SATURATION-INSENSITIVE DUELING BANDITS WITH GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study contextual dueling bandits with general function approximation under the Bradley–Terry–Luce (BTL) preference model. A key challenge in this setting is the *saturation* of the preference model: when the current reward model can already distinguish two actions with high confidence, the resulting preference feedback becomes weakly informative, making it difficult to further improve reward estimation. Consequently, existing sample-complexity analyses often depend on the inverse-derivative factor $1/\sigma'[\Delta_{r^*}]$ which can be prohibitively large when the link function σ saturates for large reward gaps Δ_{r^*}. To address this issue, we introduce `SI-CDB`, an algorithm that selects opponent arms using a carefully designed heuristic for arm selection. This design enables saturation-insensitive reward learning and recovers the near-optimal dependence for linear reward classes, eliminating the unfavorable $1/\sigma'(\cdot)$ factor. The core of our analysis is a localized Eluder dimension framework tailored to dueling bandits with general function approximation. Our theoretical results also explain why two-arm regret analysis is crucial for improving single-arm performance in dueling bandits.

1 INTRODUCTION

Dueling bandits (Yue et al., 2012; Saha, 2021; Bengs et al., 2022) model online learning from binary preference feedback: in each round, the learner selects two actions and observes which one is preferred. This framework underlies the theory of reinforcement learning from human feedback (RLHF; Zhu et al. 2023; Ji et al. 2023) and of online preference-based fine-tuning of large language models (LLMs; Ouyang et al. 2022; Xiong et al. 2024). The most widely used preference model is the Bradley–Terry–Luce (BTL) model (Hunter, 2004; Luce, 1959), under which action a is preferred to action b in context x with probability

$$\mathbb{P}(a \succ b \mid x) = \frac{\exp(r^*(x,a))}{\exp(r^*(x,a)) + \exp(r^*(x,b))} = \sigma\big(r^*(x,a) - r^*(x,b)\big), \tag{1.1}$$

where r^* is an unknown reward function and $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid *link function*.

In these settings, simply collecting more pairwise comparisons is not enough, as the comparisons must also be informative for the learning objective. Recent work on active preference learning for LLM alignment and RLHF supports this point, showing that adaptive query selection can improve label efficiency (Muldrew et al., 2024; Das et al., 2025; Melo et al., 2024; Ji et al., 2024). We study one basic mechanism behind uninformative comparisons under the BTL model: saturation of the sigmoid link. To illustrate this issue, consider a context x and two actions (a, b) such that both the ground-truth reward r^* and a current reward estimate r_t clearly distinguish between them; i.e., $r^*(x,a) \gg r^*(x,b)$ and $r_t(x,a) \gg r_t(x,b)$. In this regime, even if r_t is still substantially different from r^*, the two induced preference models may be nearly indistinguishable on the comparison pair (a,b), i.e., $\mathbb{P}^*(a \succ b|x) \approx \mathbb{P}_{r_t}(a \succ b|x) \approx 1$. As a result, observing preferences between a and b provides little information for improving r_t. This saturation effect is reflected in the inverse-curvature dependence that appears in existing analyses (Xie et al., 2025; Xiong et al., 2024). Under the BTL model, the derivative of the sigmoid link with respect to the reward gap Δ is $\sigma'(\Delta) = \sigma(\Delta)(1-\sigma(\Delta))$. When the comparison is saturated, this derivative is exponentially small in the magnitude of the reward gap. Consequently, analyses that rely on a uniform lower bound on the link curvature incur factors such as $\sup_{\Delta \in [-2B, 2B]} 1/\sigma'(\Delta) = \exp(\Theta(B))$, which can be prohibitively large in the reward scale

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.

B. We call a guarantee saturation-sensitive if its leading dependence scales with this worst-case inverse-curvature factor, and saturation-insensitive if the leading term avoids such dependence.

Such global inverse-curvature dependence appears in existing analyses of contextual dueling bandits (Saha, 2021; Bengs et al., 2022; Di et al., 2024a; Li et al., 2024). This dependence is not unavoidable, as prior work has obtained sharper regret bounds by exploiting local curvature in logistic bandits (Faury et al., 2020; Abeille et al., 2021), and a similar method has been developed for linear contextual dueling bandits (Di et al., 2025). However, these methods rely on a known linear parameterization to construct curvature-adaptive confidence sets and control exploration (see Appendix F.1 for a detailed discussion). Their analyses therefore do not directly extend to general reward classes, leaving open how to obtain saturation-insensitive guarantees beyond the linear setting.

To obtain widely applicable guarantees, we study it in the most general formulation of contextual dueling bandits considered here, allowing arbitrary realizable reward-function classes, and ask:

Is it possible to get rid of the reward saturation phenomenon in dueling bandits with general function approximation?

In this paper, we answer this question affirmatively by presenting `SI-CDB`. The core of the algorithm is to mitigate saturation through online exploration: after choosing a greedy target arm, it uses the current confidence set to select an optimistic comparator arm. This asymmetric arm-selection rule leads to a dueling-specific regret decomposition, which we combine with a localized eluder analysis over the induced excess-loss class.

We summarize the contributions of this paper as follows:

1. We propose `SI-CDB`, a saturation-insensitive algorithm for contextual dueling bandits with general function approximation. The algorithm selects actions through two asymmetric function-action pairs. It first forms a target pair (r_1, a) by choosing an empirical risk minimizer r_1 and playing its greedy action a. It then constructs, around r_1 in empirical loss, a confidence set of plausible reward functions. From this confidence set, the algorithm forms a comparator pair (r_2, b) by jointly choosing a reward function and a comparator action. Given the target pair (r_1, a) and the context x, this comparator pair is selected by maximizing $J(b, r) = 2r(x, b) - r_1(x, b) - 2r(x, a)$. The coefficients in this objective are chosen so that, at each round t, the two-arm regret decomposes into played-pair reward-gap errors of the form $|\Delta_{r_j}(x, a, b) - \Delta_{r^*}(x, a, b)|$ for $j \in \{1, 2\}$. This decomposition is the key bridge to the localized excess-loss analysis.
2. We derive a regret bound for `SI-CDB` based on the localized analysis. We partition all steps into two parts based on whether the reward error terms $|\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)|$ are small. Our localized analysis shows that (i) the total regret over steps with small reward errors, which is the main contributor to the cumulative regret bound, depends only polynomially on B, and (ii) the number of steps with large reward errors, despite the exponential dependence on B, depends only logarithmically on the number of steps T and is therefore a lower-order term when $T \to \infty$. Compared with works on both the linear reward function (Saha, 2021; Bengs et al., 2022) and the general reward function class (Li et al., 2024), the regret of `SI-CDB` does not contain terms that scale with both $\mathrm{poly}(T)$ and $e^{\Theta(B)}$.

Notations. We use normal-face letters to denote scalars, lower-case bold-face letters to denote vectors, and upper-case bold-face letters to denote matrices. Let $\mathbb{1}[\cdot]$ denote the indicator function. For a positive integer N, let $[N]$ denote the set $\{1, 2, \ldots, N\}$. We use standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, with $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ hiding logarithmic factors. For $p \in [0, 1]$, we use $\mathrm{Ber}(p)$ to denote the Bernoulli distribution with mean p. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ denotes its ℓ_2-norm.

2 Related Work

Dueling bandit and contextual preference learning. Dueling bandits extend the standard bandit framework by replacing direct scalar reward observations with pairwise preference feedback between selected actions, which is a setting initiated by Yue et al. (2012). As surveyed by Bengs et al. (2021), prior work on dueling bandits has studied a range of optimality notions, including Condorcet winners (Zoghi et al., 2014; Komiyama et al., 2015), Copeland winners (Zoghi et al., 2015; Wu & Liu,

2016; Komiyama et al., 2016), Borda winners (Jamieson et al., 2015; Falahatgar et al., 2017; Heckel et al., 2018; Saha et al., 2021), and von Neumann winners (Ramamohan et al., 2016; Dudík et al., 2015; Balsubramani et al., 2016). Contextual dueling bandits incorporate side information and typically require structure on the preference model. Saha (2021) studied stochastic contextual preference bandits with a logistic link function. More recent works provide variance-aware guarantees (Di et al., 2024a), corruption-robust guarantees (Cheng et al., 2025; Di et al., 2025), oracle-efficient algorithms under realizability (Saha & Krishnamurthy, 2022), active-query guarantees for general function class (Sekhari et al., 2023), Thompson sampling-style algorithm for dueling bandits (Li et al., 2024) and Borda-regret guarantees for generalized linear dueling bandits (Wu et al., 2024).

Logistic and generalized linear bandits. Generalized linear bandits, including logistic bandits as a canonical example, model the expected reward through a nonlinear link function applied to a linear predictor. Early UCB-style algorithms for generalized linear bandits incur curvature factors that can be large when the link derivative is small (Filippi et al., 2010; Li et al., 2017). Subsequent work on logistic bandits showed that such worst-case curvature dependence is not always unavoidable. Faury et al. (2020) used self-concordance to sharpen this dependence. Abeille et al. (2021) developed instance-dependent minimax guarantees based on local curvature. Faury et al. (2022) provided a more computationally efficient algorithm. Liu et al. (2026) further studied the simple-regret objective. They showed that the informativeness of an action depends on the local curvature of the sigmoid. Actions with the highest reward therefore need not be the most informative. The same saturation phenomenon appears in dueling bandits: a global lower bound on the sigmoid slope over the reward range may be exponentially small. Chen et al. (2026) avoid this exponential dependence in online RLHF by adopting a preference-based notion of regret. Their objective compares the optimal and learned responses against a fixed sampler policy, rather than measuring standard reward-based regret. Therefore, removing the exponential dependence from the leading regret term for general reward function classes remains an open problem.

General function approximation. The eluder dimension was introduced by Russo & Van Roy (2013) to quantify exploration complexity for general function approximation. Subsequent work on nonlinear bandits and reinforcement learning measures exploration complexity through sums of uncertainty-weighted estimation errors, leading to generalized variants of the eluder dimension (Agarwal et al., 2023; Ye et al., 2023; Di et al., 2024b; Zhao et al., 2024). However, global eluder dimensions can be pessimistic for generalized linear models because they reflect worst-case curvature. The recent localized-eluder framework of Bakhtiari et al. (2026) addresses this by measuring complexity on localized expected excess-loss scales and treating non-local rounds separately. Our work imports the localization perspective into contextual dueling bandits. The main difference is structural: rather than a single-action cost regret bridge, our algorithm is designed to yield a dueling-specific decomposition in terms of two played reward-gap errors, which are then controlled by log-loss excess and localized eluder arguments.

3 PRELIMINARIES

In this work, we study the contextual dueling bandit with a general reward function class. Let $\mathcal{X}$ be the context space and $\mathcal{A}$ be the finite action space. At each round $t \in [T]$, the learner observes the context $x_t \in \mathcal{X}$, selects two actions $a_t, b_t \in \mathcal{A}$, and observes a binary preference feedback $o_t \in \{0, 1\}$, where $o_t = 1$ means that a_t is preferred to b_t, denoted as $a_t \succ b_t$. We use $z = (x, a, b)$ to denote the context-action triplet, and write $z_t = (x_t, a_t, b_t)$ for the query played at round t. Suppose that $r^* : \mathcal{X} \times \mathcal{A} \to [-B, B]$ is the ground truth reward function. The preference probability, i.e., $\mathbb{P}(o_t = 1)$, is determined by the difference of the ground truth reward $\Delta_{r^*}(z_t) = r^*(x_t, a_t) - r^*(x_t, b_t)$ and the link function $\sigma : \mathbb{R} \to [0, 1]$, with

$$\mathbb{P}(o_t = 1|z_t) = \mathbb{P}(a_t \succ b_t|z_t) = \sigma(\Delta_{r^*}(z_t)).$$

We apply the Bradley-Terry-Luce (BTL) model (Hunter, 2004; Luce, 1959) as the link function, where the preference probability satisfies

$$\mathbb{P}(o_t = 1|z_t) = \frac{\exp(r^*(x_t, a_t))}{\exp(r^*(x_t, a_t)) + \exp(r^*(x_t, b_t))} = \frac{1}{1 + \exp(-\Delta_{r^*}(z_t))}.$$

Therefore, in the BTL model, the link function is $\sigma(v) = 1/(1 + e^{-v})$. We make the following realizability assumption on the ground-truth reward function r^*.

Assumption 3.1. There exists a reward function class $\mathcal{R} \subseteq \{r : \mathcal{X} \times \mathcal{A} \to [-B, B]\}$, such that the ground-truth reward function $r^* \in \mathcal{R}$.

Example 3.2. We consider the linear reward function class as an important case: The reward function $(x, a) \mapsto r_{\boldsymbol{\theta}}(x, a) = \langle \boldsymbol{\theta}, \boldsymbol{\mu}(x, a) \rangle$ is parameterized by $\boldsymbol{\theta} \in \Theta \subset \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq B\}$, and $\boldsymbol{\mu}(x, a) \in \mathbb{B}_2^d(0, 1)$ is the feature mapping. Suppose the ground truth reward function is $r^* = r_{\boldsymbol{\theta}^*}$ with $\boldsymbol{\theta}^* \in \Theta$. Therefore, the reward function satisfies $|r_{\boldsymbol{\theta}}(x, a)| \leq \|\boldsymbol{\theta}\|_2 \cdot \|\boldsymbol{\mu}(x, a)\|_2 \leq B$ by Cauchy-Schwarz inequality.

Beyond this linear example, our framework also accommodates nonlinear reward functions. This flexibility comes at the cost of a more involved analysis, but avoids the potentially restrictive assumption that the latent reward is linear in a known feature map.

In this work, we assume $B \geq 1$, but do not impose a constant upper bound on B. Consequently, the latent reward gap can be large: there may exist a triplet $z = (x, a, b) \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ such that

$$|\Delta_{r^*}(z)| \simeq 2B.$$

Such large gaps may drive the sigmoid link function into its saturation regime, where the gradient σ' becomes exponentially small. Indeed, when $u = \Delta_{r^*}(z)$ has large magnitude, the preference probability $\sigma(u)$ is close to either 0 or 1. Thus the comparison is nearly deterministic, while the link itself becomes nearly flat, with its derivative $\sigma'(u) = \sigma(u)(1 - \sigma(u)) \asymp e^{-|u|} \to 0$. This places us outside the regime commonly considered in analyses that assume the link function has a uniformly bounded inverse derivative. In particular, prior works often introduce a condition of the form

$$\kappa := \sup_{z \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}} \left\{ \frac{1}{\sigma'[\Delta_{r^*}(z)]} \right\} < \infty.$$

When σ is sigmoid, however, $\sigma'(u) \asymp e^{-|u|} \to 0$ when $|u| \to \infty$. Previously, prior analyses often treat κ as a benign problem-dependent constant. Thus, if $|\Delta_{r^*}(z)| \simeq 2B$, then $\kappa \simeq e^{2B}$ can be exponentially large in the reward scale.

Our goal is to minimize the cumulative two-armed regret:

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} R_t, \quad \text{where} \quad R_t = 2r^*(x_t, a_t^*) - r^*(x_t, a_t) - r^*(x_t, b_t).$$

With the time-varying context, the optimal action in step t, i.e., $a_t^* = \mathrm{argmax}_{a \in \mathcal{A}} r^*(x_t, a)$, can also be time-varying. The term R_t calculates the gaps of the reward of the optimal action a_t^* and the selected actions a_t, b_t.

4 METHODOLOGY

We now present `SI-CDB` (Algorithm 1). The key component of `SI-CDB` is the asymmetric selection of the two actions: the target action a_t exploits the current reward estimate (Lines 5–6), while the comparator action b_t is chosen optimistically to explore (Line 8).

To select the target action, the algorithm first estimates the reward function from the preference feedback collected so far. For a candidate preference probability $p \in (0, 1)$, the Bernoulli log loss is

$$\ell(o, p) = -o \log p - (1 - o) \log(1 - p). \tag{4.1}$$

Given the past queries $z_1, \ldots, z_{t-1}$ and feedback $o_1, \ldots, o_{t-1}$, the cumulative log loss of a reward function r is

$$L_t(r) = \sum_{i=1}^{t-1} \ell\big(o_i, \sigma[\Delta_r(z_i)]\big). \tag{4.2}$$

The algorithm computes the maximum likelihood estimate $r_1^t \in \mathrm{argmin}_{r \in \mathcal{R}} L_t(r)$ (Line 5). Given r_1^t, the algorithm selects the target action greedily, i.e., $a_t \in \mathrm{argmax}_{a \in \mathcal{A}} r_1^t(x_t, a)$.

To select the comparator action, the algorithm first constructs a confidence set around r_1^t using the same loss L_t:

$$\mathrm{Conf}_t = \{r \in \mathcal{R} : L_t(r) \leq L_t(r_1^t) + \beta_t\},$$

which contains the ground-truth reward r^* with high probability. Similar constructions are widely used in model-based RL (Zhan et al., 2022; Liu et al., 2023; Wang et al., 2025) for optimistic exploration with general function approximation. The algorithm then jointly selects the comparator action b_t and an optimistic reward function $r_2^t \in \mathrm{Conf}_t$ by maximizing the following objective over $\mathcal{A} \times \mathrm{Conf}_t$.

$$J_t(b, r) := 2 \cdot r(x_t, b) \; - 1 \cdot r_1^t(x_t, b) \; - 2 \cdot r(x_t, a_t), \qquad (b_t, r_2^t) \in \operatorname*{argmax}_{b \in \mathcal{A},\, r \in \mathrm{Conf}_t} J_t(b, r), \quad (4.3)$$

Maximizing over $r \in \mathrm{Conf}_t$ plays the role of an uncertainty bonus without requiring an explicit feature representation. To explain the coefficients $(2, -1, -2)$, we rewrite $J_t(b, r)$ as

$$J_t(b, r) = \underbrace{r_1^t(x_t, b) + 2[\Delta_{r_1^t}(x_t, a_t, b) - \Delta_r(x_t, a_t, b)]}_{\widetilde{J}_t(b,r)} - 2r_1^t(x_t, a_t).$$

The term $-2r_1^t(x_t, a_t)$ does not depend on b or r. Thus, maximizing $J_t(b, r)$ is equivalent to maximizing $\widetilde{J}_t(b, r)$. Its two terms correspond to exploitation and exploration, respectively:

- The first term favors a comparator with high estimated reward. It therefore encourages exploitation.
- The second term favors disagreement in the played gap among plausible models.

The coefficients are selected to produce the two-gap decomposition in Lemma B.2, not merely as a heuristic. We summarize the procedue in Algorithm 1. In Appendix F.2, we will provide a comparison of `SI-CDB` with prior work (Bengs et al., 2022; Di et al., 2025) when restricted to the linear setting.

Algorithm 1 Saturation-Insensitive Contextual Dueling Bandits (`SI-CDB`)

Input: Reward function class $\mathcal{R}$, action space $\mathcal{A}$, confidence widths $(\beta_t)_{t \le T}$.
Initialize $L_1(r) = 0$ for all $r \in \mathcal{R}$.
for $t = 1, \ldots, T$ **do**
  Observe context x_t.
  Select $r_1^t \in \operatorname{argmin}_{r \in \mathcal{R}} L_t(r)$.
  Select target action $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} r_1^t(x_t, a)$.
  Construct confidence set $\mathrm{Conf}_t = \{r \in \mathcal{R} : L_t(r) \le L_t(r_1^t) + \beta_t\}$.
  Select comparator pair of action and reward function: $(b_t, r_2^t) \in \operatorname{argmax}_{b \in \mathcal{A}, r \in \mathrm{Conf}_t} J_t(b, r)$ where $J_t(b, r)$ is defined in (4.3).
  Play (a_t, b_t), and observe preference $o_t \in \{0, 1\}$.
  Update $L_{t+1}(r) = L_t(r) + \ell(o_t, \sigma(\Delta_r(z_t)))$ for all $r \in \mathcal{R}$, where $\ell(\cdot, \cdot)$ is defined in (4.1)
end for

5 Main Results

In this section, we present the regret guarantee of `SI-CDB`. We first introduce the complexity measures used in our analysis, namely the covering number and the localized eluder dimension of the excess-loss class induced by $\mathcal{R}$ (Section 5.1). We then state the regret bound for general reward classes and specialize it to linear rewards (Section 5.2).

5.1 Covering Number and Localized Eluder Dimension

Our regret bound depends on two complexity measures of the excess-loss classes induced by the reward class: a covering number and a localized eluder dimension. We first introduce these classes, which reflect the preference feedback available to the learner.

In standard contextual bandits with reward feedback, candidate functions are distinguished through differences in predicted rewards. In dueling bandits, however, the learner observes preferences, so statistical distinguishability depends on the Bernoulli distributions induced by the reward functions through the comparison link. We capture this distinction using the excess log loss. For each $r \in \mathcal{R}$ and query $z = (x, a, b)$, define the excess log loss and its expectation under the true preference distribution as

$$\phi_r(o, z) = \ell(o, \sigma[\Delta_r(z)]) - \ell(o, \sigma[\Delta_{r^*}(z)]),$$

$$\bar{\phi}_r(z) = \mathbb{E}_{o\sim \mathrm{Ber}(\sigma[\Delta_{r^*}(z)])}[\phi_r(o,z)] = \mathrm{KL}\left(\mathrm{Ber}(\sigma[\Delta_{r^*}(z)]) \| \mathrm{Ber}(\sigma[\Delta_r(z)])\right). \tag{5.1}$$

Thus, $\bar{\phi}_r(z)$ equals the Kullback–Leibler divergence from the true preference distribution to that induced by r at query z. We define the corresponding classes of realized and expected excess loss as

$$\Phi(\mathcal{R}) = \{\phi_r : r \in \mathcal{R}\}, \quad \bar{\Phi}(\mathcal{R}) = \{\bar{\phi}_r : r \in \mathcal{R}\}. \tag{5.2}$$

The two classes enter our analysis in different ways. The covering number of $\Phi(\mathcal{R})$ to obtain concentration bounds that hold uniformly over all candidate reward functions, while the localized eluder dimension of $\bar{\Phi}(\mathcal{R})$ to bound the cumulative expected excess loss of the selected models on the queries made by the algorithm.

Covering number. Let $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|)$ denote the covering number of $\mathcal{F}$ at radius ϵ under the norm $\|\cdot\|$. For the realized excess-loss class, we use

$$N_T = \mathcal{N}(\Phi(\mathcal{R}), 1/T, \|\cdot\|_\infty). \tag{5.3}$$

Here $\|\cdot\|_\infty$ is the supremum norm over observations (o, z).

Localized eluder dimension. To avoid exponential dependence on B in the leading regret term, we seek to control regret using local rather than worst-case curvature. The localized eluder dimension of Bakhtiari et al. (2026) provides a tool for controlling expected excess loss at local scales. Here, we provide the formal definition.

Definition 5.1 (Localized Eluder Dimension, Bakhtiari et al. 2026)**.** Let $\mathcal{Z}$ be a set, Ψ be a class of real-valued functions on $\mathcal{Z}$. Let $z = (z_1, z_2, \ldots, z_n)$ be a length-n sequence in $\mathcal{Z}$. The localized Eluder dimension is defined as follows:

- We say that $x \in \mathcal{Z}$ is ϵ-independent of z with respect to Ψ if there exists $\psi \in \Psi$ such that
$$\sum_{t=1}^{n} |\psi(z_t)| \le \epsilon, \quad \text{and} \quad |\psi(x)| > \epsilon.$$
- We say that z is an ϵ-eluder sequence with respect to Ψ if, for every $t \le n$, the point z_t is ϵ-independent of $z_1, \ldots, z_{t-1}$ with respect to Ψ.
- The (ϵ, η)-localized eluder dimension $\dim^{\eta}_{\text{elud}}(\epsilon, \Psi)$ of Ψ is the maximum length of an ω-eluder sequence with respect to Ψ over all $\omega \in [\epsilon, \eta]$.

In this work, we consider $d^{\epsilon,\eta}_{\text{elud}}(\mathcal{R}) := \dim^{\eta}_{\text{elud}}\big(\epsilon, \bar{\Phi}(\mathcal{R})\big)$. We write it as $d^{\eta}_{\text{elud}}(\mathcal{R})$ when $\epsilon = 1/T$.

The upper scale η restricts the independence thresholds considered in the definition, allowing us to measure complexity at the excess-loss scales relevant to our regret analysis. However, the analysis of Bakhtiari et al. (2026) does not directly address two-arm reward regret under pairwise preference feedback. Our asymmetric selection rule and the resulting regret decomposition enable us to extend the localized analysis to this setting, as detailed in Section 6.

Remark 5.2. In prior work, complexity measures can be distinguished according to whether they are determined by the hypothesis class alone or also depend on the underlying problem instance. The first category includes the standard eluder dimension and its weighted generalization (Russo & Van Roy, 2013; Agarwal et al., 2023), which is defined by differences between functions in the hypothesis class and do not depend on the ground-truth function. The second category includes the Bellman eluder dimension (Jin et al., 2021), defined through residuals involving the *real* Bellman operator, and the complexity measure of Wang et al. (2025) for the model-based setting. Our measure belongs to the second category, as ϕ_r is defined relative to the true reward function r^*.

5.2 Regret Bound of SI-CDB

With the covering number and the localized eluder dimension, we now provide the regret upper bound of SI-CDB:

Theorem 5.3. For $\delta \in (0, 1)$, choose the confidence radius as $\beta_t = O\big(B \log(N_T/\delta)\big)$. where N_T is defined in (5.3). Then, with probability at least $1 - \delta$, the regret of Algorithm 1 can be upper bounded by

$$\mathrm{Regret}(T) \le \inf_{u>0} \left[3\sqrt{\frac{2T\Gamma_u}{\sigma'(4u)}} + 4B\left(\frac{8\beta_T}{u^2\sigma'(2B)} + 1\right) d^{\theta_u,\theta_u}_{\text{elud}}(\mathcal{R})\right] + 8B.$$

where

$$\Gamma_u = O(\beta_T \log(1+BT)\, d_{\text{elud}}^{\tau_u}(\mathcal{R}) + B)\,, \qquad \tau_u = \min\{2B, u^2/8\}, \qquad \theta_u = \frac{\sigma'(2B)}{2}u^2.$$

In Theorem 5.3, u is an analysis parameter that sets the localization scale and balances the following two contributions to regret:

- The first term controls regret on localized rounds and depends on the local curvature $\sigma'(4u)$. For $u = O(1)$, this curvature is bounded away from zero, so the term avoids the worst-case inverse-curvature factor. Its remaining dependence on the reward class is captured by Γ_u.
- The second term accounts for non-localized rounds. It retains the potentially exponential factor $1/\sigma'(2B)$, but has no explicit $\sqrt{T}$ multiplier. Its dependence on T enters through β_T and $d_{\text{elud}}^{\theta_u,\theta_u}(\mathcal{R})$.

The infimum over $u > 0$ allows us to select the best tradeoff between these two contributions. Thus, the bound separates the local-curvature contribution from an additive term that retains the worst-case curvature dependence.

To demonstrate the effectiveness of our results, we consider the linear function class. We firstly establish the bound for the covering number and the localized eluder dimension of the linear reward function class:

Proposition 5.4. The uniform-covering number can be bounded by

$$\log N_T \le d\log(1+8BT).$$

Moreover, there exists a universal constant $C > 0$ such that for every $0 < \epsilon \le \eta \le \sigma'(2B)/4$, we have

$$d_{\text{elud}}^{\epsilon,\eta}(\mathcal{R}) \le Cd\log\Big(1+\frac{B^2}{\epsilon}\Big).$$

For the derivation of these results, please refer to Appendix D.

Setting $u = \sqrt{\sigma'(2B)}$ in Theorem 5.3 ensures that both τ_u and θ_u are at most $\sigma'(2B)/4$, as required by the preceding proposition. Then, $\theta_u = [\sigma'(2B)]^2/2$. As a result, we have

Corollary 5.5. With probability at least $1-\delta$, the regret of Algorithm 1 can be upper bounded by

$$\text{Regret}(T) \le \widetilde{O}\bigg(d\sqrt{BT} + \frac{d^2B^2}{[\sigma'(2B)]^2}\bigg).$$

The dependence of d and T in the leading term $\widetilde{O}(d\sqrt{BT})$ matches the lower bound proved in Li et al. (2024). Moreover, it improves the dependence of B to $\sqrt{B}$ compared with Di et al. (2025), while our results can further extend to general function approximation.

6 Proof Overview

Before presenting the proof, we emphasize the role of two-armed regret in our analysis. The key insight is that, even when only the performance of the first arm matters, controlling two-armed regret provides the query-gap localization needed to avoid worst-case inverse-curvature dependence. We illustrate this connection using the linear reward class in Example 3.2.

Intuition. Let $\mathbf{d}_t = \boldsymbol{\mu}(x_t, a_t) - \boldsymbol{\mu}(x_t, b_t)$ denote the feature difference of the queried pair. The Hessian of its logistic loss, evaluated at $\boldsymbol{\theta}^*$, is

$$\sigma'(\mathbf{d}_t^\top \boldsymbol{\theta}^*)\, \mathbf{d}_t \mathbf{d}_t^\top.$$

The local curvature factor $\sigma'(\mathbf{d}_t^\top \boldsymbol{\theta}^*)$ can be much larger than its uniform lower bound $\sigma'(2B)$. In particular, on rounds with small two-armed regret, both queried actions have near-optimal rewards, and hence

$$\big|\mathbf{d}_t^\top \boldsymbol{\theta}^*\big| = |r^*(x_t, a_t) - r^*(x_t, b_t)| \le R_t.$$

As this gap approaches zero, $\sigma'(\mathbf{d}_t^\top \boldsymbol{\theta}^*)$ approaches $1/4$. Thus, on rounds with small two-armed regret, the local curvature is bounded away from zero even when the worst-case lower bound $\sigma'(2B)$ is exponentially small. In contrast, small one-arm regret does not ensure benign curvature: even if a_t is optimal, a suboptimal comparator b_t can leave the queried reward gap large and the comparison highly saturated.

To use this intuition in the proof, we must identify rounds on which the two-armed regret is small. Our asymmetric selection rule makes this possible by bounding R_t in terms of the gap-estimation errors at the queried pair. When these estimation errors are small, the regret bound ensures that both the true and estimated reward gaps lie in a region of benign curvature. We formalize this argument through an asymmetric regret decomposition and a localized eluder analysis.

Key Technique I: Asymmetric Regret Decomposition. Algorithm 1 maintains two reward functions with distinct roles. The MLE r_1^t determines the greedy target action a_t, while $r_2^t \in \mathrm{Conf}_t$ is selected jointly with the comparator b_t by maximizing the asymmetric objective J_t. We show how these two selection rules combine to bound two-armed regret by estimation errors at the queried pair.

Assume $r^* \in \mathrm{Conf}_t$, as justified by the concentration argument below. Then (a_t^*, r^*) is feasible for the comparator optimization, so $J_t(a_t^*, r^*) \le J_t(b_t, r_2^t)$. Expanding this inequality gives

$$\begin{aligned} &2r^*(x_t, a_t^*) - r_1^t(x_t, a_t^*) - 2r^*(x_t, a_t) \\ &\quad \le 2r_2^t(x_t, b_t) - r_1^t(x_t, b_t) - 2r_2^t(x_t, a_t). \end{aligned}$$

Substituting into the definition of R_t and using the greedy optimality condition $r_1^t(x_t, a_t^*) \le r_1^t(x_t, a_t)$, we obtain

$$\begin{aligned} R_t &\le r_1^t(x_t, a_t^*) - r_1^t(x_t, b_t) + \Delta_{r^*}(z_t) - 2\Delta_{r_2^t}(z_t) \\ &\le \Delta_{r_1^t}(z_t) + \Delta_{r^*}(z_t) - 2\Delta_{r_2^t}(z_t). \end{aligned}$$

Thus the greedy selection removes the dependence on a_t^*, and the coefficients in J_t allow the remaining terms to be written as two gap-estimation errors. Taking absolute values yields the following decomposition.

Lemma 6.1 (Informal)**.** For any round t with $r^* \in \mathrm{Conf}_t$, the actions and reward functions selected by Algorithm 1 satisfy

$$R_t \le \underbrace{\left|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)\right|}_{e_{1,t}} + 2\underbrace{\left|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)\right|}_{e_{2,t}}.$$

Crucially, both errors concern the comparison actually queried: no estimation error at the unknown optimal action remains. The decomposition therefore connects reward regret to the expected excess losses $\bar{\phi}_{r_j^t}(z_t)$ defined in (5.1), which measure how well the selected models predict the observed preferences. It also provides the bound $R_t \le 3\max_j e_{j,t}$ needed to turn small estimation errors into query-gap localization. **Thus, small gap-estimation errors ensure that both queried actions are near-optimal.** We use both consequences in the next part.

Key Technique II: Localized Eluder Analysis. We combine the regret decomposition with the localized eluder tools to control regret using the curvature of the queried comparisons. For a threshold $u > 0$, define

$$E_t := \max_{j\in\{1,2\}} e_{j,t}, \qquad I_T := \{t \in [T] : E_t \le u\}, \qquad I_T^c := [T] \setminus I_T.$$

This partition is used only in the analysis: the algorithm neither observes E_t nor needs to identify the localized rounds. The same sequence of actions is analyzed for every $u > 0$, allowing us to optimize the resulting bound over u.

On the localized rounds I_T, the regret decomposition gives $R_t \le 3E_t \le 3u$. Because R_t sums the suboptimalities of both queried actions, it also bounds their true reward gap. Consequently, for $j \in \{1, 2\}$,

$$\begin{aligned} &|\Delta_{r^*}(z_t)| \le R_t \le 3u, \\ &|\Delta_{r_j^t}(z_t)| \le |\Delta_{r^*}(z_t)| + e_{j,t} \le 4u. \end{aligned}$$

Thus both the true and estimated gaps lie in $[-4u, 4u]$. **Controlling both arms is what bounds the true query gap here.** Since the second derivative of the logistic loss is σ', its curvature between the true and estimated gaps lies between $\sigma'(4u)$ and $1/4$. A second-order expansion therefore gives

$$\frac{\sigma'(4u)}{2} e_{j,t}^2 \le \bar{\phi}_{r_j^t}(z_t) \le \frac{1}{8} e_{j,t}^2 \le \frac{u^2}{8}.$$

The lower bound converts expected excess loss into regret using local curvature, while the upper bound restricts the excess-loss scale needed for the localized eluder analysis. It remains to control the accumulation of these losses along the adaptively selected queries.

For this step, we first use the uniform concentration bound of Bakhtiari et al. (2026). With probability at least $1 - \delta$, uniformly over $r \in \mathcal{R}$ and $t \in [T]$,

$$\sum_{i<t} \bar{\phi}_r(z_i) \le 2 \sum_{i<t} \phi_r(o_i, z_i) + 2\beta_t.$$

Since expected excess losses are nonnegative, this event ensures $r^* \in \mathrm{Conf}_t$. Moreover, the definition of the confidence set implies that each selected estimator has bounded historical expected excess loss:

$$\sum_{i<t} \bar{\phi}_{r_j^t}(z_i) \le 4\beta_T, \qquad j \in \{1, 2\}.$$

Although the selected model changes across rounds, each model must remain consistent with the previous comparisons. The localized eluder dimension controls how often such a model can fit the history yet incur substantial expected excess loss on a new query. This connects the historical loss bound to the accumulation of losses along the algorithm's queries.

This bound also holds when the history is restricted to I_T. Since the excess losses on I_T are localized at scale τ_u, the historical loss bound allows us to control their accumulation using the localized eluder dimension, giving $\sum_{t \in I_T} \bar{\phi}_{r_j^t}(z_t) = O(\Gamma_u)$. The local curvature bound and Cauchy–Schwarz then give

$$\sum_{t \in I_T} R_t \le \sum_{t \in I_T} (e_{1,t} + 2e_{2,t}) = O\left(\sqrt{\frac{T\Gamma_u}{\sigma'(4u)}} \right).$$

For the remaining rounds I_T^c, we bound their number instead. On each such round, at least one error exceeds u. The global curvature bound implies that its expected excess loss exceeds $\theta_u = \sigma'(2B)u^2/2$. The historical loss bound and the localized eluder dimension then control the number of such rounds:

$$|I_T^c| = O\left(\left(1 + \frac{\beta_T}{\sigma'(2B)u^2} \right) d_{\text{elud}}^{\theta_u, \theta_u}(\mathcal{R}) + 1 \right).$$

Using $R_t \le 4B$ on these remaining rounds and optimizing over u yields the two contributions in Theorem 5.3. The worst-case inverse-curvature factor enters the bound for the remaining rounds, while the localized $\sqrt{T}$ term uses $\sigma'(4u)$ and the localized complexity encoded in Γ_u.

7 Conclusion

We studied contextual dueling bandits with general reward function approximation under BTL preference feedback. The central challenge is the saturation of the logistics link: large reward gaps produce nearly deterministic preferences and small sigmoid derivatives, causing standard analyses to pay worst-case curvature factors. We proposed `SI-CDB`, a saturation insensitive algorithm that maintains a Bernoulli log-loss confidence set, selects a greedy target action and chooses an optimistic competitor through an objective tailored to dueling feedback.

Limitations. `SI-CDB` can be computationally demanding, as each round requires empirical risk minimization and joint optimization over the action space and the confidence set. Practical approximations, including warm-started gradient updates and alternating optimization, are discussed in Appendix G, although our current guarantees assume exact optimization. Developing more efficient algorithms and tractable exact implementations that retain saturation-insensitive guarantees remains an important direction for future work.

AI USE STATEMENT

We used Claude Opus 5, GPT-5.6, and GPT-6 to assist with manuscript writing and to help check the correctness of proofs. The main proofs were developed by the authors. The authors take full responsibility for the final manuscript, including all AI-assisted content, and for the correctness of the proofs and results.

ETHICS STATEMENT

This work studies the theoretical foundations of contextual dueling bandits with general function approximation. It does not involve experiments with human participants or the collection of personal data. We do not identify specific ethical concerns arising from the theoretical results presented in this paper.

REPRODUCIBILITY STATEMENT

This work is theoretical. The problem setting and assumptions are specified in Section 3, and the algorithm is described in Section 4 with pseudocode in Algorithm 1. Detailed proofs of the main results and supporting lemmas are provided in Appendices B and C. Appendix D derives the complexity bounds for the linear reward class, while Appendix F.2 explains the linear specialization of the comparator objective. Practical approximation procedures are described in Appendix G.

REFERENCES

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 987–1063. PMLR, 2023.

Alireza Bakhtiari, Alex Ayoub, Samuel Robertson, David Janz, and Csaba Szepesvári. Eluder dimension: localise it! *Advances in Neural Information Processing Systems*, 38:62160–62200, 2026.

Akshay Balsubramani, Zohar Karnin, Robert E. Schapire, and Masrour Zoghi. Instance-dependent regret bounds for dueling bandits. In *Conference on Learning Theory*, pp. 336–360. PMLR, 2016.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022.

Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. Avoiding exp (r) scaling in rlhf through preference-based exploration. *Advances in Neural Information Processing Systems*, 38: 164474–164510, 2026.

Yuwei Cheng, Fan Yao, Xuefeng Liu, and Haifeng Xu. Learning from imperfect human feedback: A tale from corruption-robust dueling. In *International Conference on Learning Representations*, volume 2025, pp. 29795–29826, 2025.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 96–112. Springer, 2025.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *International Conference on Learning Representations*, volume 2024, pp. 36738–36761, 2024a.

Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4377–4410, 2024b.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. In *International Conference on Machine Learning*, pp. 13513–13545. PMLR, 2025.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *International Conference on Machine Learning*, pp. 1088–1096. PMLR, 2017.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 2020.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pp. 1057–1066. PMLR, 2018.

David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1): 384–406, 2004.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pp. 416–424. PMLR, 2015.

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401*, 2024.

Xiang Ji, Huazheng Wang, Minshuo Chen, Tuo Zhao, and Mengdi Wang. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, pp. 1141–1154. PMLR, 2015.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, pp. 1235–1244. PMLR, 2016.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. *arXiv preprint arXiv:2404.06013*, 2024.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 363–376, 2023.

Shuai Liu, Alireza Bakhtiari, Alex Ayoub, Botao Hao, and Csaba Szepesvári. Learning what to recommend: Minimax optimal simple regret in logistic bandits. *arXiv preprint arXiv:2601.21167*, 2026.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Luckeciano C Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37:118052–118085, 2024.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024.

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.

Siddartha Y. Ramamohan, Arun Rajkumar, and Shivani Agarwal. Dueling bandits: Beyond Condorcet winners to general tournament solutions. In *Advances in Neural Information Processing Systems*, 2016.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994. PMLR, 2022.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pp. 9235–9244. PMLR, 2021.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. In *Advances in Neural Information Processing Systems*, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Zhiyong Wang, Dongruo Zhou, John C. S. Lui, and Wen Sun. Model-based RL as a minimalist approach to horizon-free and second-order bounds. In *International Conference on Learning Representations*, 2025.

Huasen Wu and Xin Liu. Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 649–657, 2016.

Yue Wu, Tao Jin, Qiwei Di, Hao Lou, Farzad Farnoud, and Quanquan Gu. Borda regret minimization for generalized linear dueling bandits. In *International Conference on Machine Learning*, pp. 53571–53596. PMLR, 2024.

Tengyang Xie, Dylan Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed H Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. In *International Conference on Learning Representations*, volume 2025, pp. 43632–43669, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54715–54754. PMLR, 2024.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pp. 39834–39863. PMLR, 2023.

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.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 37:94684–94735, 2024.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative upper confidence bound for the K-armed dueling bandit problem. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2014.

Masrour Zoghi, Zohar Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 307–315, 2015.

A Additional Notations

The log-loss can be written in the canonical logistic form

$$\ell(o, \sigma(v)) = \psi(v) - ov, \qquad \psi(v) = \log(1 + e^v). \tag{A.1}$$

This representation is useful because the curvature of the loss, $\psi''(v) = \sigma'(v)$, explicitly captures the sensitivity of binary comparisons to the reward gap: comparisons are informative when the gap is moderate, but become statistically weak when the sigmoid link is saturated.

Using (A.1), we can represent it as a Bregman divergence, i.e.,

$$\begin{aligned}\bar{\phi}_r(z) &= \psi\big(\Delta_r(z)\big) - \psi\big(\Delta_{r^*}(z)\big) - \frac{e^{\Delta_{r^*}(z)}}{1 + e^{\Delta_{r^*}(z)}}\big(\Delta_r(z) - \Delta_{r^*}(z)\big) \\ &= B_\psi\big(\Delta_r(z), \Delta_{r^*}(z)\big)\end{aligned} \tag{A.2}$$

B Proof of Main Results

Lemma B.1. The event $\mathcal{E}$ defined in (B.1) holds with probability at least $1 - \delta$.

We firstly provide a detailed description of Lemma B.1. As in (B.1), we define the high-probability event $\mathcal{E}$, which denotes the event where for any $r \in \mathcal{R}$ and $t \in [T]$, we have

$$\sum_{i=1}^{t-1} \bar{\phi}_r(z_i) \le 2 \sum_{i=1}^{t-1} \phi_r(o_i, z_i) + 2\beta_t. \tag{B.1}$$

On the event of $\mathcal{E}$, since the left hand side is non-negative because each $\phi_r(z_i)$ is a KL-divergence, we have for any $r \in \mathcal{R}$ and $t \in [T]$,

$$L_t(r) + \beta_t - L_t(r^*) = \sum_{i=1}^{t-1} \phi_r(o_i, z_i) + \beta_t \ge 0.$$

Therefore, $r^* \in \mathrm{Conf}_t$ for any $t \in [T]$, so we can use the property $J(a_t^*, r^*) \le J(b_t, r_2^t)$ to bound the term $r^*(x_t, a_t^*)$ in the regret.

From now on, we suppose the event $\mathcal{E}$ holds. We consider the following regret decomposition:

Lemma B.2. Suppose the event $\mathcal{E}$ holds. For any t, let r_1^t, r_2^t, z_t be defined in Algorithm 1. Then, the regret R_t at step t can be decomposed as

$$R_t \le \big|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)\big| + 2\big|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)\big|.$$

To continue, we define $E_t = \max\big\{\big|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)\big|, \big|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)\big|\big\}$. For a constant $u > 0$ to be determined, we consider the following "good" index sets:

$$I_t := \big\{s \le t : E_t \le u\big\}.$$

We consider the regret on I_T and I_T^c separately, i.e.,

$$\mathrm{Regret}(T) = \underbrace{\sum_{t \in I_T} R_t}_{(I_1)} + \underbrace{\sum_{t \in I_T^c} R_t}_{(I_2)}.$$

Bounding (I_1). Using Lemma B.2, we have

$$(I_1) \le \underbrace{\sum_{t \in I_T} \big|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)\big|}_{(I_{1,1})} + 2 \underbrace{\sum_{t \in I_T} \big|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)\big|}_{(I_{1,2})}. \tag{B.2}$$

Due to the definition of I_T, for any $t \in I_T$, we have $R_t \le 3E_t \le 3u$. Moreover, we have

$$|\Delta_{r^*}(z_t)| = |r^*(x_t, a_t) - r^*(x_t, b_t)|$$

$$
\begin{aligned}
&= \big|\big(r^*(x_t, a_t) - r^*(x_t, a_t^*)\big) + \big(r^*(x_t, a_t^*) - r^*(x_t, b_t)\big)\big| \\
&\le \big|r^*(x_t, a_t) - r^*(x_t, a_t^*)\big| + \big|r^*(x_t, a_t^*) - r^*(x_t, b_t)\big| \\
&= 2r^*(x_t, a_t^*) - r^*(x_t, a_t) - r^*(x_t, b_t) = R_t.
\end{aligned}
$$

Thus, we have $|\Delta_{r^*}(z_t)| \le R_t \le 3u$. For any $j \in \{1, 2\}$, we have the following upper bound for $|\Delta_{r_j^t}(z_t)|$:

$$
\begin{aligned}
\big|\Delta_{r_j^t}(z_t)\big| &\le \big|\Delta_{r^*}(z_t)\big| + \big|\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)\big| \\
&\le 3u + E_t \le 4u.
\end{aligned}
$$

where the first inequality holds by the triangle inequality, and the second inequality holds by the definition of E_t. Next, we need the following lemma:

Lemma B.3. Let $\bar{\phi}_r$ be defined in (5.1). For any $r \in \mathcal{R}$ and $z \in \mathcal{Z}$, there exists ξ on the segment between $\Delta_{r^*}(z)$ and $\Delta_r(z)$ such that

$$
\bar{\phi}_r(z) = \tfrac{1}{2}\,\sigma'(\xi)\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2.
$$

Consequently, for any $M \ge \max\{|\Delta_r(z)|, |\Delta_{r^*}(z)|\}$, we have the following inequality:

$$
\frac{\sigma'(M)}{2}\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2 \le \bar{\phi}_r(z) \le \frac{1}{8}\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2.
$$

Using this lemma with $M = 4u$, we know that for any $j \in \{1, 2\}$,

$$
\big(\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)\big)^2 \le \frac{2}{\sigma'(4u)}\bar{\phi}_{r_j^t}(z_t) \tag{B.3}
$$

Therefore, we have

$$
\begin{aligned}
(I_{1,j}) &= \sum_{t \in I_T} \big|\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)\big| \\
&\le \sqrt{\sum_{t \in I_T} \big(\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)\big)^2} \cdot \sqrt{T} \\
&\le \sqrt{\frac{2T}{\sigma'(4u)}} \cdot \sqrt{\sum_{t \in I_T} \bar{\phi}_{r_j^t}(z_t)},
\end{aligned} \tag{B.4}
$$

where the first inequality holds due to the Cauchy-Schwarz inequality, and $|I_T| \le T$. The second inequality holds due to (B.3). The next lemma shows an upper bound of $\sum_{i=1}^{t-1} \bar{\phi}_{r_j^t}(z_i)$, which is important to further control $(I_{1,j})$.

Lemma B.4. Suppose the high-probability event $\mathcal{E}$ defined in Lemma B.1 holds. Then, for any $t \in [T+1]$, $j \in \{1, 2\}$, we have

$$
\sum_{i=1}^{t-1} \bar{\phi}_{r_j^t}(z_i) \le 4\beta_T.
$$

Note that $\bar{\phi}_r(z) \ge 0$ because it's defined as a Bregman divergence. Moreover, $I_T \subseteq [T]$. Let $I_T = \{t_1 < t_2 < \ldots < t_K\}$, where $|I_T| = K$. Thus, for any $1 \le k \le K$, we know that

$$
\sum_{l=1}^{k-1} \bar{\phi}_{r_j^{t_k}}(z_{t_l}) \le \sum_{i=1}^{t_k - 1} \bar{\phi}_{r_j^{t_k}}(z_i) \le 4\beta_T.
$$

Moreover, for any k, we have

$$
\bar{\phi}_{r_j^{t_k}}(z_{t_k}) \le \frac{1}{8}\big(\Delta_{r_j^{t_k}}(z_{t_k}) - \Delta_{r^*}(z_{t_k})\big)^2
$$

due to Lemma B.3. First, using the definition of E_t, we have

$$
\frac{1}{8}\big(\Delta_{r_j^{t_k}}(z_{t_k}) - \Delta_{r^*}(z_{t_k})\big)^2 \le \frac{1}{8}(E_{t_k})^2 \le \frac{u^2}{8},
$$

where the last inequality uses $t_k \in I_T$. Second, we have the following lemma

Lemma B.5. Let ϕ_r and $\bar{\phi}_r$ be defined in (5.1). Then, for any reward function $r \in \mathcal{R}$, any $(o, z) \in \{0,1\} \times \mathcal{X} \times \mathcal{A} \times \mathcal{A}$, the following properties hold:

$$|\phi_r(o,z)| \le 2B.$$

Additionally, we have

$$\bar{\phi}_r(z) = \mathbb{E}_o[\phi_r(o,z)] \le 2B, |\phi_r(o,z) - \bar{\phi}_r(z)| \le 4B.$$

Hence, $\bar{\phi}_{r_j^{t_k}}(z_{t_k}) \le 2B$. In conclusion, we have

$$\phi_{r_j^{t_k}}(z_{t_k}) \le \min\{2B, u^2/8\} := \tau_u.$$

Now, applying Lemma E.2, for any $k \in [K]$, $j \in \{1,2\}$, and $\epsilon > 0$, the following inequality holds: $d_{\text{elud}}^{\epsilon,\eta}(\mathcal{R})$

$$\sum_{l=1}^{k} \phi_{r_j^{t_l}}(z_{t_l}) \le \big(2B + \beta_T \log(1 + 2B/\epsilon)\big) d_{\text{elud}}^{\epsilon,\tau_u}(\mathcal{R}) + 2B + K\epsilon. \tag{B.5}$$

Now, set $\epsilon = 1/T$, and define

$$\Gamma_u = \big(2B + \beta_T \log(1 + 2BT)\big) d_{\text{elud}}^{\tau_u}(\mathcal{R}) + 2B + 1.$$

Combining (B.2), (B.4) and (B.5), we have shown that

$$(I_1) \le (I_{1,1}) + 2(I_{1,2}) \le 3\sqrt{\frac{2T\Gamma_u}{\sigma'(4u)}}. \tag{B.6}$$

Bounding (I_2). Using $r^*(\cdot) \in [-B, B]$, we directly have $(I_2) \le 4B|I_T^c|$. It suffices to bound $|I_T^c|$. Note that for any $t \in I_T^c$, $E_t = \max\big\{\big|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)\big|, \big|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)\big|\big\} > u$. Trivially, $\max\{|\Delta_r(z)|, |\Delta_{r^*}(z)|\} \le 2B$. Using Lemma B.3, for any $j \in \{1,2\}$, we have

$$\bar{\phi}_{r_j^t}(z_t) \ge \frac{\sigma'(2B)}{2}\big(\Delta_{r_j^t}(z_t) - \Delta_{r^*}(z_t)\big)^2.$$

Therefore,

$$\max_{j\in\{1,2\}} \bar{\phi}_{r_j^t}(z_t) \ge \frac{\sigma'(2B)}{2}(E_t)^2 > \frac{\sigma'(2B)}{2}u^2 := \theta_u.$$

This indicates

$$I_u^c \subseteq \bigcup_{j\in\{1,2\}} \left\{t \in [T] : \bar{\phi}_{r_j^t}(z_t) > \frac{\sigma'(2B)}{2}u^2\right\}. \tag{B.7}$$

Using Lemma B.4, for any $t \in [T+1]$, $j \in \{1,2\}$, we have

$$\sum_{i=1}^{t-1} \bar{\phi}_{r_j^t}(z_i) \le 4\beta_T.$$

Now we can apply Lemma E.2. For every $\epsilon \in (0, \theta_u]$, $j \in \{1,2\}$, and every t, we have

$$\sum_{i=1}^{t-1} \mathbb{1}\left\{\bar{\phi}_{r_j^i}(z_i) > \epsilon\right\} \le \left(\frac{4\beta_T}{\epsilon} + 1\right) d_{\text{elud}}^{\epsilon,\theta_u}(\mathcal{R}) + 1.$$

Using (B.7), we have

$$\begin{aligned} |I_u^c| &\le \sum_j \left|\left\{t \in [T] : \bar{\phi}_{r_j^t}(z_t) > \theta_u\right\}\right| \\ &= \sum_j \sum_{t\in[T]} \mathbb{1}\left\{\bar{\phi}_{r_j^t}(z_t) > \theta_u\right\} \end{aligned}$$

$$\leq 2\left(\frac{8\beta_T}{u^2\sigma'(2B)}+1\right)d_{\text{elud}}^{\theta_u,\theta_u}(\mathcal{R})+2.$$

As a result, we have

$$(I_2)\leq 4B\left(\frac{8\beta_T}{u^2\sigma'(2B)}+1\right)d_{\text{elud}}^{\theta_u,\theta_u}(\mathcal{R})+8B. \tag{B.8}$$

In conclusion, combining (B.6) and (B.8), we have

$$\text{Regret}(T)\leq 3\sqrt{\frac{2T\Gamma_u}{\sigma'(4u)}}+4B\left(\frac{8\beta_T}{u^2\sigma'(2B)}+1\right)d_{\text{elud}}^{\theta_u,\theta_u}(\mathcal{R})+8B.$$

Taking u which can achieve the minimum, we have completed the proof of Theorem 5.3.

C Proof of Lemmas in Appendix B

C.1 Proof of Lemma B.1

Proof of Lemma B.1. We define a filtration $\mathcal{F}_t$ to be the σ-algebra generated by $\{Z_s\}_{s\leq t-1}\cup z_t$. Recall that $\phi_r(o,z)=\ell(o,f_r(z))-\ell(o,f_{r^*}(z))$. Note that

$$\mathbb{E}[\phi_r(o_t,z_t)|\mathcal{F}_t]=\mathbb{E}_{o_t\sim\text{Ber}(f_{r^*}(z_t))}[\phi_r(o_t,z_t)]=\bar{\phi}_r(z_t).$$

To apply Lemma E.1, we need the following lemmas:

Lemma C.1. With the same notations as Lemma B.5, we have

$$\text{Var}_o[\phi_r(o,z)]\leq(2B+4)\bar{\phi}_r(z).$$

Proof of Lemma C.1. It follows from Proposition 15 of Bakhtiari et al. (2026). □

With Lemmas B.5 and C.1, we can apply Lemma E.1 with $b=2B$ and $c=2B+4$. Therefore, with probability at least $1-\delta$, for any $t\in[T]$ and $r\in\mathcal{R}$, we have

$$\sum_{i=1}^{t-1}\bar{\phi}_r(z_i)\leq 2\sum_{i=1}^{t-1}\phi_r(o_i,z_i)+2\beta_t.$$

Here, we define

$$\beta_t=\frac{5T\epsilon}{2}+60(B+1)\log(N_\epsilon h_T/\delta),$$

where $N_\epsilon=\mathcal{N}\big(\Phi(\mathcal{R}),\epsilon,\|\cdot\|_\infty\big)$, $h_T=e+\log(1+T)$. Moreover, if we set $\epsilon=1/T$, we complete the proof of Lemma B.1. □

C.2 Proof of Lemma B.2

Proof of Lemma B.2. Since $r^*\in\text{Conf}_t$, we can use the optimality condition of (b_t,r_2^t) to bound the term $r^*(x_t,b_t)$, i.e.,

$$2r^*(x_t,a_t^*)-r_1^t(x_t,a_t^*)-2r^*(x_t,a_t)\leq 2r_2^t(x_t,b_t)-r_1^t(x_t,b_t)-2r_2^t(x_t,a_t).$$

Rearranging terms, we have

$$2r^*(x_t,a_t^*)\leq 2r_2^t(x_t,b_t)-r_1^t(x_t,b_t)-2r_2^t(x_t,a_t)+r_1^t(x_t,a_t^*)+2r^*(x_t,a_t). \tag{C.1}$$

Substituting (C.1) into the definition of the single-step regret,we have the per-step regret bound

$$\begin{aligned}
R_t&=2r^*(x_t,a_t^*)-r^*(x_t,a_t)-r^*(x_t,b_t)\\
&\leq\big[2r_2^t(x_t,b_t)-r_1^t(x_t,b_t)-2r_2^t(x_t,a_t)+r_1^t(x_t,a_t^*)+2r^*(x_t,a_t)\big]-r^*(x_t,a_t)-r^*(x_t,b_t)\\
&=\Delta_{r^*}(z_t)-2\Delta_{r_2^t}(z_t)+\big[r_1^t(x_t,a_t^*)-r_1^t(x_t,b_t)\big]\\
&\leq\Delta_{r^*}(z_t)-2\Delta_{r_2^t}(z_t)+\big[r_1^t(x_t,a_t)-r_1^t(x_t,b_t)\big]
\end{aligned}$$

$$
\begin{aligned}
&= \Delta_{r_1^t}(z_t) - 2\Delta_{r_2^t}(z_t) + \Delta_{r^*}(z_t) \\
&\le \underbrace{|\Delta_{r_1^t}(z_t) - \Delta_{r^*}(z_t)|}_{e_{1,t}} + 2\underbrace{|\Delta_{r^*}(z_t) - \Delta_{r_2^t}(z_t)|}_{e_{2,t}},
\end{aligned} \tag{C.2}
$$

where the first inequality holds due to (C.1), the second inequality holds due to the optimality of a_t in the design of the algorithm, which indicates $r_1^t(x_t, a_t^*) \le r_1^t(x_t, a_t)$, and the last inequality holds due to the triangle inequality. □

C.3 Proof of Lemma B.3

Proof of Lemma B.3. First, recall that in (A.2)

$$\bar{\phi}_r(z) = B_\psi(\Delta_r(z), \Delta_{r^*}(z)),$$

where B_ψ denotes the Bregman divergence generated by the logistic potential $\psi(v) := \log(1 + e^v)$. Note that $\psi'(v) = \sigma(v)$ and $\psi''(v) = \sigma'(v)$. Applying Taylor's theorem to ψ, we have

$$
\begin{aligned}
\psi\big(\Delta_r(z)\big) = \psi\big(\Delta_{r^*}(z)\big) &+ \psi'\big(\Delta_{r^*}(z)\big)\big(\Delta_r(z) - \Delta_{r^*}(z)\big) \\
&+ \tfrac{1}{2}\,\psi''(\xi)\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2,
\end{aligned} \tag{C.3}
$$

where ξ is a point on the segment between $\Delta_{r^*}(z)$ and $\Delta_r(z)$. Using the definition of Bregman divergence, we have

$$
\begin{aligned}
B_\psi\big(\Delta_r(z), \Delta_{r^*}(z)\big) &= \psi\big(\Delta_r(z)\big) - \psi\big(\Delta_{r^*}(z)\big) - \psi'\big(\Delta_{r^*}(z)\big)\big(\Delta_r(z) - \Delta_{r^*}(z)\big) \\
&= \Big[\psi\big(\Delta_{r^*}(z)\big) + \psi'\big(\Delta_{r^*}(z)\big)\big(\Delta_r(z) - \Delta_{r^*}(z)\big) + \tfrac{1}{2}\,\psi''(\xi)\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2\Big] \\
&\quad - \psi\big(\Delta_{r^*}(z)\big) - \psi'\big(\Delta_{r^*}(z)\big)\big(\Delta_r(z) - \Delta_{r^*}(z)\big) \\
&= \tfrac{1}{2}\,\psi''(\xi)\big(\Delta_r(z) - \Delta_{r^*}(z)\big)^2,
\end{aligned} \tag{C.4}
$$

where we use (A.2).

For the last inequality, note that $\psi'' = \sigma'$. Thus, the upper bound suffices $\psi''(\xi) \le 1/4$. When $M \ge \max\{|\Delta_r(z)|, |\Delta_{r^*}(z)|\}$, since ξ is on the segment between $\Delta_{r^*}(z)$ and $\Delta_r(z)$, i.e., there exists $\lambda \ge 0$ such that $\xi = \lambda\Delta_{r^*}(z) + (1-\lambda)\Delta_r(z)$, we have $|\xi| \le M$. We complete the proof by noting $\psi''(\xi) = \sigma'(\xi) \ge \sigma'(M)$. □

C.4 Proof of Lemma B.4

Proof of Lemma B.4. Note that $r_2^t \in \mathrm{Conf}_t$ by the algorithm design. Moreover, $r_1^t \in \mathrm{Conf}_t$ holds trivially. Therefore, on the event $\mathcal{E}$, for any $t \in [T]$ and $j \in \{1, 2\}$, we have

$$
\begin{aligned}
\sum_{i=1}^{t-1} \bar{\phi}_{r_j^t}(z_i) &\le 2\sum_{i=1}^{t-1} \phi_{r_j^t}(o_t, z_t) + 2\beta_t \\
&= 2\big(L_t(r_j^t) - L_t(r^*) + \beta_t\big) \\
&\le 2\Big(L_t(r_j^t) - L_t(r_1^t) + \beta_t\Big) \\
&\le 4\beta_t \le 4\beta_T,
\end{aligned} \tag{C.5}
$$

where the first inequality holds due to the definition of the high-probability event, the second inequality holds due to $r_1^t \in \operatorname{argmin}_{r\in\mathcal{R}} L_t(r)$, the third inequality holds because $r_j^t \in \mathrm{Conf}_t$, and the last inequality holds due to the monotonicity of β_t. □

C.5 Proof of Lemma B.5

Proof of Lemma B.5. Recall that

$$
\begin{aligned}
\phi_r(o, z) &= \ell\big(o, f_r(z)\big) - \ell\big(o, f_{r^*}(z)\big), \\
\ell\big(o, f_r(z)\big) &= -o\log\sigma\big(\Delta_r(z)\big) - (1-o)\log\sigma\big(-\Delta_r(z)\big).
\end{aligned}
$$

Therefore, we have

$$\phi_r(o,z) = o \cdot \log \frac{1+\exp(-\Delta_r(z))}{1+\exp(-\Delta_{r^*}(z))} + (1-o)\cdot \log \frac{1+\exp(\Delta_r(z))}{1+\exp(\Delta_{r^*}(z))}.$$

We notice that for any r, $\Delta_r(z) \in [-2B, 2B]$, so

$$-2B = \log \frac{1+\exp(-2B)}{1+\exp(2B)} \le \log \frac{1+\exp(-\Delta_r(z))}{1+\exp(-\Delta_{r^*}(z))} \le \log \frac{1+\exp(2B)}{1+\exp(-2B)} = 2B.$$

Similarly, we have

$$\log \frac{1+\exp(\Delta_r(z))}{1+\exp(\Delta_{r^*}(z))} \in [-2B, 2B].$$

Therefore, $\phi_r(o,z)$ can be bounded by

$$|\phi_r(o,z)| \le \max\left\{ \left| \log \frac{1+\exp(-\Delta_r(z))}{1+\exp(-\Delta_{r^*}(z))} \right|, \left| \log \frac{1+\exp(\Delta_r(z))}{1+\exp(\Delta_{r^*}(z))} \right| \right\} \le 2B.$$

□

D Dueling Bandits with Linear Reward

In this section, we consider the specialized linear reward function class, i.e.,

$$\mathcal{R} = \Big\{ r_{\boldsymbol{\theta}} = \langle \boldsymbol{\mu}(x,a), \boldsymbol{\theta} \rangle \Big| \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d \Big\},$$

where $\boldsymbol{\mu} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a fixed feature map. We assume $\|\boldsymbol{\mu}(x,a)\|_2 \le 1$ for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, and $\|\boldsymbol{\theta}\|_2 \le B$ for any $\boldsymbol{\theta} \in \Theta$. Consider the functor Φ and $\bar{\Phi}$ defined in (5.2). We will specify the uniform covering number of $\Phi(\mathcal{R})$ and the localized Eluder dimension of $\bar{\Phi}(\mathcal{R})$.

D.1 Covering Number

Proposition D.1. Under Definition xx with the uniform norm, the ϵ-covering number of $\Phi(\mathcal{R})$ can be upper bounded by

$$\mathcal{N}\big(\Phi(\mathcal{R}), \epsilon, \|\cdot\|_\infty\big) \le \left(1 + \frac{8B}{\epsilon}\right)^d.$$

In particular, when $\epsilon = 1/T$,

$$N_T = \mathcal{N}\big(\Phi(\mathcal{R}), 1/T, \|\cdot\|_\infty\big) \le \big(1 + 8BT\big)^d.$$

Proof of Proposition D.1. The proof directly follows from the Lipschitzness of Φ. To be more specific, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have

$$\begin{aligned}
\|\Phi(r_{\boldsymbol{\theta}}) - \Phi(r_{\boldsymbol{\theta}'})\|_\infty &= \sup_{o,z} \Big| \ell\big(o, \sigma[\Delta_{r_{\boldsymbol{\theta}}}(z)]\big) - \ell\big(o, \sigma[\Delta_{r_{\boldsymbol{\theta}'}}(z)]\big) \Big| \\
&= \sup_{o,z} \Big| \psi\big(\Delta_{r_{\boldsymbol{\theta}}}(z)\big) - \psi\big(\Delta_{r_{\boldsymbol{\theta}'}}(z)\big) - o\big(\Delta_{r_{\boldsymbol{\theta}}}(z) - \Delta_{r_{\boldsymbol{\theta}'}}(z)\big) \Big| \\
&\le \sup_{o,z} \Big| \psi\big(\Delta_{r_{\boldsymbol{\theta}}}(z)\big) - \psi\big(\Delta_{r_{\boldsymbol{\theta}'}}(z)\big) \Big| + \Big| \Delta_{r_{\boldsymbol{\theta}}}(z) - \Delta_{r_{\boldsymbol{\theta}'}}(z) \Big| \\
&\le \sup_{z} 2 \Big| \Delta_{r_{\boldsymbol{\theta}}}(z) - \Delta_{r_{\boldsymbol{\theta}'}}(z) \Big|,
\end{aligned}$$

where we use $\ell\big(o, \sigma(v)\big) = \psi(v) - ov$. The first inequality holds due to the triangle inequality. The last inequality holds due to $\psi(v) = \log(1+e^v)$ is 1-Lipschitz. Moreover, we have

$$\begin{aligned}
\sup_z 2 \Big| \Delta_{r_{\boldsymbol{\theta}}}(z) - \Delta_{r_{\boldsymbol{\theta}'}}(z) \Big| &= \sup_{x,a,b} 2 \Big| \langle \boldsymbol{\mu}(x,a) - \boldsymbol{\mu}(x,b), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \Big| \\
&\le 4 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.
\end{aligned}$$

Therefore, $\|\Phi(r_{\boldsymbol{\theta}}) - \Phi(r_{\boldsymbol{\theta}'})\|_\infty \le 4\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$. Using the standard arguments, i.e, let $\{\boldsymbol{\theta}_i\}_{i=1}^N$ be an $\epsilon/4$-cover of Θ under the Euclidean norm. Then, for every $\boldsymbol{\theta} \in \Theta$, there exists some $i \in [N]$ such that

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_i\|_2 \le \epsilon/4,$$

and hence

$$\|\Phi(r_{\boldsymbol{\theta}}) - \Phi(r_{\boldsymbol{\theta}_i})\|_\infty \le 4\|\boldsymbol{\theta} - \boldsymbol{\theta}_i\|_2 \le \epsilon.$$

Thus $\{\Phi(r_{\boldsymbol{\theta}_i})\}_{i=1}^N$ forms an ϵ-cover of $\Phi(\mathcal{R})$ under $\|\cdot\|_\infty$. Therefore,

$$\mathcal{N}\big(\Phi(\mathcal{R}), \epsilon, \|\cdot\|_\infty\big) \le \mathcal{N}\big(\Theta, \epsilon/4, \|\cdot\|_2\big).$$

Since $\Theta \subseteq B_2(0, B) \subset \mathbb{R}^d$, the standard Euclidean covering bound (see e.g. Vershynin (2018)) gives

$$\mathcal{N}\big(\Theta, \epsilon/4, \|\cdot\|_2\big) \le \left(1 + \frac{8B}{\epsilon}\right)^d.$$

Consequently,

$$\mathcal{N}\big(\Phi(\mathcal{R}), \epsilon, \|\cdot\|_\infty\big) \le \left(1 + \frac{8B}{\epsilon}\right)^d.$$

□

D.2 Localized Eluder Dimension

For $z = (x, a, b)$, we use shorthand notation $\widetilde{\boldsymbol{\mu}}(z) = \boldsymbol{\mu}(x, a) - \boldsymbol{\mu}(x, b)$. In this section, we consider

$$\bar{\Phi}(\mathcal{R}) := \Big\{ z \mapsto \mathrm{KL}\big(\mathrm{Ber}\left[\sigma(\langle\widetilde{\boldsymbol{\mu}}(z), \boldsymbol{\theta}^*\rangle)\right] \big\|\, \mathrm{Ber}\left[\sigma(\langle\widetilde{\boldsymbol{\mu}}(z), \boldsymbol{\theta}\rangle)\right]\big) : \boldsymbol{\theta} \in \Theta \Big\},$$

where $\boldsymbol{\theta}^* \in \Theta$. Without loss of generality, we can assume Θ is convex. We first prove an upper bound of the localized Eluder dimension of $\bar{\Phi}(\mathcal{R})$, which follows the proof of Proposition 4 in Bakhtiari et al. (2026). The main difference lies in that we consider the abstract action set $\mathcal{X} \times \mathcal{A} \times \mathcal{A}$ via a fixed feature map $\boldsymbol{\mu}$, while they consider $\mathcal{A} \in \mathbb{R}^d$.

We have the following result:

Proposition D.2. Assume $\Theta \subseteq \mathbb{R}^d$ is convex. $\boldsymbol{\theta}^* \in \Theta$. For any $\boldsymbol{\theta} \in \Theta$, $\|\boldsymbol{\theta}\|_2 \le B$. There exists a universal constant $C > 0$ such that for every

$$0 < \epsilon \le \eta \le \frac{\sigma'(2B)}{4},$$

we have

$$\dim_{\mathrm{elud}}^{\eta}\left(\epsilon; \bar{\Phi}(\mathcal{R})\right) \le Cd \log\left(1 + \frac{B^2}{\epsilon}\right).$$

To prove this result, we first need the following definition, first proposed in Bakhtiari et al. (2026).

Definition D.3 (Witness Sequence)**.** For a function class $\mathcal{F}$, $\omega > 0$, and a sequence of actions $(z_1, \ldots, z_k) \in (\mathcal{X} \times \mathcal{A} \times \mathcal{A})^k$. A sequence $(g_1, \ldots, g_k) \in \mathcal{F}^k$ is called an ω-witness sequence for $(a_1, \ldots, a_k)$ if for every $t \in \{1, \ldots, k\}$,

$$\sum_{i=1}^{t-1} g_t(a_i) \le \omega \text{ and } g_t(a_t) \ge \omega.$$

Lemma D.4 (Adapted from Bakhtiari et al. 2026, Lem. 36)**.** Let $(z_1, \ldots, z_k) \subseteq \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ be an ω-eluder sequence with respect to $\bar{\Phi}(\mathcal{R})$. Then there exists a witness sequence $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \subseteq \Theta$ such that

$$|\langle\widetilde{\boldsymbol{\mu}}(z_i), \boldsymbol{\theta}_t - \boldsymbol{\theta}^*\rangle| \le 2\sqrt{\omega/\sigma'(2B)}, \qquad i \le t \le k. \tag{D.1}$$

In particular, if $\omega \le 1/(4/\sigma'(2B))$, then

$$|\langle\widetilde{\boldsymbol{\mu}}(z_i), \boldsymbol{\theta}_t - \boldsymbol{\theta}^*\rangle| \le 1 \text{ for all } i \le t \le k.$$

For any ϵ-eluder sequence $\{a_1, \ldots, a_k\}$, we can find a witness for it by definition. Using Lemma D.4, we can find $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ another witness sequence such that

$$|\langle \widetilde{\boldsymbol{\mu}}(z_i), \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \rangle| \leq 1 \text{ for all } i \leq t \leq k.$$

The remaining proof follows the proof of Proposition 4 in Bakhtiari et al. (2026), by considering the feature map $\widetilde{\boldsymbol{\mu}}(z_j)$.

Proof of Corollary 5.5. If suffices to combine Proposition D.1 and Proposition D.2. □

D.3 Proof of Lemma D.4

Proof of Lemma D.4. Due to the definition, we can find $\{\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_k\}$ as a witness sequence of $(z_1, \ldots, z_k)$, i.e.,

$$\bar{\phi}_{r_{\boldsymbol{\theta}'_t}}(z_t) \geq \omega,$$
$$\sum_{i=1}^{t-1} \bar{\phi}_{r_{\boldsymbol{\theta}'_t}}(z_i) \leq \omega$$

Note that $\bar{\phi}_r(z)$ is convex in r and is minimized at $r_{\boldsymbol{\theta}^*}$, with $\bar{\phi}_{r_{\boldsymbol{\theta}^*}}(z) = 0$. Therefore, we can find $\lambda_t \in [0, 1]$ such that

$$\bar{\phi}_{r_{\boldsymbol{\theta}^*} + \lambda_t (r_{\boldsymbol{\theta}'_t} - r_{\boldsymbol{\theta}^*})}(z_t) = \omega.$$

Easy to see,

$$r_{\boldsymbol{\theta}_*} + \lambda_t (r_{\boldsymbol{\theta}'_t} - r_{\boldsymbol{\theta}^*}) = r_{(\boldsymbol{\theta}^* + \lambda_t (\boldsymbol{\theta}'_t - \boldsymbol{\theta}^*))}$$

Since Θ is convex, we have $\boldsymbol{\theta}_t := \boldsymbol{\theta}^* + \lambda_t (\boldsymbol{\theta}'_t - \boldsymbol{\theta}^*) \in \Theta$. By our construction, we have

$$\bar{\phi}_{r_{\boldsymbol{\theta}_t}}(z_t) = \omega.$$

Moreover, we have

$$\sum_{i=1}^{t-1} \bar{\phi}_{r_{\boldsymbol{\theta}_t}}(z_i) \leq \sum_{i=1}^{t-1} \bar{\phi}_{r_{\boldsymbol{\theta}'_t}}(z_i) \leq \omega.$$

Therefore, $\boldsymbol{\theta}_t$ satisfies the same condition as $\boldsymbol{\theta}'_t$. And thus, $\{\boldsymbol{\theta}_t\}$ is also a witness sequence.

Using Lemma B.3, we have

$$\begin{aligned} \omega \geq \bar{\phi}_{r_{\boldsymbol{\theta}_t}}(z_j) &\geq \frac{\sigma'(2B)}{2} \left| \Delta_{\boldsymbol{\theta}_t}(z_j) - \Delta_{\boldsymbol{\theta}^*}(z_j) \right|^2 \\ &= \frac{\sigma'(2B)}{2} \left| \langle \widetilde{\boldsymbol{\mu}}(z_j), \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \rangle \right|. \end{aligned}$$

This completes the proof of the lemma. □

E Auxiliary Lemmas

Lemma E.1 (Theorem 9 in Bakhtiari et al. (2026))**.** Let $\mathcal{Z}$ be a set, $\{Z_t\}_{t \in [T]}$ be a $\mathcal{Z}$ valued process adapted to a filtration $\{\mathcal{F}_t\}_{t \in [T]}$, and $\boldsymbol{\Phi}$ a set of real-valued functions on $\mathcal{Z}$. Assume that

1. For some $b > 0$, for all $\varphi \in \Phi$ and $t \in \mathbb{N}_+$, $|\mathbb{E}[\varphi(Z_t)|\mathcal{F}_{t-1}] - \varphi(Z_t)| \leq b$.
2. For some $c > 0$, for all $\varphi \in \Phi$ and $t \in \mathbb{N}_+$, $\mathrm{Var}[\varphi(Z_t)|\mathcal{F}_{t-1}] \leq c\mathbb{E}[\varphi(Z_t)|\mathcal{F}_{t-1}]$.

Let $\delta \in (0, 1)$, $\epsilon > 0$, and N be the ϵ-covering number of $\boldsymbol{\Phi}$ in the uniform metric. For any $n \in \mathbb{N}_+$, define

$$\beta(n, \delta, \epsilon, N) = \frac{5n\epsilon}{2} + 15(b + c) \log(N h_n / \delta),$$

where $h_n = e + \log(1 + n)$. Then with probability at least $1 - \delta$, for all $\varphi \in \Phi$ and $n \in \mathbb{N}_+$,

$$\sum_{t=1}^{n} \mathbb{E}[\varphi(Z_t)|\mathcal{F}_{t-1}] \leq 2 \sum_{t=1}^{n} \phi(Z_t) + 2\beta(n, \delta, \epsilon, N).$$

Lemma E.2 (Lemma 24 in Bakhtiari et al. (2026))**.** Fix $B, \beta > 0$ and $0 < u \leq B$. Let $\mathcal{X}$ be a set, and Ψ a set of $[0, B]$-valued functions on $\mathcal{X}$. Suppose sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ and $(\psi_1, \ldots, \psi_n) \in \Psi^n$ satisfy that for all $t \in [n]$,

$$\sum_{i=1}^{t-1} \psi_t(x_i) \leq \beta.$$

Then the following hold:

1. For any $\epsilon \in (0, u]$ and any $t \in [n]$,

$$\sum_{i=1}^{t} \mathbb{1}[\psi_i(x_i) > \epsilon] \leq (1 + \beta/\epsilon) \dim^u_{\mathrm{elud}}(\epsilon, \Psi) + 1.$$

2. If $\psi_t(x_t) \leq u$ for all $t \in [n]$, then for any $\omega \in (0, u]$ and any $t \in [n]$,

$$\sum_{i=1}^{t} \psi_i(x_i) \leq \big(B + \beta \log(1 + B/\omega)\big) \dim^u_{\mathrm{elud}}(\omega, \Psi) + B + t\omega.$$

F Connections to Prior Work

F.1 Saturation Dependence in Existing Guarantees

Global inverse-curvature dependence. For the BTL model, $\sigma'(v) = \sigma(v)(1 - \sigma(v))$. When rewards lie in $[-B, B]$, the worst-case inverse curvature is

$$\kappa_B := \sup_{|v| \leq 2B} \frac{1}{\sigma'(v)} = \frac{1}{\sigma'(2B)} = 2 + e^{2B} + e^{-2B}.$$

Thus, replacing local derivatives by a uniform lower bound can introduce exponential dependence on the reward scale. Analyses of contextual preference and dueling bandits that use global derivative bounds include Saha (2021); Bengs et al. (2022); Di et al. (2024a). Reward-based guarantees for general or policy-induced reward classes also exhibit saturation dependence, as discussed in Li et al. (2024); Xie et al. (2025); Xiong et al. (2024). The precise dependence differs across these works, but the use of worst-case curvature is a common source of unfavorable reward-scale dependence.

Local curvature in logistic and linear dueling bandits. In logistic bandits, a line of work improves this dependence by retaining local curvature (Faury et al., 2020; Abeille et al., 2021; Faury et al., 2022). For feature vectors $\mathbf{v}_s$ and a linear parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, the regularized Hessian is

$$H_t(\boldsymbol{\theta}) = \lambda I + \sum_{s<t} \sigma'(\mathbf{v}_s^\top \boldsymbol{\theta}) \mathbf{v}_s \mathbf{v}_s^\top,$$

where $\lambda > 0$. Its weights preserve the curvature at individual observations, allowing confidence sets and exploration bounds to adapt to local derivatives. For example, the Logistic-UCB-2 algorithm of Faury et al. (2020) achieves

$$R_T^{\mathrm{logistic}} = \widetilde{O}\Big(d\sqrt{T} + \kappa_{\log} d^2\Big),$$

where $\kappa_{\log}$ is the worst-case inverse derivative in their logistic-bandit model; the display suppresses dependence on the fixed parameter-radius bound as well as logarithmic factors. Here regret is measured in expected Bernoulli rewards. The worst-case curvature appears in an additive term, rather than multiplying the leading $\sqrt{T}$ term. Related advances for generalized linear models include Lee et al. (2024); **?**.

In linear contextual dueling bandits, Di et al. (2025) use estimates of the local link derivative to construct a derivative-weighted covariance matrix. For sigmoid feedback, their RCDB-S algorithm has a leading term

$$\widetilde{O}\Big(dB^{3/2}\sqrt{T}\Big)$$

without the worst-case inverse-derivative factor; their full bound also accounts for adversarial feedback. These methods exploit a known linear parameterization and matrix arguments such as elliptical-potential bounds. Such tools are not directly available for general reward classes. Our analysis instead uses the complexity of the induced excess-loss class and a regret decomposition tailored to the two queried actions.

Preference-based exploration in RLHF. Chen et al. (2026) take a different route through preference-based exploration and an iteratively updated comparator sampler. For a fixed sampler π_{sam}, their POPO subroutine controls the preference-based regret

$$\mathrm{Reg}_{\mathrm{pref}}(\pi_{\mathrm{sam}}, T) = \sum_{t=1}^{T} \mathbb{E}[P^*(y^* \succ y' \mid x) - P^*(y \succ y' \mid x)],$$

where x follows the prompt distribution and, conditionally on x, $y^* \sim \pi^*$, $y \sim \pi_t$, and $y' \sim \pi_{\mathrm{sam}}$. Here P^* is the true preference probability and π^* is the reward-optimal policy. Their SE-POPO algorithm updates the sampler between successive calls to POPO and uses a preference-to-reward reduction to obtain reward guarantees with polynomial dependence on the reward scale. Thus, preference-based regret is an intermediate tool in their analysis. Our result instead controls cumulative two-armed latent-reward regret through the asymmetric selection rule and query-gap localization.

F.2 Reduction to the Linear Case

Consider the linear class in Example 3.2. Write $r_1^t = r_{\widehat{\theta}_t}$ and keep its greedy target a_t fixed. In parameter space, the loss-based confidence set is

$$\{\theta \in \Theta : L_t(r_\theta) \le L_t(r_{\widehat{\theta}_t}) + \beta_t\}.$$

To expose the connection with an explicit bonus, suppose that this set is replaced by the full ellipsoid

$$\mathrm{Conf}_t = \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|_{M_t} \le \rho_t\}, \qquad M_t \succ 0.$$

Here $\|v\|_{M_t} = (v^\top M_t v)^{1/2}$, and ρ_t denotes the ellipsoid radius, not the empirical-loss threshold β_t. For $\mathbf{d}_t(b) = \boldsymbol{\mu}(x_t, b) - \boldsymbol{\mu}(x_t, a_t)$, the comparator objective is

$$\begin{aligned} J_t(b, r_\theta) &= 2\langle \mathbf{d}_t(b), \theta - \widehat{\theta}_t\rangle + \langle \boldsymbol{\mu}(x_t, b), \widehat{\theta}_t\rangle - 2\langle \boldsymbol{\mu}(x_t, a_t), \widehat{\theta}_t\rangle \\ &\le 2\rho_t \|\mathbf{d}_t(b)\|_{M_t^{-1}} + r_1^t(x_t, b) - 2r_1^t(x_t, a_t). \end{aligned}$$

The last step is the Cauchy–Schwarz inequality in the M_t norm. Maximizing over the full ellipsoid attains this upper bound. Dropping the term independent of b therefore yields the comparator score

$$r_1^t(x_t, b) + 2\rho_t \|\boldsymbol{\mu}(x_t, b) - \boldsymbol{\mu}(x_t, a_t)\|_{M_t^{-1}}.$$

With an unweighted Gram matrix, this has the estimated-reward-plus-pairwise-width form of the CoLSTIM comparator (Bengs et al., 2022). Using a derivative-weighted matrix gives the corresponding local-curvature bonus form used by RCDB-S (Di et al., 2025). The original loss-based confidence set retains curvature through the log loss instead of replacing it with an explicitly chosen ellipsoid.

The comparison concerns the comparator score under the stated ellipsoidal replacement, not equality of the confidence sets or identity of the complete algorithms. Thus, the connection explains the implicit curvature-aware exploration in our selection rule; the localized analysis establishes its regret guarantee in the general-class formulation.

G Practical implementation of SI−CDB

We outline a practical implementation of SI−CDB that approximates its optimization steps using gradient-based updates and alternating optimization.

Reward-model update. For a parameterized nonlinear reward model, a gradient-based optimizer such as Adam can be warm-started from the previous round's model. A fixed number of updates offers a computationally inexpensive approximation to solving the cumulative-loss problem from scratch. A replay sample can also be used to approximate the historical loss, with rescaling to the cumulative-loss scale.

Target-action update. For the finite action set in our model, the target can be found by evaluating the current reward model on all candidate actions. For a large or continuous differentiable action space, a practical extension is to sample initial actions, retain the highest-scoring ones, and run multiple gradient-ascent trajectories in action space. The best resulting candidate is retained.

Comparator update. Fix the target a_t and the fitted reward model r_1^t, and write

$$g_t(r) = L_t(r) - L_t(r_1^t) - \beta_t.$$

The comparator problem maximizes $J_t(b, r)$ subject to $g_t(r) \leq 0$. Initialize the candidate model at r_1^t, which is feasible for the loss-based constraint, and initialize the comparator action. Then alternate between updating the model with the action fixed and updating the action with the model fixed. Retain only candidates that satisfy the confidence constraint.

The action update can use enumeration, or the multi-start procedure above when applicable. For the model update, the penalized objective proposed for minimization is

$$-J_t(b, r) + \lambda g_t(r) + \frac{\rho}{2}[g_t(r)]_+^2, \qquad [v]_+ = \max\{v, 0\},$$

with multiplier $\lambda \geq 0$ and penalty parameter $\rho > 0$. After a proposed update, check the confidence constraint. If it is violated, backtracking along the parameter update can be used to search for a feasible step. The multiplier update is

$$\lambda \leftarrow \max\{0, \lambda + \eta_\lambda g_t(r)\},$$

where $\eta_\lambda > 0$ is a step size.