Distributionally Robust Active Learning for Gaussian Process Regression

Shion Takeno [1] **Okura Yoshito** [1] **Yu Inatsu** [2] **Aoyama Tatsuya** [1] **Tomonari Tanaka** [1] **Akahane Satoshi** [1] **Hiroyuki Hanada** [3] **Noriaki Hashimoto** [3] **Taro Murayama** [4] **Hanju Lee** [4] **Shinya Kojima** [4] **Ichiro Takeuchi** [1 3]

Abstract

Gaussian process regression (GPR) or kernel ridge regression is a widely used and powerful tool for nonlinear prediction. Therefore, active learning (AL) for GPR, which actively collects data labels to achieve an accurate prediction with fewer data labels, is an important problem. However, existing AL methods do not theoretically guarantee prediction accuracy for target distribution. Furthermore, as discussed in the distributionally robust learning literature, specifying the target distribution is often difficult. Thus, this paper proposes two AL methods that effectively reduce the worst-case expected error for GPR, which is the worst-case expectation in target distribution candidates. We show an upper bound of the worst-case expected squared error, which suggests that the error will be arbitrarily small by a finite number of data labels under mild conditions. Finally, we demonstrate the effectiveness of the proposed methods through synthetic and real-world datasets.

1. Introduction

Active learning (AL) (Settles, 2009) is a framework for achieving high prediction performance with fewer data when labeling new data is expensive. For this purpose, AL algorithms actively acquire the label of data that improves the prediction performance of some statistical model based on *acquisition functions* (AFs). Many types of AFs have been proposed, such as uncertainty sampling (US), random sampling (RS), variance reduction, and information gain, as summarized in (Settles, 2009).

[1]Department of Mechanical Engineering, Nagoya University, Aichi, Japan [2]Department of Computer Science, Nagoya Institute of Technology, Aichi, Japan [3]RIKEN AIP, Tokyo, Japan [4]DENSO CORPORATION, Aichi, Japan. Correspondence to: Shion Takeno <takeno.s.mllab.nit@gmail.com>, Ichiro Takeuchi <takeuchi.ichiro.n6@f.mail.nagoya-u.ac.jp>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Gaussian process regression (GPR) model (Rasmussen & Williams, 2005) is often used as a base statistical model for AL algorithms due to its flexible prediction capability (Seo et al., 2000; Yu et al., 2006; Guestrin et al., 2005; Krause et al., 2008b; Hoang et al., 2014; Hübotter et al., 2024). Standard AL methods for the GPR are based on information gain (Guestrin et al., 2005; Krause et al., 2008b; Kirsch et al., 2021; Kirsch & Gal, 2022; Bickford Smith et al., 2023; Hübotter et al., 2024). Most information gain-based approaches are heuristics without theoretical guarantees. A notable exception is the work by (Guestrin et al., 2005; Krause et al., 2008b), which shows that the US for the GPR model is optimal to maximize the information gain from the obtained data labels regarding the GP prior. Furthermore, from the analysis of kernelized bandits (e.g., Srinivas et al., 2010; Salgia et al., 2024), we can see that the US and RS guarantee the convergence of the maximum of posterior variance (See Proposition 2.3 for details). Another commonly used AF is variance reduction (Seo et al., 2000; Yu et al., 2006; Shoham & Avron, 2023; Hübotter et al., 2024), which can be computed efficiently in the GPR. However, these AFs do not incorporate the importance of the unlabelled dataset, that is, the prior information regarding the target distribution. In addition, to our knowledge, except for the worst-case analysis in Proposition 2.3, there are no theoretical guarantees for the target prediction error.

Several studies have tackled the development of the target distribution-aware AL (Kirsch et al., 2021; Kirsch & Gal, 2022; Bickford Smith et al., 2023). In particular, as an extension of the distributionally robust learning (Chen et al., 2020), Frogner et al. (2021) proposed *distributionally robust AL* (DRAL), which aims to minimize the worst-case error in the set of target distributions to obtain a robust model. However, since these studies employed the heuristic AL methods based on, e.g., information gain and expected model change (Settles, 2009), the theoretical guarantee has not been shown.

This paper develops a DRAL framework for the GPR model. We aim to minimize the worst-case expected error, where the worst-case scenario and the expectation are taken regarding the target distribution candidates and chosen target distributions, respectively. Note that our formulation generalizes target distribution-aware AL since it includes the case

in which the unique target distribution can be specified. We perform the theoretical analysis under two conditions called Bayesian and frequentist assumptions (Srinivas et al., 2010), in which we leverage several useful lemmas in kernelized bandit literature (Srinivas et al., 2010; Vakili et al., 2021a;b; Kusakawa et al., 2022).

Our contributions are summarized as follows:

1. We show several properties of the worst-case squared error for the GPR model, which suggests that the error can be bounded from above using the posterior variance even if the input domain is continuous. Along the way to proving the error properties, we show the Lipschitz constant of the posterior mean of GPs in Lemmas 2.9 and 3.3, which may be of independent interest.
2. We propose two DRAL methods for the GPR model, inspired by the RS and the greedy algorithm. Our proposed methods are designed to guarantee the convergence of the (expected) posterior variance.
3. We show the probabilistic upper bounds of the error incurred by the proposed algorithm, which suggests that under mild conditions, the error can be arbitrarily small by a finite number of data labels.

Finally, we demonstrate the effectiveness of the proposed methods via synthetic and real-world regression problems.

2. Background

This section provides the known properties of the GPR.

2.1. GPR model

The GPR model (Rasmussen & Williams, 2005) is a kernel-based regression model. Let us consider that we have already obtained the training dataset of input-output pair $\mathcal{D}_t = \{(\boldsymbol{x}_i, y_{\boldsymbol{x}_i})\}_{i=1}^t$, where $\forall i, \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $y_{\boldsymbol{x}_i} \in \mathbb{R}$, and d is an input dimension. The GPR model assumes that, without loss of generality, f follows zero-mean GP, that is, $f \sim \mathcal{GP}(0, k)$, where $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a predefined positive semidefinite kernel function. In addition, the i-th observation $y_{\boldsymbol{x}_i}$ is assumed to be contaminated by i.i.d. Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ as $y_{\boldsymbol{x}_i} = f(\boldsymbol{x}_i) + \epsilon_i$. Then, the posterior distribution of f becomes again a GP, whose mean and variance are analytically derived as follows:

$$\begin{aligned} \mu_t(\boldsymbol{x}) &= \boldsymbol{k}_t(\boldsymbol{x})^\top \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_t\right)^{-1} \boldsymbol{y}_t, \\ \sigma_t^2(\boldsymbol{x}) &= k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_t(\boldsymbol{x})^\top \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}_t\right)^{-1} \boldsymbol{k}_t(\boldsymbol{x}), \end{aligned} \tag{1}$$

where $\boldsymbol{k}_t(\boldsymbol{x}) := \left(k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_t)\right)^\top \in \mathbb{R}^t$, $\boldsymbol{K} \in \mathbb{R}^{t \times t}$ is the kernel matrix whose (i, j)-element is $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{I}_t \in \mathbb{R}^{t \times t}$ is the identity matrix, and $\boldsymbol{y}_t := (y_{\boldsymbol{x}_1}, \ldots, y_{\boldsymbol{x}_t})^\top \in \mathbb{R}^t$. Finally, for later use, let the posterior variance at $\boldsymbol{x}^*$ when $\boldsymbol{x}_t = \boldsymbol{x}$ be $\sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x}) := \sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x}_t = \boldsymbol{x})$. Note that the posterior variance calculation does not require $\boldsymbol{y}_t$. Furthermore, it is known that $\mu_t(\boldsymbol{x})$ is equivalent to the kernel ridge regression estimator with regularization parameter $\lambda = \sigma^2$ (Kanagawa et al., 2018).

Maximum Information Gain (MIG): Further, we define MIG (Srinivas et al., 2010; Vakili et al., 2021b):

Definition 2.1 (Maximum information gain)**.** Let $f \sim \mathcal{GP}(0, k)$ over $\mathcal{X} \subset [0, r]^d$. Let $A = \{\boldsymbol{a}_i\}_{i=1}^T \subset \mathcal{X}$. Let $\boldsymbol{f}_A = \left(f(\boldsymbol{a}_i)\right)_{i=1}^T$, $\boldsymbol{\epsilon}_A = \left(\epsilon_i\right)_{i=1}^T$, where $\forall i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and $\boldsymbol{y}_A = \boldsymbol{f}_A + \boldsymbol{\epsilon}_A \in \mathbb{R}^T$. Then, MIG γ_T is defined as follows:

$$\gamma_T := \max_{A \subset \mathcal{X}; |A| = T} I(\boldsymbol{y}_A; \boldsymbol{f}_A),$$

where I is the Shannon mutual information.

It is known that MIG is sublinear for commonly used kernel functions, for example, $\gamma_T = \mathcal{O}\left(d \log T\right)$ for linear kernels, $\gamma_T = \mathcal{O}\left((\log T)^{d+1}\right)$ for squared exponential (SE) kernels $k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 / (2\ell^2)\right)$, and $\gamma_T = \mathcal{O}\left(T^{\frac{d}{2\nu + d}} (\log T)^{\frac{2\nu}{2\nu + d}}\right)$ for Matérn-ν kernels $k_{\mathrm{Mat}} = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\ell}\right)^\nu J_\nu \left(\frac{\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\ell}\right)$, where $\ell, \nu > 0$ are the lengthscale and smoothness parameter, respectively, and $\Gamma(\cdot)$ and J_ν are Gamma and modified Bessel functions, respectively (Srinivas et al., 2010; Vakili et al., 2021b).

Lipschitz Consatant of $\sigma_t(\boldsymbol{x})$: We will use the following useful result from Theorem E.4 in (Kusakawa et al., 2022):

Lemma 2.2 (Lipschitz constant for posterior standard deviation)**.** *Let $k(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be linear, SE, or Matérn-ν kernel and $k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$. Moreover, assume that a noise variance σ^2 is positive. Then, for any $t \geq 1$ and $\mathcal{D}_t$, the posterior standard deviation $\sigma_t(\boldsymbol{x})$ satisfies that*

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d,\ |\sigma_t(\boldsymbol{x}) - \sigma_t(\boldsymbol{x}')| \leq L_\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|_1,$$

where L_σ is a positive constant given by

$$L_\sigma = \begin{cases} 1 & \text{if } k(\boldsymbol{x}, \boldsymbol{x}') \text{ is the linear kernel}, \\ \frac{\sqrt{2}}{\ell} & \text{if } k(\boldsymbol{x}, \boldsymbol{x}') \text{ is the SE kernel}, \\ \frac{\sqrt{2}}{\ell}\sqrt{\frac{\nu}{\nu - 1}} & \text{if } k(\boldsymbol{x}, \boldsymbol{x}') \text{ is the Matérn kernel}, \end{cases}$$

where $\nu > 1$.

2.2. Uncertainty Sampling and Random Sampling

For the GPR model, the US selects the most uncertain input as t-th input:

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \sigma_{t-1}^2(\boldsymbol{x}).$$

The RS randomly selects a t-th input by a fixed probability distribution $p(\boldsymbol{x})$ over $\mathcal{X}$:

$$\boldsymbol{x}_t \sim p(\boldsymbol{x}).$$

For both algorithms, the upper bound of the maximum variance is known:

Proposition 2.3. *Assume $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. If we run the US, the following inequality holds:*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \sigma_T^2(\boldsymbol{x}) \leq \frac{C_1\gamma_T}{T},$$

where $C_1 = 1/\log(1+\sigma^{-2})$. Furthermore, if we run the RS, the following inequality holds with probability at least $1-\delta$, where $\delta \in (0,1)$, under several conditions:

$$\max_{\boldsymbol{x}\in\mathcal{X}} \sigma_T^2(\boldsymbol{x}) = \mathcal{O}\left(\frac{\sigma^2\gamma_T}{T}\right).$$

Proof. For the US, see, e.g., Eq. (16) in (Vakili et al., 2021a) and Lemma 5.4 in (Srinivas et al., 2010). For the RS, see Theorem 3.1 in (Salgia et al., 2024). □

When γ_T is sublinear, the above upper bounds suggest that the maximum variance will be arbitrarily small within the finite time horizons.

2.3. Regularity Assumptions and Predictive Guarantees

Here, we provide the details of Bayesian and frequentist assumptions and predictive guarantees for both assumptions.

2.3.1. Bayesian Assumption

We consider the following assumption:

Assumption 2.4. The function f is a sample path $f \sim \mathcal{GP}(0,k)$ and the i-th observation $y_{\boldsymbol{x}_i}$ is contaminated by i.i.d. Gaussian noise $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ as $y_{\boldsymbol{x}_i} = f(\boldsymbol{x}_i)+\epsilon_i$. In addition, the kernel function is normalized as $k(\boldsymbol{x},\boldsymbol{x}') \leq 1$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}$.

Furthermore, for continuous $\mathcal{X}$, we assume the following smoothness condition:

Assumption 2.5. Let $\mathcal{X} \subset [0,r]^d$ be a compact set, where $r > 0$. Assume that the kernel k satisfies the following condition on the derivatives of a sample path f. There exist the constants $a, b > 0$ such that,

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{\partial f(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right| > L\right) \leq a\exp\left(-\frac{L^2}{b^2}\right),$$

for all $j \in [d]$.

This assumption holds for stationary and four times differentiable kernels (Theorem 5 of Ghosal & Roy, 2006), such as SE kernel and Matérn-ν kernels with $\nu > 2$ (Section 4 of Srinivas et al., 2010). These assumptions are commonly used (Srinivas et al., 2010; Kandasamy et al., 2018; Paria et al., 2020; Takeno et al., 2023; 2024).

As with Lemma 5.1 in (Srinivas et al., 2010), the credible interval can be obtained as follows:

Lemma 2.6. *Suppose that $\mathcal{X}$ is finite and Assumption 2.4 holds. Pick $\delta \in (0,1)$ and $t \in \mathbb{N}$. Then, for any given $\mathcal{D}_t$,*

$$\Pr\left(|f(\boldsymbol{x}) - \mu_t(\boldsymbol{x})| \leq \beta_\delta^{1/2}\sigma_t(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X} \mid \mathcal{D}_t\right) \geq 1-\delta,$$

where $\beta_\delta = 2\log(|\mathcal{X}|/\delta)$.

2.3.2. Frequentist Assumption

We assume that f is an element of the reproducing kernel Hilbert space (RKHS) specified by the kernel k as with (Srinivas et al., 2010; Chowdhury & Gopalan, 2017; Vakili et al., 2021a; 2022; Li & Scarlett, 2022):

Assumption 2.7. Let f be an element of RKHS $\mathcal{H}_k$ specified by the kernel k used in the GPR model. Furthermore, the RKHS norm of f is bounded as $\|f\|_{\mathcal{H}_k} \leq B < \infty$ for some $B > 0$, where $\|\cdot\|_{\mathcal{H}_k}$ denotes the RKHS norm of $\mathcal{H}_k$. In addition, the i-th observation $y_{\boldsymbol{x}_i}$ is contaminated by independent sub-Gaussian noises $\{\epsilon_i\}_{i\in\mathbb{N}}$ as $y_{\boldsymbol{x}_i} = f(\boldsymbol{x}_i)+\epsilon_i$. That is, for all $i \in \mathbb{N}$, for all $\eta \in \mathbb{R}$, and for some $R > 0$, the moment generating function of ϵ_i satisfies $\mathbb{E}[\exp(\eta\epsilon_i)] \leq \exp\left(\frac{\eta^2R^2}{2}\right)$. Finally, the kernel function is normalized as $k(\boldsymbol{x},\boldsymbol{x}') \leq 1$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}$.

Furthermore, for continuous $\mathcal{X}$, we assume the following smoothness condition as with (Chowdhury & Gopalan, 2017; Vakili et al., 2021a; 2022):

Assumption 2.8. The kernel function k satisfies the following condition on the derivatives. There exists a constant L_k such that,

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{j\in[d]} \left|\frac{\partial^2 k(\boldsymbol{u},\boldsymbol{v})}{\partial u_j \partial v_j}\Big|_{\boldsymbol{u}=\boldsymbol{v}=\boldsymbol{x}}\right|^{1/2} \leq L_k.$$

This assumption provides the Lipschitz constant of f:

Lemma 2.9 (Lemma 5.1 in (De Freitas et al., 2012))**.** *Suppose that Assumption 2.8 holds. Then, any $g \in \mathcal{H}_k$ is Lipschitz continuous with respect to $\|g\|_{\mathcal{H}_k}L_k$.*

We rely on the confidence bounds for non-adaptive sampling methods, which is a direct consequence of Theorem 1 in (Vakili et al., 2021a) and the union bound:

Lemma 2.10. *Suppose that $\mathcal{X}$ is finite and Assumption 2.7 holds. Pick $\delta \in (0,1)$ and $t \in \mathbb{N}$. Assume that $(\boldsymbol{x}_i)_{i\in[t]}$ is*

independent of $(\epsilon_i)_{i\in[t]}$. *Then, the following holds:*

$$\Pr\left(|f(\boldsymbol{x}) - \mu_t(\boldsymbol{x})| \leq \beta_\delta^{1/2}\sigma_t(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}\right) \geq 1-\delta,$$

where $\beta_\delta = \left(B + \frac{R}{\sigma}\sqrt{2\log(\frac{2|\mathcal{X}|}{\delta})}\right)^2$.

3. Problem Statement and Its Property

This section provides details on our problem setup and its properties.

3.1. Problem Statement

We aim to minimize the worst-case expected errors regarding the GP prediction $\mu_T(\boldsymbol{x})$ after T-th function evaluations:

$$E_T \coloneqq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[(f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*))^2\right], \tag{2}$$

where $\mathcal{P}$ is a set of target distributions over the input space $\mathcal{X}$ called ambiguity set (Chen et al., 2020). We assume that $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[g(\boldsymbol{x}^*)]$ exists for any continuous function $g : \mathcal{X} \to \mathbb{R}$. This paper concentrates on the setting where the training input space from which we can obtain labels includes the test input space.

Our problem setup can be seen as the generalization of the target distribution-aware AL and the AL for the worst-case error $\max_{\boldsymbol{x}\in\mathcal{X}}(f(\boldsymbol{x}) - \mu_T(\boldsymbol{x}))^2$. This is because our problem is equivalent to the target distribution-aware AL if we set $|\mathcal{P}| = 1$ and to the worst-case error minimization if $\mathcal{P}$ includes $\{p \in \mathcal{P}_\mathcal{X} \mid \exists \boldsymbol{x} \in \mathcal{X}, p(\boldsymbol{x}) = 1\}$, where $\mathcal{P}_\mathcal{X}$ is the set of the distributions over $\mathcal{X}$.

3.2. High Probability Bound of Error

If the input space $\mathcal{X}$ is finite, we can obtain the upper bound of Eq. (2) as the direct consequence of Lemmas 2.6 and 2.10:

Lemma 3.1. *Fix* $\delta \in (0,1)$ *and* $T \in \mathbb{N}$. *Suppose that Assumption 2.4 holds and* β_δ *is set as in Lemma 2.6, or Assumption 2.7 holds and* β_δ *is set as in Lemma 2.10. Then, the following holds with probability at least* $1-\delta$*:*

$$E_T \leq \beta_\delta \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right].$$

For continuous $\mathcal{X}$, the confidence parameter $\beta_\delta \propto \log|\mathcal{X}|$ diverges if we apply Lemmas 2.6 and 2.10 directly. Therefore, in this case, the Lipschitz property is often leveraged (Chowdhury & Gopalan, 2017; Vakili et al., 2021a). The Lipschitz constant of f can be directly derived from the Assumption 2.5, or Assumption 2.8 and Lemma 2.9 (Srinivas et al., 2010; De Freitas et al., 2012).

Furthermore, we need the Lipschitz constant of μ_T. In the frequentist setting, the Lipschitz constant for μ_T can be derived as $\mathcal{O}(L_k\sqrt{t\log t})$ by Lemma 4 in (Vakili et al., 2021a) and Lemma 2.9. To obtain a slightly tighter upper bound, we show the following lemma:

Lemma 3.2. *Fix* $\delta \in (0,1)$ *and* $t \in [T]$. *Suppose that Assumptions 2.7 and 2.8 hold. Then,* $\mu_t(\cdot)$ *is Lipschitz continuous with the Lipschitz constant,*

$$L_k\left(B + \frac{R}{\sigma}\sqrt{2\gamma_t + 2\log\left(\frac{d}{\delta}\right)}\right)$$

with probability at least $1-\delta$.

We show the proof in Appendix A.1. Since the MIG γ_T is sublinear for the kernels on which we mainly focus, the upper bound $\mathcal{O}(L_k\sqrt{\gamma_t})$ is tighter than $\mathcal{O}(L_k\sqrt{t\log t})$.

In the Bayesian setting, the upper bound of the Lipschitz constant for μ_T has not been shown to our knowledge. Therefore, we show the following lemma:

Lemma 3.3. *Fix* $\delta \in (0,1)$ *and* $t \in [T]$. *Suppose that Assumptions 2.4 and 2.5 hold and the kernel has mixed partial derivative* $\frac{\partial^2 k(\boldsymbol{x},\boldsymbol{z})}{\partial x_j \partial z_j}$ *for all* $j \in [d]$. *Set* a *and* b *as in Lemma 2.5. Assume that* $(\boldsymbol{x}_i)_{i\in[t]}$ *is independent of* $(\epsilon_i)_{i\in[t]}$ *and* f. *Then,* μ_t *and* $r_t(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) - \mu_t(\boldsymbol{x})$ *satisfies the following:*

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\partial\mu_t(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right| > L\right) \leq 2a\exp\left(-\frac{L^2}{b^2}\right),$$
$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\partial r_t(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right| > L\right) \leq 2a\exp\left(-\frac{L^2}{b^2}\right),$$

for all $j \in [d]$.

See Appendix A.2 for the proof, in which we leverage Slepian's inequality (Proposition A.2.6 in van der Vaart & Wellner, 1996) and the fact that the derivative of the sample path follows GP jointly when the kernel is differentiable.

By leveraging the above results, even if $\mathcal{X}$ is continuous, we can obtain the following upper bound of Eq. (2):

Lemma 3.4. *Suppose that Assumptions 2.7 and 2.8 hold. Fix* $\delta \in (0,1)$ *and* $T \in \mathbb{N}$. *Then, the following holds with probability at least* $1-\delta$*:*

$$E_T \leq 2\beta_{\delta,T}\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] + \mathcal{O}\left(\frac{\max\{\gamma_T, \log(\frac{T}{\delta})\}}{T^2}\right).$$

where $\beta_{\delta,T} = \left(B + \frac{R}{\sigma}\sqrt{2d\log\left(Tdr+1\right) + 2\log\left(\frac{4}{\delta}\right)}\right)^2$.

Lemma 3.5. *Suppose that Assumptions 2.4 and 2.5 hold. Fix* $\delta \in (0,1)$ *and* $T \in \mathbb{N}$. *Then, the following holds with*

probability at least $1-\delta$*:*

$$E_T \leq 2\beta_{\delta,T} \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] + \mathcal{O}\left(\frac{\log(\frac{T}{\delta})}{T^2}\right),$$

where $\beta_{\delta,T} = 2d\log(Tdr+1) + 2\log(2/\delta)$.

See Appendices A.3 and A.4 for the proof.

Consequently, we can minimize Eq. (2) by minimizing $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right]$. In this perspective, the US and RS are theoretically guaranteed because of $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] \leq \max_{\boldsymbol{x}\in\mathcal{X}} \sigma_T^2(\boldsymbol{x})$ and Proposition 2.3. However, the US and RS do not incorporate the information of $\mathcal{P}$. Therefore, the practical effectiveness of the US and RS is limited.

3.3. Other Performance Mesuares

Although we mainly discuss the squared error, other measures can also be bounded from above:

Lemma 3.6. *The worst-case expected absolute error for any* $T \in \mathbb{N}$ *is bounded from above as follows:*

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[|f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*)|\right] \leq \sqrt{E_T},$$

where E_T *is defined as in Eq.* (2).

Lemma 3.7. *The worst-case expectation of entropy for any* $T \in \mathbb{N}$ *is bounded from above as follows:*

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[H\left[f(\boldsymbol{x}^*) \mid \mathcal{D}_T\right]\right] \leq \frac{1}{2}\log\left(2\pi e\tilde{E}_T\right),$$

where $\tilde{E}_T = \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right]$ *and* $H[f(\boldsymbol{x}) \mid \mathcal{D}_T] = \log\left(\sqrt{2\pi e}\sigma_T(\boldsymbol{x})\right)$ *is Shannon entropy.*

See Appendices A.5 and A.6 for the proof. Therefore, minimizing $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right]$ also provides the convergence of the absolute error and the entropy[1].

4. Proposed Methods and Analysis

We aim to design algorithms that enjoy both a similar convergence guarantee as the US and RS and practical effectiveness, incorporating the information of $\mathcal{P}$. In particular, we consider two algorithms inspired by the greedy algorithm and the RS and show theoretical guarantees. Algorithm 1 shows the pseudo-code of the proposed algorithms.

4.1. Algorithms

First, we consider the RS-based algorithm. The algorithm is straightforward as follows:

$$\boldsymbol{x}_t \sim p_t(\boldsymbol{x}), \tag{3}$$

[1]For the absolute error, we can design algorithms that directly reduce σ_t, not σ_t^2, and achieves the similar theoretical guarantee.

Algorithm 1 Proposed DRAL methods

Require: Domain $\mathcal{X}$, GP prior μ and k, ambiguity set $\mathcal{P}$
$\mathcal{D}_0 \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  Update $\sigma_{t-1}^2(\cdot)$ according to Eq. (1)
  Compute $\boldsymbol{x}_t$ according to Eq. (3) or Eq. (4)
end for
Observe $y_1, \ldots, y_T$
Update $\mu_T(\cdot)$ and $\sigma_T^2(\cdot)$ according to Eq. (1)
return $\mu_T(\cdot)$ and $\sigma_T^2(\cdot)$

where $p_t(\boldsymbol{x}) = \arg\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*)]$ and we assume that we can generate the sample from p_t. By using the worst-case distribution p_t for each iteration, this algorithm incorporates the information of $\mathcal{P}$.

Second, we consider the greedy algorithm since its practical efficiency has often been reported (e.g., Bian et al., 2017). However, in our setup, the algorithm that greedily decreases the expected posterior variance should be

$$\arg\min_{\boldsymbol{x}\in\mathcal{X}} \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x})],$$

which requires huge computational time in general due to min-max optimization. Thus, we consider an approximately greedy algorithm as follows:

$$\arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x})],$$

where $p_t(\boldsymbol{x}) = \arg\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*)]$ is the worst-case distribution defined by $(\boldsymbol{x}_i)_{i\in[t-1]}$. On the other hand, the theoretical guarantee for this algorithm is challenging for us. Hence, inspired by the fact that the US has a theoretical guarantee, we set the constraint so that the chosen input is uncertain than $\mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*)]$:

$$\boldsymbol{x}_t = \arg\min_{\boldsymbol{x}\in\mathcal{X}_t} \mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x})], \tag{4}$$

where $\mathcal{X}_t \coloneqq \{\boldsymbol{x} \in \mathcal{X} \mid \sigma_{t-1}^2(\boldsymbol{x}) \geq \mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*)]\}$. Note that $|\mathcal{X}_t| \geq 1$ holds due to the definition.

Necessity of Constraints: We considered that the constraint regarding $\sigma_{t-1}^2(\boldsymbol{x})$ makes the analysis easy since the US that maximize $\sigma_{t-1}(\boldsymbol{x}_t)$ achieves the error convergence, as shown in Proposition 2.3. Therefore, we employ the constraint on $\mathcal{X}_t$. We set the threshold of the constraint as $\mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*)]$ sake of the analysis. On the other hand, our experimental results suggest that the greedy (approximated) expected error reduction algorithm without the constraint shows superior performance. Therefore, removing or alleviating the constraint can be important future work.

Computational Complexity: The computation of the GP has $O(T^3)$ computational complexity, which can be alleviated by scalable GP learning approaches (Liu et al., 2020). However, more careful proofs incorporating an approximation error in GP learning are required to derive similar theoretical analyses as ours, as with (Vakili et al., 2022). On the other hand, the computational complexity of the maximization $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma^2_{t-1}(\boldsymbol{x}^*)]$, the expectation $\mathbb{E}_{p_t(\boldsymbol{x}^*)}[\sigma^2_{t-1}(\boldsymbol{x}^*)]$ and the sampling from $p_t(\boldsymbol{x})$ depends on the ambiguity set $\mathcal{P}$ and may increase in proportion to d. Although our experiments focus on the set of discrete probability distributions with a moderate size of $|\mathcal{X}|$, the above computations may be complicated if $\mathcal{P}$ is the set of continuous probability distributions or $|\mathcal{X}|$ is huge. Extensions to such more computationally intractable ambiguity sets $\mathcal{P}$, e.g., the ball defined by Wasserstein distance and Kullback–Leibler divergence (Hu & Hong, 2013; Frogner et al., 2021), is crucial future work.

4.2. Analysis

Here, we show the error convergence by Eqs. (3) and (4):

Theorem 4.1. *Fix $\delta \in (0,1)$. Assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact subset. If we run Algorithm 1 with Eq. (3), the following holds with probability at least $1-\delta$:*

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma^2_T(\boldsymbol{x}^*)] \leq \frac{2C_1\gamma_T}{T} + \mathcal{O}\left(\frac{\log(1/\delta)}{T}\right),$$

where $C_1 = 1/\log(1+\sigma^{-2})$.

Theorem 4.2. *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact subset. If we run Algorithm 1 with Eq. (4), the following holds:*

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma^2_T(\boldsymbol{x}^*)] \leq \frac{C_1\gamma_T}{T},$$

where $C_1 = 1/\log(1+\sigma^{-2})$.

See Appendix B for the proof, in which Lemma 3 in (Kirschner & Krause, 2018) is used to show Theorem 4.1.

Consequently, our proposed methods achieve almost the same convergence as those of the US and RS shown in Proposition 2.3. Furthermore, by combining Lemmas 3.1, 3.4, and 3.5, we can see that the upper bound of E_T:

Corollary 4.3. *Fix $\delta \in (0,1)$ and $T \in \mathbb{N}$. Then, if we run Algorithm 1, the following hold with probability at least $1-\delta$:*

1. *When Assumption 2.4 or Assumptions 2.7 holds,*

$$E_T = \mathcal{O}\left(\frac{\log(|\mathcal{X}|/\delta)\gamma_T}{T}\right);$$

2. *When Assumptions 2.4 and 2.5 or Assumptions 2.7 and 2.8 hold,*

$$E_T = \mathcal{O}\left(\frac{\log(T/\delta)\gamma_T}{T}\right),$$

Proof. We can obtain the result by combining Lemmas 3.1, 3.4, and 3.5, Theorems 4.1 and 4.2, and the union bound. Note that we assume $|\mathcal{X}| > T$. □

Thus, the error incurred by the proposed algorithms converges to 0 with high probability for discrete and continuous input domains, at least with linear, SE, and Matérn kernels.

5. Related Work

As discussed in Section 1, many AL algorithms have been developed (Settles, 2009). The AL algorithms for classification problems are heavily discussed compared with the regression problem (for example, Houlsby et al., 2011; Zhao et al., 2021; Bickford Smith et al., 2023). In particular, theoretical properties for binary classification problems are well-investigated (Hanneke et al., 2014). On the other hand, the theoretical analysis of AL for the regression problem is relatively limited.

AL is often referred to as optimal experimental design (OED) (Lindley, 1956; Cohn, 1993; Chaloner & Verdinelli, 1995; Cohn et al., 1996; Ryan & Morgan, 2007). The OED frameworks aim to reduce the uncertainty of target parameters or statistical models. For this purpose, many measures for the optimality have been proposed, such as A-optimality (average), D-optimality (determinant), and V-optimality (variance) (Pukelsheim, 2006; Allen-Zhu et al., 2017). The OED methods for various models, such as the linear model (e.g., Allen-Zhu et al., 2017), neural network (e.g., Cohn, 1993), and GPs (e.g., Yu et al., 2006), have been proposed. Our analysis concentrates on the V-optimality of the GPR (Seo et al., 2000; Yu et al., 2006; Shoham & Avron, 2023) and its DR variant, for which, to our knowledge, a theoretical guarantee has not been shown.

In OED or AL, subset selection algorithms (Das & Kempe, 2008) are often leveraged. The subset selection is a general problem whose goal is to find the subset that maximizes some set function. Therefore, the AL can be seen as the subset selection of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \in \mathcal{X}$. In this literature, the submodular property of the set function, for which the greedy algorithm can be optimal, is commonly investigated (Das & Kempe, 2008; Krause et al., 2008b; Guestrin et al., 2005; Bian et al., 2017). The criteria for the AL, such as the D-optimality of the GPR (Krause et al., 2008b; Guestrin et al., 2005), sometimes satisfy the submodular property. Furthermore, Das & Kempe (2008) have shown sufficient conditions for the greedy algorithms to be optimal in Theorem 3.4 (an assumption can be rephrased as $k(\boldsymbol{x},\boldsymbol{x}') \leq \frac{1}{4T}$ in our problem) and Section 8. However, even if we consider minimizing $\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma^2_T(\boldsymbol{x}^*)\right]$ with $|\mathcal{P}| = 1$, these conditions and the submodularity do not hold in general. Therefore, the DR maximization of submodular function (e.g., Krause et al., 2008a; Staib et al., 2019) also cannot be applied directly.

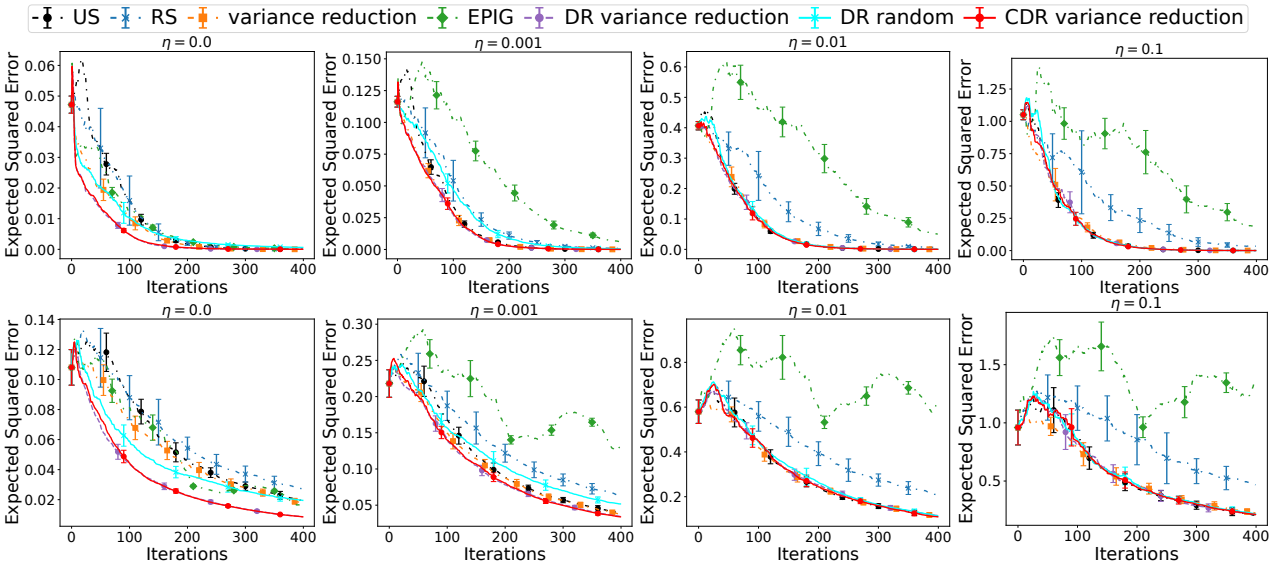

Figure 1. Result of the error E_T in the synthetic data experiments with $\eta = 0, 0.001, 0.01, 0.1$. The horizontal and vertical axes show the number of iterations and E_T, respectively. The error bar shows mean and standard errors for 20 random trials regarding the random initial point (and the algorithm's randomness). The top and bottom columns represent the results of the SE and Matérn kernels, respectively.

Several studies have discussed the target distribution-aware AL. At least from (Sugiyama, 2005), the effectiveness of AL incorporating the information of target distribution for misspecified models has been discussed. Transductive AL (Seo et al., 2000; Yu et al., 2006; Shoham & Avron, 2023) can be interpreted as the expected error minimization when the uniformly random target distribution $p(\boldsymbol{x}) = 1/|\mathcal{X}|$ is specified. Kirsch et al. (2021); Kirsch & Gal (2022); Bickford Smith et al. (2023) extended this setting so that an arbitrary target distribution can be considered. These existing methods are heuristic algorithms that do not guarantee the convergence of the error. On the other hand, Hübotter et al. (2024) show transductive AL methods with theoretical guarantees. However, for the VTR algorithm in (Hübotter et al., 2024), their analysis assumes an assumption of submodularity, which may not hold as discussed immediately after Assumption 3.2 of (Hübotter et al., 2024). Furthermore, their analysis for the VTR algorithm results in $\max_{\boldsymbol{x}\in\mathcal{X}} \sigma_T^2(\boldsymbol{x}) = O(|\mathcal{X}|\gamma_T/T)$, which is vacuous in the usual regime that $T < |\mathcal{X}|$ (see Theorems 3.3 and C.13 and Section C.6.2 of Hübotter et al. (2024)).

Frogner et al. (2021) further extended to DRAL using the AF called expected model change (Settles, 2009) for the non-Bayesian model. In addition, Liu et al. (2015) considers DRAL for non-Bayesian classification models. However, these methods are heuristic greedy algorithms and are not theoretically guaranteed for the prediction error.

The DRAL is inspired by the DR learning (DRL) (Chen & Paschalidis, 2018; Chen et al., 2020). DRL considers learning a robust statistical model by optimizing the model parameter so that the worst-case expected loss is minimized, where the worst-case is taken regarding the target distribution candidates called an *ambiguity set*. Therefore, DRAL is an intuitive extension of DRL to AL.

Another related literature is core-set selection (Sener & Savarese, 2018), which selects the subset of the training dataset to maintain prediction accuracy while reducing the computational cost. Our proposed methods can be applied to the core-set selection for the GPR. However, its effectiveness may be limited since the information on training labels is not leveraged.

Other highly relevant literature is kernelized bandits, also called Bayesian optimization (BO) (Kushner, 1964; Srinivas et al., 2010; Shahriari et al., 2016). BO aims for efficient black-box optimization using the GPR model. For this purpose, several properties of GPs, such as the confidence intervals and the MIG, have been analyzed (Srinivas et al., 2010; Vakili et al., 2021a;b). Our analyses heavily depended on the existing results in this field.

In addition, level set estimation (LSE) (Gotovos et al., 2013; Bogunovic et al., 2016; Inatsu et al., 2024) is an AL framework using the GPR model, which aims to classify the test input set by whether or not a black-box function value exceeds a given threshold. In particular, Inatsu et al. (2021) consider the variant of LSE, which aims to classify by whether or not the DR measure defined by the black-box function exceeds a given threshold. Thus, our problem setup differs from the problem of (Inatsu et al., 2021).

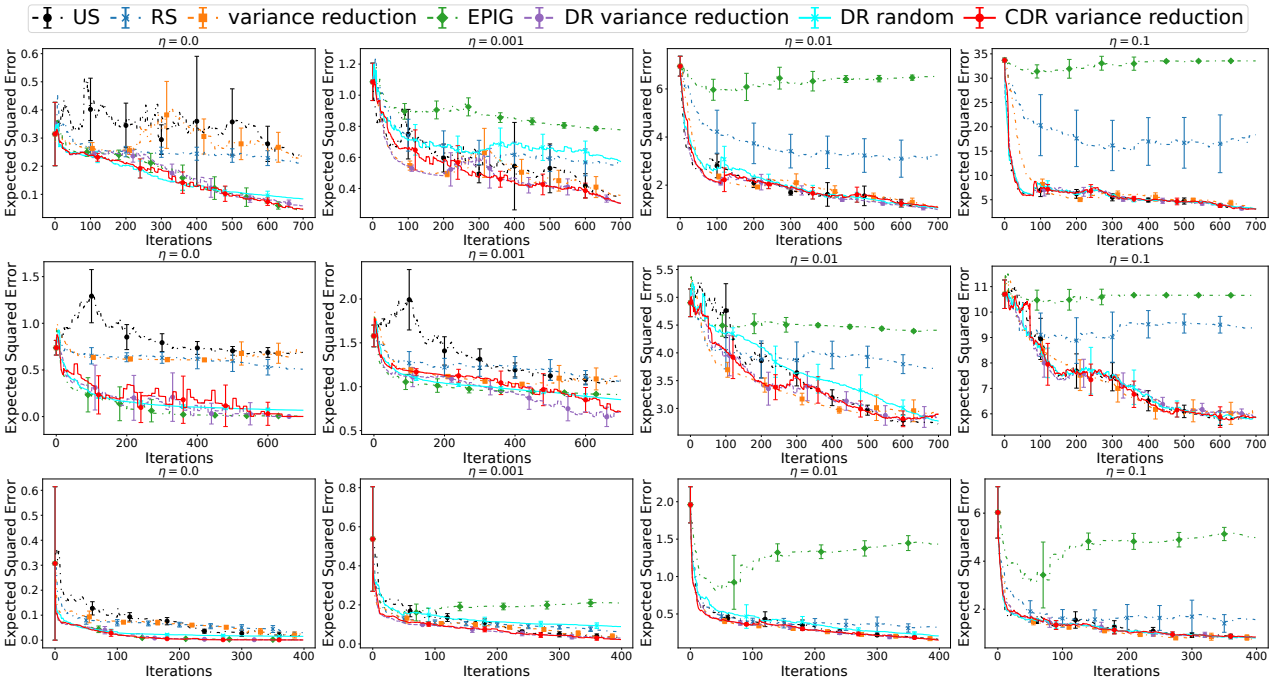

Figure 2. Result of the expected squared error E_T in the real-world data experiments with $\eta = 0, 0.001, 0.01, 0.1$. The horizontal and vertical axes show the number of iterations and E_T, respectively. The error bar shows mean and standard errors for 20 random trials regarding the random initial point (and the algorithm's randomness). The top, middle, and bottom rows represent the results of the King County house sales dataset, red wine quality dataset, and auto MPG dataset, respectively.

6. Experiments

In this section, we demonstrate the effectiveness of the proposed methods via synthetic and real-world datasets. We employ RS, US, variance reduction (Yu et al., 2006), and expected predictive information gain (EPIG) (Bickford Smith et al., 2023) as the baseline. Note that we do not employ the method of (Frogner et al., 2021) since adapting their method to GPR models is not apparent, and they focus on the ambiguity sets defined by Wasserstein distance over a continuous input domain. We show the implementation details of EPIG in Appendix C.2. Furthermore, as the ablation study, we performed the method, referred to as DR variance reduction, that greedily minimizes $E_{p_t(\boldsymbol{x}^*)}\left[\sigma_t^2(\boldsymbol{x}^* \mid \boldsymbol{x})\right]$, that is, the unconstrained version of Eq. (4). We referred to the proposed methods as DR random and constrained DR variance reduction (CDR variance reduction). We evaluate the performance by the error E_T defined in Eq. (2). Furthermore, for the synthetic dataset, we show the result of $\max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x})}\left[\sigma_t^2(\boldsymbol{x}) \mid \mathcal{D}_t\right]$ in Appendix C.1.

In these experiments, by some $\eta \geq 0$, we define the ambiguity set as follows:

$$\mathcal{P} = \{p \in \mathcal{P}_{\mathcal{X}} \mid \|p_{\mathrm{ref}} - p\|_\infty \leq \eta\},$$

where $\mathcal{P}_{\mathcal{X}}$ and p_{ref} are sets of all distributions over $\mathcal{X}$ and some reference distribution, respectively, and $\|\cdot\|_\infty$ denotes L_∞ norm. Note that $\eta = 0$ matches the case that the unique target distribution p_{ref} is specified. Since we consider the case of discrete $\mathcal{X}$, maximization over $\mathcal{P}$ can be written as linear programming for which we used CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018).

The aim of our experiments is to show that the proposed algorithms are consistently well-performing for any size of $\mathcal{P}$, i.e., η, in contrast to the baselines, which can deteriorate for some η. The same parameter η is used for the computation of the performance measure E_T and the actual algorithm. Therefore, we expect that our proposed algorithms show good performance among all $\eta = 0, 0.001, 0.01, 0.1$. On the other hand, if η is large, then $\mathcal{P}$ starts to include more distributions. Therefore, the proposed algorithms behave similarly to the US when η is large. Hence, the proposed algorithms have comparable performance to the usual AL methods when η is large, though the proposed algorithms are superior to those if η is small. On the other hand, EPIG is designed for the case of $\eta = 0$, that is, the test distribution is explicitly specified. However, EPIG aims to decrease the entropy $\mathbb{E}_{p(\boldsymbol{x})}[\log(\sigma_{t-1}(\boldsymbol{x}))]$, not $\mathbb{E}_{p(\boldsymbol{x})}[\sigma_{t-1}^2(\boldsymbol{x})]$. Therefore, EPIG is not suitable for decreasing E_T and can be inefficient in our experiments even for the case of $\eta = 0$.

6.1. Synthetic Data Experiemnts

We set $\mathcal{X} = \{-1, -0.8, \dots, 1\}^3$, where $|\mathcal{X}| = 11^3 = 1331$. The target function f is the sample path from GPs, where we use SE and Matérn-ν kernels with $\nu = 5/2$. We use the fixed hyperparameters of the kernel function in the GPR model, which is used to generate f, and fix $\sigma^2 = 10^{-4}$. The first input $\boldsymbol{x}_1$ is selected uniformly at random, and T is set to 400. Furthermore, we set $p_{\mathrm{ref}} = \mathcal{N}(\mathbf{0}, 0.2\boldsymbol{I}_3)$.

Figure 1 shows the result. We can see that DR and CDR variance reductions show superior performance consistently for all the kernel functions and η, although the DR random is often inferior to those due to the randomness. This result suggests that the DR and CDR variance reductions effectively incorporate the information of $\mathcal{P}$. Furthermore, although the constraint by $\mathcal{X}_t$ is required for the theoretical analysis in CDR variance reduction, we can confirm that it does not sacrifice the practical effectiveness. On the other hand, the usual AL methods, such as US and variance reduction, deteriorate when η is small since they do not incorporate the information of $\mathcal{P}$. When η is large, since our problem approaches the worst-case error minimization, the US and variance reduction result in relatively good results. On the other hand, the EPIG designed for the case $\eta = 0$ is inferior for all η since the EPIG is based on the entropy $\mathcal{O}\left(\log(\sigma_t(\boldsymbol{x}))\right)$, not the squared error.

6.2. Real-World Dataset Experiments

We use the King County house sales[2], the red wine quality (Cortez & Reis, 2009), and the auto MPG datasets (Quinlan, 1993) (See Appendix C.3 for details). For all experiments, we used SE kernels, where the hyperparameters ℓ and σ^2 are adaptively determined by the marginal likelihood maximization (Rasmussen & Williams, 2005) per 10 iterations. The first input is selected uniformly at random. Furthermore, we normalize the inputs and outputs of all datasets before the experiments and set $p_{\mathrm{ref}} = \mathcal{N}(\mathbf{0}, 0.3\boldsymbol{I}_d)$.

Figure 2 shows the result. We can confirm the same tendency that DR and CDR variance reductions show superior performance consistently, as the synthetic data experiments shown in Figure 1. Note that the fluctuations come from the hyperparameter estimation.

7. Conclusion

This paper investigated the DRAL problem for the GPR, in which we aim to reduce the worst-case error E_T. We first showed several properties of this problem for the GPR, which implies that minimizing the variance guarantees a decrease in E_T. Therefore, we designed two algorithms that reduce the target variance and incorporate information about target distribution candidates for practical effectiveness. Then, we proved the theoretical error convergence of the proposed methods, whose practical effectiveness is demonstrated via synthetic and real-world datasets.

Limitation and Future Work: We can consider several future research directions. First, since we do not show the optimality of the convergence rate, developing a (near) optimal algorithm for E_T is vital. For this goal, the approximate submodularity (Bian et al., 2017) may be relevant from the empirical superiority of DR variance reduction. Second, since the expectation over $p(\boldsymbol{x}^*)$ may be intractable, an analysis incorporating the approximation error or developing an efficient algorithm without expectation computation may be crucial (DR random does not require the expectation but is often inefficient). Third, although our analyses only require the existence of the maximum over $\mathcal{P}$, our experiments are limited to the discrete distribution set defined by the L_∞ ball. Thus, more general experiments regarding, e.g., the continuous probability distributions and the ambiguity sets defined by Kullback-Leibler divergence (Hu & Hong, 2013) and Wasserstein distance (Frogner et al., 2021), are interesting from the practical perspective.

Impact Statement

This paper focuses on the theoretical and algorithmic aspects of the machine learning method. We consider that some potential societal consequences of this paper need not necessarily be highlighted here.

Acknowkedgements

This work was partially supported by JST ACT-X Grant Number (JPMJAX23CD and JPMJAX24C3), JST PRESTO Grant Number JPMJPR24J6, JST CREST Grant Numbers (JPMJCR21D3 including AIP challenge program and JPMJCR22N2), JST Moonshot R&D Grant Number JPMJMS2033-05, JSPS KAKENHI Grant Number (JP20H00601, JP23K16943, JP23K19967, JP24K15080, and JP24K20847), NEDO (JPNP20006), and RIKEN Center for Advanced Intelligence Project.

References

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

Adler, R. J. *The Geometry of Random Fields*, volume 62. SIAM, 1981.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S.

[2] https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 126–135. PMLR, 2017.

Bian, A. A., Buhmann, J. M., Krause, A., and Tschiatschek, S. Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 498–507. PMLR, 2017.

Bickford Smith, F., Kirsch, A., Farquhar, S., Gal, Y., Foster, A., and Rainforth, T. Prediction-oriented Bayesian active learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348. PMLR, 2023.

Bogunovic, I., Scarlett, J., Krause, A., and Cevher, V. Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation. In *Advances in neural information processing systems 29*, pp. 1507–1515. Curran Associates, Inc., 2016.

Chaloner, K. and Verdinelli, I. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995.

Chen, R. and Paschalidis, I. C. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19 (13):1–48, 2018.

Chen, R., Paschalidis, I. C., et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4 (1-2):1–243, 2020.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853, 2017.

Cohn, D. Neural network exploration using optimal experiment design. *Advances in neural information processing systems*, 6, 1993.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Cortez, Paulo, C. A. A. F. M. T. and Reis, J. Wine Quality. UCI Machine Learning Repository, 2009.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Using data mining for wine quality assessment. In *Discovery Science: 12th International Conference*, pp. 66–79. Springer, 2009.

Costa, N. D., Pförtner, M., Costa, L. D., and Hennig, P. Sample path regularity of Gaussian processes from the covariance kernel, 2024.

Das, A. and Kempe, D. Algorithms for subset selection in linear regression. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pp. 45–54. Association for Computing Machinery, 2008.

De Freitas, N., Smola, A. J., and Zoghi, M. Exponential regret bounds for Gaussian process bandits with deterministic observations. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pp. 955–962. Omnipress, 2012.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Frogner, C., Claici, S., Chien, E., and Solomon, J. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56): 1–46, 2021.

Ghosal, S. and Roy, A. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413 – 2429, 2006.

Gotovos, A., Casati, N., Hitz, G., and Krause, A. Active learning for level set estimation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1344–1350, 2013.

Guestrin, C., Krause, A., and Singh, A. P. Near-optimal sensor placements in Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 265–272. Association for Computing Machinery, 2005.

Handel, R. V. Probability in high dimension, 2016. Lecture notes. Available in https://web.math.princeton.edu/~rvan/APC550.pdf.

Hanneke, S. et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Hoang, T. N., Low, B. K. H., Jaillet, P., and Kankanhalli, M. Nonmyopic ε-Bayes-optimal active learning of Gaussian processes. In *International conference on machine learning*, pp. 739–747. PMLR, 2014.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

Hübotter, J., Sukhija, B., Treven, L., As, Y., and Krause, A. Transductive active learning: Theory and applications. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124686–124755. Curran Associates, Inc., 2024.

Inatsu, Y., Iwazaki, S., and Takeuchi, I. Active learning for distributionally robust level-set estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4574–4584. PMLR, 2021.

Inatsu, Y., Takeno, S., Kutsukake, K., and Takeuchi, I. Active learning for level set estimation using randomized straddle algorithms. *Transactions on Machine Learning Research*, 2024.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Póczos, B. Parallelised Bayesian optimisation via Thompson sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 133–142, 2018.

Kirsch, A. and Gal, Y. Unifying approaches in active learning and active sampling via Fisher information and information-theoretic quantities. *Transactions on Machine Learning Research*, 2022. Expert Certification.

Kirsch, A., Rainforth, T., and Gal, Y. Test distribution-aware active learning: A principled approach against distribution shift and outliers. *arXiv:2106.11719*, 2021.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 358–384. PMLR, 2018.

Krause, A., McMahan, H. B., Guestrin, C., and Gupta, A. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(93):2761–2801, 2008a.

Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.*, 9:235–284, 2008b.

Kusakawa, S., Takeno, S., Inatsu, Y., Kutsukake, K., Iwazaki, S., Nakano, T., Ujihara, T., Karasuyama, M., and Takeuchi, I. Bayesian optimization for cascade-type multistage processes. *Neural Computation*, 34(12):2408–2431, 2022.

Kushner, H. J. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

Li, Z. and Scarlett, J. Gaussian process bandit optimization with few batches. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 92–107. PMLR, 2022.

Lindley, D. V. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986 – 1005, 1956.

Liu, A., Reyzin, L., and Ziebart, B. Shift-pessimistic active learning using robust bias-aware prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020.

Paria, B., Kandasamy, K., and Póczos, B. A flexible framework for multi-objective Bayesian optimization using random scalarizations. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 766–776, 2020.

Park, M., Lee, S., Hwang, S., and Kim, D. Additive ensemble neural networks. *IEEE Access*, 8:113192–113199, 2020.

Park, S. H. and Kim, S. B. Robust expected model change for active learning in regression. *Applied Intelligence*, 50: 296–313, 2020.

Pukelsheim, F. *Optimal design of experiments*. SIAM, 2006.

Quinlan, R. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Ryan, T. P. and Morgan, J. Modern experimental design. *Journal of Statistical Theory and Practice*, 1(3-4):501–506, 2007.

Salgia, S., Vakili, S., and Zhao, Q. Random exploration in Bayesian optimization: Order-optimal regret and computational efficiency. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43112–43141. PMLR, 2024.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Seo, S., Wallat, M., Graepel, T., and Obermayer, K. Gaussian process regression: active data selection and test point rejection. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pp. 241–246, 2000.

Settles, B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Shahriari, B., Swersky, K., Wang, Z., Adams, R., and De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2016.

Shoham, N. and Avron, H. Experimental Design for Overparameterized Learning With Application to Single Shot Deep Active Learning . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(10):11766–11777, 2023.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022. Omnipress, 2010.

Staib, M., Wilder, B., and Jegelka, S. Distributionally robust submodular maximization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 506–516. PMLR, 2019.

Sugiyama, M. Active learning for misspecified models. In *Advances in Neural Information Processing Systems*, volume 18, pp. 1305–1312. MIT Press, 2005.

Takeno, S., Inatsu, Y., and Karasuyama, M. Randomized Gaussian process upper confidence bound with tighter Bayesian regret bounds. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33490–33515. PMLR, 2023.

Takeno, S., Inatsu, Y., Karasuyama, M., and Takeuchi, I. Posterior sampling-based Bayesian optimization with tighter Bayesian regret bounds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47510–47534. PMLR, 2024.

Vakili, S., Bouziani, N., Jalali, S., Bernacchia, A., and Shiu, D.-s. Optimal order simple regret for Gaussian process bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21202–21215. Curran Associates, Inc., 2021a.

Vakili, S., Khezeli, K., and Picheny, V. On information gain and regret bounds in Gaussian process bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 82–90, 2021b.

Vakili, S., Scarlett, J., Shiu, D.-s., and Bernacchia, A. Improved convergence rates for sparse approximation methods in kernel-based learning. In *International Conference on Machine Learning*, pp. 21960–21983. PMLR, 2022.

van der Vaart, A. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1081–1088, 2006.

Zhao, E., Liu, A., Anandkumar, A., and Yue, Y. Active learning under label shift. In *International conference on artificial intelligence and statistics*, pp. 3412–3420. PMLR, 2021.

A. Proofs for Section 3

A.1. Proof of Lemma 3.2

From the definition of μ_t, we obtain

$$\begin{aligned}\mu_t(\cdot) &\leq \boldsymbol{v}_t^\top(\cdot)\boldsymbol{y}_t \\ &\leq \boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t + \boldsymbol{v}_t^\top(\cdot)\boldsymbol{\epsilon}_t,\end{aligned}$$

where $\boldsymbol{v}_t(\boldsymbol{x}) = \left(\boldsymbol{k}_t^\top(\boldsymbol{x})(\boldsymbol{K}_t + \sigma^2\boldsymbol{I}_t)^{-1}\right)^\top$, $\boldsymbol{f}_t = \big(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_t)\big)^\top$, and $\boldsymbol{\epsilon}_t = (\epsilon_1, \ldots, \epsilon_t)^\top$. Therefore, the Lipschitz constant of μ_t is bounded from above by the Lipscthiz constants of $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t$ and $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{\epsilon}_t$.

For the first term $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t$, we follow the proof of Lemma 4 of (Vakili et al., 2021a). Recall the RKHS-based definition of kernel ridge estimator:

$$\mu_t = \operatorname*{arg\,min}_{\mu \in \mathcal{H}_k} \sum_{i=1}^{t} \big(y_{\boldsymbol{x}_i} - \mu(\boldsymbol{x}_i)\big)^2 + \sigma^2 \|\mu\|_{\mathcal{H}_k}.$$

Therefore, we can derive

$$\begin{aligned}\min_{\mu \in \mathcal{H}_k} \sum_{i=1}^{t} \big(f(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)\big)^2 + \sigma^2\|\mu\|_{\mathcal{H}_k} &= \sum_{i=1}^{t} \big(f(\boldsymbol{x}_i) - \boldsymbol{v}_t^\top(\boldsymbol{x}_i)\boldsymbol{f}_t\big)^2 + \sigma^2 \|\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t\|_{\mathcal{H}_k} \\ &\leq \sum_{i=1}^{t} \big(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\big)^2 + \sigma^2\|f\|_{\mathcal{H}_k} \qquad (\because f \in \mathcal{H}_k) \\ &= \sigma^2 \|f\|_{\mathcal{H}_k}.\end{aligned}$$

Hence, we obtain $\|\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t\|_{\mathcal{H}_k} \leq \|f\|_{\mathcal{H}_k} \leq B$. By combining Lemma 2.9, $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t$ is BL_k Lipschitz continuous.

For the second term $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{\epsilon}_t$, we leverage the confidence bounds of kernel ridge estimator (Theorem 3.11 in Abbasi-Yadkori, 2013). Let $g : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$ as $g(\boldsymbol{x}, 0) = g(\boldsymbol{x}, 1) = 0$ and fix $j \in [d]$. Then, the zero function g belongs to the RKHS with any kernel function $\overline{k}$. Thus, we design the following kernel function $\overline{k}$:

$$\begin{aligned}\overline{k}\left((\boldsymbol{x}, 0), (\boldsymbol{z}, 0)\right) &= k(\boldsymbol{x}, \boldsymbol{z}), \\ \overline{k}\left((\boldsymbol{x}, 1), (\boldsymbol{z}, 1)\right) &= \frac{\partial^2 k(\boldsymbol{x}, \boldsymbol{z})}{\partial x_j \partial z_j}, \\ \overline{k}\left((\boldsymbol{x}, 0), (\boldsymbol{z}, 1)\right) &= \frac{\partial k(\boldsymbol{x}, \boldsymbol{z})}{\partial z_j},\end{aligned}$$

for all $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$. Note that since the kernel function, k has partial derivatives due to Assumption 2.8, the derivative of the kernel and the kernel itself are the kernels again as discussed in, e.g., Sec. 9.4 in (Rasmussen & Williams, 2005) and Sec. 2.2 in (Adler, 1981). Thus, we can interpret $\overline{\boldsymbol{v}}_t^\top(\cdot)\boldsymbol{\epsilon}_t$ as the kernel ridge estimator for $g(\boldsymbol{x}, 1)$, where $\overline{\boldsymbol{v}}_t(\boldsymbol{x}) = \left(\frac{\partial \boldsymbol{k}_t^\top(\boldsymbol{x})}{\partial x_j}(\boldsymbol{K}_t + \sigma^2\boldsymbol{I}_t)^{-1}\right)^\top$. In addition, $\|g\|_{\mathcal{H}_{\overline{k}}} = 0$. Therefore, from Theorem 3.11 in (Abbasi-Yadkori, 2013) and $|g(\boldsymbol{x}, 1) - \overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t| = |\overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t|$, we obtain

$$\Pr\left(|\overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t| \leq \overline{\sigma}_t(\boldsymbol{x})\frac{R}{\sigma}\sqrt{2\gamma_t + 2\log(1/\delta)}, \forall \boldsymbol{x} \in \mathcal{X}\right) \geq 1 - \delta,$$

where $\delta \in (0, 1)$ and $\overline{\sigma}_t(\boldsymbol{x}) = \overline{k}\left((\boldsymbol{x}, 1), (\boldsymbol{x}, 1)\right) - \frac{\partial \boldsymbol{k}_t^\top(\boldsymbol{x})}{\partial x_j}(\boldsymbol{K}_t + \sigma^2\boldsymbol{I}_t)^{-1}\frac{\partial \boldsymbol{k}_t(\boldsymbol{x})}{\partial x_j}$ is the posterior variance that corresponds to this kernel ridge estimation. Note that since the kernel matrix $\boldsymbol{K}_t$ is defined by $k(\boldsymbol{x}, \boldsymbol{z})$, the MIG is the usual one defined by $\mathcal{X}$ and t. In addition, due to the monotonic decreasing property of the posterior variance, $\overline{\sigma}_t(\boldsymbol{x}) \leq L_k$, we obtain

$$\Pr\left(|\overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t| \leq \frac{L_k R}{\sigma}\sqrt{2\gamma_t + 2\log(1/\delta)}, \forall \boldsymbol{x} \in \mathcal{X}\right) \geq 1 - \delta,$$

and thus,

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}|\overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t|\leq\frac{L_kR}{\sigma}\sqrt{2\gamma_t+2\log\left(1/\delta\right)}\right)\geq 1-\delta.$$

Consequently, by using the union bound for all $j\in[d]$, we derive

$$\Pr\left(\sup_{j\in[d]}\sup_{\boldsymbol{x}\in\mathcal{X}}|\overline{\boldsymbol{v}}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t|\leq\frac{L_kR}{\sigma}\sqrt{2\gamma_t+2\log\left(d/\delta\right)}\right)\geq 1-\delta,$$

which shows that $\boldsymbol{v}_t^\top(\boldsymbol{x})\boldsymbol{\epsilon}_t$ is $\frac{L_kR}{\sigma}\sqrt{2\gamma_t+2\log\left(d/\delta\right)}$ Lipschitz continuous.

Combining the Lipschitz constants of $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{f}_t$ and $\boldsymbol{v}_t^\top(\cdot)\boldsymbol{\epsilon}_t$, we can obtain the result.

A.2. Proof of Lemma 3.3

First, we fix $(\boldsymbol{x}_i)_{i\in[t]}$ without loss of generality since

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\partial\mu_t(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right|>L\right)=\mathbb{E}_{(\boldsymbol{x}_i)_{i\in[t]}}\left[\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\partial\mu_t(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right|>L\middle|(\boldsymbol{x}_i)_{i\in[t]}\right)\right].$$

That is, the upper bound of the conditional probability given any $(\boldsymbol{x}_i)_{i\in[t]}$ directly suggests the upper bound of the target probability on the left-hand side. Note that from the assumption $(\boldsymbol{x}_i)_{i\in[t]}$ is independent of $(\epsilon_i)_{i\in[t]}$ and f, the observations $\boldsymbol{y}_t$ follows Gaussian distribution even if $(\boldsymbol{x}_i)_{i\in[t]}$ is fixed.

We leverage Slepian's inequality shown as Proposition A.2.6 in (van der Vaart & Wellner, 1996):

Lemma A.1 (Slepian, Fernique, Marcus, and Shepp)**.** *Let X and Y be separable, mean-zero Gaussian processes indexed by a common index set T such that*

$$\mathbb{E}[(X_s-X_t)^2]\leq\mathbb{E}[(Y_s-Y_t)^2],$$

for all $s,t\in T$. Then,

$$\Pr\left(\sup_{t\in T}X_t\geq\lambda\right)\leq\Pr\left(\sup_{t\in T}Y_t\geq\lambda\right),$$

for all $\lambda>0$.

The separability (Definition 5.22 in Handel, 2016) holds commonly. As discussed in Remark 5.23 in (Handel, 2016), for example, if the sample path is almost surely continuous, then the separability holds. Furthermore, the sample path defined by the commonly used kernel functions, such as linear, SE, and Matérn-ν kernels with $\nu\geq 1$, is continuous almost surely (Costa et al., 2024). In addition, if the kernel function is continuous, the posterior mean function is also continuous, almost surely.

First, we provide the proof of the result regarding $\mu_t(\boldsymbol{x})$. Since $\boldsymbol{y}_t\mid(\boldsymbol{x}_i)_{i\in[t]}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{K}_t+\sigma^2\boldsymbol{I}_t)$, we can see that $\mu_t\mid(\boldsymbol{x}_i)_{i\in[t]}\sim\mathcal{GP}\big(0,k_{\mu_t}(\boldsymbol{x},\boldsymbol{z})\big)$, where $k_{\mu_t}(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{k}_t(\boldsymbol{x})^\top\left(\boldsymbol{K}+\sigma^2\boldsymbol{I}_t\right)^{-1}\boldsymbol{k}_t(\boldsymbol{z})$. Furthermore, it is known that if the kernel has mixed partial derivative $\frac{\partial^2k(\boldsymbol{x},\boldsymbol{z})}{\partial x_j\partial z_j}$, f and its derivative $\partial f(\boldsymbol{x})/\partial x_j$ jointly follow GPs (Rasmussen & Williams, 2005; Adler, 1981). Specifically, the derivative is distributed as

$$f^{(j)}\coloneqq\frac{\partial f(\boldsymbol{x})}{\partial x_j}\sim\mathcal{GP}\left(0,\tilde{k}(\boldsymbol{x},\boldsymbol{z})\coloneqq\frac{\partial^2k(\boldsymbol{u},\boldsymbol{v})}{\partial u_j\partial v_j}\Big|_{\boldsymbol{u}=\boldsymbol{x},\boldsymbol{v}=\boldsymbol{z}}\right),$$

for all $j\in[d]$. Note that since the prior mean of f is zero, the prior mean of $f^{(j)}$ is also zero. As with f, the derivative of μ_t is distributed as

$$\mu_t^{(j)}\coloneqq\frac{\partial\mu_t(\boldsymbol{x})}{\partial x_j}\mid(\boldsymbol{x}_i)_{i\in[t]}\sim\mathcal{GP}\left(0,\tilde{k}_{\mu_t}(\boldsymbol{x},\boldsymbol{z})\coloneqq\frac{\partial^2k_{\mu_t}(\boldsymbol{u},\boldsymbol{v})}{\partial u_j\partial v_j}\Big|_{\boldsymbol{u}=\boldsymbol{x},\boldsymbol{v}=\boldsymbol{z}}\right),$$

for all $j \in [d]$. In addition, the covariance is given as

$$\mathrm{Cov}\left(f(\boldsymbol{x}), f^{(j)}(\boldsymbol{z})\right) = \left.\frac{\partial k(\boldsymbol{x}, \boldsymbol{u})}{\partial u_j}\right|_{\boldsymbol{u}=\boldsymbol{z}}.$$

Then, we see that the posterior variance of the derivative can be obtained in the same way as the usual GP calculation, as follows:

$$\begin{aligned}\mathrm{Var}\left(f^{(j)}(\boldsymbol{x}) \mid \mathcal{D}_t\right) &= \tilde{k}(\boldsymbol{x}, \boldsymbol{x}) - \tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{x}),\\ \mathrm{Cov}\left(f^{(j)}(\boldsymbol{x}), f^{(j)}(\boldsymbol{z}) \mid \mathcal{D}_t\right) &= \tilde{k}(\boldsymbol{x}, \boldsymbol{z}) - \tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{z}).\end{aligned}$$

On the other hand, we can obtain that

$$\begin{aligned}\mathbb{E}\left[\left(f^{(j)}(\boldsymbol{x}) - f^{(j)}(\boldsymbol{z})\right)^2\right] &= \tilde{k}(\boldsymbol{x}, \boldsymbol{x}) + \tilde{k}(\boldsymbol{z}, \boldsymbol{z}) - 2\tilde{k}(\boldsymbol{x}, \boldsymbol{z}),\\ \mathbb{E}\left[\left(\mu_t^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{z})\right)^2 \mid (\boldsymbol{x}_i)_{i\in[t]}\right] &= \tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{x}) + \tilde{k}_{\mu_t}(\boldsymbol{z}, \boldsymbol{z}) - 2\tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{z}),\end{aligned}$$

for all $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$. Then, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left(f^{(j)}(\boldsymbol{x}) - f^{(j)}(\boldsymbol{z})\right)^2\right] - \mathbb{E}\left[\left(\mu_t^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{z})\right)^2 \mid (\boldsymbol{x}_i)_{i\in[t]}\right]\\ &= \tilde{k}(\boldsymbol{x}, \boldsymbol{x}) + \tilde{k}(\boldsymbol{z}, \boldsymbol{z}) - 2\tilde{k}(\boldsymbol{x}, \boldsymbol{z}) - \left(\tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{x}) + \tilde{k}_{\mu_t}(\boldsymbol{z}, \boldsymbol{z}) - 2\tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{z})\right)\\ &= \mathrm{Var}\left(f^{(j)}(\boldsymbol{x}) \mid \mathcal{D}_t\right) + \mathrm{Var}\left(f^{(j)}(\boldsymbol{z}) \mid \mathcal{D}_t\right) - 2\mathrm{Cov}\left(f^{(j)}(\boldsymbol{x}), f^{(j)}(\boldsymbol{z}) \mid \mathcal{D}_t\right) \geq 0.\end{aligned}$$

Consequently, by applying Lemma A.1, we obtain

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \mu_t^{(j)}(\boldsymbol{x}) \geq \lambda \,\middle|\, (\boldsymbol{x}_i)_{i\in[t]}\right) \leq \Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} f_t^{(j)}(\boldsymbol{x}) \geq \lambda\right).$$

Since $\mu_t^{(j)}(\boldsymbol{x})$ and $f_t^{(j)}(\boldsymbol{x})$ follow centered GPs, we obtain

$$\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\mu_t^{(j)}(\boldsymbol{x})\right| \geq \lambda \,\middle|\, (\boldsymbol{x}_i)_{i\in[t]}\right) \leq 2\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \mu_t^{(j)}(\boldsymbol{x}) \geq \lambda \,\middle|\, (\boldsymbol{x}_i)_{i\in[t]}\right) \leq 2\Pr\left(\sup_{\boldsymbol{x}\in\mathcal{X}} \left|f_t^{(j)}(\boldsymbol{x})\right| \geq \lambda\right).$$

Hence, from Lemma 2.5, we obtain the desired result.

We can obtain the result regarding $f^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{x})$ in almost the same proof. We can see that

$$f^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{x}) \sim \mathcal{GP}(0, \tilde{k}(\boldsymbol{x}, \boldsymbol{z}) - \tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{z})).$$

Then,

$$\begin{aligned}&\mathbb{E}\left[\left(f^{(j)}(\boldsymbol{x}) - f^{(j)}(\boldsymbol{z})\right)^2\right] - \mathbb{E}\left[\left(f^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{x}) - \left(f^{(j)}(\boldsymbol{x}) - \mu_t^{(j)}(\boldsymbol{x})\right)\right)^2 \,\middle|\, (\boldsymbol{x}_i)_{i\in[t]}\right]\\ &= \tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{x}) + \tilde{k}_{\mu_t}(\boldsymbol{z}, \boldsymbol{z}) - 2\tilde{k}_{\mu_t}(\boldsymbol{x}, \boldsymbol{z}) \geq 0.\end{aligned}$$

The remaining proof is the same as the case of $\mu_t^{(j)}$.

A.3. Proof of Lemma 3.4

As with the existing studies (e.g., Srinivas et al., 2010), we consider the discretization of input space. Let $\overline{\mathcal{X}} \subset \mathcal{X}$ be a finite set with each dimension equally divided into $\lceil \tau d r \rceil$, where $\tau > 0$. Therefore, $|\overline{\mathcal{X}}| = \lceil \tau d r \rceil^d$ and $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x} - [\boldsymbol{x}]\|_1 \leq \frac{1}{\tau}$, where $[\boldsymbol{x}]$ is the nearest input in $\overline{\mathcal{X}}$, that is, $[\boldsymbol{x}] = \arg\min_{\tilde{\boldsymbol{x}}\in\overline{\mathcal{X}}} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_1$. Note that we leverage $\overline{\mathcal{X}}$ purely for the analysis, and $\overline{\mathcal{X}}$ is not related to the algorithm.

From Assumption 2.8 and Lemma 2.9, we see that f is BL_k Lipschitz continuous. Furthermore, from Lemma 3.2, μ_t is $L_k\left(B + \frac{R}{\sigma}\sqrt{2\gamma_t + 2\log\left(\frac{d}{\delta}\right)}\right)$ Lipschitz continuous with probability at least $1-\delta$. Combining the above, we see that $f - \mu_t$ is $L_k\left(2B + \frac{R}{\sigma}\sqrt{2\gamma_t + 2\log\left(\frac{d}{\delta}\right)}\right)$ Lipschitz continuous with probability at least $1-\delta$.

From the above arguments, by combining Lemma 2.10 and the union bound, the following events hold simultaneously with probability at least $1-\delta$:

1. $f(\boldsymbol{x}) - \mu_T(\boldsymbol{x})$ is $L_{\mathrm{res}}(T)$ Lipschitz continuous, where $L_{\mathrm{res}}(T) = L_k\left(2B + \frac{R}{\sigma}\sqrt{2\gamma_T + 2\log\left(\frac{2d}{\delta}\right)}\right)$.
2. The confidence bounds on $\overline{\mathcal{X}}$ hold; that is,
$$\forall \boldsymbol{x} \in \overline{\mathcal{X}}, |f(\boldsymbol{x}) - \mu_T(\boldsymbol{x})| \leq \beta_{\delta,\tau}^{1/2}\sigma_T(\boldsymbol{x}),$$
where $\beta_{\delta,\tau} = \left(B + \frac{R}{\sigma}\sqrt{2d\log\left(\tau dr + 1\right) + 2\log\left(\frac{4}{\delta}\right)}\right)^2$.

Then, we can obtain the upper bound as follows:

$$\begin{aligned}
E_T &= \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left(f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*)\right)^2\right] \\
&\leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left(f([\boldsymbol{x}^*]) - \mu_T([\boldsymbol{x}^*]) + L_{\mathrm{res}}(T)\|\boldsymbol{x}^* - [\boldsymbol{x}^*]\|_1\right)^2\right] && (\because \text{The above event 1}) \\
&\leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left(f([\boldsymbol{x}^*]) - \mu_T([\boldsymbol{x}^*]) + \frac{L_{\mathrm{res}}(T)}{\tau}\right)^2\right] && (\because \text{The definition of } \overline{\mathcal{X}}) \\
&\leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left(\beta_{\delta,\tau}^{1/2}\sigma_T([\boldsymbol{x}^*]) + \frac{L_{\mathrm{res}}(T)}{\tau}\right)^2\right] && (\because \text{The above event 2}) \\
&\leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\left(\beta_{\delta,\tau}^{1/2}\sigma_T(\boldsymbol{x}^*) + \frac{\beta_{\delta,\tau}^{1/2}L_\sigma + L_{\mathrm{res}}(T)}{\tau}\right)^2\right] && (\because \text{Lemma 2.2}) \\
&\leq 2\beta_{\delta,\tau}\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] + 2\left(\frac{\beta_{\delta,\tau}^{1/2}L_\sigma + L_{\mathrm{res}}(T)}{\tau}\right)^2. && (\because (a+b)^2/2 \leq a^2 + b^2)
\end{aligned}$$

If we set $\tau = T$, noting that $L_{\mathrm{res}}(T) = \mathcal{O}(\sqrt{\gamma_T})$ and $\beta_{\delta,\tau}^{1/2} = \mathcal{O}(\log(T/\delta))$, we obtain the following:

$$E_T \leq 2\beta_{\delta,T}\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] + \mathcal{O}\left(\frac{\max\{\gamma_T, \log(T/\delta)\}}{T^2}\right).$$

Although by setting $\tau = \Omega(T)$, we can make the second term small arbitrarily, $\beta_{\delta,T} = \Theta\left(\log(T/\delta)\right)$ and $\mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_T^2(\boldsymbol{x}^*)\right] = \Omega(1/T)$ (Lemma 4.2 in Takeno et al., 2024). Therefore, since the first term is $\Omega\left(\frac{\log(T/\delta)}{T}\right)$ and $\frac{\max\{\gamma_T, \log(T/\delta)\}}{T^2} = \mathcal{O}\left(\frac{\log(1/\delta)}{T}\right)$ if γ_T is sublinear, we do not set τ more large value for simplicity.

A.4. Proof of Lemma 3.5

As with the existing studies (e.g., Srinivas et al., 2010), we consider the discretization of input space. Let $\overline{\mathcal{X}} \subset \mathcal{X}$ be a finite set with each dimension equally divided into $\lceil \tau dr \rceil$, where $\tau > 0$. Therefore, $|\overline{\mathcal{X}}| = \lceil \tau dr \rceil^d$ and $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x} - [\boldsymbol{x}]\|_1 \leq \frac{1}{\tau}$, where $[\boldsymbol{x}]$ is the nearest input in $\overline{\mathcal{X}}$, that is, $[\boldsymbol{x}] = \arg\min_{\tilde{\boldsymbol{x}}\in\overline{\mathcal{X}}} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_1$. Note that we leverage $\overline{\mathcal{X}}$ purely for the analysis, and $\overline{\mathcal{X}}$ is not related to the algorithm.

In addition, from Lemma 3.3, the following inequality holds with probability at least $1-\delta$:

$$\sup_{j\in d}\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{\partial r_t(\boldsymbol{u})}{\partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{x}}\right| \leq b\sqrt{\log(2ad/\delta)},$$

which implies that L_{res}, the Lipschitz constant of $r_t(\boldsymbol{x}) = f(\boldsymbol{x}) - \mu_T(\boldsymbol{x})$, can be bounded from above.

Then, by combining the above argument, Lemma 2.6, and the union bound, the following events hold simultaneously with probability at least $1 - \delta$:

1. $f(\boldsymbol{x}) - \mu_T(\boldsymbol{x})$ is L_{res} Lipschitz continuous, where $L_{\mathrm{res}} = b\sqrt{\log(4ad/\delta)}$.

2. The confidence bounds on $\overline{\mathcal{X}}$ hold; that is,

$$\forall \boldsymbol{x} \in \overline{\mathcal{X}}, |f(\boldsymbol{x}) - \mu_T(\boldsymbol{x})| \leq \beta_{\delta,\tau}^{1/2} \sigma_T(\boldsymbol{x}),$$

   where $\beta_{\delta,\tau} = 2d\log(\tau dr + 1) + 2\log(2/\delta)$.

Hence, we can obtain the upper bound as follows:

$$\begin{aligned}
E_T &= \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\left(f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*) \right)^2 \right] \\
&\leq \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\left(f([\boldsymbol{x}^*]) - \mu_T([\boldsymbol{x}^*]) + L_{\mathrm{res}} \|\boldsymbol{x}^* - [\boldsymbol{x}^*]\|_1 \right)^2 \right] && (\because \text{The above event 1}) \\
&\leq \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\left(f([\boldsymbol{x}^*]) - \mu_T([\boldsymbol{x}^*]) + \frac{L_{\mathrm{res}}}{\tau} \right)^2 \right] && \left(\because \text{The definition of } \overline{\mathcal{X}}\right) \\
&\leq \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\left(\beta_{\delta,\tau}^{1/2} \sigma_T([\boldsymbol{x}^*]) + \frac{L_{\mathrm{res}}}{\tau} \right)^2 \right] && (\because \text{The above event 2}) \\
&\leq \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\left(\beta_{\delta,\tau}^{1/2} \sigma_T(\boldsymbol{x}^*) + \frac{\beta_{\delta,\tau}^{1/2} L_\sigma + L_{\mathrm{res}}}{\tau} \right)^2 \right] && (\because \text{Lemma 2.2}) \\
&\leq 2\beta_{\delta,\tau} \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\sigma_T^2(\boldsymbol{x}^*) \right] + 2 \left(\frac{\beta_{\delta,\tau}^{1/2} L_\sigma + L_{\mathrm{res}}}{\tau} \right)^2. && \left(\because (a+b)^2/2 \leq a^2 + b^2\right)
\end{aligned}$$

Then, by setting $\tau = T$, we can see that

$$E_T \leq 2\beta_{\delta,T} \max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[\sigma_T^2(\boldsymbol{x}^*) \right] + \mathcal{O}\left(\frac{\log(T/\delta)}{T^2} \right).$$

Although by setting $\tau = \Omega(T)$, we can make the second term small arbitrarily, we do not do so since the first term is dominant compared with $\mathcal{O}\left(\frac{\log(T/\delta)}{T^2}\right)$ term.

A.5. Proof of Lemma 3.6

Since the maximum $\tilde{p}_T(\boldsymbol{x}^*) = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[|f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*)| \right]$ exists, we obtain

$$\begin{aligned}
\max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[|f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*)| \right] &= \mathbb{E}_{\tilde{p}_T(\boldsymbol{x}^*)} \left[|f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*)| \right], \\
&\leq \sqrt{\mathbb{E}_{\tilde{p}_T(\boldsymbol{x}^*)} \left[(f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*))^2 \right]}, \\
&\leq \sqrt{\max_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)} \left[(f(\boldsymbol{x}^*) - \mu_T(\boldsymbol{x}^*))^2 \right]},
\end{aligned}$$

where we used Jensen's inequality.

A.6. Proof of Lemma 3.7

Since the maximum $\overline{p}_T(\boldsymbol{x}^*) = \arg\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[H[f(\boldsymbol{x}) \mid \mathcal{D}_t]\right]$ exists, we obtain

$$\begin{aligned}
\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x})}[H[f(\boldsymbol{x}) \mid \mathcal{D}_t]] &= \frac{1}{2}\mathbb{E}_{\overline{p}_T(\boldsymbol{x}^*)}[\log(2\pi e\sigma_T^2(\boldsymbol{x}))] \\
&\leq \frac{1}{2}\log(2\pi e\mathbb{E}_{\overline{p}_T(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x})]) && (\because \text{Jensen's inequality}) \\
&\leq \frac{1}{2}\log\left(2\pi e \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x})}[\sigma_T^2(\boldsymbol{x})]\right).
\end{aligned}$$

B. Proofs for Section 4

B.1. Proof of Theorem 4.1

From the definition, $\sigma_t^2(\boldsymbol{x})$ is monotonically decreasing along with $\mathcal{D}_{t-1} \subset \mathcal{D}_t$. Therefore, for all $t \leq T$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$,

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_t^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t].$$

Note that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ are random variables due to the randomness of the algorithm. Hence, we obtain

$$\begin{aligned}
\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] &\leq \frac{1}{T}\sum_{t=1}^{T} \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{p_t(\boldsymbol{x}_t)}[\sigma_{t-1}^2(\boldsymbol{x}_t) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}]. && (\because \text{Definition of } p_t)
\end{aligned}$$

Then, we apply the following lemma (Lemma 3 in Kirschner & Krause, 2018):

Lemma B.1. *Let Y_t be any non-negative stochastic process adapted to a filtration $\{\mathcal{F}_t\}$, and define $m_t = \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}]$. Further assume that $Y_t \leq b_t$ for a fixed, non-decreasing sequence $(b_t)_{t\geq 1}$. Then, if $b_T \geq 1$, with probability at least $1-\delta$ for any $T \geq 1$, it holds that,*

$$\sum_{t=1}^{T} m_t \leq 2\sum_{t=1}^{T} Y_t + 4b_T \log\frac{1}{\delta} + 8b_T \log(4b_T) + 1.$$

The random variable $\mathbb{E}_{p_t(\boldsymbol{x})}[\sigma_t^2(\boldsymbol{x}) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t]$ satisfies the condition of this lemma by setting $b_t = 1$ for all $t \in [T]$. Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] &\leq \frac{1}{T}\left(2\sum_{t=1}^{T} \sigma_{t-1}^2(\boldsymbol{x}_t) + 4\log\frac{1}{\delta} + 8\log(4) + 1\right) \\
&\leq \frac{1}{T}\left(2C_1\gamma_T + 4\log\frac{1}{\delta} + 8\log(4) + 1\right).
\end{aligned}$$

Here, we use $\sum_{t=1}^{T} \sigma_{t-1}^2(\boldsymbol{x}_t) \leq C_1\gamma_T$ (Lemma 5.2 in Srinivas et al., 2010).

B.2. Proof of Theorem 4.2

From the definition, $\sigma_t^2(\boldsymbol{x})$ is monotonically decreasing along with $\mathcal{D}_{t-1} \subset \mathcal{D}_t$. Therefore, for all $t \leq T$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$,

$$\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \leq \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_t^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t].$$

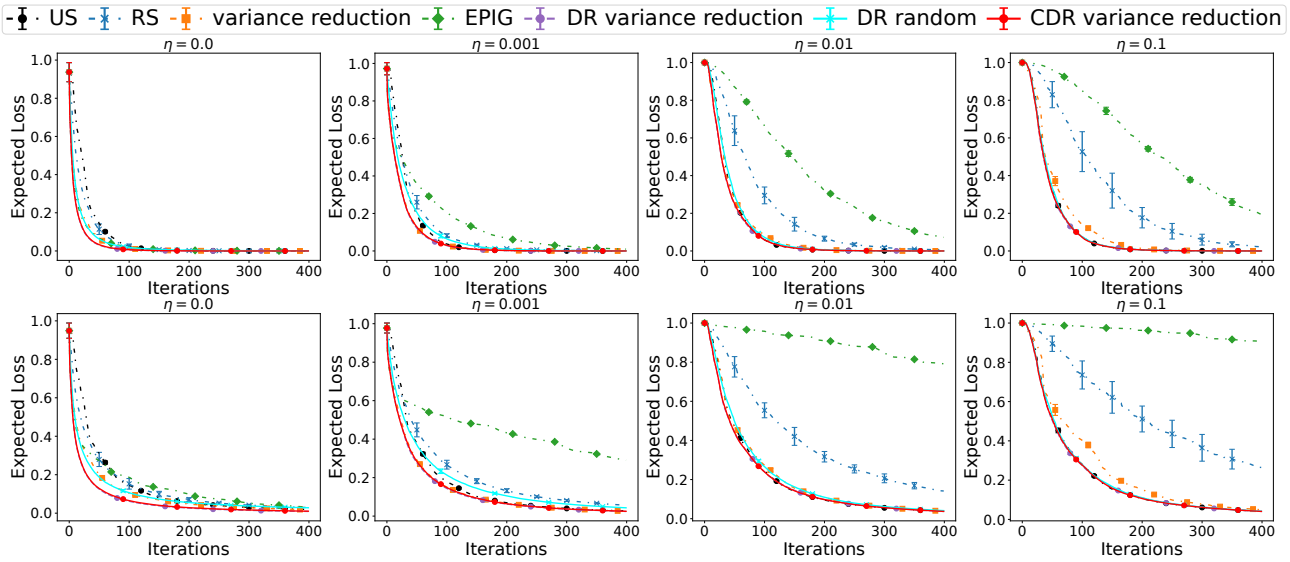

Figure 3. Result of $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_t^2(\boldsymbol{x}^*)\right]$ in the synthetic data experiments with $\eta = 0, 0.001, 0.01, 0.1$. The horizontal and vertical axes show the number of iterations and $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_t^2(\boldsymbol{x}^*)\right]$, respectively. The error bar shows mean and standard errors for 20 random trials regarding the random initial point (and the algorithm's randomness). The top and bottom rows represent the results of the GPR model with SE and Matérn kernels, respectively.

Hence, we obtain

$$\begin{aligned}
\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_T^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] &\leq \frac{1}{T}\sum_{t=1}^{T} \max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}[\sigma_{t-1}^2(\boldsymbol{x}^*) \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \sigma_{t-1}^2(\boldsymbol{x}_t) && (\because \text{Definition of } \mathcal{X}_t) \\
&\leq \frac{C_1\gamma_T}{T}. && (\because \text{Lemma 5.2 in (Srinivas et al., 2010)})
\end{aligned}$$

C. Other Experimental Settings and Results

C.1. Results for Variance

Figure 3 shows the result of $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_t^2(\boldsymbol{x}^*)\right]$, which suggests that the proposed methods effectively minimize $\max_{p\in\mathcal{P}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\sigma_t^2(\boldsymbol{x}^*)\right]$.

C.2. Details on Implementation of EPIG

EPIG is defined as follows (Bickford Smith et al., 2023):

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[H[y_{\boldsymbol{x}^*} \mid \mathcal{D}_{t-1}, \boldsymbol{x}] - H[y_{\boldsymbol{x}^*} \mid \mathcal{D}_{t-1}]\right].$$

Although Bickford Smith et al. (2023) have discussed the efficient computation for EPIG, in the regression problem, the EPIG can be computed analytically except for the expectation over $p(\boldsymbol{x}^*)$ as follows:

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{p(\boldsymbol{x}^*)}\left[\log\left(\frac{\sigma_{t-1}^2(\boldsymbol{x}, \boldsymbol{x}^*)}{\sqrt{\sigma_{t-1}^2(\boldsymbol{x}) + \sigma^2}\sqrt{\sigma_{t-1}^2(\boldsymbol{x}^*) + \sigma^2}}\right)\right],$$

where $\sigma_{t-1}^2(\boldsymbol{x}, \boldsymbol{x}^*)$ is the posterior covariance between $\boldsymbol{x}$ and $\boldsymbol{x}^*$. Since we focus on the discrete input domain in the experiments, the expectation over $p(\boldsymbol{x}^*)$ can also be computed analytically. We used the above equation for the implementation.

C.3. Details on Real-World Datasets

The King County house sales dataset is a dataset used to predict house prices in King County by 7-dimensional features, such as the area and the number of rooms. This dataset has been used for testing the regression (Park et al., 2020), and a similar dataset has also been used for the AL studies (Park & Kim, 2020). Although this dataset includes 20000 data points, we used a random sample of 1000 data points for simplicity.

Red wine quality dataset (Cortez & Reis, 2009) is a dataset used to predict the quality of wines from the wine ingredients expressed by 11-dimensional features. This dataset includes 1600 data points and has been used for the regression problem (Cortez et al., 2009).

Auto MPG dataset (Quinlan, 1993) is a dataset used to predict automobile fuel efficiency from 6-dimensional features, such as the weight of the automobile and engine horsepower. Auto MPG dataset has been used for the AL research (Park & Kim, 2020). This dataset includes 399 data points.