A Finite Time Analysis of Thompson Sampling for Bayesian Optimization with Preferential Feedback

Joseph Lazzaro
Imperial College London

Davide Buffelli
MediaTek Research

Da-shan Shiu
MediaTek Research

Sattar Vakili
MediaTek Research

Abstract

Preference feedback, in the form of pairwise comparisons rather than scalar scores, has seen increasing use in applications such as human-, laboratory-, and expert-in-the-loop design, as well as scientific discovery. We propose a Thompson Sampling (TS) approach to Bayesian optimization with preferential feedback that models comparisons using a monotone link on latent utility differences and leverages the dueling kernel induced by a base kernel. We provide a finite-time analysis showing that the performance of the proposed method matches that of standard TS for conventional Bayesian optimization with scalar feedback. The analysis exploits the anchor invariance of TS for challenger selection and introduces a double-TS pairing variant. We also demonstrate the performance of the method on both synthetic and real-world examples.

1 INTRODUCTION

Optimizing a black-box objective when only *pairwise preferences* between candidates are available, rather than calibrated scalar evaluations, arises in many modern applications. Examples include expert- and lab-in-the-loop design and evaluation, learning from human feedback, and scientific discovery settings in which absolute scoring is unavailable, too noisy, costly, or ill-defined, but relative judgments are comparatively reliable. These challenges have motivated a line of work on *Bayesian Optimization from Human Feedback* (BOHF), which models preferences via a monotone link applied to latent utility differences, most commonly through the Bradley–Terry–Luce model (Bradley and Terry, 1952). BOHF retains the sample-efficiency benefits of kernel methods while operating under strictly weaker feedback than conventional Bayesian optimization. Performance is typically measured by *regret*, which quantifies the loss relative to the globally optimal action across both elements of each queried pair (Pásztor et al., 2024; Xu et al., 2024; Kayal et al., 2025).

In this paper, we advocate *Thompson Sampling* (TS), which samples from the current posterior and uses the resulting draw as the acquisition function, as a particularly natural rule for BOHF. This contrasts with classic Bayesian optimization acquisition functions such as GP-UCB (Srinivas et al., 2010), which selects points with the largest upper confidence bound; GP-EI (Bull, 2011; Wang and de Freitas, 2014), which maximizes expected improvement; and GP-PI (Kushner, 1964), which maximizes the probability of improvement. Two properties make TS especially well suited to the preferential setting. First, *anchor independence*: a TS draw of the comparison function decomposes as a difference of latent utilities, so maximizing the draw with respect to the first argument while holding the other fixed is equivalent to maximizing a draw of the latent utility itself. This reduces pair selection to two independent single-argument maximizations without introducing anchor bias. Second, TS naturally supports drawing multiple independent samples from the same posterior surrogate, providing a simple and symmetric mechanism for selecting both elements of a queried pair at each round. These advantages contrast with acquisition functions such as UCB, whose indices couple the two arguments and are deterministic, making it difficult to select two points in a symmetric and fully sequential manner.

TS is a widely used and principled approach for sequential decision-making that naturally balances exploration and exploitation through posterior sampling. The idea is to sample a model from the posterior and select actions that are optimal under that model. This inherent randomness aligns exploration with the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

agent's uncertainty and often leads to more efficient learning compared to confidence-bound methods such as UCB (Russo and Van Roy, 2014). Empirically, TS has been observed to outperform UCB-style acquisitions across a range of bandit and Bayesian optimization problems (Agrawal and Goyal, 2012; Kaufmann et al., 2012).

Despite TS tracing back to Thompson (1933), finite-time analyses even for classical bandits emerged only in the 2010s (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Russo and Van Roy, 2014). For kernelized settings with scalar feedback, regret guarantees are known for GP-TS variants (Chowdhury and Gopalan, 2017; Vakili et al., 2021b). Preference feedback introduces additional challenges: the link nonlinearity and inference from pairwise comparisons, the design of pair-selection mechanisms rather than single-point queries, and the need to relate the resulting choices to standard steps in BO analysis. We develop an analysis of TS tailored to BOHF and show that its regret matches the rates achieved in standard BO with scalar feedback, thereby closing the gap between preferential and scalar-feedback settings. In Theorem 1, we prove that, with probability at least $1-\delta$, over T rounds the cumulative regret is

$$\tilde{\mathcal{O}}\big(\beta_T(\delta)\,\sqrt{T\,\Gamma(T)}\big), \tag{1}$$

where $\beta_T(\delta) = \mathcal{O}\left(\sqrt{\Gamma(T) + \log\frac{1}{\delta}}\right)$ is a $(1-\delta)$ confidence-width multiplier and $\Gamma(T)$ is the maximum information gain. This matches the regret bounds established by Chowdhury and Gopalan (2017) for standard TS in the conventional scalar-feedback setting.

Our work relates closely to recent BOHF literature. Pásztor et al. (2024) proposed a Stackelberg game to pick the pair (MaxMinLCB). Xu et al. (2024) presented an optimistic, UCB-based method that fixes one action to one of the points chosen in the previous round. The former attained the same theoretical guarantee as ours given in Equation (1), while the latter achieved a looser guarantee of $\tilde{\mathcal{O}}\big((T\,\Gamma(T))^{3/4}\big)$. Most recently, Kayal et al. (2025) developed a multi-round, batch exploration scheme that leverages tighter, offline-style confidence intervals to achieve an improved regret bound of $\tilde{\mathcal{O}}\big(\sqrt{T\,\Gamma(T)}\big)$, at the expense of full sequentiality. In contrast, our approach preserves the simplicity and fully sequential nature of standard acquisition-based BO, selects both points in an independent and symmetric manner, and provides finite-time guarantees that match those for scalar-feedback BO.

1.1 Contributions

In this paper, we make the following contributions. (i) We present a simple TS algorithm for BOHF that exploits anchor independence and multiple posterior samples to select *both* elements of each pair in a *fully sequential and symmetric* manner. Prior BOHF methods are either asymmetric, for example by fixing one action to a previously chosen point, or non-sequential due to batching. Our design proposes two contenders at every round without privileging either side. This is often desirable, and in some cases necessary, when eliciting preferences from human evaluators or large language model (LLM) judges. (ii) We provide a finite-time analysis establishing regret bounds that match the scalar-feedback setting. We extend the analysis of standard TS to preference feedback, which introduces additional challenges due to the link nonlinearity and the pairwise design. Our proof leverages anchor independence, proven via Proposition 1, together with novel techniques that relate the symmetrically selected pair to the uncertainty in their utility difference. Combined with information-theoretic bounds based on information gain, this yields $\tilde{\mathcal{O}}\big(\beta_T(\delta)\sqrt{T\,\Gamma(T)}\big)$ regret. (iii) We give a practical implementation recipe using kernel surrogates and validate it on synthetic benchmarks and a real catalyst design task using the OCx24 dataset (Abed et al., 2024), showing competitive or improved performance compared to existing methods.

1.2 Application and Experimental Setting

Experimental design and human-in-the-loop. In many laboratory and design workflows, absolute measurements are scarce, costly, or poorly calibrated, whereas relative judgments are cheap and reliable. For instance, in protein engineering or materials discovery one can often decide which of two candidates yields higher fluorescence, stability, or band gap under identical conditions without running (or trusting) a full calibration pipeline (Yang et al., 2019; Dang et al., 2025; Tucker et al., 2020); similar pairwise judgments arise in perceptual tuning and A/B testing. This motivates BOHF framework: a GP surrogate over pairwise differences that preserves BO's sample-efficiency while operating under strictly weaker, 1-bit feedback (Pásztor et al., 2024; Xu et al., 2024; Kayal et al., 2025).

Foundation models as preference oracles. Pretrained foundation models, most notably large language models (LLMs), can act as noisy *judges* over pairs by providing preference labels that encode broad domain knowledge (Zheng et al., 2023; Lee et al., 2024; Liu et al., 2024). This idea is central to learning from AI feedback in alignment and evaluation, where models are trained or assessed using pairwise compar-

isons rather than scalar scores (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Zheng et al., 2023; Liu et al., 2023). Combining such an LLM judge with BOHF turns preference-only optimization into a form of *knowledge discovery* from the model. In this setting, the optimizer queries comparative judgments to navigate large design spaces, while the GP surrogate aggregates these feedback signals into principled uncertainty estimates and acquisition decisions.

1.3 Related Work

Bayesian Optimization with Scalar Feedback. BO is a powerful framework for global optimization of expensive black-box functions, widely used in machine learning, science and engineering applications (Wang et al., 2023). Classical methods include Expected Improvement (Jones et al., 1998; Tran-The et al., 2022), Upper Confidence Bound (Srinivas et al., 2010), and Thompson Sampling (Chowdhury and Gopalan, 2017), which have achieved regret bounds of order $\Gamma(T)\sqrt{T}$, where $\Gamma(T)$ is the maximum information gain after T rounds. Further refinements have achieved either improved regret bounds (e.g., Whitehouse et al. 2024) or near-optimal rates, for instance through batched sampling methods (Li and Scarlett, 2022), *Sup* variants of the UCB algorithm (Valko et al., 2013), and domain-shrinking methods (Salgia et al., 2021), which attain order $\sqrt{\Gamma(T)T}$ matching the lower bounds of Scarlett et al. (2017).

Dueling Bandits. The dueling bandits problem extends the classical multi-armed bandit framework to settings where feedback is provided in the form of pairwise preferences rather than absolute rewards. In this paradigm, the learner selects two options at each round and receives feedback indicating which is preferred, rather than a numerical score. This approach is particularly relevant in applications where quantifying utility is challenging or expensive, but relative comparisons are feasible or cheap, such as recommendation systems and medical treatment design (for a survey of methods and applications, see Bengs et al., 2021). Noisy pairwise comparisons between arms are modeled by a preference matrix and the objective is typically defined relative to a Condorcet winner (an arm that beats every other with probability greater than 1/2) or, when none exists, a Copeland winner (maximizing the number of pairwise wins). Various methods have been proposed (e.g. Yue et al., 2012; Ailon et al., 2014; Zoghi et al., 2015; Jamieson et al., 2015; Ramamohan et al., 2016), including natural optimistic extensions like Relative UCB (Zoghi et al., 2014) and Double Thompson Sampling (Wu and Liu, 2016) which have been shown to perform well both numerically and theoretically with logarithmic regret. However, their performance and applicability is highly dependent on the size of the action space being moderately small.

Methods have also been proposed for larger action spaces. Firstly, linear contextual dueling bandits has been studied, where it is assumed that the arm rewards are linear across the action space (e.g. Dudík et al., 2015; Saha and Krishnamurthy, 2022; Bengs et al., 2022). A form of Thompson sampling, called Feel-Good Thompson Sampling, has also been applied to linear contextual dueling bandits (Li et al., 2024) and achieved near optimal regret of order $d\sqrt{T}$. For convex structures on the rewards, gradient descent methods have been proposed (Yue and Joachims, 2009; Saha et al., 2021). In this paper we will consider much more broad class of reward functions, which we will outline in Section 2.

Bayesian Optimization with Preference Feedback. In addition to the three works above that study the same BOHF setting as ours (Xu et al., 2024; Pásztor et al., 2024; Kayal et al., 2025), kernel-based learning from preferential feedback has also been analyzed by Xu et al. (2020) and Mehta et al. (2025), with applications to sample-efficient preference alignment in large language models. Both works, however, rely on strong assumptions: they introduce a *Borda* function—the probability that an action is preferred over a uniformly random alternative—under which the problem effectively reduces to standard Bayesian optimization.

A closely related setting is also studied by Astudillo et al. (2023), who consider Bayesian optimization from preference feedback over finite action spaces and derive asymptotic, instance-independent upper bounds on the simple regret. In contrast, we study continuous action spaces and provide non-asymptotic, instance-dependent regret guarantees.

There have also been a variety of heuristic approaches proposed for BOHF, which lack both finite-time theoretical guarantees and asymptotic convergence results (Mikkola et al., 2020; Takeno et al., 2023; Siivola et al., 2021; González et al., 2017).

2 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we overview the BOHF setup and the prediction and uncertainty tools that underlie our algorithmic design and analysis.

2.1 BOHF Framework

At each round $t = 1, 2, \ldots, T$, the agent selects a pair (x_t, x'_t) from the action domain $\mathcal{X}$. The feedback is a binary preference

$$y_t = \mathbb{1}\{x_t \succ x'_t\}, \qquad y_t \in \{0, 1\},$$

where $x_t \succ x'_t$ denotes that x_t is preferred to x'_t and $\mathbb{1}\{\cdot\}$ is the indicator. Following standard practice, for any pair $(x, x') \in \mathcal{X} \times \mathcal{X}$ the random variable $y = \mathbb{1}\{x \succ x'\}$ is modeled as

$$\Pr(y = 1 \mid x, x') = \mu\big(f(x) - f(x')\big),$$

with a known, strictly increasing link $\mu : \mathbb{R} \to [0, 1]$ satisfying $\mu(0) = \frac{1}{2}$; we adopt the logistic (sigmoid) link $\mu(u) = (1 + e^{-u})^{-1}$. The unknown latent utility $f : \mathcal{X} \to \mathbb{R}$ quantifies the value of each action. This is the Bradley–Terry–Luce (BTL) model (Bradley and Terry, 1952), widely used for bandits and RL with preference feedback (Pásztor et al., 2024; Xu et al., 2024; Zhan et al., 2024; Wu and Sun, 2024; Kayal et al., 2025). When $f(x) > f(x')$ we have $\Pr(x \succ x') > \frac{1}{2}$, and conversely if $f(x) < f(x')$.

This feedback is strictly weaker than in standard BO, where a scalar observation of f is revealed at each step. The objective is to adaptively select pairs so as to approach the globally preferred action

$$x^\star \in \arg\max_{x \in \mathcal{X}} f(x).$$

A common performance measure is the cumulative regret over T steps,

$$R(T) = \sum_{t=1}^{T} \frac{\Pr(x^\star \succ x_t) + \Pr(x^\star \succ x'_t) - 1}{2}. \tag{2}$$

This is equivalent (up to link-dependent constants) to the utility-based variant $\sum_{t=1}^{T} \big(f(x^\star) - \frac{1}{2}(f(x_t) + f(x'_t))\big)$ used by Xu et al. (2024); see Saha (2021).

2.2 Preliminaries and Assumptions

As in Pásztor et al. (2024); Xu et al. (2024); Kayal et al. (2025), we assume f lies in the reproducing kernel Hilbert space (RKHS) of a known positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This is a broad assumption: RKHSs of common kernels can approximate essentially all continuous functions on compact subsets of $\mathbb{R}^d$ (Srinivas et al., 2010). Let $\mathcal{H}_k$ be the RKHS of k, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$. The reproducing property gives $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$. Under mild conditions, Mercer's theorem yields

$$k(x, x') = \sum_{m=1}^{\infty} \gamma_m \, \varphi_m(x) \, \varphi_m(x'), \tag{3}$$

with $\gamma_m > 0$ and $\{\sqrt{\gamma_m}\varphi_m\}_m$ an orthonormal basis of $\mathcal{H}_k$. Any $f \in \mathcal{H}_k$ admits the expansion $f = \sum_{m=1}^{\infty} w_m \sqrt{\gamma_m}\varphi_m$, with $\|f\|^2_{\mathcal{H}_k} = \sum_m w_m^2$ (details in Appendix B).

Write $z = (x, x')$ and $h(z) = f(x) - f(x')$ for $(x, x') \in \mathcal{X} \times \mathcal{X}$. Following Pásztor et al. (2024), define the *dueling* kernel

$$k^\Delta(z_1, z_2) = k(x_1, x_2) + k(x'_1, x'_2) - k(x_1, x'_2) - k(x'_1, x_2), \tag{4}$$

for $z_1 = (x_1, x'_1)$ and $z_2 = (x_2, x'_2)$. One has $\|f\|_{\mathcal{H}_k} = \|h\|_{\mathcal{H}_{k^\Delta}}$ (Pásztor et al., 2024, Prop. 4).

Assumption 1. *The utility f belongs to $\mathcal{H}_k$ and satisfies $\|f\|_{\mathcal{H}_k} \le B$ for some $B > 0$. Without loss of generality, k is normalized so that $k(\cdot, \cdot) \le 1$ on $\mathcal{X} \times \mathcal{X}$.*

2.3 Preference Prediction and Uncertainty

We briefly recall the standard kernel ridge regression (KRR) predictor for scalar-feedback before turning to preferences. Given scalar data $\{(x_i, o_i)\}_{i=1}^t$ with $o_i = f(x_i) + \varepsilon_i$, KRR yields

$$\hat{f}_t(x) = k_t^\top(x) \, (K_t + \lambda I)^{-1} \boldsymbol{o}_t, \tag{5}$$
$$k_t(x, x') = k(x, x') - k_t^\top(x) \, (K_t + \lambda I)^{-1} k_t(x'),$$

where $k_t(x) = [k(x, x_i)]_{i=1}^t$, $K_t = [k(x_i, x_j)]_{i,j=1}^t$, $\lambda > 0$, and $\boldsymbol{o}_t = [o_i]_{i=1}^t$. The estimator $\hat{f}_t$ solves

$$\hat{f}_t = \arg\min_{g \in \mathcal{H}_k} \sum_{i=1}^{t} \big(g(x_i) - o_i\big)^2 + \tfrac{\lambda}{2}\|g\|^2_{\mathcal{H}_k},$$

and admits confidence bounds of the form

$$|f(x) - \hat{f}_t(x)| \le \hat{\beta}(\delta) \, \sigma_t(x),$$

where $\sigma_t^2(x) := k_t(x, x)$ denotes the posterior variance, and $\hat{\beta}(\delta)$ is a confidence-width multiplier, under standard conditions (Abbasi-Yadkori, 2013; Chowdhury and Gopalan, 2017; Vakili et al., 2021a; Whitehouse et al., 2024).

With preference feedback, there is no closed-form analogue of (5). The setting is classification-like with binary labels. For a history $H_t = \{(x_i, x'_i, y_i)\}_{i=1}^t$, consider the regularized logistic loss over $h \in \mathcal{H}_{k^\Delta}$:

$$\mathcal{L}_{k^\Delta}(h, H_t) = \sum_{i=1}^{t} \Big[-y_i \log \mu\big(h(x_i, x'_i)\big) - (1 - y_i) \log \big(1 - \mu(h(x_i, x'_i))\big) \Big] + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_{k^\Delta}}.$$

Define the predictor h_t as the minimizer

$$h_t = \arg\min_{h \in \mathcal{H}_{k^\Delta}} \mathcal{L}_{k^\Delta}(h, H_t). \tag{6}$$

By the representer theorem, h_t admits the finite expansion

$$h_t(\cdot) = \sum_{i=1}^{t} \theta_i \, k^{\Delta}\big(\cdot, (x_i, x_i')\big), \tag{7}$$

with coefficients $\boldsymbol{\theta}_t = [\theta_1, \ldots, \theta_t]^\top \in \mathbb{R}^t$. Writing $k_t^{\Delta}(z) = [k^{\Delta}(z, (x_j, x_j'))]_{j=1}^t$ and $k_t^{\Delta} = [k^{\Delta}((x_i, x_i'), (x_j, x_j'))]_{i,j=1}^t$, the loss in parameter space is

$$\mathcal{L}_{k^{\Delta}}(\boldsymbol{\theta}, H_t) = \sum_{i=1}^{t} \Big[- y_i \log \mu\big(\boldsymbol{\theta}^\top k_t^{\Delta}(x_i, x_i')\big) - (1 - y_i) \log \big(1 - \mu(\boldsymbol{\theta}^\top k_t^{\Delta}(x_i, x_i'))\big)\Big] + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2. \tag{8}$$

In analogy with (5), an uncertainty proxy is

$$k_t^{\Delta}(z, z') \;=\; k^{\Delta}(z, z') - (k_t^{\Delta})^\top(z) \left(K_t^{\Delta} + \lambda \, \kappa \, I\right)^{-1} k_t^{\Delta}(z'), \tag{9}$$

where

$$\kappa \;:=\; \sup_{(x,x') \in \mathcal{X}^2} \frac{1}{\dot{\mu}\big(f(x) - f(x')\big)} \;=\; \sup_{z \in \mathcal{X}^2} \frac{1}{\dot{\mu}\big(h(z)\big)}, \tag{10}$$

with $\dot{\mu}$ the derivative of the link μ. We use the notation $\sigma_t^2(z) = k_t^{\Delta}(z, z)$. We formally define the maximum information gain with t preference queries as

$$\Gamma(t) \;=\; \max_{(x_1, x_1'), \ldots, (x_t, x_t')} \tfrac{1}{2} \log \det\big(I + (\lambda\kappa)^{-1} K_t^{\Delta}\big) \,. \tag{11}$$

We can then provide the following confidence sequence for our estimate of the difference function h_t from (6).

Lemma 1. *Under Assumption 1, with probability at least $1 - \delta$, for all $t \geq 1$ and all $x, x' \in \mathcal{X}$,*

$$|\, h_t(x, x') - h(x, x') \,| \;\leq\; \beta_t(\delta) \, \sigma_t(x, x'),$$

where

$$\beta_t(\delta) \;=\; 4B \;+\; 2\sqrt{\tfrac{2\kappa}{\lambda} \Big(\Gamma(t) + \log(1/\delta)\Big)}. \tag{12}$$

Proof. This result follows from the proof of (Pásztor et al., 2024, Corollary 5), with the composition by the link function omitted. As a consequence, the Lipschitz constant does not appear in the scaling of $\beta_t(\delta)$. □

3 THOMPSON SAMPLING: FROM SCALAR TO PREFERENTIAL FEEDBACK

We describe a TS policy for BOHF, which we refer to as Preferential Feedback TS (PF-TS). The policy uses the pairwise predictor h_t together with the associated uncertainty (covariance) k_t^{Δ} from Section 2.3 (equations (6) and (9)) to define a Gaussian surrogate posterior on pairs, from which we draw sample paths.

Surrogate posterior on pairs. At round t, we work with the GP on $\mathcal{X} \times \mathcal{X}$ having mean function h_t and covariance k_t^{Δ}, and draw two *independent* samples

$$\tilde{h}_t^{(1)}(\cdot, \cdot), \; \tilde{h}_t^{(2)}(\cdot, \cdot) \;\sim\; \mathrm{GP}\big(h_t, v_t^2 k_t^{\Delta}\big).$$

To mitigate under-exploration, we include an exploration scale $v_t > 0$, as in standard TS for scalar-feedback Bayesian optimization, which multiplies the posterior standard deviation in the sampling step. Its schedule is discussed in Section 4, and in particular we will set $v_t = \beta_t(\delta)$ in our analysis.

Algorithm 1 Preferential Feedback Thompson Sampling (PF-TS)

Require: Horizon T
Initialize $t \leftarrow 1$.
while $t \leq T$ **do**
  Compute h_t and k_t^{Δ} according to (6) and (9).
  Sample $\tilde{h}_t^{(1)}(\cdot, \cdot) \sim \mathrm{GP}\big(h_t, v_t^2 k_t^{\Delta}\big)$.
  Sample $\tilde{h}_t^{(2)}(\cdot, \cdot) \sim \mathrm{GP}\big(h_t, v_t^2 k_t^{\Delta}\big)$.
  Pick an anchor $x_0 \in \mathcal{X}$.
  Select $x_t \in \arg\max_{x \in \mathcal{X}} \; \tilde{h}_t^{(1)}(x, x_0)$,
  Select $x_t' \in \arg\max_{x \in \mathcal{X}} \; \tilde{h}_t^{(2)}(x, x_0)$.
  Query (x_t, x_t'), observe preference $y_t \in \{0, 1\}$, and augment the data.
  Update $t \leftarrow t + 1$.
end while

3.1 Anchor independence and reduction to single-argument maximization.

Fix any anchor $x_0 \in \mathcal{X}$ as a device to score candidates via $\tilde{h}_t^{(i)}(\cdot, x_0)$. Because the comparison function is a difference, $h(x, x') = f(x) - f(x')$, maximizing over x yields

$$x^\star \;\in\; \arg\max_{x \in \mathcal{X}} \, h(x, x_0) \;=\; \arg\max_{x \in \mathcal{X}} \, f(x),$$

so the maximizer is independent of the anchor x_0. In Section 4.1 we show that any posterior Thompson sample $\tilde{h}_t$ is also separable, i.e., $\tilde{h}_t(x, x') = \tilde{f}_t(x) - \tilde{f}_t(x')$, and the same argument applies. Hence, the algorithm's performance is *anchor-independent*; any choice of x_0 yields the same selected actions.

3.2 Action selection

The pair at round t is chosen by a *double* Thompson Sampling procedure from two independently sampled difference functions. Draw two independent surrogates $\tilde{h}_t^{(1)}, \tilde{h}_t^{(2)} \sim \mathrm{GP}(h_t, v_t^2 k_t^{\Delta})$ and fix an anchor $x_0 \in \mathcal{X}$. Select

$$x_t \in \arg\max_{x \in \mathcal{X}} \, \tilde{h}_t^{(1)}(x, x_0), \qquad x_t' \in \arg\max_{x \in \mathcal{X}} \, \tilde{h}_t^{(2)}(x, x_0).$$

Finally, play the pair (x_t, x'_t), observe y_t, and recompute h_{t+1} and k^{Δ}_{t+1} via (6) and (9).

4 ANALYSIS

In this section, we analyze the performance of the PF-TS algorithm.

4.1 Anchor-independence

A key property that makes TS naturally extendable to the BOHF setting is the anchor-independence of the acquisition (TS samples), in contrast to other acquisitions such as UCB. We formalize this below.

Proposition 1. *Let k^{Δ} be a dueling kernel and let $H_t = \{(x_i, x'_i, y_i)\}_{i=1}^{t}$ be the observation set as in Section 2.1. Let h_t and k^{Δ}_t be the predictor and uncertainty estimates defined via (6) and (9). Let $\tilde{h}_t \sim GP(h_t, v_t^2 k^{\Delta}_t)$ be a Thompson sample on $\mathcal{X}^2$, where v_t is a fixed sequence. Then there exists a function $\tilde{f}_t$ such that, for all $(x, x') \in \mathcal{X}^2$, almost surely,*

$$\tilde{h}_t(x, x') \;=\; \tilde{f}_t(x) \;-\; \tilde{f}_t(x'). \tag{13}$$

Here, $\tilde{f}_t$ can be interpreted as a posterior sample of the latent utility f. A detailed proof is provided in Appendix C.

This contrasts with the UCB index, which is not separable in this way. As a result, GP-UCB is not anchor-independent: the choices of x_t and x'_t generally depend on the anchor, and the reduction used for TS does not apply. Figure 1 illustrates this by plotting the argmax of the first argument (x_t) while fixing the second argument to an anchor point x_0. We compare UCB acquisitions (constructed using Lemma 1) with Thompson samples for different anchors x_0. While x_t is invariant to the anchor under TS, it clearly depends on the anchor under UCB.

The anchor-independence property is also practically relevant. When selecting an action by maximizing the acquisition function, as described in Section 3.2, it suffices to optimize over a single parameter at a time. This stands in contrast to prior BOHF methods, which require searching over pairs of parameters and therefore incur a higher computational burden. In our current implementations, we perform this optimization by searching over a discrete set of candidate parameters; further details are provided in Appendices D and F.

Bayesian optimization settings often assume that data acquisition, such as querying a human or model-based preference oracle, is the dominant cost. As a result, computational savings in optimizing the acquisition function are typically secondary to sample efficiency,

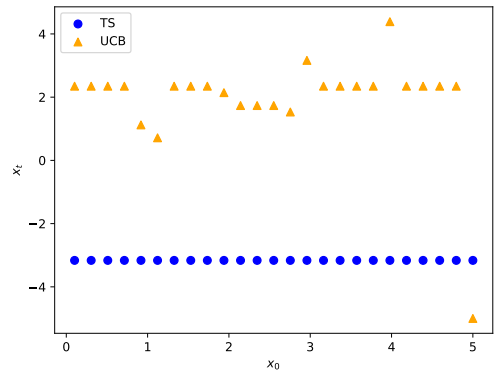


Figure 1: Selected x_t under TS (blue) and UCB (orange) as the anchor x_0 varies. Utility is the Ackley function; the history contains 200 uniformly sampled pairs. TS is anchor invariant: the chosen x_t does not change with x_0, whereas the point chosen by UCB depends on the anchor.

which is the primary focus of this paper. Nevertheless, the uncertainty estimates used by our method can themselves be approximated. For instance, posterior variances may be approximated via Nyström methods applied to the kernel matrix inverse. Such approximations preserve the regret guarantees, provided they only affect the exploration scale up to constant factors, as established in related settings by Vakili et al. (2022).

4.2 The performance analysis of PF-TS

In this section, we provide a finite-time analysis of the PF-TS algorithm. We recover the same guarantees as vanilla TS with scalar-valued feedback, despite operating under a more restrictive feedback model. In particular: (i) preference feedback supplies only minimal, 1-bit information per query; (ii) unlike vanilla TS, our selection uses an anchor point, whose effect we handle via the separability (anchor-independence) property; and (iii) we select the pair by drawing two independent Thompson samples and maximizing against the same anchor, while the observed label depends on the pair, necessitating arguments different from standard TS in conventional BO. Our proof develops techniques that relate the symmetrically selected pair to the posterior uncertainty of their utility difference and leverages bounds based on information gain.

Theorem 1. *Consider the BOHF problem introduced in Section 2.1. Under Assumption 1, the performance of the PF-TS given in Algorithm 1, with $v_t = \beta_t(\delta)$, satisfies*

$$R(T) \;=\; \tilde{\mathcal{O}}\Big(\beta_T(\delta)\,\sqrt{T\,\Gamma(T)}\Big), \tag{14}$$

with probability greater than $1-2\delta$. Here $\Gamma(T)$ denotes

the maximum information gain under the dueling kernel k^{Δ} and we have omitted logarithmic terms.

Proof sketch. On a high-probability event (uniform in t), both the surrogate mean h_{t-1} and the two TS draws $\tilde{h}_t^{(1)}, \tilde{h}_t^{(2)}$ lie within $\beta_t(\delta)$ standard deviations of the true h. Using separability $h(x, x') = f(x) - f(x')$ and any anchor x_0, write $h(x^\star, x_t) = h(x^\star, x_0) - h(x_t, x_0)$, add and subtract the TS draw used to pick x_t, and use that it maximizes $x \mapsto \tilde{h}_t^{(1)}(x, x_0)$. Choosing the anchor by *anchor independence* ($x_0 = x'_t$ for the first draw and $x_0 = x_t$ for the second) yields the per-round utility regret bounded by (up to constants)

$$\beta_t(\delta)\sqrt{\log\left(\tfrac{t}{\delta}\right)}\Big[\sigma_{t-1}(x_t, x'_t) \;+\; \sigma_{t-1}(x^\star, x_t) \;+\; \sigma_{t-1}(x^\star, x'_t)\Big],$$

where the Lipschitzness of the link converts utility gaps to probability gaps in (2).

Intuitively, the terms involving $x^\star$ are small: as the algorithm explores, points near the maximizer are repeatedly compared, so the posterior uncertainty for pairs comparing $x^\star$ shrinks quickly. We prove this by extending standard TS tools: an anti-concentration argument ensures that, with constant probability each round, the TS draw picks a point whose variance (relative to the other action as anchor) is not overly small; this lets us link the $x^\star$-terms to the variance at the played pair and bound $\sigma_{t-1}(x^\star, x_t) + \sigma_{t-1}(x^\star, x'_t)$ in expectation by $\sigma_{t-1}(x_t, x'_t)$ (up to constants). Finally, the sum of sequential standard deviations at the queried pairs is controlled by an information-gain bound (analogous to the elliptical potential lemma in linear bandits), yielding

$$\sum_{t=1}^{T} \sigma_{t-1}(x_t, x'_t) \;=\; \tilde{\mathcal{O}}\big(\sqrt{T\,\Gamma(T)}\big),$$

and thus the overall regret rate.

Remark 1. *The quantity $\Gamma(T)$ is a kernel-dependent, algorithm-independent complexity term that appears throughout BO and BOHF analyses (see, e.g., Srinivas et al., 2010; Pásztor et al., 2024; Xu et al., 2024). Its growth with T is well understood for common kernels: for linear kernels, $\Gamma(T) = \mathcal{O}(d \log T)$; for kernels with exponentially decaying Mercer eigenvalues (e.g., Squared Exponential), $\Gamma(T) = \mathrm{polylog}(T)$; and for Matérn kernels with smoothness $\nu > 1/2$ in dimension d, $\Gamma(T) = \tilde{\mathcal{O}}\big(T^{d/(2\nu+d)}\big)$ (see, e.g., Vakili et al., 2021a). Furthermore, Pásztor et al. (2024, Prop. 4) show that the dueling kernel has eigenvalues exactly twice those of the base kernel, so the* scaling *of $\Gamma(T)$ with T is identical for the dueling and base kernels. Consequently, our bound attains the same regret order as GP-TS in conventional BO with scalar feedback.*

5 EXPERIMENTS

In this section we illustrate the numerical performance of PF-TS. In the interest of space, we leave some additional experiments to be presented in Appendix E.

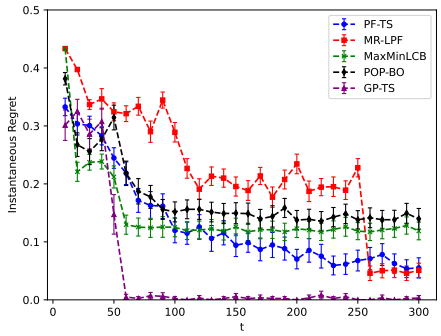
5.1 Comparison with Existing Methods

We start by considering a synthetic utility function f given by the one-dimensional Ackley function, a non-convex benchmark with many local optima, widely used to test Bayesian optimization algorithms (Jamil and Yang, 2013) and adopted in recent BOHF studies (Pásztor et al., 2024; Kayal et al., 2025). In Figures 2a and 2b, we plot the regret for PF-TS (ours), MR-LPF (Kayal et al., 2025), MaxMinLCB (Pásztor et al., 2024), and POP-BO (Xu et al., 2024). We use the same inference procedure as in Section 2.3 and a Matérn-5/2 kernel for all methods. Figure 2a reports the mean instantaneous regret, and Figure 2b the mean cumulative regret (with standard errors), over 30 runs up to $T = 300$. For comparison, we also include vanilla GP-TS (Chowdhury and Gopalan, 2017), which receives direct scalar feedback on the same f (with standard Gaussian observation noise). To place them on the same regret scale as the BOHF methods, we evaluate their regret by repeating the same action at each time t, i.e., $x'_t = x_t$.

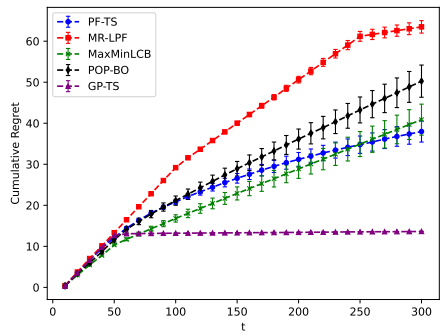
From Figure 2b, the cumulative regret of our method after 300 iterations is significantly smaller than that of MR-LPF and POP-BO, and competitive with MaxMinLCB. We also see in Figure 2a that the instantaneous (simple) regret achieved by our method at $T = 300$ is lower than that of MaxMinLCB. By design, MR-LPF attains competitive instantaneous regret at this horizon, consistent with its asymptotically optimal regret order (Kayal et al., 2025). Note, the cumulative regret curve appears smoother than the instantaneous regret curve because it aggregates the instantaneous values over time, averaging out short-term fluctuations and noise.

5.2 Catalyst Design with Preference Feedback (OCx24)

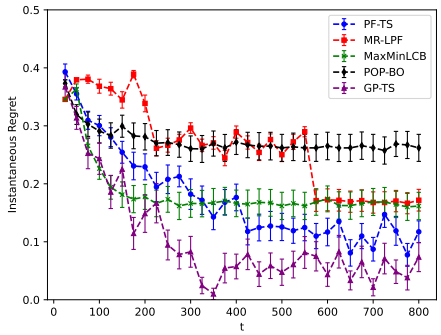
We evaluate a practical case study: optimizing catalyst composition to maximize hydrogen yield. Let $f : \mathcal{X} \to \mathbb{R}$ denote the achieved yield for a composition $x \in \mathcal{X}$. This setting naturally supports preference-based feedback: precise yield measurements can be slow or costly, whereas pairwise judgments between candidates are often much easier to obtain. In many laboratory workflows, early-time indicators (e.g., initial rates or partial conversion) are sufficiently predictive to rank two catalysts, making preference labels a



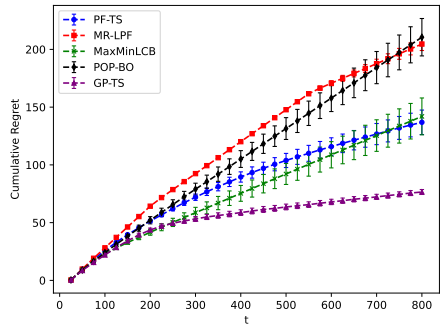
(a) **(Ackley)** Instantaneous Regret



(b) **(Ackley)** Cumulative Regret



(c) **(Catalyst)** Instantaneous Regret



(d) **(Catalyst)** Cumulative Regret

Figure 2: Regret of PF-TS, MR-LPF, MaxMinLCB, POP-BO, and GP-TS. Mean over 30 runs; bands show ± 1 standard error. (a,b) Ackley objective ($T = 300$). (c,d) Catalyst hydrogen-yield objective ($T = 800$). Left column: instantaneous (simple) regret. Right column: cumulative regret (2).

viable and efficient proxy.

We use the Open Catalyst Experiments 2024 (OCx24) dataset released by *FAIR Chemistry* (Meta AI) (Abed et al., 2024), which reports hydrogen yields for 63 catalyst compositions formed from silver (Ag), gold (Au), and zinc (Zn). Each composition is a triplet $(x_{\mathrm{Ag}}, x_{\mathrm{Au}}, x_{\mathrm{Zn}}) \in [0,1]^3$ with $x_{\mathrm{Ag}} + x_{\mathrm{Au}} + x_{\mathrm{Zn}} = 1$. We set the utility f to the recorded hydrogen yield for each composition and synthetically generate preference feedback from f in our implementation.

Figures 2c and 2d plot regret for PF-TS (ours), MR-LPF (Kayal et al., 2025), MaxMinLCB (Pásztor et al., 2024), and POP-BO (Xu et al., 2024), using the same inference procedure as Section 2.3 and a Matérn-5/2 kernel with hyperparameters same as the previous experiments. For reference, we also include cumulative regret for vanilla GP-TS with direct scalar feedback on f. Results are averaged over 30 runs. Over $T = 800$ iterations, PF-TS achieves the lowest instantaneous regret among BOHF methods and substantially lower cumulative regret than MR-LPF and POP-BO, while remaining competitive with MaxMinLCB.

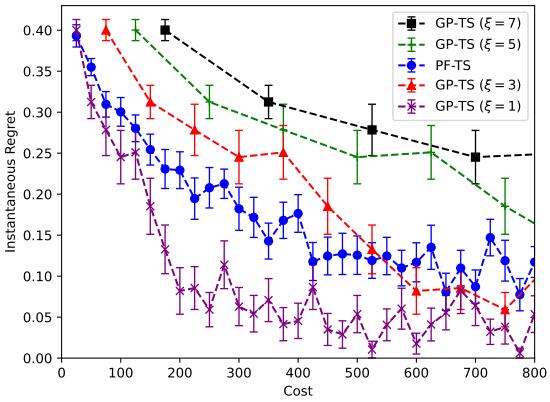


Figure 3: Instantaneous (simple) regret versus total cost c (preference units) up to 800 for PF-TS and GP-TS under cost ratios $\xi \in \{1, 3, 5, 7\}$ with Catalyst data. For GP-TS, c buys c/ξ scalar iterations; PF-TS runs c preference iterations. We plot regret and ± 1 standard error after each additional 25 iterations can be bought.

5.3 Preference vs. Scalar Feedback

Since the goal is to optimize the latent utility f, direct scalar observations are inherently more informative than pairwise preferences. Accordingly, in Figure 2 the vanilla scalar-feedback method (GP-TS) accrues markedly less regret than BOHF methods when

compared at the same *number of iterations*. However, when preference labels are substantially cheaper to obtain than scalar measurements, BOHF can be advantageous under a fixed *cost* budget.

To make this concrete, suppose a single scalar observation costs $\xi \geq 1$ units relative to one preference query (cost 1). In Figure 3 we compare our PF-TS to GP-TS on the OCx24 dataset under a total cost budget c (in preference units), for $\xi \in \{1, 3, 5, 7\}$. PF-TS runs for c preference iterations, while GP-TS is allotted only c/ξ scalar iterations. We plot the instantaneous (simple) regret of each method at cost $c \leq 800$. When preference feedback is sufficiently cheaper (larger ξ), PF-TS becomes preferable on a cost-adjusted basis.

6 CONCLUSION

We introduced PF-TS, a Thompson-sampling approach to BOHF that leverages anchor independence and draws multiple posterior samples to select both actions each round in a fully sequential and symmetric manner. By our main guarantee (Regret bound in Theorem 1), PF-TS achieves cumulative regret $\tilde{\mathcal{O}}\big(\beta_T(\delta)\sqrt{T\,\Gamma(T)}\big)$, matching the order attained by GP-TS under scalar feedback. Empirically, on both synthetic (Ackley) and real (OCx24) objectives, PF-TS delivers competitive or improved performance to recent BOHF baselines at practical horizons.

Its symmetry in pair selection and full sequentiality make PF-TS a natural fit for applications where pairwise judgments are cheap and timely, e.g., human-in-the-loop design, LLM-as-judge evaluation, and high-throughput scientific discovery. Under cost budgets where scalar labels are more expensive than preferences, PF-TS can outperform standard BO on cost-adjusted regret. Looking ahead, promising directions include cost-aware hybrids of scalar and preference feedback, *one-out-of-many* preference queries (in contrast to the *one-out-of-two* BOHF setting), and large-scale experiments with LLM judges, including active data collection for fine-tuning large models.

Acknowledgments

This work was completed while Joseph Lazzaro was an intern at MediaTek Research.

References

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

J. Abed, J. Kim, M. Shuaibi, B. Wander, B. Duijf, S. Mahesh, H. Lee, V. Gharakhanyan, S. Hoogland, E. Irtem, et al. Open catalyst experiments 2024 (ocx24): Bridging experiments and computational models. *arXiv preprint arXiv:2411.11783*, 2024.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2014.

R. Astudillo, Z. J. Lin, E. Bakshy, and P. Frazier. qeubo: A decision-theoretic acquisition function for preferential bayesian optimization. In *International conference on artificial intelligence and statistics*, 2023.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

V. Bengs, R. Busa-Fekete, A. El Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021.

V. Bengs, A. Saha, and E. Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, 2022.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 2011.

S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017.

P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

T. Dang, L.-H. Pham, S. T. Truong, A. Glenn, W. Nguyen, E. A. Pham, J. S. Glenn, S. Koyejo, and T. Luong. Preferential multi-objective bayesian optimization for drug discovery. In *ICLR 2025 Workshop on Human-AI Coevolution*, 2025.

M. Dudík, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, 2015.

J. González, Z. Dai, A. Damianou, and N. D. Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, 2017.

M. W. Hoffman, B. Shahriari, and N. de Freitas. Exploiting correlation and budget constraints in bayesian multi-armed bandit optimization. *arXiv preprint arXiv:1303.6746*, 2013.

K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, 2015.

M. Jamil and X.-S. Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 2013.

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 1998.

É. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, 2012.

A. Kayal, S. Vakili, L. Toni, D.-s. Shiu, and A. Bernacchia. Bayesian optimization from human feedback: Near-optimal regret bounds. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

H. J. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 1964.

H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. R. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

X. Li, H. Zhao, and Q. Gu. Feel-good thompson sampling for contextual dueling bandits. In *Forty-first International Conference on Machine Learning*, 2024.

Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023.

Y. Liu, H. Zhou, Z. Guo, E. Shareghi, I. Vulić, A. Korhonen, and N. Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *Conference on Language Modeling (COLM)*, 2024.

V. Mehta, S. Belakaria, V. Das, O. Neopane, Y. Dai, I. Bogunovic, B. E. Engelhardt, S. Ermon, J. Schneider, and W. Neiswanger. Sample efficient preference alignment in LLMs via active exploration. In *Second Conference on Language Modeling*, 2025.

P. Mikkola, M. Todorović, J. Järvi, P. Rinke, and S. Kaski. Projective preferential bayesian optimization. In *International Conference on Machine Learning*, 2020.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

B. Pásztor, P. Kassraie, and A. Krause. Bandits with preference feedback: A stackelberg game perspective. In *Advances in Neural Information Processing Systems 38*, 2024.

S. Y. Ramamohan, A. Rajkumar, and S. Agarwal. Dueling bandits: Beyond condorcet winners to general tournament solutions. *Advances in Neural Information Processing Systems*, 29, 2016.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 2014.

A. Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 2021.

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

A. Saha, T. Koren, and Y. Mansour. Dueling convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

S. Salgia, S. Vakili, and Q. Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 2021.

J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, 2017.

E. Siivola, A. K. Dhaka, M. R. Andersen, J. González, P. G. Moreno, and A. Vehtari. Preferential batch bayesian optimization. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

S. Takeno, M. Nomura, and M. Karasuyama. Towards practical preferential bayesian optimization with skew gaussian processes. In *International Conference on Machine Learning*, 2023.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

H. Tran-The, S. Gupta, S. Rana, and S. Venkatesh. Regret bounds for expected improvement algorithms in gaussian process bandit optimization. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

M. Tucker, E. Novoseller, C. Kann, Y. Sui, Y. Yue, J. W. Burdick, and A. D. Ames. Preference-based learning for exoskeleton gait optimization. In *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020.

S. Vakili, N. Bouziani, S. Jalali, A. Bernacchia, and D.-s. Shiu. Optimal order simple regret for gaussian process bandits. *Advances in Neural Information Processing Systems*, 2021a.

S. Vakili, H. Moss, A. Artemev, V. Dutordoir, and V. Picheny. Scalable thompson sampling using sparse gaussian process models. *Advances in neural information processing systems*, 2021b.

S. Vakili, J. Scarlett, D.-s. Shiu, and A. Bernacchia. Improved convergence rates for sparse approximation methods in kernel-based learning. In *International Conference on Machine Learning*, 2022.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

X. Wang, Y. Jin, S. Schmitt, and M. Olhofer. Recent advances in bayesian optimization. *ACM Computing Surveys*, 2023.

Z. Wang and N. de Freitas. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. In *Advances in Neural Information Processing Systems*, 2014.

J. Whitehouse, A. Ramdas, and S. Z. Wu. On the sublinear regret of gp-ucb. *Advances in Neural Information Processing Systems*, 36, 2024.

H. Wu and X. Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 2016.

R. Wu and W. Sun. Making RL with preference-based feedback efficient via randomization. In *The Twelfth International Conference on Learning Representations*, 2024.

W. Xu, W. Wang, Y. Jiang, B. Svetozarevic, and C. Jones. Principled preferential bayesian optimization. In *International Conference on Machine Learning*, 2024.

Y. Xu, A. Joshi, A. Singh, and A. Dubrawski. Zeroth order non-convex optimization with dueling-choice bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

K. K. Yang, Z. Wu, and F. H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 2019.

Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 2012.

W. Zhan, M. Uehara, N. Kallus, J. D. Lee, and W. Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

L. Zheng, W. Chiang, Y. Sheng, Z. Li, N. Du, T. Zhang, X. Ma, P. Abbeel, I. Stoica, Y. Zhuang, E. P. Li, and D. Lin. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.05685.

L. Zimmer, M. Lindauer, and F. Hutter. Auto-pytorch: Multi-fidelity metalearning for efficient and robust autodl. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

M. Zoghi, S. Whiteson, R. Munos, and M. Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

M. Zoghi, Z. S. Karnin, S. Whiteson, and M. De Rijke. Copeland dueling bandits. *Advances in neural information processing systems*, 2015.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] This has been outlined in detail in Sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] One main focus of this paper is the analysis of our proposed PF-TS algorithm, with both theoretical results in Section 4 and empirical studies in Section 5.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] This will be included in the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Assumptions have been clearly stated in Section 2 and for Theorem statements in Section 4. Any missing details are readily available in the appendix.
   (b) Complete proofs of all theoretical results. [Yes] Complete proofs can be found in the appendix.
   (c) Clear explanations of any assumptions. [Yes] Clear explanations are given in Sections 2 and 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code and directions can be found in the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] This is discussed in detail in the appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We state and include standard error bars in all plots.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] This is discussed in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Notation Summary

Notation	Description
$\mathcal{X}$	Action (design) space.
x, x', x_t, x'_t	Actions selected at a given round; (x_t, x'_t) is the queried pair at round t.
T	Total number of optimization rounds (horizon).
$f : \mathcal{X} \to \mathbb{R}$	Unknown latent utility function to be maximized.
$x^\star$	Globally optimal action, $x^\star \in \arg\max_{x\in\mathcal{X}} f(x)$.
$y_t \in \{0, 1\}$	Observed preference feedback at round t; $y_t = 1$ if $x_t \succ x'_t$.
$\mu(\cdot)$	Known monotone link function mapping utility differences to preference probabilities (e.g., logistic link).
$h(x, x')$	Latent comparison function, defined as $h(x, x') = f(x) - f(x')$.
$k(\cdot, \cdot)$	Base positive-definite kernel on $\mathcal{X} \times \mathcal{X}$.
$\mathcal{H}_k$	Reproducing kernel Hilbert space (RKHS) induced by kernel k.
$\|f\|_{\mathcal{H}_k}$	RKHS norm of the latent utility function f.
$k^\Delta(\cdot, \cdot)$	Dueling (difference) kernel on $\mathcal{X} \times \mathcal{X}$ defined from k.
$\mathcal{H}_{k^\Delta}$	RKHS induced by the dueling kernel k_Δ.
h_t	Regularized empirical risk minimizer estimating h at round t.
$k_t^\Delta(\cdot, \cdot)$	Posterior covariance (uncertainty) kernel for h after t observations.
$\sigma_t^2(z)$	Posterior variance proxy at pair $z = (x, x')$, given by $k_{\Delta,t}(z, z)$.
λ	Regularization parameter in kernelized estimation.
κ	Uniform bound on the inverse derivative of the link function μ.
$\Gamma(t)$	Maximum information gain after t preference queries under kernel k_Δ.
$\beta_t(\delta)$	Confidence-width multiplier ensuring high-probability error bounds.
$\tilde{h}_t$	Thompson sample from the Gaussian surrogate posterior over comparison functions.
$\tilde{f}_t$	Latent utility function corresponding to a separable Thompson sample $\tilde{h}_t$.
v_t	Exploration scale used in Thompson sampling (typically set to $\beta_t(\delta)$).
$R(T)$	Cumulative regret over T rounds.

Table 1: Summary of notation used throughout the main body of the paper.

B Background Material

This section provides standard background material on Bayesian Optimisation, included for completeness and to aid in the interpretation of our results.

B.1 Reproducing Kernel Hilbert Space (RKHS)

Let $\mathcal{X}$ be a nonempty set and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *positive-definite kernel*, i.e., for any $n \in \mathbb{N}$, any points $x_1, \ldots, x_n \in \mathcal{X}$ and any coefficients $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \;\geq\; 0.$$

The RKHS $\mathcal{H}_k$ associated with k is the unique Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying:

1. For each $x \in \mathcal{X}$, the *representer function* $k(\cdot, x)$ belongs to $\mathcal{H}_k$.
2. *Reproducing property:* For all $f \in \mathcal{H}_k$ and all $x \in \mathcal{X}$,

$$f(x) \;=\; \langle f,\, k(\cdot, x)\rangle_{\mathcal{H}_k}.$$

Intuitively, $\mathcal{H}_k$ is the completion of the linear span of $\{k(\cdot, x) : x \in \mathcal{X}\}$ with respect to the inner product induced by k. The RKHS norm $\|f\|_{\mathcal{H}_k}$ measures the smoothness of f relative to the kernel k, and plays a central role in the theoretical analysis of BO.

B.2 Mercer's Theorem

Mercer's theorem provides a spectral decomposition of continuous, symmetric, and positive-definite kernels on compact domains, and connects the kernel k with an orthonormal basis of the associated L^2 space.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous symmetric positive-definite kernel, with $\mathcal{X} \subset \mathbb{R}^d$ compact and equipped with a probability measure μ. Mercer's theorem states that there exists a sequence $\{\gamma_m\}_{m=1}^{\infty}$ of positive eigenvalues and an orthonormal set of eigenfunctions $\{\varphi_m\}_{m=1}^{\infty}$ in $L^2(\mathcal{X}, \mu)$ such that the kernel admits the expansion

$$k(x, x') \;=\; \sum_{m=1}^{\infty} \gamma_m \, \varphi_m(x) \, \varphi_m(x'), \tag{15}$$

where the series converges absolutely and uniformly on $\mathcal{X} \times \mathcal{X}$.

For the RKHS $\mathcal{H}_k$ induced by k, the eigenfunctions $\{\varphi_m\}$ form an orthonormal basis of $L^2(\mathcal{X}, \mu)$, but not necessarily of $\mathcal{H}_k$. Any function $f \in \mathcal{H}_k$ can be written as

$$f(x) \;=\; \sum_{m=1}^{\infty} w_m \sqrt{\gamma_m} \, \varphi_m(x), \tag{16}$$

for some coefficients $\{w_m\}_{m=1}^{\infty}$ with

$$\|f\|_{\mathcal{H}_k}^2 \;=\; \sum_{m=1}^{\infty} w_m^2.$$

This expansion makes explicit the connection between the geometry of $\mathcal{H}_k$ and the spectral properties of the kernel k.

B.3 Common Kernel Functions

The choice of kernel k in a GP determines the properties of the functions drawn from it. Two widely used kernels are described below.

Squared Exponential (RBF) Kernel. The squared exponential (SE), or radial basis function (RBF), kernel is given by

$$k_{\mathrm{SE}}(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|_2^2}{2\ell^2}\right),$$

where $\ell > 0$ is the characteristic lengthscale and σ_f^2 is the signal variance. This kernel is infinitely mean-square differentiable, encoding a prior belief in extreme smoothness.

Matérn Kernel. The Matérn kernel is a generalisation that allows control over the smoothness of sample paths via a parameter $\nu > 0$:

$$k_{\text{Matérn}}(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, r}{\ell}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\, r}{\ell}\right),$$

where $r = \|x - x'\|_2$, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind. When $\nu = 1/2$, the kernel reduces to the exponential kernel; when $\nu \to \infty$, it converges to the squared exponential kernel. Finite ν yields sample paths that are $\lceil \nu - 1 \rceil$ times mean-square differentiable.

C Proof of Proposition 1

By construction (see (6)), $h_t \in \mathcal{H}_{k^\Delta}$ is the minimizer of a regularized empirical risk; by the representer theorem there exist coefficients $\{\bar{\theta}_i\}_{i=1}^{t}$ such that

$$h_t(\cdot, \cdot) \;=\; \sum_{i=1}^{t} \bar{\theta}_i \, k^{\Delta}\big((\cdot, \cdot), (x_i, x_i')\big). \tag{17}$$

Since k^Δ is the *dueling* kernel induced by a base kernel k on $\mathcal{X}$,

$$k^\Delta\big((x,x'),(u,u')\big) \;=\; k(x,u)+k(x',u')-k(x,u')-k(x',u),$$

and therefore h_t is a *difference* function: there exists $g_t \in \mathcal{H}_k$ with

$$h_t(x,x') \;=\; g_t(x)-g_t(x') \qquad \forall (x,x')\in\mathcal{X}^2. \tag{18}$$

(Equivalently, $T:\mathcal{H}_k \to \mathcal{H}_{k^\Delta}$ defined by $(Tf)(x,x')=f(x)-f(x')$ is an isometry onto; (17) lies in its range.)

Let $\tilde{h}_t \sim \mathrm{GP}(h_t, k_t^\Delta)$ be the Thompson sample on $\mathcal{X}^2$. We will construct $\tilde{f}_t$ explicitly and then verify the identity. Fix an arbitrary anchor $x_0 \in \mathcal{X}$ and define

$$\tilde{f}_t(x) \;:=\; \tilde{h}_t(x,x_0), \qquad x\in\mathcal{X}.$$

Consider the (centered) discrepancy

$$S(x,x') \;:=\; \tilde{h}_t(x,x') - \big(\tilde{f}_t(x)-\tilde{f}_t(x')\big) \;=\; \tilde{h}_t(x,x') - \tilde{h}_t(x,x_0) + \tilde{h}_t(x',x_0).$$

Since $\tilde{h}_t$ is Gaussian, $S(x,x')$ is Gaussian for each (x,x'). We show that $S(x,x')$ is almost surely zero.

Mean of S. Using (18),

$$\mathbb{E}\,S(x,x') \;=\; h_t(x,x')-h_t(x,x_0)+h_t(x',x_0) \;=\; \big(g_t(x)-g_t(x')\big)-\big(g_t(x)-g_t(x_0)\big)+\big(g_t(x')-g_t(x_0)\big) \;=\; 0.$$

Variance of S. We next show $\mathrm{Var}(S(x,x'))=0$. It suffices to verify that the covariance kernel k_t^Δ inherits the dueling *separable-difference* structure. Write the usual kernel ridge–style posterior covariance (on any finite set) as

$$k_t^\Delta(\cdot,\cdot) \;=\; k^\Delta(\cdot,\cdot) \;-\; k_Z^\Delta(\cdot)^\top\big(K_Z^\Delta+\lambda\kappa I\big)^{-1}k_Z^\Delta(\cdot),$$

where $Z=\{(x_i,x_i')\}_{i=1}^t$, $K_Z^\Delta=[k^\Delta(z_i,z_j)]_{i,j=1}^t$ and $k_Z^\Delta(z)=[k^\Delta(z,z_i)]_{i=1}^t$. Each term on the right-hand side is a finite linear combination of *dueling* kernels, hence k_t^Δ is again dueling: there exists a positive semidefinite kernel k_t on $\mathcal{X}$ such that

$$k_t^\Delta\big((x,x'),(u,u')\big) \;=\; k_t(x,u)+k_t(x',u')-k_t(x,u')-k_t(x',u). \tag{19}$$

(One concrete choice is $k_t(x,u):=k_t^\Delta\big((x,x_0),(u,x_0)\big)$; positive semidefiniteness follows from that of k_t^Δ.)

Now expand the variance using (19):

$$\begin{aligned}
\mathrm{Var}\big(S(x,x')\big) &= k_t^\Delta\big((x,x'),(x,x')\big)+k_t^\Delta\big((x,x_0),(x,x_0)\big)+k_t^\Delta\big((x',x_0),(x',x_0)\big)\\
&\quad -2\,k_t^\Delta\big((x,x'),(x,x_0)\big)+2\,k_t^\Delta\big((x,x'),(x',x_0)\big)-2\,k_t^\Delta\big((x,x_0),(x',x_0)\big)\\
&= \big[k_t(x,x)+k_t(x',x')-2k_t(x,x')\big]+\big[k_t(x,x)+k_t(x_0,x_0)-2k_t(x,x_0)\big]\\
&\quad +\big[k_t(x',x')+k_t(x_0,x_0)-2k_t(x',x_0)\big]\\
&\quad -2\big[k_t(x,x)+k_t(x',x_0)-k_t(x,x_0)-k_t(x',x)\big]\\
&\quad +2\big[k_t(x,x')+k_t(x',x_0)-k_t(x,x_0)-k_t(x',x')\big]\\
&\quad -2\big[k_t(x,x')+k_t(x_0,x_0)-k_t(x,x_0)-k_t(x',x_0)\big]\\
&= 0,
\end{aligned}$$

by term-by-term cancellation. Hence $S(x,x')$ is Gaussian with zero mean and zero variance, so $S(x,x')=0$ almost surely.

Conclusion. For all $(x,x')\in\mathcal{X}^2$,

$$\tilde{h}_t(x,x') \;=\; \tilde{f}_t(x)-\tilde{f}_t(x') \qquad \text{a.s.},$$

with the explicit choice $\tilde{f}_t(x)=\tilde{h}_t(x,x_0)$ (unique up to an additive constant). This proves the proposition.

Remarks on representer structure: The use of the representer theorem above ensures the *mean* h_t admits a decomposition through k via (18). For the sampled component, one can also view finite-dimensional marginals of $\tilde{h}_t$ as living in the linear span of kernel sections $\{k^\Delta(\cdot,z_i)\}_{i=1}^t$; because k^Δ is separable through k, the difference property is preserved by posterior linear updates, yielding (19). The anchor construction then provides the pathwise identification $\tilde{f}_t(x)=\tilde{h}_t(x,x_0)$ and completes the argument.

D Proof of Theorem 1

Structure of the Proof. The structure of the proof is as follows: In Sections D.1–D.2, we develop high-probability concentration inequalities for posterior samples and define uniform events ensuring reliable uncertainty control. Section D.3 introduces an anti-concentration argument that guarantees the selection of 'unsaturated' actions with sufficient probability, thereby ensuring exploration. Finally, by combining these lemmas with information-gain bounds, the analysis (Sections D.4–D.6) derives the cumulative regret rate $\tilde{O}(\beta_T(\delta)\sqrt{T\Gamma(T)})$, matching that of standard GP-TS under scalar feedback. Before beginning the proof, we briefly comment on continuous action spaces.

Continuous action spaces We emphasize that the statement in Theorem 1 is valid for both discrete *and* continuous action spaces. For technical reasons, in the case of continuous action spaces we need to follow existing work analyzing TS (for vanilla BO), by discretizing the action space in each round. In particular, we will pick a decision set $\mathcal{C}_t \subset \mathcal{X}$. Namely, the actions in PF-TS will be selected as

$$x_t \in \arg\max_{x \in \mathcal{C}_t} \tilde{h}_t^{(1)}(x, x_0), \qquad x_t' \in \arg\max_{x \in \mathcal{C}_t} \tilde{h}_t^{(2)}(x, x_0).$$

For the remainder of the proof we will consider this latter setting of a continuous action space. However, we note that the same order regret is achievable for finite, constant size action spaces with only minor modifications of the proof. We will choose the same discretization as (Chowdhury and Gopalan, 2017). Namely, without loss of generality, let $\mathcal{X} = [0, w]^d \subset \mathbb{R}^d$. Then let the decision set $\mathcal{C}_t$ be evenly spaced across the action space with

$$|\mathcal{C}_t| = (BGwdt^2)^d,$$

where B is the norm bound from Assumption 1, w is the width of the action space in each dimension, d is the number of dimensions, and we define

$$G = \sup_{x \in \mathcal{X}} \sup_{j \in [d]} \left(\frac{\partial^2 k(p,q)}{\partial p_j \partial q_j}\Big|_{p=q=x} \right).$$

This choice helps us control the error caused by the discretisation. In particular, let $[x]_t$ be the closest element of the decision set to $x \in \mathcal{X}$. Then, as discussed in (Section 4.2, Chowdhury and Gopalan, 2017), selecting a decision set of this size guarantees for all $x \in \mathcal{X}$ that

$$|f(x) - f([x]_t)| < 1/t^2. \tag{20}$$

This will ensure that the decision set is sufficiently fine as to not accumulate excessive regret, as described blow. Finally, regardless of the action space we define the exploration scale

$$v_t = \beta_t(\delta).$$

Discrete regret & Lipschitz reduction We begin by reducing the regret expressed in terms of preference probabilities to a form involving utility differences, which are easier to analyze. Recall that under the BTL model the probability of preference is given by $\Pr(x \succ x') = \mu\big(h(x, x')\big)$ with an increasing L_μ–Lipschitz link. Because the regret definition depends on these probabilities, directly controlling it is inconvenient. However, the Lipschitz property of the link function allows us to relate differences in preference probabilities to differences in the latent utility function. This enables us to bound the instantaneous dueling regret in terms of utility gaps.

In addition, following standard analyses of Thompson sampling for continuous action spaces, we work with a discretized decision set and introduce a notion of discrete regret with respect to the best action in this set. We denote the instantaneous dueling regret with respect to the global optimum by r_t and the discrete regret with respect to the best action in the decision set by Δ_t. Using the Lipschitz property of the link function, we obtain the following relationship.

$$\begin{aligned} r_t &:= \frac{\mu\big(h(x^\star, x_t)\big) + \mu\big(h(x^\star, x_t')\big) - 1}{2} \;\le\; \frac{L_\mu}{2}\Big(h(x^\star, x_t) + h(x^\star, x_t')\Big), \\ \Delta_t &:= \frac{\mu\big(h([x^\star]_t, x_t)\big) + \mu\big(h([x^\star]_t, x_t')\big) - 1}{2} \;\le\; \frac{L_\mu}{2}\Big(h([x^\star]_t, x_t) + h([x^\star]_t, x_t')\Big). \end{aligned} \tag{21}$$

Furthermore, using the Lipschitz property again and as a consequence of the fineness of the decision set (20) we have

$$\begin{aligned}|r_t - \Delta_t| &\leq \frac{L_\mu}{2}|h(x^*, x_t) + h(x^*, x_t') - h([x^*]_t, x_t) + h([x^*]_t, x_t')| \\ &\leq L_\mu |f(x^*) - f([x^*]_t)| \\ &\leq \frac{L_\mu}{t^2}\end{aligned} \tag{22}$$

D.1 Concentration of a posterior sample

We have the following general concentration result for a posterior sample of a difference function, which will be useful for our analysis. Intuitively, it shows that a Thompson sample remains close to the posterior mean, with deviations controlled by the posterior standard deviation. This allows us to relate sampled comparison functions to the estimated function in a uniform way over the candidate set. Here we denote the filtration $\mathcal{F}t$ as the history up until time t, namely $Ht-1$ defined in Section 2.

Lemma 2. *Consider a posterior sample of the difference function $\tilde{h}_t \sim GP(h_t, v_t^2 k_t^\Delta)$. Let $\mathcal{C}$ be a finite subset of the action space $\mathcal{X}$ and $\delta \in (0,1)$, then*

$$\mathbb{P}\left(\forall x, x' \in \mathcal{C}, |\tilde{h}_t(x,x') - h_t(x,x')| \leq v_t\sqrt{2\log\left(|\mathcal{C}|^2/\delta\right)}\sigma_t(x,x') \middle| \mathcal{F}_t\right) \geq 1-\delta$$

Proof. As discussed in Appendix B, we know that for any $x, x' \in D'$ that the posterior sample at that pair has distribution $\tilde{h}_t(x,x') \sim N(h_t(x,x'), v_t^2\sigma_t^2(x,x'))$. Hence, from (Lemma B4, Hoffman et al., 2013) we know that for any $x, x' \in D'$.

$$\mathbb{P}\left(|\tilde{h}_t - h_t(x,x')| \leq v_t\sqrt{2\log\left(1/\delta\right)}\sigma_t(x,x') \middle| \mathcal{F}_t\right) \geq 1-\delta.$$

Hence, by taking a union bound over all $|\mathcal{C}|^2$ pairs of actions from $\mathcal{C}$, we attain the desired result. □

D.2 Uniform high probability events

We now introduce two high probability events which uniformly control the distance between the estimated difference function h_t and the true difference function h, as well as the difference between the estimated difference function h_t and each of the sampled difference functions $\tilde{h}_t^{(i)}$. Namely, define the following events, each required to hold *for all* $z \in \mathcal{X}^2$:

$$E_1: \ |h(z) - h_t(z)| \leq \beta_t(\delta)\,\sigma_t(z) \ \text{ for all } t \leq T, \quad E_{2,t}: \ \left|\tilde{h}_t^{(i)}(z) - h_t(z)\right| \leq \beta_t(\delta)\sqrt{2\log\left(|\mathcal{C}_t|^2 t^2\right)}\,\sigma_t(z) \ \text{ for } i \in \{1,2\}.$$

By Lemma 1 and Lemma 2, we know that these events hold with high probability, namely

$$\mathbb{P}\left(E_1\right) \geq 1-\delta \tag{23}$$

$$\mathbb{P}\left(E_{2,t}|\mathcal{F}_t\right) \geq 1 - \frac{2}{t^2} \tag{24}$$

Furthermore, for ease of notation, we define the constant

$$d_t = \beta_t(\delta)\left(1 + \sqrt{2\log\left(|\mathcal{C}_t|^2 t^2\right)}\right) \tag{25}$$

and observe that, under both events $E_1 \cap E_{2,t}$ we have for all $t \leq T$, all $z \in \mathcal{X}^2$, and for $i \in \{1,2\}$

$$\left|\tilde{h}_t^{(i)}(z) - h(z)\right| \leq d_t\sigma_t(z). \tag{26}$$

D.3 Anti-concentration ensures unsaturated selection

To analyze the exploration behavior of the algorithm, we partition the action space into saturated and unsaturated points relative to a chosen anchor. Intuitively, a point is saturated if its posterior uncertainty is small compared to its sub-optimality gap, meaning the algorithm is already confident it is suboptimal. Unsaturated points are those for which uncertainty remains large enough that they may still plausibly be optimal. More formally, we define the saturated set $S_t(a)$ as the set of actions whose posterior uncertainty with an anchor point a is smaller than their discrete sub-optimality gap. Namely

$$S_t(a) \;:=\; \big\{x \in \mathcal{C}_t : \; h([x^\star]_t, x) \;>\; d_t \sigma_t(x,a)\big\}. \tag{27}$$

We will refer to actions as unsaturated (with respect to a) when they are not in the saturated set $S_t(a)$.

We can then extend the Gaussian anti-concentration arguments used in (Chowdhury and Gopalan, 2017) in order to prove that we will always play an unsaturated action with some probability, through the following two lemmas. Intuitively, saturated points are those whose sub-optimality gap already exceeds their posterior uncertainty, meaning the algorithm has enough evidence that they are suboptimal. In contrast, unsaturated points still have sufficiently large uncertainty relative to their gap and therefore require further exploration. The following results show that the randomness of the Thompson sample ensures that such points are selected with non-negligible probability: Lemma 3 establishes a constant-probability anti-concentration property of the posterior sample, and Lemma 4 uses this to show that the maximizer of the sampled function lies outside the saturated set with constant probability. This guarantees that the algorithm continues to explore points whose uncertainty remains significant, which will be a key ingredient for controlling the regret in the subsequent analysis.

Lemma 3. *For any filtration $\mathcal{F}_t$ in which E_1 is true, for any posterior sample $\tilde{h}_t$ and for any $x, x' \in \mathcal{X}$ we can lower bound the probability*

$$\mathbb{P}\left(\tilde{h}_t(x,x') > h(x,x') \middle| E_1\right) \geq p$$

for some universal constant $p = \frac{1}{4e\sqrt{\pi}}$.

Proof. We first rewrite the event in terms of a standard normal random variable, recalling that a posterior sample for the difference function at $x, x' \in \mathcal{X}$ is distributed $\tilde{h}_t(x,x') \sim N(h_t(x,x'), v_t^2\sigma_t^2(x,x'))$. Let $\theta_t = \frac{|h(x,x') - h_t(x,x')|}{v_t\sigma_t(x,x')}$, then we can write

$$\begin{aligned}
\mathbb{P}\left(\tilde{h}_t(x,x') > h(x,x') \middle| E_1\right) &\geq \mathbb{P}\left(\frac{\tilde{h}_t(x,x') - h_t(x,x')}{v_t\sigma_t(x,x')} > \frac{h(x,x') - h_t(x,x')}{v_t\sigma_t(x,x')} \middle| E_1\right) \\
&\geq \mathbb{P}\left(\frac{\tilde{h}_t(x,x') - h_t(x,x')}{v_t\sigma_t(x,x')} > \frac{|h(x,x') - h_t(x,x')|}{v_t\sigma_t(x,x')} \middle| E_1\right) \\
&\geq \frac{1}{4\theta_t\sqrt{\pi}} \exp(-\theta_t^2) \\
&\geq \frac{1}{4e\sqrt{\pi}}
\end{aligned}$$

The penultimate inequality holds from via a Gaussian anti-concentration (Lemma 7, Chowdhury and Gopalan, 2017). The final inequality holds since, by definition under event E_1 (see Section D.2),

$$\theta_t = \frac{|h(x,x') - h_t(x,x')|}{v_t\sigma_t(x,x')} \leq 1$$

□

Lemma 4. *For any filtration $\mathcal{F}_t$ such that E_1 is true and for any $a, a' \in \mathcal{C}_t$,*

$$\mathbb{P}\left[x_t \in \mathcal{C}_t \setminus S_t(a) \mid \mathcal{F}_t\right] \geq p - \frac{2}{t^2}$$
$$\mathbb{P}\left[x'_t \in \mathcal{C}_t \setminus S_t(a') \mid \mathcal{F}_t\right] \geq p - \frac{2}{t^2},$$

with $p = \frac{1}{4e\sqrt{\pi}}$.

Proof. We prove the first statement. The second statement is true by a symmetric argument. By anchor independence (see Section 4.1), we know that the first action x_t is the maximizer of the single argument search

$$x_t = \arg\max_{x \in \mathcal{C}_t} \tilde{h}_t^{(1)}(x, a)$$

for any $a \in \mathcal{C}_t$, regardless of what anchor point is used in the implementation of the algorithm x_0.

If $\tilde{h}_t^{(1)}([x^\star]_t, a)$ is greater than $\tilde{h}_t^{(1)}(x, a)$ for all saturated points at round t (namely, $\tilde{h}_t^{(1)}([x^\star]_t, a) > \tilde{h}_t^{(1)}(x, a), \forall x \in S_t(a)$), then one of the unsaturated points in $\mathcal{C}_t$ (which always includes $[x^\star]_t$) must be played and hence $x_t \in \mathcal{C}_t \setminus S_t(a)$. This implies

$$\mathbb{P}\left[x_t \in \mathcal{C}_t \setminus S_t(a) \mid \mathcal{F}_t\right] \geq \mathbb{P}\left[\tilde{h}_t^{(1)}([x^\star]_t, a) > \tilde{h}_t^{(1)}(x, a), \forall x \in S_t \mid \mathcal{F}_t\right]. \tag{28}$$

Now, by the definition of the saturated set $S_t(a)$, we know $h([x^*]_t, x) > d_t \sigma_{t-1}(x, a)$, for all $x \in S_t(a)$. Furthermore, if both the events E_1 and $E_{2,t}$ are true, then $\tilde{h}_t(x, a) \leq h(x, a) + d_t \sigma_{t-1}(x, a)$, for all $x \in \mathcal{C}_t$. Thus, if events E_1 and $E_{2,t}$ hold and using the separability of h, we have for all $x \in S_t(a)$, $\tilde{h}_t(x, a) < h(x, a) + h([x^*]_t, x) = h([x^\star]_t, a)$. Therefore, for any filtration $\mathcal{F}_t$ such that E_1 is true, either $E_{2,t}$ is false, or else for all $x \in S_t(a)$, $\tilde{h}_t^{(1)}(x, a) < h([x^\star]_t, a)$. Hence, for any $\mathcal{F}_t$ such that E_1 is true,

$$\begin{aligned}
&\mathbb{P}\left[\tilde{h}_t^{(1)}([x^\star]_t, a) > \tilde{h}_t^{(1)}(x, a), \forall x \in S_t(a) \mid \mathcal{F}_t\right] \\
&\quad \geq \mathbb{P}\left[\tilde{h}_t^{(1)}([x^\star]_t, a) > h([x^\star]_t, a) \mid \mathcal{F}_t\right] - \mathbb{P}\left[E_{2,t}^C \mid \mathcal{F}_t\right] \\
&\quad \geq p - \frac{2}{t^2}
\end{aligned}$$

For the final inequality we have used Lemma 3 and (24). Now the proof follows by combining this final inequality with Equation (28). □

D.4 Control via least-variance unsaturated points

Having established that unsaturated points are selected with constant probability, we next relate the uncertainty at the played action to the smallest uncertainty among unsaturated actions. For this purpose, we introduce the notion of a least-variance unsaturated point, which is the unsaturated action whose posterior variance with respect to the anchor is minimal. Intuitively, this point represents the most "well-understood" action among those that still require exploration. The following argument shows that, because the algorithm selects unsaturated actions with constant probability, the expected posterior variance of the chosen action is lower bounded by a constant fraction of the variance of the least-variance unsaturated point. This connection allows us to control the uncertainty of the played actions in terms of the smallest remaining uncertainty among exploratory candidates, which will be crucial when bounding the regret in the subsequent steps of the proof.

Formally, for any potential anchor point $a \in \mathcal{C}_t$, define a least-variance unsaturated point as

$$\bar{x}_t(a) \in \arg\min_{x \in \mathcal{X} \setminus S_t(a)} \sigma_t(x, a).$$

Note that such a point always exists because we have $x^* \in \mathcal{C}_t \setminus S_t(a)$ for any $a \in \mathcal{C}_t$.

Lemma 5. *Condition on any filtration $\mathcal{F}_{t-1}$ such that E_1 holds. Let $a_t := x'_t$ and $a'_t := x_t$. Then there exists an absolute constant $p \in (0,1)$ (the one from Lemma 4) such that*

$$\mathbb{E}\big[\sigma_t(\bar{x}_t(a_t), a_t) \;\big|\; \mathcal{F}_t\big] \;\leq\; \frac{1}{p - \frac{2}{t^2}}\, \mathbb{E}\big[\sigma_t(x_t, a_t) \;\big|\; \mathcal{F}_t\big], \tag{29}$$

$$\mathbb{E}\big[\sigma_t(\bar{x}_t(a'_t), a'_t) \;\big|\; \mathcal{F}_t,\big] \;\leq\; \frac{1}{p - \frac{2}{t^2}}\, \mathbb{E}\big[\sigma_t(x'_t, a'_t) \;\big|\; \mathcal{F}_t\big]. \tag{30}$$

Proof. We prove (29); (30) is identical by symmetry. Note that, on the event

$$U_t(a_t) \;:=\; \{\, x_t \notin S_t(a_t) \,\},$$

we have by definition of $\bar{x}_t(a_t)$ that $\sigma_t(x_t, a_t) \;\geq\; \sigma_t(\bar{x}_t(a_t), a_t)$.

Taking conditional expectations gives us

$$\begin{aligned}
\mathbb{E}\left[\sigma_t(x_t, a_t)\middle|\mathcal{F}_t,\right] &\geq \mathbb{E}\left[\sigma_t(x_t, a_t)\mathbb{1}\{U_t(a_t)\}\middle|\mathcal{F}_t\right] \\
&= \sum_{a \in \mathcal{C}_t} \mathbb{E}\left[\sigma_t(x_t, a_t)\mathbb{1}\{U_t(a_t)\}\middle|\mathcal{F}_t, a_t = a\right] \Pr\left(a_t = a\middle|\mathcal{F}_t\right) \\
&= \sum_{a \in \mathcal{C}_t} \mathbb{E}\left[\sigma_t(x_t, a)\mathbb{1}\{U_t(a)\}\middle|\mathcal{F}_t, a_t = a\right] \Pr\left(a_t = a\middle|\mathcal{F}_t\right) \\
&= \sum_{a \in \mathcal{C}_t} \mathbb{E}\left[\sigma_t(\bar{x}_t(a), a)\mathbb{1}\{U_t(a)\}\middle|\mathcal{F}_t, a_t = a\right] \Pr\left(a_t = a\middle|\mathcal{F}_t\right) \\
&= \sum_{a \in \mathcal{C}_t} \sigma_t(\bar{x}_t(a), a) \Pr\left[U_t(a)\middle|\mathcal{F}_t, a_t = a\right] \Pr\left(a_t = a\middle|\mathcal{F}_t\right) \\
&\geq \left(p - \frac{2}{t^2}\right) \sum_{a \in \mathcal{C}_t} \sigma_t(\bar{x}_t(a), a) \Pr\left(a_t = a\middle|\mathcal{F}_t\right) \\
&= \left(p - \frac{2}{t^2}\right) \mathbb{E}\left[\sigma_t(\bar{x}_t(a_t), a_t)\middle|\mathcal{F}_t\right]
\end{aligned}$$

Where in the second line we have used the law of total expectation, the fourth line comes from the definition of $\bar{x}_t(a_t)$ as noted at the beginning of the prof, in the fifth line we can take $\sigma_t(\bar{x}_t(a), a)$ since it is deterministic given the conditions of the expectation, and in the penultimate line we have used Lemma 4. Rearranging yields (29). □

D.5 Using anchor independence

We now present a lemma that exploits the anchor-independence property of PF-TS. In particular, Proposition 1 implies that the actions selected by the algorithm are invariant to the choice of anchor. Consequently, for the purpose of the analysis we are free to select anchors that simplify the regret decomposition. In particular, when bounding the instantaneous regret in round t, we choose the anchor points $a_t = x'_t$ for the first action and $a'_t = x_t$ for the second. This choice allows us to relate the regret directly to the posterior uncertainty of the played pair.

Lemma 6. *In round t of the PF-TS algorithm, when both good events hold $E_1 \cap E_{2,t}$ we have that the instantaneous regret can be upper bounded by*

$$r_t \leq \frac{L_\mu d_t}{4}\left(2\sigma_t(\bar{x}_t, x'_t) + 2\sigma_t(\bar{x}'_t, x_t) + 2\sigma_t(x_t, x'_t)\right) + \frac{L_\mu}{t^2}$$

Proof. Firstly, we can upper bound the instantaneous regret with the discrete instantaneous regret with some error, as shown in equation (22),

$$r_t \leq \Delta_t + \frac{L_\mu}{t^2}.$$

For the rest of the proof we focus on upper bounding Δ_t. We start by defining the anchor points $a_t = x'_t$ and $a'_t = x_t$ adding and subtracting the utility of the two least-variance unsaturated points with respect to the two

anchor points, namely $\bar{x}_t = \bar{x}_t(a_t)$ and $\bar{x}'_t = \bar{x}_t(a'_t)$. Recall from the Lipschitz reduction in equation (21) that

$$\begin{aligned}
\frac{2\Delta_t}{L_\mu} &\leq f([x^*]_t) - \frac{f(x_t) + f(x'_t)}{2} \\
&= (f([x^*]_t) - \frac{f(\bar{x}_t) + f(\bar{x}'_t)}{2} + \frac{f(\bar{x}_t) + f(\bar{x}'_t)}{2} - \frac{f(x_t) + f(x'_t)}{2} \\
&= \frac{h([x^*]_t, \bar{x}_t) + h([x^*]_t, \bar{x}'_t)}{2} + \frac{f(\bar{x}_t) + f(\bar{x}'_t)}{2} - \frac{f(x_t) + f(x'_t)}{2}
\end{aligned}$$

Since we know that both $\bar{x}_t$ and $\bar{x}'_t$ are unsaturated with respect to $S_t(a_t)$ and $S_t(a'_t)$, by definition (27) we have that

$$\leq \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t)\right) + \frac{f(\bar{x}_t) + f(\bar{x}'_t)}{2} - \frac{f(x_t) + f(x'_t)}{2}$$

Now, let's add and subtract both $f(a_t)$ and $f(a'_t)$

$$\begin{aligned}
&= \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t)\right) + \frac{f(\bar{x}_t) - f(a_t) + f(\bar{x}'_t) - f(a'_t)}{2} - \frac{f(x_t) - f(a_t) + f(x'_t) - f(a'_t)}{2} \\
&= \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t)\right) + \frac{h(\bar{x}_t, a_t) + h(\bar{x}'_t, a'_t)}{2} - \frac{h(x_t, a_t) + h(x'_t, a'_t)}{2}
\end{aligned}$$

Now, we know from (26) that under event $E_1 \cap E_{2,t}$, our posterior sample is $d_t\sigma_t$ close to the true difference function. Hence,

$$\begin{aligned}
&\leq \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t)\right) \\
&\quad + \frac{\tilde{h}^{(1)}(\bar{x}_t, a_t) + d_t\sigma_t(\bar{x}_t, a_t) + \tilde{h}^{(2)}(\bar{x}'_t, a'_t) + d_t\sigma_t(\bar{x}'_t, a'_t)}{2} \\
&\quad - \frac{\tilde{h}^{(1)}(x_t, a_t) - d_t\sigma_t(x_t, a_t) + \tilde{h}^{(2)}(x'_t, a'_t) - d_t\sigma_t(x'_t, a'_t)}{2}
\end{aligned}$$

Now we use the selection rule of PF-TS together with the anchor-independence property (Proposition 1). By anchor independence, the actions chosen by the algorithm are invariant to the anchor used in the maximization. Hence, although the algorithm is implemented with a fixed anchor x_0, we may equivalently view x_t and x'_t as the maximizers of the sampled functions with respect to the arbitrary anchors a_t and a'_t, respectively. In particular, $x_t = \arg\max_x \tilde{h}^{(1)}(x, a_t)$ and $x'_t = \arg\max_x \tilde{h}^{(2)}(x, a'_t)$. Therefore the sampled values at the selected actions dominate those at any other candidate points, including $\bar{x}_t$ and $\bar{x}'_t$. Substituting these inequalities into the previous expression allows the sampled-function terms to be upper bounded by their corresponding uncertainty terms, yielding

$$\leq \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t) + \sigma_t(\bar{x}_t, a_t) + \sigma_t(\bar{x}'_t, a'_t) + \sigma_t(x_t, a_t) + \sigma_t(x'_t, a'_t)\right).$$

Now, by picking $a_t = x'_t$ and $a'_t = x_t$, we can write this as

$$\begin{aligned}
\frac{2\Delta_t}{L_\mu} &\leq \frac{d_t}{2}\left(\sigma_t(\bar{x}_t, x'_t) + \sigma_t(\bar{x}'_t, x_t) + \sigma_t(\bar{x}_t, x'_t) + \sigma_t(\bar{x}'_t, x_t) + \sigma_t(x_t, x'_t) + \sigma_t(x'_t, x_t)\right) \\
&= \frac{d_t}{2}\left(2\sigma_t(\bar{x}_t, x'_t) + 2\sigma_t(\bar{x}'_t, x_t) + \sigma_t(x_t, x'_t) + \sigma_t(x'_t, x_t)\right) \\
&= \frac{d_t}{2}\left(2\sigma_t(\bar{x}_t, x'_t) + 2\sigma_t(\bar{x}'_t, x_t) + 2\sigma_t(x_t, x'_t)\right)
\end{aligned}$$

Where the final line holds by symmetry of σ_t. The proof is then complete. □

D.6 Putting everything together

We now combine the results from the previous subsections to obtain the final regret bound. Lemma 6 shows that the instantaneous regret can be controlled by posterior uncertainty terms involving the played pair (x_t, x'_t) and the least-variance unsaturated points. The role of Lemma 5 is to relate the uncertainty of these unsaturated points to the uncertainty of the played pair: since the algorithm selects unsaturated actions with constant probability, the expected posterior variance of the chosen actions cannot be much smaller than that of the least-variance unsaturated points. Consequently, the first two terms in the bound of Lemma 6 can be controlled in expectation by $\sigma_t(x_t, x'_t)$.

This reduction is important because the cumulative sum of posterior uncertainties $\sigma_t(x_t, x'_t)$ can be bounded using standard information-theoretic arguments. In particular, the total uncertainty accumulated by the queried pairs is controlled by the maximum information gain under the dueling kernel. The following lemma formalizes this relationship and allows us to convert the variance-based regret bound into a cumulative regret bound in terms of the information gain.

Lemma 7. *We can update the conditional expectation of the instantaneous regret as follows, for any filtration $\mathcal{F}_t$ where E_1 holds.*

$$\mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] \leq C' L_\mu d_t \mathbb{E}\left[\sigma_t(x_t, x'_t)|\mathcal{F}_{t-1}\right] + \frac{L_\mu + 1}{t^2},$$

where C' is a universal constant.

Proof. By taking the conditional expectation of the instantaneous regret and multiplying by $1 = (\mathbb{1}\{E_{2,t}\} + \mathbb{1}\{E_{2,t}^C\})$, we get

$$\begin{aligned}
\mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[r_t \mathbb{1}\{E_{2,t}\}|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[r_t \mathbb{1}\{E_{2,t}^C\}|\mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[\frac{L_\mu d_t}{4}\left(2\sigma_t(\bar{x}_t, x'_t) + 2\sigma_t(\bar{x}'_t, x_t) + 2\sigma_t(x_t, x'_t)\right) + \frac{L_\mu}{t^2}\right] + \mathbb{E}\left[r_t \mathbb{1}\{E_{2,t}^C\}|\mathcal{F}_{t-1}\right] \\
&\leq \left(\frac{L_\mu d_t}{p - \frac{1}{t^2}} + \frac{L_\mu d_t}{2}\right)\mathbb{E}\left[\sigma_t(x_t, x'_t)|\mathcal{F}_{t-1}\right] + \frac{L_\mu}{t^2} + \mathbb{E}\left[r_t \mathbb{1}\{E_{2,t}^C\}|\mathcal{F}_{t-1}\right] \\
&\leq \left(\frac{L_\mu d_t}{p - \frac{1}{t^2}} + \frac{L_\mu d_t}{2}\right)\mathbb{E}\left[\sigma_t(x_t, x'_t)|\mathcal{F}_{t-1}\right] + \frac{L_\mu}{t^2} + \frac{1}{2}\Pr\left[\{E_{2,t}^C\}|\mathcal{F}_{t-1}\right] \\
&\leq \left(\frac{L_\mu d_t}{p - \frac{1}{t^2}} + \frac{L_\mu d_t}{2}\right)\mathbb{E}\left[\sigma_t(x_t, x'_t)|\mathcal{F}_{t-1}\right] + \frac{L_\mu}{t^2} + \frac{1}{t^2} \\
&\leq C' L_\mu d_t \mathbb{E}\left[\sigma_t(x_t, x'_t)|\mathcal{F}_{t-1}\right] + \frac{L_\mu}{t^2} + \frac{1}{t^2}
\end{aligned}$$

Where the second inequality uses Lemma 6, the third inequality uses Lemma 5 and the symmetry of σ_t, the fourth inequality holds since $r_t \leq 1/2$, the penultimate inequality uses equation (24), and the final inequality holds with $C' = 85$. □

For the remainder of the proof we use the upper bound on the instantaneous regret established in Lemma 7 to obtain a high-probability bound for the cumulative regret $R(T)$. Intuitively, we first control the expected instantaneous regret in terms of posterior uncertainty, and then convert this expected bound into a high-probability statement for the cumulative regret using a martingale concentration argument.

To do so, we decompose the cumulative regret into its conditional expectation and a martingale difference sequence capturing the stochastic fluctuations around this expectation. Lemma 8 formalizes this decomposition and identifies the corresponding martingale difference sequence. Lemma 9 then provides a concentration bound for this martingale term. Finally, Lemma 10 combines these results with the expected regret bound from Lemma 7 to obtain a high-probability bound on the cumulative regret.

We start by defining the following quantities.

$$\bar{r}_t = r_t \mathbb{1}\{E_1\}$$
$$X_t = \bar{r}_t - C' L_\mu d_t \sigma_t(x_t, x'_t) - \frac{L_\mu + 1}{t^2}$$
$$Y_t = \sum_{s=1}^{t} X_s$$

It is then straightforward to show that the process Y_t is a super martingale.

Lemma 8. *The process $(Y_t; t = 0, \dots, T)$ is a super martingale process with respect to filtration $\mathcal{F}_t$*

Proof. From Lemma 7 we know that

$$\mathbb{E}\left[\bar{r}_t | \mathcal{F}_{t-1}\right] \leq C' L_\mu d_t \mathbb{E}\left[\sigma_t(x_t, x'_t) | \mathcal{F}_{t-1}\right] + \frac{L_\mu + 1}{t^2}.$$

Hence, we immediately have

$$\mathbb{E}\left[Y_t - Y_{t-1} | \mathcal{F}_{t-1}\right] \leq 0,$$

namely the process $(Y_t; t = 0, \dots, T)$ is a super martingale. □

Lemma 9. *Let $\delta \in (0, 1/2)$. With probability greater than $1 - 2\delta$ the total cumulative regret can be upper bounded by*

$$R_T = \sum_{t=1}^{t} r(t) = \tilde{\mathcal{O}}\left(\beta_t(\delta) \sum_{t=1}^{T} \sigma_t(x_t, x'_t) + \sqrt{\log(1/\delta) T}\right)$$

Proof. By definition of Y_t and X_t, we have

$$\begin{aligned}
|Y_t - Y_{t-1}| &= |X_t| \\
&\leq |\bar{r}_t| + C' L_\mu d_t \sigma_t(x_t, x'_t) + \frac{L_\mu + 1}{t^2} \\
&\leq \frac{1}{2} + C' L_\mu d_t \sigma_t(x_t, x'_t) + \frac{L_\mu + 1}{t^2} \\
&\leq \frac{1}{2} + C' L_\mu d_t + \frac{L_\mu + 1}{t^2}
\end{aligned}$$

Hence, we can apply the Azuma-Hoeffding inequality (Lemma 11, Chowdhury and Gopalan, 2017) to attain the following with probability greater than $1 - \delta$

$$\begin{aligned}
\sum_{t=1}^{T} \bar{r}_T &\leq \sum_{t=1}^{T} C' L_\mu d_t \sigma_t(x_t, x'_t) + \sum_{t=1}^{T} \frac{L_\mu + 1}{t^2} + \sqrt{2 \log(1/\delta) \sum_{t=1}^{T} \left(\frac{1}{2} + C' L_\mu d_t + \frac{L_\mu + 1}{t^2}\right)^2} \\
&\leq C' L_\mu d_T \sum_{t=1}^{T} \sigma_t(x_t, x'_t) + (L_\mu + 1)\frac{\pi^2}{6} + \sqrt{2 \log(1/\delta) \cdot T \left(\frac{1}{2} + C' L_\mu d_T + L_\mu + 1\right)^2} \\
&= C' L_\mu d_T \sum_{t=1}^{T} \sigma_t(x_t, x'_t) + (L_\mu + 1)\frac{\pi^2}{6} + \left(\frac{1}{2} + C' L_\mu d_T + L_\mu + 1\right) \sqrt{2T \log(1/\delta)} \qquad (31)
\end{aligned}$$

Where in the second inequality we have used that $d_t \leq d_T$ and calculated the sums. Now, since we have defined $\bar{r}_t = r_t \mathbb{1}\{E_1\}$, we know that under event E_1 we have $\bar{r}_t = r_t$ for all $t \geq 1$. Furthermore, from (23), we know event E_1 holds with probability at least $1 - \delta$. Hence by taking a union bound over events E_1 and (31), the following holds with probability greater than $1 - 2\delta$

$$\sum_{t=1}^{T} r_t \leq C' L_\mu d_T \sum_{t=1}^{T} \sigma_t(x_t, x'_t) + (L_\mu + 1)\frac{\pi^2}{6} + \left(\frac{1}{2} + C' L_\mu d_T + L_\mu + 1\right) \sqrt{2T \log(1/\delta)}$$

Then plugging in our definition for d_t from (25) yields the desired result. □

Lemma 10. *The sum of posterior standard deviation for the action pairs played from rounds $t = 1$ to T can be upper bounded by*

$$\sum_{t=1}^{T} \sigma_t(x_t, x_t') \leq \sqrt{\frac{8T\,\Gamma(T)}{\log(1 + 4(\lambda\kappa)^{-1})}}$$

Proof. By the Cauchy Schwarz inequality we know that

$$\sum_{t=1}^{T} \sigma_t(x_t, x_t') \leq \sqrt{T \sum_{t=1}^{T} \sigma_t^2(x_t, x_t')}.$$

Note that from (Lemma 14 Pásztor et al., 2024), we have the following inequality for the sum of variances.

$$\sum_{t=1}^{T} \sigma_t^2(x_t, x_t') \leq \frac{8\,\Gamma(T)}{\log(1 + 4(\lambda\kappa)^{-1})}$$

Plugging this in to our first equation gives us the desired result. □

Finally, combining the bound on the sum of posterior standard deviations from Lemma 10 with the high-probability bound on the cumulative regret established in Lemma 9 yields the desired regret guarantee for Theorem 1. In particular, with probability at least $1 - 2\delta$, the cumulative regret satisfies

$$R_T = \tilde{\mathcal{O}}\Big(\beta_T(\delta)\sqrt{T\Gamma(T)} + \sqrt{T\log(1/\delta)}\Big)\,.$$

This completes the proof of Theorem 1.

E Additional Experiments

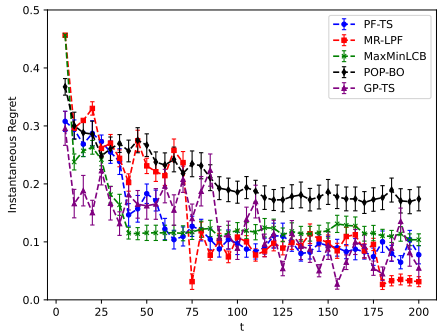
In this section, we present additional experiments that further support the findings of the main paper. In addition to the common Bayesian optimization benchmark and real-world catalyst design task considered in Section 5, we evaluate the proposed method on two new objectives that capture complementary aspects of preference-based optimization.

The first objective is a synthetic function drawn from the reproducing kernel Hilbert space (RKHS) of a Matérn-5/2 kernel. The function is generated following the procedure of Kayal et al. (2025) and provides a controlled setting that aligns closely with the kernelized assumptions underlying BOHF methods. The second objective is a hyperparameter tuning task involving the optimization of five hyperparameters (batch size, number of units per layer, learning rate, momentum, and weight decay) for a four-layer neural network trained on the MNIST classification task, using data from the LCBench benchmark (Zimmer et al., 2021). This setting models noisy preference feedback obtained from partial training runs and reflects a realistic scenario for scalable hyperparameter optimization.

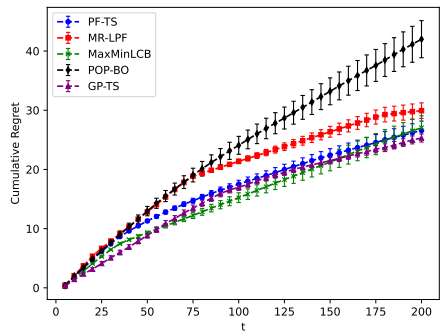
Across both objectives, we observe trends consistent with the results reported for the Catalyst and Ackley tasks. In particular, as shown in Figure 4, PF-TS exhibits a faster reduction in instantaneous regret during the early stages of optimization compared to MR-LPF, which leads to a substantial improvement in cumulative regret after 200 iterations. Final average cumulative regret values, reported in the rebuttal, further confirm that PF-TS achieves competitive or improved performance relative to existing methods on both synthetic and practical preference-based optimization problems.

F Experimental Details

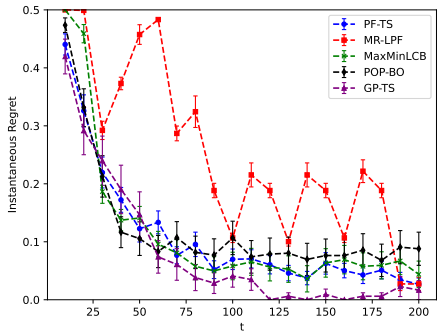
In this section we provide more details of the experiments run and displayed in the main paper. We also provide any additional, supporting results.



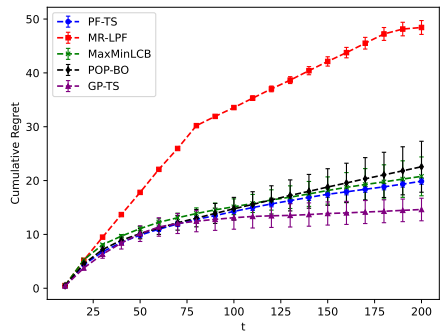
(a) **(HPO)** Instantaneous Regret



(b) **(HPO)** Cumulative Regret



(c) **(RKHS)** Instantaneous Regret



(d) **(RKHS)** Cumulative Regret

Figure 4: Regret of PF-TS, MR-LPF, MaxMinLCB, POP-BO, and GP-TS. Mean over 30 runs; shaded bands show ± 1 standard error. (a,b) Hyperparameter tuning objective ($T = 200$). (c,d) RKHS optimisation objective ($T = 200$). Left column: instantaneous (simple) regret. Right column: cumulative regret (2).

F.1 Ackley Function Experiments

The Ackley function considered in Section 5 is defined as

$$\text{Ackley}(x) = -a \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\right) + a + \exp(1) \tag{32}$$

with constants $a = 20$, $b = 0.2$, $c = 2\pi$ and d is the number of dimensions in the action space.

In Bayesian Optimization we oftentimes try to minimize the function, however in this problem formulation we try to maximize the utility. Hence, in our implementation, we flip the Ackley function and try to find the maxima. This is illustrated in Figure 32 below and was what we used in the experiments.

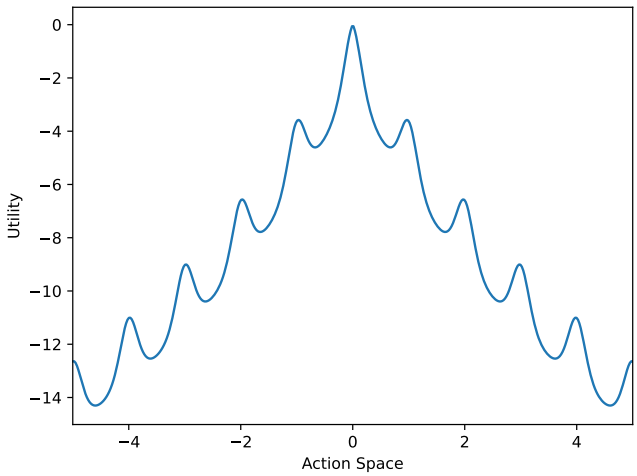


Figure 5: Visualization of Ackley utility function used in experiments.

For our simulations, we consider the one-dimensional ackley function and discretize the action space $\mathcal{X} = [-5, 5]$ into 40 evenly spaced points from which the algorithms can pick any two actions x_t, x'_t to play in each iteration. To produce the preference feedback from this utility function, we simply draw a Bernoulli random variable as

$$y_t = \mathbb{1}\{x_t \succ x'_t\} \sim \text{Bernoulli}(\sigma(f(x_t) - f(x'_t)).$$

In Figures 2a and 2b we plot the regret from using algorithms PF-TS (Ours), MR-LPF (Kayal et al., 2025), MaxMinLCB (Pásztor et al., 2024) and POP-BO (Xu et al., 2024). We use the same inference described in Section 2.3 with regularization term $\lambda = 0.05$ and we have used a Mátern 2.5 kernel for each algorithm with length scale $\ell = 0.1$. In Figure 2a we plot the mean instantaneous regret and in Figure 2b we plot the mean cumulative regret (as well as the standard errors on each) at each iteration from 30 runs of each algorithm up to horizon $T = 300$. On the same figure we have also plotted GP-TS (Chowdhury and Gopalan, 2017) with access to direct scalar feedback of the same utility function f, where we use standard Gaussian noise for the observations. Namely, we allow GP-TS in each round to select one action from the action space $x_t \in \mathcal{X}$. From this action they then observe direct scalar feedback of

$$y_t = f(x_t) + \epsilon_t,$$

where $\epsilon_t \sim N(0, 1)$. Then we record the regret in each round t to be the same as the preference feedback in equation (2) with $x'_t = x_t$. Again, this is equivalent to using the standard regret definition of $f(x^*) - f(x_t)$ up to the same constant. By using (2) we simply put the regret for all algorithms on the same scale. For confidence interval-based algorithms MR-LPF, MaxMinLCB, POP-BO, and GP-UCB we used the confidence interval width coefficient equal to 1 (as in Kayal et al., 2025). For GP-TS and TS we used the exploration bonus $v_t^2 = \sqrt{t + 1 + \log(2/0.05)}$. This performs well in practice and is a minor modification to our bonus in (12), where we have loosened the information gain to t and then taken the square-root.

F.2 Catalyst Composition Experiments

The data used for the experiments described in Section 5.2 can be found in the Open Catalyst Experiments 2024 (OCx24) dataset released by *FAIR Chemistry* (Meta AI) (Abed et al., 2024). The full dataset can be accessed publically on github (`https://github.com/facebookresearch/fairchem/tree/main/src/fairchem/applications/ocx/data/experimental_data`) which reports hydrogen yields for various different catalyst compositions, for two different experiments at different currents. One of the most popular compositions studied in this work are catalysts composed of silver (Ag), gold (Au), and zinc (Zn). Hence, to define our utility function, we fix a particular experimental set up and only vary these compositions. In particular, we consider the CO2R reaction at a 300-amp current. We then record the hydrogen yield ('fe-h2') for each of the 63 such compositions as the utility for each Ag/Au/Zn composition. This extracted data be viewed in the supplementary material. The

hydrogen yields, originally measured quantitatively in the range [0,100], were rescaled to [0,10] for the purposes of our experiments.

Figures 2c and 2d similarly plot regret for PF-TS (ours), MR-LPF (Kayal et al., 2025), MaxMinLCB (Pásztor et al., 2024), and POP-BO (Xu et al., 2024), using the same inference procedure as Section 2.3 and a Matérn-5/2 kernel with hyperparameters same as the previous experiments. For reference, we again also include cumulative regret for vanilla GP-TS (Chowdhury and Gopalan, 2017) with direct scalar feedback on f. Results are averaged over 30 runs up to a horizon of $T = 800$ iterations.

F.3 RKHS Function Experiments

For experimental results displayed in Figures 4c and 4d we use a utility function sampled from an RKHS with Matérn-5/2 kernel (similar to the process used by Kayal et al. (2025)). The sampled utility function is displayed in the following figure. When running the experiments using this utility function, we discretize the $[0, 1]$ action space into 30 evenly spaced points. We then construct any preference feedback from action pairs with the same procedure discussed in Appendix F.1.

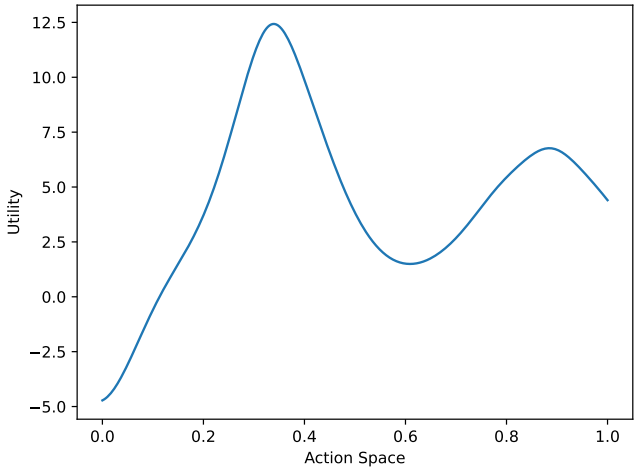


Figure 6: Visualization of utility function sampled from RKHS used in experiments.

F.4 Hyperparameter Tuning Experiments

For the results displayed in Figures 4a and 4b, we use data from the hyperparamter tuning benchmark LCBench (Zimmer et al., 2021). In particular, we look at the recorded accuracy data for different neural network architectures with the objective of image classification on the MNIST dataset. For these experiments we consider a set of 40 hyperparameter configurations for the neural networks used which have 4 layers. Hence, in this case we define a utility function f to be the accuracy achieved from each configuration and the action space $\mathcal{X}$ is 5-dimensional with entries representing batch size, number of units per layer, learning rate, momentum, and weight decay. With this utility function and action space defined, we construct the preference feedback in the same way as usual (see Appendix F.1).

F.5 Cost of Scalar vs Preference Feedback

For the results displayed in Figure 3, we use the results from the experiments from the previous Section 5.2. In particular, we take the instantaneous regret data for PF-TS (ours) using preference feedback and vanilla GP-TS which has access to direct scalar feedback from the catalyst composition utility function over a horizon of $T = 800$ iterations. We then plot a total cost budget c versus achieved instantaneous regret for our algorithm PF-TS, where one unit of cost is the cost of one preference feedback iteration (i.e. cost budget c buys c preference feedback iterations). We then consider various cost ratios $\xi \in \{1, 3, 5, 7\}$ between scalar and preference feedback (i.e. cost budget c buys c/ξ scalar feedback iterations). Under these different ratios, we plot the total cost budget c versus achieved instantaneous regret for the vanilla GP-TS algorithm using scalar feedback. Namely,

at each cost budget c, PF-TS runs for c preference iterations, while GP-TS is allotted only c/ξ scalar iterations. In Figure 3, we plot the regret and ± 1 standard error after each additional 25 iterations can be bought by each algorithm at each ratio. Note that we see that with a cost budget of $c = 800$ we can afford many more preference feedback iterations than scalar feedback iterations when the cost ratio, ξ, is high for the scalar feedback. Hence, we can achieve lower instantaneous regret in these settings.

F.6 Anchor Independence

The plot in Figure 1 was made by considering an action space of size $|\mathcal{X}| = 50$ and an Ackley utility function. We uniformly select 200 pairs of actions and then estimate h_t, σ_t using the inference presented in Section 2. We then draw a posterior sample of $\tilde{h}_t$ as

$$\tilde{h}_t(\cdot, \cdot) \sim GP(h_t, v_t^2 k_t^{\Delta}).$$

Then for each of 25 evenly spaced choices for $x_0 \in [0, 5]$, we plot the maximizer of the single-argument search for TS as

$$x_t = \arg\max_{x \in \mathcal{X}} \tilde{h}_t(x, x_0)$$

and for UCB

$$x_t = \arg\max_{x \in \mathcal{X}} h_t(x, x_0) + \sigma_t^{\Delta}(x, x_0).$$

Here we have arbitrarily selected the scaling parameter $\beta = 1$ for UCB in order to illustrate it's anchor dependence. To reiterate: for a Thompson selected action, we see from Figure 1 that any choice of anchor point leads to the same choice of action. On the other hand, the action selected by UCB can vary dependent on the choice of anchor point.

F.7 Compute Information

For the implementation of the inference methods described in Section 2, we extended the recent codebase of Kayal et al. (2025). The full implementation of our method is provided in the supplementary material. All simulations were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU (25 GB RAM) and an AMD EPYC 7542 32-core CPU.

For each preference-feedback algorithm applied to the Ackley utility function (Section 5.1), the runtime per experiment was approximately 5–10 hours. For the catalyst dataset experiments (Section 5.2), each preference-feedback algorithm required approximately 10–20 hours to complete.