On the Complexity of Preference-Based Bandits

Ahmed Ben Yahmed*
Criteo AI Lab, Paris, France
CREST, ENSAE, France
FairPlay joint team
a.benyahmed@criteo.com

Marc Abeille
Criteo AI Lab, Paris, France
FairPlay joint team
m.abeille@criteo.com

Clément Calauzènes
Criteo AI Lab, Paris, France
FairPlay joint team
c.calauzenes@criteo.com

Abstract

We study preference-based bandits with general reward function classes, where a learner sequentially selects pairs of arms and observes binary preference feedback governed by the Bradley–Terry model. This setting naturally arises in applications such as recommender systems, tournament ranking, and learning from human feedback, where relative preferences are easier to elicit than absolute rewards. The observation model inherits the logistic bandit challenge of handling the problem-dependent constant κ, which accounts for the non-linearity of the link function and can grow arbitrarily large. Moreover, prior work has predominantly focused on linear or kernelized reward models, precluding the use of richer function classes. To address these limitations, we consider general reward function classes and introduce the *locally sensitive eluder dimension*, a novel complexity measure tailored to the logistic structure of preference feedback that yields fine-grained regret guarantees without unfavorable dependence on κ. Building on this notion, we propose **GINOP** (Generic INformative OPtimism), an algorithm that constructs log-loss confidence sets and jointly selects arm pairs to balance optimism and informative exploration. We establish a first-order regret bound that, in contrast with what previous results suggest, demonstrates that learning with preference feedback is as statistically efficient as learning from direct reward observation. Finally, we corroborate our theoretical findings with empirical evaluations against competitive baselines.

1 Introduction

In contrast to classical stochastic bandits, where the learner plays arms and receives direct reward feedback (Lattimore and Szepesvári, 2020), dueling bandits – also referred to as preference-based bandits – constitute a sequential decision-making framework in which the learner plays pairs of arms (duels) and observes noisy preference feedback (Bengs et al., 2021). Specifically, at each round, the learner receives a binary signal indicating which arm in the selected pair is preferred. Preference-based bandits naturally arise in real-world applications where eliciting absolute rewards is difficult, while pairwise comparisons are more natural and reliable. Notable examples include online recommendation systems – where users tend to provide more accurate relative preferences than absolute ratings – web search ranking, fine-tuning of large language models, and comparative evaluations such as choosing between two restaurants or movies. More broadly, this framework is well-suited to any setting in which relative judgments are easier to obtain than absolute reward values.

When the decision space – comprising, for instance, items in online platforms, search results, or model parameters in large language models – is large or even infinite, it is common to assume that the reward function is parameterized by an unknown mapping, typically taken to be linear (Saha, 2021; Bengs et al., 2022; Li et al., 2024) or belonging to a reproducing kernel Hilbert space (RKHS) associated with a given kernel (Xu et al., 2024; Pásztor et al., 2024; Kayal et al., 2025). However, such parametrizations may be overly restrictive in practice and can limit the expressiveness of the

*Correspondence to: a.benyahmed@criteo.com.

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

model. To address this limitation, the present paper considers a more general setting in which the reward function belongs to a rich and generic function class, and must be learned from preference feedback obtained through pairwise comparisons of selected arms.

2 Setting

Notation. The set $\{1, \dots, T\}$ is denoted by $[T]$. The sigmoid function is defined as $\sigma(x) := (1 + e^{-x})^{-1}$, and $\dot{\sigma}$ and $\ddot{\sigma}$ denote its first and second derivatives, respectively. The log-loss function is defined as $\ell(a, b) := -a \log(b) - (1 - a) \log(1 - b)$. The inner product between two vectors u, v is denoted $\langle u, v \rangle$. We define $\|x\| := \sqrt{\langle x, x \rangle}$ and, for any positive definite matrix $V \in \mathbb{R}^{d \times d}$, $\|x\|_V := \sqrt{\langle x, Vx \rangle}$, as the Euclidean and weighted Euclidean norms of x, respectively. We denote by $\mathbb{B}_2^d(r)$ the d-dimensional Euclidean ball of radius r centered at the origin. Standard asymptotic notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ is used throughout, while $\widetilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors.

Problem formulation. We consider a stochastic dueling bandit problem defined by a decision set $\mathcal{X}$ and a reward function class $\mathcal{F}$. At each round, the learner selects a pair of arms – referred to as a *duel* – from $\mathcal{X}$ and observes preference feedback over the selected arms. This formulation contrasts with the classical stochastic bandit setting, in which the learner selects a single arm $x \in \mathcal{X}$ at each round t and observes a stochastic absolute reward $r_t(x)$. In dueling bandits, at round t, the learner selects a pair of arms $(x_t, x_t') \in \mathcal{X}^2$. Rather than observing the individual rewards $r_t(x_t)$ and $r_t(x_t')$, the learner receives stochastic preference feedback $y_t \in \{0, 1\}$, where $y_t = 1$ indicates that arm x_t is preferred over x_t', and $y_t = 0$ otherwise. Yet, the objective is to cumulate as much reward as possible. We impose the following realizability and boundedness conditions, common in the stochastic setting.

Assumption 2.1 (Realizability)**.** $\exists f^\star \in \mathcal{F},\ \forall t \in [T],\ \forall x \in \mathcal{X},\quad f^\star(x) = \mathbb{E}[r_t(x) \mid x]$.

Assumption 2.2 (Bounded reward class)**.** $\exists S > 0,\ \forall x \in \mathcal{X},\ \forall f \in \mathcal{F},\quad f(x) \in [-S, S]$.

Observation model. The link between the preference feedback and the underlying rewards is assumed to follow the Bradley–Terry (BT) model (Bradley and Terry, 1952), a standard modeling choice in the dueling bandits literature (Saha, 2021; Bengs et al., 2022; Li et al., 2024). Under this model, the probability that arm x_t is preferred over x_t' is given by

$$\mathbb{P}(x_t \succ x_t') = \mathbb{P}(y_t = 1 \mid x_t, x_t') = \sigma(f^\star(x_t) - f^\star(x_t')),$$

where $x_t \succ x_t'$ denotes that arm x_t is preferred over arm x_t'. While the BT model provides a tractable link between rewards and preferences, it introduces a problem-dependent constant

$$\kappa := \sup_{x, x' \in \mathcal{X}} \frac{1}{\dot{\sigma}(f^\star(x) - f^\star(x'))},$$

which captures the degree of non-linearity of the sigmoid function. This quantity can be prohibitively large: it scales as $\kappa \approx e^{2S}$, so that even $S = 5$ already yields $\kappa > 2.2 \times 10^4$. Consequently, it requires careful treatment in both the algorithm design and the theoretical analysis in order to avoid a detrimental impact on the resulting regret guarantees.

Performance measure. After selecting a pair of arms (x_t, x_t') at round t, the learner incurs an instantaneous regret. Two standard notions of instantaneous regret in the dueling bandits literature (Saha, 2021; Bengs et al., 2022; Li et al., 2024) are the *average* and *weak* instantaneous regrets, defined respectively as $\rho_t^a := f^\star(x^\star) - \frac{f^\star(x_t) + f^\star(x_t')}{2}$ and $\rho_t^w := f^\star(x^\star) - \max\{f^\star(x_t),\ f^\star(x_t')\}$, where $x^\star := \arg\max_{x \in \mathcal{X}} f^\star(x)$ denotes the optimal arm. After T rounds, the cumulative regret of a policy is defined as $\mathcal{R}_T^\tau := \sum_{t=1}^T \rho_t^\tau,\ \tau \in \{a, w\}$. A desirable policy achieves sublinear regret, i.e., $\lim_{T \to \infty} \frac{\mathcal{R}_T^\tau}{T} = 0$, which implies that the policy asymptotically identifies the optimal arm and selects it consistently in comparisons. By definition, $\mathcal{R}_T^w \leq \mathcal{R}_T^a$. Therefore, throughout the remainder of the paper, we restrict attention to the average cumulative regret

$$\mathcal{R}_T := \mathcal{R}_T^a = \sum_{t=1}^T \left(f^\star(x^\star) - \frac{f^\star(x_t) + f^\star(x_t')}{2} \right).$$

3 Related Work and Contributions

Dueling bandits. Learning with indirect feedback was first studied in supervised preference learning (Aiolli and Sperduti, 2004; Chu and Ghahramani, 2005), and later extended to online and

sequential settings motivated by applications such as human-provided feedback (Yue and Joachims, 2009; Yue et al., 2012; Houlsby et al., 2011). We refer the readers to Bengs et al. (2021) for a detailed survey on various works covering different settings of dueling bandits.

An initial line of work considers finite multi-armed domains and learns a preference matrix specifying pairwise relations among arms. Such approaches often rely on efficient sorting or tournament schemes based on win frequencies per action (Jamieson and Nowak, 2011; Zoghi et al., 2014b; Komiyama et al., 2015; Wu and Liu, 2016; Falahatgar et al., 2017). Rather than selecting both arms jointly, these strategies simplify the problem by choosing one arm at random (Zoghi et al., 2014a; Zimmert and Seldin, 2018), greedily (Chen and Frazier, 2017), or from the set of previously selected arms (Ailon et al., 2014). Building on an online square-loss oracle, Saha and Krishnamurthy (2022) propose an efficient and optimal algorithm when the preference matrix is well-specified by a given function class.

To handle large or even infinite decision spaces, an alternative paradigm has emerged, namely *utility-based dueling bandits*, to which the present work belongs. A prominent line of research adopts linear utility models. Saha (2021) and Bengs et al. (2022) both employ UCB-style estimators but with different decision schemes: the former constructs a set of plausible winners and selects the most informative pair among them, while the latter greedily selects the first arm and then chooses a competitive second arm to duel. Both achieve a regret of $\widetilde{\mathcal{O}}(\kappa d\sqrt{T})$, where d is the dimension of the arm space. Similarly, Di et al. (2024) present a variance-aware algorithm that proceeds in phases, maintaining an active set of arms via a UCB-style criterion, and achieves $\widetilde{\mathcal{O}}\left(d\kappa\sqrt{\sum_{t=1}^{T}\eta_t^2} + d\kappa^2\right)$, where $\{\eta_t\}$ denotes the noise variance sequence. Alternatively, Li et al. (2024) propose a Thompson sampling method that also achieves $\widetilde{\mathcal{O}}(\kappa d\sqrt{T})$. Di et al. (2025) propose RCDB, an algorithm that attains nearly minimax-optimal regret for dueling bandits in the presence of adversarial preference feedback. Furthermore, for the sigmoid link function, they develop RCDB-S that incorporates the effect of local derivatives into the maximum likelihood estimation analysis via a refined estimator of the link function's derivative which removes the dependence on κ from the regret bound in the linear setting. However, such linear methods have limited practical appeal - as they cannot capture the complex nonlinear utility functions arising in real-world problems - motivating the extension to Reproducing kernel Hilbert spaces (RKHS). Considering a kernelized logistic negative log-likelihood loss to estimate the utility, but with different decision rules, Xu et al. (2024) achieve a regret of $\widetilde{\mathcal{O}}\big((\gamma_T T)^{3/4}\big)$ – where γ_T denotes the maximum information gain – by retaining one arm from the previous round and selecting the second as the maximizer of a UCB on the pairwise preference. Pásztor et al. (2024) adopt a game-theoretic arm-selection strategy and achieve a regret of $\widetilde{\mathcal{O}}(\gamma_T\kappa^2\sqrt{T})$. Building on a more involved phased approach, Kayal et al. (2025) removed the dependence on κ compared to Pásztor et al. (2024). Finally, the recent work of Verma et al. (2025) considers neural dueling bandits in the wide-network regime (NTK), leveraging neural tangent features for preference prediction to achieve a regret of $\widetilde{\mathcal{O}}\big(\kappa\tilde{d}\sqrt{T}\big)$, with $\tilde{d}$ the effective dimension.

It is worth noting that preference-based feedback has played a central role in integrating human feedback into learning systems, yielding notable successes across a range of applications (Stiennon et al., 2022; Bai et al., 2022; Saha et al., 2023; Zhu et al., 2023; Ji et al., 2023; Munos et al., 2024).

Bradley-Terry preference model and logistic bandits. The Bradley-Terry (BT) model is a seminal framework to derive binary preferences from absolute scores: for items with scores v_1 and v_2, the preference probability is $\sigma(v_1 - v_2)$ where σ is the sigmoid function $x \mapsto (1+e^{-x})^{-1}$. Consequently, dueling bandits under the BT assumption inherit the technical challenges of logistic bandits. Early logistic bandit analyses relied on global linearization (Filippi et al., 2010), yielding regret bounds of $\mathcal{O}(\kappa\sqrt{T})$, where κ can scale exponentially with the ambient dimension. In the linear utility setting, recent advances Faury et al. (2020); Abeille et al. (2021); Faury et al. (2022) demonstrate that a careful treatment of logistic non-linearity leads to an improved regret bound of $\mathcal{O}(\sqrt{T/\kappa})$. Whether such improvements are intrinsically tied to the linear structure of the utility function, or can be extended to more general classes of reward functions, remains an open question.

Eluder Dimension. The key challenge in going beyond linear and kernelized mappings is the need to design sets of plausible models compatible with observed data – that is, to establish concentration guarantees – and to quantify how their width translates into online prediction error. The seminal work of Russo and Van Roy (2013) addresses both challenges, relating the former to covering numbers and the latter to a newly introduced complexity measure—the *eluder dimension*—which is further

studied and characterized in Li et al. (2022).This notion has been employed by Sekhari et al. (2023) in the context of preference-based bandits. However, it is primarily developed for additive noise models (e.g., building on least-squares regression), which precludes extending logistic bandits to generic function classes while preserving the sharp non-linearity treatment introduced by Faury et al. (2020). Recently, Bakhtiari et al. (2025) attempted to reconcile both approaches by introducing a *localized eluder dimension* defined for arbitrary loss functions (in line with Liu et al. (2022), as opposed to the squared loss). They further show—independently of the loss function—that a global eluder dimension necessarily entails an unfavorable dependence on κ. This observation motivates their localization approach, evaluating the eluder dimension only over a suitable subfamily $\mathcal{F}' \subset \mathcal{F}$. The localized eluder dimension is independent of κ but applies only when the confidence set is small enough, yielding an additional regret term quantifying the regret incurred before localization. Unfortunately, this regret term is uncontrolled in general and when instantiated to generalized linear models turns out to be linear in T, rendering the approach vacuous (see Section A for details).

Main contributions. This work is the first to study the complexity of the preference bandit setting with general utility classes, introducing a novel complexity measure tailored to yield fine-grained regret guarantees. Our contributions may be summarized as follows.

1. **Locally sensitive eluder dimension.** The lower bound established in (Bakhtiari et al., 2025, Theorem 2) shows that the global eluder dimension of Russo and Van Roy (2013) style necessarily scales with κ in the generalized linear setting, leading to an undesirable dependence that motivates the introduction of a refined complexity measure. We therefore introduce a new notion of eluder dimension (Definition 5.1), tailored to the preference feedback setting. This quantity properly accounts for the non-linearity induced by the logistic loss, and thereby avoids any unfavorable dependence on κ.
2. **GINOP algorithm (Section 5).** We propose a Generic INformed-OPtimistic strategy for preference bandits with general reward function classes. The algorithm takes the reward class as input and iteratively constructs adequate confidence sets, which are then leveraged to select actions (x_t, x'_t) in a correlated manner, balancing the dual objectives of optimistically collecting reward and enforcing dissimilarity to gather more informative comparisons.
3. **Regret guarantees (Theorem 5.1).** Through a careful analysis, we demonstrate that, over a time horizon T, **GINOP** achieves a first-order regret bound that is, up to logarithmic factors, of order
$$\sqrt{\underbrace{d_\sigma(\Delta\mathcal{F},\, \Delta_{f^\star},\, 1/T^2)}_{\text{locally sensitive eluder dimension}} \quad \underbrace{\log\left(\mathcal{N}_T\right)}_{\text{log-covering number}} \quad \frac{T}{\dot{\sigma}^\star}} + \kappa\, \Gamma_T,$$
where $\dot{\sigma}^\star$ denotes the local curvature of the sigmoid at the optimal choice and is a well-behaved constant, and Γ_T is a lower-order instance-dependent term. We further instantiate our results to the linear and kernelized reward function classes, and show that our eluder dimension introduces no additional κ dependence, thereby demonstrating that our approach surpasses previous work in both generality and sharpness (Section 6). From learning-theoretic standpoint, Theorem 5.1 shows that learning from preference feedback is as statistically efficient as learning from direct reward observations.
4. **Empirical evaluation (Section 7).** We corroborate our theoretical findings with empirical evaluations, comparing **GINOP** against several existing baselines.

4 Learning Process

From the history of observations $\mathcal{H}_t = \{x_1, x'_1, y_1, \ldots, x_t, x'_t, y_t\}$, the learner constructs estimators of the underlying reward function. Since rewards are only revealed through preferences, standard least-squares estimation is not applicable, precluding a direct use of Russo and Van Roy (2013)'s results. We instead leverage a procedure based on the log-loss, which yields refined guarantees.

Under the BT model, the preference feedback depends on the reward difference $f^\star(x) - f^\star(x')$ associated with a pair (x, x'), where $f^\star \in \mathcal{F}$. We introduce the *difference operator* Δ_f, the induced function class $\Delta\mathcal{F}$, and the corresponding excess loss class $\Phi(\Delta\mathcal{F})$:
$$\Delta_f : (x, x') \in \mathcal{X}^2 \mapsto f(x) - f(x') \in [-2S,\, 2S], \qquad \Delta\mathcal{F} := \big\{\Delta_f : f \in \mathcal{F}\big\},$$
$$\Phi(\Delta\mathcal{F}) := \Big\{(y, x, x') \mapsto \ell\big(y,\, \sigma(\Delta_f(x, x'))\big) - \ell\big(y,\, \sigma(\Delta_{f^\star}(x, x'))\big) : f \in \mathcal{F}\Big\}.$$

Equipped with these definitions, the true reward function $f^\star$ is estimated from $\mathcal{H}_{t-1}$ via maximum likelihood estimation (MLE). For all $t \in [T]$,

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{L}(f;\, \mathcal{H}_{t-1}), \tag{1}$$

where the negative log-likelihood is defined as $\mathcal{L}(f;\, \mathcal{H}_{t-1}) := \sum_{s=1}^{t-1} \ell\Big(y_s,\, \sigma\big(\Delta_f(x_s, x_s')\big)\Big)$.

We build confidence sets $\mathcal{F}_t \subset \mathcal{F}$ consisting of functions close to $\hat{f}_t$ under the log-loss. For all $t \geq 1$,

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \mathcal{L}(f; \mathcal{H}_{t-1}) \leq \mathcal{L}(\hat{f}_t; \mathcal{H}_{t-1}) + \beta_t(\delta) \right\},$$

$$\text{where} \quad \beta_t(\delta) = \frac{5}{2} + 60(2S+1)\, \log\left(\frac{\mathcal{N}_T(\Phi(\Delta\mathcal{F}))\big(e + \log(1+t)\big)}{\delta} \right) = \mathcal{O}\big(\log \frac{\mathcal{N}_T(\Phi(\Delta\mathcal{F}))}{\delta}\big) \tag{2}$$

The sequence $\{\beta_t(\delta)\}_t$ is non-decreasing and defines the confidence radius. It captures the complexity of the problem through $\mathcal{N}_T(\Phi(\Delta\mathcal{F}))$, the $1/T$-covering number of the class $\Phi(\Delta\mathcal{F})$ under the uniform metric. Concentration results of the form (2) hold naturally for finite function classes, with confidence widths of order $\beta_t(\delta) = \mathcal{O}(\log(|\mathcal{F}|/\delta))$. Using the covering number allows these results to be extended to infinite function classes. Lemma 4.1 - adapted from (Bakhtiari et al., 2025, Proposition 22) - ensures $\{\mathcal{F}_t\}_{t\geq 1}$ is a valid sequence of confidence sets.

Lemma 4.1. *Let $\mathcal{F}_t$ be defined as in Equation* (2) *for all $t \geq 1$. Under Assumption 2.2, with probability at least $1-\delta$, it holds that $f^\star \in \bigcap_{t\geq 1} \mathcal{F}_t$.*

Finally, while $\mathcal{F}_t$ contains models that are compatible with the observations up to time t (small empirical risk), we are also interested in the prediction error associated with $\mathcal{F}_t$. We define the worst-case uncertainty in the performance gap between two arms x, x' by

$$\omega_t(x, x') = \sup_{f, f' \in \mathcal{F}_t} \big(\Delta_f(x, x') - \Delta_{f'}(x, x')\big). \tag{3}$$

5 Algorithm and Main Results

GINOP: Generic INformative OPtimism for Preference Bandits

1 **Inputs:** $\mathcal{X}$, $\mathcal{F}$, $\{\beta_t\}$, $\delta = \frac{1}{T^2}$, $\mathcal{H}_0 = \emptyset$.
2 **for** $t = 1, \ldots, T$ **do**
3 Construct $\hat{f}_t$, ω_t from (1), (2) and (3).
4 Play: $x_t, x_t' = \operatorname*{argmax}_{x, x' \in \mathcal{X}} \hat{f}_t(x) + \hat{f}_t(x') + \omega_t(x, x')$.
5 Observe feedback y_t.
6 Update history: $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{x_t, x_t', y_t\}$.
7 **end**

We introduce the **GINOP** algorithm for preference bandits with a general reward function class. At each round t, the algorithm leverages the available historical data to compute the MLE $\hat{f}_t$ together with an associated uncertainty measure ω_t (Step 3). It then selects a pair of arms (x_t, x_t') that maximizes an *informative optimism* criterion (Step 4), which augments the estimated cumulative reward of the pair, $\hat{f}_t(x) + \hat{f}_t(x')$, with a positive bonus $\omega_t(x, x')$ that explicitly quantifies the uncertainty in the *reward difference* between the two arms, thereby promoting targeted exploration. Crucially, since $\omega_t(x, x) = 0$, the algorithm is structurally discouraged from selecting identical arms, particularly in the early rounds. In contrast to standard bandit settings with direct feedback, in which the exploration bonus typically aggregates the uncertainties of individual arms, the bonus employed here is specifically tailored to the dueling structure by prioritizing comparisons between *distinct* arms. A distinguishing feature of **GINOP** lies in its decision rule, which jointly selects both dueling arms in a single optimization step. This stands in contrast to prior approaches, which typically proceed in two stages: either (i) following a *leader–follower* paradigm, in which a leader arm is selected first and the follower is chosen conditionally on the leader (Bengs et al., 2022; Li et al., 2024; Xu et al., 2024; Pásztor et al., 2024; Verma et al., 2025); or (ii) by first constructing, at each round, a set of *plausible winners* and subsequently selecting the dueling pair from within this set (Saha, 2021; Di et al., 2024; Kayal et al., 2025).

Locally sensitive eluder dimension. As demonstrated by Russo and Van Roy (2013), traditional complexity measures for function classes – such as the VC dimension and covering numbers– do not suffice to analyze bandit settings. To address this limitation, the authors introduce the *eluder dimension*, which quantifies how the *learning error* relates to the *on-policy prediction error* - informally, how $w_t(x_t, x_t')$ reduces as $\mathcal{F}_t$ shrinks. However, their definition is tailored to additive noise models and does not correctly accommodate Bernoulli preference feedback. To overcome this limitation, we introduce the *locally sensitive eluder dimension*, a novel complexity notion better suited to the preference bandit setting.

Definition 5.1 ((σ, ε)-eluder dimension)**.** *Let $\mathcal{Z}$ be a set and $\mathcal{Q} = \{q : \mathcal{Z} \to \mathbb{R}\}$. Fix $q^\star \in \mathcal{Q}$ and let $\varepsilon > 0$. Define $\bar{\varphi}$ the excess log-loss as $\bar{\varphi}(a, b) = \ell(a, b) - \ell(a, a)$.*

1. *Given a sequence $\bar{z} = (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$, we say that $z \in \mathcal{Z}$ is (σ, ε)-independent from $\bar{z}$ w.r.t. $(\mathcal{Q}, q^\star)$ if there exist $q \in \mathcal{Q}$ s.t.*

$$\sum_{j \le n} \bar{\varphi}\big(\sigma(q^\star(z_j)), \sigma(q(z_j))\big) \le \varepsilon^2 \tag{4}$$

and

$$\dot{\sigma}(q^\star(z))\,(q^\star(z) - q(z))^2 > \varepsilon^2 \tag{5}$$

2. *The (σ, ε)-eluder dimension of $\mathcal{Q}$ localised at $q^\star$, denoted $d_\sigma(\mathcal{Q}, q^\star, \varepsilon)$, is the length d of the longest sequence $(z_1, z_2, \ldots, z_d) \in \mathcal{Z}^d$ for which $\exists \varepsilon' > \varepsilon$ s.t. for any $i \le d$, z_i is (σ, ε)-independent from $(z_1, z_2, \ldots, z_{i-1})$ w.r.t. $(\mathcal{Q}, q^\star)$.*

The notion of ε-independence used in Definition 5.1 follows the same rational as Russo and Van Roy (2013): z (resp. $\bar{z}$) has a large (resp. small) loss and hence are independent. Note that losses are considered between q and $q^\star$ rather than on pairs (q, q'). This is a minor variation that allows to fix a nominal model of interest within $\mathcal{Q}$ (see Li et al. (2022) for a more thorough discussion). The main difference lies in the fact that Russo and Van Roy (2013) defines the eluder dimension via the square loss, which limits its applicability to additive noise models, whereas – in line with Bakhtiari et al. (2025) – we use $\bar{\varphi}$ for the learning error (in Equation (4)) to obtain an eluder dimension tailored to the logistic structure of our problem. However, we depart from Bakhtiari et al. (2025) by leveraging a distinct prediction error measure. This translates in Equation (5) using $\dot{\sigma}(q^\star(z))\,(q^\star(z) - q(z))^2$ instead of $\bar{\varphi}\big(\sigma(q^\star(z)), \sigma(q(z))\big)$. As a result, local sensitivity is taken into account directly in the eluder definition while preserving the original ideas as

$$\bar{\varphi}\big(\sigma(q^\star(z)), \sigma(q(z))\big) \simeq \dot{\sigma}(q^\star(z))\,(q^\star(z) - q(z))^2 .$$

It is worth noting that this quantity appears only in the analysis and is not required for the implementation of the algorithm.

Theoretical guarantees. We are now ready to state the main result, which provides regret guarantees for the **GINOP** algorithm over a general reward function class. A high-level overview of the proof is provided in the subsequent paragraph, while the complete proof is deferred to Section C.2.

Theorem 5.1. *Under Assumption 2.1 and Assumption 2.2, letting $\varepsilon = 1/T^2$ and $\delta = 1/T^2$, the **GINOP** algorithm satisfies*

$$\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{\frac{T}{\dot{\sigma}^\star}\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta\right)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)} + \kappa\,\Gamma_T\right),$$

$$\textit{where}\quad \Gamma_T = \widetilde{\mathcal{O}}\Big(\log\big(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta\big)\, d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\Big), \quad \dot{\sigma}^\star := \dot{\sigma}(\Delta_{f^\star}(x^\star, x^\star)) = \frac{1}{4}.$$

Theorem 5.1 establishes a first-order regret upper bound. The leading $\sqrt{T}$ term depends on three quantities: (i) $\dot{\sigma}^\star$, the local curvature of the sigmoid at the optimal arms, which is a well-behaved constant in contrast to κ; (ii) the log-covering number, which controls statistical overfitting and is a standard feature of complexity notions in statistical learning theory; and (iii) the locally sensitive eluder dimension associated to $\Delta_{\mathcal{F}}$, which captures how effectively the value of unobserved actions can be inferred from observed samples. The dependence on κ is confined to the lower-order term $\kappa\,\Gamma_T$, where Γ_T is a problem-dependent quantity. That the leading-order term depends on the

local (small) curvature $\dot{\sigma}^\star$ rather than on the global one κ assesses that, from a learning-theoretic perspective, learning from preference feedback is statistically as efficient as learning from direct reward observations. While assessing the tightness of the upper bound is difficult in the general case, we instantiate our results to specific reward function classes in Section 6, enabling a more refined evaluation.

Idea of proof. For convenience, we denote by $\mathcal{GE}$ the intersection of all high-probability events required to establish our results (see Remark B.1). Complete proofs are deferred to Section B and Section C. The algorithm relies on a tailored notion of optimism in its decision rule (Step 4). In contrast to standard bandit settings with direct feedback, where the exploration bonus typically aggregates uncertainties over individual arms, the bonus here is specifically designed to capture the structure of dueling bandits by prioritizing comparisons between *distinct* arms. Although this departs from the classical formulation, the proposed informative optimism preserves the underlying principle of optimism by enabling tight round-wise regret control (see Lemma 5.1, proved in Section C.1).

Lemma 5.1 (Round-Wise Regret)**.** *Under Assumption 2.1 and Assumption 2.2, on the event $\mathcal{GE}$, the following holds:*

$$\forall t \in [T],\ \rho(t) := f^\star(x^\star) - \frac{f^\star(x_t) + f^\star(x_t')}{2} \ \leq\ \omega_t(x_t, x_t').$$

The regret bound then hinges on efficiently upper bounding the on-policy prediction error $\omega_t(x_t, x_t')$ accumulated along the trajectory, which we establish in Lemma 5.2. This result, proved in Section B.2, constitutes one of our principal theoretical contributions as it provides a locally sensitive characterization of the on-policy prediction error through the introduced eluder dimension.

Lemma 5.2. *On the event $\mathcal{GE}$, the following hold:*

$$\sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x_t'))}\, \omega_t(x_t, x_t') = \widetilde{\mathcal{O}}\Big(\sqrt{T\, \log(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\Big),$$
$$\sum_{t=1}^{T} \omega_t(x_t, x_t')^2 = \widetilde{\mathcal{O}}\Big(\kappa \log(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta)\, d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\Big).$$

Finally, we carefully decompose the regret so that it scales with the sensitivity of the sigmoid function around the optimal action $(x^\star, x^\star)$.

$$\begin{aligned}\mathcal{R}_T &\leq \frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x_t'))}\, \omega_t(x_t, x_t') + \frac{1}{\sqrt{2\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\rho(t)}\, \omega_t(x_t, x_t') \\ &\leq \underbrace{\frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x_t'))}\, \omega_t(x_t, x_t')}_{(a)} + \underbrace{\frac{1}{\sqrt{2\dot{\sigma}^\star}} \sqrt{\sum_{t=1}^{T} \omega_t(x_t, x_t')^2}}_{(b)} \sqrt{\mathcal{R}_T},\end{aligned}$$

where the second inequality follows from Taylor expansion and Cauchy–Schwarz inequality. This yields a quadratic inequality in $\mathcal{R}_T$, whose solution satisfies $\mathcal{R}_T \leq a + 2b^2$. Bounding the terms (a) and (b) using Lemma 5.2 completes the proof and yields the stated regret bound.

6 Instantiation to Familiar Utility Classes

In Section 5, the **GINOP** algorithm and its first-order regret upper bound were established for a general reward function class. To further illuminate the behavior of the algorithm and assess the sharpness of the theoretical guarantees in canonical settings, we now specialize the analysis to two classical utility models: the linear and kernelized reward function classes.

Linear utility. In this setting, $\mathcal{X} \subset \mathbb{B}_2^d(1)$ and, for any $x \in \mathcal{X}$, the reward function is given by $f^\star(x) = \langle \theta^\star, x \rangle$, where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter satisfying $\|\theta^\star\| \leq S$. Accordingly, the class of linear reward functions $\mathcal{F}_{lin}$ can be identified with the parameter set $\Theta \subset \mathbb{B}_2^d(S)$, i.e., $\mathcal{F}_{lin} = \{\langle \theta, \cdot \rangle : \theta \in \Theta\}$. Under the Bradley–Terry model, the preference feedback satisfies $y_t \sim \text{Bernoulli}(\sigma(\langle \theta^\star, x_t - x_t' \rangle))$. The algorithmic implementation of **GINOP** in the linear utility

setting, including the characterization of the MLE $\hat{f}_t$ and the uncertainty quantifier ω_t, is deferred to Section E.1. To specialize the regret bound of Theorem 5.1 to $\mathcal{F}_{lin}$, it remains to control the key complexity quantities appearing in the theorem, namely $\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))\right)$ and $d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)$, in the case $\mathcal{F} = \mathcal{F}_{lin}$.

Proposition 6.1. *In the linear reward setting, the following bounds hold:*
$\forall\, \varepsilon > 0, \quad d_\sigma(\Delta\mathcal{F}_{lin}, \Delta_{f^\star}, \varepsilon) = \mathcal{O}\big(d\log(1+1/\varepsilon)\big) \quad$ *and* $\quad \log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}_{lin}))\right) = \mathcal{O}\big(d\log T\big).$

Given Proposition 6.1, whose proof is deferred to Appendix E.1, the following corollary is immediate.

Corollary 6.2. *In the linear reward setting, letting $\varepsilon = 1/T^2$ and $\delta = 1/T^2$, the **GINOP** algorithm satisfies $\mathcal{R}_T = \widetilde{\mathcal{O}}\Big(d\sqrt{\frac{T}{\dot{\sigma}^\star}} + \kappa d^3\Big)$.*

In the linear utility setting, similarly to the standard logistic bandit framework (i.e., without preference feedback) studied in Abeille et al. (2021), **(i)** the dependence on κ is confined to lower-order terms, and **(ii)** the leading term depends on the local curvature of the sigmoid at the optimal action, $\dot{\sigma}^\star$. In particular, Abeille et al. (2021) establish a minimax regret bound of the form $\widetilde{\mathcal{O}}\big(d\sqrt{\dot{\sigma}^\star T} + \kappa\big)$, which we match up to a factor $1/\dot{\sigma}^\star$. We emphasize that this discrepancy is not a weakness of our analysis, but rather stems from a difference in the reward metric: our objective involves $f^\star(x)$, whereas their formulation considers $\sigma(f^\star(x))$. A straightforward adaptation of their lower bound techniques to our setting confirms the tightness of Corollary 6.2 with respect to $\dot{\sigma}^\star$.

The refined characterization of the scaling with respect to the sigmoid curvature allows one to fully exploit a key feature of preference feedback: actions are selected as pairs, and the observed signal depends on relative rewards. For the optimal pair $(x^\star, x^\star)$, the local curvature of the sigmoid satisfies $\dot{\sigma}^\star = \dot{\sigma}(0) = 1/4$. Consequently, the dependence on curvature disappears from the leading-order term, yielding performance comparable to that obtained under direct reward observations. In particular, the resulting guarantee aligns with that obtained from (Russo and Van Roy, 2013, Proposition 4 + Example 4) in a direct reward model.

Consequently, for dueling bandits with linear utilities, **GINOP** is minimax optimal: its regret bound matches the lower bound $\Omega(d\sqrt{T})$ established by Li et al. (2024). This stands in contrast to prior work (Saha, 2021; Bengs et al., 2022; Li et al., 2024), whose leading term carries an additional dependence on κ.

Kernelized utility. Following (Pásztor et al., 2024; Xu et al., 2024; Kayal et al., 2025), we assume that the utility function $f^\star$ belongs to a known reproducing kernel Hilbert space (RKHS). Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel, and let $\mathcal{H}_k$ denote the associated RKHS, equipped with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$. We assume $\|f^\star\|_{\mathcal{H}_k} \le S$ and $k(x,x) \le 1$ for all $x \in \mathcal{X}$. By the reproducing property, for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, $\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k} = f(x)$. By Mercer's theorem, under mild regularity conditions, the kernel admits the spectral representation $k(x,x') = \sum_{m=1}^{\infty} \gamma_m \phi_m(x)\phi_m(x')$, where $\gamma_m > 0$ and $\{\psi_m := \sqrt{\gamma_m}\phi_m\}_{m\ge1}$ forms an orthonormal basis of $\mathcal{H}_k$. In particular, any $f \in \mathcal{H}_k$ can be expressed as $f(\cdot) = \sum_{m=1}^{\infty} \theta_m \psi_m(\cdot) = \langle\theta, \Psi(\cdot)\rangle, \quad$ with $\quad \|f\|_{\mathcal{H}_k}^2 = \|\theta\|_2^2 \le S^2$. We refer to $\{\gamma_m\}$ and $\{\phi_m\}$ as the Mercer eigenvalues and eigenfunctions of k, respectively. Accordingly, the class of kernelized reward functions $\mathcal{F}_k$ can be identified with the parameter set $\Theta_k \subset \mathbb{B}_2^{\mathbb{N}}(S)$, i.e., $\mathcal{F}_k = \{\langle\theta, \Psi(\cdot)\rangle : \theta \in \Theta_k\}$, with $f^\star(\cdot) = \langle\theta^\star, \Psi(\cdot)\rangle$ and $\theta^\star \in \Theta_k$. Under the Bradley–Terry model, the preference feedback satisfies $y_t \sim \mathrm{Bernoulli}(\sigma(\langle\theta^\star, \Psi(x_t) - \Psi(x_t')\rangle))$. Define $z = (x, x') \in \mathcal{X} \times \mathcal{X}$. Following Pásztor et al. (2024); Kayal et al. (2025), we introduce the *dueling kernel* $\mathbb{k}(z_1, z_2) = k(x_1,x_2) + k(x_1',x_2') - k(x_1,x_2') - k(x_1',x_2)$, for $z_1 = (x_1, x_1')$ and $z_2 = (x_2, x_2')$. This construction satisfies $\|\Delta_f\|_{\mathcal{H}_{\mathbb{k}}} = \|f\|_{\mathcal{H}_k}$, as established in (Pásztor et al., 2024, Proposition 4). The algorithmic implementation of **GINOP** in the kernelized utility setting, including the characterization of the MLE $\hat{f}_t$ and the uncertainty quantifier ω_t, is deferred to Section E.2. We assess whether the resulting locally sensitive eluder dimension induces an undesirable dependence on κ. For any $\lambda > 0$ and $T \ge 1$, the maximum information gain after T observations is defined as $\gamma_T(\lambda;\, \mathcal{X}\times\mathcal{X}) = \max_{(x_1,x_1'),\ldots,(x_T,x_T') \in \mathcal{X}\times\mathcal{X}} \frac{1}{2}\log\Big(\det\Big(I + \lambda^{-1}\mathbb{K}_T\Big)\Big)$, with $\mathbb{K}_T = \big[\mathbb{k}\big((x_i,x_i'),(x_j,x_j')\big)\big]_{i,j=1}^T$. The quantity γ_T admits a natural geometric interpretation: it measures the logarithm of the maximum volume of the ellipsoid generated by T points in $\mathcal{X}\times\mathcal{X}$, thereby capturing the intrinsic geometric complexity of the domain. Proposition 6.3, whose proof is deferred to Appendix E.2, establishes a precise relationship between our eluder dimension and the maximum information gain.

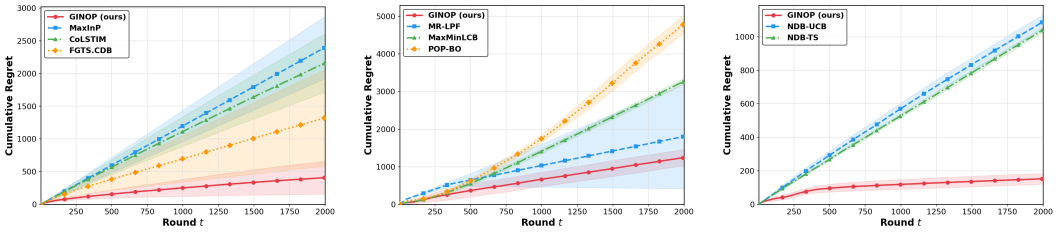
Proposition 6.3. *In the kernelized setting, the following bound holds:*

$$\forall\, \varepsilon > 0, \qquad d_\sigma(\Delta_{\mathcal{F}_k}, \Delta_{f^\star}, \varepsilon) = \mathcal{O}\Big(\gamma_T(\lambda; \mathcal{X} \times \mathcal{X})\Big), \quad \textit{with } \lambda = \frac{2\varepsilon(1+S)}{(2S)^2}.$$

It is further known that the logarithmic covering number of a bounded RKHS satisfies $\log \mathcal{N}_T \asymp \gamma_T$, as discussed in (Pásztor et al., 2024; Wainwright, 2019, Example 5.12). Building on this, we show in Section E.2.1, for commonly used kernel functions, that –in line with Kayal et al. (2025)– our results, and in particular their dependence on κ, improve upon the bounds of Xu et al. (2024) and Pásztor et al. (2024).

7 Experiments

To corroborate our theoretical results, we conduct numerical experiments evaluating the performance of **GINOP** against several existing baselines across a range of reward function classes. Additional details on the experimental setup and results are deferred to Section F.



(a) Linear utility. (b) Kernelized utility. (c) Against a NN.

Figure 1: We benchmark **GINOP** against several baselines across different settings, reporting in each case the cumulative regret over a horizon of $T = 2000$ rounds, averaged over 20 independent trials and considering an action set of 30 arms. **(a)** In the linear setting, we take $d = 10$ and benchmark against **MaxInP** (Saha, 2021), **CoLSTIM** (Bengs et al., 2022), and **FGTS.CDB** (Li et al., 2024). **(b)** In the kernelized setting, we adopt the Ackley function as the reward and employ the Matérn kernel with smoothness parameter $\nu = 2.5$; we benchmark against **POP-BO** (Xu et al., 2024), **MaxMinLCB** (Pásztor et al., 2024), and **MR-LPF** (Kayal et al., 2025). **(c)** In the neural setting, we consider a cosine reward function and benchmark against the neural-network-based approaches **NDB-UCB** and **NDB-TS** (Verma et al., 2025).

We observe that **GINOP** outperforms existing baselines, in agreement with our theoretical bounds, highlighting its favorable dependence on the curvature of the link function. Indeed, all existing approaches (but **MR-LPF**) suffer from a κ-dependence both in the algorithmic design and in the theoretical performance, which inflates their regret and induces a $\sqrt{T}$-like shape that becomes apparent only at considerably larger horizons. This stresses the importance of the fine treatment of non-linearity in preference-based bandits. The exception is **MR-LPF** in the kernelized setting, which – thanks to its multi-phase approach – is able to eliminate the κ-dependence and consequently exhibits a regret profile comparable to that of **GINOP**.

Conclusion and Future Work

By leveraging a novel eluder dimension tailored to the preference feedback setting together with an informed-optimistic exploration scheme, we show that the resulting algorithm (**GINOP**) enjoys tight regret guarantees. These results hold for general reward function classes, thereby accommodating a wide range of problem instances. In particular, this highlights that the non-linearity of the logistic model – inherited from the Bradley-Terry assumption – does not adversely impact the regret in the preference bandit setting. Consequently, our work demonstrates that learning from preference feedback can be as statistically efficient as learning from direct reward observations.

While the primary focus of this paper is theoretical, on the computational front **GINOP** matches the computational complexity of existing work in specific settings (linear and kernelized utilities). Computational efficiency for general function classes, however, remains a common problem in the literature (Sekhari et al., 2023); the design of suitable oracle-based implementations lies beyond the

scope of this paper and is left for future work. In addition, promising future work (following Bengs et al. (2022)) lies in extending the analysis to observation models beyond the Bradley–Terry model.

References

Abeille, M., Faury, L., and Calauzenes, C. (2021). Instance-wise minimax-optimal algorithms for logistic bandits. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3691–3699. PMLR. (Cited on 3, 8, 22, 24)

Ailon, N., Karnin, Z., and Joachims, T. (2014). Reducing dueling bandits to cardinal bandits. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 856–864, Bejing, China. PMLR. (Cited on 3)

Aiolli, F. and Sperduti, A. (2004). Learning preferences for multiclass problems. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press. (Cited on 2)

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. (Cited on 3)

Bakhtiari, A., Ayoub, A., Robertson, S. M., Janz, D., and Szepesvari, C. (2025). Eluder dimension: localise it! In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. (Cited on 4, 5, 6, 14, 15, 16)

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. (2021). Preference-based online learning with dueling bandits: a survey. *J. Mach. Learn. Res.*, 22(1). (Cited on 1, 3)

Bengs, V., Saha, A., and Hüllermeier, E. (2022). Stochastic contextual dueling bandits under linear stochastic transitivity models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1764–1786. PMLR. (Cited on 1, 2, 3, 5, 8, 9, 10, 22, 26)

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345. (Cited on 2)

Chen, B. and Frazier, P. I. (2017). Dueling bandits with weak regret. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 731–739. PMLR. (Cited on 3)

Chu, W. and Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 137–144, New York, NY, USA. Association for Computing Machinery. (Cited on 2)

Croissant, L., Abeille, M., and Bouchard, B. (2024). Near-continuous time Reinforcement Learning for continuous state-action spaces. In Vernade, C. and Hsu, D., editors, *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 444–498. PMLR. (Cited on 20, 21)

Di, Q., He, J., and Gu, Q. (2025). Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13513–13545. PMLR. (Cited on 3)

Di, Q., Jin, T., Wu, Y., Zhao, H., Farnoud, F., and Gu, Q. (2024). Variance-aware regret bounds for stochastic contextual dueling bandits. In *The Twelfth International Conference on Learning Representations*. (Cited on 3, 5, 22)

Falahatgar, M., Orlitsky, A., Pichapati, V., and Suresh, A. T. (2017). Maximum selection and ranking under noisy comparisons. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1088–1096. PMLR. (Cited on 3)

Faury, L., Abeille, M., Calauzenes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR. (Cited on 3, 4, 21)

Faury, L., Abeille, M., Jun, K.-S., and Calauzenes, C. (2022). Jointly efficient and optimal algorithms for logistic bandits. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR. (Cited on 3)

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc. (Cited on 3)

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. (Cited on 3)

Jamieson, K. G. and Nowak, R. D. (2011). Active ranking using pairwise comparisons. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, page 2240–2248, Red Hook, NY, USA. Curran Associates Inc. (Cited on 3)

Ji, X., Wang, H., Chen, M., Zhao, T., and Wang, M. (2023). Provable benefits of policy learning from human preferences in contextual bandit problems. (Cited on 3)

Kayal, A., Vakili, S., Toni, L., shan Shiu, D., and Bernacchia, A. (2025). Bayesian optimization from human feedback: Near-optimal regret bounds. In *Forty-second International Conference on Machine Learning*. (Cited on 1, 3, 5, 8, 9, 23, 24, 26, 27)

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1152–1161, Lille, France. PMLR. (Cited on 3)

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press. (Cited on 1)

Lee, J., Yun, S.-Y., and Jun, K.-S. (2024). A unified confidence sequence for generalized linear models, with applications to bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. (Cited on 21)

Li, G., Kamath, P., Foster, D. J., and Srebro, N. (2022). Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750. (Cited on 4, 6)

Li, X., Zhao, H., and Gu, Q. (2024). Feel-good thompson sampling for contextual dueling bandits. (Cited on 1, 2, 3, 5, 8, 9, 22, 26)

Liu, Q., Chung, A., Szepesvari, C., and Jin, C. (2022). When is partially observable reinforcement learning not scary? In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5175–5220. PMLR. (Cited on 4, 15)

Munos, R., Valko, M., Calandriello, D., Gheshlaghi Azar, M., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Fiegel, C., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D. J., Precup, D., and Piot, B. (2024). Nash learning from human feedback. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36743–36768. PMLR. (Cited on 3)

Pásztor, B., Kassraie, P., and Krause, A. (2024). Bandits with preference feedback: A stackelberg game perspective. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 11997–12034. Curran Associates, Inc. (Cited on 1, 3, 5, 8, 9, 23, 24, 26, 27)

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. (Cited on 3, 4, 6, 8, 20, 23)

Saha, A. (2021). Optimal algorithms for stochastic contextual preference bandits. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30050–30062. Curran Associates, Inc. (Cited on 1, 2, 3, 5, 8, 9, 22, 26)

Saha, A. and Krishnamurthy, A. (2022). Efficient and optimal algorithms for contextual dueling bandits under realizability. In Dasgupta, S. and Haghtalab, N., editors, *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 968–994. PMLR. (Cited on 3)

Saha, A., Pacchiano, A., and Lee, J. (2023). Dueling rl: Reinforcement learning with trajectory preferences. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6263–6289. PMLR. (Cited on 3)

Sekhari, A., Sridharan, K., Sun, W., and Wu, R. (2023). Contextual bandits and imitation learning with preference-based active queries. In *Thirty-seventh Conference on Neural Information Processing Systems*. (Cited on 4, 9)

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2022). Learning to summarize from human feedback. (Cited on 3)

Vakili, S., Khezeli, K., and Picheny, V. (2021). On information gain and regret bounds in gaussian process bandits. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 82–90. PMLR. (Cited on 25, 26)

Verma, A., Dai, Z., Lin, X., Jaillet, P., and Low, B. K. H. (2025). Neural dueling bandits: Preference-based optimization with human feedback. In *The Thirteenth International Conference on Learning Representations*. (Cited on 3, 5, 9, 27)

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. (Cited on 9)

Wu, H. and Liu, X. (2016). Double thompson sampling for dueling bandits. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. (Cited on 3)

Xu, W., Wang, W., Jiang, Y., Svetozarevic, B., and Jones, C. N. (2024). Principled preferential bayesian optimization. (Cited on 1, 3, 5, 8, 9, 23, 26, 27)

Yue, Y., Broder, J., Kleinberg, R. D., and Joachims, T. (2012). The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78:1538–1556. (Cited on 3)

Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1201–1208, New York, NY, USA. Association for Computing Machinery. (Cited on 3)

Zhu, B., Jordan, M., and Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR. (Cited on 3)

Zimmert, J. and Seldin, Y. (2018). Factored bandits. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. (Cited on 3)

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. (2014a). Relative upper confidence bound for the k-armed dueling bandit problem. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 10–18, Bejing, China. PMLR. (Cited on 3)

Zoghi, M., Whiteson, S. A., de Rijke, M., and Munos, R. (2014b). Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 73–82, New York, NY, USA. Association for Computing Machinery. (Cited on 3)

Appendices

Appendix Contents

A A Closer Look at the Localization Technique of Bakhtiari et al. (2025) **15**

B Preliminaries **15**

B.1 Concentrations 15

B.2 Bounding the On-Policy Prediction Error 16

C GINOP Performance **18**

C.1 Regret Decomposition 18

C.2 Regret Bound 18

D On the Eluder Dimension **19**

E Instantiation to Familiar Utility Classes **21**

E.1 Linear Utility 21

E.2 Kernelized Utility 23

E.2.1 Specialization to Commonly Used Kernel Functions 25

F Details of Experiments **26**

A A Closer Look at the Localization Technique of Bakhtiari et al. (2025)

Bakhtiari et al. (2025) build on the global ℓ_1-eluder dimension of Liu et al. (2022), defined as follows.

Definition A.1. *Let $\mathcal{Z}$ be a set, Ψ a class of real-valued functions on $\mathcal{Z}$, and $z = (z_1, z_2, \ldots, z_n)$ a sequence of length n in $\mathcal{Z}$. We define the following.*

1. *An element $z_t \in \mathcal{Z}$ is said to be ε-independent of $z_{1:t-1}$ with respect to Ψ if there exists $\psi \in \Psi$ such that*
$$\sum_{i=1}^{t-1} |\psi(z_i)| \leq \varepsilon \quad \textit{and} \quad |\psi(z_t)| > \varepsilon.$$
2. *The sequence z is called an ε-eluder sequence with respect to Ψ if, for every $t \leq n$, z_t is ε-independent of $z_{1:t-1}$ with respect to Ψ.*
3. *The ε-eluder dimension $\dim_{\mathrm{elud}}(\varepsilon;\, \Psi)$ of Ψ is the length of the longest ε'-eluder sequence with respect to Ψ, maximized over all $\varepsilon' \geq \varepsilon$.*

Building on their lower bound (Bakhtiari et al., 2025, Theorem 2), the authors establish that the global eluder dimension $\dim_{\mathrm{elud}}(1/T;\, \mathcal{F})$, taken over the entire class $\mathcal{F}$, necessarily incurs an undesirable dependence on κ. This observation motivates a form of *localization*: rather than considering the eluder dimension of the full class, they propose restricting it to a suitable subset $\mathcal{F}' \subseteq \mathcal{F}$ and deriving a regret upper bound in terms of $\dim_{\mathrm{elud}}(1/T;\, \mathcal{F}')$ instead of $\dim_{\mathrm{elud}}(1/T;\, \mathcal{F})$. This refinement, at least ostensibly, removes the κ-dependence from the leading-order term. However, this comes at the cost of an opaque second-order term of the form

$$\mathrm{card}\{t \in [T] : f_t \notin \mathcal{F}'\}, \tag{6}$$

which is generally intractable: the subclass $\mathcal{F}'$ is defined only implicitly and cannot be evaluated for general function classes.

To better illustrate the localization technique, the authors instantiate it in the generalized linear model setting, where they establish that the appropriate local family $\mathcal{F}'$ permitting the removal of the κ-dependence from the leading term is given by (Bakhtiari et al., 2025, Section 4.2):

$$\mathcal{F}' \equiv \Theta' = \{\theta \in \Theta : \forall x \in \mathcal{X},\ |\langle x,\, \theta - \theta^\star \rangle| \leq 1/M\}\,,$$

where $M = 1/4$ is the self-concordance constant. Hence, the challenge lies in bounding the additional term:

$$\mathrm{card}\{t \in [T] : f_t \notin \mathcal{F}'\} = \mathrm{card}\{t \in [T] : \exists x \in \mathcal{X},\ |\langle x,\, \theta_t - \theta^\star \rangle| > 1/M\} \tag{7}$$

Unfortunately, instead of upper bounding the term in (7) which requires sufficient accuracy *uniformly over arms*, (Bakhtiari et al., 2025, Proposition 5) inaccurately bound the quantity:

$$\mathrm{card}\{t \in [T] : |\langle x_t,\, \theta_t - \theta^\star \rangle| > 1/M\},$$

which differs substantially from the correct second-order term (7) in that it requires sufficient accuracy only for actions played over the trajectory. While the latter *on-policy* term is indeed logarithmic, it does not imply a bound on (7) which can be linear in T in all generality.

While we align with Bakhtiari et al. (2025) regarding the shortcomings of the standard global eluder dimension, we depart from their approach: in Definition 5.1, we introduce a novel *locally sensitive eluder dimension* that explicitly and directly encodes the sensitivity of the link function within the definition itself, rather than relying on an external localization technique.

B Preliminaries

B.1 Concentrations

We recall Lemma 4.1 stated in the main:

Lemma 4.1. *Let $\mathcal{F}_t$ be defined as in Equation (2) for all $t \geq 1$. Under Assumption 2.2, with probability at least $1 - \delta$, it holds that $f^\star \in \bigcap_{t \geq 1} \mathcal{F}_t$.*

For each $f \in \mathcal{F}$, we define the excess loss function $\varphi_f : [0,1] \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and its corresponding expected excess loss $\bar{\varphi}_f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$\varphi_f(y, x, x') = \ell(y, \sigma(\Delta_f(x, x'))) - \ell(y, \sigma(\Delta_{f^\star}(x, x'))),$$
$$\bar{\varphi}_f(x, x') = \mathbb{E}_y[\ell(y, \sigma(\Delta_f(x, x')))] - \mathbb{E}_y[\ell(y, \sigma(\Delta_{f^\star}(x, x')))].$$

With this in hand, we introduce Proposition B.1, which is a direct adaptation of Proposition 22 in Bakhtiari et al. (2025).

Proposition B.1. *With probability at least $1 - \delta$, the following holds:*

$$\forall f \in \mathcal{F},\ \forall t \in [T], \quad \sum_{s=1}^{t} \bar{\varphi}_f(x_s, x'_s) \le 2\left(\sum_{s=1}^{t} \varphi_f(y_s, x_s, x'_s) + \beta_t(\delta)\right).$$

Remark B.1. *We denote by the event $\mathcal{GE}$ the intersection of the good events in Lemma 4.1, and Proposition B.1, which occurs with probability at least $1 - 2\delta$. This event will be used extensively throughout the proofs in the appendices.*

B.2 Bounding the On-Policy Prediction Error

Lemma 5.2. *On the event $\mathcal{GE}$, the following hold:*

$$\sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x'_t))}\, \omega_t(x_t, x'_t) = \widetilde{\mathcal{O}}\Big(\sqrt{T\, \log(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\Big),$$
$$\sum_{t=1}^{T} \omega_t(x_t, x'_t)^2 = \widetilde{\mathcal{O}}\Big(\kappa \log(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta)\, d^2_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\Big).$$

Proof of Lemma 5.2. Throughout the proof we suppose we are on $\mathcal{GE}$.

Step 1. First we establish :

$$\begin{aligned}
\forall t \in [T],\ \omega_t(x_t, x'_t)^2 &= \Big(\sup_{f, f' \in \mathcal{F}_t} \Delta_f(x_t, x'_t) - \Delta_{f'}(x_t, x'_t)\Big)^2 \\
&\le \sup_{f, f' \in \mathcal{F}_t} \Big(\Delta_f(x_t, x'_t) - \Delta_{f'}(x_t, x'_t)\Big)^2 \\
&\le \sup_{f, f' \in \mathcal{F}_t} \Big(\Delta_f(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t) + \Delta_{f'}(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t)\Big)^2 \\
&\overset{(i)}{\le} \sup_{f, f' \in \mathcal{F}_t} 2\Big(\Delta_f(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t)\Big)^2 + 2\Big(\Delta_{f'}(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t)\Big)^2 \\
&\le 4 \sup_{f \in \mathcal{F}_t} \big(\Delta_f(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t)\big)^2 \\
&= 4\big(\Delta_{f_t}(x_t, x'_t) - \Delta_{f^\star}(x_t, x'_t)\big)^2 \qquad (8)
\end{aligned}$$

where in (i) we use that $(a+b)^2 \le 2a^2 + 2b^2$. For the last equality $f_t \in \mathcal{F}_t$ denotes a function attaining the supremum.

Step 2. We apply Proposition B.1, for $f_t \in \mathcal{F}$:

$$\begin{aligned}
\sum_{s=1}^{t-1} \bar{\varphi}_{f_t}(x_s, x'_s) &\le 2\left(\sum_{s=1}^{t-1} \varphi_{f_t}(y_s, x_s, x'_s) + \beta_t\right) \\
&\le 2\left(\sum_{s=1}^{t-1} \ell(y_s, \sigma(\Delta_{f_t}(x_s, x'_s))) - \sum_{s=1}^{t-1} \ell(y_s, \sigma(\Delta_{f^\star}(x_s, x'_s))) + \beta_t(\delta)\right) \quad \text{(By definition of } \varphi_{f_t}\text{)} \\
&\le 2\left(\underbrace{\sum_{s=1}^{t-1} \ell(y_s, \sigma(\Delta_{f_t}(x_s, x'_s))) - \sum_{s=1}^{t-1} \ell(y_s, \sigma(\Delta \hat{f}_t(x_s, x'_s)))}_{\le \beta_t(\delta)\text{: because } f_t \in \mathcal{F}_t} + \beta_t(\delta)\right) \quad \text{(By definition of } \varphi_{\hat{f}_t} \text{ and } f^\star \in \mathcal{F}_t\text{)} \\
&\le 4\beta_t(\delta).
\end{aligned}$$

Hence:

$$\sum_{s=1}^{t-1} \bar{\varphi}_{f_t}(x_s, x'_s) \le 4\beta_t(\delta). \tag{9}$$

Step 3. Recalling that our decision space is $\mathcal{X} \times \mathcal{X}$, we define

$$\begin{aligned}
&\Psi : \mathcal{F} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, && \phi : \mathcal{F} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, \\
&\quad (f, x, x') \mapsto \bar{\varphi}_f(x, x'), && \quad (f, x, x') \mapsto \dot{\sigma}(\Delta_{f^\star}(x, x'))\big(\Delta_f(x, x') - \Delta_{f^\star}(x, x')\big)^2.
\end{aligned}$$

For this choice of ψ and ϕ, the eluder $d_{\psi,\phi}(\mathcal{F}, \varepsilon)$ from Definition D.2 exactly coincide with $d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)$ from Definition 5.1. Using this, we apply Lemma D.3 with this choice of ψ and ϕ.

Let $\{(f_1, x_1, x'_1), \ldots, (f_T, x_T, x'_T)\}$ denote the corresponding sequence in Lemma D.3. Setting $\beta = 4\beta_t$ and $C = S$ yields that:

$$\begin{aligned}
&\sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x, x'))}\big(\Delta_f(x, x') - \Delta_{f^\star}(x, x')\big) \le 2\sqrt{4T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)} + \sqrt{S}\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon), \\
&\sum_{t=1}^{T} \dot{\sigma}(\Delta_{f^\star}(x, x'))\big(\Delta_f(x, x') - \Delta_{f^\star}(x, x')\big)^2 \le 4\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\left(2 + \log\left(\frac{TS}{16\beta_T(\delta)^2 d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\right)\right) \\
&\qquad\qquad + d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big(2 + 4\beta_T(\delta) d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big)(1 + S) \\
&\implies \sum_{t=1}^{T} \big(\Delta_f(x, x') - \Delta_{f^\star}(x, x')\big)^2 \le 4\kappa\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\left(2 + \log\left(\frac{TS}{16\beta_T(\delta)^2 d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\right)\right) \\
&\qquad\qquad + \kappa s d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big(2 + 4\beta_T(\delta) d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big)(1 + S)
\end{aligned}$$

Using (8) implies the desired result:

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x'_t))}\omega_t(x_t, x'_t) &\le 8\sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)} + 2\sqrt{S}\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\right), \\
\sum_{t=1}^{T} \omega_t(x_t, x'_t)^2 &\le 16\kappa\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\left(2 + \log\left(\frac{TS}{16\beta_T(\delta)^2 d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\right)\right) \\
&\quad + 4\kappa d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big(2 + 4\beta_T(\delta) d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\big)(1 + S) \\
&= \kappa\, \Gamma_T.
\end{aligned}$$

where

$$\Gamma_T = \widetilde{\mathcal{O}}\Big(\log(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)^2\Big).$$

□

C GINOP Performance

C.1 Regret Decomposition

Lemma 5.1 (Round-Wise Regret)**.** *Under Assumption 2.1 and Assumption 2.2, on the event $\mathcal{GE}$, the following holds:*

$$\forall t \in [T],\ \rho(t) := f^\star(x^\star) - \frac{f^\star(x_t) + f^\star(x_t')}{2}\ \leq\ \omega_t(x_t, x_t').$$

Proof of Lemma 5.1. Let x_t and x_t' be the arms selected by the algorithm at round t. Then the instantaneous regret can be decomposed as follows:

$$\begin{aligned} 2\,\rho(t) &= \Delta_{f^\star}(x^\star, x_t) + \Delta_{f^\star}(x^\star, x_t') \\ &= \Delta_{f^\star}(x^\star, x_t) - \Delta_{\hat{f}_t}(x^\star, x_t) + \Delta_{\hat{f}_t}(x^\star, x_t) + \Delta_{f^\star}(x^\star, x_t') - \Delta_{\hat{f}_t}(x^\star, x_t') + \Delta_{\hat{f}_t}(x^\star, x_t') \end{aligned}$$

By the Lemma 4.1 and the definition of ω_t, then on $\mathcal{GE}$:

$$\begin{aligned} \forall t \in [T], \quad \Delta_{f^\star}(x^\star, x_t) - \Delta_{\hat{f}_t}(x^\star, x_t) &\leq \omega_t(x^\star, x_t) \\ \Delta_{f^\star}(x^\star, x_t) - \Delta_{\hat{f}_t}(x^\star, x_t') &\leq \omega_t(x^\star, x_t') \end{aligned}$$

Hence:

$$2\,\rho(t) \leq \omega_t(x^\star, x_t) + \Delta_{\hat{f}_t}(x^\star, x_t) + \omega_t(x^\star, x_t') + \Delta_{\hat{f}_t}(x^\star, x_t'). \tag{10}$$

Given the optimization step (Step 4), we obtain:

$$\begin{aligned} \hat{f}_t(x^\star) + \hat{f}_t(x_t') + \omega_t(x^\star, x_t') &\leq \hat{f}_t(x_t) + \hat{f}_t(x_t') + \omega_t(x_t, x_t') \\ \implies \hat{f}_t(x^\star) - \hat{f}_t(x_t) &\leq -\omega_t(x^\star, x_t') + \omega_t(x_t, x_t') \\ \implies \Delta_{\hat{f}_t}(x^\star, x_t) &\leq -\omega_t(x^\star, x_t') + \omega_t(x_t, x_t') \end{aligned} \tag{11}$$

Similarly, we obtain:

$$\Delta_{\hat{f}_t}(x^\star, x_t') \leq -\omega_t(x^\star, x_t) + \omega_t(x_t, x_t') \tag{12}$$

Combining (10), (11) and (12) gives $\rho(t) \leq \omega(x_t, x_t')$. □

C.2 Regret Bound

Theorem 5.1. *Under Assumption 2.1 and Assumption 2.2, letting $\varepsilon = 1/T^2$ and $\delta = 1/T^2$, the **GINOP** algorithm satisfies*

$$\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{\frac{T}{\dot{\sigma}^\star}\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta\right)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)} + \kappa\,\Gamma_T\right),$$

where $\Gamma_T = \widetilde{\mathcal{O}}\Big(\log\big(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))/\delta\big)\, d_\sigma^2(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)\Big), \quad \dot{\sigma}^\star := \dot{\sigma}(\Delta_{f^\star}(x^\star, x^\star)) = \dfrac{1}{4}.$

Proof of Theorem 5.1.

Step.1 We work under the good event $\mathcal{GE}$. We define $\dot{\sigma}^\star = \dot{\sigma}(\Delta_{f^\star}(x^\star, x^\star)) = 1/4$.

$$\mathcal{R}_T = \sum_{t=1}^{T} \rho(t) \leq \sum_{t=1}^{T} \omega_t(x_t, x_t') = \frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}^\star}\, \omega_t(x_t, x_t').$$

Using a first order Taylor expansion on $\dot{\sigma}$:

$$\begin{aligned}
\dot{\sigma}^\star &\leq \dot{\sigma}(\Delta_{f^\star}(x_t, x_t')) + \frac{1}{4} \left|\Delta_{f^\star}(x_t, x_t') - \Delta_{f^\star}(x^\star, x^\star)\right| \qquad (|\ddot{\sigma}| \leq |\dot{\sigma}| \leq 1/4) \\
&\leq \dot{\sigma}(\Delta_{f^\star}(x_t, x_t')) + \frac{\rho(t)}{2}.
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathcal{R}_T &\leq \frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}^\star}\, \omega_t(x_t, x_t') \\
&\leq \frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x_t'))}\, \omega_t(x_t, x_t') + \frac{1}{\sqrt{2\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\rho(t)}\, \omega_t(x_t, x_t') \\
&\leq \underbrace{\frac{1}{\sqrt{\dot{\sigma}^\star}} \sum_{t=1}^{T} \sqrt{\dot{\sigma}(\Delta_{f^\star}(x_t, x_t'))}\, \omega_t(x_t, x_t')}_{a} + \underbrace{\frac{1}{\sqrt{2\dot{\sigma}^\star}} \sqrt{\sum_{t=1}^{T} \omega_t(x_t, x_t')^2}}_{b} \sqrt{\mathcal{R}_T} \quad \text{(Cauchy Schwartz)}
\end{aligned}$$

Which leads to the following quadratic inequality on $\mathcal{R}_T$:

$$\begin{aligned}
&\mathcal{R}_T \leq a + b\sqrt{\mathcal{R}_T} \\
\implies &\sqrt{\mathcal{R}_t} \leq \sqrt{a} + b \\
\implies &\mathcal{R}_t \leq 2a + 2b^2.
\end{aligned}$$

By Lemma 5.2:

$$\begin{aligned}
a &= \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{\dot{\sigma}^\star}} \sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, f^\star, \varepsilon)}\right) \\
b^2 &= \frac{\kappa}{2\dot{\sigma}^\star} \Gamma_T.
\end{aligned}$$

Hence on $\mathcal{GE}$: $\mathcal{R}_T \leq \widetilde{\mathcal{O}}\Big(\frac{\kappa}{2\dot{\sigma}^\star}\Gamma_T + \frac{1}{\sqrt{\dot{\sigma}^\star}} \sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, f^\star, \varepsilon)}\Big)$

Step 2.

$$\begin{aligned}
R_T &= \mathbb{E}\left[\mathcal{R}_T \mid \mathcal{GE}\right] \mathbb{P}(\mathcal{GE}) + \mathbb{E}\left[\mathcal{R}_T \mid \overline{\mathcal{GE}}\right] \mathbb{P}(\overline{\mathcal{GE}}) \\
&\leq \mathbb{E}\left[\mathcal{R}_T \mid \mathcal{GE}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \rho(t) \mid \overline{\mathcal{GE}}\right] 2\delta \\
&\leq \widetilde{\mathcal{O}}\Big(\frac{\kappa}{2\dot{\sigma}^\star}\Gamma_T + \frac{1}{\sqrt{\dot{\sigma}^\star}} \sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\Big) + 4ST\delta \\
&\leq \widetilde{\mathcal{O}}\Big(\frac{\kappa}{2\dot{\sigma}^\star}\Gamma_T + \frac{1}{\sqrt{\dot{\sigma}^\star}} \sqrt{T\beta_T(\delta)\, d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)}\Big) \qquad \text{(Because: } \delta = \frac{1}{T^2}\text{)}.
\end{aligned}$$

Which concludes the proof of the regret bound.

□

D On the Eluder Dimension

We state all the results of this section for a more generic version of Definition 5.1.

Definition D.1 (ε-independence)**.** *Let $\mathcal{Q}$ be a class of functions $q : \mathcal{Z} \to \mathbb{R}$. Let ψ and ϕ be two functions from $\mathcal{Q} \times \mathcal{Z} \to \mathbb{R}_+$. Given a sequence $(z_1, \dots, z_n) \in \mathcal{Z}^n$ of length $n > 0$ and $\varepsilon > 0$, we say that $z \in \mathcal{Z}$ is $(\varepsilon, \psi, \phi)$-independent from the sequence with respect to $\mathcal{Q}$ if there exists $q \in \mathcal{Q}$ such that*

$$\sum_{j \leq n} \psi(q, z_j) \leq \varepsilon^2 \quad \textit{and} \quad \phi(q, z) > \varepsilon^2.$$

Definition D.2 ($(\varepsilon, \psi, \phi)$-Eluder)**.** *Let $\mathcal{Q}$ be a class of functions $q : \mathcal{Z} \to \mathbb{R}$. Let ψ and ϕ be two functions from $\mathcal{Q} \times \mathcal{Z} \to \mathbb{R}_+$. The $(\varepsilon, \psi, \phi)$-eluder dimension $d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$ of $\mathcal{Q}$ is the length d of the longest sequence $(z_1, \dots, z_d) \in \mathcal{Z}^d$ for which there exists $\varepsilon' > \varepsilon$ such that for any $i \leq d$, z_i is $(\varepsilon', \psi, \phi)$-independent from $(z_1, \dots, z_{i-1})$ with respect to $\mathcal{Q}$.*

Lemma D.1 (Generalization of Proposition 3 of Russo and Van Roy (2013))**.** *Let $\mathcal{Z}$ be a set and $\mathcal{Q}$ be a class of functions $q : \mathcal{Z} \to [-S, S]$. Let ψ and ϕ be two functions from $\mathcal{Q} \times \mathcal{Z} \to \mathbb{R}_+$. Additionally, let's assume that ϕ is uniformly bounded above by some $C > 0$. Suppose sequences $(z_1, q_1), \dots, (z_T, q_T) \in \mathcal{Z} \times \mathcal{Q}$ and $\beta > 0$ such that, for all $t \leq T$,*

$$\sum_{j \leq t} \psi(q_t, z_j) \leq \beta\,, \tag{13}$$

Then, for any $\varepsilon > 0$,

$$\sum_{j \leq T} \mathbb{1}_{[\phi(q_j, z_j) > \varepsilon^2]} \leq \left(\frac{\beta}{\varepsilon^2} + 1\right) d_{\psi,\phi}(\mathcal{Q}, \varepsilon)\,. \tag{14}$$

Proof of Lemma D.1. The proof is essentially the one of Proposition 3 of Russo and Van Roy (2013) that adapts almost directly. It is re-derived here for the sake of completeness. For readability, we will drop the dependence on ψ and ϕ in the notation of (in)dependence.

Claim 1. Let $\varepsilon > 0$. Then if for some $t \leq T$, we have $\phi(q_t, z_t) > \varepsilon^2$, then z_t is ε-dependent with respect to $\mathcal{Q}$ on at most β/ε^2 disjoint subsequences in $(z_1, \dots, z_{t-1})$.

Proof of Claim 1. For t such that $\phi(q_t, z_t) > \varepsilon^2$, let $(z_1, \dots, z_l)$ be a subsequence of $(z_1, \dots, z_{t-1})$ such that z_t is ε-dependent on $(z_1, \dots, z_l)$ with respect to $\mathcal{Q}$. Then, necessarily, $\sum_{i \leq l} \psi(q_t, z_i) > \varepsilon^2$ (otherwise, z_t would not be ε-dependent). Then, if z_t is ε-dependent on L disjoint subsequences of $(z_1, \dots, z_{t-1})$, we have $\sum_{i<t} \psi(q_t, z_i) > \varepsilon^2 L$. Since $\sum_{i<t} \psi(q_t, z_i) \leq \beta$, we have $L < \beta/\varepsilon^2$.

Claim 2. Let $\mathcal{Q}$ be a class of functions $q : \mathcal{Z} \to \mathbb{R}$. For any $\varepsilon > 0$ and sequence $(z_1, \dots, z_\tau) \in \mathcal{Z}^\tau$, there exists $j \leq \tau$ such that z_j is ε-dependent with respect to $\mathcal{Q}$ on at least $L \geq \tau / d_{\psi,\phi}(\mathcal{Q}, \varepsilon) - 1$ disjoint subsequences in $(z_1, \dots, z_{\tau-1})$.

Proof of Claim 2. Let choose an integer L such that $L d_{\psi,\phi}(\mathcal{Q}, \varepsilon) + 1 \leq \tau \leq (L+1) d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$. Let $B_i = (z_i)$ for all $i = 1, \dots, L$. If z_{L+1} is ε-dependent on each subsequence $B_1, \dots, B_L$, the claim is established. Otherwise, choose a subsequence B_i such that z_{L+1} is ε-independent of B_i and append it $B_i := B_i \cup (z_{L+1})$. Repeat this process for elements with indices $j > L+1$ until z_j is ε-dependent on each subsequence or $j = \tau$. If $j < \tau$, the proof is finished. Let's focus on the case $j = \tau$. Since for any $i \leq L$, each element of B_i is ε-independent of its predecessors, $|B_i| \leq d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$. Moreover, $\sum_{i \leq L} |B_i| = \tau - 1 \geq L d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$. Thus, for any $i \leq L$, $|B_i| = d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$. Then z_τ must be ε-dependent on each subsequence by definition of $d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$.

Putting everything together, let $(z_{t_1}, \dots, z_{t_\tau})$ be the sub-sequence of $(z_1, \dots, z_T)$ consisting in indexes t for which $\phi(q_t, z_t) > \varepsilon^2$. By applying Claim 1 and 2 to this sub-sequence, we obtain that

$$\left\lfloor \frac{\tau}{d_{\psi,\phi}(\mathcal{Q}, \varepsilon)} \right\rfloor - 1 < \frac{\beta}{\varepsilon^2} \Leftrightarrow \tau < \left(\frac{\beta}{\varepsilon^2} + 1\right) d_{\psi,\phi}(\mathcal{Q}, \varepsilon)$$

□

We now restate (a simplified version of) Lemma 21 from Croissant et al. (2024) before applying it.

Lemma D.2 (Lemma 21 in Croissant et al. (2024)). *Let $(z_t)_{t\in\mathbb{N}} \in [0,Z]^{\mathbb{N}}$, $(b_t)_{t\in\mathbb{N}} \in (\mathbb{R}\to\mathbb{R}_+)^{\mathbb{N}}$ a sequence of non-increasing functions and $a \in (\mathbb{R}\to\mathbb{R}_+)$ non-increasing such that for any $T>0$ and $\varepsilon>0$*

$$\sum_{t=1}^{T} \mathbb{1}_{[z_t>\varepsilon]} \le \frac{b_T(\varepsilon)}{\varepsilon^2} + a(\varepsilon) \quad \textit{and} \quad \varepsilon \le \frac{b_T(\varepsilon)\wedge\sqrt{b_T(\varepsilon)}}{\sqrt{T}} \tag{15}$$

Then, the following two inequalities hold,

$$\sum_{t=1}^{T} z_t \le 2\sqrt{T b_T(\varepsilon)} + Z a(\varepsilon) \tag{16}$$

$$\sum_{t=1}^{T} z_t^2 \le b_T(\varepsilon)\left(2+\log\left(\frac{TZ^2}{b_T(\varepsilon)^2}\right)\right) + a(\varepsilon)(2+b_T(\varepsilon))(1+Z^2) \tag{17}$$

Lemma D.3. *Let $\mathcal{Z}$ be a set and $\mathcal{Q}$ be a class of functions $q:\mathcal{Z}\to[-S,S]$. Let ψ and ϕ be two functions from $\mathcal{Q}\times\mathcal{Z}\to\mathbb{R}_+$. Additionally, let's assume that ϕ is uniformly bounded above by some $C>0$. Suppose sequences $(z_1,q_1),\dots,(z_T,q_T)\in\mathcal{Z}\times\mathcal{Q}$ and $\beta>0$ such that, for all $t\le T$,*

$$\sum_{j\le t} \psi(q_t,z_j) \le \beta\,, \tag{18}$$

Then, for any $0<\varepsilon\le\frac{\beta}{\sqrt{T}}$,

$$\sum_{t=1}^{T}\sqrt{\phi(q_t,z_t)} \le 2\sqrt{T\beta d_{\psi,\phi}(\mathcal{Q},\varepsilon)} + \sqrt{C} d_{\psi,\phi}(\mathcal{Q},\varepsilon) \tag{19}$$

$$\sum_{t=1}^{T}\phi(q_t,z_t) \le \beta d_{\psi,\phi}(\mathcal{Q},\varepsilon)\left(2+\log\left(\frac{TC}{\beta^2 d^2_{\psi,\phi}(\mathcal{Q},\varepsilon)}\right)\right) + d_{\psi,\phi}(\mathcal{Q},\varepsilon)(2+\beta d_{\psi,\phi}(\mathcal{Q},\varepsilon))(1+C)\,. \tag{20}$$

Proof of Lemma D.3. Apply Lemma D.2 with $z_t=\sqrt{\phi(q_t,z_t)}$, $b_t(\varepsilon)=\beta d_{\psi,\phi}(\mathcal{Q},\varepsilon)$ and $a(\varepsilon)=d_{\psi,\phi}(\mathcal{Q},\varepsilon)$, noticing that

1. $d_{\psi,\phi}(\mathcal{Q},\varepsilon)$ is non-increasing as a function of ε,
2. as $d_{\psi,\phi}(\mathcal{Q},\cdot)\ge 1$, the condition $\varepsilon\le\frac{\beta}{\sqrt{T}}$ is sufficient to satisfy the r.h.s. of Equation (15).

□

E Instantiation to Familiar Utility Classes

E.1 Linear Utility

In this setting, $\mathcal{X}\subset\mathbb{B}_2^d(1)$ and, for any $x\in\mathcal{X}$, the reward function is given by $f^\star(x)=\langle\theta^\star,x\rangle$, where $\theta^\star\in\mathbb{R}^d$ is an unknown parameter satisfying $\|\theta^\star\|\le S$. Accordingly, the class of linear functions $\mathcal{F}_\Theta$ can be identified with the parameter set $\Theta\subset\mathbb{B}_2^d(S)$. Under the Bradley–Terry model, the preference feedback satisfies $y_t\sim\text{Bernoulli}(\sigma(\langle\theta^\star,x_t-x_t'\rangle))$.

Algorithmic Implementation. Let $\hat\theta_t$ denote the MLE given by Equation (1). The corresponding confidence set becomes: $\Theta_t=\left\{\theta\in\Theta:\mathcal{L}(\theta;\mathcal{H}_{t-1})\le\mathcal{L}(\hat\theta_t;\mathcal{H}_{t-1})+\beta_t(\delta)\right\}$. This set can be further enclosed within an ellipsoid of the following form (see Lee et al. (2024) for instance): $\Theta_t\subset\left\{\theta\in\Theta:\|\theta-\hat\theta_t\|^2_{\nabla^2\mathcal{L}(\hat\theta_t;\mathcal{H}_{t-1})}\le 2(1+2S)\beta_t(\delta)\right\}$. Moreover, the uncertainty on the performance gap in Equation (3) admits a closed-form characterization:

$$\forall(x,x')\in\mathcal{X}^2,\ \omega_t(x,x')^2\propto\beta_t\|x-x'\|^2_{(\nabla^2\mathcal{L}(\hat\theta_t;\mathcal{H}_{t-1}))^{-1}} \overset{\text{Faury et al. (2020)}}{\propto} \beta_t\|x-x'\|^2_{H_t^{-1}}$$

where $H_t=\sum_{s=1}^{t-1}\dot\sigma((x_s-x_s')^\top\hat\theta_t)(x_s-x_s')(x_s-x_s')^\top$. Therefore, our approach differs from prior work in two ways. First, regarding the decision rule, existing methods typically proceed in

two stages: either by constructing a set of plausible optimal arms and then selecting a pair that maximizes an exploration bonus (see, e.g., Saha (2021); Di et al. (2024)), or by first selecting a reference arm using an upper-confidence bound and then choosing a second arm that maximizes a relative optimistic criterion (see, e.g., Bengs et al. (2022)). In contrast, our algorithm selects the pair of arms jointly in a single step. Second, the uncertainty bonus is fundamentally different. Prior approaches (Saha, 2021; Bengs et al., 2022; Li et al., 2024) rely on quantities of the form $\|x-x'\|_{V_t^{-1}}$, where $V_t = \sum_{i=1}^{t-1}(x_i - x'_i)(x_i - x'_i)^\top$. In contrast, **GINOP** leverages a locally adaptive design matrix H_t that incorporates the curvature of the sigmoid through $\dot{\sigma}$. This refinement more accurately captures the nature of observations and avoids the undesirable dependence on κ in the leading term of the regret, as highlighted in the subsequent analysis.

To correctly assess **GINOP** regret in the case of linear utility, we need first to evaluate the key complexity quantities appearing in the theorem, namely $\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))\right)$ and $d_\sigma(\Delta\mathcal{F}, \Delta_{f^\star}, \varepsilon)$, when $\mathcal{F} = \mathcal{F}_{lin}$.

Proposition 6.1. *In the linear reward setting, the following bounds hold:*

$$\forall\, \varepsilon > 0, \quad d_\sigma(\Delta\mathcal{F}_{lin}, \Delta_{f^\star}, \varepsilon) = \mathcal{O}\big(d\log(1+1/\varepsilon)\big) \quad \textit{and} \quad \log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}_{lin}))\right) = \mathcal{O}\big(d\log T\big).$$

Proof of Proposition 6.1.

We start with $d_\sigma(\Delta\mathcal{F}_{lin}, \Delta_{f^\star}, \varepsilon)$. Let $((x_1, x'_1), \dots, (x_n, x'_n)) \in \mathcal{X}^{2n}$ and $(\theta_0, \dots, \theta_n) \subseteq \mathbb{B}_2^d(S)$ be the longest ε-independent sequence. For simplicity we denote by $z_i = x_i - x'_i$. Then there exists $\varepsilon' > \varepsilon$ such that for all $t \le n$,

$$\sum_{i=1}^{t-1} \bar{\varphi}(\sigma(z_i^\top\theta^\star), \sigma(z_i^\top\theta_t)) \le \varepsilon', \qquad \dot{\sigma}(z_t^\top\theta^\star)(z_t^\top\left(\theta^\star - \theta_t\right))^2 \ge \varepsilon'.$$

A second-order Taylor expansion yields, for all $i < t \le n$,

$$\bar{\varphi}(\sigma(z_i^\top\theta^\star), \sigma(z_i^\top\theta_t)) = \bar{\varphi}(\sigma(z_i^\top\theta^\star), \sigma(z_i^\top\theta^\star)) + \left[\sigma(z_i^\top\theta^\star) - \sigma(z_i^\top\theta^\star)\right] z_i^\top(\theta_t - \theta^\star) + \left[z_i^\top(\theta_t - \theta^\star)\right]^2 \underbrace{\int_0^1 (1-\nu)\dot{\sigma}(z_i^\top\theta^\star + \nu z_i^\top(\theta_t - \theta^\star))d\nu}_{(A)}.$$

Using Lemma 8 of Abeille et al. (2021)., we obtain

$$(A) \ge \frac{\dot{\sigma}(z_i^\top\theta^\star)}{2 + \left|z_i^T(\theta_t - \theta^\star)\right|} \ge \frac{\dot{\sigma}(z_i^\top\theta^\star)}{2(1+S)}.$$

Hence, for all $t \le n$,

$$\sum_{i=1}^{t-1} \dot{\sigma}(z_i^\top\theta^\star)(z_i^\top(\theta^\star - \theta_t))^2 \le 2(1+S)\varepsilon', \tag{21}$$

while

$$\dot{\sigma}(z_t^\top\theta^\star)(z_t^\top(\theta^\star - \theta_t))^2 \ge \varepsilon'. \tag{22}$$

Define

$$\bar{H}_t(\theta^\star) = \sum_{i=1}^{t-1} \dot{\sigma}(z_i^\top\theta^\star) z_i z_i^\top, \qquad H_t(\theta^\star) = \bar{H}_t(\theta^\star) + \lambda I,$$

with $\lambda = \varepsilon' 2(1+S)/(2S)^2$. From equation (21):

$$\|\theta^\star - \theta_t\|^2_{H_t(\theta^\star)} \le \|\theta^\star - \theta_t\|^2_{\bar{H}_t(\theta^\star)} + \lambda(2S)^2 \le 4(1+S)\varepsilon'$$

From equation (22):

$$\varepsilon' \le \dot{\sigma}(z_t^\top\theta_t)\|z_t\|^2_{H_t(\theta^\star)^{-1}}\|\theta_t - \theta^\star\|^2_{H_t(\theta^\star)} \le 4(1+S)\varepsilon'\dot{\sigma}(z_t^\top\theta^\star)\|z_t\|^2_{H_t(\theta^\star)^{-1}}$$
$$\implies \dot{\sigma}(z_t^\top\theta^\star)\|z_t\|^2_{H_t(\theta^\star)^{-1}} \ge \frac{1}{4(1+S)}.$$

Using $|H_t(\theta^\star)| = |H_{t-1}(\theta^\star)| \left(1 + \dot{\sigma}(z_t^\top \theta^\star)\|z_t\|^2_{H_t(\theta^\star)^{-1}}\right)$ we obtain:

$$|H_n(\theta^\star)| \geq \left(1 + \frac{1}{4(1+S)}\right)^{n-1} \lambda^d$$

Further:

$$|H_n(\theta^\star)| \leq \left(\frac{Tr(H_n(\theta^\star))}{d}\right)^d \leq \left(\lambda + \frac{n-1}{4d}\right)^d$$

which implies :

$$\left(1 + \frac{n-1}{4d\lambda}\right) \geq \left(1 + \frac{1}{4(1+S)}\right)^{\frac{n-1}{d}}$$

The rest of the proof directly follows Russo and Van Roy (2013)'s proof of Proposition 6 (step 3). We obtain:

$$n \leq 1 + d\left[10(1+S)\log\left(1 + \frac{2(1+S)}{\varepsilon'}\right) + \ln 5(1+S)\right]$$

And hence:

$$n = \mathcal{O}\left(d \log\left(1 + 1/\epsilon\right)\right).$$

Bounding $\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))\right)$**.** Let $\Phi(\Delta_{f_\theta}) \in \Phi(\Delta\mathcal{F}_{lin})$. For all $y \in [0,1]$ and $x, x' \in \mathcal{X}$, we have

$$\begin{aligned}
\Phi(\Delta_{f_\theta})(y, x, x') &= \ell\big(y, \sigma(\Delta_{f_\theta}(x, x'))\big) - \ell\big(y, \sigma(\Delta_{f_{\theta^\star}}(x, x'))\big), \\
\nabla_\theta \Phi(\Delta_{f_\theta})(y, x, x') &= \nabla_\theta \ell\big(y, \sigma(\Delta_{f_\theta}(x, x'))\big) \\
&= \big(\sigma(\langle \theta, x - x' \rangle) - y\big)(x - x').
\end{aligned}$$

Since $\sigma(\cdot) \in [0,1]$ and $\|x - x'\| \leq 2$, it follows that

$$\|\nabla_\theta \Phi(\Delta_{f_\theta})(y, x, x')\| \leq 2.$$

Hence, the mapping $\theta \mapsto \Phi(\Delta_{f_\theta})$ is Lipschitz continuous. By Example 3 in Russo and Van Roy (2013), we obtain

$$\log\left(\mathcal{N}_T(\Phi(\Delta\mathcal{F}))\right) = \mathcal{O}(d \log T).$$

□

E.2 Kernelized Utility

Similar to (Pásztor et al., 2024; Xu et al., 2024; Kayal et al., 2025), we assume that the utility function $f^\star$ belongs to a known Reproducing Kernel Hilbert Space (RKHS). Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel, and let $\mathcal{H}_k$ denote the associated RKHS. The space $\mathcal{H}_k$ is equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$. We assume $\|f^\star\|_{\mathcal{H}_k} \leq S$ and $k(x, x) \leq 1$ for all $x \in \mathcal{X}$. By the reproducing property, for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, it holds that $\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x)$. By Mercer's theorem, under mild conditions, the kernel admits the representation $k(x, x') = \sum_{m=1}^\infty \gamma_m \phi_m(x)\phi_m(x')$, where $\gamma_m > 0$, and $\{\psi_m := \sqrt{\gamma_m}\phi_m\}_{m \geq 1}$ forms an orthonormal basis of $\mathcal{H}_k$. In particular, any $f \in \mathcal{H}_k$ can be expressed as $f(\cdot) = \sum_{m=1}^\infty \theta_m \psi_m(\cdot) = \theta^\top \Psi(\cdot)$, with $\|f\|^2_{\mathcal{H}_k} = \sum_{m=1}^\infty \theta_m^2 \leq S$. We refer to $\{\gamma_m\}$ and $\{\phi_m\}$ as the (Mercer) eigenvalues and eigenfunctions of k, respectively.

Define $z = (x, x') \in \mathcal{X} \times \mathcal{X}$ and $\Delta_f(z) = f(x) - f(x')$. Following Pásztor et al. (2024); Kayal et al. (2025), we introduce the *dueling kernel*

$$\mathbb{k}(z_1, z_2) = k(x_1, x_2) + k(x_1', x_2') - k(x_1, x_2') - k(x_1', x_2),$$

for $z_1 = (x_1, x_1')$ and $z_2 = (x_2, x_2')$. This construction satisfies $\|\Delta_f\|_{\mathcal{H}_\mathbb{k}} = \|f\|_{\mathcal{H}_k}$ (see (Pásztor et al., 2024, Proposition 4)). Moreover, both f and Δ_f share the same Mercer coefficient vector $\theta \in \ell_2(\mathbb{N})$; that is, for all $x, x' \in \mathcal{X}$, $f(x) = \theta^\top \Psi(x)$ and $\Delta_f(x, x') = \theta^\top\big(\Psi(x) - \Psi(x')\big)$.

Algorithmic Implementation. Invoking Mercer's theorem, and following a similar approach to (Pásztor et al., 2024; Kayal et al., 2025), we consider the regularized negative log-likelihood loss

$$\mathcal{L}_{\Bbbk}(\theta; \mathcal{H}_{t-1}) := \sum_{s=1}^{t-1} \ell\Big(y_s, \sigma\big(\theta^\top(\Psi(x_s) - \Psi(x'_s))\big)\Big) + \frac{\lambda}{2}\|\theta\|^2.$$

The MLE estimator is defined as $\hat{f}_t = \hat{\theta}_t^\top \Psi$, where $\hat{\theta}_t = \operatorname{argmin}_{\theta \in \ell_2(\mathbb{N})} \mathcal{L}_{\Bbbk}(\theta; \mathcal{H}_{t-1})$. A tractable counterpart of this infinite-dimensional optimization problem is obtained by leveraging the Representer Theorem. Specifically, $\alpha_t = \operatorname{argmin}_{\alpha \in \mathbb{R}^{t-1}} \sum_{s=1}^{t-1} \ell\Big(y_s, \sigma\big(\alpha^\top \Bbbk_{t-1}(x_s, x'_s)\big)\Big) + \frac{\lambda}{2}\|\alpha\|^2$, where $\Bbbk_{t-1}(z) = \big[\Bbbk(z, (x_j, x'_j))\big]_{j=1}^{t-1}$ denotes the vector of dueling-kernel evaluations between the pair z and the past observed pairs. This yields the explicit expression

$$\forall x \in \mathcal{X}, \qquad \hat{f}_t(x) = \hat{\theta}_t^\top \Psi(x) = \big\langle \alpha_t, k_{t-1}(x) - k'_{t-1}(x) \big\rangle,$$

where $k_{t-1}(x) = \big[k(x, x_j)\big]_{j=1}^{t-1}$, $k'_{t-1}(x) = \big[k(x, x'_j)\big]_{j=1}^{t-1}$. As in the linear case, the confidence width over the pair space admits the kernelized form

$$\begin{aligned}\omega_t(x, x')^2 &\propto \beta_t \big\|\Psi(x) - \Psi(x')\big\|^2_{H_t^{-1}} \\ &\propto \frac{\beta_t}{\lambda}\Big(\Bbbk(z,z) - \Bbbk_{t-1}(z)^\top W_{t-1}^{1/2}\big(W_{t-1}^{1/2}\mathbb{K}_{t-1}W_{t-1}^{1/2} + \lambda I\big)^{-1} W_{t-1}^{1/2}\Bbbk_{t-1}(z)\Big),\end{aligned}$$

where $z = (x, x')$, $H_t = \lambda I_{\mathcal{H}_{\Bbbk}} + \sum_{s=1}^{t-1} \dot{\sigma}\big(\hat{f}_t(x_s) - \hat{f}_t(x'_s)\big)\big(\Psi(x_s) - \Psi(x'_s)\big)\big(\Psi(x_s) - \Psi(x'_s)\big)^\top$, $W = \operatorname{diag}(\dot{\sigma}(\Delta_{\hat{f}_t}(z_1)), \dots, \dot{\sigma}(\Delta_{\hat{f}_t}(z_{t-1})))$ and $\mathbb{K}_{t-1} = \big[\Bbbk\big((x_i, x'_i), (x_j, x'_j)\big)\big]_{i,j=1}^{t-1}$ is the dueling-kernel Gram matrix over the observed pairs.

Proposition 6.3. *In the kernelized setting, the following bound holds:*

$$\forall\, \varepsilon > 0, \qquad d_\sigma(\Delta_{\mathcal{F}_k}, \Delta_{f^\star}, \varepsilon) = \mathcal{O}\Big(\gamma_T(\lambda; \mathcal{X} \times \mathcal{X})\Big), \quad \textit{with } \lambda = \frac{2\varepsilon(1+S)}{(2S)^2}.$$

Proof of Proposition 6.3. Let $\big((x_1, x'_1), \dots, (x_n, x'_n)\big) \in \mathcal{X}^{2n}$ and $(f_1, \dots, f_n) \subseteq \mathcal{F}_k$ be the longest ε-independent sequence, i.e $d_\sigma(\Delta_{\mathcal{F}_k}, \Delta_{f^\star}, \varepsilon) = n$. For brevity, we denote $z_i = \Psi(x_i) - \Psi(x'_i)$ and consider $(\theta_1, \dots, \theta_n) \subseteq \ell_2(\mathbb{N})$ such that $f_i = \theta_i^\top \Psi$ for all $i \in [n]$. Then there exists $\varepsilon' > \varepsilon$ such that for all $t \le n$,

$$\sum_{i=1}^{t-1} \bar{\varphi}(\sigma(z_i^\top \theta^\star), \sigma(z_i^\top \theta_t)) \le \varepsilon', \qquad \dot{\sigma}\big(z_t^\top \theta^\star\big)\big(z_t^\top(\theta^\star - \theta_t)\big)^2 \ge \varepsilon'.$$

A second-order Taylor expansion yields, for all $i < t \le n$,

$$\begin{aligned}\bar{\varphi}(\sigma(z_i^\top \theta^\star), \sigma(z_i^\top \theta_t)) = \bar{\varphi}(\sigma(z_i^\top \theta^\star), \sigma(z_i^\top \theta^\star)) &+ \big[\sigma(z_i^\top \theta^\star) - \sigma(z_i^\top \theta^\star)\big]\, z_i^\top(\theta_t - \theta^\star) \\ &+ \big[z_i^\top(\theta_t - \theta^\star)\big]^2 \underbrace{\int_0^1 (1-\nu)\dot{\sigma}(z_i^\top \theta^\star + \nu z_i^\top(\theta_t - \theta^\star))d\nu}_{(A)}.\end{aligned}$$

Using Lemma 8 of Abeille et al. (2021)., we obtain

$$(A) \ge \frac{\dot{\sigma}(z_i^\top \theta^\star)}{2 + \big|z_i^T(\theta_t - \theta^\star)\big|} \ge \frac{\dot{\sigma}(z_i^\top \theta^\star)}{2(1+S)}.$$

Hence, for all $t \le n$,

$$\sum_{i=1}^{t-1} \dot{\sigma}(z_i^\top \theta^\star)(z_i^\top(\theta^\star - \theta_t))^2 \le 2(1+S)\varepsilon', \tag{23}$$

while

$$\dot{\sigma}(z_t^\top \theta^\star)(z_t^\top(\theta^\star - \theta_t))^2 \ge \varepsilon'. \tag{24}$$

Define

$$\bar{H}_t(\theta^\star) = \sum_{i=1}^{t-1} \dot{\sigma}(z_i^\top \theta^\star) z_i z_i^\top, \qquad H_t(\theta^\star) = \bar{H}_t(\theta^\star) + \lambda I,$$

with $\lambda = 2\varepsilon(1+S)/(2S)^2$. From equation (23):

$$\|\theta^\star - \theta_t\|^2_{H_t(\theta^\star)} \leq \|\theta^\star - \theta_t\|^2_{\bar{H}_t(\theta^\star)} + (2S)^2\lambda \leq 2(1+S)\varepsilon' + (2S)^2\lambda \leq 2(1+S)(\varepsilon' + \varepsilon)$$

From equation (24):

$$\begin{aligned}\varepsilon' \leq \dot{\sigma}(z_t^\top \theta_t)\|z_t\|^2_{H_t(\theta^\star)^{-1}}\|\theta_t - \theta^\star\|^2_{H_t(\theta^\star)} &\leq 2(1+S)(\varepsilon'+\varepsilon)\dot{\sigma}(z_t^\top\theta^\star)\|z_t\|^2_{H_t(\theta^\star)^{-1}} \\ \implies \dot{\sigma}(z_t^\top\theta^\star)\|z_t\|^2_{H_t(\theta^\star)^{-1}} &\geq \frac{\varepsilon'}{2(1+S)(\varepsilon'+\varepsilon)} \geq \frac{1}{4(1+S)}.\end{aligned}$$

Using that $\det\Big(\lambda^{-1}H_t(\theta^\star)\Big) = \det\Big(\lambda^{-1}H_{t-1}(\theta^\star)\Big)\Big(1 + \dot{\sigma}(z_t^\top\theta^\star)\|z_t\|^{-1}_{H_t(\theta^\star)}\Big)$ we obtain:

$$\det\Big(\lambda^{-1}H_n(\theta^\star)\Big) \geq \det\Big(I\Big)\Big(1 + \frac{1}{4(1+S)}\Big)^n \geq \Big(1 + \frac{1}{4(1+S)}\Big)^n$$

Given that $\dot{\sigma}(\cdot) \leq 1$ and the definition of the maximum information gain, we obtain:

$$\begin{aligned}&\gamma_T(\lambda;\, \mathcal{X}\times\mathcal{X}) \geq \log\Big(\det\Big(\lambda^{-1}H_n(\theta^\star)\Big)\Big), \\ &\implies n \leq \left(\log\left(1 + \frac{1}{4(1+S)}\right)\right)^{-1} \gamma_T(\lambda;\, \mathcal{X}\times\mathcal{X}).\end{aligned}$$

□

E.2.1 Specialization to Commonly Used Kernel Functions

Combining Theorem 5.1 and Proposition 6.3 with the fact that, for bounded RKHS, $\log \mathcal{N}_T \asymp \gamma_T$, the regret upper bound of **GINOP** in the kernelized setting becomes:

$$\mathcal{R}_T = \mathcal{O}\left(\gamma_T\sqrt{\frac{T}{\dot{\sigma}^\star}} + \kappa\,\gamma_T^3\right).$$

In what follows, we specialize this result to several commonly used kernel functions.

Linear kernel. The linear kernel is defined as $k(x,x') = x^\top x'$ where $x, x' \in \mathbb{R}^d$ and, according to Vakili et al. (2021), $\gamma_T = \mathcal{O}(d\log T)$. Hence, $\gamma_T^3 = o(\sqrt{T})$, and consequently

$$\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\gamma_T\sqrt{\frac{T}{\dot{\sigma}^\star}}\right) = \widetilde{\mathcal{O}}\Big(2d\log(T)\sqrt{T}\Big). \tag{25}$$

Squared exponential (SE) kernel. The SE kernel is defined as

$$k(x,x') = \sigma^2_{\mathrm{SE}}\exp\left(-\frac{\|x-x'\|^2}{\ell^2}\right), \qquad x, x' \in \mathbb{R}^d,$$

where σ^2_{SE} denotes the variance parameter and ℓ the length-scale parameter. According to Vakili et al. (2021), $\gamma_T = \mathcal{O}\Big(\log^{d+1}(T)\Big)$. Hence, $\gamma_T^3 = o(\sqrt{T})$, and consequently

$$\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\gamma_T\sqrt{\frac{T}{\dot{\sigma}^\star}}\right) = \widetilde{\mathcal{O}}\Big(2\log^{d+1}(T)\sqrt{T}\Big). \tag{26}$$

Matérn kernel. The Matérn kernel is defined as

$$k(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,\|x-x'\|}{\rho} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\,\|x-x'\|}{\rho} \right), \qquad x, x' \in \mathbb{R}^d,$$

where $\rho, \nu > 0$ are kernel parameters, $\Gamma(\cdot)$ denotes the gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. The parameter ν controls the smoothness of the kernel and, following Xu et al. (2024), is assumed to be large enough such that $\nu > 3d/2$. According to Vakili et al. (2021), $\gamma_T = \mathcal{O}\left(T^{\frac{d}{2\nu+d}} \log(T)\right)$. Hence, $\gamma_T^3 = o(\gamma_T \sqrt{T})$, and consequently

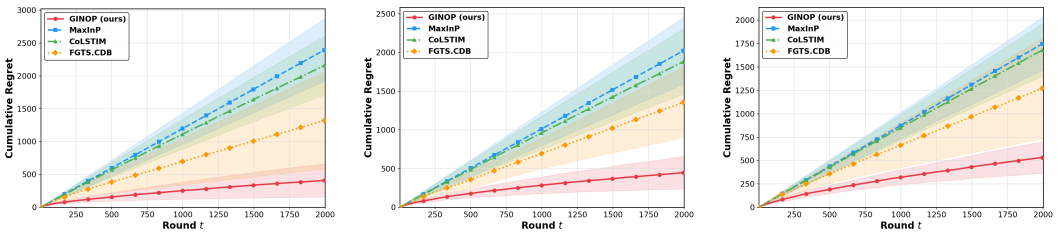
$$\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\gamma_T \sqrt{\frac{T}{\dot{\sigma}^\star}}\right) = \widetilde{\mathcal{O}}\left(2\log(T)\, T^{\frac{1}{2}+\frac{d}{2\nu+d}}\right). \tag{27}$$

As a result, by Equations (25) to (27) and in line with Kayal et al. (2025), **GINOP** exhibits the correct dependence on κ and improves upon the bounds of Xu et al. (2024) and Pásztor et al. (2024), which are of order $\widetilde{\mathcal{O}}\left((\gamma_T T)^{3/4}\right)$ and $\widetilde{\mathcal{O}}\left(\gamma_T \kappa^2 \sqrt{T}\right)$, respectively.

F Details of Experiments

To corroborate our theoretical results, we conduct numerical experiments evaluating the performance of **GINOP** against several existing baselines across a range of reward function classes. For all experiments, we report the cumulative regret over a horizon of $T = 2000$ rounds, averaged over 20 independent trials. All experiments were conducted on a machine featuring an Apple M1 chip (8 cores) and 16 GB of RAM.

Linear Setting. We consider the linear utility setting in which $f^\star(x) = \langle x, \theta^\star \rangle$ for all $x \in \mathcal{X}$, with $\theta^\star \in \mathbb{B}_2^d(S)$, $d \in \{10, 15, 20\}$, $S = 2$, and $|\mathcal{X}| = 30$, with the arms drawn uniformly at random from $\{\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$. We benchmark against **MaxInP** (Saha, 2021), **CoLSTIM** (Bengs et al., 2022), and **FGTS.CDB** (Li et al., 2024), which, despite their differing decision rules, all attain a regret bound of order $\kappa d\sqrt{T}$. While **MaxInP** requires no specific hyperparameter tuning, for **CoLSTIM** we follow the recommendations of the original paper and set $c = C_{\text{thresh}} = \eta = \sqrt{d \log T}$, $\tau = t_0 = d\,|\mathcal{X}|$, and $p_t = \min\left(1, \sqrt{\frac{d}{t-\tau}\log(dT)}\right)$; for **FGTS.CDB**, in accordance with their Theorem 5.2, we set $\eta = 0.25$ and $\mu = 1/(10e^S\sqrt{T})$. The results, reported in Figure 2, indicate that **GINOP** consistently outperforms the competing baselines: due to their adverse dependence on κ, the baselines' regret is inflated, exhibiting a $\sqrt{T}$-like shape that becomes apparent only at considerably larger horizons.



(a) $d = 10$. (b) $d = 15$. (c) $d = 20$.

Figure 2: Linear Utility.

Kernelized Setting. We adopt the Ackley function as the reward on the interval $[-5, 5]$. The Ackley function has a diverse optimization landscape, featuring multiple local minima, flat plateaus, and valleys, which makes it a popular benchmark in the non-convex optimization literature. It is defined as (here we consider $d = 1$):

$$f(x) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)\right) + 20 + \exp(1).$$

We consider 30 arms forming a uniform mesh over the input domain, and employ the Matérn kernel with smoothness parameter $\nu = 2.5$ and lengthscale $\rho = 0.1$. We benchmark against **POP-BO** (Xu et al., 2024), **MaxMinLCB** (Pásztor et al., 2024), and **MR-LPF** (Kayal et al., 2025). The results are reported in Figure 3. **GINOP** clearly outperforms **POP-BO** and **MaxMinLCB**: the former scales as $(\gamma_T T)^{3/4}$, while the latter incurs an additional adverse dependence on κ in the leading term. **MR-LPF**, in contrast, exhibits a regret profile comparable to that of **GINOP**, as its multi-phase approach allows it to eliminate the κ-dependence.

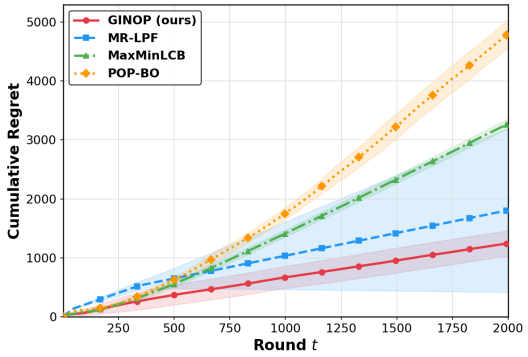


Figure 3: Kernelized Utility.

Against a Neural-Network Approach. In this setting, the reward is generated by $f^\star(x) = \cos(3\langle x, \theta^\star \rangle)$, with $\theta^\star \in \mathbb{R}^d$, where we take $d = 10$ and $|\mathcal{X}| = 30$, with the arms drawn uniformly at random from $\{-1, 1\}^d$. We benchmark against **NDB-UCB** and **NDB-TS** (Verma et al., 2025), which employ a neural network in the NTK regime to estimate the reward function and select arms accordingly. Following the original paper, we use a neural network with 2 hidden layers of width 50 and ReLU activation functions, with hyperparameters $\lambda = 1.0$, $\delta = 0.05$, and a fixed $\nu_T = \nu = 1.0$. The results, reported in Figure 4, show that **GINOP** outperforms both **NDB-UCB** and **NDB-TS**, in line with the theory: the regret bounds of **NDB-UCB** and **NDB-TS** exhibit an adverse dependence on κ.

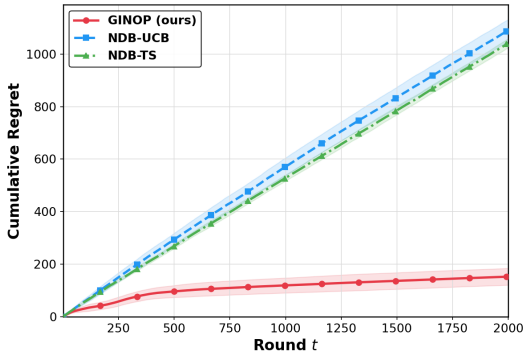


Figure 4: Against a NN.