Principled Preferential Bayesian Optimization

Wenjie Xu [1 2] **Wenbin Wang** [1] **Yuning Jiang** [1] **Bratislav Svetozarevic** [2 3] **Colin N. Jones** [1]

Abstract

We study the problem of preferential Bayesian optimization (BO), where we aim to optimize a black-box function with only *preference* feedback over a pair of candidate solutions. Inspired by the *likelihood ratio* idea, we construct a confidence set of the black-box function using only the preference feedback. An optimistic algorithm with an efficient computational method is then developed to solve the problem, which enjoys an information-theoretic bound on the total cumulative regret, a *first-of-its-kind* for preferential BO. This bound further allows us to design a scheme to report an estimated best solution, with a guaranteed convergence rate. Experimental results on sampled instances from Gaussian processes, standard test functions, and a thermal comfort optimization problem all show that our method stably achieves better or competitive performance as compared to the existing state-of-the-art heuristics, which, however, do not have theoretical guarantees on regret bounds or convergence.

1. Introduction

Bayesian optimization (BO) is a popular sample-efficient black-box optimization method (Shahriari et al., 2015; Frazier, 2018). It is widely applied to tuning hyperparameters of machine learning models (Snoek et al., 2012), optimizing the performance of control systems (Xu et al., 2022b), and discovering new drugs (Negoescu et al., 2011), etc.

The main idea of BO is based on *surrogate modeling*. That is, a learning algorithm (typically Gaussian process regression) is applied to learn the unknown black-box function using historical samples, which then outputs a learned surrogate together with uncertainty quantification. Then BO algorithms, such as the popular Expected Improvement (Jones et al., 1998) and GP-UCB algorithms (Srinivas et al., 2012), use the information of this learned surrogate and uncertainty quantification to choose the next sample point.

The conventional BO setting assumes each sample, which typically corresponds to a round of real-world experiment or software simulation in practice, returns a noisy scalar evaluation of the black-box function. However, many *human-in-the-loop* systems can not return such a scalar value, or it is much more difficult to directly obtain such a scalar evaluation from humans since humans are bad at sensing absolute magnitude (Kahneman & Tversky, 2013). In contrast, it is much easier for a human to compare a pair of solutions and report which is preferred (Lichtenstein & Slovic, 1971; Tversky & Kahneman, 1974; Kahneman & Tversky, 2013).

This gives rise to *preferential Bayesian optimization* (González et al., 2017), where the scalar evaluation of the black-box function is not available. But rather, we can query an oracle to compare a pair of solutions, or the so-called *duels*. Such settings arise widely in a broad range of applications, such as visual design optimization (Koyama et al., 2020), thermal comfort optimization (Abdelrahman & Miller, 2022) and robotic gait optimization (Li et al., 2021).

Existing preferential Bayesian optimization methods are mostly heuristic, without formal guarantees on cumulative regret or convergence to the global optimal solution. For example, (González et al., 2017) proposes several heuristic acquisition strategies, including expected improvement and Thompson sampling-based methods, for preferential Bayesian optimization. (Mikkola et al., 2020) extends the preferential Bayesian optimization to the projective setting. (Takeno et al., 2023) proposes a Thompson sampling-based method for practical preferential Bayesian optimization with skew Gaussian process. (Astudillo et al., 2023) proposes a decision theoretical acquisition strategy with a convergence rate guarantee for a finite input set. However, as far as we know, all the existing preferential Bayesian optimization methods can not provide theoretical guarantees on cumulative regret or global convergence with continuous input space, partially due to the challenge of quantifying uncertainty in a principled way.

Beyond preferential BO, optimization from preference feed-

[1]Automatic Control Laboratory, EPFL, Lausanne, Switzerland [2]Urban Energy Systems Laboratory, Empa, Zurich, Switzerland [3]The Institute for Artificial Intelligence Research and Development of Serbia, Serbia. Correspondence to: Yuning Jiang <yuning.jiang@ieee.org>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

back has also been investigated in other contexts. In the following, we first survey the related work other than preferential BO and then highlight our unique contributions.

Dueling Bandits In dueling bandits (Yue et al., 2012), the goal is to identify the best arm from a set of finite arms, using only the noisy comparison feedback. It has also been extended to adversarial (Gajane et al., 2015) and contextual (Dudík et al., 2015; Saha & Krishnamurthy, 2022) settings. One extension that is most related to this work is kernelized dueling bandits (Sui et al., 2017; 2018). However, this line of research is typically restricted to the case where the number of arms is finite, and the regret bound can blow up to infinity when the number of arms goes to infinity (e.g., Thm. 2 in (Sui et al., 2017)). A recent work (Mehta et al., 2023) proposes an offline method with suboptimality bound by learning winning probability, which, however, are not applicable to online learning problems due to linear growth of regret over the randomly sampled compared point sequences. In the existing literature, there is no *cumulative* regret bound that depends on an inherent complexity metric (such as covering number and maximum information gain (Srinivas et al., 2012)) of the black-box function with continuous input space.

Convex Optimization with Preference Feedback (Saha et al., 2021; Yue & Joachims, 2009) consider the optimization of convex functions, where only a comparison oracle of function values over different points is available. The proposed methods estimate the gradient from the preference signals. However, this line of research restricts the function to be convex, while in practice, the black-box function may be non-convex. The proposed method may get stuck in a local optimum and can be sample-inefficient since each estimate of the gradient already needs several samples.

Reinforcement Learning from Human Feedback Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Griffith et al., 2013) has recently become very popular. It has found many successes in wide applications, including training robots (Hiranaka et al., 2023), playing games (Warnell et al., 2018), and remarkably large language models (Ouyang et al., 2022). On the theoretical line of RLHF research, recent results analyze the offline learning of the implicit reward function (Zhu et al., 2023) and the model-based optimistic reinforcement learning from human feedback (Wang et al., 2023). However, the existing theoretical analysis either only deals with finite-dimensional generalized linear models or highly relies on the complexity measure of Eluder dimension (Osband & Van Roy, 2014). The existing generic theoretical analysis for RLHF can not be directly applied to the Bayesian optimization setting, where the Eluder dimension of the infinite-dimensional reproducing kernel Hilbert space is not well understood.

Optimistic Model-based Sequential Decision Making Optimism in the face of uncertainty is a widely adopted design principle for model-based sequential decision making problems, such as in Bayesian optimization/reinforcement learning (Wu et al., 2022; Xu et al., 2023; Pacchiano et al., 2021; Curi et al., 2020; Liu et al., 2023). The optimism principle has also been applied to RLHF (Wang et al., 2023) recently. However, as far as we know, there is no existing principled optimistic algorithm for preferential BO yet.

Our contributions. Guided by the optimism principle, we design a preferential Bayesian optimization algorithm that enjoys information-theoretic bounds on the cumulative regret. Specifically, our contributions include:

- **Algorithm design.** Inspired by the recent work of the confidence set based on optimistic maximum likelihood estimate (Liu et al., 2023) and the *likelihood ratio* confidence set idea (Owen, 1990; Emmenegger et al., 2023), we construct a confidence set by only using the preference feedback. We then exploit the principle of optimism in the face of uncertainty to design a **P**rincipled **O**ptimistic **P**referential **B**ayesian **O**ptimization (**POP-BO**) algorithm, together with a scheme of reporting an estimated best solution.
- **Theoretical analysis.** Under some mild regularity assumptions, we prove an information-theoretic bound on the cumulative regret of POP-BO algorithm, which is *first-of-its-kind* [1] for preferential Bayesian optimization. This is significant since previous information-theoretic regret bounds typically assume the direct scalar evaluations of black-box functions (Srinivas et al., 2012) while the recent generic theoretical results for RLHF typically rely on Eluder dimension, which is not well understood for RKHS.
- **Efficient computations.** The optimistic algorithm needs to solve bi-level optimization problems with the inner variable in an infinite-dimensional function space. We leverage the *representer theorem* (Schölkopf et al., 2001) to reduce the inner optimization problem to finite-dimensional space, which turns out to be tractable via convex optimization. This further allows efficient grid-free joint optimization.
- **Empirical validations and toolbox.** [2] Experimental results show that POP-BO consistently achieves better or competitive performance as compared to the state-of-the-art heuristic baselines and more than 10 times speed-up in computation as compared to the Thompson sampling based method. We also provide a reusable toolbox for future applications of our method.

[1] (Mehta et al., 2023) provides a bound on the partial cumulative regret, which only captures the suboptimality of one point in each compared duel. We consider stronger total cumulative regret over both points in the compared duel. See Appendix Q for a detailed discussion.

[2] Code link: https://github.com/PREDICT-EPFL/POP-BO

2. Problem Statement

We consider the maximization of a black-box function f,

$$\max_{x\in\mathcal{X}} f(x), \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ with d as the input dimension. We use $x \succ x'$ to denote the event that 'x is preferred to x''. In contrast to the standard BO setup, we assume that we can not directly evaluate the scalar value of $f(x)$ but rather, we have a comparison oracle that compares any two points x, x' and returns a preference signal $\mathbf{1}_{x\succ x'}$, which is defined as

$$\mathbf{1}_{x\succ x'} = \begin{cases} 1, & \text{if } x \text{ is preferred,} \\ 0, & \text{if } x' \text{ is preferred.} \end{cases} \tag{2}$$

Before proceeding, we state a set of common assumptions.

Assumption 2.1. $\mathcal{X}$ is compact and nonempty.

Assumption 2.1 is reasonable because, in many applications (e.g., continuous hyperparameter tuning) of Bayesian Optimization, we are able to restrict the optimization into certain ranges based on domain knowledge. Regarding the black-box function f, we assume that,

Assumption 2.2. $f \in \mathcal{H}_k$, where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a symmetric, positive semidefinite kernel function and $\mathcal{H}_k$ is the corresponding reproducing kernel Hilbert space (RKHS, see (Schölkopf et al., 2001)). Furthermore, we assume $\|f\|_k \leq B$, where $\|\cdot\|_k$ is the norm induced by the inner product in the corresponding RKHS.

Assumption 2.2 requires that the function to be optimized is regular in the sense that it has a bounded norm in the RKHS, which is a common assumption (Chowdhury & Gopalan, 2017a; Zhou & Ji, 2022). For simplicity, we will use $\mathcal{B}_f$ to denote the set $\left\{\tilde{f} \in \mathcal{H}_k | \|\tilde{f}\|_k \leq B\right\}$, which is a ball with radius B in $\mathcal{H}_k$.

Remark 2.3 (**Choice of** B). In practice, a tight norm bound B might not be known beforehand. In the theoretical analysis, we only assume that there is a finite bound B, possibly unknown beforehand. In the practical implementation of our algorithm, we can adapt B based on hypothesis testing (Newey & McFadden, 1994). For example, we can double B every time we detect a low likelihood value (See more elaboration in Appendix G.).

Assumption 2.4. $k(x, x') \leq 1, \forall x, x' \in \mathcal{X}$ and $k(x, x')$ is continuous on $\mathbb{R}^d \times \mathbb{R}^d$.

Assumption 2.4 is a commonly adopted mild assumption in the BO literature (Srinivas et al., 2012; Chowdhury & Gopalan, 2017a). It holds for most commonly used kernel functions after normalization, such as the linear kernel, the Matérn kernel, and the squared exponential kernel.

Assumption 2.5. The random preference feedback $\mathbf{1}_{x\succ x'}$ from the comparison oracle follows the Bernoulli distribution with $\mathbb{P}(\mathbf{1}_{x\succ x'} = 1) = p_{x\succ x'} = \sigma(y - y')$, where $y = f(x)$, $y' = f(x')$ and $\sigma(u) = 1/(1+e^{-u})$[3].

Assumption 2.5 equivalently assumes that,

$$\mathbb{P}(\mathbf{1}_{x\succ x'} = 1) = \frac{e^{f(x)}}{e^{f(x)} + e^{f(x')}}, \tag{3}$$

which can be observed to be the widely used Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952) for pairwise comparison. The intuition here is that the more advantage $f(x)$ has as compared to $f(x')$, the more likely x is preferred. The same comparison model is also used in, e.g., training large language models (Ouyang et al., 2022). At step t, our algorithm queries the pair (x_t, x'_t) and the comparison oracle returns the random preference $\mathbf{1}_{x_t\succ x'_t} \in \{0, 1\}$. For the simplicity of notation, we use $\mathbf{1}_\tau \in \{0, 1\}$ to denote the realization of the Bernoulli random variable $\mathbf{1}_{x_\tau\succ x'_\tau}$ when querying the comparison oracle at step τ. Based on the historical comparison results

$$\mathcal{D}_t := \{(x_\tau, x'_\tau, \mathbf{1}_\tau)\}_{\tau=1}^t, \tag{4}$$

the algorithm needs to decide the next pair of samples to compare. Without further notice, all the theoretical results in this paper are under the assumptions 2.1, 2.2, 2.4, 2.5, and all the corresponding proofs are in the appendices.

Notations. The probability, denoted as $\mathbb{P}(\cdot)$, is taken over the randomness of the preference feedback generated by the comparison oracle and the randomness generated by the algorithm. Let the filtration $\mathcal{F}_t$ capture all the randomness up to step t. $\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)$ denotes the standard covering number (Zhou, 2002) of the function space ball $\mathcal{B}_f$ with the covering balls' radius ϵ and the infinity norm $\|\cdot\|_\infty$. We will also use $[\tau]$ to denote the set $\{1, \cdots, \tau\}$.

3. High Confidence Set

3.1. Likelihood-based Confidence Set

We first introduce the function,

$$\begin{aligned} p_{\hat{f}}(x_\tau, x'_\tau, \mathbf{1}_\tau) :=& \mathbf{1}_\tau \sigma(\hat{f}(x_\tau) - \hat{f}(x'_\tau)) + \\ & (1 - \mathbf{1}_\tau)\left(1 - \sigma(\hat{f}(x_\tau) - \hat{f}(x'_\tau))\right), \end{aligned} \tag{5}$$

which is the likelihood of $\hat{f}$ over the event $\mathbf{1}_{x_\tau\succ x'_\tau} = \mathbf{1}_\tau$ under the Bernoulli preference model in Assumption 2.5.

We can then derive the likelihood function of a fixed function

[3]We mainly consider the widely used sigmoid function here. Our result can be extended to more general σ under mild regularity conditions.

$\hat{f}$ over the historical preference dataset $\mathcal{D}_t$ [4].

$$\mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) := \prod_{\tau=1}^t p_{\hat{f}}(x_\tau, x'_\tau, \mathbf{1}_\tau) \tag{6}$$

Taking log gives the log-likelihood function,

$$\begin{aligned}\ell_t(\hat{f}) &:= \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) = \sum_{\tau=1}^t \log p_{\hat{f}}(x_\tau, x'_\tau, \mathbf{1}_\tau) \\ &= \sum_{\tau=1}^t \log \left(\frac{e^{z_\tau}\mathbf{1}_\tau + e^{z'_\tau}(1-\mathbf{1}_\tau)}{e^{z_\tau}+e^{z'_\tau}} \right) \\ &= \sum_{\tau=1}^t (z_\tau \mathbf{1}_\tau + z'_\tau(1-\mathbf{1}_\tau)) - \sum_{\tau=1}^t \log\left(e^{z_\tau}+e^{z'_\tau}\right),\end{aligned} \tag{7}$$

where $z_\tau = \hat{f}(x_\tau), z'_\tau = \hat{f}(x'_\tau)$, $\mathbf{1}_\tau \in \{0,1\}$ is the data realization of $\mathbf{1}_{x_\tau \succ x'_\tau}$, and the last equality can be checked correct for either $\mathbf{1}_\tau = 1$ or $\mathbf{1}_\tau = 0$.

A common method for statistical estimation is by maximizing the likelihood. Hence, we introduce the maximum likelihood estimator (MLE),

$$\hat{f}_t^{\text{MLE}} \in \arg\max_{\tilde{f}\in\mathcal{B}_f} \log \mathbb{P}_{\tilde{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t). \tag{8}$$

With the maximum likelihood estimator introduced, the posterior high confidence set can be derived as shown in Thm. 3.1 using the maximum log-likelihood value.

Theorem 3.1 (Likelihood-based Confidence Set). *$\forall \epsilon, \delta > 0$, let,*

$$\mathcal{B}_f^{t+1} := \{\tilde{f} \in \mathcal{B}_f | \ell_t(\tilde{f}) \geq \ell_t(\hat{f}_t^{\text{MLE}}) - \beta_1(\epsilon, \delta, t)\}, \tag{9}$$

where $\beta_1(\epsilon,\delta,t) := \sqrt{32tB^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{6\delta}} + C_L \epsilon t = \mathcal{O}\left(\sqrt{t \log \frac{t\mathcal{N}(\mathcal{B}_f,\epsilon,\|\cdot\|_\infty)}{\delta}} + \epsilon t\right)$*, with C_L a constant independent of δ, t and ϵ. We have,*

$$\mathbb{P}\left(f \in \mathcal{B}_f^{t+1}, \forall t \geq 1\right) \geq 1-\delta. \tag{10}$$

Intuitively, the confidence set $\mathcal{B}_f^{t+1}$ includes the functions with the log-likelihood value that is only 'a little worse' than the maximum likelihood estimator. It turns out that by correctly setting the 'worse' level β_1, the confidence set $\mathcal{B}_f^{t+1}$ contains the ground-truth function f with high probability. This is reasonable because the preference data is generated with the ground-truth function, and thus the likelihood of the ground-truth function will not be too much lower than the maximum likelihood estimator.

[4] Note that $\mathbb{P}_{\hat{f}}(\cdot)$ is the likelihood function in $\hat{f}$ over the historical data $\mathcal{D}_t$, not the probability taken over the data/algorithm randomness.

Remark 3.2 (**Choice of** ϵ). In Thm. 3.1, $\beta_1(\epsilon,\delta,t)$ also depends on a small positive value ϵ, which is to be chosen. In the theoretical analysis, it will be seen that ϵ can be selected to be $1/T$, where T is the algorithm's running horizon.

Remark 3.3 (**Likelihood Ratio Idea**). The confidence set $\mathcal{B}_f^{t+1}$ contains the functions $\tilde{f}$ that satisfy,

$$\frac{\mathbb{P}_{\tilde{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)}{\mathbb{P}_{\hat{f}_t^{\text{MLE}}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)} \geq e^{-\beta_1(\epsilon,\delta,t)}, \tag{11}$$

which is the likelihood ratio confidence set (Owen, 1990).

Remark 3.4. Surrogate-based black-box optimization with kernel method is often referred to as Bayesian optimization due to its close relations to Bayesian Gaussian process model. Hence, we refer to our method as preferential BO.

Based on the confidence set in Thm. 3.1, we can derive the pointwise confidence range for the black-box function.

$$\inf_{\tilde{f}\in\mathcal{B}_f^t} \tilde{f}(x) \leq f(x) \leq \sup_{\tilde{f}\in\mathcal{B}_f^t} \tilde{f}(x). \tag{12}$$

Fig. 1 demonstrates the maximum likelihood estimate function and the confidence range with the ground truth function sampled from a Gaussian process, random comparison inputs, and $\beta_1(\epsilon,\delta,t)$ set to be a constant 1.0. It can be seen that the maximum likelihood estimate approximates the ground truth better and better with the confidence range shrinking, as we have more and more comparison data.

3.2. Bound Duel-wise Error

Thm. 3.1 gives a high confidence set based on the likelihood function. However, it is not straightforward how the likelihood bounds lead to the error bounds on function value differences over a compared pair (x, x'), which determines the preference distribution. The following theorem further gives such a bound over the historical samples.

Lemma 3.5 (Elliptical Bound). *For any estimate $\hat{f}_{t+1} \in \mathcal{B}_f^{t+1}$ that is measurable with respect to the filtration $\mathcal{F}_t$, we have, with probability at least $1-\delta$, $\forall t \geq 1$,*

$$\begin{aligned}\sum_{\tau=1}^t \left(\left(\hat{f}_{t+1}(x_\tau) - \hat{f}_{t+1}(x'_\tau)\right) - (f(x_\tau) - f(x'_\tau))\right)^2 \\ \leq \beta(\epsilon, \delta/2, t),\end{aligned} \tag{13}$$

and

$$f \in \mathcal{B}_f^{t+1}, \tag{14}$$

where $\beta(\epsilon,\delta/2,t) = \frac{\sigma'^2}{\bar{H}_\sigma}(\beta_2(\epsilon,\delta/2,t) + 2\beta_1(\epsilon,\delta/2,t)) = \mathcal{O}\left(\sqrt{t\log\frac{t\mathcal{N}(\mathcal{B}_f,\epsilon,\|\cdot\|_\infty)}{\delta}} + \epsilon t + \epsilon^2 t\right)$*, with* $\beta_2(\epsilon,\delta,t) = 8H_\sigma \bar{\sigma'}^2 \epsilon^2 t + 2C_L \epsilon t + \sqrt{8tB_p^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f,\epsilon,\|\cdot\|_\infty)}{3\delta}}$ *and the constants $\underline{\sigma'}, H_\sigma, \bar{\sigma'}, B_p$ as defined in Appendix B.*

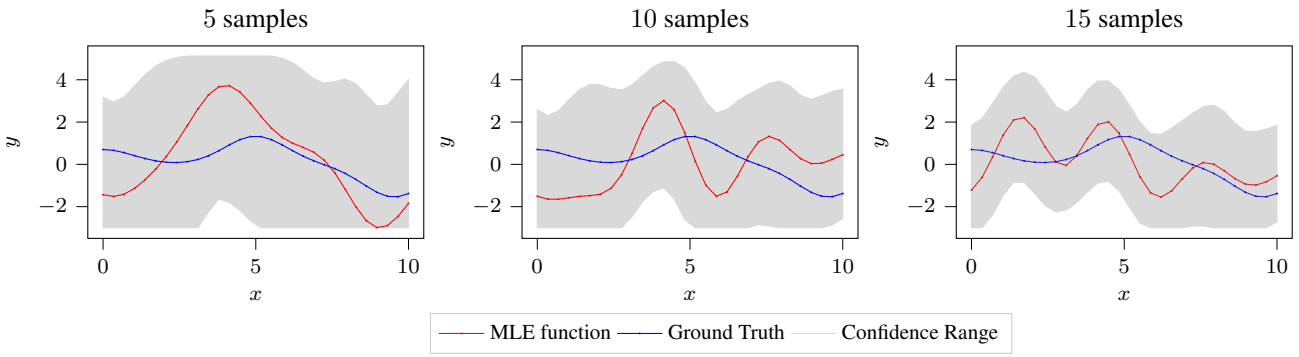

Figure 1. Demonstration of the maximum likelihood function and the confidence set based on likelihood. The results are derived using random sequential comparisons (that is, comparing x_t to x_{t-1}), where each x_t is uniformly randomly sampled from the input set.

Lem. 3.5 highlights that with high probability, all the functions in the confidence set have difference values over the historical sample points that lie in a ball with the ground-truth function difference value as the center and $\sqrt{\beta(\epsilon, \delta/2, t)}$ as the radius. Lem. 3.5 indicates that our likelihood-based learning scheme can gradually learn the function differences $f(x_\tau) - f(x'_\tau)$ but not the absolute value $f(x_\tau)$. This is reasonable since shifting f by a constant will not change the distribution of preference feedback.

Furthermore, to derive an error bound over a new pair (x, x'), we need to quantify the uncertainty of $\tilde{f}(x) - \tilde{f}(x')$, where $\tilde{f} \in \mathcal{B}_f$. Since $-\tilde{f} \in \mathcal{B}_f$ by the definition of $\mathcal{B}_f$, it can be seen that $\tilde{f}(x) - \tilde{f}(x') \in \mathcal{B}_{ff'}$, where

$$\mathcal{B}_{ff'} := \{F(x, x') = \tilde{f}(x) + \tilde{f}'(x') | \tilde{f}, \tilde{f}' \in \mathcal{B}_f\}. \quad (15)$$

Indeed, $\mathcal{B}_{ff'}$ is the ball with radius $2B$ in the RKHS equipped with the additive kernel function $k^{ff'}((x, x'), (\bar{x}, \bar{x}')) := k(x, \bar{x}) + k(x', \bar{x}')$, which we term as the augmented RKHS here, and inner product $\langle f_1 + f'_1, f_2 + f'_2 \rangle_{k^{ff'}} := \langle f_1, f_2 \rangle_k + \langle f'_1, f'_2 \rangle_k$. The readers are referred to (Christmann & Hable, 2012; Kandasamy et al., 2015) for more details of the additive kernel and the corresponding RKHS. To quantify the uncertainty of a new pair (x, x'), we further introduce the function,

$$\left(\sigma_t^{ff'}(\omega)\right)^2 = k^{ff'}(\omega, \omega) \quad (16)$$
$$-k^{ff'}(\omega_{1:t-1}, \omega)^\top \left(K_{t-1}^{ff'} + \lambda I\right)^{-1} k^{ff'}(\omega_{1:t-1}, \omega),$$

where $\omega := (x, x')$, $\omega_{1:t-1} := ((x_\tau, x'_\tau))_{\tau=1}^{t-1}$, $K_{t-1}^{ff'} := (k^{ff'}((x_{\tau_1}, x'_{\tau_1}), (x_{\tau_2}, x'_{\tau_2})))_{\tau_1 \in [t-1], \tau_2 \in [t-1]}$, and λ is a positive regularization constant.

Theorem 3.6 (Duel-wise Error Bound). *For any estimate $\hat{f}_{t+1} \in \mathcal{B}_f^{t+1}$ measurable with respect to $\mathcal{F}_t$, we have, with probability at least $1 - \delta$, $\forall t \geq 1, (x, x') \in \mathcal{X} \times \mathcal{X}$,*

$$\left|(\hat{f}_{t+1}(x) - \hat{f}_{t+1}(x')) - (f(x) - f(x'))\right|$$
$$\leq 2\left(2B + \lambda^{-1/2}\sqrt{\beta(\epsilon, \delta/2, t)}\right)\sigma_{t+1}^{ff'}((x, x')). \quad (17)$$

Remark 3.7. In preferential BO, we do not get the scalar value of $f(x) - f(x')$. Hence, $\sigma_t^{ff'}$ can not be interpreted as the posterior standard deviation as in (Srinivas et al., 2012). However, it turns out that $\sigma_t^{ff'}$, as a measure of uncertainty, still accounts for a factor of the duel-wise error.

To characterize the complexity of this augmented RKHS, we use the maximum information gain (Srinivas et al., 2012),

$$\gamma_T^{ff'} := \max_{\Omega \subset \mathcal{X} \times \mathcal{X}; |\Omega| = T} \frac{1}{2} \log \left| I + \lambda^{-1} K_\Omega^{ff'} \right|, \quad (18)$$

where $K_\Omega^{ff'} = \left(k^{ff'}((x, x'), (\bar{x}, \bar{x}'))\right)_{(x,x'),(\bar{x},\bar{x}') \in \Omega}$.

4. Algorithm

4.1. Principled Optimistic Algorithm

We are now ready to give the optimistic algorithm in Alg. 1.

Algorithm 1 **P**rincipled **O**ptimistic **P**referential **B**ayesian **O**ptimization (**POP-BO**).

1: Given the initial point $x_0 \in \mathcal{X}$ and set $\mathcal{B}_f^1 = \mathcal{B}_f$.
2: **for** $t \in [T]$ **do**
3: Set the reference point $x'_t = x_{t-1}$.
4: Compute
$$x_t \in \arg\max_{x \in \mathcal{X}} \max_{\tilde{f} \in \mathcal{B}_f^t} (\tilde{f}(x) - \tilde{f}(x'_t)),$$
with the inner optimal function denoted as $\tilde{f}_t$.
5: Query the comparison oracle to get the feedback result $\mathbf{1}_t$ and append the new data to $\mathcal{D}_t$.
6: Update the maximum likelihood estimator $\hat{f}_t^{\text{MLE}}$ and the posterior confidence set $\mathcal{B}_f^{t+1}$.
7: **end for**

The key to Alg. 1 is line 4. The idea is to maximize the

optimistic advantage of $\tilde{f}(x)$ as compared to $\tilde{f}(x'_t)$ with the uncertainty of the black-box function $\tilde{f} \in \mathcal{B}_f^t$.

In line 3, we set the reference point x'_t as the last generated point x_{t-1}. In practice, this may correspond to two possible scenarios. In the first, each comparison requires one experiment, such as image quality comparison. In this case, we only need to set one of the compared pair as the last newly generated solution. While in the other scenario, comparing x_t and x'_t needs separate experiments for x_t and x'_t. For example, when optimizing the building thermal comfort, the occupants need to experience both thermal conditions to report preference. If at step t, the oracle still has memory about the experience with input x_{t-1}, we can directly compare x_t and x_{t-1}. In this case, setting x'_t to be x_{t-1} saves the experimental expense with x'_t.

For online applications, cumulative regret is more of our interest. However, for an offline optimization setting, it may be of more interest to identify one near-optimal solution to report. Unlike in the scalar evaluation setting, where we can directly use the scalar value to report the best observed solution, we can not directly identify the best sampled solution in the preferential Bayesian optimization scenario. To address this issue, we report the solution $x_{t^\star}$, where

$$t^\star \in \arg\min_{t\in[T]} 2\left(2B + \lambda^{-1/2}\sqrt{\beta(\epsilon, \delta/2, t)}\right)\sigma_t^{ff'}((x_t, x'_t)). \tag{19}$$

The idea is that although the best sample may not be known, we can derive a solution by minimizing the known term $2(2B+\lambda^{-1/2}\sqrt{\beta(\epsilon,\delta/2,t)})\sigma_t^{ff'}((x_t,x'_t))$ to find a solution $x_{t^\star}$ to report. Indeed, this term upper bounds the uncertainty of the optimistic advantage (as shown in Thm. 3.6). Hence, the smaller it is, the more certain that $f(x_t)$ is close to the ground-truth optimal value. At step t, we can report the current estimated solution with index $\tau^\star(t)$ satisfying a similar formula to Eq. (19).

4.2. Efficient Computations

Line 4 in Alg. 1 requires solving a nested optimization problem with inner variables in an infinite-dimensional function space. The update of the maximum likelihood estimator also requires solving an optimization problem with an infinite-dimensional function as the decision variable. These are in general not tractable in their current forms. Fortunately, we can reduce the infinite-dimensional problems to finite-dimensional ones, thanks to the structures of the problem and the representer theorem (Schölkopf et al., 2001).

Maximum likelihood estimation. Since the log-likelihood function

$$\ell_t(\tilde{f}) = \log \mathbb{P}_{\tilde{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \tag{20}$$

$$= \sum_{\tau=1}^t (z_\tau \mathbf{1}_\tau + z'_\tau(1-\mathbf{1}_\tau)) - \sum_{\tau=1}^t \log\left(e^{z_\tau} + e^{z'_\tau}\right)$$

only depends on the function value $(z_\tau, z'_\tau) = (\tilde{f}(x_\tau), \tilde{f}(x'_\tau))$, we only need to optimize over (z_τ, z'_τ) subject to that they are functions in $\mathcal{H}_k$ with norm less or equal to B. Furthermore, Alg. 1 sets $x'_\tau = x_{\tau-1}$ and thus $z'_\tau = z_{\tau-1}$. So we can reduce the optimization variables to only $(z_\tau)_{\tau=0}^t$. Hence, Eq. (20) is reduced to the following log-likelihood function that only depends on $(z_\tau)_{\tau=0}^t$,

$$\ell(Z_{0:t}|\mathcal{D}_t) \tag{21}$$

$$:= Z_{1:t}^\top \mathbf{1}_{1:t} + {Z_{0:t-1}}^\top(1-\mathbf{1}_{1:t}) - \sum_{\tau=1}^t \log\left(e^{z_\tau} + e^{z_{\tau-1}}\right),$$

where $Z_{0:t} := (z_\tau)_{\tau=0}^t$, $Z_{1:t} := (z_\tau)_{\tau=1}^t$, $Z_{0:t-1} := (z_\tau)_{\tau=0}^{t-1}$ and $\mathbf{1}_{1:t} = (\mathbf{1}_\tau)_{\tau=1}^t$.

By the representer theorem (Schölkopf et al., 2001), the maximum likelihood estimation problem can be solved via,

$$\begin{aligned} \ell_t(\hat{f}_t^{\mathrm{MLE}}) = \max_{Z_{0:t}\in\mathbb{R}^{t+1}} \quad & \ell(Z_{0:t}|\mathcal{D}_t) \\ \text{subject to} \quad & Z_{0:t}^\top K_{0:t}^{-1} Z_{0:t} \le B^2, \end{aligned} \tag{22}$$

where $K_{0:t} := (k(x_{\tau_1}, x_{\tau_2}))_{\tau_1\in\{0\}\cup[t],\tau_2\in\{0\}\cup[t]}$. The constraint restricts that the function values need to come from a function inside the function space ball $\mathcal{B}_f$, where the left-hand side is indeed the minimum norm square of the possible interpolant through $\{(x_\tau, z_\tau)\}_{\tau=0}^t$ as shown in (Wendland, 2004). It can be checked that the maximization problem in Eq. (22) has a concave objective (as shown in Appendix A) with a convex feasible set. Thus, the problem in Eq. (22) is tractable via convex optimization.

Generating new sample point. On the line 4 of Alg. 1, a bi-level optimization problem needs to be solved, where the inner-level part has an infinite-dimensional function variable. The inner optimization problem has the form,

$$\begin{aligned} \max_{\tilde{f}} \quad & \tilde{f}(x) - \tilde{f}(x_t) \\ \text{subject to} \quad & \tilde{f} \in \mathcal{B}_f, \\ & \ell_t(\tilde{f}) \ge \ell_t(\hat{f}_t^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t), \end{aligned} \tag{23}$$

where $\beta_1(\epsilon, \delta, t)$ is as given in Thm. 3.1. Similar to the representer theorem, we have,

Lemma 4.1. *Prob. (23) can be equivalently reduced to,*

$$\begin{aligned} \max_{Z_{0:t}\in\mathbb{R}^{t+1}, z\in\mathbb{R}} \quad & z - z_t \\ \textit{subject to} \quad & \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix} \le B^2, \\ & \ell(Z_{0:t}|\mathcal{D}_t) \ge \ell_t(\hat{f}_t^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t), \end{aligned} \tag{24}$$

where

$$K_{0:t,x} = \begin{bmatrix} K_{0:t} & (k(x_\tau, x))_{\tau=0}^{t} \\ (k(x_\tau, x))_{\tau=0}^{t}{}^\top & k(x,x) \end{bmatrix}. \quad (25)$$

Similarly, it can be checked that the Prob. (24) is convex.

For low-dimensional x, the outer-level problem can be solved via grid search. For medium-dimensional problems, we can optimize the inner/outer variables using a gradient-based/zero-order optimization method. Alternatively, we can jointly optimize $x, Z_{0:t}$, and z by a nonlinear programming solver from multiple random initial conditions. That is, we add x as another optimization variable as shown in the Prob. (26),

$$\begin{aligned} \max_{Z_{0:t}\in\mathbb{R}^{t+1}, z\in\mathbb{R}, x\in\mathcal{X}} \quad & z - z_t \\ \text{subject to} \quad & \text{Constraints of Prob. (24).} \end{aligned} \quad (26)$$

More details on this joint optimization approach is in Appendix H.

Remark 4.2. We add a matrix $\epsilon_K I$ to $K_{0:t}$ and $K_{0:t,x}$ before inversion to avoid numerical issue, where $\epsilon_K > 0$ is small.

Remark 4.3. In this paper, we mainly consider the setting where in each step, the preference is queried over two candidate points. Our Alg. 1 and the efficient computation schemes in this section can be easily extended to multiple-choice setting, where in each step, the best or most preferred point is queried over a batch of candidates. The detailed discussion is in Appendix I.

5. Theoretical Analysis

We first introduce the performance metrics to use. As in the scalar Bayesian optimization setting ((Srinivas et al., 2012)), cumulative regret is used as defined in Eq. (27),

$$R_T := \sum_{t=1}^{T} \left(f(x^\star) - f(x_t)\right), \quad (27)$$

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

Remark 5.1. The cumulative regret R_T as defined in Eq. (27) does not explicitly consider the sub-optimality of the reference point x'_t. However, since $x'_t = x_{t-1}$, the cumulative regret of the reference points is the same as R_T in Eq. (27), up to the difference of the first/last term.

Cumulative regret is of interest in the online setting. In the offline optimization setting, it is of more interest to analyze the sub-optimality of the final reported solution, i.e.,

$$f(x^\star) - f(x_{t^\star}), \quad (28)$$

where $x_{t^\star}$ is the final reported solution as defined in Eq. (19).

5.1. Regret Bound and Convergence Rate

Theorem 5.2 (Cumulative Regret Bound). *With probability at least $1-\delta$, the cumulative regret of Alg. 1 satisfies,*

$$R_T = \mathcal{O}\left(\sqrt{\beta_T \gamma_T^{ff'} T}\right), \quad (29)$$

where

$$\beta_T = \beta(1/T, \delta, T) = \mathcal{O}\left(\sqrt{T \log \frac{T\mathcal{N}(\mathcal{B}_f, 1/T, \|\cdot\|_\infty)}{\delta}}\right).$$

Remark 5.3 (**Differentiate from GP-UCB regret**). Our bound has a similar form as compared to the well-known regret bound for standard GP-UCB type algorithms (Srinivas et al., 2012; Chowdhury & Gopalan, 2017a). However, the β_T term here is significantly different from that in the existing literature (e.g., in Thm. 3 in (Srinivas et al., 2012)). It is derived specifically for the preferential BO and will lead to a bit larger bound for specific kernels in Sec. 5.2.

We highlight that Thm. 5.2 provides the *first-of-its-kind* information-theoretic bound on the cumulative regret of preferential BO, which further allows us to derive a convergence rate for the reported solution $x_{t^\star}$ in Thm. 5.4.

Theorem 5.4 (Convergence Guarantee). *Let $t^\star$ be defined as in Eq. (19). With probability at least $1-\delta$,*

$$f(x^\star) - f(x_{t^\star}) \le \mathcal{O}\left(\frac{\sqrt{\beta_T \gamma_T^{ff'}}}{\sqrt{T}}\right). \quad (30)$$

Thm 5.4 highlights that by minimizing the known term $2\big(2B + \lambda^{-1/2}\sqrt{\beta(\epsilon, \frac{\delta}{2}, t)}\big)\sigma_t^{ff'}((x_t, x'_t))$, the reported final solution $x_{t^\star}$ has a guaranteed convergence rate.

5.2. Kernel-Specific Bounds and Rates

In this section, we show kernel-specific bounds for the regret and convergence rate for the reported solution. The explicit forms of the considered kernels are given in Appendix L.

Theorem 5.5 (Kernel-Specific Regret Bounds). *Setting $\epsilon = 1/T$ and running our POP-BO algorithm in Alg. 1,*

1. *If $k(x,y) = \langle x, y\rangle$, we have,*

$$R_T = \mathcal{O}\left(T^{3/4}(\log T)^{3/4}\right). \quad (31)$$

2. *If $k(x,y)$ is a squared exponential kernel, we have,*

$$R_T = \mathcal{O}\left(T^{3/4}(\log T)^{3/4(d+1)}\right). \quad (32)$$

3. *If $k(x,y)$ is a Matérn kernel, we have,*

$$R_T = \mathcal{O}\left(T^{3/4}(\log T)^{3/4} T^{\frac{d}{\nu}\left(\frac{1}{4} + \frac{d+1}{4+2(d+1)d/\nu}\right)}\right), \quad (33)$$

where ν is the smooth parameter of the Matérn kernel that is assumed to be large enough such that $\nu > \frac{d}{4}(3 + d + \sqrt{d^2 + 14d + 17}) = \Theta(d^2)$.

Remark 5.6 (**Comparison to GP-UCB with Scalar Feedback**). Interestingly, as compared to the kernel-specific bounds in the scalar evaluation-based optimization (Fig. 1 in (Srinivas et al., 2012)), the regret bound of preferential Bayesian optimization approximately has an additional factor of $T^{1/4}$. This is reasonable since intuitively, scalar evaluation can imply preference, but not vice versa. Therefore, preference feedback contains less information and thus may suffer from higher regret. Fig. 2 in Sec. 6.1 and Fig. 4 in Appendix N empirically verify our bounds here.

We then derive the kernel-specific convergence rates for the reported solution $x_{t^\star}$, as shown in Tab. 3 in the Appendix O.

6. Experimental Results

In this section, we compare our method to the state-of-the-art preferential BO methods on sampled instances from Gaussian process, standard test functions, and a thermal comfort optimization problem. The comparison outcome is sampled as assumed in Assump. 2.5. We implement our algorithm based on the Gaussian process package GPy (GPy, since 2012). The optimization problems for MLE and generating new samples are formulated and solved using CasADi (Andersson et al., 2019) and Ipopt (Wächter & Biegler, 2006). We compare our methods to three baseline methods: dueling Thompson sampling (González et al., 2017), skew-GP based preferential BO (Takeno et al., 2023), and the qEUBO (Astudillo et al., 2023). The dueling Thompson sampling method (González et al., 2017) derives the next pair to compare by maximizing the soft-Copeland's score. The skew-GP based method (Takeno et al., 2023) applies standard BO algorithms conditioned on the Thompson sampling results on the historical sample points that are consistent with the historical preference feedbacks. The qEUBO (Astudillo et al., 2023) method uses the expected utility of the best option as an acquisition function. More experimental details and results on thermal comfort optimization are put in the Appendix P.

6.1. Sampled Instances from Gaussian Process

In this section, we sample the black-box function f from a Gaussian process with the squared exponential kernel as shown in Appendix L where the variance parameter is 9.0 and the lengthscale is 1.0. We sampled 30 instances in total.

Fig. 2 shows the performance comparisons with baselines. Our method achieves the lowest sublinear growth in cumulative regret. It also achieves better/competitive convergence speed for the reported solution as compared to the DTS method, while outperforming the SGP.

However, our method only uses less than 10% of the computation time as compared to the DTS as shown in Tab. 1. The SGP method gets stuck in local optimum because it overly trusts the random preference feedback (hard constraint when doing Thompson sampling). Although the qEUBO method performs slightly better in the reported solution, it suffers from more than 2.5 times the cumulative regret as compared to ours. Similar to qEUBO (reporting posterior mean maximizer), we can report the maximizer of the minimum-norm $\hat{f}_t^{\mathrm{MLE}}$ (POP-BO max-MLE in Fig. 2) instead of $x_{t^\star}$ in Eq. (19), and achieves faster convergence than qEUBO.

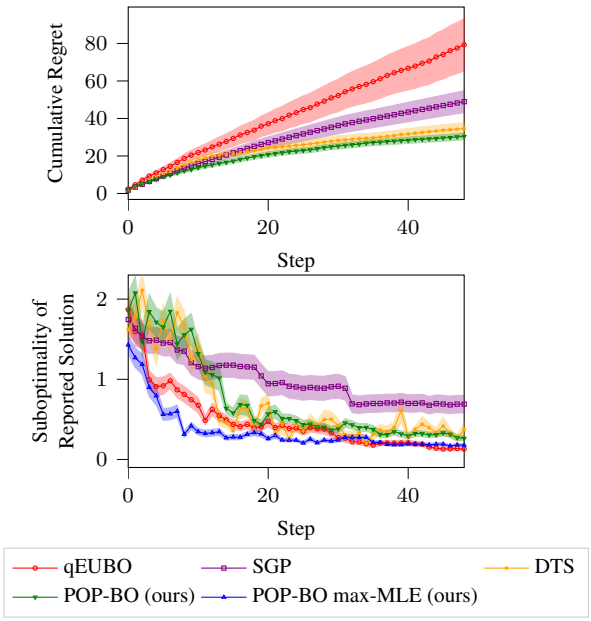

Figure 2. Cumulative regret and the suboptimality of the reported solution, where the shaded areas represent ± 0.1 standard deviation. qEUBO represents the method in (Astudillo et al., 2023), which reports the solution that maximizes the expected objective value conditioned on the historical samples. SGP represents the skew-GP based method (Takeno et al., 2023), which reports the first point of the duel proposed by the algorithm in the last step. DTS represents the dueling Thompson sampling method in (González et al., 2017), which reports the Condorcet winner.

Table 1. Computation time normalized against the DTS method.

DTS	qEUBO	SGP	POP-BO (ours)
1.0	0.21	0.07	0.09

6.2. Test Function Optimization

In this section, we compare our method to several well-known global optimization test functions (Dixon, 1978; Molga & Smutnicki, 2005), which are divided by the standard deviation of samples over a grid. We run our method multiple times from different random initial points. Tab. 2

shows that POP-BO consistently finds better or comparable solutions as compared to other baselines.

Table 2. Suboptimality for the final reported solution after 30 steps. The results (mean $\pm$ standard deviation) are taken over 30 runs with random starting points.

Problem	DTS	qEUBO	SGP	**POP-BO** (ours)
Beale	0.84 ± 0.52	0.15 ± 0.52	0.10 ± 0.19	$\mathbf{0.008 \pm 0.025}$
Branin	1.35 ± 1.16	0.71 ± 1.16	2.20 ± 0.81	$\mathbf{0.31 \pm 0.29}$
Bukin	1.45 ± 1.13	0.59 ± 1.20	1.27 ± 0.80	$\mathbf{0.92 \pm 0.54}$
Cross-in-Tray	1.56 ± 1.39	2.03 ± 1.82	1.79 ± 1.49	$\mathbf{1.38 \pm 0.97}$
Eggholder	3.08 ± 0.55	3.11 ± 0.55	1.87 ± 0.94	$\mathbf{1.83 \pm 0.96}$
Holder Table	3.21 ± 1.38	3.20 ± 1.38	1.56 ± 1.62	$\mathbf{1.22 \pm 1.01}$
Levy13	2.36 ± 1.22	1.06 ± 1.22	1.29 ± 1.00	$\mathbf{0.35 \pm 0.31}$

6.3. Scalability to Higher Dimension

To demonstrate the computational scalability of our joint optimization approach (as shown in Prob. (26)), we consider a set of higher dimensional problems. Due to space limitation, we show the results for the optimization of 12-dimensional black-box function sampled from a Gaussian process with squared exponential kernel function. More results can be found in Appendix P.1 and Appendix P.2. The optimization domain is set to be $[0, 10]^{12}$. We run 10 randomly sampled instances for 100 steps. The average update time per step is only 18.0 seconds on a personal computer with one `Intel64 Family 6 Model 142 Stepping 12 GenuineIntel 1803 Mhz` processor and `16.0 GB RAM`. This is comparably very small considering that each query to the comparison oracle can be very expensive in practice (e.g., heating the room up to a certain temperature to evaluate occupant comfort, which may take tens of minutes). We compare our method to the SGP baseline, which is one of the state-of-the-art computationally practical preferential Bayesian optimization method. Fig. 3 shows the cumulative regret (in log scale) and the suboptimality of the reported solution for the problem. It can be seen that our algorithm still achieves sublinear regret growth and good convergence for the suboptimality of the reported solution within 100 steps in this 12-dimensional problem. Fig. 3 also shows that our POP-BO has faster convergence speed in higher dimensional problem and thus scales better than the SGP method.

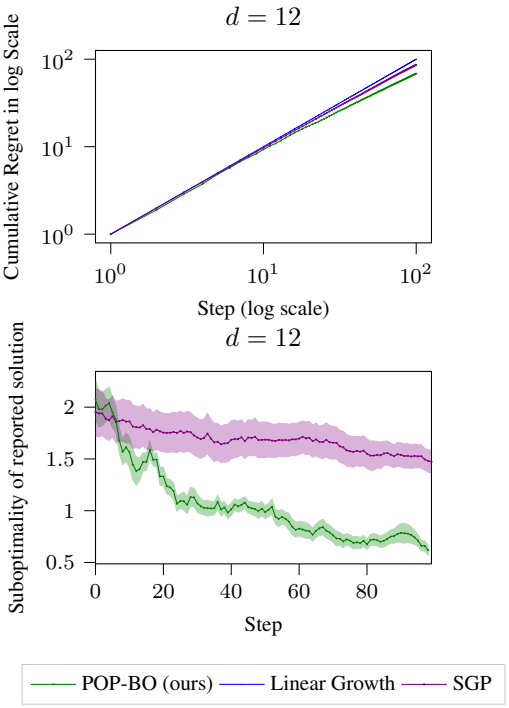

Figure 3. Cumulative regret in log scale and the suboptimality of the reported solution in linear scale for a 12-dimensional problem sampled from Gaussian process. For reference purpose, we also plot T in the cumulative regret plot in log scale, where the shaded areas represent ± 0.2 standard deviation.

7. Conclusion and Future Work

In this paper, we have presented a principled optimistic preferential Bayesian optimization algorithm, based on the likelihood-based confidence set. An efficient computational method is developed to implement the algorithm. We further show an information-theoretic bound on the cumulative regret, a *first-of-its-kind* for preferential BO. We also design a scheme to report an estimated optimal solution, with a guaranteed convergence rate. Experimental results show that our method achieves better or competitive performance as compared to the state-of-the-art heuristics, which, however, do not have theoretical guarantees on regret. Future works include the extension to the safety-critical problem (Berkenkamp et al., 2016; Guo et al., 2023) and game theoretical setting. The likelihood-based confidence set and the error bound in Sec. 3 can also be applied to more scenarios with preference feedback.

Acknowledgements

This research was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545, the Swiss Federal Office of Energy SFOE as part of the SWEET consortium SWICE, and in part by the Swiss Data Science Center, grant agreement C20-13.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abdelrahman, M. M. and Miller, C. Targeting occupant feedback using digital twins: Adaptive spatial–temporal thermal preference sampling to optimize personal comfort models. *Building and Environment*, 218:109090, 2022.

Andersson, J. A., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. CasADi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36, 2019.

Astudillo, R., Lin, Z. J., Bakshy, E., and Frazier, P. qEUBO: A decision-theoretic acquisition function for preferential Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1093–1114. PMLR, 2023.

Berkenkamp, F., Schoellig, A. P., and Krause, A. Safe controller optimization for quadrotors with Gaussian processes. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 491–496. IEEE, 2016.

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

Bull, A. D. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017a.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*, 2017b.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Christmann, A. and Hable, R. Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis*, 56(4):854–873, 2012.

Curi, S., Berkenkamp, F., and Krause, A. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Dixon, L. C. W. The global optimization problem: an introduction. *Towards Global Optimiation 2*, pp. 1–15, 1978.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Edmunds, D. E. and Triebel, H. *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge Univ Pr, 1996.

Emmenegger, N., Mutny, M., and Krause, A. Likelihood ratio confidence sets for sequential decision making. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Fanger, P. O. et al. Thermal comfort. analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

Frazier, P. I. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Gajane, P., Urvoy, T., and Clérot, F. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*, pp. 218–227. PMLR, 2015.

González, J., Dai, Z., Damianou, A., and Lawrence, N. D. Preferential Bayesian optimization. In *International Conference on Machine Learning*, pp. 1282–1291. PMLR, 2017.

GPy. GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy, since 2012.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in Neural Information Processing Systems*, 26, 2013.

Guo, B., Jiang, Y., Kamgarpour, M., and Ferrari-Trecate, G. Safe zeroth-order convex optimization using quadratic local approximations. In *2023 European Control Conference (ECC)*, pp. 1–8. IEEE, 2023.

Hiranaka, A., Hwang, M., Lee, S., Wang, C., Fei-Fei, L., Wu, J., and Zhang, R. Primitive skill-based robot learning from human evaluative feedback. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7817–7824. IEEE, 2023.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pp. 99–127. World Scientific, 2013.

Kandasamy, K., Schneider, J., and Póczos, B. High dimensional Bayesian optimisation and bandits via additive models. In *International Conference on Machine Learning*, pp. 295–304. PMLR, 2015.

Koyama, Y., Sato, I., and Goto, M. Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics (TOG)*, 39(4):88–1, 2020.

Lalley, S. P. Concentration inequalities. *Lecture notes, University of Chicago*, 2013.

Li, K., Tucker, M., Bıyık, E., Novoseller, E., Burdick, J. W., Sui, Y., Sadigh, D., Yue, Y., and Ames, A. D. ROIAL: Region of interest active learning for characterizing exoskeleton gait preference landscapes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3212–3218. IEEE, 2021.

Lichtenstein, S. and Slovic, P. Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, 89(1):46, 1971.

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

Lyu, J., Shi, Y., Du, H., and Lian, Z. Sex-based thermal comfort zones and energy savings in spaces with joint operation of air conditioner and fan. *Building and Environment*, 246:111002, 2023. ISSN 0360-1323. doi: https://doi.org/10.1016/j.buildenv.2023.111002. URL https://www.sciencedirect.com/science/article/pii/S0360132323010296.

Maddalena, E. T., Scharnhorst, P., and Jones, C. N. Deterministic error bounds for kernel-based learning techniques under bounded noise. *Automatica*, 134:109896, 2021.

Mehta, V., Neopane, O., Das, V., Lin, S., Schneider, J., and Neiswanger, W. Kernelized offline contextual dueling bandits. *arXiv preprint arXiv:2307.11288*, 2023.

Mikkola, P., Todorović, M., Järvi, J., Rinke, P., and Kaski, S. Projective preferential Bayesian optimization. In *International Conference on Machine Learning*, pp. 6884–6892. PMLR, 2020.

Molga, M. and Smutnicki, C. Test functions for optimization needs. *Test functions for optimization needs*, 101:48, 2005.

Negoescu, D. M., Frazier, P. I., and Powell, W. B. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

Newey, W. K. and McFadden, D. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111–2245, 1994.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the Eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Owen, A. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 1413–1423. PMLR, 2021.

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

Saha, A., Koren, T., and Mansour, Y. Dueling convex optimization. In *International Conference on Machine Learning*, pp. 9245–9254. PMLR, 2021.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pp. 416–426. Springer, 2001.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Inf. Process. Syst.*, 25, 2012.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

Sui, Y., Burdick, J., Yue, Y., et al. Stage-wise safe Bayesian optimization with Gaussian processes. In *Proc. of the Int. Conf. on Mach. Learn.*, pp. 4781–4789, 2018.

Takeno, S., Nomura, M., and Karasuyama, M. Towards practical preferential Bayesian optimization with skew Gaussian processes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 33516–33533, 2023.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185 (4157):1124–1131, 1974.

Wächter, A. and Biegler, L. T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

Wang, Y., Liu, Q., and Jin, C. Is RLHF more difficult than standard RL? A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Warnell, G., Waytowich, N., Lawhern, V., and Stone, P. Deep TAMER: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Wu, C., Li, T., Zhang, Z., and Yu, Y. Bayesian optimistic optimization: Optimistic exploration for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14210–14223, 2022.

Wu, Y. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.

Xu, W., Jiang, Y., Maddalena, E. T., and Jones, C. N. Lower bounds on the worst-case complexity of efficient global optimization. *arXiv preprint arXiv:2209.09655*, 2022a.

Xu, W., Jones, C. N., Svetozarevic, B., Laughman, C. R., and Chakrabarty, A. VABO: Violation-aware Bayesian optimization for closed-loop control performance optimization with unmodeled constraints. In *2022 American Control Conference (ACC)*, pp. 5288–5293. IEEE, 2022b.

Xu, W., Jiang, Y., Svetozarevic, B., and Jones, C. Constrained efficient global optimization of expensive black-box functions. In *International Conference on Machine Learning*, pp. 38485–38498. PMLR, 2023.

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

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Zhang, H., Lee, S., and Tzempelikos, A. Bayesian meta-learning for personalized thermal comfort modeling. *Building and Environment*, 249:111129, February 2024. ISSN 03601323. doi: 10.1016/j.buildenv.2023.111129. URL https://linkinghub.elsevier.com/retrieve/pii/S0360132323011563.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Zhou, X. and Ji, B. On kernelized multi-armed bandits with constraints. *Advances in Neural Information Processing Systems*, 35, 2022.

Zhu, B., Jordan, M., and Jiao, J. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 43037–43067, 23–29 Jul 2023.

Without further notice, all the results shown in this appendix are under the assumptions 2.1, 2.2, 2.4, and 2.5.

A. Preliminaries

To prepare for the proofs of the main results shown in this paper, we first state several useful lemmas.

Lemma A.1. *The function $\psi(y, y') = \log(e^y + e^{y'})$ is convex in (y, y').*

Proof. We calculate the Hessian of the function ψ and derive

$$\nabla^2\psi = \frac{e^{y+y'}}{(e^y + e^{y'})^2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \succcurlyeq 0. \tag{34}$$

Hence, ψ is convex.

□

Therefore, we can see $\ell(Z_{0:t}|\mathcal{D}_t)$ is concave in $Z_{0:t}$.

Lemma A.2. $\forall \tilde{f} \in \mathcal{B}_f, x \in \mathcal{X}, \tilde{f}(x) \in [-B, B]$.

Proof. $|\tilde{f}(x)| = |\langle \tilde{f}, k(x, \cdot)\rangle| \leq \left\|\tilde{f}\right\| \|k(x, \cdot)\| \leq B\sqrt{k(x, x)} \leq B$, where the first inequality follows by Cauchy–Schwarz inequality, the second inequality follows by Assump. 2.2, and the last inequality follows by Assump. 2.4. □

B. Properties of the Function $\sigma(\cdot)$

When applying the function σ to the difference of objective function $\tilde{f}(x) - \tilde{f}(x'), \forall \tilde{f} \in \mathcal{B}_f$, we have the calculations by single variable calculus,

$$\begin{aligned} u &:= \tilde{f}(x) - \tilde{f}(x') \in [-2B, 2B], \\ \sigma(u) &\in [\underline{\sigma}, \bar{\sigma}], \\ \sigma'(u) &= \frac{1}{2 + e^u + e^{-u}} \in [\underline{\sigma'}, \bar{\sigma'}], \end{aligned}$$

where $\underline{\sigma} = 1/(1+e^{2B}), \bar{\sigma} = 1/(1+e^{-2B})$ and $\underline{\sigma'} = 1/(2+e^{2B}+e^{-2B}), \bar{\sigma'} = 1/4$. We also introduce some constants $B_p = \frac{\bar{\sigma}}{\underline{\sigma}} - \frac{\underline{\sigma}}{\bar{\sigma}}$, $H_\sigma = \frac{1}{2\bar{\sigma}^2}$ and $C_L = 1 + \frac{2}{1+e^{-2B}}$, which will be used in the proof.

C. Proof of Thm. 3.1

To prepare for the proof of the theorem, we first prove several lemmas.

Lemma C.1. *For any fixed $\hat{f} \in \mathcal{B}_f$, we have,*

$$\mathbb{P}\left(\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sqrt{32tB^2 \log \frac{1}{\delta_t}}\right) \geq 1 - \delta_t, \tag{35}$$

where f is the ground-truth function.

Proof. We use y_τ (y'_τ resp.) to denote $f(x_\tau)$ ($f(x'_\tau)$ resp.). We use z_τ (z'_τ resp.) to denote $\hat{f}(x_\tau)$ ($\hat{f}(x'_\tau)$ resp.). And we use p_τ to denote $\sigma(y_\tau - y'_\tau)$.

$$\begin{aligned} &\mathbb{P}\left(\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \xi\right) \\ =&\mathbb{P}\left(\sum_{\tau=1}^t ((z_\tau - y_\tau)\mathbf{1}_\tau + (z'_\tau - y'_\tau)(1 - \mathbf{1}_\tau)) - \sum_{\tau=1}^t \log\left(e^{z_\tau} + e^{z'_\tau}\right) + \sum_{\tau=1}^t \log\left(e^{y_\tau} + e^{y'_\tau}\right) \leq \xi\right) \end{aligned}$$

$$=\mathbb{P}\left(\sum_{\tau=1}^{t}((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-\sum_{\tau=1}^{t}((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\le\xi'\right)$$

where $\xi'=\xi+\sum_{\tau=1}^{t}\log\left(e^{z_\tau}+e^{z'_\tau}\right)-\sum_{\tau=1}^{t}\log\left(e^{y_\tau}+e^{y'_\tau}\right)-\sum_{\tau=1}^{t}((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))$, and the probability $\mathbb{P}$ is taken with respect to the randomness from the comparison oracle and the randomness from the algorithm.

It can be checked that $\psi_\tau(y,y'):=\log\left(e^{y}+e^{y'}\right)-p_\tau y-(1-p_\tau)y'$ is a convex function and $\nabla\psi_\tau(y_\tau,y'_\tau)=0$. This implies that (y_τ,y'_τ) achieves the minimum for the convex function ψ_τ. Therefore,

$$\log\left(e^{y_\tau}+e^{y'_\tau}\right)-p_\tau y_\tau-(1-p_\tau)y'_\tau\le\log\left(e^{z_\tau}+e^{z'_\tau}\right)-p_\tau z_\tau-(1-p_\tau)z'_\tau.$$

Rearrangement gives,

$$\log\left(e^{z_\tau}+e^{z'_\tau}\right)-\log\left(e^{y_\tau}+e^{y'_\tau}\right)-((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\ge 0.$$

Hence, $\xi'\ge\xi$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\log\mathbb{P}_{\hat{f}}((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})-\log\mathbb{P}_f((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})\le\xi\right)\\
=&\mathbb{P}\left(\sum_{\tau=1}^{t}((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-\sum_{\tau=1}^{t}((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\le\xi'\right)\\
\ge&\mathbb{P}\left(\sum_{\tau=1}^{t}((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-\sum_{\tau=1}^{t}((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\le\xi\right)
\end{aligned}$$

We further notice that

$$\mathbb{E}[((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\,|\mathcal{F}_{\tau-1}]=0,\tag{36}$$

and with probability one,

$$|((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))|=|(z_\tau-y_\tau-z'_\tau+y'_\tau)(\mathbf{1}_\tau-p_\tau)|\le 4B.\tag{37}$$

We can thus apply the Azuma-Hoeffding inequality (see, e.g., (Lalley, 2013)). By Azuma–Hoeffding inequality,

$$\begin{aligned}
&\mathbb{P}\left(\log\mathbb{P}_{\hat{f}}((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})-\log\mathbb{P}_f((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})\le\xi\right)\\
\ge&\mathbb{P}\left(\sum_{\tau=1}^{t}((z_\tau-y_\tau)\mathbf{1}_\tau+(z'_\tau-y'_\tau)(1-\mathbf{1}_\tau))-\sum_{\tau=1}^{t}((z_\tau-y_\tau)p_\tau+(z'_\tau-y'_\tau)(1-p_\tau))\le\xi\right)\\
\ge&1-\exp\left\{-\frac{\xi^2}{32tB^2}\right\}.
\end{aligned}$$

Set $\exp\left\{-\frac{\xi^2}{32tB^2}\right\}=\delta_t$. That is, $\xi=\sqrt{32tB^2\log\frac{1}{\delta_t}}$. We then get the desired result. □

We then have the following high probability confidence set lemma.

Lemma C.2. *For any fixed $\hat{f}\in\mathcal{B}_f$ that is independent of $((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})$, we have, with probability at least $1-\delta$,*

$$\log\mathbb{P}_{\hat{f}}((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})-\log\mathbb{P}_f((x_\tau,x'_\tau,\mathbf{1}_\tau)_{\tau=1}^{t})\le\sqrt{32tB^2\log\frac{\pi^2t^2}{6\delta}},\ \forall t\ge 1.\tag{38}$$

Proof. We use $\mathcal{E}_t$ to denote the event $\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sqrt{32tB^2 \log \frac{1}{\delta_t}}$. We pick $\delta_t = {}^{(6\delta)}/_{(\pi^2 t^2)}$ and have,

$$
\begin{aligned}
&\mathbb{P}\left(\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sqrt{32tB^2 \log \frac{1}{\delta_t}}, \forall t \geq 1\right) \\
=&1 - \mathbb{P}\left(\overline{\cap_{t=1}^\infty \mathcal{E}_t}\right) \\
=&1 - \mathbb{P}\left(\cup_{t=1}^\infty \overline{\mathcal{E}_t}\right) \\
\geq&1 - \sum_{t=1}^\infty \mathbb{P}\left(\overline{\mathcal{E}_t}\right) \\
=&1 - \sum_{t=1}^\infty \left(1 - \mathbb{P}\left(\mathcal{E}_t\right)\right) \\
=&1 - \sum_{t=1}^\infty \left(1 - \mathbb{P}\left(\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sqrt{32tB^2 \log \frac{1}{\delta_t}}\right)\right) \\
\geq&1 - \sum_{t=1}^\infty \delta_t \\
=&1 - \frac{6\delta}{\pi^2} \sum_{t=1}^\infty \frac{1}{t^2} \\
=&1 - \delta.
\end{aligned}
$$

□

We then have a lemma to bound the difference of log likelihood when two functions are close in infinity-norm sense.

Lemma C.3. *There exists an independent constant $C_L > 0$, such that, $\forall \epsilon > 0$, $\forall f_1, f_2 \in \mathcal{B}_f$ that satisfies $\|f_1 - f_2\|_\infty \leq \epsilon$, we have,*

$$
\log \mathbb{P}_{f_1}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{f_2}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq C_L \epsilon t. \tag{39}
$$

Proof. We use $z_{i,\tau}$ ($z'_{i,\tau}$, resp.) to denote $f_i(x_\tau)$ ($f_i(x'_\tau)$, resp.), $\forall i \in \{0, 1\}$.

$$
\begin{aligned}
&\log \mathbb{P}_{f_1}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{f_2}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \\
=&\sum_{\tau=1}^t \left((z_{1,\tau} - z_{2,\tau})\mathbf{1}_\tau + (z'_{1,\tau} - z'_{2,\tau})(1 - \mathbf{1}_\tau)\right) - \sum_{\tau=1}^t \log\left(e^{z_{1,\tau}} + e^{z'_{1,\tau}}\right) + \sum_{\tau=1}^t \log\left(e^{z_{2,\tau}} + e^{z'_{2,\tau}}\right) && (40) \\
\leq&\epsilon t + \sum_{\tau=1}^t \max_{z,z' \in [-B,B]} \left\|\nabla_{z,z'} \log\left(e^z + e^{z'}\right)\right\| \left\|(z_{1,\tau}, z'_{1,\tau}) - (z_{2,\tau}, z'_{2,\tau})\right\| && (41) \\
\leq&\epsilon t + \sum_{\tau=1}^t \frac{\sqrt{2}}{1 + e^{-2B}} \sqrt{2}\epsilon && (42) \\
=&\left(1 + \frac{2}{1 + e^{-2B}}\right) \epsilon t, && (43)
\end{aligned}
$$

where the equality (40) follows by the definition of log-likelihood function, and the inequality (41) follows by the assumption and the mean-value theorem. The conclusion follows by setting $C_L = 1 + \frac{2}{1+e^{-2B}}$.

□

Main proof: We use $\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)$ to denote the covering number of the set $\mathcal{B}_f$, with $(f_i^\epsilon)_{i=1}^{\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}$ be a set of ϵ-covering for the set $\mathcal{B}_f$. Reset the 'δ' in Lem. C.2 as ${}^{\delta}/_{\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}$ and applying the probability union bound, we have,

with probability at least $1-\delta$, $\forall f_i^\epsilon, t \geq 1$,

$$\log \mathbb{P}_{f_i^\epsilon}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sqrt{32tB^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{6\delta}}. \tag{44}$$

By the definition of ϵ-covering, there exists $j \in [\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)]$, such that,

$$\|\hat{f}_t^{\mathrm{MLE}} - f_j^\epsilon\|_\infty \leq \epsilon. \tag{45}$$

Hence, with probability at least $1-\delta$,

$$\begin{aligned}
&\log \mathbb{P}_{\hat{f}_t^{\mathrm{MLE}}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \\
=&\log \mathbb{P}_{\hat{f}_t^{\mathrm{MLE}}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{f_j^\epsilon}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) + \log \mathbb{P}_{f_j^\epsilon}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \\
\leq& C_L \epsilon t + \sqrt{32tB^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{6\delta}},
\end{aligned}$$

where the inequality follows by Lem. C.3 and the inequality (44).

D. Proof of Lem. 3.5

We first have a lemma.

Lemma D.1. *We have,*

$$\log \hat{p} - \log p \leq \frac{1}{p}(\hat{p} - p) - H_\sigma (\hat{p} - p)^2, \forall p, \hat{p} \in [\underline{\sigma}, \bar{\sigma}], \tag{46}$$

where $H_\sigma = \frac{1}{2\bar{\sigma}^2}$.

Proof. Let $\zeta(\hat{p}) = \log \hat{p} - \log p - \frac{1}{p}(\hat{p} - p) + H_\sigma (\hat{p} - p)^2, \forall p, \hat{p} \in [\underline{\sigma}, \bar{\sigma}]$. We have,

$$\zeta'(\hat{p}) = \frac{1}{\hat{p}} - \frac{1}{p} + 2H_\sigma(\hat{p} - p) = (\hat{p} - p)\left(\frac{1}{\bar{\sigma}^2} - \frac{1}{\hat{p}p}\right), \forall \hat{p} \in [\underline{\sigma}, \bar{\sigma}].$$

Since $\forall p, \hat{p} \in [\underline{\sigma}, \bar{\sigma}]$, we have $\frac{1}{\bar{\sigma}^2} - \frac{1}{\hat{p}p} \leq 0$. Hence, $\zeta'(\hat{p}) \geq 0, \forall \hat{p} \in [\underline{\sigma}, p]$ and $\zeta'(\hat{p}) \leq 0, \forall \hat{p} \in [p, \bar{\sigma}]$. Therefore, $\zeta(\hat{p})$ achieves the maximum over $[\underline{\sigma}, \bar{\sigma}]$ at the point p. So $\zeta(\hat{p}) \leq \zeta(p) = 0$. Rearrangement then gives the desired result. □

For any fixed function $\hat{f} \in \mathcal{B}_f$, we use the notations $\hat{p}_\tau = \sigma(\hat{f}(x_\tau) - \hat{f}(x'_\tau)) \in [\underline{\sigma}, \bar{\sigma}]$ and $p_\tau = \sigma(f(x_\tau) - f(x'_\tau)) \in [\underline{\sigma}, \bar{\sigma}]$. We have,

$$\begin{aligned}
&\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \\
=&\sum_{\tau=1}^t \Big(\log p_{\hat{f}}(x_\tau, x'_\tau, \mathbf{1}_\tau) - \log p_f(x_\tau, x'_\tau, \mathbf{1}_\tau)\Big) \\
=&\sum_{\tau=1}^t \left(\mathbf{1}_\tau \left(\log \hat{p}_\tau - \log p_\tau\right) + (1 - \mathbf{1}_\tau)\left(\log\left(1 - \hat{p}_\tau\right) - \log(1 - p_\tau)\right)\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \\
=&\sum_{\tau=1}^t \left(\mathbf{1}_\tau \left(\log \hat{p}_\tau - \log p_\tau\right) + (1 - \mathbf{1}_\tau)\left(\log\left(1 - \hat{p}_\tau\right) - \log(1 - p_\tau)\right)\right) \\
\leq&\sum_{\tau=1}^t \left(\mathbf{1}_\tau \left(\frac{\hat{p}_\tau - p_\tau}{p_\tau} - H_\sigma\left(\hat{p}_\tau - p_\tau\right)^2\right) + (1 - \mathbf{1}_\tau)\left(\frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} - H_\sigma\left(\hat{p}_\tau - p_\tau\right)^2\right)\right)
\end{aligned}$$

Rearrangement gives,

$$H_\sigma \sum_{\tau=1}^{t} (\hat{p}_\tau - p_\tau)^2 + \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \leq \sum_{\tau=1}^{t} \left(\mathbf{1}_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - \mathbf{1}_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} \right). \tag{47}$$

We then have the following lemma,

Lemma D.2. *For any fixed $\hat{f} \in \mathcal{B}_f$ and $\forall t \geq 1$, we have, with probability at least $1 - \delta_t$,*

$$\mathbb{P}\left(H_\sigma \sum_{\tau=1}^{t} (\hat{p}_\tau - p_\tau)^2 \leq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) + \sqrt{2tB_p^2 \log \frac{1}{\delta_t}} \right) \geq 1 - \delta_t. \tag{48}$$

Proof. Since $\mathbb{E}\left[\mathbf{1}_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - \mathbf{1}_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} | \mathcal{F}_{\tau-1} \right] = \mathbb{E}\left[p_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - p_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} | \mathcal{F}_{\tau-1} \right] = 0$ and with probability one,

$$\left| \mathbf{1}_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - \mathbf{1}_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} \right| \leq \mathbf{1}_\tau \left| \frac{\hat{p}_\tau - p_\tau}{p_\tau} \right| + (1 - \mathbf{1}_\tau) \left| \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} \right| \tag{49}$$

$$= \mathbf{1}_\tau \left| \frac{\hat{p}_\tau}{p_\tau} - 1 \right| + (1 - \mathbf{1}_\tau) \left| \frac{1 - \hat{p}_\tau}{1 - p_\tau} - 1 \right| \tag{50}$$

$$\leq \frac{\bar{\sigma}}{\underline{\sigma}} - \frac{\underline{\sigma}}{\bar{\sigma}} = B_p, \tag{51}$$

where the last inequality follows by that $\hat{p}_\tau, p_\tau, 1 - \hat{p}_\tau, 1 - p_\tau \in [\underline{\sigma}, \bar{\sigma}]$. Thus we can apply the Azuma–Hoeffding inequality. By Azuma–Hoeffding inequality, we have,

$$\mathbb{P}\left(\sum_{\tau=1}^{t} \left(\mathbf{1}_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - \mathbf{1}_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} \right) \leq \xi \right) \geq 1 - \exp\left\{ -\frac{\xi^2}{2tB_p^2} \right\}. \tag{52}$$

We set $\exp\left\{ -\frac{\xi^2}{2tB_p^2} \right\} = \delta_t$, and derive

$$\mathbb{P}\left(\sum_{\tau=1}^{t} \left(\mathbf{1}_\tau \frac{\hat{p}_\tau - p_\tau}{p_\tau} + (1 - \mathbf{1}_\tau) \frac{p_\tau - \hat{p}_\tau}{1 - p_\tau} \right) \leq \sqrt{2tB_p^2 \log \frac{1}{\delta_t}} \right) \geq 1 - \delta_t. \tag{53}$$

Combining the inequality (47) and the inequality (53), the desired result is derived.

□

Lemma D.3. *For any fixed $\hat{f} \in \mathcal{B}_f$, we have, with probability at least $1 - \delta$,*

$$H_\sigma \sum_{\tau=1}^{t} (\hat{p}_\tau - p_\tau)^2 \leq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) + \sqrt{2tB_p^2 \log \frac{\pi^2 t^2}{6\delta}}, \quad \forall t \geq 1. \tag{54}$$

Proof. We use $\mathcal{E}_t$ [5] to denote the event $H_\sigma \sum_{\tau=1}^{t} (\hat{p}_\tau - p_\tau)^2 \leq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) + \sqrt{2tB_p^2 \log \frac{1}{\delta_t}}$ and pick $\delta_t = {}^{(6\delta)}/_{(\pi^2 t^2)}$. We have,

$$\begin{aligned}
&\mathbb{P}\left(H_\sigma \sum_{\tau=1}^{t} (\hat{p}_\tau - p_\tau)^2 \leq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) - \log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) + \sqrt{2tB_p^2 \log \frac{1}{\delta_t}}, \forall t \geq 1 \right) \\
=&1 - \mathbb{P}\left(\overline{\cap_{t=1}^{\infty} \mathcal{E}_t} \right) \\
=&1 - \mathbb{P}\left(\cup_{t=1}^{\infty} \overline{\mathcal{E}_t} \right)
\end{aligned}$$

[5] With abuse of notation here. $\mathcal{E}_t$ is only a local notation in this proof here.

$$
\begin{aligned}
\geq & 1-\sum_{t=1}^{\infty} \mathbb{P}\left(\overline{\mathcal{E}_t}\right) \\
= & 1-\sum_{t=1}^{\infty}\left(1-\mathbb{P}\left(\mathcal{E}_t\right)\right) \\
= & 1-\sum_{t=1}^{\infty}\left(1-\mathbb{P}\left(H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau-p_\tau\right)^2 \leq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)-\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)+\sqrt{2tB_p^2 \log \frac{1}{\delta_t}}\right)\right) \\
\geq & 1-\sum_{t=1}^{\infty} \delta_t \\
= & 1-\frac{6\delta}{\pi^2} \sum_{t=1}^{\infty} \frac{1}{t^2} \\
= & 1-\delta.
\end{aligned}
$$

□

Main Proof: Resetting the 'δ' in Lem. D.3 to be $\delta / \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)$, we can guarantee the Eq. (54) holds for all the function in an ϵ-covering of $\mathcal{B}_f$ with probability at least $1-\delta$, by applying the probability union bound.

For any $\hat{f}_{t+1} \in \mathcal{B}_f^{t+1} \subset \mathcal{B}_f$, there exists a function in the ϵ-covering of $\mathcal{B}_f$, which we set to be $\hat{f}$, such that $\|\hat{f}_{t+1}-\hat{f}\|_\infty \leq \epsilon$. We also use $\hat{p}_\tau^{t+1}$ to denote $\sigma(\hat{f}_{t+1}(x_\tau)-\hat{f}_{t+1}(x'_\tau))$. Thus, we have,

$$
\begin{aligned}
& H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau^{t+1}-p_\tau\right)^2 && (55)\\
\leq & 2H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau^{t+1}-\hat{p}_\tau\right)^2+2H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau-p_\tau\right)^2 && (56)\\
\leq & 2H_\sigma \bar{\sigma'}^2 \sum_{\tau=1}^{t}\left((\hat{f}_{t+1}(x_\tau)-\hat{f}_{t+1}(x'_\tau))-(\hat{f}(x_\tau)-\hat{f}(x'_\tau))\right)^2+2H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau-p_\tau\right)^2 && (57)\\
\leq & 8H_\sigma \bar{\sigma'}^2 \sum_{\tau=1}^{t} \epsilon^2+2H_\sigma \sum_{\tau=1}^{t}\left(\hat{p}_\tau-p_\tau\right)^2 && (58)\\
\leq & 8H_\sigma \bar{\sigma'}^2 \epsilon^2 t+\sqrt{8tB_p^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{6\delta}}+2\left(\log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)-\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)\right) && (59)\\
\leq & C(\epsilon, \delta, t)+2\left(\log \mathbb{P}_{\hat{f}_t^{\mathrm{MLE}}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)-\log \mathbb{P}_{\hat{f}_{t+1}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)\right) && (60)\\
& +2\left(\log \mathbb{P}_{\hat{f}_{t+1}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)-\log \mathbb{P}_{\hat{f}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t)\right) \\
\leq & C(\epsilon, \delta, t)+2C_L \epsilon t+2\beta_1(\epsilon, \delta, t) && (61)\\
= & \beta_2(\epsilon, \delta, t)+2\beta_1(\epsilon, \delta, t), && (62)
\end{aligned}
$$

where $C(\epsilon, \delta, t)=8H_\sigma \bar{\sigma'}^2 \epsilon^2 t+\sqrt{8tB_p^2 \log \frac{\pi^2 t^2 \mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{6\delta}}$ and $\beta_2(\epsilon, \delta, t)=C(\epsilon, \delta, t)+2C_L \epsilon t$. The inequality (56) follows by the fact that $(a+b)^2 \leq 2a^2+2b^2, \forall a, b \in \mathbb{R}$. The inequality (58) follows because $\|\hat{f}_{t+1}-\hat{f}\|_\infty \leq \epsilon$. The inequality (59) follows by Lem. D.3 (with reset of 'δ'). The inequality (60) follows by that

$$
\log \mathbb{P}_{\hat{f}_t^{\mathrm{MLE}}}((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t) \geq \log \mathbb{P}_f((x_\tau, x'_\tau, \mathbf{1}_\tau)_{\tau=1}^t).
$$

The inequality (61) follows by the fact that $\hat{f}_{t+1} \in \mathcal{B}_f^{t+1}$ and Lem. C.3.

Furthermore,

$$
\sum_{\tau=1}^{t}\left(\hat{p}_\tau^{t+1}-p_\tau\right)^2=\sum_{\tau=1}^{t}\left(\sigma\left(\hat{f}_{t+1}(x_\tau)-\hat{f}_{t+1}(x'_\tau)\right)-\sigma\left(f(x_\tau)-f(x'_\tau)\right)\right)^2 \tag{63}
$$

$$\geq \sum_{\tau=1}^{t} \underline{\sigma'}^2 \left(\left(\hat{f}_{t+1}(x_\tau) - \hat{f}_{t+1}(x'_\tau) \right) - (f(x_\tau) - f(x'_\tau)) \right)^2, \tag{64}$$

where the inequality follows by mean value theorem. The conclusion then follows.

E. Proof of Thm. 3.6

Before we proceed to prove Thm. 3.6, we first conduct a black-box analysis in Sec. E.1 to bound the pointwise error for a generic RKHS with a generic learning scheme, which we think can be of independent interest.

E.1. Black-box Analysis on the Pointwise Inference Error in a Generic RKHS

Suppose we have a generic RKHS $\tilde{\mathcal{H}}$ with a generic positive semidefinite kernel function $\tilde{k}(\cdot,\cdot)$. After obtaining some information (preference information in this paper) on a sequence $\tilde{x}_1, \tilde{x}_2, \cdot, \tilde{x}_{t-1}$, a learning scheme outputs a learnt uncertainty set,

$$\mathcal{S}_t = \{\tilde{h} \in \mathcal{B} | \sum_{\tau=1}^{t-1} \left(\tilde{h}(\tilde{x}_\tau) - h(\tilde{x}_\tau) \right)^2 \leq \tilde{\beta}_t \}, \tag{65}$$

where $\mathcal{B}$ is a function space ball with radius $\tilde{B}$ in $\tilde{\mathcal{H}}$, $h \in \mathcal{B}$ is the ground truth function and $\tilde{\beta}_t$ quantifies the size of this confidence set. Let $\tilde{\mathcal{X}}$ denote the function input set, which is assumed to be compact. We introduce the function,

$$\tilde{\sigma}_t^2(\tilde{x}) = \tilde{k}\left(\tilde{x}, \tilde{x}\right) - \tilde{k}(\tilde{x}_{1:t-1}, \tilde{x})^\top \left(\tilde{K}_{t-1} + \lambda I \right)^{-1} \tilde{k}\left(\tilde{x}_{1:t-1}, \tilde{x}\right), \tag{66}$$

where λ is a positive constant and $\tilde{K}_{t-1} = (\tilde{k}(\tilde{x}_i, \tilde{x}_j))_{i\in[t-1], j\in[t-1]}$. We have the following theorem.

Theorem E.1. *$\forall \tilde{h} \in \mathcal{S}_{t+1}, \tilde{x} \in \tilde{\mathcal{X}}$, we have,*

$$|\tilde{h}(\tilde{x}) - h(\tilde{x})| \leq 2(\tilde{B} + \lambda^{-1/2}\tilde{\beta}_{t+1}^{1/2})\tilde{\sigma}_{t+1}(\tilde{x}). \tag{67}$$

Proof. For simplicity, we use $\phi(\tilde{x})$ to denote the function $\tilde{k}(\tilde{x}, \cdot)$, where $\phi : \mathbb{R}^{\tilde{d}} \to \tilde{\mathcal{H}}$ maps a finite dimensional point $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ to the RKHS $\tilde{\mathcal{H}}$. For simplicity, we use $h_1^\top h_2$ to denote the inner product of two functions h_1, h_2 from the RKHS $\tilde{\mathcal{H}}$. Therefore, $h(\tilde{x}) = \langle h, \tilde{k}(\tilde{x}, \cdot)\rangle_{\tilde{k}} = h^\top \phi(\tilde{x})$ and $\tilde{k}(\tilde{x}, \tilde{\tilde{x}}) = \langle \tilde{k}(\tilde{x}, \cdot), \tilde{k}(\tilde{\tilde{x}}, \cdot)\rangle = \phi(\tilde{x})^\top \phi(\tilde{\tilde{x}})$, $\forall \tilde{x}, \tilde{\tilde{x}} \in \tilde{\mathcal{X}}$. We can introduce the feature map

$$\Phi_t := \left[\phi(\tilde{x}_1)^\top, \ldots, \phi(\tilde{x}_t)^\top\right]^\top,$$

we then get the kernel matrix $\tilde{K}_t = \Phi_t \Phi_t^\top = (\tilde{k}(\tilde{x}_i, \tilde{x}_j))_{i,j\in[t]}$, $\tilde{k}_t(\tilde{x}) = \Phi_t \phi(\tilde{x}) = (\tilde{k}(\tilde{x}, \tilde{x}_i))_{i\in[t]}$ for all $\tilde{x} \in \tilde{\mathcal{X}}$ and $h_{1:t} = \Phi_t h$.

Note that when the Hilbert space $\tilde{\mathcal{H}}$ is finite-dimensional, Φ_t is interpreted as the normal finite-dimensional matrix. In the more general setting where $\tilde{\mathcal{H}}$ can be an infinite-dimensional space, Φ_t is the evaluation operator $\tilde{\mathcal{H}} \to \mathbb{R}^t$ defined as $\Phi_t h := [h(\tilde{x}_1), \cdots, h(\tilde{x}_t)]^\top, \forall h \in \tilde{\mathcal{H}}$, with $\Phi_t^\top : \mathbb{R}^t \to \tilde{H}$ as its adjoint operator. For the simplicity of notation, we abuse the notation I to denote the identity mapping in both the RKHS $\tilde{\mathcal{H}}$ and $\mathbb{R}^t$. The specific meaning of I depends on the context.

Since the matrices/operators $(\Phi_t^\top \Phi_t + \lambda I)$ and $(\Phi_t \Phi_t^\top + \lambda I)$ are strictly positive definite and

$$(\Phi_t^\top \Phi_t + \lambda I)\Phi_t^\top = \Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I),$$

we have

$$\Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} = (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top. \tag{68}$$

Also from the definitions above $(\Phi_t^\top \Phi_t + \lambda I)\phi(\tilde{x}) = \Phi_t^\top \tilde{k}_t(\tilde{x}) + \lambda \phi(\tilde{x})$, and thus $\phi(\tilde{x}) = (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top \tilde{k}_t(\tilde{x}) + \lambda (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x})$. Hence, from Eq. (68) we deduce that

$$\phi(\tilde{x}) = \Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} \tilde{k}_t(\tilde{x}) + \lambda (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x}), \tag{69}$$

which gives

$$\phi(\tilde{x})^\top \phi(\tilde{x}) = \tilde{k}_t(\tilde{x})^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} \tilde{k}_t(\tilde{x}) + \lambda \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x}), \tag{70}$$

by multiplying both sides of Eq. (69) with $\phi(\tilde{x})^\top$. This implies

$$\lambda \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x}) = \tilde{k}(\tilde{x}, \tilde{x}) - \tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} \tilde{k}_t(\tilde{x}) = \tilde{\sigma}_{t+1}^2(\tilde{x}), \tag{71}$$

where the second equality follows by the definition of $\tilde{\sigma}_{t+1}(\tilde{x})$. Now observe that $\forall \tilde{h} \in \mathcal{B}$,

$$\begin{aligned}
&|\tilde{h}(\tilde{x}) - \tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} \tilde{h}_{1:t}| && (72)\\
=&|\phi(\tilde{x})^\top \tilde{h} - \phi(\tilde{x})^\top \Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} \Phi_t \tilde{h}| && (73)\\
=&|\phi(\tilde{x})^\top \tilde{h} - \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top \Phi_t \tilde{h}| && (74)\\
=&|\phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} (\Phi_t^\top \Phi_t + \lambda I) \tilde{h} - \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top \Phi_t \tilde{h}| && (75)\\
=&|\lambda \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \tilde{h}| && (76)\\
\leq& \left\|\lambda (\Phi_t^T \Phi_t + \lambda I)^{-1} \phi(\tilde{x})\right\|_{\tilde{k}} \left\|\tilde{h}\right\|_{\tilde{k}} && (77)\\
=& \left\|\tilde{h}\right\|_{\tilde{k}} \sqrt{\lambda \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \lambda I (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x})} && (78)\\
\leq& \tilde{B} \sqrt{\lambda \phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} (\Phi_t^\top \Phi_t + \lambda I) (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x})} && (79)\\
=& \tilde{B}\, \tilde{\sigma}_{t+1}(\tilde{x}), && (80)
\end{aligned}$$

where the equality (74) uses Eq. (68), the inequality (77) is by Cauchy-Schwartz, the inequality (79) follows by the assumption that $\|\tilde{h}\|_{\tilde{k}} \leq \tilde{B}$ and that $\Phi_t^\top \Phi_t$ is positive semidefinite, and the equality (80) is from Eq. (71). We define $\Delta_{1:t} = \tilde{h}_{1:t} - h_{1:t}$,

$$\begin{aligned}
&|\tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} \Delta_{1:t}| && (81)\\
=&|\phi(\tilde{x})^\top \Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} \Delta_{1:t}| && (82)\\
=&|\phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top \Delta_{1:t}| && (83)\\
\leq& \left\|(\Phi_t^\top \Phi_t + \lambda I)^{-1/2} \phi(\tilde{x})\right\|_{\tilde{k}} \left\|(\Phi_t^\top \Phi_t + \lambda I)^{-1/2} \Phi_t^\top \Delta_{1:t}\right\|_{\tilde{k}} && (84)\\
=& \sqrt{\phi(\tilde{x})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \phi(\tilde{x})} \sqrt{(\Phi_t^\top \Delta_{1:t})^\top (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top \Delta_{1:t}} && (85)\\
=& \lambda^{-1/2} \tilde{\sigma}_{t+1}(\tilde{x}) \sqrt{\Delta_{1:t}^\top \Phi_t \Phi_t^\top (\Phi_t \Phi_t^\top + \lambda I)^{-1} \Delta_{1:t}} && (86)\\
=& \lambda^{-1/2} \tilde{\sigma}_{t+1}(\tilde{x}) \sqrt{\Delta_{1:t}^\top \tilde{K}_t (\tilde{K}_t + \lambda I)^{-1} \Delta_{1:t}} && (87)\\
\leq& \lambda^{-1/2} \tilde{\sigma}_{t+1}(\tilde{x}) \sqrt{\Delta_{1:t}^\top \Delta_{1:t}} && (88)\\
\leq& \lambda^{-1/2} \tilde{\beta}_{t+1}^{1/2} \tilde{\sigma}_{t+1}(\tilde{x}) && (89)
\end{aligned}$$

where the equality (83) is from Eq. (68), the inequality (84) is by Cauchy-Schwartz and the equality (86) uses both Eq. (68) and Eq. (71). We can finally derive,

$$\begin{aligned}
&\left|\tilde{h}(\tilde{x}) - h(\tilde{x})\right| && (90)\\
=& \left|\tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} (\tilde{h}_{1:t} - h_{1:t}) - \left(h(\tilde{x}) - \tilde{k}_t(\tilde{x})^T (\tilde{K}_t + \lambda I)^{-1} h_{1:t}\right) + \left(\tilde{h}(\tilde{x}) - \tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} \tilde{h}_{1:t}\right)\right| && (91)\\
\leq& \left|\tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} (\tilde{h}_{1:t} - h_{1:t})\right| + \left|h(\tilde{x}) - \tilde{k}_t(\tilde{x})^T (\tilde{K}_t + \lambda I)^{-1} h_{1:t}\right| + \left|\tilde{h}(\tilde{x}) - \tilde{k}_t(\tilde{x})^\top (\tilde{K}_t + \lambda I)^{-1} \tilde{h}_{1:t}\right| && (92)\\
\leq& \left(2\tilde{B} + \lambda^{-1/2} \tilde{\beta}_{t+1}^{1/2}\right) \tilde{\sigma}_{t+1}(\tilde{x}), && (93)
\end{aligned}$$

where the equality (91) follows by splitting, the inequality (92) follows by triangle inequality, the last inequality follows by combining the inequality (80) and the inequality (89). The conclusion then follows. □

Remark E.2. The proof idea is inspired by the proof of Thm. 2 in (Chowdhury & Gopalan, 2017b).

E.2. Main Proof of Thm. 3.6

We set the generic RKHS $\tilde{\mathcal{H}}$ to be the augmented RKHS with the additive kernel function $k^{ff'}$, the function space ball to be $\mathcal{B}_{ff'}$, $\tilde{B} = 2B$ and the confidence set as,

$$\mathcal{S}_t := \left\{ \tilde{f}(x) - \tilde{f}(x') | \tilde{f} \in \mathcal{B}_f, \sum_{\tau=1}^{t-1} \left((\tilde{f}(x_\tau) - \tilde{f}(x'_\tau)) - (f(x_\tau) - f(x'_\tau)) \right)^2 \leq \beta(\epsilon, \delta/2, t-1) \right\} \subset \mathcal{B}_{ff'}.$$

The desired result then follows by applying Thm. E.1.

F. Proof of Lem. 4.1

It suffices to prove that for any feasible solution of Prob. (23), we can find a corresponding feasible solution of Prob. (24) with the same objective value and that the inverse direction also holds.

1. In this part, we first show that for any feasible solution of Prob. (23), we can find a corresponding feasible solution of Prob. (24) with the same objective value. Let $\tilde{f}$ be a feasible solution of Prob. (23). We construct $\tilde{Z}_{0:t} = (\tilde{f}(x_\tau))_{\tau=0}^t$ and $\tilde{z} = \tilde{f}(x)$. Consider the minimum-norm interpolation problem,

$$\begin{aligned} \min_{s \in \mathcal{B}_f} \quad & \|s\|^2 \\ \text{subject to} \quad & s(x_\tau) = \tilde{z}_\tau, \forall \tau \in \{0\} \cup [t], \\ & s(x) = \tilde{z}. \end{aligned} \tag{94}$$

   By representer theorem, the Prob. (94) admits an optimal solution with the form $\alpha^\top k_{0:t,x}(\cdot)$, where $k_{0:t,x} := (k(w, \cdot))_{w \in \{x_0, \cdots, x_t, x\}}$. So Prob. (94) can be reduced to

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^{t+2}} \quad & \alpha^\top K_{0:t,x} \alpha \\ \text{subject to} \quad & K_{0:t,x} \alpha = \begin{bmatrix} \tilde{Z}_{0:t} \\ \tilde{z} \end{bmatrix}. \end{aligned} \tag{95}$$

   Hence, by solving Prob. (95), we can derive the minimum norm square with interpolation constraints as

$$\begin{bmatrix} \tilde{Z}_{0:t} \\ \tilde{z} \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} \tilde{Z}_{0:t} \\ \tilde{z} \end{bmatrix}.$$

   Since $\tilde{f}$ itself is an interpolant by construction of $(\tilde{Z}_{0:t}, \tilde{z})$. We have

$$\begin{bmatrix} \tilde{Z}_{0:t} \\ \tilde{z} \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} \tilde{Z}_{0:t} \\ \tilde{z} \end{bmatrix} \leq \|\tilde{f}\|^2 \leq B^2.$$

   And since the log-likelihood only depends on $\tilde{Z}_{0:t}$, it holds that

$$\ell(\tilde{Z}_{0:t} | \mathcal{D}_t) = \ell_t(\tilde{f}) \geq \ell_t(\hat{f}_t^{\text{MLE}}) - \beta_1(\epsilon, \delta, t).$$

   And the objectives satisfy,

$$\tilde{z} - \tilde{z}_t = \tilde{f}(x) - \tilde{f}(x_t).$$

   Therefore, $(\tilde{Z}_{0:t}, \tilde{z})$ is a feasible solution for Prob. (24) with the same objective as $\tilde{f}$ for Prob. (23).

2. We then show that for any feasible solution of Prob. (24), we can find a corresponding feasible solution of Prob. (23) with the same objective value. Let $(Z_{0:t}, z)$ be a feasible solution of Prob. (24). We construct

$$\tilde{f}_z = \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top K_{0:t,x}^{-1} k_{0:t,x}(\cdot).$$

Hence,

$$\|\tilde{f}_z\|^2 = \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix} \leq B^2.$$

And it can be checked that $\tilde{f}_z(x_\tau) = z_\tau, \forall \tau \in \{0\} \cup [t]$ and $\tilde{f}_z(x) = z$. So $\ell_t(\tilde{f}_z) = \ell(Z_{0:t}|\mathcal{D}_t) \geq \ell_t(\hat{f}_t^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t)$. And the objectives satisfy $\tilde{f}_z(x) - \tilde{f}_z(x_t) = z - z_t$. So it is proved that for any feasible solution of Prob. (24), we can find a corresponding feasible solution of Prob. (23) with the same objective value.

The desired result then follows.

G. Elaboration on Remark 2.3

By assumption 2.2, we assume that there exists a large enough constant B that upper bounds the norm of the ground-truth black-box function f. However, the exact value of this upper bound may be unknown to us in practice, while the execution of our algorithm relies on the knowledge of B (in Problem (23), B is a key parameter). So we need to guess the value of B. Suppose our guess is $\hat{B}$. It is possible that $\hat{B}$ is even smaller than the ground-truth function norm $\|f\|$. To detect this wrong guess, we observe that, with the correct setting of B such that $B \geq \|f\|$, we have that by Thm. 3.1 and the definition of maximum likelihood estimate, with high probability,

$$\ell_t(\hat{f}_{t|B}^{\mathrm{MLE}}) \geq \ell_t(f) \geq \ell_t(\hat{f}_{t|B}^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t|B),$$

where $\hat{f}_{t|B}^{\mathrm{MLE}}$ is the maximum likelihood estimate function with function norm bound B and $\beta_1(\epsilon, \delta, t|B)$ is the corresponding parameter as defined in Thm. 3.1 with norm bound B. We also have $2B$ is a valid upper bound on $\|f\|$ and thus,

$$\ell_t(\hat{f}_{t|2B}^{\mathrm{MLE}}) \geq \ell_t(f) \geq \ell_t(\hat{f}_{t|2B}^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t|2B).$$

Hence,

$$\ell_t(\hat{f}_{t|B}^{\mathrm{MLE}}) \geq \ell_t(f) \geq \ell_t(\hat{f}_{t|2B}^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t|2B).$$

That is to say, $\ell_t(\hat{f}_{t|B}^{\mathrm{MLE}})$ needs to be greater than or equal to $\ell_t(\hat{f}_{t|2B}^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t|2B)$ when B is a valid upper bound on $\|f\|$.

Therefore, we can use the heuristic: every time we find that

$$\ell_t(\hat{f}_{t|\hat{B}}^{\mathrm{MLE}}) < \ell_t(\hat{f}_{t|2\hat{B}}^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t|2\hat{B}),$$

we double the upper bound guess $\hat{B}$.

H. Jointly Optimize x, $Z_{0,t}$ and z for the Problem (24).

For medium-dimensional problems ($d > 4$), we can jointly optimize x, $Z_{0:t}$, and z by a nonlinear programming solver from multiple random initial conditions. That is, we can also treat x in the problem (23) as an optimization variable. In this way, we lose convexity but only need to solve the problem (23) for only *once* in each step t.

More specifically, we solve the optimization problem (96).

$$\begin{aligned} \max_{x \in \mathbb{R}^d, Z_{0:t} \in \mathbb{R}^{t+1}, z \in \mathbb{R}} \quad & z - z_t \\ \text{subject to} \quad & \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix} \leq B^2, \\ & \ell(Z_{0:t}|\mathcal{D}_t) \geq \ell_t(\hat{f}_t^{\mathrm{MLE}}) - \beta_1(\epsilon, \delta, t), \end{aligned} \tag{96}$$

The only constraint that involves x is

$$\begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top K_{0:t,x}^{-1} \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix} \leq B^2. \tag{97}$$

Applying matrix inversion, we derive that the left-hand side is equal to,

$$Z_{0:t}^\top K_{0:t}^{-1} Z_{0:t} + \frac{1}{k(x,x) - k_t(x)^\top K_{0:t}^{-1} k_t(x)} \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}^\top \begin{bmatrix} K_{0:t}^{-1} k_t(x) \\ -1 \end{bmatrix} \begin{bmatrix} K_{0:t}^{-1} k_t(x) \\ -1 \end{bmatrix}^\top \begin{bmatrix} Z_{0:t} \\ z \end{bmatrix}, \tag{98}$$

where $k_t(x) := (k(x_\tau, x))_{\tau=0}^t$.

We can then apply a nonlinear programming solver such as Ipopt to solve the problem (96) from randomly sampled initial points. Then the best converged solution is set to be the next sample point x_t.

I. Extension to the Multiple-Choice Setting

In this paper, we mainly consider the setting where human expresses preference over only two choices, because of its low cognitive burden to the human user and simplicity of theoretical analysis. However, we can extend POP-BO to the multiple-choice setting where human can compare multiple choices and express the favorite one.

Suppose that in each step τ, we aim to generate a batch of q points. Then we can mix the new batch with the old batch generated in step $\tau - 1$, and query the comparison oracle to report the favorite point among the $2q$ points.

Firstly, the confidence set of functions can be similarly constructed using the likelihood ratio idea and the multiple-choice probabilistic preference model as in (Astudilo et al. 2023),

$$\mathbb{P}\left(x_r \text{ is the favorite}\right) = \frac{e^{f(x_r)}}{\sum_{x \in \{\text{last batch and the new batch}\}} e^{f(x)}}. \tag{99}$$

Secondly, to generate the new batch, the basic idea is that we can apply a 'bootstrap'-type technique. More specifically, we can sequentially generate the new batch $x^1, x^2, \cdots, x^q$. When generating the new point x^{r+1}, we maximize its corresponding optimistic advantage of z^{r+1} as compared to the maximum of $z_{t-q+1:t}, z^1, \cdots, z^r$ by solving a similar problem to Problem (23). That is, we solve the Problem (100) to generate the new point x^{r+1} in the same batch,

$$\begin{aligned} \max_{x \in \mathbb{R}^d, z \in \mathbb{R}, z^{1:r} \in \mathbb{R}^r, Z_{0:t} \in \mathbb{R}^{t+1}} \quad & z - \max\{z_{t-q+1}, \cdots, z_t, z^1, \cdots, z^r\} \\ \text{subject to} \quad & \begin{bmatrix} Z_{0:t} \\ z^{1:r} \\ z \end{bmatrix}^\top K_{0:t,x^{1:r},x}^{-1} \begin{bmatrix} Z_{0:t} \\ z^{1:r} \\ z \end{bmatrix} \leq B^2, \\ & \ell(Z_{0:t} | \mathcal{D}_t) \geq \ell_t(\hat{f}_t^{\text{MLE}}) - \beta_t, \end{aligned} \tag{100}$$

which is equivalent to

$$\begin{aligned} \max_{x \in \mathbb{R}^d, z \in \mathbb{R}, v \in \mathbb{R}, z^{1:r} \in \mathbb{R}^r, Z_{0:t} \in \mathbb{R}^{t+1}} \quad & z - v \\ \text{subject to} \quad & \begin{bmatrix} Z_{0:t} \\ z^{1:r} \\ z \end{bmatrix}^\top K_{0:t,x^{1:r},x}^{-1} \begin{bmatrix} Z_{0:t} \\ z^{1:r} \\ z \end{bmatrix} \leq B^2, \\ & \ell(Z_{0:t} | \mathcal{D}_t) \geq \ell_t(\hat{f}_t^{\text{MLE}}) - \beta_t, \\ & v \geq z_{t-i+1}, i \in [q], \\ & v \geq z^j, j \in [r], \end{aligned} \tag{101}$$

by introducing an auxiliary variable $v \in \mathbb{R}$. Problem (101) can be efficiently solved by the nonlinear programming solver Ipopt.

J. Proof of Thm. 5.2

To prepare for the following analysis, we first give a useful lemma.

Lemma J.1 (Lemma 4, (Chowdhury & Gopalan, 2017b))**.**

$$\sum_{t=1}^{T} \sigma_t^{ff'}((x_t, x_t')) \leq \sqrt{4(T+2)\gamma_T^{ff'}}, \tag{102}$$

where $\sigma_t^{ff'}$ is as defined in Eq. (16) and $\gamma_T^{ff'}$ is as defined in Eq. (18).

Proof. Apply the Lemma 4 in (Chowdhury & Gopalan, 2017b) by setting the kernel function as $k^{ff'}$. □

For convenience, we use β_t to denote $\beta(\epsilon, \delta/2, t)$. We can then analyze the regret of the optimistic algorithm.

$$\begin{aligned}
R_T =& \sum_{t=1}^{T} [f(x^\star) - f(x_t)] \\
=& \sum_{t=1}^{T} [(f(x^\star) - f(x_t')) - (f(x_t) - f(x_t'))] \\
\leq& \sum_{t=1}^{T} [(\tilde{f}_t(x_t) - \tilde{f}_t(x_t')) - (f(x_t) - f(x_t'))] \\
\leq& \sum_{t=1}^{T} 2(2B + \lambda^{-1/2}\beta_t^{1/2})\sigma_t^{ff'}((x_t, x_t')),
\end{aligned}$$

where the first inequality follows by the optimality of $(x_t, \tilde{f}_t)$ for the optimization problem in line 4 of the Alg. 1, and the second inequality follows by Thm. 3.6 (Note that $\beta(\epsilon, \delta/2, t-1) \leq \beta_t = \beta(\epsilon, \delta/2, t)$). Hence,

$$\begin{aligned}
R_T \leq& \sum_{t=1}^{T} 2(2B + \lambda^{-1/2}\beta_t^{1/2})\sigma_t^{ff'}((x_t, x_t')) \\
\leq& 2(2B + \lambda^{-1/2}\beta_T^{1/2}) \sum_{t=1}^{T} \sigma_t^{ff'}((x_t, x_t')) \\
\leq& 2(2B + \lambda^{-1/2}\beta_T^{1/2}) \sqrt{4(T+2)\gamma_T^{ff'}} \\
=& \mathcal{O}\left(\sqrt{\beta_T T \gamma_T^{ff'}}\right).
\end{aligned}$$

K. Proof of Thm. 5.4

We have

$$\begin{aligned}
f(x^\star) - f(x_{t^\star}) =& (f(x^\star) - f(x_{t^\star}')) - (f(x_{t^\star}) - f(x_{t^\star}')) \\
\leq& (\tilde{f}_{t^\star}(x_{t^\star}) - \tilde{f}_{t^\star}(x_{t^\star}')) - (f(x_{t^\star}) - f(x_{t^\star}')) \\
\leq& 2(2B + \lambda^{-1/2}\beta_{t^\star}^{1/2})\sigma_{t^\star}^{ff'}((x_{t^\star}, x_{t^\star}')),
\end{aligned}$$

where $\sigma_{t^\star}^{ff'}$ is as given in Eq. (16) with the kernel function as $k^{ff'}((x_1, x_1'), (x_2, x_2')) = k(x_1, x_2) + k(x_1', x_2')$ and $\beta_{t^\star} = \beta(\epsilon, \delta/2, t^\star)$. Furthermore, by the definition of $t^\star$,

$$\begin{aligned}
2(2B + \lambda^{-1/2}\beta_{t^\star}^{1/2})\sigma_{t^\star}^{ff'}((x_{t^\star}, x_{t^\star}')) \leq& \frac{1}{T} \sum_{t=1}^{T} 2(2B + \lambda^{-1/2}\beta_t^{1/2})\sigma_t^{ff'}((x_t, x_t')) \\
\leq& \frac{2}{T}(2B + \lambda^{-1/2}\beta_T^{1/2}) \sum_{t=1}^{T} \sigma_t^{ff'}((x_t, x_t'))
\end{aligned}$$

$$\leq \frac{2}{T}(2B + \lambda^{-1/2}\beta_T^{1/2})\sqrt{4(T+2)\gamma_T^{ff'}}$$
$$= \mathcal{O}\left(\frac{\sqrt{\beta_T \gamma_T^{ff'}}}{\sqrt{T}}\right).$$

The conclusion then follows.

L. Commonly Used Specific Kernel Functions

- Linear:
$$k(x, \bar{x}) = x^\top \bar{x}.$$
- Squared Exponential (SE):
$$k(x, \bar{x}) = \sigma_{\text{SE}}^2 \exp\left\{-\frac{\|x - \bar{x}\|^2}{l^2}\right\},$$
where σ_{SE}^2 is the variance parameter and l is the lengthscale parameter.
- Matérn:
$$k(x, \bar{x}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|x - \bar{x}\|}{\rho}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{\|x - \bar{x}\|}{\rho}\right),$$
where ρ and ν are the two positive parameters of the kernel function, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind. ν captures the smoothness of the kernel function.

M. Proof of Thm. 5.5

Recall that

$$\beta(\epsilon, \delta/2, t) = \frac{\sigma'^2}{H_\sigma}\left(\beta_2(\epsilon, \delta, t) + 2\beta_1(\epsilon, \delta, t)\right) = \mathcal{O}\left(\sqrt{t \log \frac{t\mathcal{N}(\mathcal{B}_f, \epsilon, \|\cdot\|_\infty)}{\delta}} + \epsilon t + \epsilon^2 t\right).$$

We pick $\epsilon = 1/T$, and can thus derive,

$$\beta_T = \beta(T^{-1}, \delta/2, T) = \mathcal{O}\left(\sqrt{T \log \frac{T\mathcal{N}(\mathcal{B}_f, T^{-1}, \|\cdot\|_\infty)}{\delta}}\right).$$

1. k is a linear kernel, then the corresponding RKHS is a finite-dimensional space and $\log \mathcal{N}(\mathcal{B}_f, T^{-1}, \|\cdot\|_\infty) = \mathcal{O}\left(\log \frac{1}{\epsilon}\right) = \mathcal{O}\left(\log T\right)$ (see, e.g., (Wu, 2017)). The corresponding $k^{ff'}((x, x'), (y, y')) = x^\top y + x'^\top y' = \langle (x, x'), (y, y') \rangle$, which is also linear. Thus, by Thm. 5 in (Srinivas et al., 2012),
$$\gamma_T^{ff'} = \mathcal{O}(\log T).$$
Hence,
$$R_T = \mathcal{O}\left((T \log T)^{1/4+1/2}\right) = \mathcal{O}\left(T^{3/4}(\log T)^{3/4}\right).$$
2. k is a squared exponential kernel, then $\log \mathcal{N}(\mathcal{B}_f, T^{-1}, \|\cdot\|_\infty) = \mathcal{O}\left((\log \frac{1}{\epsilon})^{d+1}\right) = \mathcal{O}\left((\log T)^{d+1}\right)$ (Example 4, (Zhou, 2002)). By Thm. 4 in (Kandasamy et al., 2015), we have,
$$\gamma_T^{ff'} = \mathcal{O}((\log T)^{d+1}).$$
Hence,
$$R_T = \mathcal{O}\left(T^{3/4}(\log T)^{3/4(d+1)}\right).$$

3. k is a Matérn kernel. Lem. 3 in (Bull, 2011) implies the equivalence between RKHS and Sobolev Hilbert space. We can then apply the rich results on the bound of covering number of Sobolev Hilbert space (Edmunds & Triebel, 1996). So $\log \mathcal{N}(\mathcal{B}_f, T^{-1}, \|\cdot\|_\infty) = \mathcal{O}\left((\frac{1}{\epsilon})^{d/\nu} \log \frac{1}{\epsilon}\right) = \mathcal{O}\left(T^{d/\nu} \log T\right)$ (by combing the lower bound in Thm. 5.1 (Xu et al., 2022a) and the convergence rate in Thm. 1 (Bull, 2011)). By Thm. 4 in (Kandasamy et al., 2015), we have,

$$\gamma_T^{ff'} = \mathcal{O}\left(T^{\frac{d(d+1)}{2\nu+d(d+1)}} \log T\right).$$

Hence,

$$R_T = \mathcal{O}\left(T^{3/4}(\log T)^{3/4} T^{\frac{d}{\nu}\left(\frac{1}{4}+\frac{d+1}{4+2(d+1)^d/\nu}\right)}\right) \leq \mathcal{O}\left(T^{3/4}(\log T)^{3/4} T^{\frac{1}{4}\frac{d(d+2)}{\nu}}\right).$$

N. Empirical Evidence for the Order of The Cumulative Regret

Fig. 4 shows the cumulative regret of POP-BO algorithm. The experimental conditions are the same as in Sec. 6.1. Note that both horizontal and vertical axes in Fig. 4 are in log scale, and thus the slope of the curve roughly represents the power of the cumulative regret. It can be clearly seen that the order of the cumulative regret is between $\sqrt{T}$ and T (indeed, close to $T^{\frac{3}{4}}$ by checking the slope in log scale), which verifies our theoretical results in Thm. 5.5.

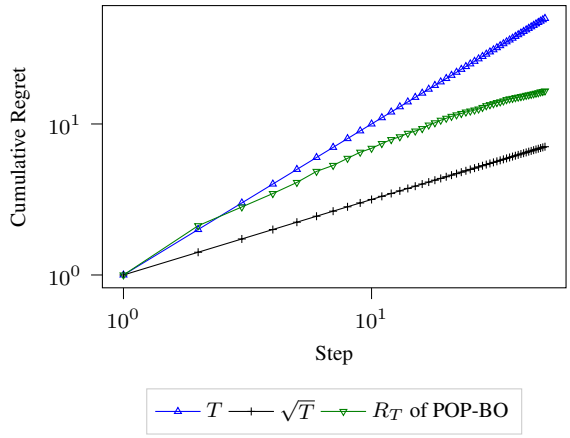

Figure 4. Cumulative regret of our algorithm in log scale. For reference purpose, we also plot $\sqrt{T}$ and T in log scale.

O. Kernel-Specific Convergence Rate

Similar to the bounds in the Appendix M, we can plug in the kernel-specific covering number and maximum information gain to derive the kernel-specific convergence rate in Tab. 3.

Table 3. Kernel-specific convergence rate for $x_{t^\star}$.

Kernel	Linear	Squared Exponential	Matérn $\left(\nu > \frac{d}{4}(3+d+\sqrt{d^2+14d+17}) = \Theta(d^2)\right)$
$f(x^\star) - f(x_{t^\star})$	$\mathcal{O}\left(\frac{(\log T)^{3/4}}{T^{1/4}}\right)$	$\mathcal{O}\left(\frac{(\log T)^{3/4(d+1)}}{T^{1/4}}\right)$	$\mathcal{O}\left(\frac{(\log T)^{3/4} T^{\frac{d}{\nu}\left(\frac{1}{4}+\frac{d+1}{4+2(d+1)^d/\nu}\right)}}{T^{1/4}}\right)$

P. More Experimental Results and Details

Selection of Hyperparameters. Three key hyperparameters that influence the performance of POP-BO are the kernel lengthscale, the norm bound and the confidence level term β as shown in Thm. 3.1. We set $\beta = \beta_0\sqrt{t}$, where β_0 is set to 1.0 by default. For the sampled instances from Gaussian processes, the lengthscale is set to be the ground truth and the norm bound is set to be 1.1 times the ground truth. For the test function examples, we choose the lengthscale by maximizing the

likelihood value over a set of randomly sampled data and set the norm bound to be 6 by default (with the test functions all normalized).

Details on Sampled Instances from Gaussian Process. Specifically, we randomly sample some knot points from a joint Gaussian distribution marginalized from the Gaussian process, and then construct its corresponding minimum-norm interpolant (Maddalena et al., 2021) as the ground truth function.

Empirical Method for Reporting a Solution. In the experiment of test function optimization, we report the point that maximizes the minimum norm maximum likelihood estimator $\hat{f}_t^{\text{MLE}}$, which achieves better empirical performance.

Solution Report Method for Baselines. The approach to reporting a solution is the same as in the original paper of the baseline algorithm if it is mentioned. Therefore, for the baseline qEUBO (Astudillo et al., 2023), we report the solution that maximizes the expected objective value conditioned on the historical samples. For the baseline SGP (Takeno et al., 2023), we report the first point of the duel proposed by the algorithm in step t. For the baseline DTS (González et al., 2017), we report the Condorcet winner.

Effect of Hyperparameters. We conducted more experiments to assess the effect of hyperparameters. We observe that the hyperparameters with most influence are the norm bound B and the confidence level β_t. The larger the norm bound B is, the more variance the estimate function has. If B is set too large, the convergence for the suboptimality of the reported solution tends to be slower. β_t can be set to be $\beta_0\sqrt{t}$ in practice and determines the level of exploration, where β_0 is a fixed constant. The larger β_0 is, the more explorative the algorithm is and may have higher cumulative regret. But setting β_0 to be very small may also cause weak exploration and make the suboptimality of the reported solution converge slower.

P.1. Experimental Results for Higher-Dimensional Problems

Higher-Dimensional Problems Sampled from Gaussian Process. We consider the optimization of 7-dimensional black-box function sampled from a Gaussian process with kernel function as shown in Eq. (103),

$$k(x, \bar{x}) = \sigma_{\text{SE}}^2 \exp\left\{-\frac{\|x - \bar{x}\|^2}{l^2}\right\} \tag{103}$$

where $\sigma_{\text{SE}}^2 = 9.0$ and $l = 5\sqrt{7}$. The optimization domain is set to be $[0, 10]^7$. We run 20 randomly sampled instances for 100 steps. The average update time for each step t is only 11.0 seconds on a personal computer with one `Intel64 Family 6 Model 142 Stepping 12 GenuineIntel 1803 Mhz` processor and `16.0 GB RAM`. This is comparably very small considering that each query to the comparison oracle can be very expensive in practice (e.g., heating the room up to a certain temperature to evaluate occupant comfort, which may take tens of minutes). We compare our method to the SGP baseline.

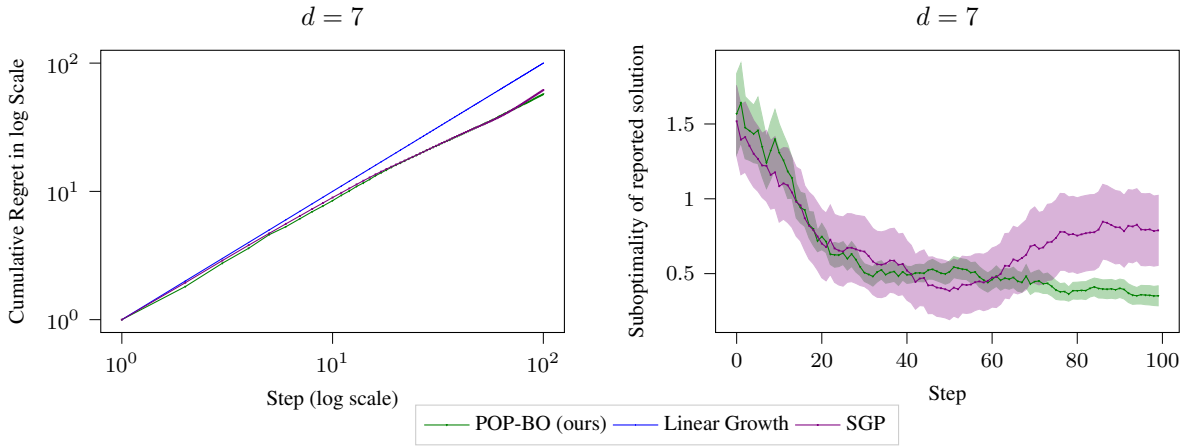

Figure 5. Cumulative regret in log scale and the suboptimality of the reported solution in linear scale for a 7-dimensional problem sampled from Gaussian process. For reference purpose, we also plot T in the cumulative regret plot in log scale, where the shaded areas represent ± 0.2 standard deviation.

Fig. 5 shows the cumulative regret (in log scale) and the suboptimality of the reported solution for our POP-BO algorithm, where the reported solution is derived by maximizing the maximum likelihood estimate function. It can be clearly seen that our algorithm achieves both sublinear regret growth and fast convergence for the suboptimality of the reported solution in this 7-dimensional problem. Interestingly, the suboptimality of SGP converges similarly to our method before 50 steps, but get even worse after 50 steps. This is because SGP ignores the randomness in the preference feedback, which leads to misbelief in the function difference value, and such misbelief is more significant when the function difference value is small.

Higher-Dimensional Test Problem. In this section, we further consider the optimization of the 6-dimensional Ackley function as shown in (Astudillo et al., 2023). For this problem, we compare POP-BO algorithm to the qEUBO algorithm proposed in (Astudillo et al., 2023). Fig. 6 shows the cumulative regret and the suboptimality of the reported solution. In this particular problem, qEUBO performs better than our POP-BO algorithm in terms of cumulative regret, while our POP-BO algorithm performs slightly better than qEUBO in terms of the suboptimality of the reported solution.

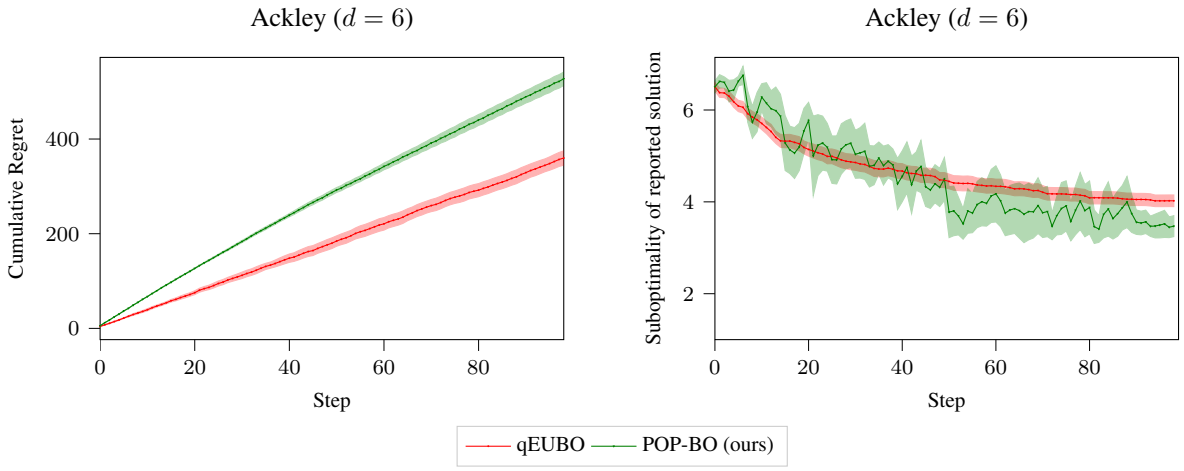

Figure 6. Cumulative regret and the suboptimality of the reported solution for the 6-dimensional Ackley function optimization problem, where the shaded areas represent ± 0.5 standard deviation.

P.2. Occupant Thermal Comfort Optimization

Two-Dimensional Comfort Optimization. An accurate model of human thermal comfort is crucial for improving occupants' comfort while saving energy in buildings. However, establishing such a model has proven to be a complex and challenging task (Zhang et al., 2024) and standard offline models ignore the individual differences among occupants. In this section, we consider the real-world problem of maximizing occupant thermal comfort directly from thermal preference feedback. To emulate real human thermal sensation, we use the well-known and widely adopted Predicted Mean Vote (PMV) model (Fanger et al., 1970) as the ground truth and generate the preference feedback according to the Bernoulli model as assumed in Assumption 2.5. We optimize the indoor air temperature and air speed, which are the two major factors that influence thermal comfort and are controllable by HVAC (Heating, Ventilation, and Air Conditioning) systems and fans. Indeed, tuning these two factors has been proven effective in providing thermal comfort while minimizing energy consumption (Lyu et al., 2023). The result is shown in Fig. 7 where the mean is taken over 30 instances of simulation. It can be seen that our method stably achieves superior performance in optimizing human thermal comfort, which implies its potential to deal with preferential feedback in real-world applications. It is also noticeable that although qEUBO achieves slightly better performance in terms of the convergence of the reported solution, the cumulative regret of qEUBO is almost twice of POP-BO's cumulative regret. This means our method is more favorable in applications where online performance during the optimization is also critical, such as online tuning of HVAC systems.

Scalability to Higher Dimension. Additionally, to demonstrate the scalability of POP-BO in this real-world comfort optimization problem, we additionally tune the mean radiant temperature and relative humidity, which results in a four-dimensional black-box optimization problem. The result is shown in Fig. 8. It can be observed that increasing the dimensionality does not drastically decrease the convergence rate of our method. Furthermore, the baseline method qEUBO can decrease the objective value very fast in the initial steps, but seems to be still very oscillatory after 10 steps. In contrast,

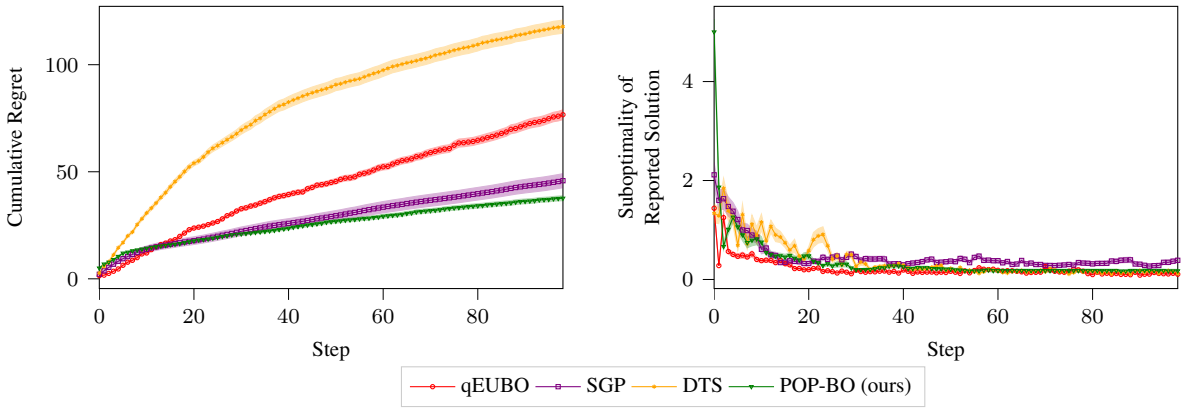

Figure 7. Cumulative regret and the suboptimality of the reported solution of different algorithms for thermal comfort optimization.

our method converges faster than SGP without the oscillation issue like qEUBO.

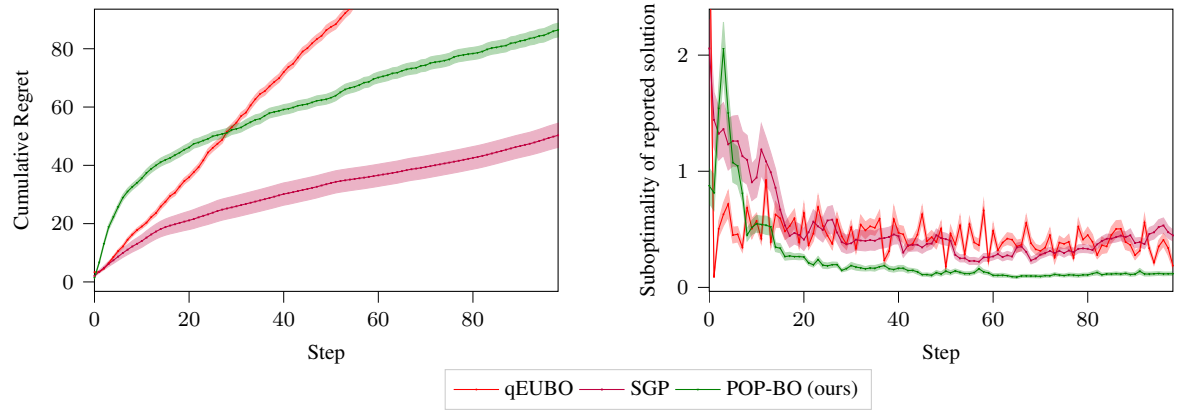

Figure 8. Cumulative regret and the suboptimality of the reported solution of different algorithms for the four-dimensional thermal comfort optimization problem.

P.3. Details About the Results in Tab. 2

The cumulative regret and evolution of suboptimality for the different test problems in Tab. 2 are shown in Fig. 9. Since the considered problems only have 2-dimensional input and in the applications of Bayesian optimization, it is typically desired to obtain a set of solution with objective value as close to the optimal value as possible. So we only consider 30 steps here. Other baselines can make limited progress in terms of the suboptimality of the reported solution within only 30 steps (partially also due to the 'adversarial' property of the test functions, i.e., severe non-convexity and multiple local maxima) as shown in Tab. 2. To the sharp contrast, our POP-BO algorithm makes significant progress in reducing the suboptimality of the reported solution by balancing exploration and exploitation, and estimating the best solution in a principled way.

To provide more insights into POP-BO's performance across different settings, we compare our algorithm's evolution of cumulative regret and suboptimality to other baseline methods for each test problem in Fig. 10 and Fig. 11. It can be observed that our method may perform slightly worse than some baselines in certain problems. For example, our method performs slightly worse than qEUBO in the Bukin problem in terms of suboptimality. However, our method performs stably and is consistently one of the best in all the test problems in terms of the suboptimality.

Q. Additional Contributions as Compared to (Mehta et al., 2023)

Notably, (Mehta et al., 2023) proposes Borda-AE algorithm, which directly learns the winning probability function using kernel ridge regression. This key design allows the authors to derive an information-theoretic convergence rate and efficient

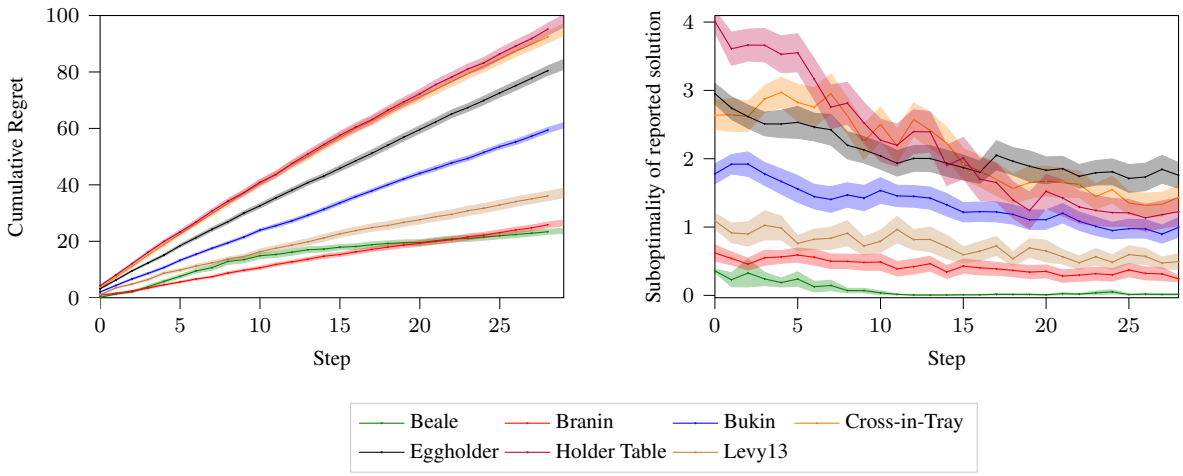

Figure 9. Cumulative regret and the suboptimality of the reported solution of POP-BO algorithm for the different test problems in Tab. 2.

computation method without diving into the learning of the underlying reward function.

However, (Mehta et al., 2023) has key limitations and our paper makes additional contributions in the following two aspects.

1. **Cumulative regret bound.** There are two possible ways to define cumulative regret. One way is that we can define the (partial) cumulative regret as the summation of the suboptimality of *only* x_t (that is, $\sum_{t=1}^{T}(f(x^\star) - f(x_t))$). With this (partial) cumulative regret definition, Borda-AE algorithm can provide a sublinear (partial) cumulative regret bound, although it has linear growth in the cumulative regret of the compared point sequence $\{x'_t\}_{t=1}^{T}$. However, in many practical online learning applications, it is desired to control the suboptimality of both x_t and x'_t sequences. For example, when tuning the thermal/visual comfort of room occupants, we require the occupants to experience both x_t and x'_t conditions for comparison purposes and the suboptimality (links to discomfort) caused by both x_t and x'_t need to be managed.

   Therefore, it is more practically relevant to define (total) cumulative regret as the total cumulative suboptimality of both x_t and x'_t sequences (that is, $\sum_{t=1}^{T}(f(x^\star) - f(x_t)) + \sum_{t=1}^{T}(f(x^\star) - f(x'_t))$). Interestingly, since $x'_t = x_{t-1}$ by the design of our POP-BO algorithm, this (total) cumulative regret bound reduces to $2\sum_{t=1}^{T}(f(x^\star) - f(x_t))$, for which we provide our sublinear cumulative regret bound. As such, the (total) cumulative regret bound provided by our paper is stronger than the (partial) cumulative regret bound that could be obtained by (Mehta et al., 2023).

2. **Applicability to online learning problem.** Following the last point, (Mehta et al., 2023) is not applicable to the online learning problem since in line 6 of the Borda-AE algorithm, a'_t is uniformly sampled from the action space, which leads to a linear growth of cumulative regret. This means Borda-AE has very poor online performance and can not be applied to an online learning problem. For example, in building thermal comfort tuning, we also want to control the discomfort caused during the tuning process. In contrast, our POP-BO algorithm has good online performance with both a theoretical bound on cumulative regret (Thm. 5.2) and empirical evidence on small cumulative regret (Fig. 2).

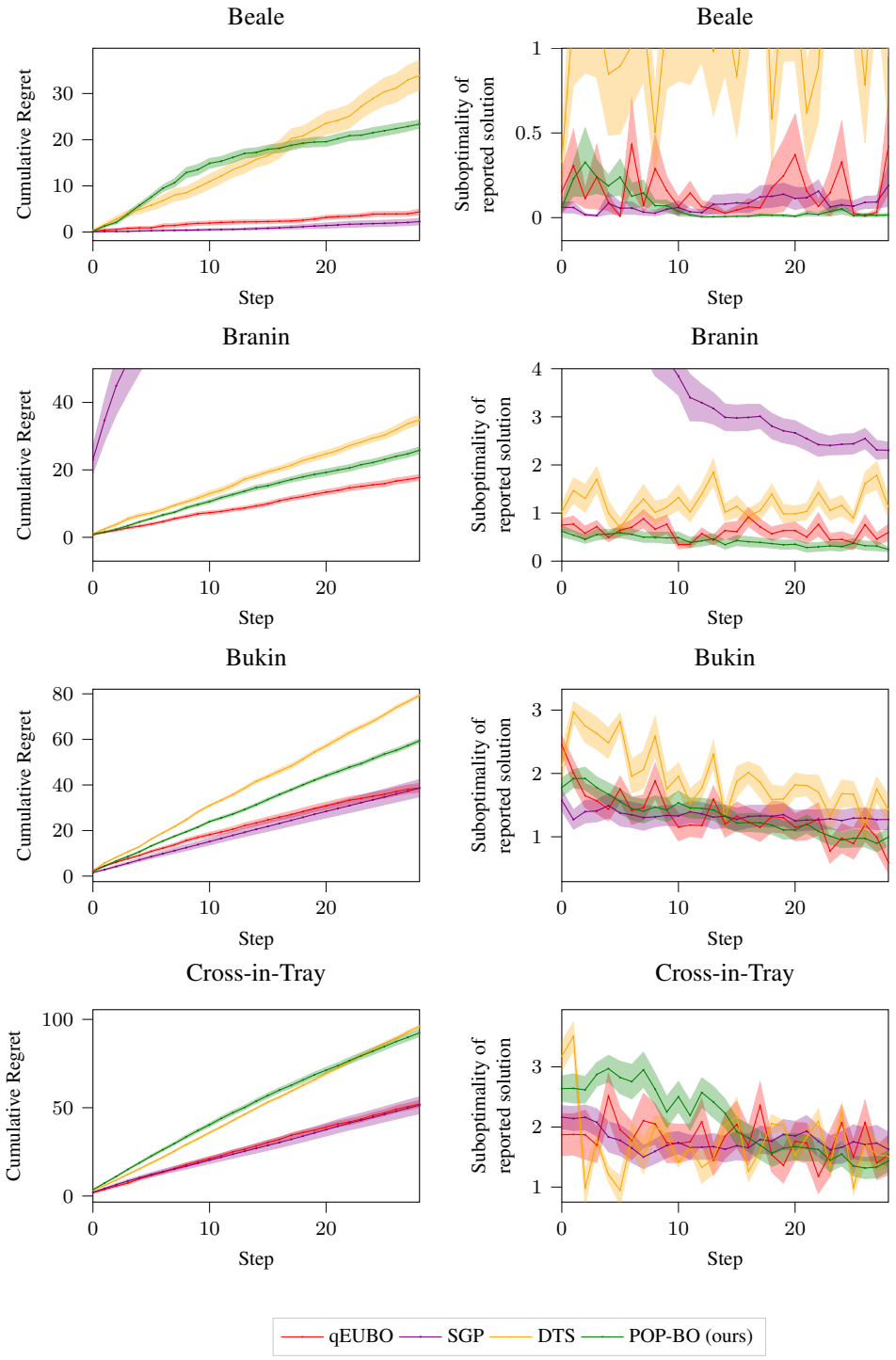

Figure 10. Cumulative regret and the suboptimality of the reported solution of different algorithms for the test problems Beale, Branin, Bukin, and Cross-in-Tray in Tab. 2.

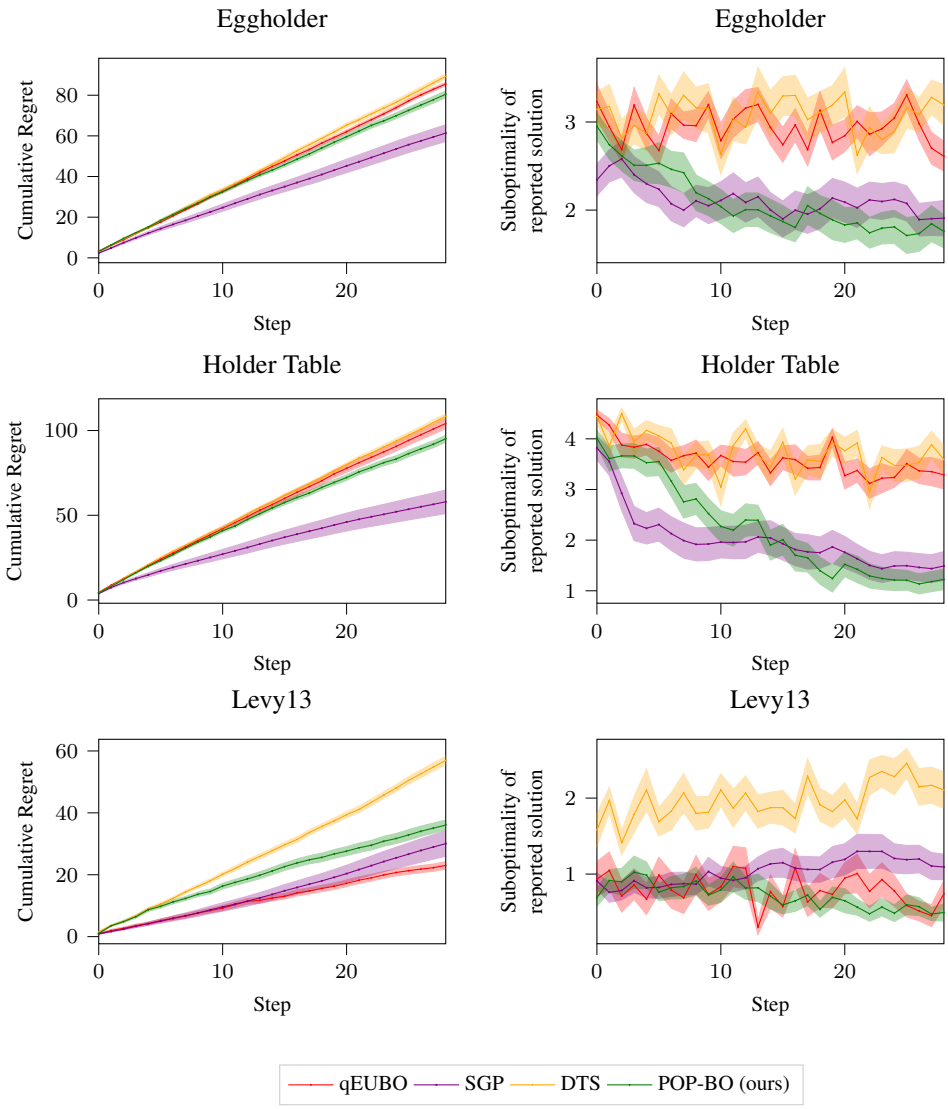

Figure 11. Cumulative regret and the suboptimality of the reported solution of different algorithms for the test problems Eggholder, Holder Table, and Levy13 in Tab. 2.