frac{\lambda_{\min}(\boldsymbol{G})}{(K-1)\lambda_{\min}(\boldsymbol{A})}.$$

Since the best arm is locally unchanged and a fixed spanning design has a finite continuous objective, this bound confines minimizing sequences near $\boldsymbol{\theta}^\star$ to a common compact subset of positive-definite designs. Continuity, convexity in the design weights, and the maximum theorem then give the claims, on either compact design domain. □

Under the design regularity in Lemma A.3, consistency makes the target designs approach $\mathcal{W}^\star(\boldsymbol{\theta}^\star)$. The deterministic set-tracking argument of Jedra and Proutiere [2020, Lemma 6 and Proposition 1], applied to the tracking subsequence, therefore gives

$$\operatorname{dist}\left(\frac{\boldsymbol{N}^{\mathcal{T}}(t)}{n_t}, \mathcal{W}^\star(\boldsymbol{\theta}^\star)\right) \longrightarrow 0 \qquad \text{a.s.}$$

By continuity of the information weights and inverse matrices near this compact set of positive-definite optimal designs, $\boldsymbol{A}_t \succeq \boldsymbol{A}_t^{\mathcal{T}}$ and $\frac{n_t}{t} \to 1$ imply

$$\limsup_{t\to\infty} t \max_{i\neq i^\star} \|\boldsymbol{x}_{i^\star,i}\|^2_{\boldsymbol{A}_t^{-1}} \le \Delta_{\min}^2 T^\star(\boldsymbol{\theta}^\star) \qquad \text{a.s.} \tag{7}$$

Proof of Theorem 4.5. Fix a trajectory on which the preceding almost-sure conclusions hold. Consistency and the unique best arm imply that, after a finite random time independent of δ, $\hat{i}_t^\star = i^\star$ and $\boldsymbol{x}_{i^\star,i}^\top \hat{\boldsymbol{\theta}}_t > \frac{\Delta_i}{2}$ for every $i \neq i^\star$. Thus the stopping rule holds whenever

$$\beta_t(\delta) \max_{i\neq i^\star} \|\boldsymbol{x}_{i^\star,i}\|^2_{\boldsymbol{A}_t^{-1}} < \frac{\Delta_{\min}^2}{4}.$$

For any fixed $C > 4$, set $t_\delta = \lceil CT^\star(\boldsymbol{\theta}^\star)\log(1/\delta)\rceil$. Equation (3) and $L_t = O(t)$ imply $\frac{\beta_{t_\delta}(\delta)}{\log(1/\delta)} \to 1$; hence Equation (7) makes the preceding stopping condition hold at t_δ for all sufficiently small δ. Consequently, $\tau \le t_\delta$ eventually and

$$\limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \le 4T^\star(\boldsymbol{\theta}^\star) \qquad \text{a.s.},$$

which proves the claimed bound. □

A.4 Proof of Theorem 4.7

Proof. Apply the same set-tracking argument to the normalized cost-optimal designs, using Lemma A.3 on $\{\boldsymbol{p} \ge 0 : \langle \boldsymbol{c}, \boldsymbol{p}\rangle = 1\}$. Writing $P_s = \sum_a p_a^\star(s-1)$, the cost-normalized cumulative target is

$$\bar{\boldsymbol{p}}_t = \frac{\sum_{s\in\mathcal{T}_t} \boldsymbol{p}^\star(s-1)/P_s}{\sum_{s\in\mathcal{T}_t} 1/P_s},$$

a convex combination of the target designs, which therefore approaches the convex cost-optimal set. Since $\frac{1}{P_s} \in [c_{\min}, c_{\max}]$, bounded tracking errors and negligible forced exploration give $\frac{B_t}{\sum_{s\in\mathcal{T}_t}(1/P_s)} \to 1$ a.s., so the empirical cost-normalized design has the same limiting optimal set. Thus the analogue of Equation (7) is

$$\limsup_{t\to\infty} B_t \max_{i\neq i^\star} \|\boldsymbol{x}_{i^\star,i}\|^2_{\boldsymbol{A}_t^{-1}} \le \Delta_{\min}^2 T^\star_{\text{cost}}(\boldsymbol{\theta}^\star) \qquad \text{a.s.}$$

For $C > 4$, let s_δ be the first time the cumulative cost exceeds $b_\delta = CT^\star_{\text{cost}}(\boldsymbol{\theta}^\star)\log(1/\delta)$. The bounded overshoot gives $b_\delta < B_{s_\delta} \le b_\delta + c_{\max}$ and $s_\delta = O(\log(1/\delta))$, so $\frac{\beta_{s_\delta}(\delta)}{\log(1/\delta)} \to 1$ a.s. The same pathwise stopping condition therefore yields $\tau \le s_\delta$ for all sufficiently small δ, and hence

$$\limsup_{\delta\to 0} \frac{B_\tau}{\log(1/\delta)} \le 4T^\star_{\text{cost}}(\boldsymbol{\theta}^\star) \qquad \text{a.s.},$$

as claimed. □

B KL Lower Bound and Gap-Ratio Comparison

Let $\mathcal{M} \subseteq \Theta$ be the set of parameters with a unique best arm. Define

$$\text{Alt}(\boldsymbol{\theta}^\star) = \{\boldsymbol{\lambda} \in \mathcal{M} : i^\star(\boldsymbol{\lambda}) \neq i^\star(\boldsymbol{\theta}^\star)\}.$$

Write $d_a(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for the KL divergence from $p_{m(a)}(\cdot \mid \boldsymbol{x}_a^\top \boldsymbol{\theta})$ to $p_{m(a)}(\cdot \mid \boldsymbol{x}_a^\top \boldsymbol{\lambda})$.

Theorem B.1 (Information-theoretic lower bound for hybrid GLM BAI). *For $\boldsymbol{\theta}^\star \in \mathcal{M}$, define*

$$T_{\mathrm{LB}}(\boldsymbol{\theta}^\star) = \left[\sup_{\boldsymbol{w} \in \Delta_{\mathcal{A}}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\theta}^\star)} \sum_{a \in \mathcal{A}} w_a d_a(\boldsymbol{\theta}^\star, \boldsymbol{\lambda}) \right]^{-1}.$$

For $\delta \in (0, 1/2)$, any δ-correct strategy on $\mathcal{M}$ satisfies

$$\mathbb{E}_{\boldsymbol{\theta}^\star}[\tau_\delta] \geq T_{\mathrm{LB}}(\boldsymbol{\theta}^\star)\, \mathrm{kl}(1-\delta, \delta),$$

where kl *is binary relative entropy, with* $\inf \varnothing = +\infty$, $1/(+\infty) = 0$, *and* $1/0 = +\infty$.

Proof. This follows directly from the standard change-of-measure lower bound [Kaufmann et al., 2016, Degenne and Koolen, 2019], applied to the joint action space $\mathcal{A}$. □

Linear–Gaussian comparison. Consider the homoscedastic linear–Gaussian benchmark with distinct arm features and parameter space $\mathbb{R}^d$: $Z \mid a \sim \mathcal{N}(\boldsymbol{x}_a^\top \boldsymbol{\theta}, \sigma^2)$, including real-valued utility-difference feedback for pair queries. Here the existing design matrix reduces to $\boldsymbol{A}(\boldsymbol{w}) = (2\sigma^2)^{-1} \sum_{a \in \mathcal{A}} w_a \boldsymbol{x}_a \boldsymbol{x}_a^\top$. Writing $\Delta_i = \boldsymbol{x}_{i^\star,i}^\top \boldsymbol{\theta}^\star$, the exact quadratic KL divergence gives [Jedra and Proutiere, 2020]

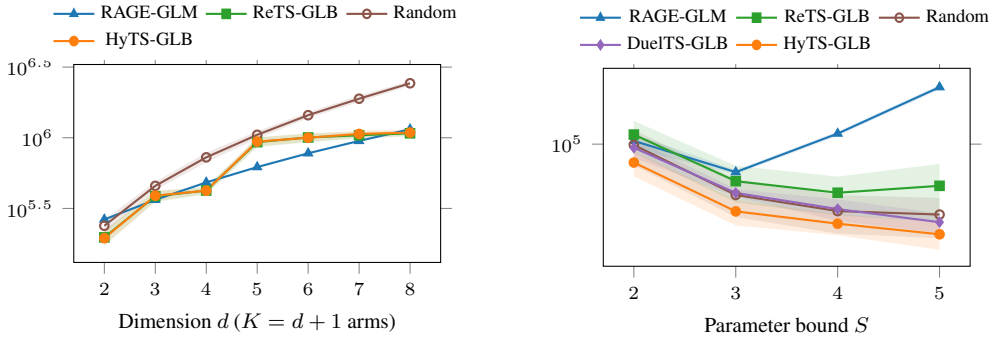
$$T_{\mathrm{LB}}(\boldsymbol{\theta}^\star) = \inf_{\boldsymbol{w} \in \Delta_{\mathcal{A}}} \max_{i \neq i^\star} \frac{\|\boldsymbol{x}_{i^\star,i}\|^2_{\boldsymbol{A}(\boldsymbol{w})^{-1}}}{\Delta_i^2}.$$

Compared with $T^\star$ in Definition 4.4, the only change is the use of individual gaps instead of the common $\Delta_{\min}$. Thus, with $\Delta_{\max} = \max_{i \neq i^\star} \Delta_i$, pointwise comparison of the design objectives yields

$$1 \leq \frac{T^\star(\boldsymbol{\theta}^\star)}{T_{\mathrm{LB}}(\boldsymbol{\theta}^\star)} \leq \left(\frac{\Delta_{\max}}{\Delta_{\min}} \right)^2.$$

Equal suboptimal gaps therefore make the two characteristic times coincide in this benchmark.

C Additional Experiments



(a) Basis-plus-rotated arms, $S = 5$

(b) Random arms, $(K, d) = (5, 4)$

Figure 2: Mean stopping queries over 30 runs, with shading showing one standard deviation.

We present two additional synthetic experiments that complement the random-arm results in Section 5. All experiments use the same logistic observation model, confidence level, implementation of the stopping rule, and Frank–Wolfe design solver as in the main text.

In Figure 2(a), the arm set is

$$\mathcal{X} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d, \cos(0.1)\boldsymbol{e}_1 + \sin(0.1)\boldsymbol{e}_2\},$$

with $\boldsymbol{\theta} = (S-1)\boldsymbol{e}_1$, $d \in \{2, 3, 4, 5, 6, 7, 8\}$ and $S = 5$. This instance creates a hard comparison between the best basis arm and a nearby rotated arm. The relevant separating direction has a small component along $\boldsymbol{e}_1 - \cos(0.1)\boldsymbol{e}_1$ and a larger component along $\boldsymbol{e}_2$, so reward queries on basis arms

can be highly informative even when the final decision is pairwise. We do not report the dueling-only variant because it required more than $20{,}000{,}000$ samples in this construction, exceeding our computational limit. HyTS-GLB has mean stopping times close to those of ReTS-GLB and below the random hybrid baseline across the displayed dimensions. Its mean stopping time is lower than RAGE-GLM at $d = 2, 4, 8$.

In Figure 2(b), we return to the random-arm construction from the main text, fix $(K, d) = (5, 4)$, and vary $S \in \{2, 3, 4, 5\}$. Changing S changes the magnitude of the linear scores and therefore the local curvature of the logistic model. HyTS-GLB achieves the smallest mean stopping time for all tested values of S, with the largest relative gains at larger S.

D Concrete Complexity Comparisons

In this section, we present two concrete examples showing that either reward or dueling feedback can be more informative, depending on the arm geometry. These examples motivate adaptively allocating queries between the two feedback channels. Both channels are logistic with unit dispersion. Let $\Theta = \mathcal{B}^2(S)$ and $\boldsymbol{\theta}^\star = S\boldsymbol{e}_1$, with $S \geq 4$, $B_{\mathrm{c}} = 2(1 + S)$, and $B_{\mathrm{d}} = 2(1 + 2S)$. We compare the characteristic times $T_{\mathrm{R}}, T_{\mathrm{D}}$ in Definition 4.4 as $S \to \infty$.

Case 1: Dueling Feedback is More Informative.

Let

$$\boldsymbol{x}_1 = \tfrac{1}{2}\boldsymbol{e}_1, \qquad \boldsymbol{x}_2 = (\tfrac{1}{2} - \tfrac{1}{S})\boldsymbol{e}_1 + \tfrac{1}{2}\boldsymbol{e}_2, \qquad \boldsymbol{x}_3 = (\tfrac{1}{2} - \tfrac{1}{S})\boldsymbol{e}_1 - \tfrac{1}{2}\boldsymbol{e}_2.$$

Arm 1 is optimal, and both linear-score gaps equal 1.

Reward-only. All reward curvatures are $\Theta(e^{-S/2})$. Equally querying arms 2 and 3 attains comparison variance $O(B_{\mathrm{c}}e^{S/2})$; bounded second-coordinate information makes this order necessary. Thus

$$T_{\mathrm{R}} = \Theta\Big(B_{\mathrm{c}}e^{S/2}\Big).$$

Dueling-only. Both best-arm comparisons have predictor 1 and constant curvature. Equal sampling attains comparison variance of order B_{d}, with a matching lower bound from bounded second-coordinate information. Hence

$$T_{\mathrm{D}} = \Theta(B_{\mathrm{d}}), \qquad \frac{T_{\mathrm{R}}}{T_{\mathrm{D}}} = \Theta\left(\frac{B_{\mathrm{c}}}{B_{\mathrm{d}}}e^{S/2}\right).$$

Case 2: Reward Feedback is More Informative.

Let

$$\boldsymbol{x}_1 = \tfrac{1}{2}\boldsymbol{e}_1, \qquad \boldsymbol{x}_2 = -\tfrac{1}{2}\boldsymbol{e}_1 + \tfrac{1}{2}\boldsymbol{e}_2, \qquad \boldsymbol{x}_3 = -\tfrac{1}{2}\boldsymbol{e}_1 - \tfrac{1}{2}\boldsymbol{e}_2.$$

Arm 1 remains optimal, and both linear-score gaps equal S.

Reward-only. All reward curvatures equal $\sigma'(S/2) = \Theta(e^{-S/2})$. Equally querying arms 2 and 3 is optimal and estimates both coordinates, yielding

$$T_{\mathrm{R}} = \Theta\left(\frac{B_{\mathrm{c}}e^{S/2}}{S^2}\right).$$

Dueling-only. Only pairs $(1, 2)$ and $(1, 3)$ provide first-coordinate information, with curvature $\sigma'(S) = \Theta(e^{-S})$. This limits every design; equally sampling the two pairs attains the same order. Hence

$$T_{\mathrm{D}} = \Theta\left(\frac{B_{\mathrm{d}}e^{S}}{S^2}\right), \qquad \frac{T_{\mathrm{R}}}{T_{\mathrm{D}}} = \Theta\left(\frac{B_{\mathrm{c}}}{B_{\mathrm{d}}}e^{-S/2}\right).$$

Since $B_{\mathrm{c}} \asymp B_{\mathrm{d}}$, the two geometries favor opposite modalities.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper investigates best-arm identification in generalized linear bandits with hybrid feedback, develops time-uniform confidence sequences, and proposes a hybrid Track-and-Stop framework. We establish δ-correctness, provide high-probability stopping-time upper bounds, and validate the method empirically.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Section 6.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We propose the time-uniform confidence sequences for hybrid Generalized linear bandits and the hybrid Track-and-Stop algorithm. The main paper presents our principal theoretical results, namely Theorems 3.3, 4.5 and 4.7. Detailed proofs are deferred to the appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive pseudocode and specify all corresponding hyperparameters to ensure reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have released our code and raw data on GitHub.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We discussed main experimental settings in Section 5. We also add additional experiments and their setting in the appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, our figures in Section 5 report error bars. The results are consistent with the theory.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Given the lightweight nature of our algorithms, they are executable on most machines without specialized hardware, so detailed CPU or GPU specifications are unnecessary.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is a theoretical and empirical study of bandit algorithms; it does not involve human subjects, private data, or deployment of a harmful system, and we follow the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: As the work is primarily theoretical and methodological, it does not have immediate societal implications. However, future applications of the proposed methods may warrant further consideration of their broader impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release models or datasets with high misuse risk; the experiments use synthetic simulations.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We only use synthetic data in the experiment.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, models, or software assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The experiments do not involve crowdsourcing or human-subject studies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve human subjects, so IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were only used for writing.