Efficient Black-box Adversarial Attacks via Bayesian Optimization Guided by a Function Prior

Shuyu Cheng [* 1 2] **Yibo Miao** [* 3 4] **Yinpeng Dong** [1 5] **Xiao Yang** [1] **Xiao-Shan Gao** [3 4 6] **Jun Zhu** [1 5]

Abstract

This paper studies the challenging black-box adversarial attack that aims to generate adversarial examples against a black-box model by only using output feedback of the model to input queries. Some previous methods improve the query efficiency by incorporating the gradient of a surrogate white-box model into query-based attacks due to the adversarial transferability. However, the localized gradient is not informative enough, making these methods still query-intensive. In this paper, we propose a Prior-guided Bayesian Optimization (P-BO) algorithm that leverages the surrogate model as a *global function prior* in black-box adversarial attacks. As the surrogate model contains rich prior information of the black-box one, P-BO models the attack objective with a Gaussian process whose mean function is initialized as the surrogate model's loss. Our theoretical analysis on the regret bound indicates that the performance of P-BO may be affected by a bad prior. Therefore, we further propose an adaptive integration strategy to automatically adjust a coefficient on the function prior by minimizing the regret bound. Extensive experiments on image classifiers and large vision-language models demonstrate the superiority of the proposed algorithm in reducing queries and improving attack success rates compared with the state-of-the-art black-box attacks. Code is available at https://github.com/yibo-miao/PBO-Attack.

1. Introduction

A longstanding problem of deep learning is the vulnerability to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015), which are generated by imposing small perturbations to natural examples but can mislead the target model. To identify the weaknesses of deep learning models and evaluate their robustness, adversarial attacks are widely studied to generate the worst-case adversarial examples. Some methods (Goodfellow et al., 2015; Kurakin et al., 2016; Carlini & Wagner, 2017) perform gradient-based optimization to maximize the classification loss, which inevitably requires access to the model architecture and parameters, known as *white-box attacks*. On the other hand, *black-box attacks* (Papernot et al., 2017) assume limited knowledge of the target model, which are more practical in real-world applications.

Tremendous efforts have been made to develop black-box adversarial attacks (Papernot et al., 2017; Chen et al., 2017; Brendel et al., 2018; Dong et al., 2018; Ilyas et al., 2018; Nitin Bhagoji et al., 2018; Tu et al., 2019; Ilyas et al., 2019; Dong et al., 2022), which can be generally categorized into *transfer-based* attacks and *query-based* attacks. In transfer-based attacks, adversarial examples generated for a surrogate model are probable to fool the target model based on the adversarial transferability (Papernot et al., 2016; Liu et al., 2017). Despite the recent improvements (Dong et al., 2018; 2019; Xie et al., 2019; Lin et al., 2020), the success rate of transfer-based attacks is still limited for diverse models. This is attributed to the inherent dependence on the unknown similarity between the surrogate model and the target model, lacking an adjustment process. Differently, query-based attacks generate adversarial examples by leveraging the query feedback of the black-box model. The most prevalent approaches involve estimating the true gradient through zeroth-order optimization (Chen et al., 2017; Nitin Bhagoji et al., 2018; Tu et al., 2019; Ilyas et al., 2018). There are also heuristic algorithms that do not rely on gradient estimation (Alzantot et al., 2019; Guo et al., 2019a; Al-Dujaili et al., 2020; Andriushchenko et al., 2020). The main limitation of these methods is that they inevitably require a tremendous number of queries to perform a successful attack.

To improve the attack success rate and query efficiency, several methods (Cheng et al., 2019; Guo et al., 2019b; Du et al.,

---
[*]Equal contribution [1]Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, BNRist Center, Tsinghua University, Beijing, 100084, China [2]JQ Investments [3]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China [4]University of Chinese Academy of Sciences, Beijing, 100049, China [5]RealAI [6]Kaiyuan International Mathematical Sciences Institute. Correspondence to: Yinpeng Dong <dongyinpeng@tsinghua.edu.cn>.

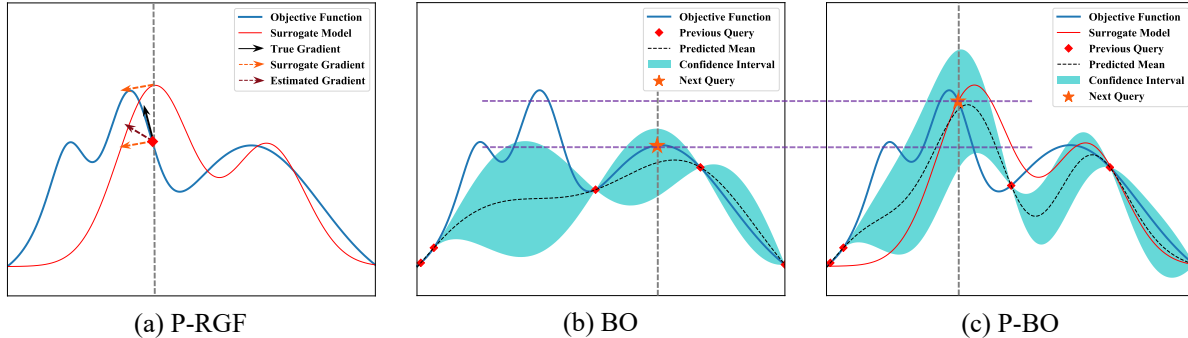(a) P-RGF   (b) BO   (c) P-BO

Figure 1. An illustration of the Prior-guided Random Gradient-Free (P-RGF) (Dong et al., 2022), Bayesian Optimization (BO), and Prior-guided Bayesian Optimization (P-BO) algorithms. The previous approaches, exemplified by P-RGF, adopt a local gradient of the surrogate model for gradient estimation. BO typically employs a zero-mean Gaussian process to approximate the unknown objective function, without leveraging any prior information. Our proposed P-BO algorithm integrates the surrogate model as a function prior into BO, which can better approximate the objective function and thus improve the query efficiency of black-box adversarial attacks.

2020; Yang et al., 2020; Dong et al., 2022) have been proposed to integrate transfer-based attacks with query-based attacks to have the best of both worlds. They typically leverage the input gradient of a surrogate white-box model as the transfer-based prior to improve query-based optimization. However, the surrogate gradient is localized and may not be informative enough, as illustrated in Fig. 1(a). These previous methods are unable to sufficiently exploit the global information of the surrogate model as a *function prior* in the entire space. Consequently, they still require hundreds of queries to successfully attack the target model. Intuitively, the function prior can provide more abundant information of the black-box model, thus can further improve the black-box attack performance if appropriately utilized. Therefore, we explore how to leverage the *global function prior* instead of the *local gradient prior* for improving the efficiency of black-box attacks in this paper.

Bayesian Optimization (BO) (Jones et al., 1998) is a classic black-box optimization approach of finding global optima of the objective functions, which can seamlessly integrate prior information over functions. In recent years, several studies (Zhao et al., 2019; Shukla et al., 2019; Ru et al., 2020) have applied Bayesian optimization to black-box attacks. These methods adopt a Bayesian statistical model (e.g., Gaussian process) to approximate the attack objective and update a posterior distribution according to which the next point to query is chosen, as shown in Fig. 1(b). Although these methods are effective with low query budgets, they usually adopt a zero-mean Gaussian process, which does not leverage any prior information, leaving room for improvement.

To address the aforementioned issues and improve black-box attacks, we propose a **Prior-guided Bayesian Optimization (P-BO)** algorithm, which integrates the surrogate model [1]

as a function prior into Bayesian optimization, as shown in Fig. 1(c). Specifically, P-BO initializes the mean function of the Gaussian process with the surrogate model's loss and updates the posterior distribution given the observed values of the objective function at the sampled locations. We theoretically analyze the regret bound of P-BO, which is closely related to the optimization error and convergence speed of the algorithm. Based on the theoretical analysis, we notice that the straightforward integration of the surrogate model into the Gaussian process may lead to a worse regret bound. Therefore, we further propose an *adaptive integration* strategy, which sets an adjustable coefficient on the surrogate model controlling the strength of utilizing the function prior. The optimal value of the coefficient is determined according to the quality of the prior by minimizing the regret bound. With this technique, P-BO can largely prevent performance degradation when the function prior is useless.

We conduct extensive experiments on CIFAR-10 and ImageNet to confirm the superiority of our method. The results demonstrate that P-BO significantly outperforms the previous state-of-the-art black-box attacks in terms of the success rate and query efficiency. For example, P-BO needs less than 20 queries on average to obtain 100% attack success rates on CIFAR-10, greatly outperforming the existing methods. Furthermore, we conduct experiments on large vision-language models (VLMs) to validate the effectiveness and practicality of our method for attacking the prevailing multimodal foundation models.

2. Preliminaries

In this section, we introduce the preliminary knowledge of black-box attacks and Bayesian optimization. More discussions on related work can be found in Appendix B.

---

[1]To avoid ambiguity, in this paper we use the term "surrogate model" to indicate the white-box model in attacks, rather than the

Bayesian statistical model (also called statistical surrogate) in BO.

2.1. Black-box Adversarial Attacks

Given a natural input $\boldsymbol{x}^{nat} \in \mathbb{R}^D$ and its ground-truth class c, adversarial attack aims to generate an adversarial example $\boldsymbol{x}^{adv}$ by solving a constrained optimization problem:

$$\boldsymbol{x}^{adv} = \arg\max_{\boldsymbol{x} \in A} f(\boldsymbol{x}, c), \tag{1}$$

where f is an attack objective function on top of the target model (e.g., cross-entropy loss, CW loss (Carlini & Wagner, 2017)), $A = \{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{x}^{nat}\|_\infty \leq \epsilon\}$ is the allowed space of adversarial examples and we consider the ℓ_∞ norm with the perturbation budget ϵ in this paper. In the following, we omit the class c in f for simplicity.

Several white-box attacks (Goodfellow et al., 2015; Kurakin et al., 2016; Carlini & Wagner, 2017; Madry et al., 2018) have been proposed to solve problem (1) by gradient-based optimization. The typical projected gradient descent method (PGD) (Madry et al., 2018) performs iterative update as

$$\boldsymbol{x}^{adv}_{t+1} = \Pi_A(\boldsymbol{x}^{adv}_t + \eta \cdot \text{sign}(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{adv}_t))), \tag{2}$$

where Π_A projects the adversarial example onto the ℓ_∞ ball around $\boldsymbol{x}^{nat}$ with radius ϵ, η is the step size, and $\text{sign}(\cdot)$ is the sign function to normalize the gradient.

On the other hand, black-box attacks assume limited knowledge about the target model, which can be challenging yet practical in various real-world applications. Black-box attacks can be roughly divided into transfer-based attacks and query-based attacks. Transfer-based attacks craft adversarial examples for a surrogate model f', which are probable to fool the black-box model f due to the transferability (Papernot et al., 2016; Liu et al., 2017). Some query-based attacks estimate the gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ by zeroth-order optimization methods, when the loss values could be obtained through model queries. For example, the random gradient-free (RGF) method (Ghadimi & Lan, 2013; Duchi et al., 2015; Nesterov & Spokoiny, 2017) estimates the gradient as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \approx \frac{1}{q} \sum_{i=1}^{q} \frac{f(\boldsymbol{x} + \sigma \boldsymbol{u}_i) - f(\boldsymbol{x})}{\sigma} \cdot \boldsymbol{u}_i, \tag{3}$$

where $\{\boldsymbol{u}_i\}_{i=1}^q$ are random directions. Query-based attacks usually demand hundreds of or thousands of queries to successfully generate an adversarial example due to the high-dimensional search space.

There is also plenty of work (Cheng et al., 2019; Guo et al., 2019b; Du et al., 2020; Yang et al., 2020; Ma et al., 2020; Huang & Zhang, 2020; Dong et al., 2022; Yin et al., 2023) integrating transfer-based attacks with query-based attacks to achieve high attack success rate and high query efficiency simultaneously. One straightforward idea is to utilize the gradient of the surrogate model as a transfer-based prior to obtain a more accurate gradient estimate (Cheng et al., 2019;

Dong et al., 2022) or restrict the search space spanned by the surrogate gradient(s) (Guo et al., 2019b; Ma et al., 2020; Yang et al., 2020). Although these methods are effective in expediting convergence and reducing the number of queries, the surrogate gradient is localized and can be misleading, limiting their effectiveness. Besides, other methods learn a generalizable model-based prior based on the surrogate model (Du et al., 2020; Huang & Zhang, 2020; Yin et al., 2023). But these methods require an additional dataset to train the attack generator, which is not applicable when the data is scarce. In this paper, we aim to develop an efficient and elegant black-box attack to leverage the global function prior of the surrogate model without additional data.

2.2. Bayesian Optimization

Bayesian Optimization (BO) (Jones et al., 1998) is an efficient method for finding global optima of black-box optimization problems. BO consists of two key components: a Bayesian statistical model, such as a Gaussian process (GP) (Rasmussen, 2003), which approximates the unknown objective function f; and an acquisition function $\alpha(\cdot)$, which is maximized to recommend the next query location by balancing exploitation and exploration with the most promising expected improvement in function values.

Specifically, assume f a priori follows the Gaussian process with mean 0 (which is common in previous work (Frazier, 2018)) and kernel function $k(\cdot, \cdot)$, denoted as $f \sim \text{GP}(0, k)$. Given the observation data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)\}$ where $y_i = f(\boldsymbol{x}_i)$, the predictive posterior distribution for $f(\boldsymbol{x})$ at a test point $\boldsymbol{x}$, denoted as $p(f(\boldsymbol{x})|\boldsymbol{x}; \mathcal{D})$, follows a Gaussian distribution $\mathcal{N}(\mu_T(\boldsymbol{x}), \sigma_T^2(\boldsymbol{x}))$, where

$$\mu_T(\boldsymbol{x}) = \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1} \mathbf{y}_T, \tag{4}$$

$$\sigma_T^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1} \mathbf{k}_T(\boldsymbol{x}), \tag{5}$$

where $\mathbf{k}_T(\boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}), k(\boldsymbol{x}_2, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_T, \boldsymbol{x})]^\top$, $\mathbf{K}_T$ is a $T \times T$ matrix with its (i, j)-th element being $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\mathbf{y}_T = [y_1, \ldots, y_T]^\top$.

Based on the posterior distribution, we can construct the acquisition function $\alpha(\cdot)$, such as Expected Improvement (EI) (Močkus, 1975), Probability of Improvement (PI) (Jones, 2001), Upper-Confidence Bounds (UCB) (Srinivas et al., 2010), and entropy-based methods (Hernández-Lobato et al., 2014). In this work, we choose UCB as the acquisition function to analyze the regret bound, which is expressed as

$$\alpha(\boldsymbol{x}) = \mu_T(\boldsymbol{x}) + \beta \sigma_T(\boldsymbol{x}), \tag{6}$$

where the coefficient β balances exploration (i.e., encouraging queries in regions with high predictive variance) and exploitation (i.e., encouraging queries in regions with high predictive mean). Subsequently, the next point $\boldsymbol{x}_{T+1}$ is selected by maximizing the acquisition function as $\boldsymbol{x}_{T+1} = \arg\max_{\boldsymbol{x} \in A} \alpha(\boldsymbol{x})$ and then used to query the objective f.

In recent years, several studies have applied BO to black-box attacks. The pioneering work (Suya et al., 2017) initially applies BO to attacking spam email classifiers. Subsequently, Zhao et al. (2019); Shukla et al. (2019); Ru et al. (2020); Miao et al. (2022) extend the application of BO to attacking deep models, providing empirical evidence of its effectiveness. However, to date, there is a notable absence of research integrating prior information into BO for black-box attacks. Beyond black-box attacks, some methods (Souza et al., 2021; Hvarfner et al., 2022) try different ways of integrating prior information into BO, but they do not utilize a deterministic function prior as in our work, and they are not applied to black-box attacks (see Appendix B for details).

3. Methodology

In this section, we introduce the Prior-guided Bayesian Optimization (P-BO) algorithm and the adaptive integration strategy in Sec. 3.1 and Sec. 3.2, respectively.

3.1. Prior-guided Bayesian Optimization

As discussed above, our main motivation is to improve the efficiency of black-box attacks by leveraging the global function prior of a surrogate model. As Bayesian optimization (BO) enables global optimization of the black-box objective function by building a probabilistic model, it can seamlessly integrate prior information over functions. Therefore, we propose a novel **Prior-guided Bayesian Optimization (P-BO)** algorithm for more efficient black-box attacks.

Specifically, we consider optimizing the objective function $f(\boldsymbol{x})$ in Eq. (1) for attacking a black-box model. We assume that we have a surrogate white-box model and can obtain its objective function f' in the same form of f. Due to the similarity between the surrogate model and the black-box one, f' could exhibit some similarities to f in the function space. And that is why adversarial examples generated by optimizing f' can have a certain probability to also mislead f (i.e., adversarial transferability) (Papernot et al., 2017; Liu et al., 2017; Dong et al., 2018). Therefore, we regard f' as a function prior to optimize the black-box objective f. P-BO initializes the mean function of the Gaussian process with the function prior f', so f a priori follows $f \sim \mathrm{GP}(f', k)$. Similar to Eq. (4) and Eq. (5), the posterior distribution $p(f(\boldsymbol{x})|\boldsymbol{x}; \mathcal{D})$ given the observation data $\mathcal{D}$ also follows a Gaussian distribution $\mathcal{N}(\mu_T(\boldsymbol{x}), \sigma_T^2(\boldsymbol{x}))$, where

$$\mu_T(\boldsymbol{x}) = \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1}(\mathbf{y}_T - \mathbf{y}_T') + f'(\boldsymbol{x}), \quad (7)$$

$$\sigma_T^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1} \mathbf{k}_T(\boldsymbol{x}). \quad (8)$$

where $\mathbf{k}_T(\boldsymbol{x})$, $\mathbf{K}_T$, $\mathbf{y}_T$ are the same with those in Eq. (4) and Eq. (5), and $\mathbf{y}_T' = [f'(\boldsymbol{x}_1), \dots, f'(\boldsymbol{x}_T)]^\top$.

By comparing Eq. (7)-(8) of P-BO with Eq. (4)-(5) of BO, we notice that only the mean $\mu_T(\boldsymbol{x})$ changes after the function prior is introduced. Hence P-BO can be viewed as first

modeling the residual $f - f'$ using its observed values with a zero-mean Gaussian process $\mathrm{GP}(0, k)$, and then adding f' to form the model of f. Intuitively, the approximation error of f becomes dependent on the magnitude of $f - f'$. When $f = f'$, we could extract an accurate estimation of f without the requirement of observed values since $\mu_T = f$. It is reasonable to anticipate that when f' is closer to f, the performance of P-BO will be better.

In the following, we theoretically analyze the regret bound of our proposed P-BO algorithm. Following Srinivas et al. (2010), we define the instantaneous regret $r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$ at the t-th iteration, where $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in A} f(\boldsymbol{x})$ is the global maximum. The cumulative regret R_T after T iterations is $R_T = \sum_{t=1}^{T} r_t$. In the realm of black-box optimization, we are typically concerned with the optimization error $\min_{1 \le t \le T} r_t$, which can be upper bounded by $\frac{R_T}{T}$. Therefore, the convergence rate of the optimization algorithm is closely related to the regret R_T. Below, we derive the regret bound of P-BO with the statistical model $\mathrm{GP}(f', k)$ and the UCB acquisition function. For technical reasons, we shall consider an observation noise $\mathcal{N}(0, \sigma^2)$ in the Gaussian process modeling procedure (replacing $\mathbf{K}_T$ in Eq. (7)-(8) with $\mathbf{K}_T + \sigma^2 \mathbf{I}$, see e.g. Theorem 3.1 of Kanagawa et al. (2018)), but the actual observed value of f is deterministic and without noise.

Theorem 3.1. *(Proof in Appendix A.1) Assume f and f' lie in the Reproducing Kernel Hilbert Space (RKHS) corresponding to kernel k, and let $\|\cdot\|_k$ denote the RKHS norm. In Bayesian optimization, suppose we model f by $\mathrm{GP}(f', k)$ with observation noise $\mathcal{N}(0, \sigma^2)$, and we use the UCB acquisition function defined in Eq. (6) where $\beta = \|f - f'\|_k$. Then the regret R_T satisfies*

$$R_T \le \|f - f'\|_k \sqrt{\frac{8}{\log(1 + \sigma^{-2})} T \gamma_T}, \quad (9)$$

where $\gamma_T = \frac{1}{2} \max_{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T \in A} \log |\mathbf{I} + \sigma^{-2} \mathbf{K}_T|$ in which $\mathbf{K}_T$ is the covariance matrix with its (i, j)-th element being $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Remark 3.2. Intuitively, incorporating the function prior f' into the modeling of f using the Gaussian process $\mathrm{GP}(f', k)$ involves substituting $\|f\|_k$ with $\|f - f'\|_k$ in the upper bound of the regret R_T in BO (Srinivas et al., 2010). Consequently, when $f' \approx f$, employing the Gaussian process $\mathrm{GP}(f', k)$ to model f significantly lowers the regret.

Theorem 3.1 indicates that the regret bound is proportional to $\|f - f'\|_k$. To achieve better performance of using P-BO, we desire a small value of $\|f - f'\|_k$, at least satisfying $\|f - f'\|_k \le \|f\|_k$. Although this requirement holds when f' approximates f well, for functions defined in a high-dimensional space, this requirement is quite strong. In Appendix C, we provide an example showing that when f

and f' are both random linear functions, $\|f - f'\|_k > \|f\|_k$ with high probability. This implies that when the target and prior functions are not closely aligned, directly modeling f with $\mathrm{GP}(f', k)$ could usually degrade the performance of P-BO compared with the vanilla BO. Therefore, we further propose an adaptive integration strategy, as detailed next.

3.2. Adaptive Integration Strategy

To address the aforementioned problem, our fundamental idea is to set a coefficient λ on the function prior and automatically adjust λ according to the quality of the prior. We aim to ensure that the performance of P-BO would not be affected by a useless prior, and to optimize λ to achieve better performance in the presence of a useful prior.

Based on our theoretical analysis in Theorem 3.1, we employ a straightforward approach by replacing f' with $\lambda f'$, where $\lambda \in \mathbb{R}$ can be interpreted as the weight of integrating the function prior. Therefore, P-BO models the objective as $f \sim \mathrm{GP}(\lambda f', k)$, whose posterior distribution can be similarly derived. Note that $\lambda = 0$ corresponds to the zero-mean Gaussian process and $\lambda = 1$ corresponds to the case presented in Sec. 3.1. The motivation behind this design is:

$$\arg\min_{\lambda} \|f - \lambda f'\|_k^2 = \|f\|_k^2 - \frac{\langle f, f' \rangle_k^2}{\|f'\|_k^2} \leq \|f\|_k^2, \quad (10)$$

where the minimum is achieved at $\lambda^* = \frac{\langle f, f' \rangle_k}{\|f'\|_k^2}$, and $\langle \cdot, \cdot \rangle_k$ denotes the inner product in the RKHS corresponding to the kernel k. Thus, by selecting an appropriate coefficient λ, we can ensure that the regret bound in Theorem 3.1 is reduced.

Although the optimal solution $\lambda^* = \frac{\langle f, f' \rangle_k}{\|f'\|_k^2}$ has an analytical form, the function f is unknown in black-box optimization. The inner product and norm of functions in the RKHS are also difficult to compute accurately. Consequently, calculating the optimal coefficient λ^* poses a significant challenge. To address this, we develop a heuristic approximation method based on the following insightful proposition to estimate the norm $\|f - \lambda f'\|_k$ in the RKHS. This estimate is then utilized to further determine the optimal coefficient λ minimizing $\|f - \lambda f'\|_k^2$.

Proposition 3.3. *(Proof in Appendix A.2) Given the function μ_T defined in Eq. (4), which is the predictive posterior mean of the objective function modeled by $\mathrm{GP}(0, k)$ given observation data $\mathcal{D}$, and let $\mathcal{H}$ be the RKHS corresponding to the kernel k, then we have*

$$\mu_T = \arg\min_{h \in \mathcal{H}} \|h\|_k, \ \ \text{s.t.} \ \ \forall 1 \leq t \leq T, h(\boldsymbol{x}_t) = y_t, \ (11)$$

and $\|\mu_T\|_k^2 = \mathbf{y}_T^\top \mathbf{K}_T^{-1} \mathbf{y}_T$.

Remark 3.4. μ_T could be viewed as the "smoothest" (in the sense of minimizing the RKHS norm) interpolation for

---

Algorithm 1 Prior-guided Bayesian Optimization (P-BO)

1: **Input:** Objective function f; function prior f'; search space A; number of queries N; kernel function k; balancing coefficient β.
2: **Output:** Approximate solution $\boldsymbol{x}^*$ maximizing f.
3: Construct an initial dataset: $\mathcal{D} \leftarrow \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_S, y_S)\}$;
4: **for** $T = S$ **to** $N - 1$ **do**
5:    Normalize $\mathbf{y}_T$, $\mathbf{y}'_T$, and f' (see text for details);
6:    Solve λ^* by maximizing $\log p(\mathcal{D}|\lambda)$;
7:    Compute posterior $p(f(\boldsymbol{x})|\boldsymbol{x}; \mathcal{D}) = \mathcal{N}(\mu_T(\boldsymbol{x}), \sigma_T^2(\boldsymbol{x}))$ of $\mathrm{GP}(\lambda f', k)$, where

$$\mu_T(\boldsymbol{x}) = \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1}(\mathbf{y}_T - \lambda^* \mathbf{y}'_T) + \lambda^* f'(\boldsymbol{x}),$$

$$\sigma_T^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_T(\boldsymbol{x})^\top \mathbf{K}_T^{-1} \mathbf{k}_T(\boldsymbol{x});$$

8:    Compute acquisition function: $\alpha(\boldsymbol{x}) = \mu_T(\boldsymbol{x}) + \beta \sigma_T(\boldsymbol{x})$;
9:    Obtain the next query point: $\boldsymbol{x}_{T+1} = \arg\max_{\boldsymbol{x} \in A} \alpha(\boldsymbol{x})$;
10:   Query f at $\boldsymbol{x}_{T+1}$ to obtain $y_{T+1} = f(\boldsymbol{x}_{T=1})$;
11:   Update the dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_{T+1}, y_{T+1})\}$;
12: **end for**
13: **return** $\boldsymbol{x}_N$.

---

the given dataset $\mathcal{D}$. Using its norm to estimate $\|h\|_k$, it can be regarded that $\|h\|_k^2 \approx \|\mu_T\|_k^2 = \mathbf{y}_T^\top \mathbf{K}_T^{-1} \mathbf{y}_T$. This estimate might be coarse when the size of $\mathcal{D}$ is small since it is a lower bound, but it is almost the best we can do, since giving an upper bound of $\|h\|_k$ with the only restriction $\mathcal{D}$ is intuitively infeasible without more assumptions.

To optimize $\|f - \lambda f'\|_k^2$, whose values at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ can be stacked into the vector $\mathbf{y}_T - \lambda \mathbf{y}'_T$, we can approximate $\|f - \lambda f'\|_k^2$ using Remark 3.4 with $h = f - \lambda f'$:

$$\begin{aligned} \|f - \lambda f'\|_k^2 &\approx (\mathbf{y}_T - \lambda \mathbf{y}'_T)^\top \mathbf{K}_T^{-1}(\mathbf{y}_T - \lambda \mathbf{y}'_T) \\ &= -\log \mathcal{N}(\mathbf{y}_T|\lambda \mathbf{y}'_T, \mathbf{K}_T) + \text{const} \quad (12) \\ &= -\log p(\mathcal{D}|\lambda) + \text{const}, \end{aligned}$$

where the const term is independent of λ, and $\log p(\mathcal{D}|\lambda)$ represents the log-likelihood of $\mathcal{D}$ under the modeling of $\mathrm{GP}(\lambda f', k)$. Notably, solving $\arg\min_\lambda \|f - \lambda f'\|_k^2$ can be approximated as maximizing the log-likelihood, i.e., solving $\arg\max_\lambda \log p(\mathcal{D}|\lambda)$. Therefore, for the sake of convenience, in implementation we can treat the integration coefficient λ as a hyperparameter in the Gaussian process model, and adaptively adjust λ together with other hyperparameters (e.g., the hyperparameters of the kernel) by maximizing the log-likelihood of the dataset.

In summary, the algorithm of P-BO is outlined in Alg. 1. First, we construct an initial dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^S$ by randomly sampling $\boldsymbol{x}_i$ in the search space A and obtaining $y_i = f(\boldsymbol{x}_i)$. At each iteration, we normalize $\mathbf{y}_T$, $\mathbf{y}'_T$, and f' to a good and similar numerical range: $\mathbf{y}_T \leftarrow \frac{\mathbf{y}_T - \mu}{\sigma}$, $\mathbf{y}'_T \leftarrow \frac{\mathbf{y}'_T - \mu'}{\sigma'}$, $f' \leftarrow \frac{f' - \mu'}{\sigma'}$, where μ and σ represent the mean and standard deviation of $\mathbf{y}_T$, and μ' and σ' represent the mean and standard deviation of $\mathbf{y}'_T$. Subsequently, we obtain an optimal integration coefficient λ^* through maximum likelihood. Following this, we model f by Gaussian

Table 1. The experimental results of black-box attacks against DenseNet-121, ResNet-50, and SENet-18 under the ℓ_∞ norm on CIFAR-10. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**.

Method	ResNet-50			DenseNet-121			SENet-18		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
NES (Ilyas et al., 2018)	79.9%	353	306	75.5%	358	306	80.2%	335	255
Bandits$_T$ (Ilyas et al., 2019)	92.0%	214	130	90.7%	240	158	93.4%	203	128
$\mathcal{N}$ATTACK (Li et al., 2019)	94.3%	296	204	93.9%	309	255	94.9%	270	204
SignHunter (Al-Dujaili et al., 2020)	84.4%	267	180	85.5%	258	158	86.1%	251	156
Square (Andriushchenko et al., 2020)	94.2%	206	100	94.3%	216	124	94.9%	197	97
NPAttack (Bai et al., 2023)	98.5%	151	75	96.0%	86	50	95.5%	124	50
RGF (Cheng et al., 2019)	86.9%	232	149	83.1%	254	179	87.6%	223	149
P-RGF (Dong et al., 2022)	98.4%	55	25	95.8%	76	27	98.3%	49	23
BO (Ru et al., 2020)	99.6%	83	44	99.7%	93	51	99.7%	81	41
P-BO ($\lambda = 1$, **ours**)	99.9%	16	**11**	99.7%	25	**11**	99.9%	**14**	**11**
P-BO (λ^*, **ours**)	**100.0%**	**15**	**11**	**100.0%**	**19**	**11**	**100.0%**	**14**	**11**

process $GP(\lambda^* f', k)$ and compute the posterior distribution $p(f(\boldsymbol{x})|\boldsymbol{x}; \mathcal{D})$. Based on the posterior distribution, we compute the acquisition function $\alpha(\boldsymbol{x}) = \mu_T(\boldsymbol{x}) + \beta\sigma_T(\boldsymbol{x})$ and maximize the acquisition function $\alpha(\boldsymbol{x})$ to obtain the next query point $\boldsymbol{x}_{T+1}$. Finally, we query the objective function f to obtain $y_{T+1} = f(\boldsymbol{x}_{T+1})$ and update the dataset $\mathcal{D}$. In the algorithm, the maximization problems of finding λ^* and $\boldsymbol{x}_{T+1}$ are solved by gradient-based methods (Frazier, 2018).

4. Experiments

In this section, we present the empirical results to demonstrate the effectiveness of the proposed methods on attacking black-box models. We perform untargeted attacks under the ℓ_∞ norm on CIFAR-10 (Krizhevsky & Hinton, 2009) and ImageNet (Russakovsky et al., 2015) for image classifiers, and MS-COCO (Lin et al., 2014) for large Vision-Language Models (VLMs). We first specify the experimental setting in Sec. 4.1. Then we show the results on CIFAR-10 in Sec. 4.2, ImageNet in Sec. 4.3, and VLMs in Sec. 4.4, respectively. We also conduct experiments on defense models in Sec. 4.5 and show the performance of adaptive integration strategy in Sec. 4.6. For further details, including results of targeted attacks, performance under the ℓ_2 norm, comparative analyses across different surrogate models, comparisons with expanded baseline and more ablation studies, please refer to Appendix D.

4.1. Experimental Settings

CIFAR-10 (Krizhevsky & Hinton, 2009). We adopt all the 10,000 test images for evaluations, which are in $[0, 1]$. We consider 3 black-box target models: ResNet-50 (He et al., 2016a), DenseNet-121 (Huang et al., 2017), and SENet-18 (Hu et al., 2018). We adopt a Wide ResNet model (WRN-34-10) (Zagoruyko & Komodakis, 2016) as the surrogate model. For BO and P-BO, we set an initial dataset $\mathcal{D}$ containing $S = 10$ randomly sampled points, and use the Matern-5/2 kernel. The loss function f is the CW loss (Carlini & Wagner, 2017) since it performs better than the cross-entropy

loss on CIFAR-10. Following Ru et al. (2020), we set the query budget $N = 1000$. The perturbation size is $\epsilon = \frac{8}{255}$ under the ℓ_∞ norm. As in previous work (Andriushchenko et al., 2020; Dong et al., 2022), we adopt the attack success rate (ASR), the mean and median number of queries of successful attacks to evaluate the performance.

ImageNet (Russakovsky et al., 2015). We choose 1,000 images randomly from the ILSVRC 2012 validation set to perform evaluations. Those images are normalized to $[0, 1]$. The black-box target models include Inception-v3 (Szegedy et al., 2016), MobileNet-v2 (Sandler et al., 2018), and ViT-B/16 (Dosovitskiy et al., 2021). We use the ResNet-152 (He et al., 2016b) as the surrogate model to provide the prior information. Similar to CIFAR-10, we set $S = 10$, $N = 1000$, and use the Matern-5/2 kernel. The perturbation size under ℓ_∞ norm is $\epsilon = \frac{8}{255}$. When employing dimensionality reduction of the search space, we use $56 \times 56 \times 3$, where the original dimensionality is $224 \times 224 \times 3$. The subscript "D" denotes the methods with dimensionality reduction.

MS-COCO (Lin et al., 2014). We select 120 image-caption pairs randomly from MS-COCO for the image captioning task. The images are normalized to $[0, 1]$. We consider three black-box VLMs: InstructBLIP (Dai et al., 2023), mPLUG-Owl (Ye et al., 2023), and VPGTrans (Zhang et al., 2023). We use MiniGPT-4 (Zhu et al., 2023) as the surrogate model. The experimental settings are the same as ImageNet except $\epsilon = \frac{16}{255}$. We conduct untargeted attacks on a selected word or phrase describing the main object of each image-caption pair, utilizing the prompt, "Write a caption for this image". We use log-likelihood between generated caption and selected word as y_t to update the model. A successful attack is reported if the generated caption does not contain the specified word or similar words (measured by CLIP-score (Radford et al., 2021) greater than 0.95).

4.2. Experimental Results on CIFAR-10

We compare P-BO with the fixed function prior $\lambda = 1$ and adaptive coefficient λ^* with various baselines – NES (Ilyas

Table 2. The experimental results of black-box attacks against Inception-v3, MobileNet-v2, and ViT-B/16 under the ℓ_∞ norm on ImageNet. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**. The subscript "D" denotes the methods with dimensionality reduction.

Method	Inception-v3			MobileNet-v2			ViT-B/16		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
NES (Ilyas et al., 2018)	53.6%	348	306	61.5%	320	255	41.4%	339	255
Bandits$_T$ (Ilyas et al., 2019)	45.8%	300	209	67.4%	262	141	42.3%	303	200
$\mathcal{N}$ATTACK (Li et al., 2019)	78.7%	347	255	87.1%	335	255	63.4%	352	306
SignHunter (Al-Dujaili et al., 2020)	80.3%	223	128	81.7%	252	156	73.0%	212	96
Square (Andriushchenko et al., 2020)	76.8%	194	100	97.8%	95	**17**	71.0%	188	**74**
NPAttack (Bai et al., 2023)	75.0%	381	300	86.9%	428	400	63.6%	374	300
RGF (Cheng et al., 2019)	54.5%	260	167	65.2%	206	107	44.2%	269	185
P-RGF (Dong et al., 2022)	72.9%	186	99	84.9%	159	71	52.7%	222	129
BO (Ru et al., 2020)	89.2%	169	105	97.8%	133	85	77.5%	219	140
P-BO ($\lambda = 1$, **ours**)	60.8%	186	75	78.7%	136	26	43.9%	223	115
P-BO (λ^*, **ours**)	91.4%	115	58	98.7%	95	54	83.6%	189	93
Bandits$_{TD}$ (Ilyas et al., 2019)	60.3%	260	142	76.5%	233	110	54.2%	267	148
RGF$_D$ (Cheng et al., 2019)	73.6%	205	107	69.8%	181	89	55.8%	232	149
P-RGF$_D$ (Dong et al., 2022)	80.1%	194	113	87.5%	162	75	61.3%	236	145
BO$_D$ (Ru et al., 2020)	94.1%	104	58	98.2%	102	61	86.7%	170	116
P-BO$_D$ ($\lambda = 1$, **ours**)	85.4%	182	90	86.8%	193	56	67.5%	236	240
P-BO$_D$ (λ^*, **ours**)	**94.4%**	**81**	**45**	**98.8%**	**94**	60	**88.2%**	**148**	81

et al., 2018), Bandits$_T$ (Ilyas et al., 2019), $\mathcal{N}$ATTACK (Li et al., 2019), SignHunter (Al-Dujaili et al., 2020), Square attack (Andriushchenko et al., 2020), NPAttack (Bai et al., 2023), RGF (Cheng et al., 2019), P-RGF (Dong et al., 2022), and BO (i.e., GP-BO in Ru et al. (2020)). For all methods, we restrict the maximum number of queries for each image as $1,000$. We report a successful attack if a method generates an adversarial example within $1,000$ queries and the size of perturbation is smaller than the budget (i.e., $\epsilon = \frac{8}{255}$).

Table 1 shows the results, where we report the attack success rate and the average/median number of queries needed to successfully generate an adversarial example. Note that for BO and P-BO, the query count includes $S = 10$ random queries constructing the initial dataset $\mathcal{D}$. We have the following observations. First, compared with the state-of-the-art attacks, the proposed P-BO generally leads to higher attack success rates and requires much fewer queries. While most attack methods can achieve success rates above 90%, only P-BO achieves an absolute 100% attack success rate. Moreover, P-BO requires less than 20 queries, significantly improving the efficiency. Second, the function prior provides useful prior information for black-box attacks since P-BO outperforms the vanilla BO. Third, P-BO outperforms P-RGF, demonstrating the effectiveness of leveraging the surrogate model as a function prior rather than the gradient prior. Fourth, the use of the adaptive coefficient λ^* in P-BO enhances the attack success rates and reduces the average number of queries compared with $\lambda = 1$, highlighting the effectiveness of using an adaptive coefficient derived by the proposed adaptive integration strategy.

4.3. Experimental Results on ImageNet

Similar to the experiments on CIFAR-10, we also compare the performance of P-BO with these baselines on ImageNet.

We also restrict the maximum number of queries for each image to be $1,000$. For some methods including Bandits, RGF, P-RGF, BO, and P-BO, we incorporate the data-dependent prior (Ilyas et al., 2019) into these methods by employing dimensionality reduction of the search space for comparison (which are denoted by adding a subscript "D").

The black-box attack performance of those methods is presented in Table 2. It can be seen that our P-BO achieves the highest ASR with the lowest average query count, outperforming the existing methods. Furthermore, within the P-BO algorithm, fixing the adaptive integration coefficient $\lambda = 1$ results in inferior performance compared with the baseline BO algorithm without incorporating priors. As discussed in Sec. 3.2, this phenomenon is attributed to the condition $\|f - f'\|_k > \|f\|_k$ in Theorem 3.1, which may be due to the lower similarity between models on the ImageNet dataset. This underscores the importance of adaptive tuning of the integration coefficient λ: when λ is adaptive, the P-BO algorithm exhibits higher success rates and significantly reduces queries, demonstrating stability in dealing with varying levels of prior's effectiveness. Additionally, the results also validate that the data-dependent prior is orthogonal to the proposed function prior, since integrating the data-dependent prior leads to better results.

4.4. Experimental Results on Vision-Language Models

Large Vision-Language Models (VLMs) have achieved unprecedented performance in response generation, especially with visual inputs, enabling more creative and adaptable interaction. Nonetheless, multimodal generation exacerbates safety concerns, since adversaries may successfully evade the entire system by subtly manipulating the most vulnerable modality (e.g., vision) (Dong et al., 2023; Zhao et al., 2023). In this section, we present the results of black-box

Table 3. The experimental results of black-box attacks against InstructBLIP, mPLUG-Owl, and VPGTrans under the ℓ_∞ norm on MS-COCO. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**.

Method	InstructBLIP			mPLUG-Owl			VPGTrans		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
RGF$_D$ (Cheng et al., 2019)	35.0%	315	292	46.7%	312	317	28.3%	248	113
P-RGF$_D$ (Dong et al., 2022)	54.2%	358	306	34.2%	274	91	39.2%	296	170
BO$_D$ (Ru et al., 2020)	68.3%	96	25	79.2%	58	23	45.8%	84	35
P-BO$_D$ ($\lambda = 1$, **ours**)	**95.8%**	16	**11**	83.3%	**24**	**14**	91.7%	**21**	**12**
P-BO$_D$ (λ^*, **ours**)	**95.8%**	**13**	**11**	**90.8%**	27	16	**96.7%**	24	**12**

Table 4. The experimental results of black-box attacks against defense models under the ℓ_∞ norm on CIFAR-10. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**.

| Method	Rice et al. (2020)			Zhang et al. (2019)			Rebuffi et al. (2021)		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
NES (Ilyas et al., 2018)	9.2%	324	255	9.5%	321	281	8.3%	328	255
$\mathcal{N}$ATTACK (Li et al., 2019)	18.7%	369	306	19.4%	357	306	15.7%	315	255
Square (Andriushchenko et al., 2020)	13.2%	340	185	15.0%	319	181	11.6%	269	126
RGF (Cheng et al., 2019)	9.3%	268	215	9.2%	259	212	8.4%	281	221
P-RGF (Dong et al., 2022)	23.4%	38	33	22.8%	40	31	19.1%	34	31
BO (Ru et al., 2020)	26.1%	232	129	23.7%	354	288	20.4%	180	112
P-BO ($\lambda = 1$, **ours**)	**36.2%**	35	**11**	32.2%	38	**11**	25.9%	**22**	**11**
P-BO (λ^*, **ours**)	**36.2%**	**27**	**11**	**33.9%**	**31**	**11**	**27.2%**	**22**	**11**

Clean Image	Adversarial Image
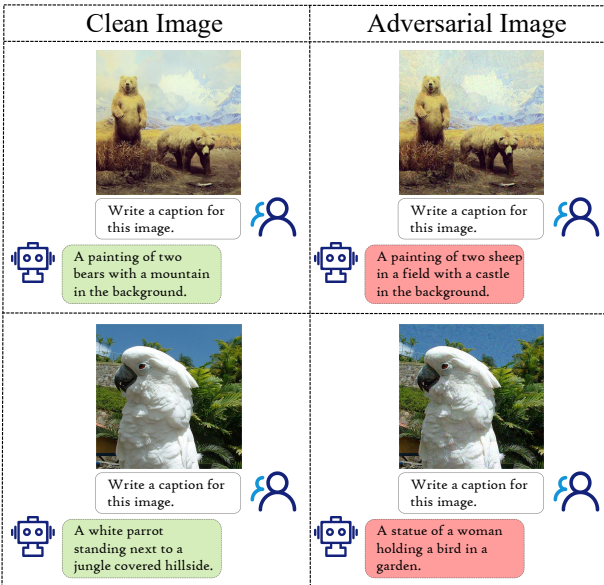
Figure 2. We show two adversarial examples against InstructBLIP. They mislead the VLM to output wrong descriptions.

adversarial attacks against VLMs. We compare the performance of P-BO with three strong baselines: RGF (Cheng et al., 2019), P-RGF (Dong et al., 2022), and BO (Ru et al., 2020). These methods adopt the dimensionality reduction technique. For all methods, we limit the maximum query count for each image to $1,000$.

Table 3 shows the attack results against InstructBLIP (Dai et al., 2023), mPLUG-Owl (Ye et al., 2023), and VPGTrans (Zhang et al., 2023). Our P-BO algorithm achieves at least 90% success rates for all target VLMs, with a low average

query count, demonstrating high query efficiency. Although P-BO ($\lambda = 1$) has a slightly lower average query count than P-BO (λ^*), there is a noticeable gap in the attack success rate. Fig. 2 shows two adversarial examples generated by P-BO, which mislead InstructBLIP to incorrectly describe the image contents. These results showcase the generalizability of P-BO in conducting black-box attacks against large VLMs, emphasizing the need for a more thorough examination of their potential security flaws before deployment.

4.5. Experimental Results on Defense Models

We include three defense models (Rice et al., 2020; Zhang et al., 2019; Rebuffi et al., 2021) on CIFAR-10 as the targets to perform black-box adversarial attacks. They are all based on adversarial training. The experimental settings are the same as those in Sec. 4.2. We adopt Zhang et al. (2019) as the surrogate model when attacking the others, while adopting Rice et al. (2020) as the surrogate model when attacking Zhang et al. (2019), since a normally trained model can be hardly useful for attacking defenses (Dong et al., 2020).

We compare the performance of P-BO with six baselines, including NES (Ilyas et al., 2018), $\mathcal{N}$ATTACK (Li et al., 2019), RGF (Cheng et al., 2019), P-RGF (Dong et al., 2022), Square attack (Andriushchenko et al., 2020), and BO (Ru et al., 2020). The attack results are presented in Table 4. Similar to the results on the normal models, the proposed P-BO method can achieve higher success rates and require much less queries than the other baselines. The experiments on adversarial defenses consistently validate the effectiveness of our proposed methods.
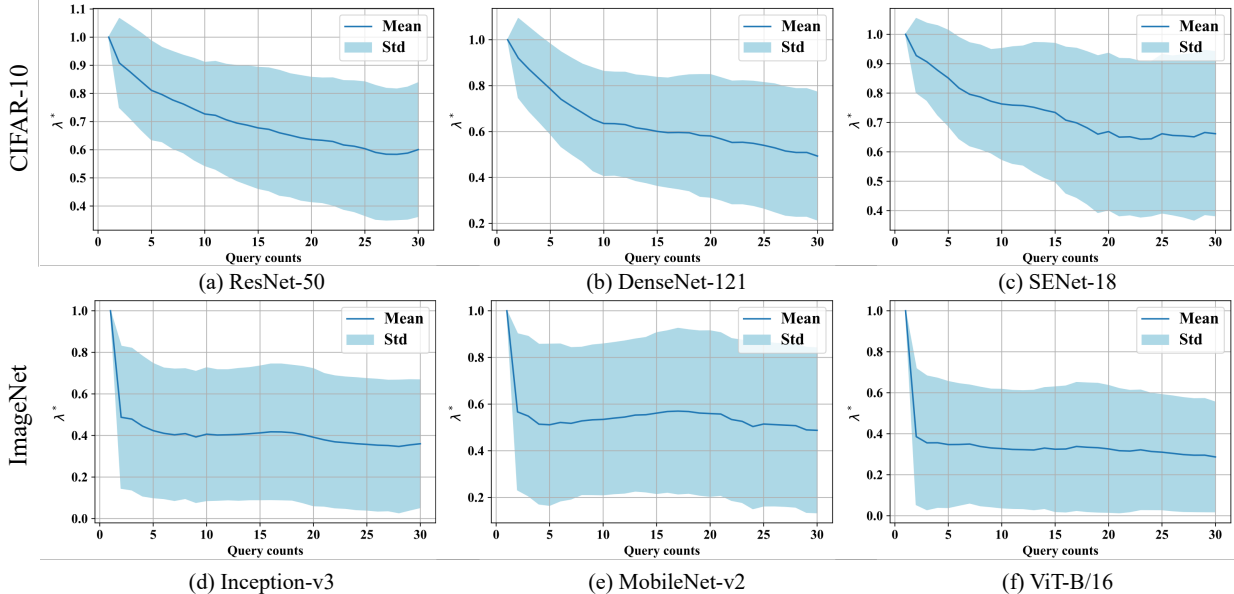
Figure 3. The mean and standard deviation of λ^* over the first 30 iterations of P-BO applied to different target models on CIFAR-10 and ImageNet. λ^* on ImageNet is substantially lower than that on CIFAR-10. This implies a lower similarity between different ImageNet models, and thus the function prior might be less useful.

4.6. Performance of Adaptive Integration Strategy

We conduct experiments to investigate the trends of λ^* across different target models on CIFAR-10 and ImageNet. Specifically, we show the mean and standard deviation of λ^* over the first 30 iterations of P-BO. Note that λ^* is initialized as 1 in the first iteration. As shown in Fig. 3, λ^* on ImageNet is significantly smaller than that on CIFAR-10, e.g., at the 5-th iteration, $\lambda^* \approx 0.8$ for various target models on CIFAR-10, whereas $\lambda^* \approx 0.4$ on ImageNet. This implies a lower similarity between different ImageNet models, and thus the function prior might be less useful. This can also explain why P-BO with $\lambda = 1$ has inferior performance than BO since λ is too large in this case. However, the adaptive integration coefficient λ^* in P-BO can be dynamically decreased to ensure that its performance remains unaffected by a useless prior, leading to consistent improvements over BO. Besides, λ^* for ViT-B/16 is lower than that for Inception-v3 and MobileNet-v2 on ImageNet. This discrepancy is caused by different architectures of ViT-B/16 (transformer) and the surrogate model (CNN).

5. Conclusion

In this paper, we propose a Prior-guided Bayesian Optimization (P-BO) method for more query-efficient black-box adversarial attacks. P-BO models the attack objective function with a Gaussian process, whose mean function is initialized by a function prior, i.e., the surrogate model's loss function. After updating the posterior distribution given the observations, the next query point is chosen by maximizing an acquisition function. We analyze the regret bound of P-BO, which is proportional to the RKHS norm between the objective function and the function prior. To avoid performance degradation in case of a bad prior, we further propose an adaptive integration strategy which automatically adjusts a coefficient on the function prior. Extensive experiments consistently demonstrate the effectiveness of P-BO in reducing the number of queries and improving attack success rates.

Impact Statement

A potential negative societal impact of our work is that malicious adversaries may adopt our method to efficiently query actual victim systems for generating adversarial samples in the real world, which can cause severe security/safety consequences for real-world applications. Thus it is imperative to develop more robust models against our attack, which we leave to future work.

Acknowledgements

References

Al-Dujaili, A., O'Reilly, U.-M., and O'Reilly, U.-M. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.

Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.-J., and Srivastava, M. B. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1111–1119, 2019.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *Proceedings of the European Conference on Computer Vision*, pp. 484–501, 2020.

Bai, Y., Wang, Y., Zeng, Y., Jiang, Y., and Xia, S.-T. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 133:109037, 2023.

Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C. J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Cheng, S., Dong, Y., Pang, T., Su, H., and Zhu, J. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, pp. 10934–10944, 2019.

Cheng, S., Wu, G., and Zhu, J. On the convergence of prior-guided zeroth-order optimization algorithms. In *Advances in Neural Information Processing Systems*, pp. 14620–14631, 2021.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 49250–49267, 2023.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Dong, Y., Pang, T., Su, H., and Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 321–331, 2020.

Dong, Y., Cheng, S., Pang, T., Su, H., and Zhu, J. Query-efficient black-box adversarial attacks guided by a transfer-based prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9536–9548, 2022.

Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., and Zhu, J. How robust is google's bard to adversarial image attacks? In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Du, J., Zhang, H., Zhou, J. T., Yang, Y., and Feng, J. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Feng, Y., Wu, B., Fan, Y., Liu, L., Li, Z., and Xia, S.-T. Boosting black-box attack with partially transferred conditional adversarial distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15095–15104, 2022.

Feurer, M., Letham, B., and Bakshy, E. Scalable meta-learning for bayesian optimization using ranking-weighted gaussian process ensembles. In *AutoML Workshop at ICML*, 2018.

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

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1487–1495, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Guo, C., Gardner, J., You, Y., Wilson, A. G., and Weinberger, K. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pp. 2484–2493, 2019a.

Guo, Y., Yan, Z., and Zhang, C. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In *Advances in Neural Information Processing Systems*, pp. 3825–3834, 2019b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*, pp. 630–645, 2016b.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pp. 918–926, 2014.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Huang, Z. and Zhang, T. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020.

Huang, Z., Huang, Y., and Zhang, T. Corrattack: Black-box adversarial attack with structured search. *arXiv preprint arXiv:2010.01250*, 2020.

Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F., and Nardi, L. πBO: Augmenting acquisition functions with user beliefs for bayesian optimization. In *International Conference of Learning Representations*, 2022.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146, 2018.

Ilyas, A., Engstrom, L., and Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019.

Jones, D. R. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21:345–383, 2001.

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

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

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Lee, D., Moon, S., Lee, J., and Song, H. O. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning*, pp. 12478–12497, 2022.

Lee, D., Lee, J., Ha, J.-W., Kim, J.-H., Lee, S.-W., Lee, H., and Song, H. O. Query-efficient black-box red teaming via bayesian optimization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 11551–11574, 2023.

Li, C., Santu, R., Gupta, S., Nguyen, V., Venkatesh, S., Sutti, A., Rubin, D., Slezak, T., Height, M., Mohammed, M., et al. Accelerating experimental design by incorporating experimenter hunches. In *IEEE International Conference on Data Mining*, pp. 257–266, 2018.

Li, C., Gupta, S., Rana, S., Nguyen, V., Robles-Kelly, A., and Venkatesh, S. Incorporating expert prior knowledge into experimental design via posterior sampling. *arXiv preprint arXiv:2002.11256*, 2020a.

Li, C., Yao, W., Wang, H., Jiang, T., and Zhang, X. Bayesian evolutionary optimization for crafting high-quality adversarial examples with limited query budget. *Applied Soft Computing*, 142:110370, 2023.

Li, J., Ji, R., Liu, H., Liu, J., Zhong, B., Deng, C., and Tian, Q. Projection & probability-driven black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 362–371, 2020b.

Li, Y., Li, L., Wang, L., Zhang, T., and Gong, B. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *International Conference on Machine Learning*, pp. 3866–3876, 2019.

Liang, S., Wu, B., Fan, Y., Wei, X., and Cao, X. Parallel rectangle flip attack: A query-based black-box attack against object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7677–7687, 2021.

Lin, J., Song, C., He, K., Wang, L., and Hopcroft, J. E. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Lord, N. A., Mueller, R., and Bertinetto, L. Attacking deep networks with surrogate-based adversarial black-box methods is easy. In *International Conference on Learning Representations*, 2022.

Ma, C., Cheng, S., Chen, L., Zhu, J., and Yong, J. Switching transferable gradient directions for query-efficient black-box adversarial attacks. *arXiv preprint arXiv:2009.07191*, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Miao, Y., Dong, Y., Zhu, J., and Gao, X.-S. Isometric 3d adversarial examples in the physical world. In *Advances in Neural Information Processing Systems*, pp. 19716–19731, 2022.

Močkus, J. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404, 1975.

Mohaghegh Dolatabadi, H., Erfani, S., and Leckie, C. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. In *Advances in Neural Information Processing Systems*, pp. 15871–15884, 2020.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Nitin Bhagoji, A., He, W., Li, B., and Song, D. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European Conference on Computer Vision*, pp. 154–169, 2018.

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, 2017.

Poloczek, M., Wang, J., and Frazier, P. Multi-information source optimization. In *Advances in Neural Information Processing Systems*, pp. 4291–4301, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ramachandran, A., Gupta, S., Rana, S., Li, C., and Venkatesh, S. Incorporating expert prior in bayesian optimisation via space warping. *Knowledge-Based Systems*, 195:105663, 2020.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71, 2003.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104, 2020.

Ru, B., Cobb, A., Blaas, A., and Gal, Y. Bayesopt adversarial attack. In *International Conference on Learning Representations*, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Shen, Y., Li, Y., Zheng, J., Zhang, W., Yao, P., Li, J., Yang, S., Liu, J., and Cui, B. Proxybo: Accelerating neural architecture search via bayesian optimization with zero-cost proxies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9792–9801, 2023.

Shukla, S. N., Sahu, A. K., Willmott, D., and Kolter, J. Z. Black-box adversarial attacks with bayesian optimization. *arXiv preprint arXiv:1909.13857*, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Snoek, J., Swersky, K., Zemel, R., and Adams, R. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pp. 1674–1682, 2014.

Souza, A., Nardi, L., Oliveira, L. B., Olukotun, K., Lindauer, M., and Hutter, F. Bayesian optimization with a prior for the optimum. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 265–296, 2021.

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

Suya, F., Tian, Y., Evans, D., and Papotti, P. Query-limited black-box attacks to classifiers. *arXiv preprint arXiv:1712.08713*, 2017.

Suya, F., Chi, J., Evans, D., and Tian, Y. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. In *29th USENIX Security Symposium*, pp. 1327–1344, 2020.

Swersky, K., Snoek, J., and Adams, R. P. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, pp. 2004–2012, 2013.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114, 2019.

Tighineanu, P., Skubch, K., Baireuther, P., Reiss, A., Berkenkamp, F., and Vinogradska, J. Transfer learning with gaussian processes for bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 6152–6181, 2022.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 742–749, 2019.

Wan, X., Kenlay, H., Ru, R., Blaas, A., Osborne, M. A., and Dong, X. Adversarial attacks on graph classifiers via bayesian optimisation. In *Advances in Neural Information Processing Systems*, pp. 6983–6996, 2021.

Wistuba, M., Schilling, N., and Schmidt-Thieme, L. Scalable gaussian process-based transfer surrogates for hyperparameter optimization. *Machine Learning*, 107(1): 43–78, 2018.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., and Yuille, A. L. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Yang, J., Jiang, Y., Huang, X., Ni, B., and Zhao, C. Learning black-box attackers with transferable priors and query feedback. In *Advances in Neural Information Processing Systems*, pp. 12288–12299, 2020.

Yatsura, M., Metzen, J., and Hein, M. Meta-learning the search distribution of black-box random search based adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 30181–30195, 2021.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yin, F., Zhang, Y., Wu, B., Feng, Y., Zhang, J., Fan, Y., and Yang, Y. Generalizable black-box adversarial attack with meta learning. *IEEE transactions on pattern analysis and machine intelligence*, 46(3):1804–1818, 2023.

Yogatama, D. and Mann, G. Efficient transfer learning method for automatic hyperparameter tuning. In *Artificial Intelligence and Statistics*, pp. 1077–1085, 2014.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., and Chua, T.-S. Vpgtrans: Transfer visual prompt generator across llms. In *Advances in Neural Information Processing Systems*, pp. 20299–20319, 2023.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

Zhao, P., Liu, S., Chen, P., Hoang, N., Xu, K., Kailkhura, B., and Lin, X. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. In *IEEE/CVF International Conference on Computer Vision*, pp. 121–130, 2019.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, pp. 54111–54138, 2023.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

A. Proofs

A.1. Proof of Theorem 3.1

Proof. We provide a proof based on that of Srinivas et al. (2010). Let $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in A} f(\boldsymbol{x})$ be the global maximum, and $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x} \in A} \mu_{t-1}(\boldsymbol{x}) + \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x})$ be the next query point. We first prove that for $\forall \boldsymbol{x} \in A, t \leq T$, we have

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}). \tag{13}$$

We first consider the case of zero-mean Gaussian process, i.e. $f' \equiv 0$. Due to the reproducing property of RKHS, we have $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \sigma_{t-1}(\boldsymbol{x})\|f - \mu_{t-1}\|_k$. Then we prove that $\|f - \mu_{t-1}\|_k \leq \|f\|_k$. Let $(a_1, \ldots, a_{t-1})^\top = \mathbf{K}_{t-1}^{-1} \mathbf{y}_{t-1}$. Then $\mu_{t-1} = \sum_{i=1}^{t-1} a_i k(\boldsymbol{x}_i, \cdot)$. Thus, $\langle f, \mu_{t-1} \rangle_k = \langle \mu_{t-1}, \mu_{t-1} \rangle_k = \mathbf{y}_{t-1}^\top \mathbf{K}_{t-1}^{-1} \mathbf{K}_{t-1} \mathbf{K}_{t-1}^{-1} \mathbf{y}_{t-1} = \mathbf{y}_{t-1}^\top \mathbf{K}_{t-1}^{-1} \mathbf{y}_{t-1}$. So, we have

$$\|f - \mu_{t-1}\|_k^2 = \|f\|_k^2 - 2\langle f, \mu_{t-1} \rangle_k + \|\mu_{t-1}\|_k^2 = \|f\|_k^2 - \mathbf{y}_{t-1}^\top \mathbf{K}_{t-1}^{-1} \mathbf{y}_{t-1}. \tag{14}$$

Thus $\|f - \mu_{t-1}\|_k \leq \|f\|_k$, and $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \|f\|_k \sigma_{t-1}(\boldsymbol{x})$.

Next, we consider the general case, where $f' \neq 0$. Let $\mu'_{t-1}(\boldsymbol{x}) := \mu_{t-1}(\boldsymbol{x}) - f'(\boldsymbol{x})$ and $\sigma'_{t-1}(\boldsymbol{x}) := \sigma_{t-1}(\boldsymbol{x})$. By Eq. (4)-(5) and Eq. (7)-(8), we can see that $\mu'_{t-1}(\boldsymbol{x})$ and $\sigma'_{t-1}(\boldsymbol{x})$ is the posterior mean and posterior standard deviation of $f - f'$ under a zero-mean Gaussian process. Using the result in zero-mean case, we have

$$|(f - f')(\boldsymbol{x}) - \mu'_{t-1}(\boldsymbol{x})| \leq \|f - f'\|_k \sigma'_{t-1}(\boldsymbol{x}). \tag{15}$$

Therefore, substituting $\mu'_{t-1}(\boldsymbol{x})$ and $\sigma'_{t-1}(\boldsymbol{x})$, we obtain $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x})$.

Then we prove that the instantaneous regret $r_t \leq 2\|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}_t)$. By the UCB acquisition function defined in Eq. (6) and definition of $\boldsymbol{x}_t$, we have

$$\mu_{t-1}(\boldsymbol{x}_t) + \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}_t) \geq \mu_{t-1}(\boldsymbol{x}^*) + \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}^*). \tag{16}$$

By plugging in Eq. (13) at $\boldsymbol{x}^*$:

$$|f(\boldsymbol{x}^*) - \mu_{t-1}(\boldsymbol{x}^*)| \leq \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}^*), \tag{17}$$

we have that $\mu_{t-1}(\boldsymbol{x}^*) + \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^*)$. So the instantaneous regret has

$$r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \mu_{t-1}(\boldsymbol{x}_t) + \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t). \tag{18}$$

Through a recursive application of Eq. (13) at $\boldsymbol{x}_t$:

$$|f(\boldsymbol{x}_t) - \mu_{t-1}(\boldsymbol{x}_t)| \leq \|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}_t), \tag{19}$$

we arrive at the revelation that

$$r_t \leq 2\|f - f'\|_k \sigma_{t-1}(\boldsymbol{x}_t). \tag{20}$$

In the third step, we prove that

$$\sum_{t=1}^{T} \sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \frac{2}{\log(1 + \sigma^{-2})} \gamma_T, \tag{21}$$

where $\gamma_T = \frac{1}{2} \max_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in A} \log |\mathbf{I} + \sigma^{-2} \mathbf{K}_T|$.

We have $\sigma_{t-1}(\boldsymbol{x}_t)^2 = \sigma^2 (\sigma^{-2} \sigma_{t-1}(\boldsymbol{x}_t)^2)$. Since $x^2 \leq C \log(1 + x^2)$ for $x \in [0, \sigma^{-2}]$, $C \geq 1$, and $\sigma^{-2} \sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \sigma^{-2} k(\boldsymbol{x}_t, \boldsymbol{x}_t) \leq \sigma^{-2}$, $\frac{\sigma^{-2}}{\log(1+\sigma^{-2})} \geq 1$, we have

$$\sigma^{-2} \sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \frac{\sigma^{-2}}{\log(1 + \sigma^{-2})} \log(1 + \sigma^{-2} \sigma_{t-1}(\boldsymbol{x}_t)^2). \tag{22}$$

15

Thus, $\sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \frac{1}{\log(1+\sigma^{-2})} \log(1 + \sigma^{-2}\sigma_{t-1}(\boldsymbol{x}_t)^2)$. Then we can derive that

$$\sum_{t=1}^{T} \sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \frac{2}{\log(1+\sigma^{-2})} \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}(\boldsymbol{x}_t)^2). \tag{23}$$

On the other hand, we also have

$$\frac{1}{2} \log |\mathbf{I} + \sigma^{-2}\mathbf{K}_T| = \mathrm{H}(\mathcal{N}(\mathbf{y}'_T, \mathbf{K}_T)) - \mathrm{H}(\mathcal{N}(\mathbf{y}_T, \sigma^2 \mathbf{I})), \tag{24}$$

where $\mathrm{H}(\cdot)$ represents the entropy of a distribution. Hence,

$$\frac{1}{2} \log |\mathbf{I} + \sigma^{-2}\mathbf{K}_T| = \mathrm{H}(\mathcal{N}(\mathbf{y}'_T, \mathbf{K}_T)) - \mathrm{H}(\mathcal{N}(\mathbf{y}_T, \sigma^2 \mathbf{I})) \tag{25}$$

$$= \mathrm{H}(\mathcal{N}(\mathbf{y}'_T, \mathbf{K}_T)) - \frac{1}{2} \log |2\pi e \sigma^2 \mathbf{I}| \tag{26}$$

$$= \mathrm{H}(\mathcal{N}(\mathbf{y}'_{T-1}, \mathbf{K}_{T-1})) + \mathrm{H}(\mathcal{N}(\mathbf{y}'_T | \mathbf{y}'_{T-1})) - \frac{1}{2} \log |2\pi e \sigma^2 \mathbf{I}| \tag{27}$$

$$= \mathrm{H}(\mathcal{N}(\mathbf{y}'_{T-1}, \mathbf{K}_{T-1})) + \frac{1}{2} \log \left(2\pi e(\sigma^2 + \sigma_{T-1}(x_T)^2)\right) - \frac{1}{2} \log |2\pi e \sigma^2 \mathbf{I}| . \tag{28}$$

By means of induction, we obtain

$$\frac{1}{2} \log |\mathbf{I} + \sigma^{-2}\mathbf{K}_T| = \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}(\boldsymbol{x}_t)^2). \tag{29}$$

Substituting Eq. (29) into Eq. (23), we will obtain the proof that

$$\sum_{t=1}^{T} \sigma_{t-1}(\boldsymbol{x}_t)^2 \leq \frac{2}{\log(1+\sigma^{-2})} \gamma_T. \tag{30}$$

Since $R_T = \sum_{t=1}^{T} r_t$, using Eq. (20), Eq. (30) and Cauchy Schwarz inequality, we can derive that

$$R_T \leq \sqrt{T \sum_{t=1}^{T} r_t} \tag{31}$$

$$\leq 2\|f - f'\|_k \sqrt{T \sum_{t=1}^{T} \sigma_{t-1}(\boldsymbol{x}_t)^2} \tag{32}$$

$$\leq \|f - f'\|_k \sqrt{\frac{8}{\log(1+\sigma^{-2})} T \gamma_T}. \tag{33}$$

$\square$

A.2. Proof of Proposition 3.3

Proof. We provide a proof based on that of Kanagawa et al. (2018). Let

$$\mathcal{F}_k := \left\{ \sum_{t=1}^{T} \beta_t k(\boldsymbol{x}_t, \cdot) | \forall \beta_1, \dots, \beta_T \in \mathbb{R} \right\}, \tag{34}$$

$$\mathcal{H}_k := \{h \in \mathcal{H} | h(\boldsymbol{x}_t) = y_t, \forall t \leq T\}. \tag{35}$$

We first prove that $\{\mu_T\} = \mathcal{F}_k \cap \mathcal{H}_k$. Notice that $\mu_T = \sum_{t=1}^T a_t k(\boldsymbol{x}_t, \cdot)$, where $(a_1, \ldots, a_T)^\top = \mathbf{K}_T^{-1} \mathbf{y}_T$, therefore, $\mu_T \in \mathcal{F}_k$. Also, $\mu_T(\boldsymbol{x}_t) = y_t$, then $\mu_T \in \mathcal{H}_k$. Hence, $\mu_T \in \mathcal{F}_k \cap \mathcal{H}_k$.

Next, we will prove that if $f \in \mathcal{F}_k \cap \mathcal{H}_k$, then $f = \mu_T$. Assume that $f \in \mathcal{F}_k \cap \mathcal{H}_k$. let $f = \sum_{t=1}^T \gamma_t k(\boldsymbol{x}_t, \cdot)$, $\gamma_t \in \mathbb{R}$. Then

$$\mu_T - f = \sum_{t=1}^T (a_t - \gamma_t) k(\boldsymbol{x}_t, \cdot) \in \mathcal{F}_k. \tag{36}$$

Also, since $f \in \mathcal{H}_k$, then $f(\boldsymbol{x}_t) = y_t$, we have

$$\langle \mu_T - f, k(\boldsymbol{x}_t, \cdot) \rangle_k = \mu_T(\boldsymbol{x}_t) - f(\boldsymbol{x}_t) = y_t - y_t = 0. \tag{37}$$

so $\mu_T - f \perp \mathcal{F}_k$, which implies $\mu_T - f \in \mathcal{F}_k \cap \mathcal{F}_k^\perp = \{0\}$. Thus, $\{\mu_T\} = \mathcal{F}_k \cap \mathcal{H}_k$.

Then we prove that $\mu_T = \arg\min_{h \in \mathcal{H}_k} \|h\|_k$. Since $\mathcal{H}_k$ is convex and closed, there exists an solution $h^* \in \mathcal{H}_k$ such that $h^* = \arg\min_{h \in \mathcal{H}_k} \|h\|_k$. For $\forall g \perp \mathcal{F}_k$, we have

$$\langle h^* + g, k(\boldsymbol{x}_t, \cdot) \rangle_k = \langle h^*, k(\boldsymbol{x}_t, \cdot) \rangle_k = h^*(\boldsymbol{x}_t) = y_t. \tag{38}$$

So $h^* + g \in \mathcal{H}_k$. Since $\|h^*\|_k \leq \|h^* + g\|_k$ and $\mathcal{F}_k$ is closed, we have $h^* \in (\mathcal{F}_k^\perp)^\perp = \mathcal{F}_k$. So $h^* \in \mathcal{F}_k \cap \mathcal{H}_k$, and $\mathcal{F}_k \cap \mathcal{H}_k = \{\mu_T\}$, thus, $h^* = \mu_T$. Since $\langle k(\boldsymbol{x}_{t_i}, \cdot), k(\boldsymbol{x}_{t_j}, \cdot) \rangle_k = k(\boldsymbol{x}_{t_i}, \boldsymbol{x}_{t_j})$ and $\mu_T = \sum_{t=1}^T a_t k(\boldsymbol{x}_t, \cdot)$, we have

$$\|\mu_T\|_k^2 = \langle \mu_T, \mu_T \rangle_k = \mathbf{y}_T^\top \mathbf{K}_T^{-1} \mathbf{K}_T \mathbf{K}_T^{-1} \mathbf{y}_T = \mathbf{y}_T^\top \mathbf{K}_T^{-1} \mathbf{y}_T. \tag{39}$$

$\square$

B. Related Work

Query-based attack methods. Query-based methods generate adversarial examples by leveraging the query feedback of the black-box model. ZOO (Chen et al., 2017) uses symmetric difference to estimate gradient for every pixel. NES (Ilyas et al., 2018) utilizes natural evolution strategy to estimate gradients. Bandits (Ilyas et al., 2019) improves the NES method by incorporating data and temporal priors into the gradient estimation. $\mathcal{N}$ATTACK (Li et al., 2019) introduces a gradient estimation framework to improve the attack success over defensive models. RGF (Cheng et al., 2019) utilizes random gradient-free method to estimate gradients. SimBA (Guo et al., 2019a) adapts a greedy strategy to update the query samples. Square Attack (Andriushchenko et al., 2020) introduces highly-efficient greedy random search for black-box adversarial attack. SignHunter (Al-Dujaili et al., 2020) adapts the gradient sign rather than the gradient as the search direction. NPAttack (Bai et al., 2023) explore the distribution of adversarial examples around benign inputs with the help of image structure information characterized by a Neural Process. BayesOpt (Ru et al., 2020) and similar methods (Huang et al., 2020; Li et al., 2023) utilize Bayesian optimization for black-box attacks. AdvFlow (Mohaghegh Dolatabadi et al., 2020) approximates the adversarial distribution with the clean data distribution PPBA (Li et al., 2020b) shrinks the solution space of possible adversarial inputs to those which contain low- frequency perturbations. PRFA (Liang et al., 2021) proposes to parallelly attack multiple rectangles for better efficiency. Additionally, several studies apply BO-based black-box attacks to various specific application scenarios, including graph classification, natural language processing, and protein classification. Wan et al. (2021) introduce a novel BO-based attack method tailored for graph classification models. They provide valuable insights into the relationship between changes in graph topology and model robustness. Lee et al. (2022) put forward a query-efficient black-box attack leveraging BO, which dynamically determines crucial positions using an automatic relevance determination (ARD) categorical kernel. In a subsequent work, Lee et al. (2023) propose innovative query-efficient black-box red teaming methods based on BO, specifically targeting large-scale generative models. The main limitation of these methods is that they inevitably require a tremendous number of queries to perform a successful attack, leading to a low attack success rate given a limited query budget.

Combination-based attack methods. Combination-based methods (Cheng et al., 2019; Guo et al., 2019b; Du et al., 2020; Yang et al., 2020; Ma et al., 2020; Huang & Zhang, 2020; Suya et al., 2020; Yatsura et al., 2021; Lord et al., 2022; Feng et al., 2022; Dong et al., 2022; Yin et al., 2023) integrate transfer-based attacks with query-based attacks to achieve high attack success rate and high query efficiency simultaneously. Cheng et al. (2019); Dong et al. (2022) propose P-RGF, utilizing the gradient of the surrogate model as a transfer-based prior to obtain a more accurate gradient estimate. Guo et al. (2019b);

Ma et al. (2020); Yang et al. (2020) restrict the search space spanned by the surrogate gradients. Subspace attack (Guo et al., 2019b) regards surrogates' transfer-prior as subspaces to reduce search space of random vectors with gradients. LeBA (Guo et al., 2019b) proposes to learn the victim's estimated gradients via high-order computation graph. Although these methods are effective in expediting convergence and reducing the number of queries, the surrogate gradient is localized and can be misleading, limiting their effectiveness. Besides, other methods learn a generalizable model-based prior based on the surrogate model (Du et al., 2020; Huang & Zhang, 2020; Yin et al., 2023). Meta attack (Du et al., 2020) adopts meta-learning to approximate the victim. TREMBA (Huang & Zhang, 2020) treats the projection from a low-dimensional space to the original space as the prior, such that the perturbation could be search in the low-dimensional space. MCG (Yin et al., 2023) trains a meta generator to produce perturbations conditioned on benign examples. But these methods require an additional dataset to train the attack generator, which is not applicable when the data is scarce.

Bayesian optimization with prior. In Bayesian optimization, the incorporation of prior information can be broadly classified into three types (Hvarfner et al., 2022). The first type of prior information pertains to the distribution of the optimal solution's location. Snoek et al. (2014); Ramachandran et al. (2020) enhance exploration in local regions and suppress exploration in unimportant areas by warping the input space. In contrast, Hvarfner et al. (2022); Li et al. (2020a); Souza et al. (2021) introduce a prior distribution on the location of the optimal solution and compute the posterior distribution of the optimal solution's location given observed data. Such prior information is typically proposed by experienced practitioners or domain experts, and its form is often straightforward. The second type of prior information concerns the structure of the objective function. For instance, Li et al. (2018) utilize the monotonic trend of the independent variables to model. These methods necessitate a concrete understanding or assumption about the nature of the objective function. The third type of prior information is derived from datasets obtained from similar unknown objective functions, and the methods utilizing such prior information are referred to as transfer learning in Bayesian optimization. There are generally two approaches (Tighineanu et al., 2022; Shen et al., 2023): the first involves jointly modeling datasets corresponding to both prior information and the data obtained from the target function (Swersky et al., 2013; Yogatama & Mann, 2014; Poloczek et al., 2017), while the second involves separately modeling the dataset corresponding to the prior information to aid in modeling the target function (Golovin et al., 2017; Feurer et al., 2018; Wistuba et al., 2018). The latter approach often involves fitting a Gaussian process on the prior dataset, treating the predictive distribution of the Gaussian process as a prior with uncertainty. Specifically, Feurer et al. (2018); Wistuba et al. (2018) weightedly average the predicted mean and variance of the Gaussian process fitted on the prior dataset with those fitted on the objective function dataset, with heuristic weights lacking theoretical analysis. Golovin et al. (2017) employ the difference between the fitted Gaussian process predicted mean on the target function dataset and that on the prior dataset for fusion but do not incorporate adaptive weights. According to our analysis of the regret upper bound, non-adaptive fusion might degrade algorithm performance. Additionally, this method's setting for predictive variance is not suitable for the deterministic function prior described in our black-box attack scenario. None of the aforementioned algorithms analyze the regret upper bound of Bayesian optimization algorithms, and their settings differ from the scenario where a deterministic function prior is directly provided. Moreover, these methods have not been applied to black-box attacks.

C. A Case of Random Linear Function

In light of Theorem 3.1, it is evident that although convergence is guaranteed when modeling with $\text{GP}(f', k)$, achieving improved algorithmic performance is desirable to have a small value for $\|f - f'\|_k$, at least satisfying $\|f - f'\|_k \leq \|f\|_k$. While this naturally holds when f' closely approximates f, in the case of functions defined in high-dimensional spaces, this condition is quite stringent. As follows, we provide a natural counterexample, demonstrating the challenge in meeting this condition, when the functions are defined in high-dimensional spaces, exhibiting a tendency towards orthogonality. Let f and f' both be linear functions, where $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, $f'(\boldsymbol{x}) = \boldsymbol{w'}^\top \boldsymbol{x}$. Consequently, $(f - f')(\boldsymbol{x}) = (\boldsymbol{w} - \boldsymbol{w'})^\top \boldsymbol{x}$. Assuming the kernel function k is isotropic and stationary, i.e., $k(\boldsymbol{x}, \boldsymbol{x'})$ depends only on $\|\boldsymbol{x} - \boldsymbol{x'}\|_2$, in accordance with the absolute homogeneity of norms, $\|f - f'\|_k \leq \|f\|_k$ is equivalent to $\|\boldsymbol{w} - \boldsymbol{w'}\|_2 \leq \|\boldsymbol{w}\|_2$. Assuming $\|\boldsymbol{w}\|_2 = \|\boldsymbol{w'}\|_2 = 1$, this requires $\boldsymbol{w}^\top \boldsymbol{w'} \geq \frac{1}{2}$. Considering that if $\boldsymbol{w}, \boldsymbol{w'}$ are uniformly sampled from the unit hypersphere in $\mathbb{R}^d$, the expected value $\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{w'})^2] = \frac{1}{d}$. It becomes evident that the probability of $\boldsymbol{w}^\top \boldsymbol{w'} \geq \frac{1}{2}$ is indeed small. This implies that when the target and prior functions are not closely aligned, a direct $\text{GP}(f', k)$ modeling approach may lead to $\|f - f'\|_k > \|f\|_k$. Such a scenario could result in the prior-guided Bayesian optimization algorithm's performance degradation compared to the approach in which prior information is not incorporated.

Table 5. The experimental results of black-box targeted attacks against DenseNet-121, ResNet-50, and SENet-18 under the ℓ_∞ norm on CIFAR-10. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**.

Methods	ResNet-50			DenseNet-121			SENet-18		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
RGF (Cheng et al., 2019)	47.6%	389	325	39.7%	392	322	44.1%	363	316
P-RGF (Dong et al., 2022)	72.2%	184	70	60.6%	219	102	71.8%	167	60
BO (Ru et al., 2020)	82.7%	223	140	74.1%	244	155	81.4%	225	135
P-BO$_\text{D}$ ($\lambda = 1$, **ours**)	96.9%	42	**21**	91.2%	97	**26**	96.8%	45	**16**
P-BO$_\text{D}$ (λ^*, **ours**)	**98.7%**	**34**	**21**	**97.1%**	**55**	28	**99.2%**	**33**	21

D. Supplementary Experimental Results

D.1. More Experimental Details

We conduct the experiments on NVIDIA 2080 Ti (for CIFAR-10 and ImageNet) and A100 (for VLMs) GPUs. The source code of P-BO is submitted in the supplementary material and will be released after the review process. For the baseline attacks, we adopt the official implementations of NES, Bandits, $\mathcal{N}$ATTACK, SignHunter, Square attack, and NPAttack. For RGF and P-RGF, we adopt a better implementation: we make the random directions $\{u_i\}_{i=1}^q$ orthogonal following Cheng et al. (2021); we set $q = 5$, $lr = 0.1$, $\sigma = 0.05$ on CIFAR-10 and $\sigma = 0.5$ on ImageNet and MS-COCO.[2] We adopt the gradient averaging method in P-RGF due to its better performance (Dong et al., 2022). We implement highly efficient BO and P-BO algorithms in PyTorch by referring to the implementation of Ru et al. (2020) based on SciPy (and NumPy). One main difference is that when optimizing the acquisition function, we use the PGD algorithm with learning rate 0.05 and 50 iterations of optimization instead of the L-BFGS method, and we can try multiple random starts in parallel. We initialize the length scale in the Matern-5/2 kernel as $\sqrt{D}$.[3] The balancing coefficient β is set to 3. Our implementation of BO has similar performance to Ru et al. (2020).

D.2. Targeted Attacks on CIFAR-10

In this section, we perform black-box targeted adversarial attacks against three CIFAR-10 models, including ResNet-50 (He et al., 2016a), DenseNet-121 (Huang et al., 2017), and SENet-18 (Hu et al., 2018). We select 100 correctly classified images and target at all other 9 classes, leading to 900 trials. We compare the performance of P-BO with three strong baselines — RGF (Cheng et al., 2019), P-RGF (Dong et al., 2022), and BO (Ru et al., 2020). For all methods, we restrict the maximum number of queries for each image to be 1,000, and the experimental settings are aligned with those outlined in Sec. 4.2.

Table 5 shows the results, where we report the success rate of black-box targeted attacks and the average/median number of queries needed to generate an adversarial example over successful attacks. Compared with the state-of-the-art attacks, the proposed method P-BO generally leads to higher attack success rates and requires much fewer queries. The P-BO algorithm demonstrates a notable improvement over conventional BO algorithms, attaining remarkably high success rates with only a few dozen average query counts, showcasing the effectiveness and practicality of P-BO. The adaptive coefficient λ^* in P-BO, in comparison to a fixed coefficient $\lambda = 1$, significantly enhances the attack success rate and reduces the average query count, underscoring the efficacy of utilizing adaptive fusion weights. Additionally, the improvement of P-BO over the baseline BO algorithm is more pronounced than the enhancement observed in P-RGF over RGF, indicating that the P-BO algorithm efficiently utilizes useful function prior information to a considerable extent.

D.3. Experimental Results under the ℓ_2 Norm on CIFAR-10

We further conduct experiments under the ℓ_2 norm on CIFAR-10. The experimental settings are the same as those in Sec. Sec. 4.2 except that we set the perturbation size as $\epsilon = \sqrt{0.001 \cdot D}$ following Dong et al. (2022) where D is the input dimension. We compare P-BO (λ^*) and P-BO ($\lambda = 1$) with six strong baselines: NES (Ilyas et al., 2018), $\mathcal{N}$ATTACK (Li et al., 2019), RGF (Cheng et al., 2019), P-RGF (Dong et al., 2022), Square attack (Andriushchenko et al., 2020), and BO

---

[2]In our implementation of RGF, P-RGF, BO and P-BO, instead of directly optimizing w.r.t. x, we reparametrize x with $x = x^{nat} + \epsilon \cdot \delta$ and optimize w.r.t. δ where $\delta \in [-1, 1]^D$. Therefore, the mentioned σ and learning rate lr in RGF and P-RGF are under the context of the search space $[-1, 1]^D$.

[3]The mentioned learning rate and length scale in BO and P-BO are under the context of the search space $[-1, 1]^D$.

Table 6. The experimental results of black-box attacks against DenseNet-121, ResNet-50, and SENet-18 under the ℓ_2 norm on CIFAR-10. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**.

Method	ResNet-50			DenseNet-121			SENet-18		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
NES (Ilyas et al., 2018)	96.5%	335	306	96.8%	321	255	96.9%	307	255
Bandits$_T$ (Ilyas et al., 2019)	97.3%	190	116	98.1%	165	96	98.3%	142	86
$\mathcal{N}$ATTACK (Li et al., 2019)	99.2%	249	204	98.4%	246	204	99.9%	221	203
Square (Andriushchenko et al., 2020)	98.6%	117	56	98.1%	123	62	98.7%	103	49
RGF (Cheng et al., 2019)	99.7%	137	89	98.9%	122	65	99.3%	115	71
P-RGF (Dong et al., 2022)	99.9%	30	21	99.9%	28	21	99.9%	27	21
BO (Ru et al., 2020)	99.9%	98	62	99.9%	80	46	99.5%	72	42
P-BO ($\lambda = 1$, **ours**)	**100.0%**	13	**11**	**100.0%**	12	**11**	**100.0%**	13	**11**
P-BO (λ^*, **ours**)	**100.0%**	**12**	**11**	**100.0%**	**12**	**11**	**100.0%**	**12**	**11**

Table 7. The experimental results of black-box attacks under the ℓ_∞ norm on CIFAR-10 using different surrogate models. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks.

Surrogate Model	Method	ResNet-50			DenseNet-121			SENet-18		
		ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
-	BO (Ru et al., 2020)	99.6%	83	44	99.7%	93	51	99.7%	81	41
WRN-34-10	P-BO ($\lambda = 1$)	99.9%	16	11	99.7%	25	11	99.9%	14	11
	P-BO (λ^*)	100.0%	15	11	100.0%	19	11	100.0%	14	11
VGG-16	P-BO ($\lambda = 1$)	99.4%	25	11	99.9%	23	11	99.6%	19	11
	P-BO (λ^*)	100.0%	20	11	100.0%	17	11	100.0%	16	11
EfficientNet-B0	P-BO ($\lambda = 1$)	99.5%	21	11	99.6%	19	11	99.7%	17	11
	P-BO (λ^*)	100.0%	19	11	100.0%	16	11	100.0%	17	11
TRADES	P-BO ($\lambda = 1$)	99.7%	84	36	99.7%	73	33	99.0%	66	31
	P-BO (λ^*)	99.9%	66	34	99.8%	59	32	99.9%	62	32

(Ru et al., 2020). We report the attack success rate and the average/median number of queries in Table 6. It can be seen that our P-BO generally leads to higher attack success rates and requires much fewer queries under the ℓ_2 norm. This implies that our P-BO method is universal for both ℓ_∞ and ℓ_2 norms.

D.4. Experimental Results on Different Surrogate Models

We further conduct experiments under the ℓ_∞ norm on CIFAR-10 and ImageNet using different surrogate models. For CIFAR-10, We adopt VGG-16 (Simonyan & Zisserman, 2014), EfficientNet-B0 (Tan & Le, 2019), and TRADES (Zhang et al., 2019) as the surrogate models. The other experimental settings are the same as those in Sec. 4.2. We report the results of P-BO (λ^*) and P-BO ($\lambda = 1$) in Table 7. It is worth noting that adopting TRADES as the alternative surrogate model is not a preferable choice. In this scenario, although our P-BO (λ^*) can adaptively optimize λ^* to achieve performance surpassing the BO baseline, it still significantly lags behind the efficiency of standard trained surrogate models. This is primarily due to the substantial disparity in loss landscapes between adversarially trained models and standard trained models, resulting in low similarity between them. For ImageNet, We use ResNeXt-50 (Xie et al., 2017), ResNet-50 (He et al., 2016a), and Swin Transformer (Liu et al., 2021) as the surrogate models. The other experimental settings are the same as those in Sec. 4.3. We report the results P-BO (λ^*) and P-BO ($\lambda = 1$) in Table 8. Compared with the results in Table 1 and Table 2 in the paper, it can be seen that our P-BO generally leads to higher attack success rates and requires much fewer queries using different surrogate models. This indicates that our P-BO is not sensitive to the selection of surrogate models, enabling consistent improvements in both the attack success rate and query efficiency across different surrogate models.

D.5. Experimental Results Compared with π-BO

π-BO (Hvarfner et al., 2022) incorporates prior beliefs about the location of the optimum in the form of a probability distribution on the acquisition function. We extend our experiments under the ℓ_∞ norm on CIFAR-10 using π-BO. We employ a Gaussian distribution as the probability distribution, with the mean derived from the optimal point obtained through PGD attack on the surrogate model and a constant variance that encapsulates prior beliefs regarding the optimal location.

Table 8. The experimental results of black-box attacks under the ℓ_∞ norm on ImageNet using different surrogate models. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. The subscript "D" denotes the methods with dimensionality reduction.

Surrogate Model	Method	Inception-v3			MobileNet-v2			ViT-B/16		
		ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
-	BO$_D$ (Ru et al., 2020)	94.1%	104	58	98.2%	102	61	86.7%	170	116
ResNet-152	P-BO$_D$ ($\lambda = 1$)	85.4%	182	90	86.8%	193	56	67.5%	236	240
	P-BO$_D$ (λ^*)	94.4%	81	45	98.8%	94	60	88.2%	148	81
ResNeXt-50	P-BO$_D$ ($\lambda = 1$)	81.8%	265	248	87.5%	231	57	64.4%	264	235
	P-BO$_D$ (λ^*)	96.4%	92	50	98.3%	84	51	89.5%	157	91
ResNet-50	P-BO$_D$ ($\lambda = 1$)	72.9%	264	270	79.4%	206	31	58.7%	304	294
	P-BO$_D$ (λ^*)	95.3%	91	51	98.5%	91	60	89.3%	165	87
Swin Transformer	P-BO$_D$ ($\lambda = 1$)	84.5%	309	242	89.8%	160	67	76.4%	269	137
	P-BO$_D$ (λ^*)	97.6%	90	59	99.4%	90	49	90.4%	152	64

Table 9. The experimental results of black-box attacks compared with π-BO under the ℓ_∞ norm on CIFAR-10. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks.

Method	ResNet-50			DenseNet-121			SENet-18		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
BO (Ru et al., 2020)	99.6%	83	44	99.7%	93	51	99.7%	81	41
π-BO (Hvarfner et al., 2022)	99.9%	78	12	99.9%	61	11	99.4%	49	11
P-BO (λ^*, **ours**)	100.0%	15	11	100.0%	19	11	100.0%	14	11

Table 10. The experimental results of P-BO variants against Inception-v3, MobileNet-v2, and ViT-B/16 under the ℓ_∞ norm on ImageNet. We report the attack success rate (ASR) and the average/median number of queries (AVG. Q/MED. Q) needed to generate an adversarial example over successful attacks. We mark the best results in **bold**. The subscript "D" denotes the methods with dimensionality reduction.

Method	Inception-v3			MobileNet-v2			ViT-B/16		
	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q	ASR	AVG. Q	MED. Q
P-BO$_D$ w/o BO	39.8%	55	2	64.8%	102	11	25.6%	96	6
P-BO$_D$ ($\lambda = 1$)	85.4%	182	90	86.8%	193	56	67.5%	236	240
P-BO$_D$ (λ^*)	94.4%	81	45	98.8%	94	60	88.2%	148	81

The remaining experimental configurations remain consistent with those outlined in Section 4.2. We compare π-BO with BO and P-BO in Table 9. While π-BO exhibits some improvements over BO, there remains a significant performance gap compared to our P-BO method. This is because the surrogate model is more appropriately utilized as a function prior and P-BO can effectively integrate the prior information for black-box attacks.

D.6. Ablation Study

Here, we conduct an ablation study to validate the necessity of Bayesian optimization (BO). We directly employ the function prior as the acquisition function α in Algorithm 1, independent of observed values from the target function, denoted as PBO w/o BO. This is equivalent to repeatedly conducting transfer attacks until succeeded. We adopt the experimental setup identical to Sec. 4.3 on ImageNet, and the results for the PBO w/o BO are presented in Table 10. It can be observed that this approach also demonstrates promising performance, often achieving successful attacks with only a few queries. However, the success rate is significantly lower compared to the P-BO algorithm. This suggests that the integration of function prior and observed values from the target function for optimization exploration, as employed by Bayesian optimization, is a crucial and effective approach.