Bandits with Preference Feedback: A Stackelberg Game Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bandits with preference feedback present a powerful tool for optimizing unknown target functions when only pairwise comparisons are allowed instead of direct value queries. This model allows for incorporating human feedback into online inference and optimization and has been employed in systems for tuning large language models. The problem is well understood in simplified settings with linear target functions or over finite small domains that limit practical interest. Taking the next step, we consider infinite domains and nonlinear (kernelized) rewards. In this setting, selecting a pair of actions is quite challenging and requires balancing exploration and exploitation at two levels: within the pair, and along the iterations of the algorithm. We propose MAXMINLCB, which emulates this trade-off as a zero-sum Stackelberg game, and chooses action pairs that are informative and yield favorable rewards. MAXMINLCB consistently outperforms existing algorithms and satisfies an anytime-valid rate-optimal regret guarantee. This is due to our novel preference-based confidence sequences for kernelized logistic estimators.

1 Introduction

In standard bandit optimization, a learner repeatedly interacts with an unknown environment that gives numerical feedback on the chosen actions according to a utility function f. However, in applications such as fine-tuning large language models, drug testing, or search engine optimization, the quantitative value of design choices or test outcomes are either not directly observable, or are known to be inaccurate, or systematically biased, e.g., if they are provided by human feedback [Casper et al., 2023]. A solution is to optimize for the target based on comparative feedback provided for a pair of queries, which is proven to be more robust to certain biases and uncertainties in the queries [Ji et al., 2023].

Bandits with preference feedback, or *dueling* bandits, address this problem and propose strategies for choosing query/action pairs that yield a high utility over the horizon of interactions. At the core of such strategies is uncertainty quantification and inference for f in regions of interest, which is closely tied to exploration and exploitation dilemma over a course of queries. Observing only comparative feedback poses an additional challenge, as we now need to balance this trade-off *jointly* over two actions. This challenge is further exacerbated when optimizing over vast or infinite action domains. As a remedy, prior work often *grounds* one of the actions by choosing it either randomly or greedily, and tries to balance exploration-exploitation for the second action as a reaction to the first [Ailon et al., 2014, Zoghi et al., 2014a, Kirschner and Krause, 2021, Mehta et al., 2023b]. This approach works well for simple utility functions over low-dimensional domains, however does not scale to more complex problems.

Aiming to solve this problem, we focus on continuous domains in the Euclidean vector space and complex utility functions that belong to the Reproducing Kernel Hilbert Space (RKHS) of a potentially non-smooth kernel. We propose MAXMINLCB, a sample-efficient algorithm, that at every step chooses the actions *jointly*, by playing a zero-sum Stackelberg (Leader-Follower) game. We choose the Lower Confidence Bound (LCB) of f as the objective of this game which the Leader aims to

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

maximize and the Follower to minimize. The equilibrium of this game yields an action pair in which the first action is a favorable candidate to maximize f and the second action is the strongest competitor against the first. Our choice of using the LCB as the objective leads to robustness against uncertainty when selecting the first action. Moreover, it makes the second action an optimistic choice as a competitor, from its own perspective. We observe empirically that this approach creates a natural exploration scheme, and in turn, yields a more sample-efficient algorithm compared to standard baselines.

Our game-theoretic strategy leads to an efficient bandit solver, if the LCB is a valid and tight lower bound on the utility function. To this end, we construct a confidence sequence for f given pairwise preference feedback, by modeling the noisy comparative observations with a logistic-type likelihood function. Our confidence sequence is anytime valid and holds uniformly over the domain, under the assumption that f resides in an RKHS. We improve prior work by removing or relaxing assumptions on the utility while maintaining the same rate of convergence. This result allows us to prove a sublinear regret bound for MaxMinLCB, and may be of independent interest, as it targets the loss function that is typically used for Reinforcement Learning with Human Feedback.

Contributions Our main contributions are:

- We propose a novel game-theoretic acquisition function for pairwise action selection with preference feedback.
- We construct preference-based confidence sequences for kernelized utility functions that are tight and anytime valid.
- Together this creates MaxMinLCB, an algorithm for bandit optimization with preference feedback over continuous domains. MaxMinLCB satisfies $\mathcal{O}(\gamma_T\sqrt{T})$ regret, where T is the horizon and γ_T is the *information gain* of the kernel.
- We benchmark MaxMinLCB over a set of standard optimization problems and consistently outperform the most common action selection algorithms from the literature.

2 Related Work

Learning with indirect feedback was first studied in supervised preference learning [Aiolli and Sperduti, 2004, Chu and Ghahramani, 2005]. Subsequently, online and sequential settings were considered, motivated by applications in which the feedback is provided in an online manner, e.g., by a human [Yue et al., 2012, Yue and Joachims, 2009, Houlsby et al., 2011]. Bengs et al. [2021] surveys this field comprehensively; here we include a brief background.

Referred to as dueling bandits, a rich body of work considers (finite) multi-armed domains and learns a preference matrix specifying the relation among the arms. Such work often relies on efficient sorting or tournament systems based on the frequency of wins for each action [Jamieson and Nowak, 2011, Zoghi et al., 2014b, Falahatgar et al., 2017]. Rather than jointly selecting the arms, such strategies often simplify the problem by selecting one 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]. In contrast, we jointly optimize both actions by choosing them as the equilibrium of a two-player zero-sum Stackelberg game, enabling a more efficient exploration/exploitation trade-off.

The multi-armed dueling setting, which is reducible to multi-armed bandits [Ailon et al., 2014], naturally fails to scale to infinite compact domains, since regularity among "similar" arms is not exploited. To go beyond finite domains, *utility-based* dueling bandits consider an unknown latent function that captures the underlying preference, instead of relying on a preference matrix. The preference feedback is then modeled as the difference in the utility of two chosen actions passed through a link function. Early work is limited to convex domains and imposes strong regularity assumptions [Yue and Joachims, 2009, Kumagai, 2017]. These assumptions are then relaxed to general compact domains if the utility function is linear [Dudík et al., 2015, Saha, 2021, Saha and Krishnamurthy, 2022]. Constructing valid confidence sets from comparative preference feedback is a challenging task. However, it is strongly related to uncertainty quantification with direct logistic feedback, which is extensively analyzed by the literature on logistic and generalized linear bandits [Filippi et al., 2010, Faury et al., 2020, Foster and Krishnamurthy, 2018, Beygelzimer et al., 2019, Faury et al., 2022, Lee et al., 2024].

Preference-based bandit optimization with linear utility functions is fairly well understood and even extends to reinforcement learning with preference feedback on trajectories [Saha et al., 2023, Zhan et al., 2023, Zhu et al., 2023, Ji et al., 2023]. However, such approaches have limited practical interest, since they cannot capture real-world problems with complex nonlinear utility functions. Alternatively, Reproducing Kernel Hilbert Spaces (RKHS) provide a rich model class for the utility, e.g., if the chosen kernel is universal. Many have proposed heuristic algorithms for bandits and Bayesian optimization in kernelized settings, albeit without providing theoretical guarantees Brochu et al. [2010], González et al. [2017], Sui et al. [2017], Tucker et al. [2020], Mikkola et al. [2020], Takeno et al. [2023].

There have been attempts to prove convergence of kernelized algorithms for preference-based bandits [Xu et al., 2020, Kirschner and Krause, 2021, Mehta et al., 2023b,a]. Such works employ a regression likelihood model which requires them to assume that both the utility and the probability of preference, as a function of actions, lie in an RKHS. In doing so, they use a regression model for solving a problem that is inherently of a classification nature. While the model is valid, it does not result in a sample-efficient algorithm. In contrast, we use a kernelized logistic negative log-likelihood loss to infer the utility function, and provide confidence sets for its minimizer. In a concurrent work, Xu et al. [2024] also consider the kernelized logistic likelihood model and propose a variant of the MULTISBM algorithm [Ailon et al., 2014] using likelihood ratio-based confidence sets. The theoretical approach and resulting algorithm bear significant differences, and the regret guarantee has a strictly worse dependency on the time horizon T, by a factor of $T^{1/4}$. This is discussed in more detail in Section 5.

3 Problem Setting

Consider an agent which repeatedly interacts with an environment: at step t the agent selects two actions $\boldsymbol{x}_t$, $\boldsymbol{x}'_t \in \mathcal{X}$ and only observes stochastic binary feedback $y_t \in [0,1]$ indicating if $\boldsymbol{x}_t \succ \boldsymbol{x}'_t$, that is, if action $\boldsymbol{x}_t$ is *preferred* over action $\boldsymbol{x}'_t$. More formally, $\mathbb{P}(y_t = 1|\boldsymbol{x}_t, \boldsymbol{x}'_t) = \mathbb{P}(\boldsymbol{x}_t \succ \boldsymbol{x}'_t)$, and $y_t = 0$ with probability $1 - \mathbb{P}(\boldsymbol{x}_t \succ \boldsymbol{x}'_t)$. Based on the preference history $H_t = \{(\boldsymbol{x}_1, \boldsymbol{x}'_1, y_1), \dots (\boldsymbol{x}_t, \boldsymbol{x}'_t, y_t)\}$, the agent aims to sequentially select favorable action pairs. Over a horizon of T steps, the success of the agent is measured through the *cumulative dueling regret*

$$R^{\mathrm{D}}(T) = \sum_{t=1}^{T} \frac{\mathbb{P}(\boldsymbol{x}^\star \succ \boldsymbol{x}_t) + \mathbb{P}(\boldsymbol{x}^\star \succ \boldsymbol{x}'_t) - 1}{2}, \tag{1}$$

which is the average sub-optimality gap between the chosen pair and the globally preferred action $\boldsymbol{x}^\star$. To better understand this notion of regret, consider the scenario where actions $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$ are both optimal. Then the probabilities are equal to 0.5 and the dueling regret will not grow further, since the regret incurred at step t is zero. This formulation of $R^{\mathrm{D}}(T)$ is commonly used in the literature of dueling Bandits and RL with preference feedback [Urvoy et al., 2013, Saha et al., 2023, Zhu et al., 2023] and is adapted from Yue et al. [2012]. Our goal is to design an algorithm that satisfies a *sublinear* dueling regret, where $R^{\mathrm{D}}(T)/T \to 0$ as $T \to \infty$. This implies that given enough evidence, the algorithm will converge to the globally preferred action. To this end, we take a utility-based approach and consider an unknown utility function $f : \mathcal{X} \to \mathbb{R}$, which encodes absolute preference, i.e., $\boldsymbol{x}_t \succ \boldsymbol{x}'_t$ if and only if $f(\boldsymbol{x}_t) > f(\boldsymbol{x}'_t)$. We model the dependency of the stochastic binary feedback y_t on f using the Bradley-Terry model [Bradley and Terry, 1952]

$$\mathbb{P}(y_t = 1|\boldsymbol{x}_t, \boldsymbol{x}'_t) \coloneqq s\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)\right) \tag{2}$$

where $s : \mathbb{R} \to [0,1]$ is the sigmoid function, i.e. $s(a) = (1 + e^{-a})^{-1}$. This probabilistic model for binary feedback is widely used in the literature for logistic and generalized bandits [Filippi et al., 2010, Faury et al., 2020]. Under the utility-based model, $\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ and we can draw connections to a classic bandit problem with direct feedback over a reward f. In particular, Saha [2021] shows that the dueling regret is *equivalent* up to constant factors, to the average *utility* regret of the two actions, that is $\sum_{t=1}^{T} f(\boldsymbol{x}^\star) - [f(\boldsymbol{x}_t) + f(\boldsymbol{x}'_t)]/2$.

Throughout this paper, we make two key assumptions over the environment. We assume that the domain $\mathcal{X} \subset \mathbb{R}^{d_0}$ is compact, and that the utility function lies in $\mathcal{H}_k$, a Reproducing Kernel Hilbert Space corresponding to some kernel function $k \in \mathcal{X} \times \mathcal{X} \to R$ with a bounded RKHS $\|f\|_k \leq B$. Without a loss of generality, we further suppose that the kernel function is normalized and $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ everywhere in the domain. Our set of assumptions extends the prior literature on logistic bandits and dueling bandits from linear rewards or finite action spaces, to continuous domains with non-parametric rewards.

4 Kernelized Confidence Sequences with Direct Logistic Feedback

As a warm-up, we consider a hypothetical scenario where $\boldsymbol{x}'_t = \boldsymbol{x}_{\mathrm{null}}$ for all $t \geq 1$ such that $f(\boldsymbol{x}_{\mathrm{null}}) = 0$. Therefore at every step, we suggest an action $\boldsymbol{x}_t$ and receive a noisy binary feedback y_t, which is equal to one with probability $s(f(\boldsymbol{x}_t))$. The conditional expectation of the feedback is then characterized as $\mathbb{E}(y_t|H_{t-1}) = s(f(\boldsymbol{x}_t))$. This example reduces our problem to logistic bandits which has been previously analyzed for linear rewards [Filippi et al., 2010, Faury et al., 2020]. We extend prior work to the non-parametric setting by proposing a tractable loss function for estimating the utility function, a.k.a. reward. We present novel confidence intervals that quantify the uncertainty over the logistic predictions *uniformly* over the action domain. In doing so, we propose confidence sequences for the kernelized logistic likelihood model that are of independent interest for developing sample-efficient solvers for online and active classification.

Since the feedback y_t is a Bernoulli random variable, its likelihood depends on the utility function as $s(f(\boldsymbol{x}_t))^{y_t}[1-s(f(\boldsymbol{x}_t))]^{1-y_t}$. Then given history H_t, we can estimate f by f_t, the minimizer of the regularized negative log-likelihood loss

$$\mathcal{L}_k^{\mathrm{L}}(f; H_t) \coloneqq \sum_{\tau=1}^{t} -y_\tau \log\left[s(f(\boldsymbol{x}_\tau))\right] - (1-y_\tau)\log\left[1-s(f(\boldsymbol{x}_\tau))\right] + \frac{\lambda}{2}\|f\|_k^2 \tag{3}$$

where $\lambda > 0$ is the regularization coefficient. The regularization term ensures that $\|f_t\|_k$ is finite and bounded. For simplicity, we assume throughout the main text that $\|f_t\|_k \leq B$. However, we do not need to rely on this assumption. In the appendix, we present a more rigorous analysis by projecting f_t back into the RKHS ball of radius B to ensure that the B-boundedness condition is met, instead of assuming it. We do not perform this projection in our experiments.

Solving for f_t may seem intractable at first glance since the loss is defined over functions in the large space of $\mathcal{H}_k$. However, it is common knowledge that the solution has a parametric form and may be calculated by using gradient descent. This is a simple application of the Representer Theorem [Schölkopf et al., 2001] and is detailed in Proposition 1.

Proposition 1 (Logistic Representer Theorem)**.** *The regularized negative log-likelihood loss of* $\mathcal{L}_k^{\mathrm{L}}(f; H_t)$ *has a unique minimizer* f_t*, which takes the form* $f_t(\cdot) = \sum_{\tau=1}^{t} \alpha_s k(\cdot, \boldsymbol{x}_\tau)$ *where* $(\alpha_1, \ldots \alpha_t) =: \boldsymbol{\alpha}_t \in \mathbb{R}^t$ *is the minimizer of the strictly convex loss*

$$\mathcal{L}_k^{\mathrm{L}}(\boldsymbol{\alpha}; H_t) = \sum_{\tau=1}^{t} -y_\tau \log\left[s(\boldsymbol{\alpha}^\top \boldsymbol{k}_t(\boldsymbol{x}_\tau))\right] - (1-y_\tau)\log\left[1-s(\boldsymbol{\alpha}^\top \boldsymbol{k}_t(\boldsymbol{x}_\tau))\right] + \frac{\lambda}{2}\|\alpha\|_2^2$$

with $\boldsymbol{k}_t(\boldsymbol{x}) = (k(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_t, \boldsymbol{x})) \in \mathbb{R}^t$.

Given f_t, we may predict the expected feedback for a point $\boldsymbol{x}$ as $s(f_t(\boldsymbol{x}))$. Centered around this prediction, we construct confidence sets of the form $[s(f_t(\boldsymbol{x})) \pm \beta_t(\delta)\sigma_t(\boldsymbol{x})]$, and show their uniform anytime validity. The width of the sets are characterized by $\sigma_t(\boldsymbol{x})$ defined as

$$\sigma_t^2(\boldsymbol{x}) \coloneqq k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_t^\top(\boldsymbol{x})(K_t + \lambda\kappa \boldsymbol{I}_t)^{-1}\boldsymbol{k}_t(\boldsymbol{x}) \tag{4}$$

where $\kappa = \sup_{a \leq B} 1/\dot{s}(a)$ and $K_t \in \mathbb{R}^{t \times t}$ is the kernel matrix satisfying $[K_t]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Our first main result shows that for a careful choice of $\beta_t(\delta)$, these sets contain $s(f(\boldsymbol{x}))$ simultaneously for all $\boldsymbol{x} \in \mathcal{X}$ and $t \geq 1$ with probability greater than $1-\delta$.

Theorem 2 (Kernelized Logistic Confidence Sequences)**.** *Assume* $f \in \mathcal{H}_k$ *and* $\|f\|_k \leq B$*. Consider any* $0 < \delta < 1$ *and set*

$$\beta_t(\delta) \coloneqq 4LB + 2L\sqrt{\frac{2\kappa}{\lambda}(\gamma_t + \log 1/\delta)}, \tag{5}$$

where $\gamma_t \coloneqq \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t} \frac{1}{2}\log\det(\boldsymbol{I}_t + (\lambda\kappa)^{-1}K_T)$*, and* $L \coloneqq \sup_{a \leq B} \dot{s}(a)$*. Then*

$$\mathbb{P}\left(\forall t \geq 1, \boldsymbol{x} \in \mathcal{X} :\ |s\left(f_t(\boldsymbol{x})\right) - s\left(f(\boldsymbol{x})\right)| \leq \beta_t(\delta)\sigma_t(\boldsymbol{x})\right) \geq 1-\delta.$$

Function-valued confidence sets around the kernelized ridge estimator are analyzed and used extensively to design bandit algorithms with noisy feedback on the true reward values [Valko et al., 2013, Srinivas et al., 2010, Chowdhury and Gopalan, 2017, Whitehouse et al., 2023]. However, under noisy logistic feedback, this literature falls short since the proposed confidence sets are no longer valid for the kernelized logistic estimator f_t. One could still estimate f using a kernelized ridge estimator estimator and benefit from this line of work. However, as empirically demonstrated in Figure 1a, this will not be a sample-efficient approach.

Proof Sketch. When minimizing the kernelized logistic loss, we do not have a closed-form solution for f_t, and can only formulate it using the fact that the gradient of the loss evaluated at f_t is the null operator, i.e., $\nabla\mathcal{L}(f_t; H_t) : \mathcal{H} \to \mathcal{H} = \boldsymbol{0}$. The key idea of our proof is to construct confidence intervals as $\mathcal{H}$-valued ellipsoids in the *gradient space* and show that the gradient operator evaluated at f belongs to it with high probability (c.f. Lemma 8). We then translate this back into intervals around point estimates $s(f_t(\boldsymbol{x}))$ uniformly for all points $\boldsymbol{x} \in \mathcal{X}$. The complete proof is deferred to Appendix B, and builds on the results of Whitehouse et al. [2023] and Faury et al. [2020].

Logistic Bandits. Such confidence sets are an integral tool for action selection under uncertainty, and bandit algorithms often rely on them to balance exploration against exploitation. To demonstrate how Theorem 2 may be used for bandit optimization with direct logistic feedback, we consider the kernelized Logistic GP-UCB algorithm. Presented in Algorithm 2, this algorithm extends LGP-UCB of Faury et al. [2020] from the linear to the kernelized setting, by using the confidence bound of Theorem 2 to calculate an optimistic estimate of the reward. We proceed to show that LGP-UCB attains a sublinear logistic regret, which is commonly defined as

$$R^{\mathrm{L}}(T) = \sum_{i=1}^{T} s(f(\boldsymbol{x}^\star)) - s(f(\boldsymbol{x}_t)).$$

To the best of our knowledge, the following corollary presents the first regret bound for logistic bandits in the kernelized setting and may be of independent interest.

Corollary 3. *Let $\delta \in (0,1]$ and choose the exploration coefficients $\beta_t(\delta)$ as described in Theorem 2 for all $t \geq 0$. Then* LGP-UCB *satisfies the anytime cumulative regret guarantee of*

$$\mathbb{P}\left(\forall T \geq 0 : R^{\mathrm{L}}(T) \leq C_L \beta_T(\delta)\sqrt{T\gamma_t}\right) \geq 1 - \delta.$$

where $C_L := \sqrt{8/\log(1+(\lambda\kappa)^{-1})}$.

5 Main Results: Bandits with Preference Feedback

We return to our main problem setting in which a pair of actions, $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$, are chosen and the feedback is a noisy binary indicator of $\boldsymbol{x}_t$ yielding a higher utility than $\boldsymbol{x}'_t$. While this type of feedback is more consistent in practice, it creates quite a challenging problem compared to the logistic case of Section 4. The search space for action pairs $\mathcal{X} \times \mathcal{X}$ is significantly larger than $\mathcal{X}$, and the observed preference feedback of $s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t))$ conveys only relative information between two actions rather than absolute as in the logistic feedback case. We start by presenting a solution to estimate f and obtain valid confidence sets under preference feedback. Using these confidence sets we then propose the MaxMinLCB algorithm which chooses action pairs that are not only favorable, i.e., yield high utility, but are also informative and help to improve utility confidence estimates.

5.1 Preference-based Confidence Sets

We consider the probabilistic model of y_t as stated in (2), and write the corresponding regularized negative loglikelihood loss as

$$\begin{aligned}\mathcal{L}^{\mathrm{D}}_k(f; H_t) := \sum_{\tau=1}^{t} & -y_\tau \log\left[s(f(\boldsymbol{x}_\tau) - f(\boldsymbol{x}'_\tau))\right] \\ & - (1-y_\tau)\log\left[1 - s\left(f(\boldsymbol{x}_\tau) - f(\boldsymbol{x}'_\tau)\right)\right] + \tfrac{\lambda}{2}\|f\|_k^2.\end{aligned} \tag{6}$$

Naturally, this loss may be optimized over different function classes and is commonly used for linear dueling bandits [e.g., Saha, 2021], and has been notably successful in reinforcement learning with human feedback [Christiano et al., 2017]. We proceed to show that the preference-based loss $\mathcal{L}^{\mathrm{D}}_k$ is equivalent to $\mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}$, the standard logistic loss (3) invoked with a specific kernel function k^{D}. This will allow us to cast the problem of inference with preference feedback as a kernelized logistic regression problem. To this end, we define the *dueling kernel* as

$$k^{\mathrm{D}}\left((\boldsymbol{x}_1, \boldsymbol{x}'_1), (\boldsymbol{x}_2, \boldsymbol{x}'_2)\right) := k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}'_1, \boldsymbol{x}'_2) - k(\boldsymbol{x}_1, \boldsymbol{x}'_2) - k(\boldsymbol{x}'_1, \boldsymbol{x}_2)$$

for all $(\boldsymbol{x}_1, \boldsymbol{x}'_1), (\boldsymbol{x}_2, \boldsymbol{x}'_2) \in \mathcal{X} \times \mathcal{X}$, and let $\mathcal{H}_{k^{\mathrm{D}}}$ be the RKHS corresponding to it. While the two function spaces $\mathcal{H}_{k^{\mathrm{D}}}$ and $\mathcal{H}_k$ are defined over different input domains, we can show that they are isomorphic, under simple regularity conditions.

Proposition 4. *Consider a kernel k and the sequence of its eigenfunctions $(\phi_i)_{i=1}^\infty$. Assume the eigenfunctions are zero-mean, i.e. $\int_{\boldsymbol{x}\in\mathcal{X}} \phi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$, and let $f : \mathcal{X} \to \mathbb{R}$. Then $f \in \mathcal{H}_k$, if and only if there exists $h \in \mathcal{H}_{k^{\mathrm{D}}}$ such that $h(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x}) - f(\boldsymbol{x}')$. Moreover, $\|h\|_{k^{\mathrm{D}}} = \|f\|_k$.*

The proof is left to Appendix D.1. The assumption on eigenfunctions in Proposition 4 is primarily made to simplify the equivalence class. In particular, the relative feedback function h can only capture the utility f up to a bias, i.e., if a constant bias b is added to all values of f, the corresponding h function will not change. The value of b may not be recovered by drawing queries from h, however, this will not cause issues in terms of identifying $\arg\max$ of f through querying values of h. Therefore, without loss of generality, we set $b = 0$ by assuming that eigenfunctions of k are zero-mean. This assumption automatically holds for all kernels that are translation or rotation invariant over symmetric domains, since their eigenfunctions are periodic $L_2(\mathcal{X})$ basis functions, e.g., Matérn kernels and sinusoids.

Proposition 4 allows us to re-write the preference-based loss function of (6) as a logistic-type loss

$$\mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}(h; H_t) = \sum_{\tau=1}^{t} -y_\tau \log\left[s(h(\boldsymbol{x}_\tau, \boldsymbol{x}'_\tau))\right] - (1-y_\tau)\log\left[1 - s\left(h(\boldsymbol{x}_\tau, \boldsymbol{x}'_\tau)\right)\right] + \frac{\lambda}{2}\|h\|^2_{k^{\mathrm{D}}},$$

that is equivalent to (3) up to the choice of kernel. We define the minimizer $h_t := \arg\min \mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}(h; H_t)$ and use it to construct anytime valid confidence sets for the utility f given only preference feedback.

Corollary 5 (Kernelized Preference-based Confidence Sequences)**.** *Assume $f \in \mathcal{H}_k$ and $\|f\|_k \le B$. Choose $0 < \delta < 1$ and set $\beta^{\mathrm{D}}_t(\delta)$ and σ^{D}_t as in equations (4) and (5), with k^D used as the kernel function. Then,*

$$\mathbb{P}\left(\forall t \ge 1, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}:\ |s\left(h_t(\boldsymbol{x}, \boldsymbol{x}')\right) - s\left(f(\boldsymbol{x}) - f(\boldsymbol{x}')\right)| \le \beta^{\mathrm{D}}_t(\delta)\sigma^{D}_t(\boldsymbol{x}, \boldsymbol{x}')\right) \ge 1 - \delta.$$

where $h_t = \arg\min \mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}(h; H_t)$.

Corollary 5 gives valid confidence sets for kernelized utility functions under preference feedback and may be of independent interest. This confidence bound immediately improves prior results on linear dueling bandits and kernelized dueling bandits with regression-type loss, to kernelized setting with logistic-type likelihood. To demonstrate this, in Appendix D.3 we present the kernelized extensions of MaxInP (Saha [2021], Algorithm 3), and IDS (Kirschner and Krause [2021], Algorithm 4) and prove the corresponding regret guarantees (cf. Theorems 15 and 16). This corollary holds almost immediately by invoking Theorem 2 with the dueling kernel k^{D} and applying Proposition 4. A proof is provided in Appendix D.1 for completeness.

Comparison to Prior Work. A line of previous work assumes that both f and the probability $s(f(\boldsymbol{x}))$ are B-bounded members of $\mathcal{H}_k$. This allows them to directly estimate $s(f(\boldsymbol{x}))$ via kernelized linear regression [Xu et al., 2020, Mehta et al., 2023b, Kirschner and Krause, 2021]. The resulting confidence intervals are then around the minimum least squares estimator, which does not align with the logistic estimator f_t. This model does not encode the fact that $s(f(\boldsymbol{x}))$ only takes values in $[0, 1]$ and considers a sub-gaussian distribution for y_t, instead of the Bernoulli formulation. Therefore, the resulting algorithms require more samples to learn an accurate reward estimate. In a concurrent work, Xu et al. [2024] address the preference-based loss function of Equation (6) and present anytime valid likelihood-ratio confidence sets for the minimizer of this loss. The width of such sets at time T, scale with $\sqrt{T \log \mathcal{N}(\mathcal{H}_k; 1/T)}$ where the second term is the metric entropy of the B-bounded RKHS at resolution $1/T$, that is, the log-covering number of this function class, using balls of radius $1/T$. It is known that $\log \mathcal{N}(\mathcal{H}_k; 1/T) \asymp \gamma_T$ as defined in Theorem 2. This may be easily verified using Wainwright [2019, Example 5.12] and [Vakili et al., 2021, Definition 1]. Noting the definition of β^{D}_t, we see that likelihood ratio sets of Xu et al. [2024] are wider than Corollary 5. Consequently, the presented regret guarantee in this work is looser by a factor of $T^{1/4}$ compared to our bound in Theorem 6.

5.2 Action Selection Strategy

We propose MaxMinLCB in Algorithm 1 for the preference feedback bandit problem that selects $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$ jointly in each time step t as

$$\begin{aligned} \boldsymbol{x}_t &= \arg\max_{\boldsymbol{x} \in \mathcal{M}_t} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}'(\boldsymbol{x})) \quad &\text{(Leader)} \\ \text{s.t. } \boldsymbol{x}'(\boldsymbol{x}) &= \arg\min_{\boldsymbol{x}' \in \mathcal{M}_t} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') \quad &\text{(Follower)} \end{aligned} \tag{7}$$

where the lower-confidence bound $\mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') = s(h_t(\boldsymbol{x}, \boldsymbol{x}')) - \beta^{\mathrm{D}}_t \sigma^{\mathrm{D}}_t(\boldsymbol{x}, \boldsymbol{x}')$ presents a pessimistic estimate of h and $\mathcal{M}_t = \{\boldsymbol{x} \in \mathcal{X} | \forall \boldsymbol{x}' \in \mathcal{X}:\ s(h_t(\boldsymbol{x}, \boldsymbol{x}')) + \beta^{\mathrm{D}}_t \sigma^{\mathrm{D}}_t(\boldsymbol{x}, \boldsymbol{x}') \ge 0.5\}$ is the set of potentially optimal actions. The second action is chosen as $\boldsymbol{x}'_t = \boldsymbol{x}'(\boldsymbol{x}_t)$. Equation (7) forms a zero-sum Stackelberg (Leader–Follower) game where the actions $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$ are chosen

Algorithm 1 MAXMINLCB

Input $(\beta_t^{\mathrm{D}})_{t\geq 1}$.
for $t \geq 1$ **do**
Play the most potent pair $(\boldsymbol{x}_t, \boldsymbol{x}'_t)$ according to the Stackelberg game

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{M}_t} s(h_t(\boldsymbol{x}, \boldsymbol{x}'(\boldsymbol{x}))) - \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}'(\boldsymbol{x}))$$
$$\text{s.t. } \boldsymbol{x}'(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}'\in\mathcal{M}_t} s(h_t(\boldsymbol{x}, \boldsymbol{x}')) - \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}')$$
$$\text{and } \boldsymbol{x}'_t = \boldsymbol{x}'(\boldsymbol{x}_t).$$

Observe y_t and append history.
Update h_{t+1} and $\sigma_{t+1}^{\mathrm{D}}$ and the set of plausible maximizers

$$\mathcal{M}_{t+1} = \{\boldsymbol{x} \in \mathcal{X} \,|\, \forall \boldsymbol{x}' \in \mathcal{X}:\, s(h_{t+1}(\boldsymbol{x}, \boldsymbol{x}')) + \beta_{t+1}^{\mathrm{D}}\sigma_{t+1}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}') \geq 0.5\}.$$

end for

sequentially [Stackelberg et al., 1952]. First, the Leader selects $\boldsymbol{x}_t$, then the Follower selects $\boldsymbol{x}'_t$ depending on the choice of $\boldsymbol{x}_t$. Importantly, the sequential nature of action selections is known and $\boldsymbol{x}_t$ is chosen by the Leader such that the Follower's action selection function, $\boldsymbol{x}'(\cdot)$, is accounted for in the selection of $\boldsymbol{x}_t$. Sequential optimization problems are known to be computationally NP-hard even for linear functions [Jeroslow, 1985]. However, due to their importance in practical applications, there are algorithms that can efficiently approximate a solution over large domains [Sinha et al., 2017, Ghadimi and Wang, 2018, Dagréou et al., 2022, Camacho-Vallejo et al., 2023].

MAXMINLCB builds on a simple insight: if the utility f is known, both the Leader and the Follower will choose $\boldsymbol{x}^\star$ yielding an objective value 0.5 for both players, and zero dueling regret. Since MAXMINLCB has no access to f, it leverages the confidence sets of Corollary 5 and uses a pessimistic approach by considering the LCB instead. There are two crucial properties of the Follower specific to this game. First, the Follower can not do worse than the Leader with respect to the LCB_t. In any scenario, the Follower can match the Leader's action which results in $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) = 0.5$. Second, for sufficiently tight confidence sets, the Follower will not select sub-optimal actions. In this case, the Leader's best action must be optimal as it anticipates the Follower's response and Equation (7) recovers the optimal actions. Therefore, the objective value of the game considered in Equation (7) is always less than, or equal to the objective of the game with known utility function f, i.e., $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 0.5 = f(\boldsymbol{x}^\star, \boldsymbol{x}^\star)$ and the gap shrinks with the confidence sets. Overall, the Stackelberg game in Equation (7) can be considered as a lower approximation of the game played with known utility function f.

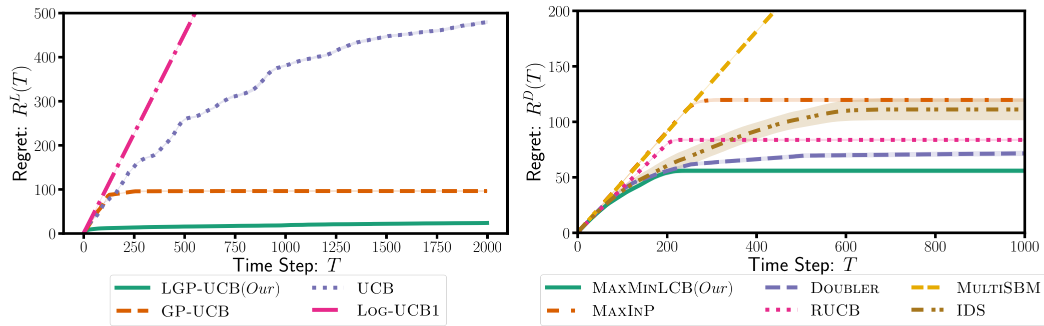
The primary challenge for MAXMINLCB is to sample action pairs that sufficiently shrink the confidence sets for the optimal actions without accumulating too much regret. MAXMINLCB balances this exploration-exploitation trade-off naturally with its game theoretic formulation. We view the selection of $\boldsymbol{x}_t$ to be exploitative by trying to maximize the unknown utility $f(\boldsymbol{x}_t)$ and minimizing regret. On the other hand, $\boldsymbol{x}'_t$ is chosen to be the most competitive opponent to $\boldsymbol{x}_t$, i.e., testing whether the condition $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \geq 0.5$ holds. Note that LCB_t is pessimistic concerning $\boldsymbol{x}_t$ making it robust against the uncertainty in the confidence set estimation. At the same time, LCB_t is an optimistic estimate for $\boldsymbol{x}'_t$ encouraging exploration. In our main theoretical result, we prove that under the assumptions of Corollary 5, MAXMINLCB achieves sublinear regret on the dueling bandit problem.

Theorem 6. *Suppose the utility function f lies in $\mathcal{H}_k$ with a norm bounded by B, and that kernel k satisfies the assumption of Proposition 4. Let $\delta \in (0, 1]$ and choose the exploration coefficient $\beta_t^{\mathrm{D}}(\delta)$ as in Corollary 5. Then* MAXMINLCB *satisfies the anytime dueling regret of*

$$\mathbb{P}\left(\forall T \geq 0: R^{\mathrm{D}}(T) \leq C_3\beta_T^{\mathrm{D}}(\delta)\sqrt{T\gamma_T^{\mathrm{D}}} = \mathcal{O}(\gamma_T^{\mathrm{D}}\sqrt{T})\right) \geq 1 - \delta$$

where γ_T^{D} is the T-step information gain of kernel k^{D} and $C_3 = (8 + 2\kappa)/\sqrt{\log(1 + 4(\lambda\kappa)^{-1})}$.

The proof is left to Appendix D.2. The information gain γ_T^{D} in Theorem 6 quantifies the structural complexity of the RKHS corresponding to k^{D} and its dependence on T is fairly understood for kernels commonly used in applications of bandit optimization. As an example, for a Matérn kernel



(a) Logistic feedback

(b) Preference feedback

Figure 1: Regret of learning the Ackley function with logistic and preference feedback. **(a)** Same UCB algorithms, each using a different confidence set. LGP-UCB performs best, showcasing the power of Theorem 2. **(b)**: Algorithms with different acquisition functions, all using our confidence sets. MAXMINLCB is more sample-efficient.

of smoothness ν defined over a d-dimensional domain, $\gamma_T = \tilde{\mathcal{O}}(T^{d/(2\nu+d)})$ [Remark 2 Vakili et al., 2021] and the corresponding regret bound grows sublinearly with T.

Restricting the optimization domain to $\mathcal{M}_t \subset \mathcal{X}$ is common in the literature [Zoghi et al., 2014a, Saha, 2021] despite being challenging in applications with large or continuous domains. We conjecture that MAXMINLCB would enjoy similar regret guarantees without restricting the selection domain to $\mathcal{M}_t$ as done in Equation (7). This claim is supported by our experiments in Section 6.2 which are carried out without such restriction on the optimization domain.

6 Experiments

Our experiments are on finding the maxima of test functions commonly used in (non-convex) optimization literature [Jamil and Yang, 2013] given preference feedback. These functions cover challenging optimization landscapes including several local optima, plateaus, and valleys, allowing us to test the versatility of MAXMINLCB. We use the Ackley function for illustration in the main text and provide the regret plots for the remainder of the functions in Appendix E.2. For all experiments, we set the horizon $T = 2000$ and evaluate all algorithms on a uniform mesh over the input domain of size 100. All experiments are run across 20 random seeds and reported values are averaged over the seeds, together with standard error. Details of implementation [1] are deferred to Appendix E.1.

6.1 Benchmarking Confidence Sets

Performance of MAXMINLCB relies on validity and tightness of the LCB. We evaluate the quality of our kernelized confidence sets, using the potentially simpler task of bandit optimization given logistic feedback. We fix the acquisition function via the celebrated principle of *optimism-in-the-face-of-uncertainty* (OfU), and choose the action that maximizes the upper confidence bound (UCB). This comparison highlights the separate benefits of LGP-UCB. We refer to the UCB algorithm instantiated with the confidence sets of Theorem 2 as LGP-UCB, and consider three baselines. UCB assumes that actions are uncorrelated, and maintains an independent confidence interval for each action as in Lattimore and Szepesvári [2020, Algorithm 3]. This demonstrates how LGP-UCB utilizes the correlation between actions. We also implement LOG-UCB1 [Faury et al., 2020] that assumes that f is a linear function, i.e., $f(\boldsymbol{x}) = \theta^T \boldsymbol{x}$ to highlight the improvements gained by kernelization. Last, we compare LGP-UCB with GP-UCB [Srinivas et al., 2010] that estimates probabilities $s(f(\cdot))$ via a kernelized ridge regression task. This comparison highlights the benefits of using our kernelized logistic estimator (Proposition 1) over regression-based approaches [Xu et al., 2020, Kirschner and Krause, 2021, Mehta et al., 2023b,a]. Figure 1a shows that the cumulative regret of LGP-UCB is the lowest among the selected algorithms. GP-UCB performs closest to LGP-UCB, however, it accumulates regret linearly during the initial steps. Note that GP-UCB and LGP-UCB differ in the estimation of the utility function f_t while estimating the width of the confidence bounds similarly. This result suggests that using the logistic-type loss (3) to infer the utility function is advantageous. As expected, UCB converges at a slower rate than either LGP-UCB or GP-UCB due to omitting the correlation between arms while

[1] We implemented the environments and algorithms end-to-end in JAX [Bradbury et al., 2018].

Table 1: Benchmarking R_T^{D} for a variety of test utility functions, $T = 2000$.

f	MAXMINLCB	DOUBLER	MULTISBM	MAXINP	RUCB	IDS
Ackley	$\mathbf{54} \pm 3$	67 ± 3	453 ± 58	112 ± 5	79 ± 3	99 ± 10
Branin	$\mathbf{63} \pm 10$	79 ± 8	213 ± 28	197 ± 23	$\mathbf{63} \pm 11$	86 ± 17
Eggholder	$\mathbf{100} \pm 7$	132 ± 8	435 ± 56	179 ± 21	155 ± 24	123 ± 13
Hoelder	$\mathbf{107} \pm 16$	132 ± 8	460 ± 59	169 ± 15	153 ± 18	119 ± 15
Matyas	81 ± 8	87 ± 8	209 ± 27	100 ± 8	79 ± 7	$\mathbf{58} \pm 8$
Michalewicz	$\mathbf{108} \pm 10$	149 ± 11	473 ± 61	196 ± 25	184 ± 28	154 ± 19
Rosenbrock	$\mathbf{18} \pm 3$	24 ± 8	131 ± 17	76 ± 6	38 ± 6	34 ± 9

LOG-UCB1's regret grows linearly as the Ackley function violates the assumption of linearity. We defer the results on the rest of the utility functions to Table 2 in Appendix E.2 and the figures therein.

6.2 Benchmarking Acquisition Functions

In this section, we compare MAXMINLCB with other utility-based bandit algorithms. To isolate the benefits of our acquisition function, we instantiate other algorithms using our confidence sets Corollary 5. Our implementation then differs from the corresponding references, while we refer to them by their original name. We consider the following baselines. DOUBLER and MULTISBM [Ailon et al., 2014] who choose $\boldsymbol{x}_t$ as a *reference* action from the recent history of actions and pair it with $\boldsymbol{x}'_t$ which maximizes the joint UCB (cf. Algorithm 5 and 6). RUCB [Zoghi et al., 2014a] similarly relies on OfU, however, it selects the reference action uniformly at random from $\mathcal{M}_t$ (Algorithm 7). MAXINP [Saha, 2021] also maintains the set of plausible maximizers $\mathcal{M}_t$, however, in each time step, it selects the pair of actions that maximize $\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}')$ (Algorithm 3). IDS [Kirschner and Krause, 2021] selects the reference action greedily by maximizing f_t, and pairs it with an informative action (Algorithm 4). Notably, all algorithms, with the except of MAXINP, choose one of the actions independently and use it as a reference point when selecting the other one. Figure 2 in Appendix A illustrates the differences in action selection between the OfU, maximum information, and MAXMINLCB approaches

Figure 1b benchmarks the algorithms using the Ackley utility function, where MAXMINLCB outperforms the baselines. All algorithms suffer from close-to-linear regret during the first phase of the learning suggesting that there is an inevitable exploration phase. Notably, MAXMINLCB, IDS, and DOUBLER are the first to select actions with high utility, while RUCB and MAXINP explore for longer. Table 1 shows the dueling regret for all utility functions. MAXMINLCB performs consistently among the best two algorithms across the analyzed functions and achieves a low standard error supporting its efficiency in balancing exploration and exploitation in the preference feedback setting. While MAXMINLCB consistently outperforms the baselines, we do not observe a clear ranking among the rest. For instance, IDS achieves the smallest regret for optimizing Matyas, while RUCB excels on the Branin function. This indicates the challenges each function offers and the performance of the action selection is task dependent. The consistent performance of MAXMINLCB demonstrates its robustness against the underlying unknown utility function.

7 Conclusion

We addressed the problem of bandit optimization with preference feedback over large domains and complex targets. We propose MAXMINLCB, which takes a game-theoretic approach to the problem of action selection under comparative feedback, and naturally balances exploration and exploitation by constructing a zero-sum Stackelberg game between the action pairs. MAXMINLCB achieves a sublinear regret for kernelized utilities, and performs competitively across a range of experiments. Lastly, by uncovering the equivalence of learning with logistic or comparative feedback, we propose kernelized preference-based confidence sets, which may employed in adjacent problems, such as reinforcement learning with human feedback. The technical setup considered in this work serves as a foundation for a number of applications in mechanism design, such as preference elicitation and welfare optimization from multiple feedback sources for social choice theory, which we leave as future work.

References

Yasin Abbasi-Yadkori. *Online learning for linearly parametrized control problems*. PhD thesis, University of Alberta, 2013.

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

Fabio Aiolli and Alessandro Sperduti. Learning preferences for multiclass problems. *Advances in neural information processing systems*, 17, 2004.

Sheldon Axler. *Measure, integration & real analysis*. Springer Nature, 2020.

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

Alina Beygelzimer, David Pal, Balazs Szorenyi, Devanathan Thiruvenkatachari, Chen-Yu Wei, and Chicheng Zhang. Bandit multiclass linear classification: Efficient algorithms for the separable case. In *International Conference on Machine Learning*, pages 624–633. PMLR, 2019.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint*, 2010.

José-Fernando Camacho-Vallejo, Carlos Corpus, and Juan G Villegas. Metaheuristics for bilevel optimization: A comprehensive review. *Computers & Operations Research*, page 106410, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Bangrui Chen and Peter I Frazier. Dueling bandits with weak regret. In *International Conference on Machine Learning*, pages 731–739. PMLR, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

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

Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

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

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

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

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

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

Dylan J Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pages 1282–1291. PMLR, 2017.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2): 150–194, 2013.

Robert G Jeroslow. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical programming*, 32(2):146–164, 1985.

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

Johannes Kirschner and Andreas Krause. Bias-robust bayesian optimization via dueling bandits. In *International Conference on Machine Learning*. PMLR, 2021.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.

Wataru Kumagai. Regret analysis for continuous dueling bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Peter D Lax. *Functional analysis*, volume 55. John Wiley & Sons, 2002.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4482. PMLR, 2024.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint*, 2023a.

Viraj Mehta, Ojash Neopane, Vikramjeet Das, Sen Lin, Jeff Schneider, and Willie Neiswanger. Kernelized offline contextual dueling bandits. *arXiv preprint*, 2023b.

Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, and Samuel Kaski. Projective preferential bayesian optimization. In *International Conference on Machine Learning*. PMLR, 2020.

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

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

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE transactions on evolutionary computation*, 22(2): 276–295, 2017.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010.

Heinrich von Stackelberg et al. Theory of the market economy. 1952.

Yanan Sui, Vincent Zhuang, Joel W Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

Shion Takeno, Masahiro Nomura, and Masayuki Karasuyama. Towards practical preferential bayesian optimization with skew gaussian processes. In *International Conference on Machine Learning*, pages 33516–33533. PMLR, 2023.

Maegan Tucker, Ellen Novoseller, Claudia Kann, Yanan Sui, Yisong Yue, Joel W Burdick, and Aaron D Ames. Preference-based learning for exoskeleton gait optimization. In *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020.

Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In *International Conference on Machine Learning*. PMLR, 2013.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Improved self-normalized concentration in hilbert spaces: Sublinear regret for gp-ucb. *arXiv preprint arXiv:2307.07539*, 2023.

Wenjie Xu, Wenbin Wang, Yuning Jiang, Bratislav Svetozarevic, and Colin N Jones. Principled preferential bayesian optimization. *arXiv preprint arXiv:2402.05367*, 2024.

Yichong Xu, Aparna Joshi, Aarti Singh, and Artur Dubrawski. Zeroth order non-convex optimization with dueling-choice bandits. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 2012.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. *arXiv preprint*, 2023.

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

Julian Zimmert and Yevgeny Seldin. Factored bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*. PMLR, 2014a.

Masrour Zoghi, Shimon A Whiteson, Maarten De Rijke, and Remi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 2014b.

Contents of Appendix

A Illustration of main concepts

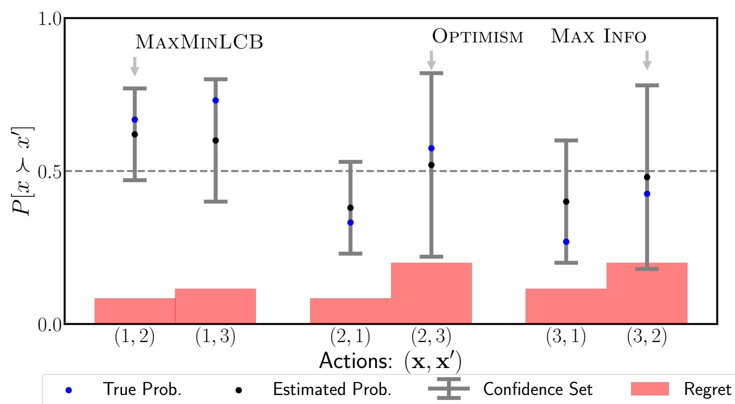


Figure 2: Confidence sets for an illustrative problem with 3 arms at a single time step. Annotated arrows highlight the action selection for three common approaches. MaxMinLCB selects the action pair $(1, 2)$ with the least regret. Upper-bound maximization (Optimism) and information maximization (Max Info) choose sub-optimal arms due to the large width of the sets that have higher regrets.

B Proofs for Bandtis with Logistic Feedback

While we have written the algorithm in terms of the kernel matrix and function evaluations, for the purpose of the proof, we mainly rely on entities in the Hilbert space. Consider the operator $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$ which corresponds to kernel k and satisfies $k(\boldsymbol{x}, \cdot) = \boldsymbol{\phi}(\boldsymbol{x})$. Then by Mercer's theorem, any $f \in \mathcal{H}_k$ may be written as $f = \boldsymbol{\theta}^\top \boldsymbol{\phi}$, where $\boldsymbol{\theta} \in \ell_2(\mathbb{N})$ and has a Bbounded ℓ_2 norm. For a sequence of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_t \in \mathcal{X}$, we define the infinite dimensional feature map $\Phi_t = [\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_t)]^\top$, which gives rise to the kernel matrix $K_t : \mathbb{R}^t \to \mathbb{R}^t$ and the covariance operator $S_t : \mathcal{H} \to \mathcal{H}$, respectively defined as $K_t = \Phi_t \Phi_t^\top$ and $S_t = \Phi_t^\top \Phi_t$. Let $\boldsymbol{I}_t$ denote the t-dimensional identity matrix, and $\boldsymbol{I}_\mathcal{H}$ be the identity operator on the RKHS. Then it is widely known that the covariance and kernel operators are connected via $\det(\boldsymbol{I}_\mathcal{H} + \rho^{-2} S_t) = \det(\boldsymbol{I}_t + \rho^{-2} K_t)$ for any $t \geq 1$ and $\rho \neq 0$. For operators on the Hilbert space, $\det(A)$ refer to a Fredholm determinant [c.f. Lax, 2002].

To analyze our function-valued confidence sequences, we start by re-writing the logistic loss function

$$\mathcal{L}(\boldsymbol{\theta}; H_t) = \sum_{s=1}^{t} -y_s \log s\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_s)\right) - \sum_{s=1}^{t} (1 - y_s) \log\left(1 - s\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_s)\right)\right) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

which is strictly convex in $\boldsymbol{\theta}$ and has a unique minimizer $\boldsymbol{\theta}_t$ which satisfies

$$\nabla \mathcal{L}(\boldsymbol{\theta}_t; H_t) = \sum_{s=1}^{t} -y_s \boldsymbol{\phi}(\boldsymbol{x}_s) + g_t(\boldsymbol{\theta}_t) = 0$$

where $g_t(\boldsymbol{\theta}) : \mathcal{H} \to \mathcal{H}$ is a linear operator defined as

$$g_t(\boldsymbol{\theta}) \coloneqq \sum_{s=1}^{t} \boldsymbol{\phi}(\boldsymbol{x}_s) s(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_s)) + \lambda \boldsymbol{\theta}.$$

In the main text, we assumed that minimizer of $\mathcal{L}$ satisfies the norm boundedness condition. Here, we present a more rigorous analysis which does not assume so. Instead, we work with a projected estimator defined via

$$\boldsymbol{\theta}_t^P = \underset{\|\boldsymbol{\theta}\|_2 \leq B}{\arg\min} \|\boldsymbol{g}_t(\boldsymbol{\theta}) - \boldsymbol{g}_t(\boldsymbol{\theta}_t)\|_{V_t^{-1}}. \tag{8}$$

where $V_t = S_t + \kappa\lambda \boldsymbol{I}_{\mathcal{H}}$ and $\boldsymbol{\theta}_t$ is the minimizer of $\mathcal{L}(\boldsymbol{\theta}; H_t)$. Roughly put, $\boldsymbol{\theta}_t^P \in \ell_2(\mathbb{N})$ is a norm bounded alternative to $\boldsymbol{\theta}_t$, which satisfies a small $\nabla \mathcal{L}$, and therefore, is expected to result in an accurate decision boundary. We will present our proof in terms of $\boldsymbol{\theta}_t^P$. This also proves the results in the main text, since $\boldsymbol{\theta}_t^P = \boldsymbol{\theta}_t$ if $\boldsymbol{\theta}_t$ itself happens to have a B-bounded norm, as assumed in the main text.

Our analysis relies on a concentration bound for $\mathcal{H}$-valued martingale sequences stated in Abbasi-Yadkori [2013] and later in Whitehouse et al. [2023]. Below, we have adapted the statement to match our notation.

Lemma 7 (Corollary 1 Whitehouse et al. [2023])**.** *Suppose the sequence $(\boldsymbol{x}_t)_{t\geq 1}$ is $(\mathcal{F}_t)_{t\geq 1}$-predictable, where $\mathcal{F}_t \coloneqq \sigma(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t, \varepsilon_1, \dots, \varepsilon_{t-1})$ and ε_t is i.i.d. zero-mean σ-subGaussian noise. Consider the RKHS $\mathcal{H}$ corresponding to a kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}')$. Then, for any $\rho > 0$ and $\delta \in (0,1)$, we have that, with probability at least $1-\delta$, simultaneously for all $t \geq 0$,*

$$\left\| \sum_{s\leq t} \varepsilon_s \boldsymbol{\phi}(\boldsymbol{x}_s) \right\|_{V_t^{-1}} \leq \sigma \sqrt{2 \log\left(\frac{1}{\delta} \sqrt{\det(\boldsymbol{I}_t + \rho^{-2} K_t)} \right)}$$

where $V_t = S_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}$.

The following lemma, which extends Faury et al. [2020][Lemma 8] to $\mathcal{H}$-valued operators, expresses the closeness of $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^\star$ in the gradient space, with respect to the norm of the covariance matrix.

Lemma 8 (Gradient Space Confidence Bounds)**.** *Set $0 < \delta < 1$. Then,*

$$\mathbb{P}\left(\forall t \geq 0 : \; \|\boldsymbol{g}_t(\boldsymbol{\theta}_t) - \boldsymbol{g}_t(\boldsymbol{\theta}^\star)\|_{V_t^{-1}} \leq \frac{1}{2}\sqrt{2\log 1/\delta + 2\gamma_T} + \sqrt{\frac{\lambda}{\kappa}} B \right) \geq 1 - \delta$$

where $V_t = S_t + \kappa\lambda \boldsymbol{I}_{\mathcal{H}}$.

Proof of Lemma 8. Recall that $\boldsymbol{g}_t(\boldsymbol{\theta}) \coloneqq \sum_{s\leq t} s(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_s))\boldsymbol{\phi}(\boldsymbol{x}_s) + \lambda\boldsymbol{\theta}$. Then it is straighforward to show that

$$\nabla \mathcal{L}(\boldsymbol{\theta}; H_t) = \sum_{s\leq t} y_s \boldsymbol{\phi}(\boldsymbol{x}_s) - g_t(\boldsymbol{\theta}).$$

Then since $\boldsymbol{\theta}_t$ is a minimizer of $\mathcal{L}_t$, it holds that $\boldsymbol{g}_t(\boldsymbol{\theta}_t) = \sum_{s\leq t} y_s \boldsymbol{\phi}(\boldsymbol{x}_s)$. This allows us to write,

$$\begin{aligned}
\|\boldsymbol{g}_t(\boldsymbol{\theta}_t) - \boldsymbol{g}_t(\boldsymbol{\theta}^\star)\|_{V_t^{-1}} &= \left\| \sum_{s\leq t} \left(y_s - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_s)\boldsymbol{\theta}^\star) \right) \boldsymbol{\phi}(\boldsymbol{x}_s) - \lambda\boldsymbol{\theta}^\star \right\|_{V_t^{-1}} \\
&\leq \left\| \sum_{s\leq t} \varepsilon_s \boldsymbol{\phi}(\boldsymbol{x}_s) \right\|_{V_t^{-1}} + \lambda \|\boldsymbol{\theta}^\star\|_{V_t^{-1}}
\end{aligned} \tag{9}$$

where $\varepsilon_s := y_s - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_s)\boldsymbol{\theta}^\star)$ is a history dependent random variable in $[0,1]$, due to our data model. To bound the first term, we recognize that any random variable in $[0,1]$ is σ-subGaussian with $\sigma = 0.5$ and apply Lemma 7 . We obtain that for all $t \geq 0$, with probability greater than $1-\delta$

$$\begin{aligned}\left\|\sum_{s\leq t} \varepsilon_s \boldsymbol{\phi}(\boldsymbol{x}_s)\right\|_{V_t^{-1}} &\leq \frac{1}{2}\sqrt{2\log\left(\frac{1}{\delta}\sqrt{\det(\boldsymbol{I}_t + (\lambda\kappa)^{-1}K_t)}\right)}\\ &\leq \frac{1}{2}\sqrt{2\log 1/\delta + 2\gamma_T}\end{aligned}$$

since $\gamma_t(\rho) = \sup_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_t} \frac{1}{2}\log\det(\boldsymbol{I}_t + \rho^{-2}K_t))$. To bound the second term in (9), note that $S_t = \Phi_t^\top\Phi_t$ is PSD and therefore $V_t \geq \kappa\lambda \boldsymbol{I}_{\mathcal{H}}$. Then

$$\lambda\|\boldsymbol{\theta}^\star\|_{V_t^{-1}} \leq \frac{\lambda}{\sqrt{\lambda\kappa}}\|\boldsymbol{\theta}^\star\|_2 \leq \sqrt{\frac{\lambda}{\kappa}}B.$$

concluding the proof. □

The following lemma shows the validity of our parameter-space confidence sets.

Lemma 9. *Set $0 < \delta < 1$ and consider the confidence sets*

$$\Theta_t(\delta) := \left\{\|\boldsymbol{\theta}\| \leq B, \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_t^P\right\|_{V_t} \leq 2\sqrt{\lambda\kappa}B + \kappa\sqrt{2\log 1/\delta + 2\gamma_T}\right\}.$$

Then,

$$\mathbb{P}\left(\forall t \geq 0:\ \boldsymbol{\theta}^\star \in \Theta_t(\delta)\right) \geq 1-\delta$$

Proof of Lemma 9. From construction of $\mathcal{E}_t(\delta)$ we have,

$$\begin{aligned}\left\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_t^P\right\|_{V_t} &\leq \kappa\left\|g_t(\boldsymbol{\theta}^\star) - g_t(\boldsymbol{\theta}_t^P)\right\|_{V_t^{-1}} && \text{(Lem. 12)}\\ &\leq \kappa\left(\|g_t(\boldsymbol{\theta}^\star) - g_t(\boldsymbol{\theta}_t)\|_{V_t^{-1}} + \left\|g_t(\boldsymbol{\theta}_t) - g_t(\boldsymbol{\theta}_t^P)\right\|_{V_t^{-1}}\right)\\ &\leq 2\kappa\|g_t(\boldsymbol{\theta}^\star) - g_t(\boldsymbol{\theta}_t)\|_{V_t^{-1}} && \text{Eq (8)}\\ &\leq \kappa\sqrt{2\log 1/\delta + 2\gamma_T} + 2\sqrt{\lambda\kappa}B && \text{(Lem. 8)}\end{aligned}$$

□

Lastly, we prove an extension of Theorem 2.

Theorem 10 (Theorem 2 - Formal)**.** *Set $0 < \delta < 1$ and consider the confidence sets $\mathcal{E}_t(\delta) \subset \mathcal{H}$ where*

$$\mathcal{E}_t(\delta) = \left\{f(\cdot) = \boldsymbol{\theta}^\top\boldsymbol{\phi}(\cdot):\ \boldsymbol{\theta} \in \Theta_t(\delta)\right\}.$$

Then, simultanously for all $\boldsymbol{x} \in \mathcal{X}$, $f \in \mathcal{E}_t(\delta)$ and $t \geq 0$

$$|s(f(\boldsymbol{x})) - s(f^\star(\boldsymbol{x}))| \leq \beta_t(\delta)\sigma_t(\boldsymbol{x})$$

with probability greater than $1-\delta$, where

$$\beta_t(\delta) := 4LB + 2L\sqrt{\frac{\kappa}{\lambda}}\sqrt{2\log 1/\delta + 2\gamma_T}$$

Proof of Theorem 10. For simplicity in notation let us define

$$\tilde{\beta}_t(\delta) := 2\sqrt{\lambda\kappa}B + \kappa\sqrt{2\log 1/\delta + 2\gamma_t}.$$

Algorithm 2 LGP-UCB

Initialize Set $(\beta_t)_{t\geq 1}$ according to Theorem 2.
for $t \geq 1$ **do**
Choose an optimistic action via

$$x_t = \arg\max_{x\in\mathcal{X}} s(f_{t-1}(\boldsymbol{x})) + \beta_{t-1}(\delta)\sigma_{t-1}(\boldsymbol{x})$$

Observe y_t and append history.
Calculate f_t acc. to Proposition 1 and update σ_t acc. to Theorem 2.
end for

Suppose $f = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\cdot)$ is in $\mathcal{E}_t(\delta)$. Then

$$\begin{aligned}
\left|s(\boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta}^\star) - s(\boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta})\right| &= \left|\alpha(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\theta}^\star)\boldsymbol{\phi}^\top(\boldsymbol{x})[\boldsymbol{\theta}-\boldsymbol{\theta}^\star]\right| && \text{Lem. 11}\\
&\leq L\left|\boldsymbol{\phi}^\top(\boldsymbol{x})[\boldsymbol{\theta}-\boldsymbol{\theta}^\star]\right| && s \text{ is } L\text{-Lipschitz}\\
&\leq L\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_{V_t}\\
&\leq L\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}}\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_t^P\right\|_{V_t} + \left\|\boldsymbol{\theta}_t^P-\boldsymbol{\theta}^\star\right\|_{V_t}\right)\\
&\leq L\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}}\left(\tilde{\beta}_t(\delta) + \left\|\boldsymbol{\theta}_t^P-\boldsymbol{\theta}^\star\right\|_{V_t}\right) && \boldsymbol{\theta}\in\Theta_t(\delta)\\
&\overset{\text{w.h.p.}}{\leq} 2L\tilde{\beta}_t(\delta)\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}} && \text{Lem. 9}\\
&\leq \frac{2L\tilde{\beta}_t(\delta)}{\sqrt{\lambda\kappa}}\sigma_t(\boldsymbol{x}) && \text{Lem. 13}\\
&= \sigma_t(\boldsymbol{x})\left(4LB + 2L\sqrt{\frac{\kappa}{\lambda}}\sqrt{2\log 1/\delta + 2\gamma_T}\right)
\end{aligned}$$

where the third to last inequality holds with probability greater than $1-\delta$, but the rest of the inequalities hold deterministically. □

Given the confidence set of Theorem 2, we give extend the LGP-UCB algorithm of Faury et al. to the kernelized setting (c.f. Algorithm 2) and prove that it satisfies sublinear regret.

Proof of Corollary 3. Recall that if $\boldsymbol{x}_t$ is the maximizer of the UCB, then

$$s(\boldsymbol{\phi}^\top(\boldsymbol{x}^\star)\boldsymbol{\theta}_t^P) - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\theta}_t^P) \leq \sigma_t(\boldsymbol{x}_t)\beta_t(\delta) - \sigma_t(\boldsymbol{x}^\star)\beta_t(\delta)$$

Then using Theorem 10, we obtain the following for the regret at step t,

$$\begin{aligned}
r_t &= s(\boldsymbol{\phi}^\top(\boldsymbol{x}^\star)\boldsymbol{\theta}^\star) - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\theta}^\star)\\
&= s(\boldsymbol{\phi}^\top(\boldsymbol{x}^\star)\boldsymbol{\theta}^\star) - s(\boldsymbol{\phi}^\top(\boldsymbol{x}^\star)\boldsymbol{\theta}_t^P) + s(\boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\theta}_t^P) - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\theta}^\star)\\
&\qquad + s(\boldsymbol{\phi}^\top(\boldsymbol{x}^\star)\boldsymbol{\theta}_t^P) - s(\boldsymbol{\phi}^\top(\boldsymbol{x}_t)\boldsymbol{\theta}_t^P)\\
&\leq \sigma_t(\boldsymbol{x}^\star)\beta_t(\delta) + \sigma_t(\boldsymbol{x}_t)\beta_t(\delta) + \sigma_t(\boldsymbol{x}_t)\beta_t(\delta) - \sigma_t(\boldsymbol{x}^\star)\beta_t(\delta)\\
&\leq 2\beta_t(\delta)\sigma_t(\boldsymbol{x}_t)
\end{aligned}$$

with probability greater than $1-\delta$ for all $t \geq 0$. Which allows us to bound the cumulative regret as,

$$\begin{aligned}
R_T = \sum_{t=1}^T r_t &\leq \sqrt{T\sum_{t=1}^T r_t^2}\\
&\leq 2\beta_T(\delta)\sqrt{T\sum_{t=1}^T \sigma_t^2(\boldsymbol{x}_t)} && \beta_t(\delta)\leq\beta_T(\delta)\\
&\leq C_1\beta_T(\delta)\sqrt{T\gamma_t} && \text{Lem. 14}
\end{aligned}$$

where $C_1 := \sqrt{8/\log(1+(\lambda\kappa)^{-1})}$. □

C Helper Lemmas for Appendix B

Lemma 11 (Mean-Value Theorem)**.** *For any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \ell_2(\mathbb{N})$ it holds that*

$$s(\boldsymbol{\theta}_2^\top \boldsymbol{\phi}(\boldsymbol{x})) - s(\boldsymbol{\theta}_1^\top \boldsymbol{\phi}(\boldsymbol{x})) = \alpha(\boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^\top \boldsymbol{\phi}(\boldsymbol{x})$$

where

$$\alpha(\boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \int_0^1 \dot{s}(\nu \boldsymbol{\theta}_2^\top \boldsymbol{\phi}(\boldsymbol{x}) + (1-\nu)\boldsymbol{\theta}_1^\top \boldsymbol{\phi}(\boldsymbol{x})) \mathrm{d}\nu$$

Proof of Lemma 11. For any differentiable function $s : \mathbb{R} \to \mathbb{R}$ by the fundamental theorem of calculus we have

$$s(z_2) - s(z_1) = \int_{z_1}^{z_2} \dot{s}(z) \mathrm{d}z.$$

Define $\nu = (z - z_1)/(z_2 - z_1)$, then $z = \nu z_2 + (1-\nu) z_1$ and re-writing the integral in terms of ν gives,

$$s(z_2) - s(z_1) = (z_2 - z_1) \int_0^1 \dot{s}(\nu z_2 + (1-\nu) z_1) \mathrm{d}\nu.$$

Letting $z_1 = \boldsymbol{\theta}_1^\top \boldsymbol{\phi}(\boldsymbol{x})$ and $z_2 = \boldsymbol{\theta}_2^\top \boldsymbol{\phi}(\boldsymbol{x})$ concludes the proof. □

Lemma 12 (Gradients to Parameters Conversion)**.** *For all $t \geq 0$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \ell_2(\mathbb{N})$ it holds that,*

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{V_t} \leq \kappa \|g_t(\boldsymbol{\theta}_1) - g_t(\boldsymbol{\theta}_2)\|_{V_t^{-1}}$$

Proof of Lemma 12. We proof the lemma through an auxiliary operator $G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) : \mathcal{H} \to \mathcal{H}$ defined as

$$G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \lambda \boldsymbol{I}_{\mathcal{H}} + \sum_{s \leq t} \alpha(\boldsymbol{x}_s; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\phi}(\boldsymbol{x}_s) \boldsymbol{\phi}^\top(\boldsymbol{x}_s).$$

Step 1. First we establish how we can go back and forth between the operator norms defined based on G_t and V_t. Recall that $\kappa = \sup_z \frac{1}{\dot{s}(z)}$. Therefore, $\kappa^{-1} \leq \dot{s}(z)$ for all $z \in \mathbb{R}$, implying that $\alpha(\boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq \int_0^1 \kappa^{-1} \mathrm{d}\nu = \kappa^{-1}$. We can then conclude,

$$G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq \lambda \boldsymbol{I}_{\mathcal{H}} + \sum_{s \leq t} \kappa^{-1} \boldsymbol{\phi}(\boldsymbol{x}_s) \boldsymbol{\phi}^\top(\boldsymbol{x}_s) = \kappa^{-1} V_t. \tag{10}$$

Step 2. Now by the definition of $g_t(\boldsymbol{\theta})$,

$$\begin{aligned} g_t(\boldsymbol{\theta}_2) - g_t(\boldsymbol{\theta}_1) &= \lambda(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) + \sum_{s \leq t} \boldsymbol{\phi}(\boldsymbol{x}_s) \left[s(\boldsymbol{\theta}_2^\top \boldsymbol{\phi}(\boldsymbol{x}_s)) - s(\boldsymbol{\theta}_1^\top \boldsymbol{\phi}(\boldsymbol{x}_s)) \right] \\ &= \lambda(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) + \sum_{s \leq t} \boldsymbol{\phi}(\boldsymbol{x}_s) \left[\alpha(\boldsymbol{x}_s; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\phi}^\top(\boldsymbol{x}_s)(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \right] & \text{(Lem. 11)} \\ &= \left(\lambda \boldsymbol{I}_{\mathcal{H}} + \sum_{s \leq t} \alpha(\boldsymbol{x}_s; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\phi}(\boldsymbol{x}_s) \boldsymbol{\phi}^\top(\boldsymbol{x}_s) \right) (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \\ &= G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \end{aligned}$$

Therefore,

$$\begin{aligned} \|g_t(\boldsymbol{\theta}_2) - g_t(\boldsymbol{\theta}_1)\|_{G_t^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} &= \left[g_t(\boldsymbol{\theta}_2) - g_t(\boldsymbol{\theta}_1) \right]^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\ &= (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^\top G_t\, (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \\ &= \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}. \end{aligned} \tag{11}$$

Step 3. Putting together the previous two steps, we can bound the V_t-norm over the parameters to the V_t^{-1} role in the gradients,

$$\begin{aligned} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{V_t} &\overset{(10)}{\leq} \sqrt{\kappa} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{G_t^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\ &\overset{(11)}{\leq} \sqrt{\kappa} \|g_t(\boldsymbol{\theta}_1) - g_t(\boldsymbol{\theta}_2)\|_{G_t^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\ &\overset{(10)}{\leq} \kappa \|g_t(\boldsymbol{\theta}_1) - g_t(\boldsymbol{\theta}_2)\|_{V_t^{-1}} \end{aligned}$$

concluding the proof. $\square$

The following two lemmas are standard results in kernelized bandits [Srinivas et al., 2010, Chowdhury and Gopalan, 2017, e.g.,]. We include it here for completeness.

Lemma 13. *For any $\boldsymbol{x} \in \mathcal{X}$ and $\rho > 0$ it holds that $\sqrt{\lambda\kappa}\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}} = \sigma_t(\boldsymbol{x})$, where*

$$\sigma_t^2(\boldsymbol{x}, \rho) = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_t^\top(\boldsymbol{x})(K_t + \rho^2 \boldsymbol{I}_t)^{-1}\boldsymbol{k}_t(\boldsymbol{x})$$

with $\boldsymbol{k}_t(\boldsymbol{x}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}, \boldsymbol{x}_t)) \in \mathbb{R}^t$.

Proof of Lemma 13. We start by stating some identities which will later be of use. First note that

$$\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)\Phi_t^\top = \Phi_t^\top \left(\Phi_t \Phi_t^\top + \rho^2 \boldsymbol{I}_t\right)$$

which gives

$$\Phi_t^\top \left(\Phi_t \Phi_t^\top + \rho^2 \boldsymbol{I}_t\right)^{-1} = \left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)^{-1}\Phi_t^\top. \tag{12}$$

Moreover, by definition of $\boldsymbol{k}_t$ we have

$$\boldsymbol{k}_t(\boldsymbol{x}) = \Phi_t \boldsymbol{\phi}(\boldsymbol{x}) \tag{13}$$

which allow us to write,

$$\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)\boldsymbol{\phi}(\boldsymbol{x}) = \Phi_t^\top \boldsymbol{k}_t(\boldsymbol{x}) + \rho^2 \boldsymbol{\phi}(\boldsymbol{x})$$

and obtain,

$$\begin{aligned}\boldsymbol{\phi}(\boldsymbol{x}) &= \left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)^{-1}\Phi_t^\top \boldsymbol{k}_t(\boldsymbol{x}) + \rho^2\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x})\\ &\overset{(12)}{=} \Phi_t^\top \left(\Phi_t \Phi_t^\top + \rho^2 \boldsymbol{I}_t\right)^{-1}\boldsymbol{k}_t(\boldsymbol{x}) + \rho^2\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x}).\end{aligned}$$

Given the above equation, we conclude the proof by the following chain of equations:

$$\begin{aligned}k(\boldsymbol{x}, \boldsymbol{x}) &= \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\\ &= \left(\Phi_t^\top \left(\Phi_t \Phi_t^\top + \rho^2 \boldsymbol{I}_t\right)^{-1}\boldsymbol{k}_t(\boldsymbol{x}) + \rho^2\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x})\right)^\top \boldsymbol{\phi}(\boldsymbol{x})\\ &= \boldsymbol{k}_t^\top(\boldsymbol{x})\left(\Phi_t \Phi_t^\top + \rho^2 \boldsymbol{I}_t\right)^{-1}\Phi_t \boldsymbol{\phi}(\boldsymbol{x}) + \rho^2 \boldsymbol{\phi}^\top(\boldsymbol{x})\left(\Phi_t^\top \Phi_t + \rho^2 \boldsymbol{I}_{\mathcal{H}}\right)^{-1}\boldsymbol{\phi}(\boldsymbol{x})\\ &\overset{(13)}{=} \boldsymbol{k}_t^\top(\boldsymbol{x})\left(K_t + \rho^2 \boldsymbol{I}_t\right)^{-1}\boldsymbol{k}_t(\boldsymbol{x}) + \rho^2 \boldsymbol{\phi}^\top(\boldsymbol{x}) V_t^{-1}\boldsymbol{\phi}(\boldsymbol{x})\end{aligned}$$

To obtain the third equation we have used the fact that for bounded operators on hilbert spaces, the inverse of the adjoint is equal to the adjoint of the inverse [e.g., Theorem 10.19 Axler, 2020]. $\square$

Lemma 14. *For all $T \geq 1$,*

$$\sum_{t=1}^T \sigma_t^2(\boldsymbol{x}_t) \leq \frac{2\gamma_T}{\log(1 + (\lambda\kappa)^{-1})}, \quad \sum_{t=1}^T (\sigma_t^{\mathrm{D}}(\boldsymbol{x}_t))^2 \leq \frac{8\gamma_T^{\mathrm{D}}}{\log(1 + 4(\lambda\kappa)^{-1})}.$$

Proof of Lemma 14. By Srinivas et al. [2010, Lemma 5.3],

$$\gamma_T = \max_{\boldsymbol{x}_1, \dots \boldsymbol{x}_T} \frac{1}{2}\sum_{t=1}^T \log(1 + (\lambda\kappa)^{-1}\sigma_{t-1}^2(\boldsymbol{x}_t)).$$

Following the technique in Srinivas et al. [2010, Lemma 5.4], since $\sigma_t^2 \leq 1$, then $\rho^{-1}\sigma_t^2 \in [0, \rho^{-1}]$. Now for any $z \in [0, \rho^{-1}]$, $z \leq C\log(1+z)$ where $C = 1/(\rho\log(1+\rho^{-1}))$. We then may write,

$$\begin{aligned}\sum_{t=1}^T \sigma_t^2(\boldsymbol{x}_t) &= \sum_{t=1}^T \lambda\kappa(\lambda\kappa)^{-1}\sigma_t^2(\boldsymbol{x}_t)\\ &\leq \sum_{t=1}^T \lambda\kappa C\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^2(\boldsymbol{x}_t)\right)\\ &= \sum_{t=1}^T \frac{\log(1 + (\lambda\kappa)^{-1}\sigma_t^2(\boldsymbol{x}_t))}{\log\left(1 + (\lambda\kappa)^{-1}\right)}\end{aligned}$$

Putting both together proves the first inequality of the lemma. As for the dueling case, we can easily check that $\sigma_t^{\mathrm{D}} \leq 2$, and a similar argument yields the second inequality. $\square$

D Proofs for Bandits with Preference Feedback

This section presents the proof of main results in Section 5, and our additional contributions in the kernelized Preference-based setting.

D.1 Equivalence of Preference-based and Logistic Losses

We start by establishing the equivalence between the logistic loss of Equation (3) and dueling loss Equation (6).

Proof of Proposition 4. By Mercer's theorem, we know that the kernel function k has eigenvalue eigenfunction pairs $(\sqrt{\lambda_i}, \tilde{\phi}_i)$ for $i \geq 1$ where $\tilde{\boldsymbol{\phi}}_i$ are orthonormal. Then $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i\geq 1} \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}')$ with $\phi_i(\boldsymbol{x}) = \sqrt{\lambda_i}\tilde{\phi}_i(\boldsymbol{x})$. Now applying the definition of k^{D}, it holds that $k^{\mathrm{D}}(\boldsymbol{z}, \boldsymbol{z}') = \sum_{i\geq 1} \psi_i^\top(\boldsymbol{z})\psi_i(\boldsymbol{z}')$ where $\psi_i(\boldsymbol{z}) = \sqrt{\lambda_i}(\phi_i(\boldsymbol{x}) - \phi_i(\boldsymbol{x}'))$. It is straighforward to check that ψ_i are the eigenfunctions of k^{D}, however, they may not be orthonormal. We have,

$$
\begin{aligned}
\langle \psi_i, \psi_i \rangle_{L_2} &= 2\lambda_i(1 - b_i^2) \\
\langle \psi_i, \psi_j \rangle_{L_2} &= -2\sqrt{\lambda_i \lambda_j} b_i b_j
\end{aligned}
$$

where $b_i = \int \tilde{\phi}_i(\boldsymbol{x})\mathrm{d}(\boldsymbol{x})$. By the assumption of the proposition, we have $b_i = 0$. However, this assumption holds automatically for all kernels commonly used in applications, e.g. any translation invariant kernel, over many domains, since $\tilde{\boldsymbol{\phi}}_i$ for such kernels are a sine basis.

Now since $f \in \mathcal{H}_k$, it may be decomposed $f = \sum_{i\geq 1} \beta_i \phi_i$ and $\|f\|_k^2 = \sum_{i\geq 1} \beta_i^2 \leq \infty$. Therefore for the difference function we may write $h(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i\geq 1} \beta_i \psi_i(\boldsymbol{z})$. We can then bound the RKHS norm of h w.r.t. the kernel k^{D} as follows

$$
\begin{aligned}
\|h\|_{k^{\mathrm{D}}}^2 &= \sum_{i\geq 1} \left(\frac{\langle h, \psi_i \rangle_{L_2}}{\langle \psi_i, \psi_i \rangle_{L_2}} \right)^2 \\
&= \sum_{i\geq 1} \left(\frac{\sum_{j\geq 1} \beta_j \langle \psi_j, \psi_i \rangle_{L_2}}{2\lambda_i(1 - b_i)} \right)^2 \\
&= \sum_{i\geq 1} \left(\beta_i - \frac{b_i}{\sqrt{\lambda_i}(1 - b_i)} \sum_{j\neq i} \beta_j b_j \sqrt{\lambda_j} \right)^2 \\
&\overset{b_i=0}{=} \|f\|_k^2 \leq B^2.
\end{aligned}
$$

Now by Mercer's theorem, $h \in \mathcal{H}_{k^{\mathrm{D}}}$ since it is decomposable as a sum of k^{D} eigenfunctions, and attains a B-bounded k^{D}-norm which we showed to be equal to $\|f\|_k$. ☐

Proof of Corollary 5. Consider the utility function f and define $h(\boldsymbol{x}, \boldsymbol{x}') \coloneqq f(\boldsymbol{x}) - f(\boldsymbol{x}')$. Then by Proposition 4, h is in RKHS of k^{D} with a k^{D}-norm bounded by B. We may estimate h by minimizing $\mathcal{L}_{k^{\mathrm{D}}}^{\mathrm{L}}(\cdot; H_t)$. Now invoking Theorem 2 with the dueling kernel we have,

$$
\mathbb{P}\left(\forall t \geq 1, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} : \left| s\left(h_t(\boldsymbol{x}, \boldsymbol{x}')\right) - s\left(h(\boldsymbol{x}, \boldsymbol{x}')\right) \right| \leq \beta_t^{\mathrm{D}}(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}') \right) \geq 1 - \delta
$$

concluding the proof by definition of h. ☐

D.2 Proof of the Preference-based Regret Bound

Recall Corollary 5, which states

$$
|s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t))| \leq \beta_t^{\mathrm{D}}(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}')
$$

with high probability simultaneously for all $(\boldsymbol{x}, \boldsymbol{x}')$ and $t \geq 1$. For simplicity in notation, we define $\omega_t(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \beta_t^{\mathrm{D}}(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}')$ and use it for the remainder of this section. Note that $\omega_t(\boldsymbol{x}, \boldsymbol{x}') = \omega_t(\boldsymbol{x}', \boldsymbol{x})$ by the symmetry of the dueling kernel k^{D}.

Proof of Theorem 6. Using Corollary 5, define

$$
\begin{aligned}
\mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') &\coloneqq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) - \omega_t(\boldsymbol{x}, \boldsymbol{x}'), \\
\mathrm{UCB}_t(\boldsymbol{x}, \boldsymbol{x}') &\coloneqq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) + \omega_t(\boldsymbol{x}, \boldsymbol{x}').
\end{aligned}
$$

We start by observing that since $s(z) = 1 - s(-z)$ then $\mathrm{LCB}(\boldsymbol{x}, \boldsymbol{x}') = 1 - \mathrm{UCB}(\boldsymbol{x}', \boldsymbol{x})$ and

$$\arg\max_{\boldsymbol{x}} \arg\min_{\boldsymbol{x}'} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') = \arg\min_{\boldsymbol{x}} \arg\max_{\boldsymbol{x}'} \mathrm{UCB}_t(\boldsymbol{x}', \boldsymbol{x})$$

where the max-min calculations are carried out sequentially, as in Equation (7).

Furthermore, we note that by the definition of $\boldsymbol{x}'_t$ in Equation (7), $LCB_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq LCB_t(\boldsymbol{x}_t, \boldsymbol{x}')$ for all $\boldsymbol{x}' \in \mathcal{M}_t$. Also, for all $\boldsymbol{x}$, $h_t(\boldsymbol{x}, \boldsymbol{x}) = 0.5$ and $\omega_t(\boldsymbol{x}, \boldsymbol{x}) = 0.0$ since it is known that $s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_t)) = s(0) = 0.5$, i.e., $LCB_t(\boldsymbol{x}, \boldsymbol{x}) = 0.5 = UCB_t(\boldsymbol{x}, \boldsymbol{x})$. Then, since $\boldsymbol{x}_t \in \mathcal{M}_t$, we get that $LCB_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq LCB_t(\boldsymbol{x}_t, \boldsymbol{x}_t) \leq 0.5$.

Step 1: First, we show the following inequality

$$s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}'_t)) \leq (1 + h(\boldsymbol{x}_t, \boldsymbol{x}'_t)) s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))$$

where $h(\boldsymbol{x}_t, \boldsymbol{x}'_t) = f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)$. Note that $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}'_t)) \geq 0.5$, $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) \geq 0.5$ and the sigmoid function s is concave on the interval $[0.5, \infty)$, i.e.,

$$\begin{aligned}
s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}'_t)) &\leq s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + s'(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)) \\
&= s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*))(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)) \\
&\leq (1 + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2}) s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))
\end{aligned}$$

where the first line is the definition of concavity, the second comes from the derivative of the sigmoid function, $s'(x) = s(x)(1 - s(x)) = s(x)s(-x)$, and in the last line we use $s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leq 0.5$.

Using this inequality, we can upper bound the average regret of the two arms with the regret of $\boldsymbol{x}_t$ as

$$\begin{aligned}
2r_t^D &= s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}'_t)) - 1 \\
&\leq s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + (1 + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2}) s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 1 \\
&\leq 2s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 1 + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t))
\end{aligned}$$

Step 2: Next, we show that the regret is bounded by $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$.

First,

$$\begin{aligned}
s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 0.5 &\leq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) - 0.5 && \text{Corollary 5} \\
&\leq 0.5 - s(h_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)) + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) && \text{Sigmoid equality} \\
&\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)
\end{aligned}$$

In the last inequality, we used that $\boldsymbol{x}_t \in \mathcal{M}_t$ implying that $0.5 - s(h_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)) \leq \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$. It implies then that

$$2r_t^D \leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) + h(\boldsymbol{x}_t, \boldsymbol{x}'_t) s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t))$$

Now, we bound $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$ by $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$. If $\boldsymbol{x}_t = \boldsymbol{x}^\star$, then $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) = 0 \leq \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$. If $\boldsymbol{x}'_t = \boldsymbol{x}^\star$, then the two expressions are equivalent. Now, assume that $\boldsymbol{x}_t \neq \boldsymbol{x}^*$ and $\boldsymbol{x}'_t \neq \boldsymbol{x}^\star$ and consider $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$.

Case 1: Assume that $\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \leq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$. Then,

$$\begin{aligned}
2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^*) &= \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\
&\leq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\
&\leq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \\
&= 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)
\end{aligned}$$

where we used the definition of $\boldsymbol{x}'_t$ in the second inequality.

Case 2: Assume that $\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \geq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$. First note that the assumption implies that $\mathrm{LCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) \geq \mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t)$ since $\mathrm{UCB}_t(\boldsymbol{x}, \boldsymbol{x}') = 1 - \mathrm{LCB}_t(\boldsymbol{x}', \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. Similarly, $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$ implies $\mathrm{UCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \geq \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)$.

$$\begin{aligned}
2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^*) &= \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) \\
&\leq \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \\
&\leq \mathrm{UCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \\
&= 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)
\end{aligned}$$

Therefore,

$$2r_t^D \leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t))$$

Step 3: Next, we upper bound the second term. Define $\Delta \coloneqq h(\boldsymbol{x}_t, \boldsymbol{x}'_t)$. By the Mean-Value Theorem, $\exists z \in [0, \Delta]$ such that

$$\dot{s}(z)(\Delta - 0) = s(\Delta) - f(0)$$

Now since $\kappa = \sup_z 1/\dot{s}(z)$ then,

$$\Delta \leq \kappa(s(\Delta) - 0.5) \leq \kappa/2 \tag{14}$$

Next, we consider the right-hand side of the inequality. Note that $\boldsymbol{x}_t, \boldsymbol{x}'_t \in \mathcal{M}_t$ implies that

$$\begin{aligned} \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) &\geq 0.5 \\ s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) &\geq 0.5 - \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned}$$

additionally $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 0.5$ implies that

$$\begin{aligned} \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) &\leq 0.5 \\ s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) &\leq 0.5 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t). \end{aligned}$$

From these two inequalities, it follows that

$$|s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5| \leq \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$$

furthermore,

$$\begin{aligned} \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - 0.5 &= s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \\ &\leq |s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5| + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \\ &\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned}$$

and similarly

$$0.5 - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$$

From these upper bounds on the distance between the ends of the confidence interval and the middle point of 0.5, it follows that

$$\begin{aligned} |s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)) - 0.5| &\leq \max\{\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - 0.5, 0.5 - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)\} && \text{Corollary 5} \\ &\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned}$$

Combining this inequality with Equation (14) and using the fact that $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) \leq 1$, we get that

$$2r_t^D \leq (4 + \kappa)\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t). \tag{15}$$

Therefore, for the cumulative dueling regret it holds

$$\begin{aligned} R^{\mathrm{D}}(T) = \sum_{t=1}^T r_t^{\mathrm{D}} &\leq \sqrt{T \sum_{t=1}^T (r_t^{\mathrm{D}})^2} \\ &\leq (2 + \kappa/2)\beta_T^{\mathrm{D}}(\delta)\sqrt{T \sum_{t=1}^T (\sigma_t^{\mathrm{D}})^2(\boldsymbol{x}_t, \sqrt{\lambda\kappa})} && \beta_t(\delta) \leq \beta_T^{\mathrm{D}}(\delta) \\ &\leq C_3 \beta_T^{\mathrm{D}}(\delta)\sqrt{T\gamma_t^{\mathrm{D}}} && \text{Lem. 14} \end{aligned}$$

with probability greater than $1 - \delta$ for all $T \geq 1$.

□

D.3 Extension of Linear Dueling bandits to Kernelized Utilities

Maximum Informative Pair Algorithm. Propose in Saha [2021] for linear utilities, the MAXINP algorithm similarly maintains a set of plausible maximizer arms, and picks the pair of actions that have the largest joint uncertainty, and therefore are expected to be informative. Algorithm 3 present the kernelized variant of this algorithm in detail. Using Corollary 5, we can show that the kernelized MAXINP also satisfies a $\tilde{\mathcal{O}}(\gamma_T\sqrt{T})$ regret.

Theorem 15. *Let $\delta \in (0,1]$ and choose the exploration coefficient $\beta_t^{\mathrm{D}}(\delta)$ as defined in Corollary 5. Then* MAXINP *satisfies the anytime dueling regret guarantee of*

$$\mathbb{P}\left(\forall T \geq 0 : R^{\mathrm{D}}(T) \leq C_2 \beta_T^{\mathrm{D}}(\delta)\sqrt{T\gamma_T^{\mathrm{D}}}\right) \geq 1-\delta$$

where γ_T^{D} is the T-step information gain of kernel k^{D} and $C_2 = 4/\sqrt{\log(1+4(\lambda\kappa)^{-1})}$.

Proof of Theorem 15. When selecting $(\boldsymbol{x}_t, \boldsymbol{x}'_t)$ according to Algorithm 3, we choose the pair via

$$\boldsymbol{x}_t, \boldsymbol{x}'_t = \operatorname*{arg\,max}_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{M}_t} \omega_t(\boldsymbol{x}, \boldsymbol{x}') \tag{16}$$

where action space is restricted

$$\mathcal{M}_t = \{\boldsymbol{x} \in \mathcal{X} \text{ s.t. } s(h_t(\boldsymbol{x},\boldsymbol{x}')) + \omega_t(\boldsymbol{x},\boldsymbol{x}') \geq 1/2\}\,.$$

Since $\boldsymbol{x}_t, \boldsymbol{x}'_t \in \mathcal{M}_t$, then

$$\begin{aligned} s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) &\leq 1/2 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\ s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)) &\leq 1/2 + \omega_t(\boldsymbol{x}'_t, \boldsymbol{x}^\star) \end{aligned} \tag{17}$$

where we have used the identity $s(-z) = 1 - s(z)$. Simultaneously for all $t \geq 1$, we can bound the single-step dueling regret with probability greater than $1-\delta$

$$\begin{aligned} 2r_t^{\mathrm{D}} &= s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}'_t)) - 1 \\ &\leq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) + s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t) - 1 && \text{(w.h.p.)} \\ &\leq 2\left(\omega_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)\right) && \text{Eq. (17)} \\ &\leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) && \text{Eq. (16)} \end{aligned}$$

where for the first inequality we have invoked Corollary 5. Therefore, for the cumulative dueling regret it holds that

$$\begin{aligned} R^{\mathrm{D}}(T) = \sum_{t=1}^{T} r_t^{\mathrm{D}} &\leq \sqrt{T\sum_{t=1}^{T}(r_t^{\mathrm{D}})^2} \\ &\leq 2\beta_T^{\mathrm{D}}(\delta)\sqrt{T\sum_{t=1}^{T}(\sigma_t^{\mathrm{D}})^2(\boldsymbol{x}_t, \sqrt{\lambda\kappa})} && \beta_t(\delta) \leq \beta_T^{\mathrm{D}}(\delta) \\ &\leq C_2\beta_T^{\mathrm{D}}(\delta)\sqrt{T\gamma_t^{\mathrm{D}}} && \text{Lem. 14} \end{aligned}$$

with probability greater than $1-\delta$ for all $T \geq 1$. □

Dueling Information Directed Sampling (IDS) Algorithm. To choose actions at each iteration t, MAXINP and MAXMINLCB require solving an optimization problem on $\mathcal{X} \times \mathcal{X}$. The Dueling IDS approach addresses this issue and presents a algorithm which requires solving an optimization problem on $\mathcal{X} \times [0,1]$, which is computationally more efficient when $d_0 > 1$. This work considers kernelized utilities, however, assumes the probability of preference itself is in an RKHS and solves a kernelized ridge regression problem to estimate the probability $s(h(\boldsymbol{x},\boldsymbol{x}')$. In the following, we present an improved version of this algorithm, by considering the preference-based loss (6) for estimating the utility function. We modify the algorithm and the theoretical analysis to accomodate for this change.

Consider the suboptimality gap $\Delta(\boldsymbol{x}) \coloneqq h(\boldsymbol{x}^\star, \boldsymbol{x})$ for an action $\boldsymbol{x} \in \mathcal{X}$. We may estimate this gap using the reward estimate maximizer $\hat{\boldsymbol{x}}_t^\star \coloneqq \arg\max_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x})$. Suppose we choose $\hat{\boldsymbol{x}}_t^\star$ as one of

Algorithm 3 MAXINP- Kernelized Variant

Input $(\beta_t^{\mathrm{D}})_{t\geq 1}$.
for $t \geq 1$ **do**
  Play the most informative pair via

$$\boldsymbol{x}_t, \boldsymbol{x}'_t = \arg\max_{\boldsymbol{x},\boldsymbol{x}' \mathcal{M}_t} \sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}')$$

  Observe y_t and append history.
  Update h_{t+1} and $\sigma_{t+1}^{\mathrm{D}}$ and the set of plausible maximizers

$$\mathcal{M}_{t+1} = \{\boldsymbol{x} \in \mathcal{X} | \forall \boldsymbol{x}' \in \mathcal{X} :\ s(h_{t+1}(\boldsymbol{x}, \boldsymbol{x}')) + \beta_{t+1}^{\mathrm{D}} \sigma_{t+1}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}') > 1/2\}.$$

end for

Algorithm 4 Dueling IDS - Kernelized Logistic Varniant

Initialize Set $(\beta_t)_{t\geq 1}$ according to Theorem 2.
for $t \geq 1$ **do**
  Find a greedy action via fixing any point $x_{\mathrm{null}} \in \mathcal{X}$ and maximizing

$$\boldsymbol{x}_t^{(1)} = \hat{\boldsymbol{x}}_t^\star = \arg\max_{x\in\mathcal{X}} h_t(x, x_{\mathrm{null}})$$

  Update u_t and $\hat{\Delta}_t(\boldsymbol{x})$ acc. to (18)
  Find an informative action and the probability of selection via

$$\boldsymbol{x}_t^{(2)}, p_t = \arg\min_{\substack{\boldsymbol{x}\in\mathcal{X}\\ p\in[0,1]}} \frac{\left((1-p)u_t + p\hat{\Delta}_t(\boldsymbol{x})\right)^2}{p \log\left(1 + (\lambda\kappa)^{-1}\left(\sigma_t^D(\boldsymbol{x}_t^{(1)}, \boldsymbol{x})\right)^2\right)}$$

  Draw $\alpha_t \sim \mathrm{Bern}(p_t)$.
  if $\alpha_t = 1$ **then**
    Choose action pair $(\boldsymbol{x}_t, \boldsymbol{x}'_t) = (\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)})$
  else
    Choose action pair $(\boldsymbol{x}_t, \boldsymbol{x}'_t) = (\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(1)})$
  end if
  Observe y_t and append history.
  Update h_{t+1} and $\sigma_{t+1}^{\mathrm{D}}$.
end for

the actions. Then u_t, as defined below, shows an optimistic estimate of the highest obtainable reward at this step

$$u_t := \max_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}, \hat{\boldsymbol{x}}_t^\star) + \tilde{\beta}_t \sigma_t^D(\boldsymbol{x}, \boldsymbol{x}_t^\star).$$

where $\tilde{\beta}_t$ is the exploration coefficient. We bound $\Delta(\boldsymbol{x})$ by the estimated gap

$$\hat{\Delta}_t(\boldsymbol{x}) := u_t + h_t(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) \tag{18}$$

and show its uniform validity in Lemma 17. Given this gap estimate, we propose the Kernelized Logistic IDS algorithm with dueling feedback in Algorithm 4, as a variant of the algorithm of Kirschner and Krause.

Theorem 16. *Let $\delta \in (0,1]$ and for all $t \geq 1$, set the exploration coefficient as $\tilde{\beta}_t = \beta_t^{\mathrm{D}}(\delta)/L$. Then Algorithm 4 satisfies the anytime cumulative dueling regret guarantee of*

$$\mathbb{P}\left(\forall T \geq 0 : R^{\mathrm{D}}(T) = \mathcal{O}\left(\beta_T^{\mathrm{D}}(\delta)\sqrt{T(\gamma_T + \log 1/\delta)}\right)\right) \geq 1 - \delta.$$

Proof of Theorem 16. The proof closely follows proof of Kirschner and Krause [2021, Theorem 1]. Define the expected average gap for a policy $\mu \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$

$$\hat{\Delta}_t(\mu) := \frac{1}{2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\mu}\hat{\Delta}_t(\boldsymbol{x}) + \hat{\Delta}_t(\boldsymbol{x}')$$

and the expected information ratio as

$$\Xi_t(\mu) = \frac{\hat{\Delta}_t^2(\mu)}{\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\mu} \log\left(1 + (\lambda\kappa)^{-1}\big(\sigma_t^D(\boldsymbol{x},\boldsymbol{x}')\big)^2\right)}.$$

Consider Algorithm 4, and for $t \geq 1$ let $\mu_t = (1-p_t)\delta_{(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(1)})} + p_t\delta_{(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(2)})}$ denote the action-taking policy μ_t defined over $\mathcal{X}\times\mathcal{X}$. Here $\delta_{(\boldsymbol{x},\boldsymbol{x}')}$ denotes a Direct delta. Then by Kirschner et al. [2020, Lemma 1],

$$\frac{1}{2}\sum_{t=1}^T h(\boldsymbol{x}^\star,\boldsymbol{x}_t) + h(\boldsymbol{x}^\star,\boldsymbol{x}_t') \leq \sqrt{\sum_{t=1}^T \Xi_t(\mu_t)\left(\gamma_T + \mathcal{O}(\log 1/\delta)\right) + \mathcal{O}(\log T/\delta)}$$

which allows us to bound the regret with probability greater than $1-\delta$ as

$$R^{\mathrm{D}}(T) \leq L\sqrt{\sum_{t=1}^T \Xi_t(\mu_t)\left(\gamma_T + \mathcal{O}(\log 1/\delta)\right) + \mathcal{O}(L\log T/\delta)} \tag{19}$$

since $s(\cdot)$ with its domain restricted to $[-2B, 2B]$ is L-Lipschitz. It remains to bound $\Xi_t(\mu_t)$, the expected information ratio for Algorithm 4. Now by definition of μ_t

$$\begin{aligned}
2\hat{\Delta}_t(\mu_t) &= (2-p_t)\hat{\Delta}_t(\boldsymbol{x}_t^{(1)}) + p_t\Delta_t(\boldsymbol{x}_t^{(2)}) \\
&= (2-p_t)\left(u_t + h_t(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(1)})\right) + p_t\Delta_t(\boldsymbol{x}_t^{(2)}) \\
&= 2(1-p_t)u_t + p_t(\hat{\Delta}_t(\boldsymbol{x}_t^{(2)}) + u_t),
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{E}_{\mu_t}\log\left(1 + \tfrac{\sigma_t^D(\boldsymbol{x},\boldsymbol{x}')^2}{\lambda\kappa}\right) &= (1-p_t)\log\left(1 + \tfrac{\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(1)})^2}{\lambda\kappa}\right) + p_t\log\left(1 + \tfrac{\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(2)})^2}{\lambda\kappa}\right) \\
&= p_t\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(2)})^2\right) && (\sigma_t^D(\boldsymbol{x},\boldsymbol{x}) = 0)
\end{aligned}$$

allowing us to re-write the expected information ratio as

$$\begin{aligned}
\Xi_t(\mu_t) &= \frac{\left(2(1-p_t)u_t + p_t(\hat{\Delta}_t(\boldsymbol{x}_t^{(2)}) + u_t)\right)^2}{4p_t\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(2)})^2\right)} \\
&\leq \frac{\left((1-p_t)u_t + p_t\hat{\Delta}_t(\boldsymbol{x}_t^{(2)})\right)^2}{p_t\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x}_t^{(2)})^2\right)} && (u_t \leq \hat{\Delta}_t(\boldsymbol{x})) \\
&= \min_{\boldsymbol{x},p}\frac{\left((1-p)u_t + p\hat{\Delta}_t(\boldsymbol{x})\right)^2}{p\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x})^2\right)} && \text{Def. } (p_t, \boldsymbol{x}_t^{(2)}) \\
&\leq \min_{\boldsymbol{x}}\frac{\hat{\Delta}_t^2(\boldsymbol{x})}{\log\left(1 + (\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x})^2\right)}. && \text{Set } p = 1
\end{aligned}$$

Now consider the definition of u_t and let $\boldsymbol{z}_t$ denote the action for which u_t is achieves, i.e. $\boldsymbol{z}_t = \arg\max h(\boldsymbol{x},\hat{\boldsymbol{x}}_t^\star) + \bar{\beta}_t(\delta)\sigma_t^{\mathrm{D}}(\boldsymbol{x},\hat{\boldsymbol{x}}_t^\star)$. Then

$$\hat{\Delta}_t(\boldsymbol{z}_t) = h(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{z}_t) + \bar{\beta}_t(\delta)\sigma_t^{\mathrm{D}}(\boldsymbol{z}_t,\hat{\boldsymbol{x}}_t^\star) + h(\boldsymbol{z}_t,\hat{\boldsymbol{x}}_t^\star) = \bar{\beta}_t(\delta)\sigma_t^{\mathrm{D}}(\boldsymbol{x},\hat{\boldsymbol{x}}_t^\star),$$

therefore using the above chain of equations we may write

$$
\begin{aligned}
\Xi_t(\mu_t) &\le \min_{\boldsymbol{x}} \frac{\hat{\Delta}_t^2(\boldsymbol{x})}{\log\left(1+\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x})^2\right)} \\
&\le \frac{\hat{\Delta}_t^2(\boldsymbol{z}_t)}{\log\left(1+(\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{z}_t)^2\right)} \\
&\le \frac{\bar{\beta}_t^2(\delta)\sigma_t^{\mathrm{D}}(\boldsymbol{z}_t\hat{\boldsymbol{x}}_t^\star)^2}{\log\left(1+(\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{z}_t)^2\right)} \\
&\le \frac{4\bar{\beta}_t^2(\delta)}{\log\left(1+4(\lambda\kappa)^{-1}\right)} \qquad (20)
\end{aligned}
$$

where last inequality holds due to the following argument. Recall that $k(\boldsymbol{x},\boldsymbol{x}) \le 1$, implying that $\sigma_t^{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')^2 \le 4$ and therefore $\log(1+\sigma_t^{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')^2) \ge \log(1+(\lambda\kappa)^{-1})\sigma_t^{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')^2/4$, similar to Lemma 14. To conclude the proof, from (19) and (20) it holds that

$$
\begin{aligned}
R^{\mathrm{D}}(T) &\le L\sqrt{\sum_{t=1}^T \Xi_t(\mu_t)\left(\gamma_T+\mathcal{O}(\log 1/\delta)\right)+\mathcal{O}(L\log T/\delta)} \\
&\le L\sqrt{\sum_{t=1}^T \frac{4\bar{\beta}_t^2(\delta)}{\log\left(1+4(\lambda\kappa)^{-1}\right)}\left(\gamma_T+\mathcal{O}(\log 1/\delta)\right)+\mathcal{O}(L\log T/\delta)} \\
&\le L\sqrt{\frac{4T\bar{\beta}_T^2(\delta)}{\log\left(1+4(\lambda\kappa)^{-1}\right)}\left(\gamma_T+\mathcal{O}(\log 1/\delta)\right)+\mathcal{O}(L\log T/\delta)} \\
&= \mathcal{O}\left(\beta_T^{\mathrm{D}}(\delta)\sqrt{T(\gamma_T+\log 1/\delta)}\right)
\end{aligned}
$$

with probability greater than $1-\delta$, simultaneously for all $T \ge 1$. □

D.3.1 Helper Lemmas for Appendix D.3

Lemma 17. *Assume $f \in \mathcal{H}_k$. Suppose $\sup_{a\le B} \dot{s}(a) = L$ and $\sup_{a\le B} 1/\dot{s}(a) = \kappa$. Then for any $0<\delta<1$*

$$
\mathbb{P}(\forall t \ge 0, \boldsymbol{x}\in\mathcal{X}:\ \Delta(\boldsymbol{x}) \le 2\hat{\Delta}_t(\boldsymbol{x})) \ge 1-\delta.
$$

Proof of Lemma 17. Note that for any three inputs $\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3$

$$
h(\boldsymbol{x}_1,\boldsymbol{x}_3) = h(\boldsymbol{x}_1,\boldsymbol{x}_2)+h(\boldsymbol{x}_2,\boldsymbol{x}_3). \qquad (21)
$$

Therefore, from the definition of the estimated gap get

$$
\begin{aligned}
\hat{\Delta}_t(\boldsymbol{x}) &= \max_{\boldsymbol{z}\in\mathcal{X}} h(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star)+h_t(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x})+\bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star) \\
&= \max_{\boldsymbol{z}\in\mathcal{X}} h(\boldsymbol{z},\boldsymbol{x})+\bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star) \\
&\ge h(\boldsymbol{x},\boldsymbol{x})+\bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x},\boldsymbol{x}_t^\star) \\
&= \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x},\boldsymbol{x}_t^\star). \qquad (22)
\end{aligned}
$$

Then going back to the definition of the true gap we may write

$$
\begin{aligned}
\Delta(\boldsymbol{x}) &= \max_{\boldsymbol{z}\in\mathcal{X}} h(\boldsymbol{z},\boldsymbol{x}) \\
&= \max_{\boldsymbol{z}\in\mathcal{X}} h(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star)+h(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x}) && \text{Eq. (21)} \\
&\overset{\text{w.h.p.}}{\le} \max_{\boldsymbol{z}\in\mathcal{X}} h_t^P(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star)+h_t(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x})+\bar{\beta}_t(\delta)\left(\sigma_t^D(\boldsymbol{z},\hat{\boldsymbol{x}}_t^\star)+\sigma_t^D(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x})\right) && \text{Lem. 18} \\
&= u_t+h_t^P(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x})+\bar{\beta}_t(\delta)\sigma_t^D(\hat{\boldsymbol{x}}_t^\star,\boldsymbol{x}) && \text{Def. } u_t \\
&= \hat{\Delta}_t(\boldsymbol{x})+\bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}_t^\star,\boldsymbol{x}) && \text{Def. } \hat{\Delta}_t(\boldsymbol{x}) \\
&\le 2\hat{\Delta}_t(\boldsymbol{x}) && \text{Eq. (22)}
\end{aligned}
$$

with probability greater than $1-\delta$. □

Lemma 18. *Assume $f \in \mathcal{H}_k$. Suppose $\sup_{a \le B} 1/\dot{s}(a) = \kappa$. Then for any $0 < \delta < 1$*

$$\mathbb{P}\left(\forall t \geq 1, x \in \mathcal{X}: \left|h(\boldsymbol{x}, \boldsymbol{x}') - h_t^P(\boldsymbol{x}, \boldsymbol{x}')\right| \leq \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}'; \sqrt{\lambda\kappa})\right) \geq 1 - \delta$$

where

$$\bar{\beta}_t(\delta) := 2B + \sqrt{\frac{\kappa}{\lambda}}\sqrt{2\log 1/\delta + 2\gamma_t(\sqrt{\lambda\kappa})}$$

Proof of Lemma 18.

$$\begin{aligned}
\left|h(\boldsymbol{x}, \boldsymbol{x}') - h_t^P(\boldsymbol{x}, \boldsymbol{x}')\right| &= \left|f(\boldsymbol{x},) - f(\boldsymbol{x}') - (f_t^P(\boldsymbol{x},) - f_t^P(\boldsymbol{x}'))\right| \\
&= \left|\boldsymbol{\psi}^\top(\boldsymbol{x}, \boldsymbol{x}')(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_t^P)\right| \\
&\leq \|\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x}')\|_{(V_t^D)^{-1}} \left\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_t^P\right\|_{V_t^D} \\
&\overset{\text{w.h.p.}}{\leq} \sqrt{\lambda\kappa}\bar{\beta}_t(\delta)\|\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x}')\|_{(V_t^D)^{-1}} && \text{Lem. 9} \\
&\leq \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}, \sqrt{\lambda\kappa}) && \text{Lem. 13}
\end{aligned}$$

where the third to last inequality holds with probability greater than $1 - \delta$, but the rest of the inequalities hold deterministically. □

E Numerical Experiments

E.1 Implementation Details

E.1.1 Optimization Functions

We use the following functions that are standard in the optimization literature [Jamil and Yang, 2013] to evaluate the robustness of MaxMinLCB and report the results in Table 1 and Table 2. We present the functions as loss functions as common in the literature, however, for the experiments we negate them all to get utilities. We use a uniform grid of 100 points over their specified domains and scale the utility values to the range $[-3, 3]$.

- Ackley: $\mathcal{X} = [-5, 5]^2$
$$f(\boldsymbol{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)\right) + 20 + \exp(1)$$
- Branin: $\mathcal{X} = [-5, 10] \times [0, 15]$
$$f(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10$$
- Eggholder: $\mathcal{X} = [-512, 512]^2$
$$f(x_1, x_2) = -(x_2 + 47)\sin\left(\sqrt{|x_2 + \frac{x_1}{2} + 47|}\right) - x_1\sin\left(\sqrt{|x_1 - (x_2 + 47)|}\right)$$
- Hoelder: $\mathcal{X} = [-10, 10]^2$
$$f(x_1, x_2) = -|\sin(x_1)\cos(x_2)\exp\left(|1 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}|\right)|$$
- Matyas: $\mathcal{X} = [-10, 10]^2$
$$f(x_1, x_2) = 0.26(x_1^2 + x_2^2) - 0.48x_1x_2$$
- Michalewicz: $\mathcal{X} = [0, \pi]^2$
$$f(\boldsymbol{x}) = -\sum_{i=1}^{d}\sin(x_i)\sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$$
where $m = 10$ and d is the dimension of the input vector $\boldsymbol{x}$.
- Rosenbrock: $\mathcal{X} = [-5, 10]^2$
$$f(\boldsymbol{x}) = \sum_{i=1}^{d-1}\left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right]$$

Algorithm 5 DOUBLER [Ailon et al., 2014]

```
Input (β_t^D)_{t≥1}.
Let 𝓛 be any action from 𝒳
for t ≥ 1 do
    for j = 1, ..., 2^t do
        Select x'_t uniformly randomly from 𝓛
        Select x_t = arg max_{x∈𝓜_t} s(h_t(x, x'_t)) + β_t^D σ_t^D(x, x'_t)
        Observe y_t and append history.
        Update h_{t+1} and σ_{t+1}^D
    end for
    𝓛 ← the multi-set of actions played as x'_t in the last for-loop over index j
end for
```

Algorithm 6 MULTISBM [Ailon et al., 2014]

Input $(\beta_t^{\mathrm{D}})_{t\geq 1}$.
for $t \geq 1$ **do**
Set $\boldsymbol{x}_t \leftarrow \boldsymbol{x}'_{t-1}$
Select $\boldsymbol{x}'_t = \arg\max_{\boldsymbol{x}\in\mathcal{M}_t} s(h_t(\boldsymbol{x}, \boldsymbol{x}_t)) + \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}_t)$
Observe y_t and append history.
Update h_{t+1} and $\sigma_{t+1}^{\mathrm{D}}$ and the set of plausible maximizers

$$\mathcal{M}_{t+1} = \{\boldsymbol{x} \in \mathcal{X} | \, \forall \boldsymbol{x}' \in \mathcal{X} : \, s(h_{t+1}(\boldsymbol{x}, \boldsymbol{x}')) + \beta_{t+1}^{\mathrm{D}}\sigma_{t+1}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}') > 1/2\}.$$

end for

E.1.2 Action Selection Optimization

To eliminate additional noise in our comparisons coming from approximate solvers, we use an exhaustive search over the domain for the action selection of LGP-UCB, MAXMINLCB, and other presented algorithms. For the numerical experiments presented in this paper, we do not consider this as a practical limitation. Due to our efficient implementation in JAX, this optimization step can be carried out in parallel and seamlessly support accelerator devices such as GPUs and TPUs.

E.1.3 Logistic Bandits

Hyperparameters. We set $\delta = 0.1$ for all algorithms. For the GP-UCB and LGP-UCB algorithms, we use $\beta = 1$, 0.25 for the noise variance, and the Radial Basis Function (RBF) kernel. We choose the variance and length scale parameters from $[0.1, 1.0]$ to optimize their performance separately. For LGP-UCB, we tuned λ, the $L2$ penalty coefficient in Proposition 1, on the grid $[0.0, 0.1, 1.0, 5.0]$ and B on $[1.0, 2.0, 3.0]$. The hyperparameter selections were done for each utility function and algorithm separately.

E.1.4 Preference Feedback Bandits

Hyperparameters. We tune the same parameters of LGP-UCB for the preference feedback bandit problem on the following grid: $\lambda \in [0, 0.1, 1]$, $B \in [1, 2, 3]$, and $[0.1, 1]$ for the kernel variance and length scale.

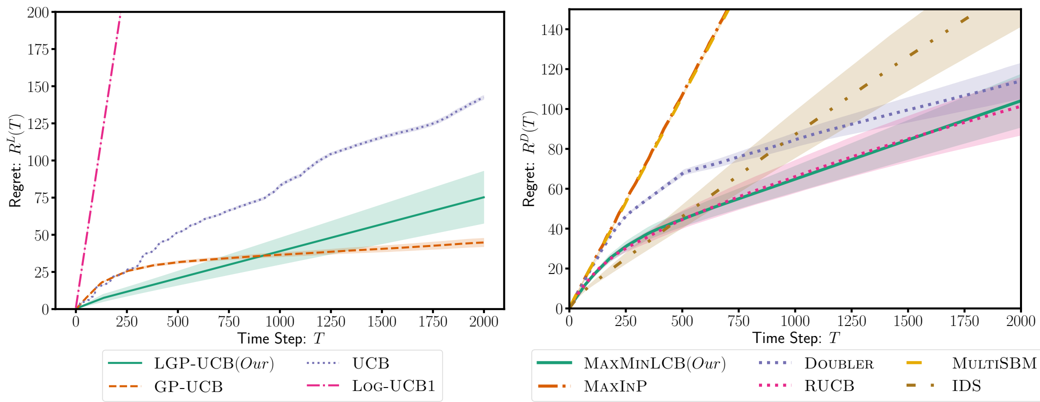
Comparison algorithms. Algorithm 5, Algorithm 6, and Algorithm 7 described the algorithms used for comparison in Section 6.2. MAXINP and IDS are defined in Algorithm 3 and Algorithm 4, respectively, in Appendix D.3 alongside with their theoretical analysis. We note that DOUBLER includes an internal for-loop, therefore, we adjusted the time-horizon T such that it observes the same number of feedback y_t as the other algorithms for a fair comparison.

E.2 Additional Experiments

In this section, we provide Table 2 that details the performance of the algorithm on the Logistic Dueling problem complementing the results in Section 6.1.

E.3 Computational Resources and Costs

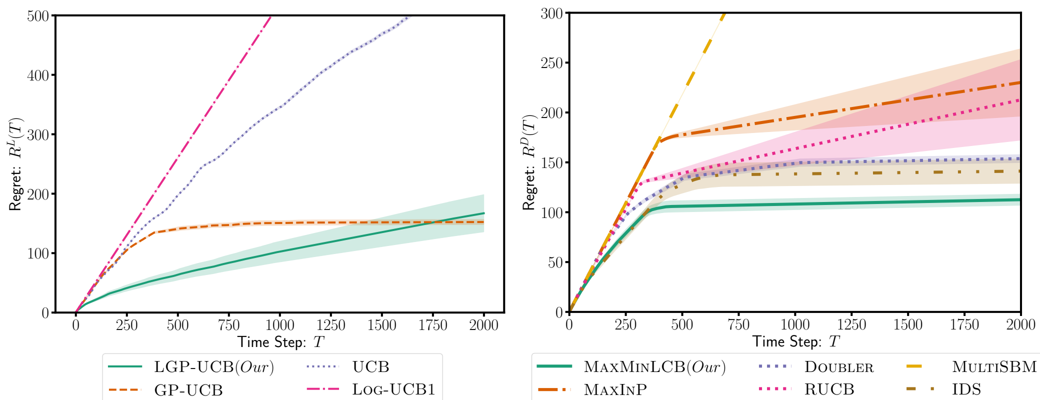
We ran our experiments on a shared cluster equipped with various NVIDIA GPUs and AMD EPYC CPUs. Our default configuration for all experiments was a single GPU with 24 GB of memory, 16



(a) Logistic feedback

(b) Preference feedback

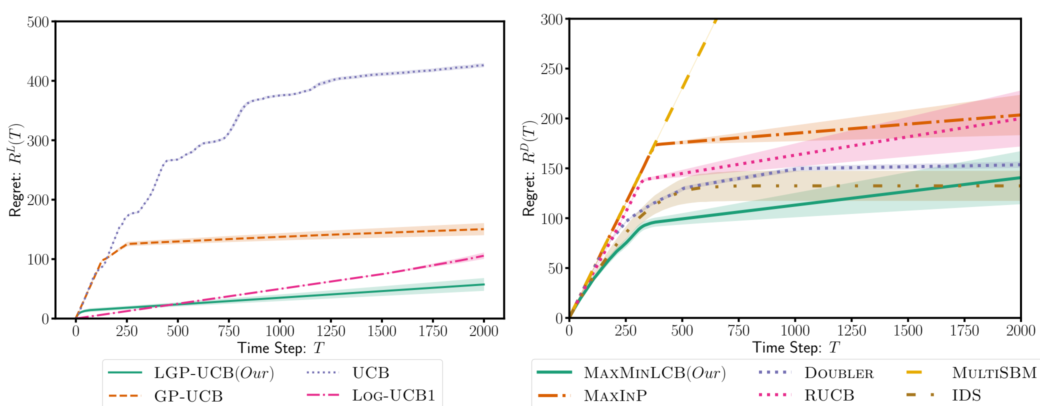
Figure 3: Regret with Branin utility function with logistic and preference feedback for horizon $T = 2000$.



(a) Logistic feedback

(b) Preference feedback

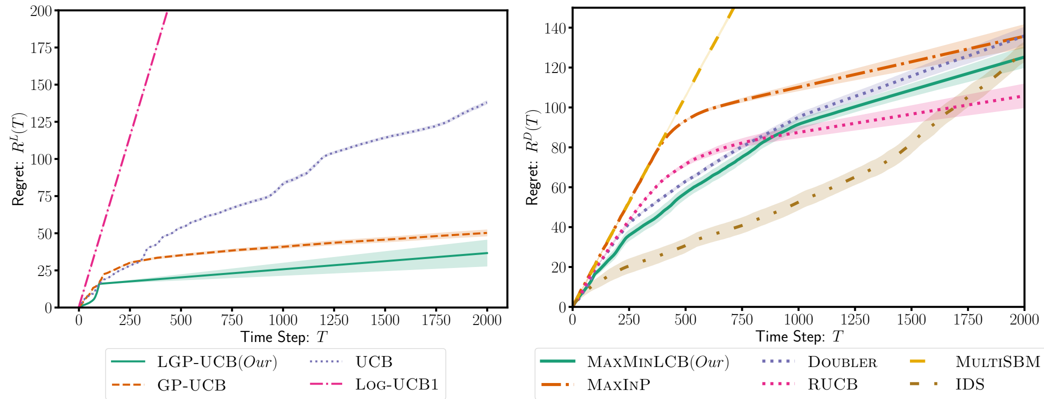
Figure 4: Regret with Eggholder utility function with logistic and preference feedback for horizon $T = 2000$.



(a) Logistic feedback

(b) Preference feedback

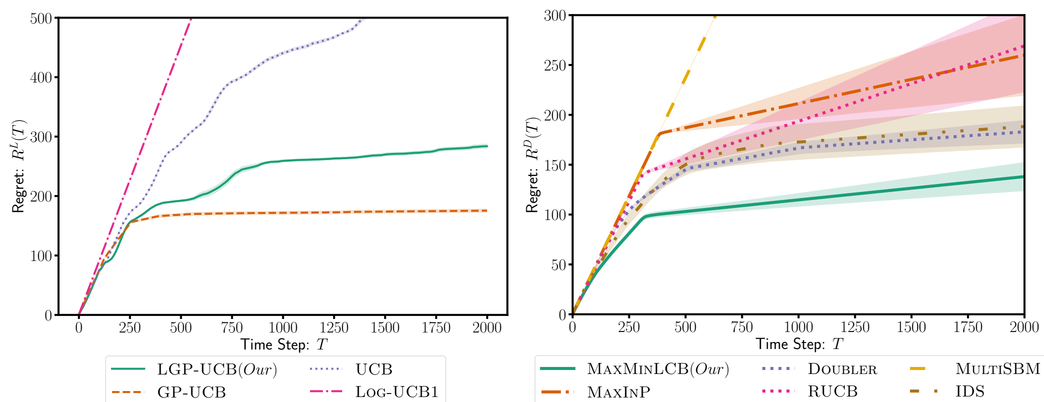
Figure 5: Regret with Hoelder utility function with logistic and preference feedback for horizon $T = 2000$.



(a) Logistic feedback

(b) Preference feedback

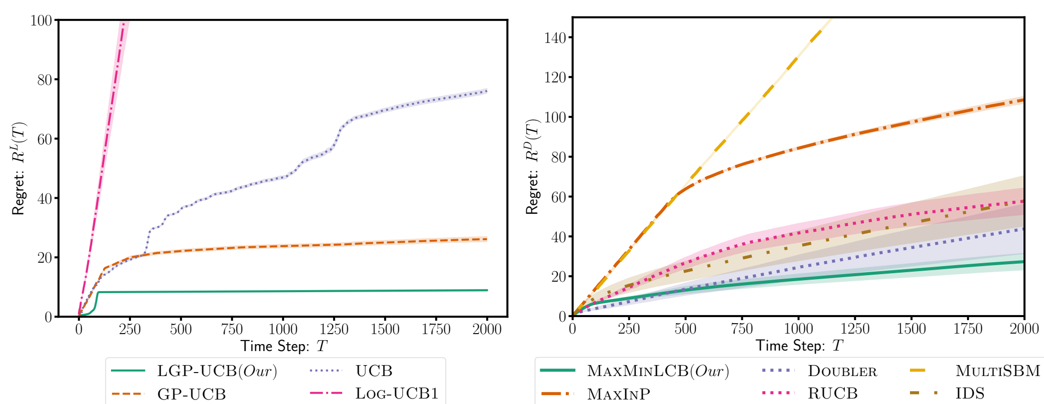
Figure 6: Regret with Matyas utility function with logistic and preference feedback for horizon $T = 2000$.



(a) Logistic feedback

(b) Preference feedback

Figure 7: Regret with Michalewicz utility function with logistic and preference feedback for horizon $T = 2000$.



(a) Logistic feedback

(b) Preference feedback

Figure 8: Regret with Rosenbrock utility function with logistic and preference feedback for horizon $T = 2000$.

Algorithm 7 RUCB [Zoghi et al., 2014a]

Input $(\beta^{\mathrm{D}}_t)_{t\geq 1}$.
for $t \geq 1$ **do**
  Select $\boldsymbol{x}'_t$ uniformly randomly from $\mathcal{M}_t$
  Select $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{M}_t} s(h_t(\boldsymbol{x}, \boldsymbol{x}'_t)) + \beta^{\mathrm{D}}_t \sigma^{\mathrm{D}}_t(\boldsymbol{x}, \boldsymbol{x}'_t)$
  Observe y_t and append history.
  Update h_{t+1} and $\sigma^{\mathrm{D}}_{t+1}$ and the set of plausible maximizers

$$\mathcal{M}_{t+1} = \{\boldsymbol{x} \in \mathcal{X} \,|\, \forall \boldsymbol{x}' \in \mathcal{X}:\ s(h_{t+1}(\boldsymbol{x}, \boldsymbol{x}')) + \beta^{\mathrm{D}}_{t+1}\sigma^{\mathrm{D}}_{t+1}(\boldsymbol{x}, \boldsymbol{x}') > 1/2\}.$$

end for

Table 2: Benchmarking R^{L}_T for a variety of test utility functions, $T = 2000$.

f	LGP-UCB	GP-UCB	UCB	LOG-UCB1
Ackley	$\mathbf{23.97} \pm 1.54$	96.35 ± 1.27	479.63 ± 3.42	1810.30 ± 0.00
Branin	75.23 ± 17.51	$\mathbf{44.81} \pm 2.81$	142.37 ± 1.33	1810.30 ± 0.00
Eggholder	167.11 ± 31.26	$\mathbf{152.34} \pm 4.28$	559.56 ± 4.15	1041.00 ± 0.00
Hoelder	$\mathbf{57.35} \pm 10.23$	150.41 ± 9.64	426.28 ± 2.94	105.64 ± 4.88
Matyas	$\mathbf{36.64} \pm 8.77$	50.21 ± 2.07	137.98 ± 1.21	920.48 ± 0.57
Michalewicz	283.85 ± 3.62	$\mathbf{175.46} \pm 2.86$	566.36 ± 3.75	1810.30 ± 0.00
Rosenbrock	$\mathbf{8.92} \pm 0.33$	26.14 ± 0.87	76.13 ± 0.84	897.04 ± 120.68

CPU cores, and 16 GB of RAM. Each experiment of the 11 configurations reported in Section 6.2 ran for about 12 hours and the experiment reported in Section 6.1 ran for 5 hours. The total computational cost to reproduce our results is around 140 hours of the default configuration. Our total computational costs including the failed experiments are estimated to be 2-3 times more.