Bandits with Preference Feedback: A Stackelberg Game Perspective

Barna Pásztor[⋆,1,2] **Parnian Kassraie**[⋆,1] **Andreas Krause**[1,2]
[1]ETH Zurich [2]ETH AI Center
{bpasztor, pkassraie, krausea}@ethz.ch

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 fine-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 algorithms 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 (a.k.a Leader-Follower) game. We choose the Lower Confidence Bound (LCB) of f as the objective of this game which the Leader aims

[⋆]Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0383

to 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 on preference-based confidence sequences 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 common baselines from the literature.

2 Related Work

Learning with indirect feedback was first studied in the supervised preference learning setting [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, Komiyama et al., 2015, Wu and Liu, 2016, 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 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 well understood and is even extended to reinforcement learning with preference feedback on trajectories [Saha et al., 2023, Zhan et al., 2023, Zhu et al., 2023, Ji et al., 2023, Munos 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 a classification. 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] which uses likelihood ratio 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 $\boldsymbol{x}'_t$. 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 a 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 a 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 reflects the preference via

$$\mathbb{P}(\boldsymbol{x}_t \succ \boldsymbol{x}'_t) \coloneqq s\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)\right) \tag{2}$$

where $s : \mathbb{R} \to [0, 1]$ is the sigmoid function[1], i.e. $s(a) = (1 + e^{-a})^{-1}$. Referred to as the Bradley-Terry model [Bradley and Terry, 1952], 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 the utility f. In particular, Saha [2021] shows that the dueling regret of (1) 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 $k \in \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with a bounded RKHS norm $\|f\|_k \leq B$. 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.

While our theoretical framework targets euclidean domains, our methodology may be used on general domains of text or images, given vector embeddings obtained via unsupervised learning. Solving a bandit problem on top of embeddings from a pretrained language model is common practice in

[1] May be generalized to differentiable monotonically increasing functions satisfying $s(x) + s(-x) = 1$.

https://doi.org/10.52202/079017-0383

further fine-tuning of such models [e.g., Nguyen et al., 2024, Mehta et al., 2023a], and is further demonstrated in our Yelp experiment (c.f Section 6.3). Lastly, we highlight that our results may be smoothly extended to contextual bandits with stochastic context, by simply modifying the signature of the kernel function to $k'(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{z}) : \mathcal{X} \times \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, where $\boldsymbol{z} \in \mathcal{Z}$ denotes the context. This setting accommodates applications in active learning for fine-tuning of large language models, where the context is the prompt and the two actions are two alternative responses.

4 Kernelized Confidence Sequences with Direct Logistic Feedback

As a warm-up, we consider a hypothetical scenario where $\boldsymbol{x}'_t = \boldsymbol{x}_{\text{null}}$ for all $t \geq 1$ such that $f(\boldsymbol{x}_{\text{null}}) = 0$. 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 $s(f(\boldsymbol{x}_t))$. This example reduces our problem to logistic bandits which has been rigorously 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 on 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.

The feedback y_t is a Bernoulli random variable, and 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

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

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 to give theoretical guarantees. 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 direct 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* (3) *has a unique minimizer f_t, which takes the form $f_t(\cdot) = \sum_{\tau=1}^{t} \alpha_\tau k(\cdot, \boldsymbol{x}_\tau)$ where $(\alpha_1, \dots \alpha_t) =: \boldsymbol{\alpha}_t \in \mathbb{R}^t$ is the minimizer of the strictly convex loss*

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

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

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 $\kappa = \sup_{a \leq B} 1/\dot{s}(a)$, with $\dot{s}(a) = s(a)(1 - s(a))$ denoting the derivative of the sigmoid function, 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 $\|f\|_k \leq B$. Let $0 < \delta < 1$ and set*

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

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

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

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 as 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 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., $\nabla \mathcal{L}(f_t; H_t) : \mathcal{H} \to \mathcal{H} = \mathbf{0}$. 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 A, and builds on the results of Faury et al. [2020] and Whitehouse et al. [2023].

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 LGP-UCB, the kernelized Logistic GP-UCB algorithm. Presented in Algorithm 2, it extends the optimistic algorithm 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 $C_L \coloneqq \sqrt{8/\log(1 + (\lambda\kappa)^{-1})}$.

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 observed response 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 problem 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. We start by presenting a solution to estimate f and obtain valid confidence sets under preference feedback. Using these confidence sets we then design the MaxMinLCB algorithm which chooses action pairs that are not only favorable, i.e., yield high utility, but are also informative for improving the utility confidence sets.

5.1 Preference-based Confidence Sets

We consider the probabilistic model of y_t as stated in Section 3, and write the corresponding regularized negative loglikelihood loss as

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

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 $\mathcal{L}_k^{\mathrm{D}}$ is equivalent to $\mathcal{L}_{k^{\mathrm{D}}}^{\mathrm{L}}$, the standard logistic loss (3) invoked with a specific kernel function k^{D}. This will

https://doi.org/10.52202/079017-0383

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

$$k^{\mathrm{D}}\big((\boldsymbol{x}_1, \boldsymbol{x}_1'), (\boldsymbol{x}_2, \boldsymbol{x}_2')\big) \coloneqq k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}_1', \boldsymbol{x}_2') - k(\boldsymbol{x}_1, \boldsymbol{x}_2') - k(\boldsymbol{x}_1', \boldsymbol{x}_2)$$

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. *Let $f : \mathcal{X} \to \mathbb{R}$. Consider a kernel k and the sequence of its eigenfunctions $(\phi_i)_{i=1}^{\infty}$. Assume the eigenfunctions are zero-mean, i.e. $\int_{\boldsymbol{x} \in \mathcal{X}} \phi_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0$. 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 C.1. The assumption on eigenfunctions in Proposition 4 is primarily made to simplify the equivalence class. In particular, the relative preference 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, 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 $h_t \coloneqq \arg\min \mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}(h; H_t)$ 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 $\|f\|_k \leq B$. 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,*

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

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 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 C.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). Corollary 5 holds almost immediately by invoking Theorem 2 with the dueling kernel k^{D} and applying Proposition 4. A proof is provided in Appendix C.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 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 when calculating the likelihood. Therefore, the resulting algorithms require more samples to learn an accurate reward estimate. In a concurrent work, Xu et al. [2024] also consider the loss function of Equation (6) and present likelihood-ratio confidence sets. The width of the sets at time T, scales 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.

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

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

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

5.2 Action Selection Strategy

Let $\boldsymbol{x}'(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}'\in\mathcal{M}_t} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}')$ denote a *response* function. We propose MAXMINLCB in Algorithm 1 for the preference feedback bandit problem that selects $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$ jointly as

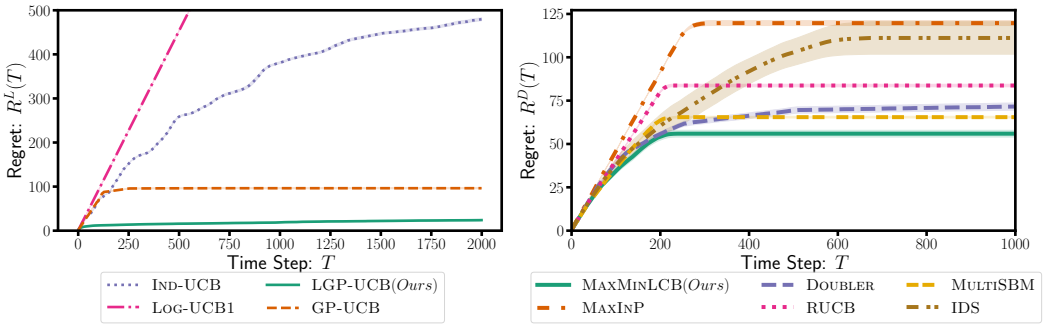
$$\begin{aligned} \boldsymbol{x}_t &= \arg\max_{\boldsymbol{x}\in\mathcal{M}_t} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}'(\boldsymbol{x})) \quad \text{(Leader)} \\ \boldsymbol{x}'_t &= \boldsymbol{x}'(\boldsymbol{x}_t) \qquad\qquad\qquad\quad \text{(Follower)} \end{aligned} \tag{7}$$

where the lower-confidence bound $\mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') = s(h_t(\boldsymbol{x}, \boldsymbol{x}')) - \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}')$ presents a pessimistic estimate of h and $\mathcal{M}_t = \{\boldsymbol{x} \in \mathcal{X} | \, \forall \boldsymbol{x}' \in \mathcal{X} : \, s(h_t(\boldsymbol{x}, \boldsymbol{x}')) + \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}') \geq 0.5\}$ 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 sequentially [Stackelberg, 1952]. First, the Leader selects $\boldsymbol{x}_t$, then the Follower selects $\boldsymbol{x}'_t$ depending on the choice of $\boldsymbol{x}_t$. Importantly, due to the sequential nature of action selection, $\boldsymbol{x}_t$ 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.

https://doi.org/10.52202/079017-0383



(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.

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 C.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 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 only 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. For all experiments, we set the horizon $T = 2000$ and evaluate all algorithms on a uniform mesh over the input domain of size 100. Additionally, we conducted experiments on the Yelp restaurant review dataset to demonstrate the applicability of MAXMINLCB on real-world data and its scaling to larger domains. All experiments are run across 20 random seeds and reported values are averaged over the seeds, together with standard error. The environments and algorithms are implemented[2] end-to-end in JAX [Bradbury et al., 2018].

6.1 Benchmarking Confidence Sets

Performance of MAXMINLCB relies on validity and tightness of the LCB. We evaluate the quality of our kernelized confidence bounds, using the potentially simpler task of bandit optimization given logistic feedback. To this end, we fix the acquisition function for the logistic bandit algorithms to the Upper Confidence Bound (UCB) function, and benchmark different methods for calculating the confidence bound. We refer to the algorithm instantiated with the confidence sets of Theorem 2 as LGP-UCB (c.f. Algorithm 2). The IND-UCB approach assumes that actions are uncorrelated, and

[2]The code is made available at github.com/lasgroup/MaxMinLCB.

Table 1: Benchmarking R_T^{D} for a variety of test utility functions, $T = 2000$. The top 3 rows show results for smoother functions without steep gradients and local optima while the bottom 5 rows show the results for more challenging problems.

f	MAXMINLCB	DOUBLER	MULTISBM	MAXINP	RUCB	IDS
Branin	$\mathbf{104} \pm 13$	114 ± 9	$\mathbf{89} \pm 13$	340 ± 2	$\mathbf{101} \pm 14$	163 ± 22
Matyas	125 ± 5	136 ± 4	$\mathbf{106} \pm 7$	136 ± 6	$\mathbf{106} \pm 6$	128 ± 5
Rosenbrock	$\mathbf{27} \pm 4$	44 ± 12	$\mathbf{25} \pm 5$	109 ± 2	58 ± 7	58 ± 13
Ackley	$\mathbf{56} \pm 2$	72 ± 2	65 ± 0.5	120 ± 1	84 ± 0.7	111 ± 9
Eggholder	$\mathbf{113} \pm 6$	154 ± 4	134 ± 3	230 ± 34	213 ± 40	141 ± 12
Hoelder	$\mathbf{141} \pm 26$	$\mathbf{154} \pm 3$	$\mathbf{136} \pm 15$	204 ± 20	200 ± 28	$\mathbf{132} \pm 15$
Michalewicz	$\mathbf{138} \pm 14$	183 ± 11	$\mathbf{155} \pm 10$	260 ± 40	269 ± 46	188 ± 21
Yelp	$\mathbf{175} \pm 22$	263 ± 28	$\mathbf{199} \pm 25$	409 ± 15	214 ± 22	255 ± 22

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 baselines. 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, IND-UCB converges at a slower rate than LGP-UCB and GP-UCB due to ignoring the correlation between arms while LOG-UCB1's regret grows linearly as the Ackley function is misspecified under the assumption of linearity. We defer the results on the rest of the utility functions to Table 2 in Appendix E 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 all algorithms with the same confidence sets, and use our improved preferred-based bound of Corollary 5. Therefore, our implementation differs from the corresponding references, while we refer to the algorithms by their original name. We consider the following baselines. DOUBLER and MULTISBM [Ailon et al., 2014] 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] chooses $\boldsymbol{x}'_t$ similarly, 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 $\mathcal{M}_t$, and at 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 exception of MAXINP, choose one of the actions independently and use it as a reference point when selecting the other one. Figure 3 illustrates the differences in action selection between UCB, maximum information, and MAXMINLCB approaches. We note that POP-BO [Xu et al., 2024] and MULTISBM only differ in the estimation of the confidence set. Since we deploy the same confidence set for all acquisition functions, the two algorithms are equivalent and we use MULTISBM in our results, however, comparisons hold for POP-BO as well.

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 initial stages of 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 consistently outperforms the baselines across the analyzed functions and achieves a low standard error, supporting its efficiency in balancing exploration and exploitation in the preference feedback setting. Only MULTISBM shows comparable performance to MAXMINLCB and even outperforms it on the Matyas function which is a relatively smooth function posing a simple optimization problem. However, MAXMINLCB achieves lower regret on the Ackley and Eggholder functions which obtain many local optima and steeper

https://doi.org/10.52202/079017-0383

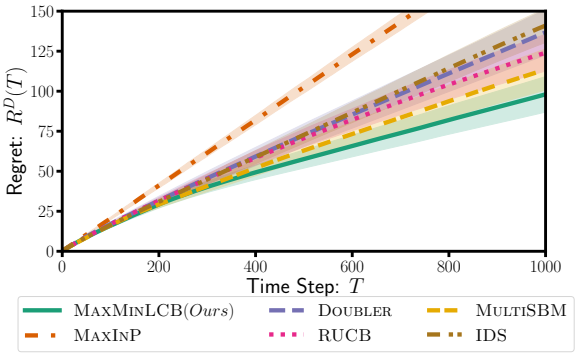


Figure 2: LGP-UCB is more sample-efficient when making restaurant recommendations based on Yelp open dataset with preference feedback. All baselines use the confidence sets of Cor. 5.

gradients showing that MaxMinLCB is preferable for challenging optimization problems. Other acquisition functions work well only in certain cases, e.g., IDS achieves the smallest regret for optimizing Matyas, while RUCB excels on the Branin function. This indicates the challenges each utility 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.

6.3 Real-world Experiment

To demonstrate the scalability and applicability of MaxMinLCB, we conduct an experiment on the Yelp dataset of restaurant reviews, which consists of 275 restaurants and 20 users after the pre-processing stage. For each user, we want to find the restaurants that best fit their preference, via making sequential recommendations and receiving comparative feedback. We define the action space $\mathcal{X}$ by assigning to each restaurant their respective 32-dimensional embedding of their reviews, i.e., $\mathcal{X} \subseteq \mathbb{R}^{32}$. The dataset provides utility values for users in the form of ratings on the scale of 1 to 5, however, not all users rated every restaurant.

We estimate missing ratings using collaborative filtering [Schafer et al., 2007]. Further details on the data processing are deferred to Appendix D.1. Figure 2 shows that the results of this larger problem align with previous conclusions. MaxMinLCB remains the best-performing algorithm with MultiSBM following second. The cumulative regret is also reported in Table 1. Note that neither of the algorithms is tuned or modified for this experiment. These results are only intended to demonstrate that 1) the computations easily scale, and 2) the kernelized approach is still applicable in a text-based domain, by using high-quality vector embeddings.

7 Conclusion

We addressed the problem of bandit optimization with preference feedback over large domains and complex targets. We proposed 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 be 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.

Acknowledgments and Disclosure of Funding

We thank Scott Sussex for his thorough feedback. This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and Innovation Program Grant agreement No. 815943. Barna Pásztor was supported by an ETH AI Center doctoral fellowship, and Parnian Kassraie by a Google PhD Fellowship.

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, et al. JAX: composable transformations of Python+ NumPy programs, 2018. URL http://github.com/google/jax.

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.

 https://doi.org/10.52202/079017-0383

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.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015.

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.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint*, 2023.

Tung Nguyen, Qiuyi Zhang, Bangding Yang, Chansoo Lee, Jorg Bornschein, Yingjie Miao, Sagi Perel, Yutian Chen, and Xingyou Song. Predicting from strings: Language model embeddings for bayesian optimization. *arXiv preprint*, 2024.

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.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

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. *Theory of the market economy*. Oxford University Press, 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.

 https://doi.org/10.52202/079017-0383

Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

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 Proofs for Bandits with Logistic Feedback

We have written the equations in the main text in terms of kernel matrices and function evaluations, for easier readability. For the purpose of the proof however, 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 B-bounded 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 $S_t : \mathcal{H}_k \to \mathcal{H}_k$, 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 acting 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]. Lastly, in the appendix, we refer to the true unknown utility function as $f^\star(\boldsymbol{x}) = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta}^\star$. In the main text, the true utility is simply referred to as f.

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

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

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

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

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

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

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

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

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 B-bounded alternative to $\boldsymbol{\theta}_t$, who also 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.

https://doi.org/10.52202/079017-0383

Lemma 7 (Corollary 1 Whitehouse et al. [2023])**.** *Suppose the sequence $(\boldsymbol{x}_t)_{t\geq 1}$ is $(\mathcal{F}_t)_{t\geq 1}$-adapted, where $\mathcal{F}_t \coloneqq \sigma(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \varepsilon_1, \ldots, \varepsilon_{t-1})$ and ε_t are 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 $t \geq 0$,*

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

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 $0 < \delta < 1$. Under the assumptions of Theorem 2*

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

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

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

$$\nabla \mathcal{L}(\boldsymbol{\theta}; H_t) = \sum_{\tau \leq t} y_\tau \boldsymbol{\phi}(\boldsymbol{x}_\tau) - g_t(\boldsymbol{\theta}).$$

Since $\boldsymbol{\theta}_t$ is a minimizer of $\mathcal{L}_t$, it holds that $\boldsymbol{g}_t(\boldsymbol{\theta}_t) = \sum_{\tau \leq t} y_\tau \boldsymbol{\phi}(\boldsymbol{x}_\tau)$. This allows us to write,

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

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

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

since $\gamma_t(\rho) = \sup_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t} \frac{1}{2} \log\det(\boldsymbol{I}_t + \rho^{-2} K_t))$. 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*

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

Then,

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

Proof of Lemma 9. We check if $\boldsymbol{\theta}^\star \in \Theta_t(\delta)$ by bounding the following norm

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

□

Lastly, we prove a formal extension of Theorem 2.

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

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

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

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

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

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

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,

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

https://doi.org/10.52202/079017-0383

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

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

B Helper Lemmas for Appendix A

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

$$\alpha(\boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \int_0^1 \dot{s}(\nu\boldsymbol{\theta}_2^\top \boldsymbol{\phi}(\boldsymbol{x}) + (1-\nu)\boldsymbol{\theta}_1^\top \boldsymbol{\phi}(\boldsymbol{x}))\mathrm{d}\nu \tag{10}$$

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

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

We change the variable of integration to $\nu = (z - z_1)/(z_2 - z_1)$, 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 norm B-bounded $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \ell_2(\mathbb{N})$*

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

Proof of Lemma 12. We prove the lemma through an auxiliary operator $G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ operating on $\mathcal{H}_k$

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

where α is defined in (10).

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 $\kappa = \sup_{z\leq B} \frac{1}{\dot{s}(z)}$. Therefore, $\kappa^{-1} \leq \dot{s}(z)$ for all $z < B$, 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,

$$G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq \lambda \boldsymbol{I}_{\mathcal{H}} + \sum_{\tau\leq t} \kappa^{-1}\boldsymbol{\phi}(\boldsymbol{x}_\tau)\boldsymbol{\phi}^\top(\boldsymbol{x}_\tau) = \kappa^{-1}V_t. \tag{11}$$

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

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

Therefore,

$$\begin{aligned} \|g_t(\boldsymbol{\theta}_2) - g_t(\boldsymbol{\theta}_1)\|_{G_t^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} &= \left[g_t(\boldsymbol{\theta}_2) - g_t(\boldsymbol{\theta}_1) \right]^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\ &= (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^\top G_t \, (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \\ &= \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}. \end{aligned} \tag{12}$$

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,

$$\begin{aligned} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{V_t} &\overset{(11)}{\le} \sqrt{\kappa} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{G_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\ &\overset{(12)}{\le} \sqrt{\kappa} \|g_t(\boldsymbol{\theta}_1) - g_t(\boldsymbol{\theta}_2)\|_{G_t^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\ &\overset{(11)}{\le} \kappa \|g_t(\boldsymbol{\theta}_1) - g_t(\boldsymbol{\theta}_2)\|_{V_t^{-1}} \end{aligned}$$

concluding the proof. □

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. *Let σ_t be as defined in (4). Then $\sqrt{\lambda\kappa}\|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}} = \sigma_t(\boldsymbol{x})$, for any $\boldsymbol{x} \in \mathcal{X}$.*

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

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

which gives

$$\Phi_t^\top \left(\Phi_t \Phi_t^\top + \lambda\kappa \boldsymbol{I}_t \right)^{-1} = \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I}_{\mathcal{H}} \right)^{-1} \Phi_t^\top. \tag{13}$$

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

$$\boldsymbol{k}_t(\boldsymbol{x}) = \Phi_t \boldsymbol{\phi}(\boldsymbol{x}) \tag{14}$$

which allow us to write

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

and obtain

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{x}) &= \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I}_{\mathcal{H}} \right)^{-1} \Phi_t^\top \boldsymbol{k}_t(\boldsymbol{x}) + \lambda\kappa \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I} \right)^{-1} \boldsymbol{\phi}(\boldsymbol{x}) \\ &\overset{(13)}{=} \Phi_t^\top \left(\Phi_t \Phi_t^\top + \lambda\kappa \boldsymbol{I}_t \right)^{-1} \boldsymbol{k}_t(\boldsymbol{x}) + \lambda\kappa \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I} \right)^{-1} \boldsymbol{\phi}(\boldsymbol{x}). \end{aligned}$$

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

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}) &= \boldsymbol{\phi}^\top(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \\ &= \left(\Phi_t^\top \left(\Phi_t \Phi_t^\top + \lambda\kappa \boldsymbol{I}_t \right)^{-1} \boldsymbol{k}_t(\boldsymbol{x}) + \lambda\kappa \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I}_{\mathcal{H}} \right)^{-1} \boldsymbol{\phi}(\boldsymbol{x}) \right)^\top \boldsymbol{\phi}(\boldsymbol{x}) \\ &= \boldsymbol{k}_t^\top(\boldsymbol{x}) \left(\Phi_t \Phi_t^\top + \lambda\kappa \boldsymbol{I}_t \right)^{-1} \Phi_t \boldsymbol{\phi}(\boldsymbol{x}) + \lambda\kappa \boldsymbol{\phi}^\top(\boldsymbol{x}) \left(\Phi_t^\top \Phi_t + \lambda\kappa \boldsymbol{I}_{\mathcal{H}} \right)^{-1} \boldsymbol{\phi}(\boldsymbol{x}) \\ &\overset{(14)}{=} \boldsymbol{k}_t^\top(\boldsymbol{x}) \left(K_t + \lambda\kappa \boldsymbol{I}_t \right)^{-1} \boldsymbol{k}_t(\boldsymbol{x}) + \lambda\kappa \boldsymbol{\phi}^\top(\boldsymbol{x}) V_t^{-1} \boldsymbol{\phi}(\boldsymbol{x}) \end{aligned}$$

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]. Re-ordering the equation above we obtain $\sigma_t^2(\boldsymbol{x}) = \lambda\kappa \|\boldsymbol{\phi}(\boldsymbol{x})\|_{V_t^{-1}}^2$, concluding the proof. □

https://doi.org/10.52202/079017-0383

Lemma 14 (Controlling posterior variance with information gain)**.** *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],

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

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

C 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.

C.1 Equivalence of Preference-based and Logistic Losses

We start by establishing the equivalence between the logistic loss (3) and dueling loss (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,

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

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 $\|f\|_k^2 = \sum_{i\geq 1}\beta_i^2 \leq \infty$. And set the difference function to $h(\boldsymbol{x},\boldsymbol{x}') = \sum_{i\geq 1}\beta_i\psi_i(\boldsymbol{z})$. 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 $\|f\|_k$. The other direction of the statement is proved the same way. □

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 $\mathcal{L}^{\mathrm{L}}_{k^{\mathrm{D}}}(\cdot; H_t)$. 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. □

C.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 in the rest of this section, we define $\omega_t(\boldsymbol{x},\boldsymbol{x}') \coloneqq \beta_t^{\mathrm{D}}(\delta)\sigma_t^D(\boldsymbol{x},\boldsymbol{x}')$ and

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

Note that $\omega_t(\boldsymbol{x},\boldsymbol{x}') = \omega_t(\boldsymbol{x}',\boldsymbol{x})$ by the symmetry of the dueling kernel k^{D}. Furthermore, recall the notation $h(\boldsymbol{x},\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x})$.

Proof of Theorem 6. **Step 1:** First, we connect the term of $\boldsymbol{x}_t'$ in the dueling regret defined in Equation (1) to that of $\boldsymbol{x}_t$. Note that both $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t'))$ and $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))$ are greater than 0.5 due to the optimality of $\boldsymbol{x}^\star$ and the sigmoid function s is concave on the interval $[0.5,\infty)$. Using the definition of concavity, we get

$$\begin{aligned}
s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t')) &\leq s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + \dot{s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_t')) \\
&= s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t))s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*))(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_t')) \\
&\leq \left(1 + \frac{h(\boldsymbol{x}_t,\boldsymbol{x}_t')}{2}\right) s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) \qquad (15)
\end{aligned}$$

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

Using Equation (15), we can upper bound the dueling regret in Equation (1) as

$$\begin{aligned}
2r_t^D &= s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t')) - 1 \\
&\leq s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + \left(1 + \frac{h(\boldsymbol{x}_t,\boldsymbol{x}_t')}{2}\right) s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 1 \\
&\leq 2s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 1 + \frac{h(\boldsymbol{x}_t,\boldsymbol{x}_t')}{2} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) \qquad (16)
\end{aligned}$$

https://doi.org/10.52202/079017-0383

Step 2: Next, we show that the single-step regret is bounded by $\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$.
First, consider the term $s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t))$

$$\begin{aligned} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) - 0.5 &\leq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) - 0.5 && \text{Corollary 5} \\ &\leq 0.5 - s(h_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)) + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) && \text{Sigmoid equality} \\ &\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \end{aligned} \tag{17}$$

In the last inequality, we used that $\boldsymbol{x}_t \in \mathcal{M}_t$ implying that $0.5 - s(h_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)) \leq \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$. Combining Equation (16) and Equation (17) implies that

$$2r_t^D \leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) \tag{18}$$

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,

$$\begin{aligned} 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^*) &= \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\ &\leq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\ &\leq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) = 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned} \tag{19}$$

where we used the assumption in the first inequality and 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 the following connection between LCB_t and UCB_t.

$$\begin{aligned} \mathrm{LCB}_t(\boldsymbol{x}, \boldsymbol{x}') &= s(h_t(\boldsymbol{x}, \boldsymbol{x}')) - \omega_t(\boldsymbol{x}, \boldsymbol{x}') \\ &= 1 - s(h_t(\boldsymbol{x}', \boldsymbol{x})) - \omega_t(\boldsymbol{x}', \boldsymbol{x}) \\ &= 1 - \mathrm{UCB}_t(\boldsymbol{x}', \boldsymbol{x}) \end{aligned} \tag{20}$$

where we used the equality $s(z) = 1 - s(-z)$ in the second equality. Using Equation (20), the assumption of Case 2 can be rewritten as $\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \geq \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)$, implying that

$$\mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \geq \mathrm{LCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t). \tag{21}$$

Similarly, $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}^\star)$ implies

$$\mathrm{UCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \geq \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) \tag{22}$$

Combining Equation (21) and Equation (22), we get

$$\begin{aligned} 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}^*) &= \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) \\ &\leq \mathrm{UCB}_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) \\ &\leq \mathrm{UCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) - \mathrm{LCB}_t(\boldsymbol{x}'_t, \boldsymbol{x}_t) = 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned} \tag{23}$$

Combining Equation (19) and Equation (23), we can rewrite the first term in Equation (18) to get

$$2r_t^D \leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) + \frac{h(\boldsymbol{x}_t, \boldsymbol{x}'_t)}{2} s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)). \tag{24}$$

It remains to bound the second term of Equation (24). By the Mean-Value Theorem, $\exists z \in [0, h(\boldsymbol{x}_t, \boldsymbol{x}'_t)]$ such that

$$\dot{s}(z)(h(\boldsymbol{x}_t, \boldsymbol{x}'_t) - 0) = s(h(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - f(0)$$

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

$$h(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq \kappa(s(h(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5) \tag{25}$$

Next, we consider the term $s(h(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5$ in Equation (25). Note that $\boldsymbol{x}_t, \boldsymbol{x}'_t \in \mathcal{M}_t$ implies that

$$\begin{aligned} &\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \geq 0.5 \\ &s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) \geq 0.5 - \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned} \tag{26}$$

Additionally note that $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}_t) = 0.5$ for all $\boldsymbol{x}_t$, therefore, by the definition of $\boldsymbol{x}'_t$, $\mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 0.5$. It implies that

$$\begin{aligned} \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) &\leq 0.5 \\ s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) &\leq 0.5 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t). \end{aligned} \tag{27}$$

From Equation (26) and Equation (27), it follows that

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

furthermore,

$$\begin{aligned} \mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - 0.5 &= s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \\ &\leq |s(h_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) - 0.5| + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \\ &\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \end{aligned} \tag{28}$$

and similarly

$$0.5 - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t). \tag{29}$$

From Equation (28) and Equation (29), it follows that

$$\begin{aligned} |s(f(\boldsymbol{x}_t) - f(\boldsymbol{x}'_t)) - 0.5| &\leq \max\{\mathrm{UCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) - 0.5, 0.5 - \mathrm{LCB}_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)\} \quad \text{Corollary 5} \\ &\leq 2\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t). \end{aligned} \tag{30}$$

Combining Equation (30) with Equation (25), it follows

$$h(\boldsymbol{x}_t, \boldsymbol{x}'_t) \leq 2\kappa\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t) \tag{31}$$

Using Equation (31) in Equation (24) and the fact that $s(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)) \leq 1$, we get

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

Therefore, for the cumulative dueling regret it holds

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

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

□

C.3 Extending Algorithms for Linear Dueling Bandits to Kernelized Setting

Maximum Informative Pair Algorithm. Proposed 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. 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 $C_2 = 4/\sqrt{\log(1 + 4(\lambda\kappa)^{-1})}$.

https://doi.org/10.52202/079017-0383

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

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

$$\boldsymbol{x}_t, \boldsymbol{x}'_t = \arg\max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{M}_t} \omega_t(\boldsymbol{x}, \boldsymbol{x}') \tag{32}$$

where action space is restricted to $\mathcal{M}_t$ and therefore,

$$\begin{aligned} s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) &\leq 1/2 + \omega_t(\boldsymbol{x}_t, \boldsymbol{x}^\star) \\ s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)) &\leq 1/2 + \omega_t(\boldsymbol{x}'_t, \boldsymbol{x}^\star) \end{aligned} \tag{33}$$

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 $1 - \delta$

$$\begin{aligned} 2r_t^{\mathrm{D}} &= s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + s(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}'_t)) - 1 \\ &\leq s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}_t)) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) + s(h_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t) - 1 && \text{(w.h.p.)} \\ &\leq 2\left(\omega_t(\boldsymbol{x}^\star, \boldsymbol{x}_t) + \omega_t(\boldsymbol{x}^\star, \boldsymbol{x}'_t)\right) && \text{Eq. (33)} \\ &\leq 4\omega_t(\boldsymbol{x}_t, \boldsymbol{x}'_t)) && \text{Eq. (32)} \end{aligned}$$

where for the first inequality we have invoked Corollary 5. Then for the regret satisfies

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

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 an algorithm which requires solving an optimization problem on $\mathcal{X} \times [0, 1]$ and 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 accommodate this.

Consider the sub-optimality 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 the actions. Then u_t shows an optimistic estimate of the highest obtainable reward at this step:

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

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

$$\hat{\Delta}_t(\boldsymbol{x}) \coloneqq u_t + h_t(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) \tag{34}$$

Algorithm 4 Preference-based IDS - Kernelized Logistic Variant

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

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

Update u_t and $\hat{\Delta}_t(\boldsymbol{x})$ acc. to (34).
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 $\alpha_t \sim \text{Bern}(p_t)$.
if $\alpha_t = 1$ **then** choose pair $(\boldsymbol{x}_t, \boldsymbol{x}_t') = (\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)})$ **else** choose $(\boldsymbol{x}_t, \boldsymbol{x}_t') = (\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(1)})$.
Observe y_t and append history.
Update h_{t+1} and σ_{t+1}^{D}.
end for

and show its uniform validity in Lemma 17. We can now propose the Kernelized Logistic IDS algorithm with preference 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 $\tilde{\beta}_t = \beta_t^{\text{D}}(\delta)/L$. Then Algorithm 4 satisfies the anytime cumulative dueling regret guarantee of*

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

Proof of Theorem 16. Our approach closely follows the proof of Kirschner and Krause [2021, Theorem 1]. Let $\mathcal{P}(\cdot)$ show the set of continuous probability distributions over a domain. Define the expected average gap for a policy $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$

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

and the expected information ratio as

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

Algorithm 4 draws actions via $\mu_t = (1-p_t)\delta_{(\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(1)})} + p_t\delta_{(\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)})}$, where $\delta_{(\boldsymbol{x},\boldsymbol{x}')}$ 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

$$R^{\text{D}}(T) \leq L\sqrt{\sum_{t=1}^{T} \Xi_t(\mu_t)\left(\gamma_T + \mathcal{O}(\log 1/\delta)\right)} + \mathcal{O}(L \log T/\delta) \tag{35}$$

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

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

https://doi.org/10.52202/079017-0383

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

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

Now consider the definition of u_t and let $\boldsymbol{z}_t$ denote the action for which u_t is achieved, 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

$$\begin{aligned}\Xi_t(\mu_t) &\le \min_{\boldsymbol{x}}\frac{\hat{\Delta}_t^2(\boldsymbol{x})}{\log\left(1+\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{x})^2\right)}\\ &\le \frac{\hat{\Delta}_t^2(\boldsymbol{z}_t)}{\log\left(1+(\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{z}_t)^2\right)}\\ &\le \frac{\bar{\beta}_t^2(\delta)\sigma_t^{\mathrm{D}}(\boldsymbol{z}_t\hat{\boldsymbol{x}}_t^\star)^2}{\log\left(1+(\lambda\kappa)^{-1}\sigma_t^D(\boldsymbol{x}_t^{(1)},\boldsymbol{z}_t)^2\right)}\\ &\le \frac{4\bar{\beta}_t^2(\delta)}{\log\left(1+4(\lambda\kappa)^{-1}\right)} && (36)\end{aligned}$$

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 (35) and (36) 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 $T\ge 1$. □

C.4 Helper Lemmas for Appendix C.3

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

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

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

$$h(\boldsymbol{x}_1, \boldsymbol{x}_3) = h(\boldsymbol{x}_1, \boldsymbol{x}_2) + h(\boldsymbol{x}_2, \boldsymbol{x}_3). \tag{37}$$

Therefore, from the definition of the estimated gap get

$$\begin{aligned}
\hat{\Delta}_t(\boldsymbol{x}) &= \max_{\boldsymbol{z} \in \mathcal{X}} h(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) + h_t(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) + \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) \\
&= \max_{\boldsymbol{z} \in \mathcal{X}} h(\boldsymbol{z}, \boldsymbol{x}) + \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) \\
&\ge h(\boldsymbol{x}, \boldsymbol{x}) + \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}_t^\star) \\
&= \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}, \boldsymbol{x}_t^\star).
\end{aligned} \tag{38}$$

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

$$\begin{aligned}
\Delta(\boldsymbol{x}) &= \max_{\boldsymbol{z} \in \mathcal{X}} h(\boldsymbol{z}, \boldsymbol{x}) \\
&= \max_{\boldsymbol{z} \in \mathcal{X}} h(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) + h(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) && \text{Eq. (37)} \\
&\overset{\text{w.h.p.}}{\le} \max_{\boldsymbol{z} \in \mathcal{X}} h_t^P(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) + h_t(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) + \bar{\beta}_t(\delta)\Big(\sigma_t^D(\boldsymbol{z}, \hat{\boldsymbol{x}}_t^\star) + \sigma_t^D(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x})\Big) && \text{Lem. 18} \\
&= u_t + h_t^P(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) + \bar{\beta}_t(\delta)\sigma_t^D(\hat{\boldsymbol{x}}_t^\star, \boldsymbol{x}) && \text{Def. } u_t \\
&= \hat{\Delta}_t(\boldsymbol{x}) + \bar{\beta}_t(\delta)\sigma_t^D(\boldsymbol{x}_t^\star, \boldsymbol{x}) && \text{Def. } \hat{\Delta}_t(\boldsymbol{x}) \\
&\le 2\hat{\Delta}_t(\boldsymbol{x}) && \text{Eq. (38)}
\end{aligned}$$

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* $0 < \delta < 1$

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

where

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

Proof of Lemma 18. This lemma is effectively a weaker parallel of Corollary 5. We have

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

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

https://doi.org/10.52202/079017-0383

D Details of Experiments

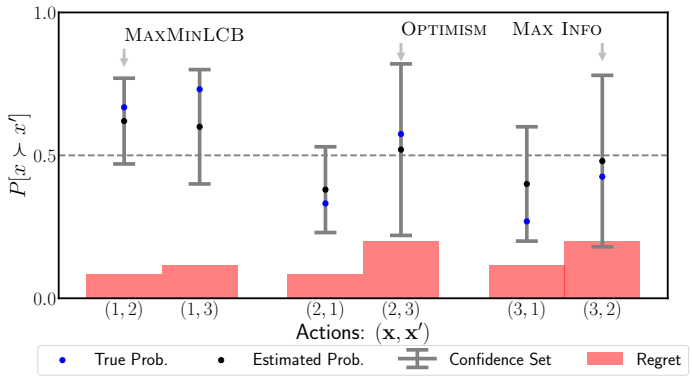


Figure 3: Confidence sets for an illustrative problem with 3 arms at a single time step. Annotated arrows highlight the action selection for three common approaches. MaxMinLCB selects the action pair $(1, 2)$ with the least regret. Upper-bound maximization (Optimism) and information maximization (Max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

Test Environments. We use a wide range of target functions common to the optimization literature [Jamil and Yang, 2013], to evaluate the robustness of MaxMinLCB. The results are reported in Table 1 and Table 2. Note that 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 $[-3, 3]$.

- Ackley: $\mathcal{X} = [-5, 5]^d, d = 2$

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

- Branin: $\mathcal{X} = [-5, 10] \times [0, 15]$

$$f(\boldsymbol{x}) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10$$

- Eggholder: $\mathcal{X} = [-512, 512]^2$

$$f(\boldsymbol{x}) = -(x_2 + 47)\sin\left(\sqrt{\left|x_2 + \frac{x_1}{2} + 47\right|}\right) - x_1 \sin\left(\sqrt{|x_1 - (x_2 + 47)|}\right)$$

- Hölder: $\mathcal{X} = [-10, 10]^2$

$$f(\boldsymbol{x}) = -\left|\sin(x_1)\cos(x_2)\exp\left(\left|1 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}\right|\right)\right|$$

- Matyas: $\mathcal{X} = [-10, 10]^2$

$$f(\boldsymbol{x}) = 0.26(x_1^2 + x_2^2) - 0.48x_1x_2$$

- Michalewicz: $\mathcal{X} = [0, \pi]^d, d = 2, m = 10$

$$f(\boldsymbol{x}) = -\sum_{i=1}^{d}\sin(x_i)\sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$$

- 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 $(\beta_t^{\mathrm{D}})_{t\geq 1}$.
Let $\mathcal{L}$ be any action from $\mathcal{X}$
for $t \geq 1$ **do**
  for $j = 1, \ldots, 2^t$ **do**
    Select $\boldsymbol{x}'_t$ uniformly randomly from $\mathcal{L}$
    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 $\sigma_{t+1}^{\mathrm{D}}$
  end for
  $\mathcal{L} \leftarrow$ the multi-set of actions played as $\boldsymbol{x}'_t$ 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

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

Input $(\beta_t^{\mathrm{D}})_{t\geq 1}$.
for $t \geq 1$ **do**
  Select $\boldsymbol{x}'_t$ uniformly randomly from $\mathcal{M}_t$
  Select $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{M}_t} s(h_t(\boldsymbol{x}, \boldsymbol{x}'_t)) + \beta_t^{\mathrm{D}}\sigma_t^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}'_t)$
  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

Acquisition Function Maximization. In our computations, to eliminate additional noise 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.

Hyper-parameters for Logistic Bandits. We set $\delta = 0.1$ for all algorithms. For GP-UCB and LGP-UCB, we set $\beta = 1$, and 0.25 for the noise variance. We use the Radial Basis Function (RBF) kernel and 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 hyper-parameter selections were done for each algorithm separately to create a fair comparison.

Hyper-parameters for Preference-based Bandits. 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. The same hyper-parameters are tuned separately for every baseline .

Pseudo-code for Baselines. Algorithm 5, Algorithm 6, and Algorithm 7 described the baselines used for the benchmark of Section 6.2. MAXINP and IDS are defined in Algorithm 3 and Algorithm 4,

https://doi.org/10.52202/079017-0383

respectively, in Appendix C.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.

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 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.

D.1 Yelp Experiment

We filter the Yelp data[3] for restaurants in Philadelphia, USA with at least 500 reviews and users who reviewed at least 90 restaurants. The final dataset includes 275 restaurants, 20 users, and a total of 2563 reviews. We define the action space $\mathcal{X}$ by assigning to each restaurant their respective 32-dimensional embedding of their reviews, i.e., $\mathcal{X} \subseteq \mathbb{R}^{32}$. For each restaurant, we concatenate all reviews in the filtered dataset and we use the TEXT-EMBEDDING-3-LARGE OpenAI embedding model [4] to retrieve the embeddings. The Yelp dataset provides utility values for users in the form of ratings on the scale of 1 to 5, however, not all users rated every restaurant. We use collaborative filtering to estimate the missing reviews [Schafer et al., 2007]. For each user, these values are used as the utilities for restaurants during the simulation. Experiments are conducted separately for each user, therefore, utility values are not aggregated but the action space is identical for each experiment.

Note that we do not assume any explicit functional form of the utility functions f that we calibrate to this data. Instead, the actions space $\mathcal{X}$ and utility values are derived separately from the dataset. Regardless, the results presented in Section 6.3 show that our kernelized method achieves good performance on this task.

E Additional Experiments

In this section, we provide Table 2 that details our logistic bandit benchmark, complementing the results in Section 6.1. Figure 4 and Figure 5 show the logistic and dueling regret of additional test functions, complementing the results of Table 1.

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

f	LGP-UCB	GP-UCB	IND-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

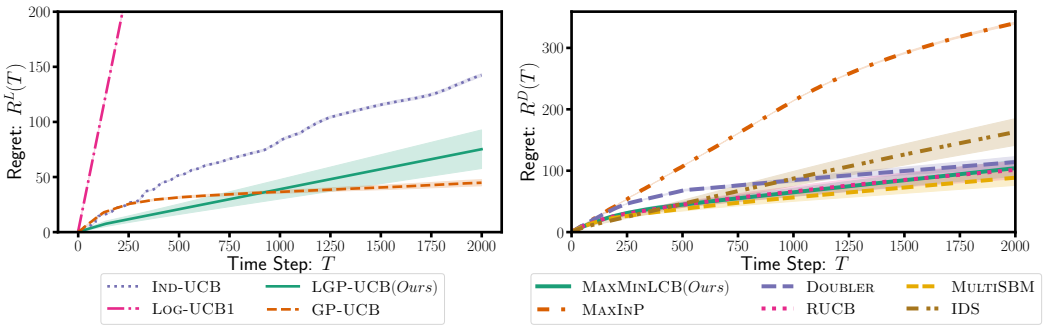


Figure 4: Regret with Branin utility function with logistic (left) and preference (right) feedback.

[3] Source: https://www.yelp.com/dataset
[4] Source: https://platform.openai.com/docs/guides/embeddings

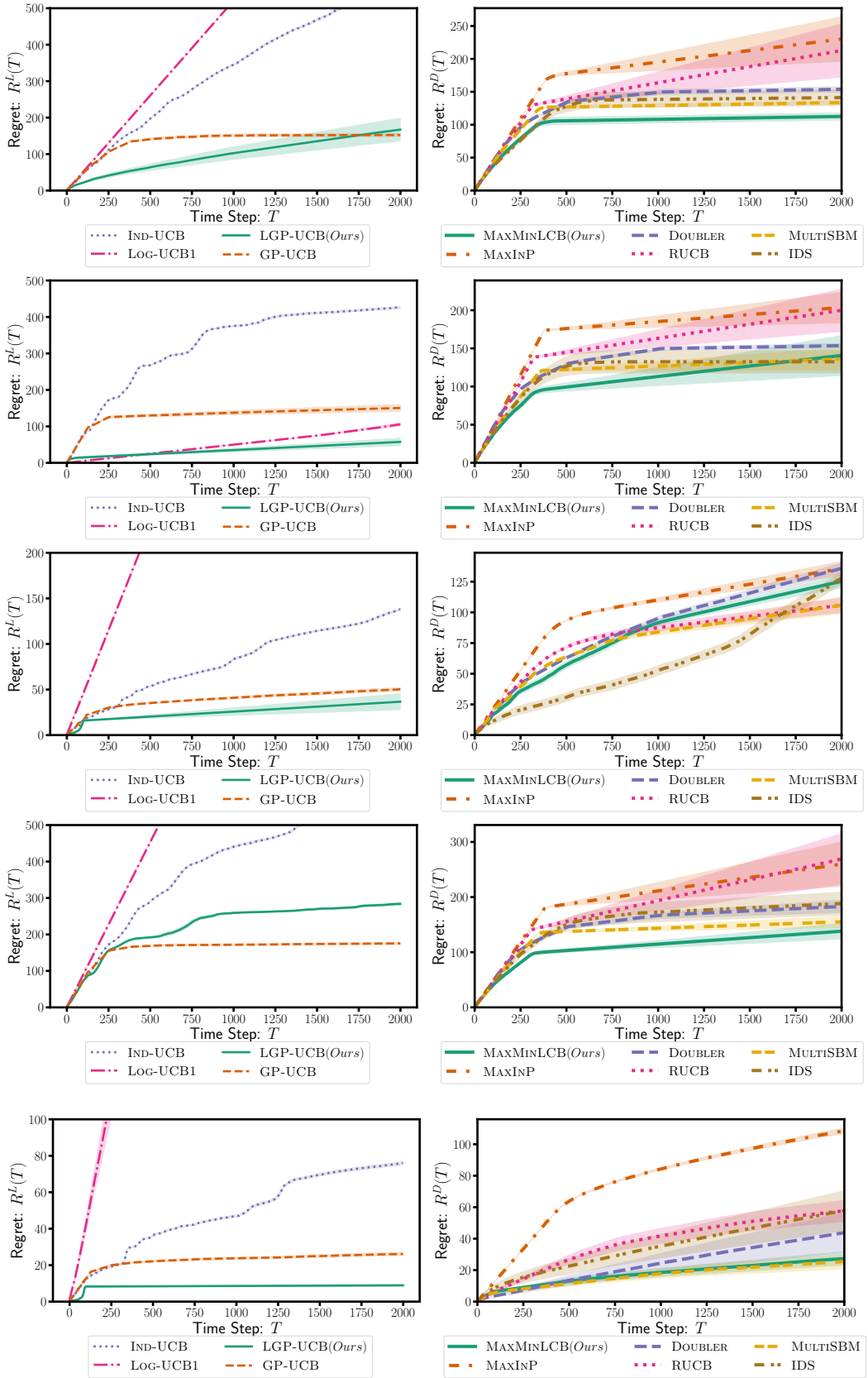


Figure 5: Top to bottom Regret for Eggholder, Hölder, Matyas Michalewicz, Rosenbrock functions, with logistic (left) and preference (right) feedback.

https://doi.org/10.52202/079017-0383

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We provide theoretical guarantees of the proposed algorithms in Section 4 and Section 5. In Section 6, we provide the results for our numerical experiments supporting our claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper is of theoretical nature. All theoretical assumptions are presented in the text, and we discuss how limiting they are. This is mainly in the Problem Setting section, or theorem statements.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: Proposition 1, Theorem 2, Corollary 3, Proposition 4, Corollary 5, Theorem 6 describe our theoretical proofs with detailed explanation of the assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We describe the experimental setting in Section 6 while providing further details on the hyperparameter selection, pseudocode, and utility function definition in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

https://doi.org/10.52202/079017-0383

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We include the code used to carry out the experiments in the Supplementary Material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We provide comprehensive explanations of the setting, hyperparameters, algorithms, and functions used in Section 6 and in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: We report all results in our figures and tables as the average and standard error over 20 random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We describe computation resources and time used for the experiments in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and conducted the research following the guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

    Answer: [NA]

https://doi.org/10.52202/079017-0383

Justification: The research presented in this paper is theoretical without any immediate societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: To the best of the authors' knowledge, there is no such risk involved.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Assets used to conduct the research presented in this paper are always cited and properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

https://doi.org/10.52202/079017-0383

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.