BOOSTING RAY SEARCH PROCEDURE OF HARD-LABEL ATTACKS WITH TRANSFER-BASED PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the most practical and challenging types of black-box adversarial attacks is the hard-label attack, where only top-1 predicted labels are available. One effective approach is to search for the optimal ray direction from the benign image that minimizes the ℓ_p norm distance to the adversarial region. The unique advantage of this approach is that it transforms the hard-label attack into a continuous optimization problem. The objective function value is the ray's radius and can be obtained through a binary search with high query cost. Existing methods use a "sign trick" in gradient estimation to reduce queries. In this paper, we theoretically analyze the quality of this gradient estimation, proposing a novel prior-guided approach to improve ray search efficiency, based on theoretical and experimental analysis. Specifically, we utilize the transfer-based priors from surrogate models, and our gradient estimators appropriately integrate them by approximating the projection of the true gradient onto the subspace spanned by these priors and some random directions, in a query-efficient way. We theoretically derive the expected cosine similarity between the obtained gradient estimators and the true gradient, and demonstrate the improvement brought by using priors. Extensive experiments on the ImageNet and CIFAR-10 datasets show that our approach significantly outperforms 11 state-of-the-art methods in query efficiency. Code will be released.

1 INTRODUCTION

Adversarial attacks represent a major security threat to deep neural networks (DNNs), where subtle, imperceptible perturbations are crafted to cause misclassifications. To assess DNN robustness and uncover vulnerabilities, the research community has developed various adversarial attack strategies. As a result, adversarial attacks and defenses have become a focal point in AI security research.

Based on the available information about the target model, adversarial attacks can be broadly classified into white-box and black-box types. White-box attacks, such as those in Madry et al. (2018); Moosavi-Dezfooli et al. (2016), rely on the target model's gradients with respect to the input, making them less practical in real-world applications. Black-box attacks, by contrast, are often more feasible, as they do not require knowledge of model parameters or gradients. A subset of black-box attacks, known as transfer-based attacks, generates adversarial examples using white-box models in an attempt to generalize to other models. While transfer-based attacks do not involve querying the target model, their success rate is inconsistent. Alternatively, query-based black-box attacks iteratively interact with the target model to achieve higher success rates. These attacks can be categorized into two subtypes: score-based and decision-based (also known as hard-label) attacks. Score-based attacks utilize the model's output logits to guide the attack, whereas hard-label attacks rely solely on top-1 predicted labels, making them particularly practical when only label information is accessible. In this work, we focus on the problem of reducing query complexity in hard-label attacks.

The difficulty of hard-label attacks is that the labels can only be flipped near the classification decision boundary, and thus the objective function is discontinuous. As a result, the attack requires solving a high-dimensional combinatorial optimization problem, which is challenging. Common approaches Chen et al. (2020); Brendel et al. (2018) start with a sample containing large adversarial perturbations and iteratively reduce the distortion by moving along the decision boundary towards a benign image. However, these methods lack convergence guarantees. To reformulate the problem as a continuous optimization task, direction optimization-based methods have been introduced. Typical approaches such as OPT Cheng et al. (2019a), Sign-OPT Cheng et al. (2020), and RayS Chen & Gu (2020) aim to minimize an objective function $g(\theta)$, which is defined as the shortest ℓ_p norm distance alone the ray direction θ from the benign image to the adversarial region. This function value can be evaluated using a binary search. Leveraging the smooth and continuous nature of decision boundaries, $g(\theta)$ is also locally continuous, making it amenable to zeroth-order (ZO) optimization with a gradient estimator. OPT employs a random gradient-free (RGF) estimator, but it incurs high query cost due to the binary search in finite differences. Sign-OPT reduces the query complexity by using the sign of directional derivative in gradient estimation, but it significantly sacrifices gradient accuracy.

To solve this problem and improve query efficiency, we employ the same objective function $g(\theta)$, and propose incorporating the transfer-based priors into gradient estimation. An ideal prior is the gradient of $g(\theta)$ from a surrogate model, but it cannot be easily obtained since $g(\theta)$ is non-differentiable due to the binary search. Instead, we propose a surrogate loss, whose gradient is proportional to that of $g(\theta)$, to obtain the prior. Once the transfer-based priors are obtained, we must design better gradient estimators that effectively integrate these priors. This is particularly challenging under the hard-label restriction, as accurately determining the value of $g(\theta)$ is costly. As a result, previous prior-guided methods of score-based attacks such as PRGF Cheng et al. (2019b); Dong et al. (2022) are not suitable in this context. Thus, we need to explore how to improve the gradient estimator with additional priors while minimizing queries. To achieve this, we propose two algorithms: Prior-Sign-OPT and Prior-OPT. They estimate the gradient in a query-efficient manner by approximating the projection of the true gradient onto a subspace spanned by priors and randomly sampled vectors. We provide a thorough theoretical analysis to validate their effectiveness and offer theoretical comparisons between Sign-OPT and our approach. In particular, Prior-OPT achieves a better approximation of the subspace projection with only slightly more queries, and can adaptively adjust a prior's weight based on its quality, striking a balance between gradient accuracy and query efficiency. While several methods Brunner et al. (2019); Shi et al. (2023) attempt to combine transfer- and decision-based attacks, they lack theoretical guarantees and often perform poorly. Crucially, in the hard-label setting, these approaches fail to effectively address the challenge of appropriately weighing the prior when it deviates significantly from the true gradient. Our approach resolves this issue and naturally scales to priors from multiple surrogate models, demonstrating further improvement in attack performance.

To summarize, our main contributions are as follows.

1. **Novelty in hard-label attacks.** We address the problem of introducing the transfer-based priors into hard-label attacks by employing the subspace projection approximation, which significantly improves the accuracy of gradient estimation with slightly more queries. Our approach not only strikes a balance between gradient estimation and query efficiency, but also elegantly integrates priors from multiple surrogate models to further improve performance.
2. **Novelty in theoretical analysis.** We analyze the quality of our gradient estimators and the (orthogonal variant of) Sign-OPT, enabling theoretical comparisons. To our knowledge, this is the first time to derive the expected cosine similarity between estimators of the Sign-OPT family and the true gradient, offering a theoretical guarantee of performance improvement.
3. **Extensive experiments**. Extensive experiments conducted on the ImageNet and CIFAR-10 datasets show that our approach outperforms 11 state-of-the-art methods significantly.

2 Related Work

The hard-label attacks can be categorized as boundary-search and direction-optimization approaches.

The boundary-search approaches start from a large perturbation or an image of the target class, and then reduce distortions by moving along the decision boundary, thereby gradually making it closer to the original image. Boundary Attack (BA) Brendel et al. (2018) is an early representative method, and its query efficiency is relatively low. HopSkipJumpAttack (HSJA) Chen et al. (2020) estimates a gradient at the decision boundary as the direction of the first update, and then find the next boundary point towards the benign image. Tangent Attack (TA) and Generalized Tangent Attack (G-TA) Ma et al. (2021) find an optimal tangent point on a virtual hemisphere or semi-ellipsoid to reach the adversarial example. CGBA Reza et al. (2023) conducts a boundary search along a semicircular path on a restricted 2D plane to find the boundary point. To eliminate the gradient estimation, SurFree Maho et al. (2021) and Triangle Attack Wang et al. (2022) find the adversarial example in a DCT space to improve query efficiency. Evolutionary Dong et al. (2019) finds better search directions by modeling their local geometries. Adaptive History-driven Attack (AHA) Li et al. (2021a) gathers data of previous queries as the prior for current sampling, which improves the random walk optimization.

The direction-optimization approaches aim to find an optimal direction θ reaching the nearest adversarial region. As mentioned in Section 1, it is challenging to address both the high query complexity issue of OPT and the low estimation accuracy issue of Sign-OPT. RayS Chen & Gu (2020) eliminates the gradient estimation and employs a hierarchical search step to efficiently find the optimal direction. However, RayS only supports ℓ_∞ norm untargeted attacks. The query efficiency of previous direction-optimization approaches does not surpass that of boundary-search approaches, making these approaches less studied. We note that the mechanisms of OPT and Sign-OPT remain poorly understood, and their inefficiency stems from the limited precision in gradient estimation.

Several methods attempt to combine transfer- and decision-based attacks, but the critical issue, namely how to weigh the prior when it deviates significantly from the true gradient, has not been well addressed. For example, BBA Brunner et al. (2019), CISA Shi et al. (2023) and SQBA Park et al. (2024) set the prior's coefficient empirically rather than through theoretical analysis. Other methods Yan et al. (2021); Li et al. (2021b) train a model to simulate a specified target model, making them difficult to attack different models. In contrast, our method dynamically calculates optimal coefficients and theoretically improves the accuracy of the estimated gradient.

3 THE PROPOSED METHOD

3.1 THE GOAL OF HARD-LABEL ATTACKS

Given a k-classes classifier $f : \mathbb{R}^d \to \mathbb{R}^k$ and a benign image $\mathbf{x} \in [0,1]^d$ which is correctly classified by f, the adversary aims to find an adversarial example $\mathbf{x}_{\text{adv}}$ with the minimum perturbation such that $f(\mathbf{x}_{\text{adv}})$ outputs incorrect prediction. Formally, we formulate the attack goal as:

$$\min_{\mathbf{x}_{\text{adv}}} d(\mathbf{x}_{\text{adv}}, \mathbf{x}) \quad \text{s.t.} \quad \Phi(\mathbf{x}_{\text{adv}}) = 1, \tag{1}$$

where $d(\mathbf{x}_{\text{adv}}, \mathbf{x}) := \|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_p$ is the ℓ_p norm distortion, and $\Phi(\cdot)$ is a success indicator function as:

$$\Phi(\mathbf{x}_{\text{adv}}) := \begin{cases} 1 & \text{if } \hat{y} = y_{\text{adv}} \text{ in the targeted attack,} \\ & \quad \text{or } \hat{y} \neq y \text{ in the untargeted attack,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\hat{y} = \arg\max_{i \in \{1,\cdots,k\}} f(\mathbf{x}_{\text{adv}})_i$ is the top-1 predicted label of f, $y \in \mathbb{R}$ is the true label of $\mathbf{x}$, and $y_{\text{adv}} \in \mathbb{R}$ is a target class label. In this study, we follow Cheng et al. (2019a; 2020) to reformulate the problem (1) as the problem of finding the ray direction of the shortest distance from $\mathbf{x}$ to the adversarial region:

$$\min_{\theta} g(\theta) \quad \text{where} \quad g(\theta) = \arg\min_{\lambda > 0} \left(\Phi(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}) = 1 \right). \tag{3}$$

Finally, the adversarial example is $\mathbf{x}^* = \mathbf{x} + g(\theta^*)\frac{\theta^*}{\|\theta^*\|}$, and θ^* is the optimal solution of problem (3).

3.2 THE OPTIMIZATION OF SEARCHING RAY DIRECTIONS

The previous works Cheng et al. (2019a; 2020) attempt to optimize the problem (3) by using ZO methods. However, the restriction of hard label access results in a high query cost of the gradient estimation, because obtaining a single value of $g(\theta)$ requires performing a binary search function with multiple queries, and the gradient estimation with finite difference requires multiple computations of $g(\theta)$. Sign-OPT Cheng et al. (2020) replaces $g(\theta + \sigma\mathbf{u}) - g(\theta)$ of the finite difference with $\text{sign}(g(\theta + \sigma\mathbf{u}) - g(\theta))$, which improves query efficiency by only using a single query (Eq. (8)). However, it also significantly reduces the accuracy of the gradient estimation. We propose to incorporate transfer-based priors to improve the accuracy of gradient estimation without excessively increasing query complexity, thus achieving an optimal balance between query complexity and estimation accuracy.

The first challenge is how to obtain a transfer-based prior $\nabla g_{\hat{f}}(\theta)$ from a surrogate model $\hat{f}$, where $g_{\hat{f}}(\theta)$ represents the shortest distance along the direction θ to the adversarial region of $\hat{f}$. This is challenging because $g_{\hat{f}}(\theta)$ is typically evaluated using binary search, making it non-differentiable. To address this, for any non-zero vector $\theta_0 \in \mathbb{R}^d$ such that $g_{\hat{f}}(\theta_0) < +\infty$, we define a surrogate function $h(\theta, \lambda)$ such that $\nabla g_{\hat{f}}(\theta_0) = k \cdot \nabla_\theta h(\theta_0, \lambda_0)$, where $\lambda_0 = g_{\hat{f}}(\theta_0)$ is treated as a constant during differentiation. Here, λ is a scalar, and k is a non-zero constant.

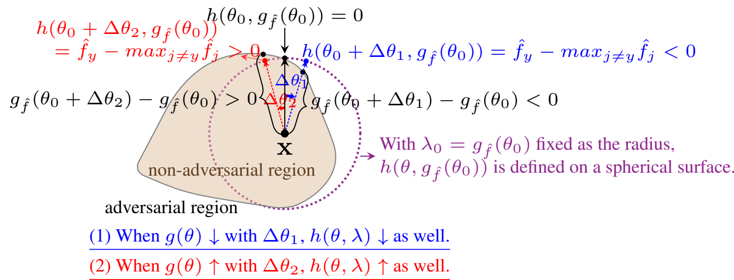


Figure 1: Geometrical explanation of $\nabla g_{\hat{f}}(\theta_0) \propto \nabla_\theta h(\theta_0, \lambda_0)$ by taking an untargeted attack as an example. When $g_{\hat{f}}(\theta)$ reduces/increases with a small $\Delta\theta$, $h(\theta, \lambda)$ changes at a similar rate. The formal proof is in Appendix C.

The surrogate function $h(\theta, \lambda)$ is defined as the negative C&W loss function of $\hat{f}$:

$$h(\theta, \lambda) = \begin{cases} \hat{f}_y - \max_{j \neq y} \hat{f}_j, & \text{if untargeted attack;} \\ \max_{j \neq \hat{y}_{\text{adv}}} \hat{f}_j - \hat{f}_{\hat{y}_{\text{adv}}}, & \text{if targeted attack;} \end{cases} \tag{4}$$

where $\hat{f}_i := \hat{f}(\mathbf{x} + \lambda \cdot \frac{\theta}{\|\theta\|})_i$ is an abbreviation for the i-th element of the output of $\hat{f}$, and $\mathbf{x}$ is the original image. Any non-zero scalar can be used as λ, but the specific value $\lambda_0 = g_{\hat{f}}(\theta_0)$ yields a gradient proportional to $\nabla g_{\hat{f}}(\theta_0)$. The value λ_0 is obtained through binary search, where $h(\theta_0, \lambda_0)$ represents the negative C&W loss at the decision boundary of the surrogate model $\hat{f}$. The geometric explanation and formal proof of $\nabla g_{\hat{f}}(\theta_0) = k \cdot \nabla_\theta h(\theta_0, \lambda_0)$ are presented in Fig. 1 and Appendix C.

In targeted attacks, determining an appropriate λ_0 value becomes a challenging task. This is because the spatial distribution of classification locations in the original image space varies across different models. For example, an adversarial image classified as a cat in the ResNet-101 model may occupy a position corresponding to a different class, such as rabbit, in the ViT model. Therefore, although we can locate the region corresponding to the predefined target class y_{adv} along the θ direction in the target model f, the same direction may not lead to the region of class y_{adv} in a surrogate model $\hat{f}$. Therefore, we must set a new target class $\hat{y}_{\text{adv}}$ before determining λ_0 and computing Eq. (4). The detailed steps for this procedure are provided in Appendix B. Given s surrogate models with priors, we present an efficient method to incorporate them into gradient estimation.

Given s non-zero vectors $\mathbf{q}_1, \cdots, \mathbf{q}_s$ representing the transfer-based priors and $q - s$ randomly sampled vectors $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \cdots, q-s$, our objective is to estimate a gradient $\mathbf{v}^* \approx \nabla g(\theta)$ as accurately as possible using these q vectors. In *the score-based attack setting*, there is a subspace projection estimator theory Meier et al. (2019); Cheng et al. (2021) that can solve this problem. Based on this theory, the optimal estimated gradient $\mathbf{v}^*$ that maximizes its similarity with the true gradient is given by Proposition 3.1 in the score-based setting.

Proposition 3.1. *(Optimality of the subspace projection estimator) Let $\mathbf{q}_1, \cdots, \mathbf{q}_s$ and $\mathbf{r}_1, \cdots, \mathbf{r}_{q-s}$ be defined above, and $S := \operatorname{span}\{\mathbf{q}_1, \cdots, \mathbf{q}_s, \mathbf{r}_1, \cdots, \mathbf{r}_{q-s}\}$ be the subspace spanned by these vectors, then the optimal $\mathbf{v}^*$ in S that maximizes $\overline{\nabla g(\theta)}^\top \mathbf{v}$ subject to $\|\mathbf{v}\| = 1$ is the ℓ_2 normalized projection of $\nabla g(\theta)$ onto S, denoted as $\mathbf{v}^* = \overline{\nabla g(\theta)_S}$.*

According to Proposition 3.1, finding the optimal approximate gradient is equivalent to finding a projection of the true gradient onto a low-dimensional subspace S spanned by all available vectors. The projection of a vector onto a subspace S can be calculated by summing its projections onto the orthonormal basis of S. To achieve this, we construct an orthonormal basis of S by using the Gram-Schmidt orthogonalization, which transforms $\mathbf{q}_1, \cdots, \mathbf{q}_s, \mathbf{r}_1, \cdots, \mathbf{r}_{q-s}$ into an orthonormal basis $\mathbf{p}_1, \cdots, \mathbf{p}_s, \mathbf{u}_1, \cdots, \mathbf{u}_{q-s}$. Note that $\mathbf{p}_1, \cdots, \mathbf{p}_s$ correspond to $\mathbf{q}_1, \cdots, \mathbf{q}_s$, and $\mathbf{u}_1, \cdots \mathbf{u}_{q-s}$ correspond to $\mathbf{r}_1, \cdots, \mathbf{r}_{q-s}$. Then, we can compute the projection of $\nabla g(\theta)$ onto S by Eq. (5):

$$\mathbf{v}^* = \sum_{i=1}^{s} \nabla g(\theta)^\top \mathbf{p}_i \cdot \mathbf{p}_i + \sum_{i=1}^{q-s} \nabla g(\theta)^\top \mathbf{u}_i \cdot \mathbf{u}_i. \tag{5}$$

With the queried function values, $\nabla g(\theta)^\top \mathbf{u}$ for the unit ℓ_2 norm $\mathbf{u}$ can be approximated by the finite difference method, without requiring backpropagation:

$$\nabla g(\theta)^\top \mathbf{u} \approx \frac{g(\theta + \sigma \mathbf{u}) - g(\theta)}{\sigma}, \tag{6}$$

where σ is a small positive number. By plugging Eq. (6) into Eq. (5), we can easily calculate $\mathbf{v}^*$ in the score-based setting as $\mathbf{v}^* = \sum_{i=1}^{s} \frac{g(\theta+\sigma\mathbf{p}_i)-g(\theta)}{\sigma} \cdot \mathbf{p}_i + \sum_{i=1}^{q-s} \frac{g(\theta+\sigma\mathbf{u}_i)-g(\theta)}{\sigma} \cdot \mathbf{u}_i$. However, in *the hard-label setting*, the finite difference requires a large number of queries due to the binary search of $g(\cdot)$. We propose two algorithms to reduce query cost by computing the approximate projection, *i.e.,* Prior-Sign-OPT and Prior-OPT. With s priors, Prior-Sign-OPT can use Eq. (7) to save queries:

$$\mathbf{v}^* = \sum_{i=1}^{s} \operatorname{sign}(g(\theta + \sigma \mathbf{p}_i) - g(\theta)) \cdot \mathbf{p}_i + \sum_{i=1}^{q-s} \operatorname{sign}(g(\theta + \sigma \mathbf{u}_i) - g(\theta)) \cdot \mathbf{u}_i. \tag{7}$$

Eq. (7) is similar to the formula of Sign-OPT, benefiting from using only a single query to calculate the sign of the directional derivative Cheng et al. (2020):

$$\operatorname{sign}(g(\theta + \sigma \mathbf{u}_i) - g(\theta)) = \begin{cases} +1, & f\left(\mathbf{x} + g(\theta)\frac{\theta+\sigma\mathbf{u}_i}{\|\theta+\sigma\mathbf{u}_i\|}\right) = y, \\ -1, & \text{Otherwise.} \end{cases} \tag{8}$$

The accuracy of the estimated gradient is crucial in optimization. A natural metric for measuring its accuracy is $\mathbb{E}[\gamma]$ and $\mathbb{E}[\gamma^2]$, where γ is its cosine similarity with the true gradient. We develop an innovative approach for computing $\mathbb{E}[\gamma]$ and $\mathbb{E}[\gamma^2]$ of Sign-OPT, Prior-Sign-OPT and Prior-OPT. Our baseline Sign-OPT is based on orthogonal random vectors, which is a variant of Cheng et al. (2020).

Theorem 3.2. *For the Sign-OPT estimator approximated by Eq.* (6) *(defined as Eq.* (38)*), we let* $\gamma = \overline{\mathbf{v}}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}} := \frac{\mathbf{v}}{\|\mathbf{v}\|}$*, then*

$$\mathbb{E}[\gamma] = \sqrt{q}\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}}, \tag{9}$$

$$\mathbb{E}[\gamma^2] = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right). \tag{10}$$

The proof of Theorem 3.2 is included in the Appendix A.1. For Prior-Sign-OPT, we have Theorem 3.3.

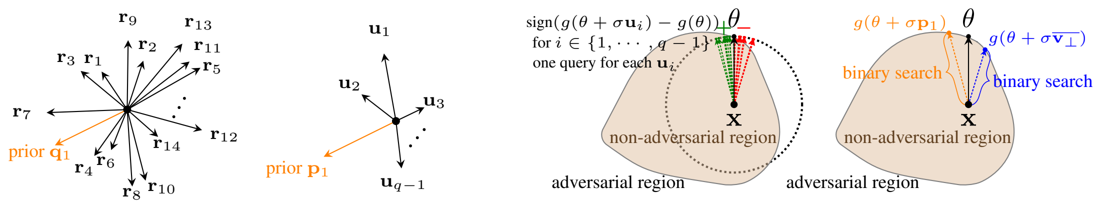
Theorem 3.3. *For the Prior-Sign-OPT estimator approximated by Eq.* (6) *(defined as Eq.* (76)*), we let* $\gamma = \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$*, then*

$$\mathbb{E}[\gamma] = \frac{1}{\sqrt{q}}\left[\sum_{i=1}^{s}|\alpha_i| + (q-s)\sqrt{1-\sum_{i=1}^{s}\alpha_i^2}\cdot\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}\right], \tag{11}$$

$$\begin{aligned}\mathbb{E}[\gamma^2] = \frac{1}{q}\Bigg[&\left(\sum_{i=1}^{s}|\alpha_i|\right)^2 + \frac{q-s}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)\left(1-\sum_{i=1}^{s}\alpha_i^2\right) \\ &+ 2\left(\sum_{i=1}^{s}|\alpha_i|\right)(q-s)\sqrt{1-\sum_{i=1}^{s}\alpha_i^2}\cdot\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}\Bigg],\end{aligned} \tag{12}$$

where $\alpha_i := \mathbf{p}_i^\top \overline{\nabla g(\theta)}$ *is the cosine similarity between the* i*-th prior and the true gradient.*

The proof of Theorem 3.3 is presented in the Appendix A.2. Now we can compare $\mathbb{E}[\gamma]$ of Sign-OPT (Eq. (9)) and Prior-Sign-OPT (Eq. (11)). In Sign-OPT, $\mathbb{E}[\gamma] \le \sqrt{\mathbb{E}[\gamma^2]} < \sqrt{2(q-1)/(\pi d)}$. When $q \ll d$, $\mathbb{E}[\gamma]$ becomes very small. In Prior-Sign-OPT, we can conclude that $\mathbb{E}[\gamma] \ge (\sum_{i=1}^{s}|\alpha_i|)/\sqrt{q}$ based on Eq. (11). Benefit from using priors, if $(\sum_{i=1}^{s}|\alpha_i|)/\sqrt{q} > \sqrt{2(q-1)/(\pi d)}$, where α_i and d are large, and q is small, then Prior-Sign-OPT outperforms Sign-OPT. In some cases, Prior-Sign-OPT may be worse than Sign-OPT. For example, when $\sum_{i=1}^{s}\alpha_i^2 = 1$ and q is large in Eq. (11), $\mathbb{E}[\gamma]$ becomes very small. Intuitively, random vectors $\mathbf{u}_1, \dots, \mathbf{u}_{q-s}$ are orthogonal to $\{\mathbf{p}_1, \cdots, \mathbf{p}_s\}$ and less correlated with the true gradient, but Eq. (7) assigns coefficients of all vectors to the same



(a) Get prior $\mathbf{q}_1$ and sample $\mathbf{r}_i$. (b) Orthonormal basis. (c) Estimate a sign-based $\mathbf{v}_\perp$. (d) Estimate $\mathbf{v}^*$ with $\mathbf{v}_\perp$ and $\mathbf{p}_1$.

Figure 2: Simplified two-dimensional illustration of the gradient estimation of Prior-OPT with a single transfer-based prior. In Fig. 2a, we first sample random vectors $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \cdots, q-1$ and obtain a transfer-based prior $\mathbf{q}_1$ using $\nabla_\theta h(\theta, \lambda)$, where $h(\theta, \lambda)$ is defined in Eq. (4). Then, as shown in Fig. 2b, we perform Gram-Schmidt orthogonalization on these vectors to obtain an orthonormal basis $\mathbf{p}_1, \mathbf{u}_1, \cdots, \mathbf{u}_{q-1}$, where $\mathbf{p}_1 = \mathbf{q}_1$. Next, we estimate $\mathbf{v}_\perp$ based on Eq. (14) with $\mathbf{u}_1, \cdots, \mathbf{u}_{q-1}$ (Fig. 2c), where each $\text{sign}(g(\theta + \sigma\mathbf{u}_i) - g(\theta))$ requires only a single query (Eq. (8)). Finally, as shown in Fig. 2d, we estimate a gradient $\mathbf{v}^*$ based on Eq. (13) with $\mathbf{p}_1$ and $\mathbf{v}_\perp$, where the values of $g(\theta + \sigma\mathbf{p}_1)$ and $g(\theta + \sigma\overline{\mathbf{v}_\perp})$ are obtained via the binary search with multiple queries.

magnitude (just $+1$ or -1), rather than their respective inner products with the true gradient, which makes its estimation less accurate. Therefore, we propose Prior-OPT to address this problem. Fig. 2 illustrates the process of gradient estimation in Prior-OPT, which is based on the following formula:

$$\mathbf{v}^* = \sum_{i=1}^{s} \frac{g(\theta + \sigma\mathbf{p}_i) - g(\theta)}{\sigma} \cdot \mathbf{p}_i + \frac{g(\theta + \sigma\overline{\mathbf{v}_\perp}) - g(\theta)}{\sigma} \cdot \overline{\mathbf{v}_\perp}, \tag{13}$$

where $\overline{\mathbf{v}_\perp}$ is the ℓ_2 normalization of $\mathbf{v}_\perp$, and $\mathbf{v}_\perp$ is obtained by:

$$\mathbf{v}_\perp = \sum_{i=1}^{q-s} \text{sign}(g(\theta + \sigma\mathbf{u}_i) - g(\theta)) \cdot \mathbf{u}_i. \tag{14}$$

$\mathbf{u}_1, \cdots, \mathbf{u}_{q-s}$ are random vectors, which contain less information. We aggregate them into a single $\mathbf{v}_\perp$ using a less accurate formula Eq. (14), which is orthogonal to all priors. Compared with Eq. (7), Eq. (13) is a more accurate approximation of the projection. Then, we have Theorem 3.4 as follows.

Theorem 3.4. *For the Prior-OPT estimator approximated by Eq.* (6) *(defined as Eq.* (107)*), we let* $\gamma = \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where* $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$*, then*

$$\mathbb{E}[\gamma] \geq \sqrt{\sum_{i=1}^{s} \alpha_i^2 + \frac{(q-s)(1 - \sum_{i=1}^{s} \alpha_i^2)}{\pi} \left(\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})} \right)^2}, \tag{15}$$

$$\mathbb{E}[\gamma] \leq \sqrt{\sum_{i=1}^{s} \alpha_i^2 + \frac{1}{d-s} \left(\frac{2}{\pi}(q-s-1) + 1 \right) \left(1 - \sum_{i=1}^{s} \alpha_i^2 \right)}, \tag{16}$$

$$\mathbb{E}[\gamma^2] = \sum_{i=1}^{s} \alpha_i^2 + \frac{1}{d-s} \left(\frac{2}{\pi}(q-s-1) + 1 \right) \left(1 - \sum_{i=1}^{s} \alpha_i^2 \right), \tag{17}$$

where $\alpha_i := \mathbf{p}_i^\top \overline{\nabla g(\theta)}$ *is the cosine similarity between the* i*-th prior and the true gradient.*

The proof of Theorem 3.4 is included in the Appendix A.3. $\mathbb{E}[\gamma^2]$ is the second-order moment, which reflects the magnitude of γ in a statistical sense. Under certain assumptions, $\mathbb{E}[\gamma^2]$ directly affects the convergence rate of optimization algorithms, as shown in Cheng et al. (2021). Intuitively, a larger γ indicates more accurate gradient estimation, leading to faster optimization and improved query efficiency. By Jensen's inequality ($(\mathbb{E}[\gamma])^2 \leq \mathbb{E}[\gamma^2]$), Eq. (16) holds. We next compare $\mathbb{E}[\gamma^2]$ for Prior-OPT (Eq. (17)) and Sign-OPT (Eq. (10)). When each $\alpha_i = 0$ and $s \ll q$, Eq. (17) $\approx$ Eq. (10). However, when $s \ll q \ll d$ and $\sum_{i=1}^{s} \alpha_i^2 > \frac{2s}{\pi d}$, with $\frac{2s}{\pi d}$ being an approximate value, Prior-OPT outperforms Sign-OPT by leveraging priors with $\overline{\alpha^2} > \frac{2}{\pi d}$. See Appendix D for details.

Algorithm 1 presents the algorithmic process. The initialization of θ_0 has two options in *untargeted attacks*: (1) θ_0^{RND}: we select the best direction with the shortest distortion from 100 random directions

Algorithm 1 Prior-Sign-OPT and Prior-OPT attack

Input: benign image $\mathbf{x}$, objective function $g(\cdot)$, attack success indicator $\Phi(\cdot)$ defined in Eq. (2), iteration T, method $m \in \{$ `Prior-OPT`, `Prior-Sign-OPT`$\}$, initialization strategy of untargeted attacks $\in \{\theta_0^{\text{PGD}}, \theta_0^{\text{RND}}\}$, the maximum gradient norm $\mathbf{g}_{\text{max}}$, attack norm $p \in \{2, \infty\}$, surrogate models $\mathbb{S} = \{\hat{f}_1, \cdots, \hat{f}_s\}$.
Output: adversarial example $\mathbf{x}^*$ that satisfies $\Phi(\mathbf{x}^*) = 1$.
$\tilde{\mathbf{x}}_0 \leftarrow$ PGD$(\mathbf{x}, \hat{f}_1)$ if initialization $= \theta_0^{\text{PGD}}$, otherwise a random $\tilde{\mathbf{x}}_0$ that satisfies $\Phi(\tilde{\mathbf{x}}_0) = 1$ is selected, which is θ_0^{RND} strategy; ▷ the targeted attack selects an image from the target class as $\tilde{\mathbf{x}}_0$.
$\theta_0 \leftarrow \frac{\tilde{\mathbf{x}}_0 - \mathbf{x}}{\|\tilde{\mathbf{x}}_0 - \mathbf{x}\|}, \quad d_0 \leftarrow \|\tilde{\mathbf{x}}_0 - \mathbf{x}\|_p$;
for t **in** $1, \ldots, T$ **do**
  for $\hat{f}_i$ **in** $\mathbb{S}$ **do**
    $\lambda_{t-1} \leftarrow$ BinarySearch$(\mathbf{x}, \theta_{t-1}, \hat{f}_i, \Phi)$;
    $\mathbf{q}_i \leftarrow \nabla_\theta h(\theta_{t-1}, \lambda_{t-1})$ on $\hat{f}_i$ with λ_{t-1} treated as a constant in differentiation; ▷ obtain s transfer-based priors
  end for
  $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \cdots, q - s$;
  $\mathbf{p}_1, \cdots, \mathbf{p}_s, \mathbf{u}_1, \cdots, \mathbf{u}_{q-s} \leftarrow$ Gram-Schmidt Orthogonalization$(\{\mathbf{q}_1, \cdots, \mathbf{q}_s, \mathbf{r}_1, \cdots, \mathbf{r}_{q-s}\})$;
  Estimate a gradient $\mathbf{v}^*$ using Eq. (7) if $m =$ `Prior-Sign-OPT`, otherwise using Eq. (13);
  $\mathbf{v}^* \leftarrow$ ClipGradNorm$(\mathbf{v}^*, \mathbf{g}_{\text{max}})$;
  $\alpha^* \leftarrow$ LineSearch$(\mathbf{x}, \mathbf{v}^*, \Phi, d_{t-1})$; ▷ search step size.
  $\theta_t \leftarrow \theta_{t-1} - \alpha^* \mathbf{v}^*, \quad d_t = \|\mathbf{x}_t - \mathbf{x}\|_p$;
end for
return $\mathbf{x}^* \leftarrow \mathbf{x} + g(\theta_T) \frac{\theta_T}{\|\theta_T\|}$;

as θ_0; (2) θ_0^{PGD}: we apply PGD Madry et al. (2018) to attack against a surrogate model $\hat{f}_0$ to initialize θ_0, which uses the transfer-based attack as initialization. In *targeted attacks*, we initialize θ_0 with an image $\tilde{\mathbf{x}}_0$ selected from the target class in the training set. In each iteration, it first calculates the gradient of Eq. (4) on each surrogate model $\hat{f}_i$ in $\mathbb{S}$ to obtain the priors $\mathbf{q}_1, \cdots, \mathbf{q}_s$. Then, we combine these priors and the randomly sampled vectors $\mathbf{r}_1, \cdots, \mathbf{r}_{q-s}$ into a list $\mathbb{L}$, where the priors are positioned ahead of the random vectors. After performing Gram-Schmidt orthogonalization on $\mathbb{L}$, the orthogonal vectors $\mathbf{p}_1, \cdots, \mathbf{p}_s, \mathbf{u}_1, \cdots, \mathbf{u}_{q-s}$ are obtained, representing an orthonormal basis of the subspace. With these orthogonal vectors, we compute Eq. (7) or Eq. (13) to estimate a gradient $\mathbf{v}^*$ in Prior-Sign-OPT and Prior-OPT, respectively. Then, we employ the gradient clipping technique to address the large norm gradient problem caused by finite difference. Finally, we use the line search to find the best step size α^* and perform gradient descent step to minimize $g(\theta)$.

4 Experiments

4.1 Experimental Setting

Datasets. All experiments are conducted on two datasets, i.e., CIFAR-10 Krizhevsky (2009) and ImageNet Deng et al. (2009), and the image sizes are $32 \times 32 \times 3$ and $299 \times 299 \times 3$ or $224 \times 224 \times 3$, respectively. We randomly selected $1{,}000$ images from the validation sets for experiments. In targeted attacks, we use the same image $\tilde{x}_0$ as the initialization for each target class across all methods, setting the target label as $y_{adv} = (y + 1) \mod k$, where y is the true label and k is the number of classes. Due to space limitations, the CIFAR-10 and ℓ_∞ norm attack results are provided in Appendix F.2.

Method Setting. The hyperparameter setting of our method is included in Table 3 of Appendix E. We denote the surrogate models in the subscript of our method in experiments. For instance, "Prior-$\text{OPT}_{\text{ResNet50\&ConViT}}$" means using the ResNet-50 and ConViT as the surrogate models for Prior-OPT, and "Prior-$\text{OPT}_{\theta_0^{\text{PGD}} + \text{ResNet50}}$" applies PGD attack on a surrogate model ResNet-50 to initialize θ_0.

Compared Methods. To provide a comprehensive comparison, we select 11 state-of-the-art hard-label attacks, including Sign-OPT, SVM-OPT Cheng et al. (2020), HSJA Chen et al. (2020), Triangle Attack Wang et al. (2022), TA, G-TA Ma et al. (2021), SurFree Maho et al. (2021), GeoDA Rahmati et al. (2020), Evolutionary Dong et al. (2019), BBA Brunner et al. (2019), and SQBA Park et al.

(2024). Sign-OPT is adopted as the baseline. SQBA, Triangle Attack, GeoDA, and our θ_0^{PGD} initialization strategy (denoted as Prior-OPT$_{\theta_0^{\text{PGD}}}$ and Prior-Sign-OPT$_{\theta_0^{\text{PGD}}}$) only support untargeted attacks. BBA and SQBA use surrogate models, denoted by subscripts in the experimental results.

Target Models and Surrogate Models. In the ImageNet dataset, we select 8 neural network architectures as the target models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The selected target models are Inception-v3 Szegedy et al. (2016), Inception-v4 Szegedy et al. (2017), ResNet-101 He et al. (2016), SENet-154 Hu et al. (2018), ResNeXt-101 ($64 \times 4d$) Xie et al. (2017), Vision Transformer (ViT) Dosovitskiy et al. (2021), Swin Transformer Liu et al. (2021), and Global Context Vision Transformer (GC ViT) Hatamizadeh et al. (2023). The Inception-V3 and Inception-V4 require a resolution of 299×299 for the input images, and we select InceptionResNet-V2 (IncResV2) and Xception as the surrogate models. ResNet-50 and ConViT d'Ascoli et al. (2021) are selected as the surrogate models for other models. In the attacks of the defense models, we use the adversarially trained (AT) surrogate models, denoted as "AT(ResNet50)" in the subscript of Prior-OPT in experimental results.

Evaluation Metric. All methods are evaluated using the mean distortion over $1{,}000$ images as $\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}(\|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_p)$ under different query budgets, where $\mathbf{X}$ is the test set and $p \in \{2, \infty\}$ is the attack norm. We also report the attack success rate (ASR) under the specific query budget, which is defined as the percentage of samples with distortions below a threshold ϵ. In ℓ_2 norm attacks, we set the threshold $\epsilon = \sqrt{0.001 \times d}$ on the ImageNet dataset, where d is the image dimension. Following Li et al. (2021a), we calculate the area under the curve (AUC) of ℓ_2 distortions versus queries.

4.2 Comparison with State-of-the-Art Attacks

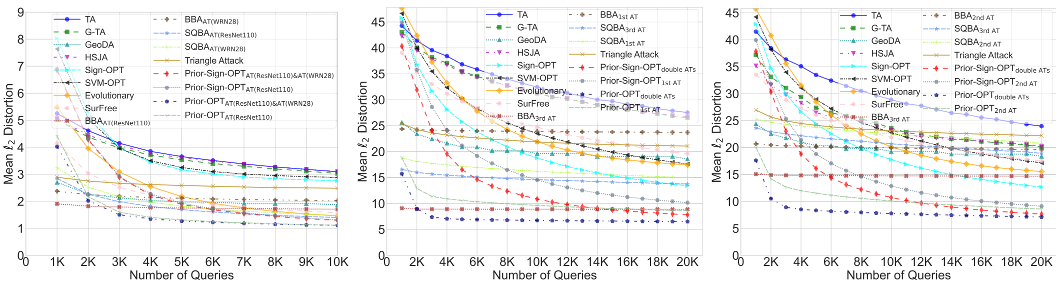
Table 1: Mean ℓ_2 distortions of different query budgets on the ImageNet dataset.

Target Model	Method	Untargeted Attack @1K	@2K	@5K	@8K	@10K	Targeted Attack @1K	@2K	@5K	@8K	@10K	@15K	@20K
Inception-v4	HSJA Chen et al. (2020)	75.392	44.530	20.567	14.194	11.645	95.876	79.001	52.176	39.190	32.951	24.546	19.522
	TA Ma et al. (2021)	67.496	42.233	20.352	14.175	11.694	78.883	61.990	40.669	31.506	27.111	21.079	17.319
	G-TA Ma et al. (2021)	67.842	41.946	19.962	13.865	11.448	79.297	62.291	40.529	30.941	26.427	20.268	16.569
	Sign-OPT Cheng et al. (2020)	86.716	48.233	18.258	11.067	8.786	80.366	65.200	42.866	32.104	27.526	20.394	16.281
	SVM-OPT Cheng et al. (2020)	89.863	47.914	18.297	11.091	8.839	79.807	65.590	43.426	33.090	28.797	22.354	18.795
	GeoDA Rahmati et al. (2020)	29.157	20.119	12.487	11.010	9.688	-	-	-	-	-	-	-
	Evolutionary Dong et al. (2019)	61.966	42.665	20.815	13.382	10.839	81.761	65.060	43.021	32.120	27.385	19.942	15.610
	SurFree Maho et al. (2021)	51.685	38.482	22.845	16.374	13.818	84.925	74.887	55.991	44.475	39.004	29.354	23.153
	Triangle Attack Wang et al. (2022)	27.217	25.853	23.743	22.581	22.132	-	-	-	-	-	-	-
	SQBA$_{\text{IncResV2}}$ Park et al. (2024)	26.134	19.035	11.189	8.432	7.417	-	-	-	-	-	-	-
	SQBA$_{\text{Xception}}$ Park et al. (2024)	23.672	17.424	10.502	8.036	7.115	-	-	-	-	-	-	-
	BBA$_{\text{IncResV2}}$ Brunner et al. (2019)	38.782	28.437	18.757	15.474	14.191	66.746	56.283	41.324	34.066	30.942	25.757	22.630
	BBA$_{\text{Xception}}$ Brunner et al. (2019)	43.317	31.519	20.504	16.712	15.282	63.069	53.363	39.740	33.166	30.221	25.438	22.561
	Prior-Sign-OPT$_{\text{IncResV2}}$	81.991	42.403	12.835	7.365	5.842	74.597	55.421	31.856	22.958	19.513	14.361	11.665
	Prior-Sign-OPT$_{\text{IncResV2\&Xception}}$	77.683	37.099	9.058	5.195	4.199	69.526	49.368	**26.882**	**19.324**	**16.697**	**12.821**	**10.769**
	Prior-Sign-OPT$_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	23.596	15.347	8.074	5.729	4.863	-	-	-	-	-	-	-
	Prior-OPT$_{\text{IncResV2}}$	49.279	18.135	5.718	4.451	4.027	67.300	49.842	33.477	27.602	25.281	21.837	19.800
	Prior-OPT$_{\text{IncResV2\&Xception}}$	42.541	13.418	**3.919**	**3.321**	**3.119**	**60.211**	**42.631**	27.547	23.011	21.441	19.193	17.983
	Prior-OPT$_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	**22.852**	**12.194**	6.568	5.114	4.548	-	-	-	-	-	-	-
ViT	HSJA Chen et al. (2020)	37.813	19.386	9.031	6.604	5.637	61.491	44.853	23.947	16.926	14.152	10.791	8.922
	TA Ma et al. (2021)	37.923	19.867	9.078	6.636	5.674	**52.110**	**36.455**	**20.536**	15.145	12.885	10.158	8.609
	G-TA Ma et al. (2021)	37.425	19.347	8.948	6.496	5.643	52.550	36.720	20.857	15.436	13.255	10.490	8.933
	Sign-OPT Cheng et al. (2020)	51.120	25.290	8.559	5.482	4.572	55.941	41.867	23.784	16.541	13.873	10.129	8.267
	SVM-OPT Cheng et al. (2020)	55.802	26.580	9.242	5.988	5.070	56.002	41.899	23.909	17.273	14.848	11.739	10.320
	GeoDA Rahmati et al. (2020)	18.880	12.904	8.039	7.153	6.313	-	-	-	-	-	-	-
	Evolutionary Dong et al. (2019)	40.382	25.709	11.925	7.974	6.719	57.141	40.187	21.782	15.191	12.795	9.677	8.311
	SurFree Maho et al. (2021)	28.228	19.016	10.194	7.321	6.303	70.337	53.129	30.054	20.595	16.908	11.794	9.204
	Triangle Attack Wang et al. (2022)	**12.789**	12.144	11.064	10.411	10.097	-	-	-	-	-	-	-
	SQBA$_{\text{ResNet50}}$ Park et al. (2024)	21.741	14.004	7.738	5.861	5.201	-	-	-	-	-	-	-
	SQBA$_{\text{ConViT}}$ Park et al. (2024)	12.886	9.762	6.240	4.947	4.452	-	-	-	-	-	-	-
	BBA$_{\text{ResNet50}}$ Brunner et al. (2019)	29.755	20.053	12.580	10.375	9.567	43.231	33.365	21.889	17.635	16.046	13.726	12.463
	BBA$_{\text{ConViT}}$ Brunner et al. (2019)	22.716	16.153	10.893	9.193	8.595	45.588	35.227	22.865	18.325	16.614	14.028	12.623
	Prior-Sign-OPT$_{\text{ResNet50}}$	50.161	27.953	9.474	5.872	4.850	55.095	40.480	22.354	15.626	13.201	9.789	8.048
	Prior-Sign-OPT$_{\text{ResNet50\&ConViT}}$	46.196	23.869	7.327	4.694	3.967	53.925	38.418	20.673	**14.422**	**12.153**	**9.090**	**7.544**
	Prior-Sign-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	29.912	18.425	7.848	5.175	4.331	-	-	-	-	-	-	-
	Prior-OPT$_{\text{ResNet50}}$	42.838	22.704	8.848	6.024	5.195	54.348	40.930	24.408	18.117	15.803	12.638	11.070
	Prior-OPT$_{\text{ResNet50\&ConViT}}$	26.495	**11.287**	**4.929**	**3.937**	**3.609**	53.369	40.002	24.706	19.148	17.116	14.114	12.650
	Prior-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	29.099	17.754	8.208	5.782	5.009	-	-	-	-	-	-	-

Table 2: Mean ℓ_2 distortions of the different numbers of priors on the ImageNet dataset.

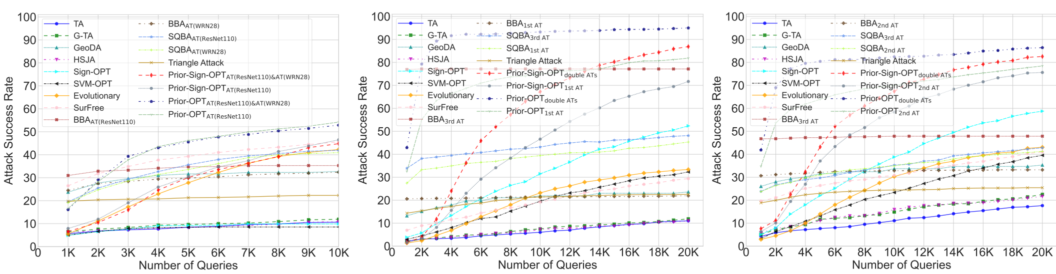
Method	Priors	Target Model: ResNet-101[1]					Target Model: Swin Transformer[2]					Target Model: GC ViT[2]				
		@1K	@2K	@5K	@8K	@10K	@1K	@2K	@5K	@8K	@10K	@1K	@2K	@5K	@8K	@10K
Sign-OPT	no prior	37.248	21.235	8.982	5.811	4.754	86.373	53.399	20.686	12.406	9.899	57.903	35.762	14.763	9.047	7.185
Prior-Sign-OPT	1 prior	34.150	18.733	6.111	3.718	3.019	84.124	52.882	20.344	11.880	9.254	57.171	36.949	14.963	8.931	6.899
	2 priors	32.848	17.548	5.121	3.136	2.593	77.459	43.062	13.614	7.903	6.331	54.896	32.418	11.012	6.651	5.342
	3 priors	31.156	15.455	4.074	2.527	2.122	73.110	37.852	10.264	5.939	4.778	52.744	28.939	8.707	5.245	4.215
	4 priors	29.984	14.707	3.698	2.333	1.989	70.246	34.470	8.526	5.066	4.169	50.256	26.027	6.435	3.804	3.212
	5 priors	**29.601**	**14.195**	**3.573**	**2.275**	**1.951**	**67.616**	**32.225**	**7.321**	**4.219**	**3.467**	**48.935**	**24.821**	**6.123**	**3.601**	**2.893**
Prior-OPT	1 prior	18.355	7.100	2.840	2.324	2.158	69.432	39.447	16.536	11.241	9.625	50.467	29.091	11.537	7.311	5.948
	2 priors	17.373	6.465	2.454	2.096	1.979	41.152	17.977	7.289	5.453	4.896	36.055	16.176	6.094	4.413	3.747
	3 priors	15.373	5.350	1.919	1.714	1.653	**36.636**	13.877	5.166	4.008	3.687	**33.181**	13.005	4.702	3.644	3.264
	4 priors	**15.422**	**5.220**	**1.849**	**1.654**	**1.596**	38.343	12.650	3.784	3.027	2.850	34.396	10.994	3.047	2.356	2.171
	5 priors	15.556	5.395	1.881	1.672	1.605	37.712	**12.070**	**3.488**	**2.747**	**2.577**	33.351	**10.369**	**2.921**	**2.329**	**2.159**

[1] Five surrogate models: ResNet-50, SENet-154, ResNeXt-101 (64 × 4d), VGG-13, SqueezeNet v1.1
[2] Five surrogate models: ResNet-50, ConViT, CrossViT, MaxViT, ViT



(a) TRADES$_{\epsilon_\infty=8/255}$ (CIFAR-10) (b) AT$_{\epsilon_\infty=4/255}$ (ImageNet)[1] (c) AT$_{\epsilon_\infty=8/255}$ (ImageNet)[2]

Figure 3: Mean distortions of untargeted attacks on the defense models equipped with the ResNet-50.



(a) TRADES (CIFAR-10) (b) AT$_{\epsilon_\infty=4/255}$ (ImageNet)[1] (c) AT$_{\epsilon_\infty=8/255}$ (ImageNet)[2]

Figure 4: Attack success rates of untargeted attacks with ℓ_2 norm constraint against defense models.

Results of Attacks against Undefended Models. Table 1 shows the results of attacks against undefended models on the ImageNet dataset. Additional results are in Appendix F.2. In summary:

(1) In untargeted attacks (Table 1 and Fig. 7), the performance of Prior-OPT significantly surpasses that of all methods, and using multiple surrogate models performs better than using a single surrogate model. In addition, the PGD initialization of θ (θ_0^{PGD}) is effective at the beginning of attacks, because it ensures the algorithm's initial attack direction θ_0 is already good by using transfer-based attacks.

(2) In targeted attacks, Prior-OPT outperforms Prior-Sign-OPT when the query budget is below $5{,}000$, while Prior-Sign-OPT performs better in later iterations with more queries.

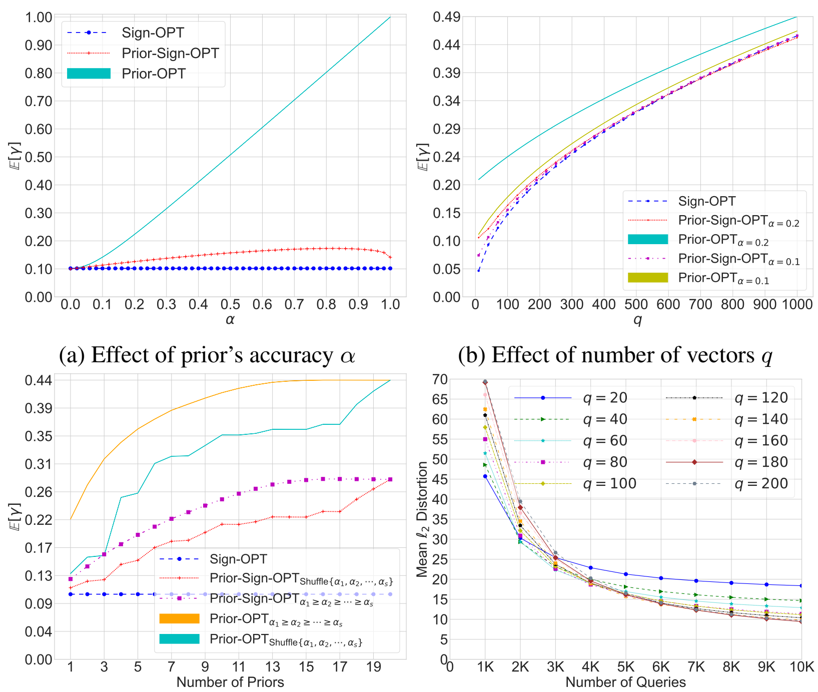
(3) Table 2 and Fig. 5c demonstrate that using more surrogate models (priors) can boost performance.

[1] 1st AT: AT(ResNet-50, $\epsilon_{\ell_\infty} = 8/255$), 3rd AT: AT(ResNet-50, $\epsilon_{\ell_2} = 3$), double ATs: combination of both
[2] 2nd AT: AT(ResNet-50, $\epsilon_{\ell_\infty} = 4/255$), 3rd AT: AT(ResNet-50, $\epsilon_{\ell_2} = 3$), double ATs: combination of both

Results of Attacks against Defense Models. We conduct the experiments on untargeted attacks against two types of defense models, *i.e.,* adversarial training (AT) Madry et al. (2018) and TRADES Zhang et al. (2019). Figs. 3 and 4 show that Prior-OPT with two surrogate models (Prior-OPT$_{\text{double ATs}}$) performs the best on the ImageNet dataset and the CIFAR-10 dataset.

4.3 Comprehensive Understanding of Prior-OPT

In the ablation studies, we perform complete control experiments based on theoretical analysis result and actual image attacks (Fig. 5). In Figs 5a, 5b, and 5c, we set the dimension of image $d = 3072$ and use $\mathbb{E}[\gamma]$ (Eq. (9) for Sign-OPT, Eq. (11) for Prior-Sign-OPT, Eq. (15) and Eq. (16) for the lower and upper bound of Prior-OPT) as the metric of gradient estimation's accuracy, where $\gamma = \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$. Figs. 5a, 5c are based on $q = 50$. Fig. 5a uses one prior and shows that Prior-OPT and Prior-Sign-OPT outperform Sign-OPT with different α. Fig. 5a also shows that Prior-Sign-OPT performs well when α is small and $\mathbb{E}[\gamma]$ decreases when α is close to 1. This is because when we set $\alpha = 1$, $\mathbb{E}[\gamma] = 1/\sqrt{q}$ in Eq. (11). Fig. 5b shows that $\mathbb{E}[\gamma]$ monotonically increases with q for each method, and Prior-Sign-OPT performs worse than Sign-OPT when $q > 500$. Fig. 5c validates that the performance can be improved when more priors are available, and using surrogate models with larger α values first yields better results. Fig. 5c is consistent with the conclusion of the experimental results in Table 2. Fig. 5d shows the untargeted attack results of Prior-OPT against Swin Transformer with varying q on ImageNet. In the early stage of iterations, a lower value of q leads to better performance. However, in the late stage of iterations with large queries, a smaller value of q performs worse.



(a) Effect of prior's accuracy α

(b) Effect of number of vectors q

(c) Effect of number of priors

(d) Prior-OPT with different q

Figure 5: Results of ablation experiments. Figs. 5a, 5b, 5c are based on Eqs. (9), (11), (15), (16).

5 Conclusion

In this paper, we propose novel hard-label attacks (*i.e.,* Prior-OPT and Prior-Sign-OPT) that incorporate transfer-based priors into the gradient estimation of ray's direction, which significantly boost the attack performance. Through theoretical analysis, the effectiveness of our approach is proved: we provide solutions for the expectations of the similarity between the estimated gradients and the true gradient, allowing for theoretical comparison with the baseline. Therefore, we gain a comprehensive understanding of Prior-OPT and Prior-Sign-OPT. Lastly, we evaluate our approach through extensive experiments and show its superior performance compared with state-of-the-art methods.

Ethics Statement

We affirm our commitment to the ICLR Code of Ethics, ensuring that our research on adversarial examples and AI security adheres to the highest ethical standards. Our work aims to identify potential vulnerabilities in AI systems, with the intention of enhancing their security and resilience. Specifically, our method can be integrated into the evaluation process for a model's robustness, enabling the study and implementation of targeted defense strategies, and providing effective support in strengthening the security specifications of artificial intelligence models. We acknowledge the dual-use nature of this research and have taken steps to responsibly communicate our findings, encouraging their use for developing robust defense mechanisms. We remain open to discussions about any ethical concerns that may arise and are dedicated to contributing positively to the field of AI security.

Reproducibility Statement

We have taken several steps to ensure the reproducibility of our research. The appendix provides comprehensive details on the algorithm settings, computational resources, and theoretical proofs. Additionally, the supplementary materials include the code necessary to reproduce our main evaluation results. We have also included thorough descriptions of our methodology and referenced all key components of our approach within the main text. To further support reproducibility, we will provide the complete attack code for our method and all the baseline methods upon acceptance of the paper. These resources are intended to enable others to replicate our experiments.

References

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, pp. 1–12, 2018.

T. Brunner, F. Diehl, M. T. Le, and A. Knoll. Guessing smart: biased sampling for efficient black-box adversarial attacks. In *IEEE/CVF International Conference on Computer Vision*, pp. 4957–4965, 2019.

Jianbo Chen, Jordan Michael I., and Wainwright Martin J. HopSkipJumpAttack: a query-efficient decision-based adversarial attack. In *IEEE Symposium on Security and Privacy*, pp. 1277–1294, 2020.

Jinghui Chen and Quanquan Gu. RayS: a ray searching method for hard-label adversarial attack. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1739–1747, 2020.

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: an optimization-based approach. In *International Conference on Learning Representations*, pp. 1–14, 2019a.

Minhao Cheng, Simranjit Singh, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: a query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, pp. 1–16, 2020.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, pp. 10934–10944, 2019b.

Shuyu Cheng, Guoqiang Wu, and Jun Zhu. On the convergence of prior-guided zeroth-order optimization algorithms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14620–14631. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/7aaece81f2d731fbf8ee0ad3521002ac-Paper.pdf`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722, 2019.

Yinpeng Dong, Shuyu Cheng, Tianyu Pang, Hang Su, and Jun Zhu. 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. doi: 10.1109/TPAMI.2021.3126733.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pp. 2286–2296. PMLR, 2021.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5927–5935, 2017.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pp. 12633–12646. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009.

Jie Li, Rongrong Ji, Peixian Chen, Baochang Zhang, Xiaopeng Hong, Ruixin Zhang, Shaoxin Li, Jilin Li, Feiyue Huang, and Yongjian Wu. Aha! adaptive history-driven attack for decision-based black-box models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16168–16177, 2021a.

Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su, and Hui Xue. QAIR: Practical query-efficient black-box attacks for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3330–3339, 2021b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Chen Ma, Xiangyu Guo, Li Chen, Jun-Hai Yong, and Yisen Wang. Finding optimal tangent points for reducing distortions of hard-label attacks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=g0wang64Zjd`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, pp. 1–28, 2018.

Thibault Maho, Teddy Furon, and Erwan Le Merrer. SurFree: a fast surrogate-free black-box attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10430–10439, 2021.

Florian Meier, Asier Mujika, Marcelo Matheus Gauy, and Angelika Steger. Improving gradient estimation in evolutionary strategies with past descent directions. *arXiv preprint arXiv:1910.05268*, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Jeonghwan Park, Paul Miller, and Niall McLaughlin. Hard-label based small query black-box adversarial attack. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3986–3995, January 2024.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: a geometric framework for black-box adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8446–8455, 2020.

Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, and Huaiyu Dai. Cgba: Curvature-aware geometric black-box attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 124–133, 2023.

Oswaldo Rio Branco de Oliveira. The implicit and the inverse function theorems: easy proofs. *arXiv e-prints*, pp. arXiv–1212, 2012.

Yucheng Shi, Yahong Han, Qinghua Hu, Yi Yang, and Qi Tian. Query-efficient black-box adversarial attack with customized iteration and sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2226–2245, 2023. doi: 10.1109/TPAMI.2022.3169802.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, pp. 4278–4284, 2017.

Michail Tsagris, Christina Beneki, and Hossein Hassani. On the folded normal distribution. *Mathematics*, 2(1):12–28, 2014.

Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *European Conference on Computer Vision*, pp. 156–174. Springer, 2022.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. Shakedrop regularization for deep residual learning. *IEEE Access*, 7:186126–186136, 2019.

Ziang Yan, Yiwen Guo, Jian Liang, and Changshui Zhang. Policy-driven attack: learning to query for hard-label black-box adversarial examples. In *International Conference on Learning Representations*, pp. 1–15, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pp. 87.1–87.12, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

APPENDIX

A THEORETICAL ANALYSIS OF SIGN-OPT, PRIOR-SIGN-OPT AND PRIOR-OPT

Lemma A.1. *Suppose* $\mathbf{u} \sim \mathcal{U}(\mathbb{S}_{d-1})$ *where* $\mathcal{U}(\mathbb{S}_{d-1})$ *denotes the uniform distribution on the unit sphere in* $\mathbb{R}^d$*. Suppose* $\mathbf{g}$ *is a fixed vector in* $\mathbb{R}^d$ *with* $\|\mathbf{g}\| = 1$*. Let* $\beta = \mathbf{u}^\top \mathbf{g}$*. Then*

$$\mathbb{E}[|\beta|] = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}}, \tag{18}$$

$$\mathbb{E}[\beta^2] = \frac{1}{d}, \tag{19}$$

where Γ *is the gamma function.*

Proof. Let $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^d$, then we can let $\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$. Hence $\beta = \frac{\mathbf{a}^\top \mathbf{g}}{\|\mathbf{a}\|}$. We note that $\frac{\mathbf{a}}{\|\mathbf{a}\|}$ and $\|\mathbf{a}\|$ are independent because the distribution of $\frac{\mathbf{a}}{\|\mathbf{a}\|}$ is always the uniform distribution on the unit sphere given any restriction to the value of $\|\mathbf{a}\|$. Therefore $\beta = \frac{\mathbf{a}^\top \mathbf{g}}{\|\mathbf{a}\|}$ and $\|\mathbf{a}\|$ are also independent, so $|\beta|$ and $\|\mathbf{a}\|$ are independent. Noting that $|\beta|\|\mathbf{a}\| = |\mathbf{a}^\top \mathbf{g}|$, we have

$$\mathbb{E}[|\mathbf{a}^\top \mathbf{g}|] = \mathbb{E}[|\beta|]\mathbb{E}[\|\mathbf{a}\|]. \tag{20}$$

Since $\mathbf{a}^\top \mathbf{g}$ is a affine transformation of the multivariate Gaussian variable $\mathbf{a}$, $\mathbf{a}^\top \mathbf{g}$ also has a Gaussian distribution with the mean 0 and the variance $\mathbf{g}^\top \mathbf{I} \mathbf{g} = 1$, so $\mathbf{a}^\top \mathbf{g} \sim \mathcal{N}(0, 1)$. Therefore, $|\mathbf{a}^\top \mathbf{g}|$ follows the folded normal distribution (actually its special case: half-normal distribution), and by the formula in Tsagris et al. (2014),

$$\mathbb{E}[|\mathbf{a}^\top \mathbf{g}|] = \frac{\sqrt{2}}{\sqrt{\pi}}. \tag{21}$$

$\|\mathbf{a}\|$ follows the chi distribution with d degrees of freedom, so by the formula of its mean

$$\mathbb{E}[\|\mathbf{a}\|] = \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}. \tag{22}$$

Substituting Eq. (21) and Eq. (22) into Eq. (20), we have proved Eq. (18).

Since β and $\mathbf{a}$ are independent, similarly to Eq. (20), we have

$$\mathbb{E}[(\mathbf{a}^\top \mathbf{g})^2] = \mathbb{E}[\beta^2]\mathbb{E}[\|\mathbf{a}\|^2]. \tag{23}$$

Since $\mathbf{a} \sim \mathcal{N}(0, \mathbb{I})$, $\mathbb{E}[\mathbf{a}\mathbf{a}^\top] = \mathbb{I}$. Hence $\mathbb{E}[(\mathbf{a}^\top \mathbf{g})^2] = \mathbb{E}[\mathbf{g}^\top \mathbf{a} \cdot \mathbf{a}^\top \mathbf{g}] = \mathbf{g}^\top \mathbb{E}[\mathbf{a}\mathbf{a}^\top]\mathbf{g} = \|\mathbf{g}\|^2 = 1$, and $\mathbb{E}[\|\mathbf{a}\|^2] = \mathbb{E}[\mathrm{Tr}(\mathbf{a}\mathbf{a}^\top)] = \mathrm{Tr}(\mathbb{E}[\mathbf{a}\mathbf{a}^\top]) = d$. By Eq. (23), Eq. (19) has been proved.

□

Lemma A.2. *Suppose* $\mathbf{g}$ *is a fixed vector in* $\mathbb{R}^d$ *with* $\|\mathbf{g}\| = 1$*. Suppose* $\mathbf{p}$ *is another fixed vector in* $\mathbb{R}^d$ *with* $\|\mathbf{p}\| = 1$*, and let* $\beta_p := \mathbf{p}^\top \mathbf{g}$*. Let* $\mathbf{u}$ *be a random vector uniformly sampled from the unit hypersphere in the* $(d-1)$*-dimensional space orthogonal to* $\mathbf{p}$*. Specifically,* $\mathbf{u}$ *can be constructed as* $\mathbf{u} = \overline{\boldsymbol{\xi} - \boldsymbol{\xi}^\top \mathbf{p} \cdot \mathbf{p}}$ *where* $\boldsymbol{\xi} \sim \mathcal{U}(\mathbb{S}_{d-1})$*. Let* $\beta_\perp := \mathbf{u}^\top \mathbf{g}$*, then*

$$\mathbb{E}[|\beta_\perp|] = \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}}\sqrt{1-\beta_p^2}. \tag{24}$$

Proof. To observe the property of $\beta_\perp$, since $\mathbf{u}$ is orthogonal to $\mathbf{p}$, we do the following decomposition for $\mathbf{g}$:

$$\mathbf{g} = \mathbf{g}^\top \mathbf{p} \cdot \mathbf{p} + \mathbf{g}_\perp = \beta_p \mathbf{p} + \mathbf{g}_\perp, \tag{25}$$

where $\mathbf{g}_\perp := \mathbf{g} - \mathbf{g}^\top \mathbf{p} \cdot \mathbf{p}$ denotes the projection of $\mathbf{g}$ to the $(d-1)$-dimensional space orthogonal to $\mathbf{p}$. By expanding the inner product we have

$$\|\mathbf{g}_\perp\|^2 = 1 - 2\beta_p^2 + \beta_p^2 = 1 - \beta_p^2, \tag{26}$$

so $\|\mathbf{g}_\perp\| = \sqrt{1-\beta_p^2}$. Meanwhile,

$$\beta_\perp = \mathbf{g}^\top \mathbf{u} = (\mathbf{g}^\top \mathbf{p} \cdot \mathbf{p} + \mathbf{g}_\perp)^\top \mathbf{u} = \mathbf{g}_\perp^\top \mathbf{u}. \tag{27}$$

Therefore, $\beta_\perp$ is essentially the inner product between a random vector uniformly sampled from the unit hypersphere in a $(d-1)$-dimensional space and a fixed vector of norm $\sqrt{1-\beta_p^2}$ in this space. In Eq. (18), replacing d with $d-1$ and multiplying the result by the norm $\sqrt{1-\beta_p^2}$, the proof is completed. □

Lemma A.3. *Let β be as defined in Lemma A.1, then the probability density function of β is (note that $-1 \le \beta \le 1$)*

$$p(\beta) = \frac{(\sqrt{1-\beta^2})^{d-3}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})}, \tag{28}$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function.

Proof. We note that when $-1 \le x \le 0$, $P(\beta \le x)$ is equal to the ratio of the surface area of the hyperspherical cap of a hyper sphere in $\mathbb{R}^d$ to the surface area of the hyper sphere. For a hyperspherical cap with height h on a unit hypersphere, its surface area is $\frac{1}{2}A_d I_{2h-h^2}(\frac{d-1}{2}, \frac{1}{2})$, where A_d is the surface area of the unit hypersphere in $\mathbb{R}^d$ and $I(\cdot,\cdot)$ is the regularized incomplete beta function. To compute $P(\beta \le x)$ for $-1 \le x \le 0$, substituting $h = x+1$ and dividing the area by A_d, we have

$$P(\beta \le x) = \frac{1}{2} I_{1-x^2}(\frac{d-1}{2}, \frac{1}{2}), \tag{29}$$

where I is the regularized incomplete beta function, defined as

$$I_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t}{\mathrm{B}(a,b)}. \tag{30}$$

Hence the probability density function is

$$\begin{aligned} p_\beta(x) &= \frac{\partial}{\partial x} P(\beta \le x) && (31)\\ &= \frac{1}{2} \frac{-2x}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} (1-x^2)^{\frac{d-1}{2}-1} (1-(1-x^2))^{\frac{1}{2}-1} && (32)\\ &= \frac{-x}{|x|} \frac{(\sqrt{1-x^2})^{d-3}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} && (33)\\ &= \frac{(\sqrt{1-x^2})^{d-3}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})}. && (34) \end{aligned}$$

Note that the last equality is because $x \le 0$.

Therefore, we have proven Eq. (28) for $\beta \le 0$. When $\beta > 0$ the formula is the same due to the symmetry. The proof is completed. □

A.1 Analysis for Sign-OPT

We can compute the projection of $\nabla g(\theta)$ onto S with $s = 0$ by summing over all its projection onto the orthonormal basis:

$$\mathbf{v} = \sum_{i=1}^{q} \frac{g(\theta + \sigma \mathbf{u}_i) - g(\theta)}{\sigma} \cdot \mathbf{u}_i, \tag{35}$$

where $\{\mathbf{u}_1, \ldots, \mathbf{u}_q\}$ is a randomly uniformly distributed orthonormal set of q vectors in $\mathbb{R}^d$, so $\mathbf{u}_i \sim \mathcal{U}(\mathbb{S}_{d-1})$ for any $i \leq q$. However, in hard-label attacks the coefficients above for each basis vector are costly to estimate. In the Sign-OPT estimator, each coefficient is replaced by its sign which is much easier to obtain using hard-label queries:

$$\mathbf{v} = \sum_{i=1}^{q} \text{sign}(g(\theta + \sigma \mathbf{u}_i) - g(\theta)) \cdot \mathbf{u}_i. \tag{36}$$

In the following analysis, we assume that g is differentiable at θ so that we have $g(\theta + \sigma \mathbf{u}) - g(\theta) = \sigma \cdot \nabla g(\theta)^\top \mathbf{u} + o(\sigma)$ where $\lim_{\sigma \to 0} \frac{o(\sigma)}{\sigma} = 0$ for any unit vector $\mathbf{u}$. We further assume that σ is sufficiently small so that we can omit $o(\sigma)$. In practice, if the target model is deterministic, picking a small σ is feasible until the numerical error dominates. Therefore, in the following analysis we assume that

$$\text{sign}(g(\theta + \sigma \mathbf{u}) - g(\theta)) \approx \text{sign}(\nabla g(\theta)^\top \mathbf{u}) \tag{37}$$

where $\mathbf{u}$ is a unit vector in $\mathbb{R}^d$. Now we can write Sign-OPT in the following form:

$$\mathbf{v} = \sum_{i=1}^{q} \text{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i. \tag{38}$$

Now, we present the proof of Theorem 3.2 for the Sign-OPT estimator defined in Eq. (38).

Proof. Since $\{\mathbf{u}_i\}_{i=1}^q$ are orthonormal, we have $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^q (\text{sign}(\nabla g(\theta)^\top \mathbf{u}_i))^2} = \sqrt{q}$. We note that $\text{sign}(\nabla g(\theta)^\top \mathbf{u}_i) = \text{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i)$, so

$$\mathbf{v} = \sum_{i=1}^{q} \text{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i) \cdot \mathbf{u}_i. \tag{39}$$

Hence

$$\gamma = \frac{\mathbf{v}^\top \overline{\nabla g(\theta)}}{\|\mathbf{v}\|} = \frac{1}{\sqrt{q}} \sum_{i=1}^{q} \text{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i) \cdot (\overline{\nabla g(\theta)}^\top \mathbf{u}_i) \tag{40}$$

$$= \frac{1}{\sqrt{q}} \sum_{i=1}^{q} |\overline{\nabla g(\theta)}^\top \mathbf{u}_i|. \tag{41}$$

Since $\nabla f(\theta)$ is a fixed vector w.r.t. the randomness of $\{\mathbf{u}_i\}_{i=1}^q$, and the marginal distribution of $\mathbf{u}_i$ is $\mathcal{U}(\mathbb{S}_{d-1})$ for any i, by Eq. (18) we have

$$\mathbb{E}[\gamma] = \frac{1}{\sqrt{q}} q \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}} = \sqrt{q} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}}. \tag{42}$$

Computing $\mathbb{E}[\gamma^2]$ is more complicated. First we have

$$\gamma^2 = \frac{(\mathbf{v}^\top \overline{\nabla g(\theta)})^2}{\|\mathbf{v}\|^2} = \frac{1}{q} (\sum_{i=1}^{q} |\overline{\nabla g(\theta)}^\top \mathbf{u}_i|)^2 \tag{43}$$

$$= \frac{1}{q} \sum_{i=1}^{q} (\overline{\nabla g(\theta)}^\top \mathbf{u}_i)^2 + \frac{1}{q} \sum_{i \neq j} |\overline{\nabla g(\theta)}^\top \mathbf{u}_i| \cdot |\overline{\nabla g(\theta)}^\top \mathbf{u}_j| \tag{44}$$

For the first part, by Eq. (19) we have

$$\forall i, \mathbb{E}[(\overline{\nabla g(\theta)}^\top \mathbf{u}_i)^2] = \frac{1}{d}. \tag{45}$$

For the second part, let us denote $\beta_i := \overline{\nabla g(\theta)}^\top \mathbf{u}_i$ and $\beta_j := \overline{\nabla g(\theta)}^\top \mathbf{u}_j$. Then we need to compute for $i \neq j$:

$$\mathbb{E}[|\beta_i| \cdot |\beta_j|] = \mathbb{E}_{\mathbf{u}_i}[\mathbb{E}[(|\beta_i| \cdot |\beta_j|) \big| \mathbf{u}_i]] \tag{46}$$

$$= \mathbb{E}_{\mathbf{u}_i}[|\beta_i| \mathbb{E}[|\beta_j| \big| \mathbf{u}_i]]. \tag{47}$$

Next we should compute $\mathbb{E}[|\beta_j| \big| \mathbf{u}_i]$. Since $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthonormal and by the sampling symmetry, given $\mathbf{u}_i$, the distribution of $\mathbf{u}_j$ is uniform on the unit hypersphere in the $(d-1)$-dimensional space orthogonal to $\mathbf{u}_i$. When calculating the conditional expectation, we consider $\mathbf{u}_i$ to be fixed and use Lemma A.2. Specifically, in Lemma A.2 we let $\mathbf{g}$ be $\overline{\nabla g(\theta)}$ and let $\mathbf{p}$ be $\mathbf{u}_i$. Then we have

$$\mathbb{E}[|\beta_j| \big| \mathbf{u}_i] = \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}} \sqrt{1-\beta_i^2}. \tag{48}$$

Substituting Eq. (48) into Eq. (47), we have

$$\mathbb{E}[|\beta_i| \cdot |\beta_j|] = \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}} \mathbb{E}[|\beta_i|\sqrt{1-\beta_i^2}]. \tag{49}$$

Here the distribution of β_i is the same as that of β in Lemma A.1, and we need to compute $\mathbb{E}[|\beta|\sqrt{1-\beta^2}]$. By Eq. (18) and Eq. (28), we have

$$\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}} = \mathbb{E}[|\beta|] = \int_{-1}^{1} p(\beta)|\beta| \mathrm{d}\beta \tag{50}$$

$$= \int_{-1}^{1} |\beta| \frac{(\sqrt{1-\beta^2})^{d-3}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} \mathrm{d}\beta, \tag{51}$$

so

$$\int_{-1}^{1} |\beta|(\sqrt{1-\beta^2})^{d-3} \mathrm{d}\beta = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}} \mathrm{B}(\frac{d-1}{2}, \frac{1}{2}). \tag{52}$$

Hence

$$\mathbb{E}[|\beta|\sqrt{1-\beta^2}] = \int_{-1}^{1} |\beta|\sqrt{1-\beta^2} p(\beta) \mathrm{d}\beta \tag{53}$$

$$= \frac{1}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} \int_{-1}^{1} |\beta|(\sqrt{1-\beta^2})^{d-2} \mathrm{d}\beta \tag{54}$$

$$= \frac{1}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d+2}{2})\sqrt{\pi}} \mathrm{B}(\frac{d}{2}, \frac{1}{2}), \tag{55}$$

where the last equality is obtained by setting d in Eq. (52) to $d+1$. Therefore, by Eq. (49) we have

$$\mathbb{E}[|\beta_i| \cdot |\beta_j|] = \frac{1}{\pi} \frac{\mathrm{B}(\frac{d}{2}, \frac{1}{2})}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d+2}{2})} \tag{56}$$

$$= \frac{1}{\pi} \frac{\Gamma(\frac{d}{2})\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{d-1}{2})} \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d+2}{2})} \tag{57}$$

$$= \frac{1}{\pi} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+2}{2})} \tag{58}$$

$$= \frac{2}{\pi d}. \tag{59}$$

Here, the second equality is due to the identity $\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, and the last equality is due to the identity $\Gamma(a+1) = a\Gamma(a)$.

Taking expectation to both sides of Eq. (44) and using Eq. (45) and Eq. (59), we have

$$\mathbb{E}[\gamma^2] = \frac{1}{q} \cdot q \cdot \frac{1}{d} + \frac{1}{q} \cdot q(q-1) \cdot \frac{2}{\pi d} \tag{60}$$

$$= \frac{1}{d} + \frac{2(q-1)}{\pi d} = \frac{1}{d}(\frac{2}{\pi}(q-1)+1). \tag{61}$$

The proof is completed. $\square$

A.2 Analysis for Prior-Sign-OPT

The Prior-Sign-OPT estimator is defined in Eq. (7). Note that there are s priors $\{\mathbf{p}_1, \cdots, \mathbf{p}_s\}$ (we assume that they have been normalized so that they have unit norm), and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q-s}\}$ is a randomly uniformly distributed orthonormal set of $q-s$ vectors in the $(d-s)$-dimensional subspace orthogonal to $\{\mathbf{p}_1, \cdots, \mathbf{p}_s\}$. For convenience we first consider the case of $s=1$, and the analysis could be easily generalized to the case of $s>1$.

A.2.1 The Case of $s=1$

When $s=1$, we write the Prior-Sign-OPT estimator in the following form:

$$\mathbf{v}^* = \mathrm{sign}(\nabla g(\theta)^\top \mathbf{p}) \cdot \mathbf{p} + \sum_{i=1}^{q-1} \mathrm{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i, \tag{62}$$

where $\mathbf{p}$ is the prior vector with $\|\mathbf{p}\|=1$, and $\{\mathbf{u}_i\}_{i=1}^{q-1}$ are the $q-1$ random orthonormal basis of the $(d-1)$-dimensional space orthogonal to $\mathbf{p}$. Note that we also employ the directional derivative approximation as in Eq. (37).

Theorem A.4. *For the Prior-Sign-OPT estimator defined in Eq. (62), we let $\gamma = \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ be its cosine similarity to the true gradient, where the notation $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$ is defined to be the ℓ_2 normalization of the corresponding vector, then*

$$\mathbb{E}[\gamma] = \frac{1}{\sqrt{q}} \left[|\alpha| + (q-1)\sqrt{1-\alpha^2} \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}} \right], \tag{63}$$

$$\mathbb{E}[\gamma^2] = \frac{1}{q} \left[\alpha^2 + \frac{q-1}{d-1}(\frac{2}{\pi}(q-2)+1)(1-\alpha^2) + 2|\alpha|(q-1)\sqrt{1-\alpha^2} \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}} \right], \tag{64}$$

where $\alpha := \mathbf{p}^\top \overline{\nabla g(\theta)}$ is the cosine similarity between the prior and the true gradient.

Proof. Note that the property of sign function (*e.g.* $\mathrm{sign}(\nabla g(\theta)^\top \mathbf{p}) = \mathrm{sign}(\overline{\nabla g(\theta)}^\top \mathbf{p})$), in Eq. (62), we denote

$$\mathbf{v}_\perp := \sum_{i=1}^{q-1} \mathrm{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i) \cdot \mathbf{u}_i, \tag{65}$$

then $\mathbf{v}^* = \mathrm{sign}(\alpha)\mathbf{p} + \mathbf{v}_\perp$. Now

$$\gamma = \frac{(\mathbf{v}^*)^\top \overline{\nabla g(\theta)}}{\|\mathbf{v}^*\|} \tag{66}$$

$$= \frac{1}{\sqrt{q}}(|\alpha| + \mathbf{v}_\perp^\top \overline{\nabla g(\theta)}). \tag{67}$$

The following argument is similar to that in the proof of Lemma A.2. Let $\mathbf{g} := \overline{\nabla g(\theta)}$, and let $\mathbf{g}_\perp := \mathbf{g} - \mathbf{g}^\top \mathbf{p} \cdot \mathbf{p}$ denotes the projection of $\mathbf{g}$ to the $(d-1)$-dimensional space orthogonal to $\mathbf{p}$. Then $\mathbf{v}_\perp^\top \mathbf{g} = \mathbf{v}_\perp^\top \mathbf{g}_\perp$. Meanwhile, since $\{\mathbf{u}_i\}_{i=1}^{q-1}$ are orthogonal to $\mathbf{p}$, $\mathbf{v}_\perp = \sum_{i=1}^{q-1} \mathrm{sign}(\mathbf{g}_\perp^\top \mathbf{u}_i) \cdot \mathbf{u}_i$. Since $\{\mathbf{u}_i\}_{i=1}^{q-1}$ are uniformly distributed on the unit hypersphere in the $(d-1)$-dimensional space orthogonal to $\mathbf{p}$ and $\mathbf{g}_\perp$ also lives in this space, $\mathbf{v}_\perp$ could be viewed as the Sign-OPT estimator for $\mathbf{g}_\perp$ as in Eq. (38), while q should be replaced with $q-1$ and the effective dimension is $d-1$ instead of d. By Eq. (9), we have

$$\mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp}] = \sqrt{q-1} \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}}. \tag{68}$$

Noting that $\|\mathbf{g}_\perp\| = \sqrt{1-\alpha^2}$ by Eq. (26) and $\|\mathbf{v}_\perp\| = \sqrt{q-1}$, we have

$$\mathbb{E}[\mathbf{v}_\perp^\top \mathbf{g}] = \mathbb{E}[\mathbf{v}_\perp^\top \mathbf{g}_\perp] = \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp} \|\mathbf{v}_\perp\| \|\mathbf{g}_\perp\|] \tag{69}$$

$$= (q-1)\sqrt{1-\alpha^2}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}}. \tag{70}$$

Taking expectation to both sides of Eq. (67) and plugging in Eq. (70), Eq. (63) has been proved.

Next we derive $\mathbb{E}[\gamma^2]$. By Eq. (67) we have

$$\gamma^2 = \frac{1}{q}(|\alpha| + \mathbf{v}_\perp^\top \mathbf{g})^2 \tag{71}$$

$$= \frac{1}{q}(\alpha^2 + (\mathbf{v}_\perp^\top \mathbf{g})^2 + 2|\alpha|\mathbf{v}_\perp^\top \mathbf{g}). \tag{72}$$

As mentioned in the discussion before Eq. (68), $\mathbf{v}_\perp$ could be viewed as the Sign-OPT estimator for $\mathbf{g}_\perp$ as in Eq. (38), while q should be replaced with $q-1$ and the effective dimension is $d-1$ instead of d. By Eq. (10), we have

$$\mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp})^2] = \frac{1}{d-1}(\frac{2}{\pi}(q-2)+1). \tag{73}$$

Noting that $\|\mathbf{g}_\perp\| = \sqrt{1-\alpha^2}$ by Eq. (26) and $\|\mathbf{v}_\perp\| = \sqrt{q-1}$, we have

$$\mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g})^2] = \mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g}_\perp)^2] = \mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp})^2 \|\mathbf{v}_\perp\|^2 \|\mathbf{g}_\perp\|^2] \tag{74}$$

$$= \frac{q-1}{d-1}(\frac{2}{\pi}(q-2)+1)(1-\alpha^2). \tag{75}$$

Taking expectation to both sides of Eq. (72) and plugging in Eq. (70) and Eq. (75), Eq. (64) has been proved. □

A.2.2 The Case of $s > 1$

In the case of $s > 1$, we write the Prior-Sign-OPT estimator in the following form:

$$\mathbf{v}^* = \sum_{i=1}^{s} \mathrm{sign}(\nabla g(\theta)^\top \mathbf{p}_i) \cdot \mathbf{p}_i + \sum_{i=1}^{q-s} \mathrm{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i, \tag{76}$$

Now, we present the proof of Theorem 3.3 for the Prior-Sign-OPT estimator defined in Eq. (76).

Proof. The following argument is similar to that in the proof of Theorem A.4. Now we have

$$\mathbf{v}_\perp := \sum_{i=1}^{q-s} \mathrm{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i) \cdot \mathbf{u}_i, \tag{77}$$

then $\mathbf{v}^* = \sum_{i=1}^{s} \mathrm{sign}(\alpha_i)\mathbf{p}_i + \mathbf{v}_\perp$. And

$$\gamma = \frac{(\mathbf{v}^*)^\top \overline{\nabla g(\theta)}}{\|\mathbf{v}^*\|} \tag{78}$$

$$= \frac{1}{\sqrt{q}}(\sum_{i=1}^{s}|\alpha_i| + \mathbf{v}_\perp^\top \overline{\nabla g(\theta)}). \tag{79}$$

Let $\mathbf{g} := \overline{\nabla g(\theta)}$, and let $\mathbf{g}_\perp := \mathbf{g} - \sum_{i=1}^{s} \mathbf{g}^\top \mathbf{p}_i \cdot \mathbf{p}_i$ denotes the projection of $\mathbf{g}$ to the $(d-s)$-dimensional space orthogonal to $\{\mathbf{p}_i\}_{i=1}^{s}$. Then $\mathbf{v}_\perp^\top \mathbf{g} = \mathbf{v}_\perp^\top \mathbf{g}_\perp$. Using the similar analysis to that in the case of $s=1$, when $s>1$ we have

$$\mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp}] = \sqrt{q-s}\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}. \tag{80}$$

Similar to the derivation of Eq. (26), we can derive that $\|\mathbf{g}_\perp\| = \sqrt{1-\sum_{i=1}^s \alpha_i^2}$. Since $\|\mathbf{v}_\perp\| = \sqrt{q-s}$, we have

$$\mathbb{E}[\mathbf{v}_\perp^\top \mathbf{g}] = \mathbb{E}[\mathbf{v}_\perp^\top \mathbf{g}_\perp] = \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp} \|\mathbf{v}_\perp\| \|\mathbf{g}_\perp\|] \tag{81}$$

$$= (q-s)\sqrt{1-\sum_{i=1}^s \alpha_i^2} \frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}. \tag{82}$$

Taking expectation to both sides of Eq. (79) and plugging in Eq. (82), Eq. (11) has been proved.

Next we derive $\mathbb{E}[\gamma^2]$. By Eq. (79) we have

$$\gamma^2 = \frac{1}{q}(\sum_{i=1}^s |\alpha_i| + \mathbf{v}_\perp^\top \mathbf{g})^2 \tag{83}$$

$$= \frac{1}{q}((\sum_{i=1}^s |\alpha_i|)^2 + (\mathbf{v}_\perp^\top \mathbf{g})^2 + 2 \cdot (\sum_{i=1}^s |\alpha_i|) \cdot \mathbf{v}_\perp^\top \mathbf{g}). \tag{84}$$

Similar to the case of $s = 1$, when $s > 1$ we have

$$\mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp})^2] = \frac{1}{d-s}(\frac{2}{\pi}(q-s-1)+1). \tag{85}$$

Noting that $\|\mathbf{g}_\perp\| = \sqrt{1-\sum_{i=1}^s \alpha_i^2}$ and $\|\mathbf{v}_\perp\| = \sqrt{q-s}$, we have

$$\mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g})^2] = \mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g}_\perp)^2] = \mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp})^2 \|\mathbf{v}_\perp\|^2 \|\mathbf{g}_\perp\|^2] \tag{86}$$

$$= \frac{q-s}{d-s}(\frac{2}{\pi}(q-s-1)+1)(1-\sum_{i=1}^s \alpha_i^2). \tag{87}$$

Taking expectation to both sides of Eq. (84) and plugging in Eq. (82) and Eq. (87), Eq. (12) has been proved. □

A.3 Analysis for Prior-OPT

The Prior-OPT estimator is defined in Eq. (13). Note that there are s priors $\{\mathbf{p}_1, \cdots, \mathbf{p}_s\}$ (we assume that they have been normalized so that they have unit norm), and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q-s}\}$ is a randomly uniformly distributed orthonormal set of $q-s$ vectors in the $(d-s)$-dimensional subspace orthogonal to $\{\mathbf{p}_1, \cdots, \mathbf{p}_s\}$. In the Prior-OPT estimator for analysis, we use the convenient definition $\mathbf{v}_\perp$ in Eq. (77) (noting that $\text{sign}(\overline{\nabla g(\theta)}^\top \mathbf{u}_i) = \text{sign}(\nabla g(\theta)^\top \mathbf{u}_i)$):

$$\mathbf{v}_\perp := \sum_{i=1}^{q-s} \text{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i, \tag{88}$$

which is consistent with the definition in Eq. (14) under the directional derivative approximation.

For convenience we first consider the case of $s = 1$, and the analysis could be easily generalized to the case of $s > 1$.

A.3.1 The Case of $s = 1$

When $s = 1$, we write the Prior-OPT estimator in the following form:

$$\mathbf{v}^* = \nabla g(\theta)^\top \mathbf{p} \cdot \mathbf{p} + \nabla g(\theta)^\top \overline{\mathbf{v}_\perp} \cdot \overline{\mathbf{v}_\perp}, \tag{89}$$

where

$$\mathbf{v}_\perp := \sum_{i=1}^{q-1} \text{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i. \tag{90}$$

Here $\mathbf{p}$ is the prior vector with $\|\mathbf{p}\| = 1$, and $\{\mathbf{u}_i\}_{i=1}^{q-1}$ are the $q-1$ random orthonormal basis of the $(d-1)$-dimensional space orthogonal to $\mathbf{p}$. Note that we employ the directional derivative approximation as in Eq. (37).

Theorem A.5. *For the Prior-OPT estimator defined in Eq.* (89)*, we let* $\gamma = \overline{\mathbf{v}^*}^\top \overline{\nabla g(\theta)}$ *be its cosine similarity to the true gradient, where the notation* $\overline{\mathbf{v}^*} := \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|}$ *is defined to be the* ℓ_2 *normalization of the corresponding vector, then*

$$\mathbb{E}[\gamma] \geq \sqrt{\alpha^2 + \frac{(q-1)(1-\alpha^2)}{\pi}\left(\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}\right)^2}, \tag{91}$$

$$\mathbb{E}[\gamma^2] = \alpha^2 + \frac{1}{d-1}(\frac{2}{\pi}(q-2)+1)(1-\alpha^2), \tag{92}$$

where $\alpha := \mathbf{p}^\top \overline{\nabla g(\theta)}$ *is the cosine similarity between the prior and the true gradient.*

Proof. Let $\mathbf{g} := \overline{\nabla g(\theta)}$. Then $\mathbf{v}^* = \|\nabla g(\theta)\|(\mathbf{g}^\top \mathbf{p} \cdot \mathbf{p} + \mathbf{g}^\top \overline{\mathbf{v}_\perp} \cdot \overline{\mathbf{v}_\perp})$. We also note that $\mathbf{v}_\perp$ is a linear combination of $\mathbf{u}_1$ to $\mathbf{u}_{q-1}$, all of which are orthogonal to $\mathbf{p}$, so $\mathbf{v}_\perp$ is also orthogonal to $\mathbf{p}$. Therefore, $\|\mathbf{v}^*\| = \|\nabla g(\theta)\|\sqrt{(\mathbf{p}^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}$. Hence

$$\gamma = \frac{(\mathbf{v}^*)^\top \overline{\nabla g(\theta)}}{\|\mathbf{v}^*\|} \tag{93}$$

$$= \frac{(\mathbf{p}^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}{\sqrt{(\mathbf{p}^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}} \tag{94}$$

$$= \sqrt{(\mathbf{p}^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}. \tag{95}$$

We denote a new estimator

$$\widetilde{\mathbf{v}^*} = \nabla g(\theta)^\top \mathbf{p} \cdot \mathbf{p} + \mathbb{E}[\nabla g(\theta)^\top \overline{\mathbf{v}_\perp}] \cdot \overline{\mathbf{v}_\perp}. \tag{96}$$

$$= \|\nabla g(\theta)\|(\mathbf{g}^\top \mathbf{p} \cdot \mathbf{p} + \mathbb{E}[\mathbf{g}^\top \overline{\mathbf{v}_\perp}] \cdot \overline{\mathbf{v}_\perp}). \tag{97}$$

Let $\widetilde{\gamma}$ be the cosine similarity between $\widetilde{\mathbf{v}^*}$ and $\nabla g(\theta)$. Then

$$\widetilde{\gamma} = \frac{(\widetilde{\mathbf{v}^*})^\top \overline{\nabla g(\theta)}}{\|\widetilde{\mathbf{v}^*}\|} \tag{98}$$

$$= \frac{(\mathbf{p}^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]\overline{\mathbf{v}_\perp}^\top \mathbf{g}}{\sqrt{(\mathbf{p}^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}} \tag{99}$$

Therefore,

$$\mathbb{E}[\widetilde{\gamma}] = \mathbb{E}\left[\frac{(\mathbf{p}^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]\overline{\mathbf{v}_\perp}^\top \mathbf{g}}{\sqrt{(\mathbf{p}^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}}\right] \tag{100}$$

$$= \frac{\alpha^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}{\sqrt{\alpha^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}} \tag{101}$$

$$= \sqrt{\alpha^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2} \tag{102}$$

Note that $\mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}] = \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp}\|\mathbf{g}_\perp\|] = \sqrt{q-1}\sqrt{1-\alpha^2}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})\sqrt{\pi}}$. Hence,

$$\mathbb{E}[\widetilde{\gamma}] = \sqrt{\alpha^2 + \frac{(q-1)(1-\alpha^2)}{\pi}\left(\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}\right)^2} \tag{103}$$

The remaining part is to show the relationship between $\mathbb{E}[\gamma]$ and $\mathbb{E}[\widetilde{\gamma}]$. Note that $\mathbf{v}^*$ is the projection of $\nabla g(\theta)$ on the 2-dimensional subspace spanned by $\mathbf{p}$ and $\overline{\mathbf{v}_\perp}$. By Proposition 1 in Meier et al. (2019), among all the vectors in the subspace spanned by $\mathbf{p}$ and $\overline{\mathbf{v}_\perp}$, $\mathbf{v}^*$ has the largest cosine similarity with $\nabla g(\theta)$. Since $\widetilde{\mathbf{v}^*}$ is a linear combination of $\mathbf{p}$ and $\mathbf{v}_\perp$, $\gamma \geq \widetilde{\gamma}$ always holds. Therefore, $\mathbb{E}[\gamma] \geq \mathbb{E}[\widetilde{\gamma}]$.

Next, we derive $\mathbb{E}[\gamma^2]$. By Eq. (93) we have

$$\mathbb{E}[\gamma^2] = \alpha^2 + \mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2]. \tag{104}$$

Since $\|\mathbf{v}_\perp\| = \sqrt{q-1}$, using Eq. (75) we have

$$\mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2] = \frac{1}{\|\mathbf{v}_\perp\|^2}\mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g})^2] \tag{105}$$

$$= \frac{1}{d-1}(\frac{2}{\pi}(q-2)+1)(1-\alpha^2). \tag{106}$$

Plugging Eq. (106) into Eq. (104), the proof is completed. $\square$

A.3.2 The Case of $s > 1$

In the case of $s > 1$, we write the Prior-OPT estimator in the following form:

$$\mathbf{v}^* = \sum_{i=1}^{s} \nabla g(\theta)^\top \mathbf{p}_i \cdot \mathbf{p}_i + \nabla g(\theta)^\top \overline{\mathbf{v}_\perp} \cdot \overline{\mathbf{v}_\perp}, \tag{107}$$

where

$$\mathbf{v}_\perp := \sum_{i=1}^{q-s} \operatorname{sign}(\nabla g(\theta)^\top \mathbf{u}_i) \cdot \mathbf{u}_i. \tag{108}$$

Now, we present the proof of Theorem 3.4 for the Prior-OPT estimator defined in Eq. (107).

Proof. Let $\mathbf{g} := \overline{\nabla g(\theta)}$. Then

$$\mathbf{v}^* = \|\nabla g(\theta)\|(\sum_{i=1}^{s} \mathbf{g}^\top \mathbf{p}_i \cdot \mathbf{p}_i + \mathbf{g}^\top \overline{\mathbf{v}_\perp} \cdot \overline{\mathbf{v}_\perp}). \tag{109}$$

We also note that $\mathbf{v}_\perp$ is a linear combination of $\mathbf{u}_1$ to $\mathbf{u}_{q-s}$, all of which are orthogonal to $\{\mathbf{p}_i\}_{i=1}^s$, so $\mathbf{v}_\perp$ is also orthogonal to $\{\mathbf{p}_i\}_{i=1}^s$. Therefore, $\|\mathbf{v}^*\| = \|\nabla g(\theta)\|\sqrt{\sum_{i=1}^s (\mathbf{p}_i^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}$. Hence

$$\gamma = \frac{(\mathbf{v}^*)^\top \overline{\nabla g(\theta)}}{\|\mathbf{v}^*\|} \tag{110}$$

$$= \sqrt{\sum_{i=1}^{s} (\mathbf{p}_i^\top \mathbf{g})^2 + (\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2}. \tag{111}$$

We denote a new estimator

$$\widetilde{\mathbf{v}^*} = \sum_{i=1}^{s} \nabla g(\theta)^\top \mathbf{p}_i \cdot \mathbf{p}_i + \mathbb{E}[\nabla g(\theta)^\top \overline{\mathbf{v}_\perp}] \cdot \overline{\mathbf{v}_\perp}. \tag{112}$$

$$= \|\nabla g(\theta)\|(\sum_{i=1}^{s} \mathbf{g}^\top \mathbf{p}_i \cdot \mathbf{p}_i + \mathbb{E}[\mathbf{g}^\top \overline{\mathbf{v}_\perp}] \cdot \overline{\mathbf{v}_\perp}). \tag{113}$$

Let $\widetilde{\gamma}$ be the cosine similarity between $\widetilde{\mathbf{v}^*}$ and $\nabla g(\theta)$. Then

$$\widetilde{\gamma} = \frac{(\widetilde{\mathbf{v}^*})^\top \overline{\nabla g(\theta)}}{\|\widetilde{\mathbf{v}^*}\|} \tag{114}$$

$$= \frac{\sum_{i=1}^s (\mathbf{p}_i^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]\overline{\mathbf{v}_\perp}^\top \mathbf{g}}{\sqrt{\sum_{i=1}^s (\mathbf{p}_i^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}} \tag{115}$$

Therefore,

$$\mathbb{E}[\widetilde{\gamma}] = \mathbb{E}\left[\frac{\sum_{i=1}^{s}(\mathbf{p}_i^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]\overline{\mathbf{v}_\perp}^\top \mathbf{g}}{\sqrt{\sum_{i=1}^{s}(\mathbf{p}_i^\top \mathbf{g})^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}}\right] \tag{116}$$

$$= \frac{\sum_{i=1}^{s}\alpha_i^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}{\sqrt{\sum_{i=1}^{s}\alpha_i^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2}} \tag{117}$$

$$= \sqrt{\sum_{i=1}^{s}\alpha_i^2 + \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}]^2} \tag{118}$$

Note that $\mathbb{E}[\overline{\mathbf{v}_\perp}^\top \mathbf{g}] = \mathbb{E}[\overline{\mathbf{v}_\perp}^\top \overline{\mathbf{g}_\perp}\|\mathbf{g}_\perp\|] = \sqrt{q-s}\sqrt{1-\sum_{i=1}^{s}\alpha_i^2}\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})\sqrt{\pi}}$. Hence,

$$\mathbb{E}[\widetilde{\gamma}] = \sqrt{\sum_{i=1}^{s}\alpha_i^2 + \frac{(q-s)(1-\sum_{i=1}^{s}\alpha_i^2)}{\pi}\left(\frac{\Gamma(\frac{d-s}{2})}{\Gamma(\frac{d-s+1}{2})}\right)^2} \tag{119}$$

The remaining part is to show the relationship between $\mathbb{E}[\gamma]$ and $\mathbb{E}[\widetilde{\gamma}]$. We note that $\mathbf{v}^*$ is the projection of $\nabla g(\theta)$ on the $(s+1)$-dimensional subspace spanned by $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_s, \overline{\mathbf{v}_\perp}\}$. By Proposition 1 in Meier et al. (2019), among all the vectors in the subspace spanned by $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_s, \overline{\mathbf{v}_\perp}\}$, $\mathbf{v}^*$ has the largest cosine similarity with $\nabla g(\theta)$. Since $\widetilde{\mathbf{v}^*}$ also lies in this subspace, $\gamma \geq \widetilde{\gamma}$ always holds. Therefore, $\mathbb{E}[\gamma] \geq \mathbb{E}[\widetilde{\gamma}]$.

Next, we derive $\mathbb{E}[\gamma^2]$. By Eq. (110) we have

$$\mathbb{E}[\gamma^2] = \sum_{i=1}^{s}\alpha_i^2 + \mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2]. \tag{120}$$

Since $\|\mathbf{v}_\perp\| = \sqrt{q-s}$, using Eq. (87) we have

$$\mathbb{E}[(\overline{\mathbf{v}_\perp}^\top \mathbf{g})^2] = \frac{1}{\|\mathbf{v}_\perp\|^2}\mathbb{E}[(\mathbf{v}_\perp^\top \mathbf{g})^2] \tag{121}$$

$$= \frac{1}{d-s}(\frac{2}{\pi}(q-s-1)+1)(1-\sum_{i=1}^{s}\alpha_i^2). \tag{122}$$

Plugging Eq. (122) into Eq. (120), the proof is completed. □

A.4 Discussions

A.4.1 Prior's Accuracy α_i

$\alpha_i = \mathbf{p}_i^\top \overline{\nabla g(\theta)}$ is the cosine similarity between the i-th surrogate model's gradient (the i-th prior) and the true gradient. The value of α_i is only used in the theoretical analysis and not required for practical algorithm. Algorithm 1 does not require any α_i or the true gradient to run. We assume that α_i is known in the theoretical analysis for analyzing its impact on the expectation of the final estimated gradient's cosine similarity γ to the true gradient, which derives the solutions of $\mathbb{E}[\gamma]$ and $\mathbb{E}[\gamma^2]$. Figs. 5, 13 demonstrate the quantitative analysis of $\mathbb{E}[\gamma]$ and $\mathbb{E}[\gamma^2]$, respectively.

A.4.2 Issue of Sign-OPT

The problem of Sign-OPT is that its accuracy of gradient estimation is too low to effectively reduce distortion, because the query efficiency is not the only goal of the hard-label attack, and it is meaningless to spend a small number of queries without reducing the distortion of adversarial examples. Our objective is to minimize ℓ_p norm distortion under the same query number budget. The key is to spend similar queries to obtain better gradient estimation, allowing for fewer iterations to attack successfully, thereby achieving the effect of saving queries.

(1) Prior-Sign-OPT spends the same queries with Sign-OPT in gradient estimation, which also uses the the sign function to estimate gradients. But it uses the prior directions, which are more useful

than the randomly sampled directions of Sign-OPT, and thus it estimates a more accurate gradient than Sign-OPT.

(2) Prior-OPT only uses slightly more queries than Sign-OPT to calculate the weighted average of directional derivatives in Eq. (13), which obtains significantly more accurate gradients, thereby resulting in a rapid decrease in ℓ_p norm distortion.

A.4.3 PRACTICALITY OF THEORY

No matter how complex a real-world situation, the generation of adversarial example is all about the gradient vectors, because it can increase the classification loss that leads to the misclassification. Our theory is about the similarity of two vectors: the estimated gradient and the true gradient, and it has universality on all image classifiers. One of the most common cases in real environments is the large difference between models, which have different gradients. We address this case by using the variable α_i in Eqs. (11), (12), (15), (16), (17), which is the cosine similarity between the i-th prior and the true gradient, and assuming to be known in the theoretical analysis. In summary, our theory can be applied to real-world environments and has universality.

B STEPS FOR SETTING λ_0 AND $\hat{y}_{\text{ADV}}$ IN OBTAINING TRANSFER-BASED PRIORS

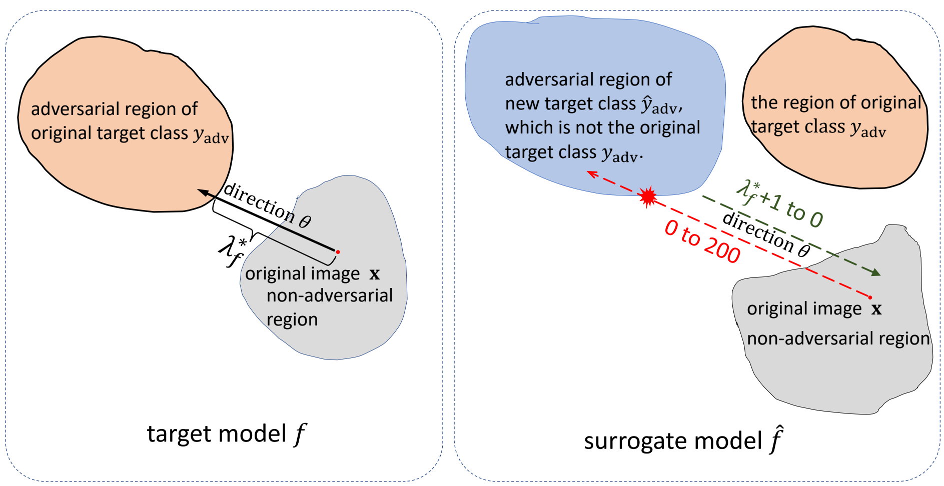


Figure 6: Illustration of how to set the new target class $\hat{y}_{\text{adv}}$ and λ_0.

As shown in Fig. 6, the steps of setting λ_0 and $\hat{y}_{\text{adv}}$ before computing Eq. (4) in Section 3.2 are as follows.

(1) We send a ray $\mathbf{r}$ from $\mathbf{x}$ along the direction θ in the surrogate model $\hat{f}$, and find the nearest decision boundary $\mathcal{B}$ at distances ranging from $\lambda_f^* + 1$ to 0 along the ray $\mathbf{r}$. The adversarial region beyond this decision boundary $\mathcal{B}$ is denoted as R_{adv}. Here, λ_f^* is the shortest distance from $\mathbf{x}$ along θ to the region corresponding to y_{adv} in the target model f.

(2) If an adversarial region is not found in the first step, we expand the search scope to locate the first decision boundary $\mathcal{B}$ of $\hat{f}$ at distances ranging from 0 to 200 along the ray $\mathbf{r}$. The adversarial region beyond this boundary $\mathcal{B}$ is denoted as R_{adv}.

Finally, the shortest distance from $\mathbf{x}$ along the direction θ to R_{adv} is set as λ_0, and the classification label of R_{adv} is denoted as $\hat{y}_{\text{adv}}$.

C PROOF FOR SURROGATE GRADIENT COMPUTATION

For a surrogate model $\hat{f}$, we define f as the negative C&W loss function as

$$f(\mathbf{x}') = \begin{cases} \hat{f}_y - \max_{j \neq y} \hat{f}_j, & \text{if untargeted attack;} \\ \max_{j \neq \hat{y}_{\text{adv}}} \hat{f}_j - \hat{f}_{\hat{y}_{\text{adv}}}, & \text{if targeted attack;} \end{cases} \tag{123}$$

where $\hat{f}_i := \hat{f}(\mathbf{x}')_i$ is an abbreviation for the i-th element of the output of $\hat{f}$. Now we consider $g(\theta)$, the distance from the benign image $\mathbf{x}$ to the adversarial region along the ray direction θ, as defined in Eq. (3). For any $\mathbf{x}'$, $\Phi(\mathbf{x}') = 1 \Leftrightarrow f(\mathbf{x}') \leq 0$, so we have

$$g(\theta) = \arg\min_{\lambda > 0} \left(f(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}) \leq 0 \right). \tag{124}$$

In the theoretical analysis let us rigorously specify the meaning of $\arg\min$ above. Let A_θ denote the set $\{\lambda : \lambda > 0, f(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}) \leq 0\}$. Then we define

$$g(\theta) = \begin{cases} \inf(A_\theta) & A_\theta \neq \emptyset \\ +\infty & A_\theta = \emptyset \end{cases}, \tag{125}$$

where $\inf(\cdot)$ denotes the infimum of a subset of $\mathbb{R}$. Now we can prove

Proposition C.1. *If f is continuous, $f(\mathbf{x}) > 0$, then given any $\theta \neq 0$ s.t. $0 < g(\theta) < +\infty$, we have $f\left(\mathbf{x} + g(\theta) \frac{\theta}{\|\theta\|}\right) = 0$.*

Proof. Since f is continuous, the function $f_\theta(\lambda) := f\left(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\right)$ is continuous w.r.t. λ. We want to prove $f_\theta(g(\theta)) = 0$.

If $f_\theta(g(\theta)) < 0$, then since f_θ is continuous and $g(\theta) > 0$, there exists $\epsilon > 0$ such that $f_\theta(g(\theta) - \epsilon) < 0$ and $g(\theta) - \epsilon > 0$. Therefore, we have $g(\theta) - \epsilon \in A_\theta$, which implies that $g(\theta) > \inf(A_\theta)$. This contradicts the definition $g(\theta) = \inf(A_\theta)$.

If $f_\theta(g(\theta)) > 0$, then there exists $\epsilon > 0$ such that $f_\theta(\lambda) > 0$ holds for all $g(\theta) \leq \lambda \leq g(\theta) + \epsilon$. This means that $[g(\theta), g(\theta) + \epsilon] \cap A_\theta = \emptyset$. Noting that $g(\theta)$ is a lower bound of A_θ, this implies that $g(\theta) + \epsilon$ is also a lower bound of A_θ, which contradicts the definition $g(\theta) = \inf(A_\theta)$.

Therefore $f_\theta(g(\theta)) = 0$. $\square$

Next we show how to calculate $\nabla g(\theta)$ based on some weak assumptions.

Theorem C.2. *Suppose f is continuously differentiable*[1] *and $f(\mathbf{x}) > 0$. Let $h(\theta, \lambda) := f\left(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\right)$. For any $\theta_0 \neq 0$ s.t. $g(\theta_0) < +\infty$, let $\lambda_0 = g(\theta_0)$, and assume that $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$, then we have: $g(\theta)$ is differentiable at θ_0, and*

$$\nabla g(\theta_0) = -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)} \nabla_\theta h(\theta_0, \lambda_0). \tag{126}$$

Remark C.3. The assumptions in the theorem are rather weak. $f(\mathbf{x}) > 0$ (the unperturbed sample can be successfully classified) is standard; $g(\theta) < +\infty$ is a common assumption, necessary for ray search procedure to work; f is continuously differentiable almost everywhere using common network architectures. The only special condition required here is that $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$, and it is usually true unless a particular f is constructed against it. (Imagine that increasing λ makes the function value decrease from a positive value to a non-positive value, and the derivative w.r.t. λ is usually not 0 when the function value crosses 0.)

[1] Continuously differentiability (also called C^1) means that all partial derivatives of the function exist and are continuous.

Proof. Since $(\theta, \lambda) \mapsto \mathbf{x} + \lambda \frac{\theta}{\|\theta\|}$ is continuously differentiable at $\{(\theta, \lambda) : \theta \in \mathbb{R}^d, \lambda \in \mathbb{R}, \theta \neq 0\}$ and f is continuously differentiable everywhere, h is continuously differentiable when $\theta \neq 0$ by the chain rule. By Proposition C.1 $h(\theta_0, \lambda_0) = 0$. And since $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$, by the Implicit Function Theorem (see e.g. Theorem 1 in Rio Branco de Oliveira (2012)), there exists a neighborhood $\Theta \subseteq \mathbb{R}^d$ of θ_0 and an open interval $\Lambda := (\lambda_0 - \eta, \lambda_0 + \eta)$ such that for each $\theta \in \Theta$, there exists a unique $\lambda \in \Lambda$ s.t. $h(\theta, \lambda) = 0$. Since θ uniquely determines λ we define $\tilde{g} : \Theta \to \Lambda$ satisfying $h(\theta, \tilde{g}(\theta)) = 0$ for all $\theta \in \Theta$. Moreover, the Implicit Function Theorem tells us that $\tilde{g}$ is continuously differentiable, and

$$\nabla \tilde{g}(\theta_0) = -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)} \nabla_\theta h(\theta_0, \lambda_0). \tag{127}$$

Now it suffices to prove $\nabla g(\theta_0) = \nabla \tilde{g}(\theta_0)$. We shall prove that there exists a neighborhood of θ_0 in which g and $\tilde{g}$ are equal. Since $h(\theta, \tilde{g}(\theta)) = 0$, from the definition of g, we have $g(\theta) \leq \tilde{g}(\theta) < +\infty$ for all $\theta \in \Theta$. By Proposition C.1, $h(\theta, g(\theta)) = 0$, so the uniqueness in Implicit Function Theorem tells us that $\forall \theta \in \Theta$, if $\lambda_0 - \eta < g(\theta) < \lambda_0 + \eta$, then $g(\theta) = \tilde{g}(\theta)$. Since $g(\theta) \leq \tilde{g}(\theta) < \lambda_0 + \eta$, it suffices to prove that $g(\theta) > \lambda_0 - \eta$.

Now we prove that there exists a neighborhood Θ' of θ_0 such that $\forall \theta \in \Theta'$, $\forall \lambda \in [0, \lambda_0 - \eta]$, $h(\theta, \lambda) > 0$ (so $\forall \theta \in \Theta', g(\theta) > \lambda_0 - \eta$). To prove that we first note that $\forall \lambda \in [0, \lambda_0 - \eta]$, $h(\theta_0, \lambda) > 0$ since $g(\theta_0) = \lambda_0 > \lambda_0 - \eta$. Since $h(\theta_0, \lambda)$ is continuous w.r.t. λ, by the Extreme Value Theorem, $h(\theta_0, \lambda)$ on $\lambda \in [0, \lambda_0 - \eta]$ could attain the minimum $h(\theta_0, \lambda^*)$ which is positive, so there exists $\epsilon > 0$ such that $\forall \lambda \in [0, \lambda_0 - \eta], h(\theta_0, \lambda) \geq \epsilon$. We pick a bounded closed neighborhood of θ_0, denoted by Θ'' such that $0 \notin \Theta''$. h is continuous on the compact set $\{(\theta, \lambda) : \theta \in \Theta'', \lambda \in [0, \lambda_0 - \eta]\}$, so by Heine-Cantor Theorem, h is uniformly continuous on the same set. This implies that there exists $\delta > 0$ such that $\forall \theta \in \Theta''$ such that $\|\theta - \theta_0\| < \delta$, we have $|h(\theta, \lambda) - h(\theta_0, \lambda)| < \epsilon$ and hence $h(\theta, \lambda) > 0$ for all $\lambda \in [0, \lambda_0 - \eta]$. Setting $\Theta' = \Theta'' \cap \{\theta : \|\theta - \theta_0\| < \delta\}$, we have $\forall \theta \in \Theta', \forall \lambda \in [0, \lambda_0 - \eta], h(\theta, \lambda) > 0$, and thus the proposition at the beginning of this paragraph is proven. Finally, using the definition of $g(\theta)$ (Eq. (124)) together with the conditions of $h(\theta, \lambda) > 0$, we have $g(\theta) > \lambda_0 - \eta$ for $\forall \theta \in \Theta'$.

Therefore, we have proven that there exists a neighborhood of θ_0, $\Theta \cap \Theta'$, in which g and $\tilde{g}$ are equal. Since the gradient only relies on the function value in a neighborhood, $\nabla g(\theta_0) = \nabla \tilde{g}(\theta_0)$ and by Eq. (127) the proof is completed. □

D Derivation of Conditions for Prior-OPT to Outperform Sign-OPT

With the formulas of $\mathbb{E}[\gamma^2]$ of Sign-OPT (Eq. (10)) and Prior-OPT (Eq. (17)), we now derive the exact value of α_i for which Prior-OPT can outperform Sign-OPT on gradient estimation.

Now, we rewrite the formulas of $\mathbb{E}[\gamma^2]$ of Sign-OPT and Prior-OPT as following:

$$\mathbb{E}[\gamma^2]_{\text{Sign-OPT}} = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right), \tag{128}$$

$$\mathbb{E}[\gamma^2]_{\text{Prior-OPT}} = \sum_{i=1}^{s}, \alpha_i^2 + \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)\left(1-\sum_{i=1}^{s}\alpha_i^2\right). \tag{129}$$

We need to find the value of α_i such that $\mathbb{E}[\gamma^2]_{\text{Prior-OPT}} > \mathbb{E}[\gamma^2]_{\text{Sign-OPT}}$.

Let $A = \sum_{i=1}^{s} \alpha_i^2$, and the inequality becomes:

$$A + (1-A)\cdot\frac{1}{d-s}\left(\frac{2}{\pi}(q-s+1)+1\right) > \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right) \tag{130}$$

Now let us simplify the left side of Eq. (130) to $\mathbb{E}[\gamma^2]_{\text{Prior-OPT}} = A + (1-A)C_2$, where $C_2 = \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)$.

Then, let us simplify the right side of Eq. (130) to $\mathbb{E}[\gamma^2]_{\text{Sign-OPT}} = C_1$, where $C_1 = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right)$.

The inequality of Eq. (130) becomes:

$$A + (1 - A)C_2 > C_1. \tag{131}$$

We rearrange the above inequality as $A(1 - C_2) + C_2 > C_1$, and the we solve for A:

$$A > \frac{C_1 - C_2}{1 - C_2} \tag{132}$$

Substituting the formulas of A, C_1, and C_2 into Eq. (132), we have:

$$\sum_{i=1}^{s} \alpha_i^2 > \frac{\frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right) - \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)}{1 - \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right)} \tag{133}$$

This is the condition of $\sum_{i=1}^{s} \alpha_i^2$ for Prior-OPT to outperform Sign-OPT. But this inequality is complex, next we show how to further simplify this inequality. We can simplify the above inequality under the reasonable assumptions that $s \ll q \ll d$ (the number of total vectors is much larger than the number of priors, and input dimension is much larger than the number of total vectors).

We first approximate denominator of Eq. (133), note that when $s \ll d$, we have $d - s \approx d$. Similarly, when $s \ll q$, we have $q - s - 1 \approx q - 1$. Therefore, the denominator simplifies to:

$$D = 1 - \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right) \approx 1 - \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right). \tag{134}$$

Since $\frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right)$ is a small number because $q \ll d$ (denote it as ϵ), the denominator becomes $D \approx 1 - \epsilon \approx 1$. Next, we simplify the numerator as:

$$\begin{aligned}
N &= \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right) - \frac{1}{d-s}\left(\frac{2}{\pi}(q-s-1)+1\right), \\
&\approx \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right) - \frac{1}{d}\left(\frac{2}{\pi}(q-s-1)+1\right), \\
&= \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1 - \left(\frac{2}{\pi}(q-s-1)+1\right)\right), \\
&= \frac{2s}{\pi d}.
\end{aligned} \tag{135}$$

Now, let us substitute the simplified N and D into the right side of Eq. (133), we have

$$\sum_{i=1}^{s} \alpha_i^2 > \frac{N}{D} \approx \frac{2s}{\pi d}. \tag{136}$$

This is the simplified condition of $\sum_{i=1}^{s} \alpha_i^2$ for Prior-OPT to outperform Sign-OPT.

Dividing both sides by s, we get the condition for the average squared cosine similarity $\overline{\alpha^2} = \frac{1}{s}\sum_{i=1}^{s} \alpha_i^2 > \frac{2}{\pi d}$. Since $\frac{2}{\pi d}$ is typically a very small value due to the large input dimension d, this threshold is relatively easy to satisfy. Therefore, Prior-OPT generally outperforms Sign-OPT when the priors have even a minimal level of informativeness (non-zero α_i).

E EXPERIMENTAL SETTING

In this section, we provide the hyperparameter settings of our approach and compared methods, which include HSJA, TA, G-TA, GeoDA, Evolutionary, Triangle Attack, SurFree, Sign-OPT, SVM-OPT, SQBA, and BBA.

Experimental Equipment. The experiments of all methods are conducted by using PyTorch 1.7.1 framework on NVIDIA V100 and A100. NVIDIA A100 GPU has TensorFloat-32 (TF32) tensor cores to achieve better performance (*i.e.,* the computation speed), and enabling TF32 tensor cores causes a large relative error compared to double precision, especially in attacks of ViTs. Therefore, in all experiments, we set `torch.backends.cuda.matmul.allow_tf32 = False` and `torch.backends.cudnn.allow_tf32 = False` to obtain higher precision.

CIFAR-10 dataset. In the CIFAR-10 dataset, we select 4 networks as target models, including a 272-layer PyramidNet+ShakeDrop network (PyramidNet-272) Han et al. (2017); Yamada et al. (2019), two wide residual networks with 28 and 40 layers (WRN-28 and WRN-40) Zagoruyko & Komodakis (2016), and DenseNet-BC-190 ($k = 40$) Huang et al. (2017). We adopt the ResNet-110 as the surrogate model in the CIFAR-10 dataset.

Prior-OPT and Prior-Sign-OPT. Hyperparameters of Prior-OPT and Prior-Sign-OPT are listed in Table 3. Our implementation source code is based on the PyTorch framework. In the experiments of targeted attacks with the same target class, we set the initial direction θ_0 of Prior-OPT and Prior-Sign-OPT to the direction of the same image of the target class used by other methods.

Table 3: The hyperparameters of Prior-OPT and Prior-Sign-OPT.

Dataset	Hyperparameter	Value
CIFAR-10	q, total number of vectors for estimating a gradient, including priors and random vectors	200
	α, the update step size of the direction θ	0.2
	β, used for the gradient estimation and the stopping threshold of binary search	0.001
	the binary search's stopping threshold	$\frac{\beta}{500}$
	number of iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm for the gradient clipping operation	0.1
ImageNet	q, total number of vectors for estimating a gradient, including priors and random vectors	200
	α, the update step size of the direction θ	0.2
	β, used for the gradient estimation of θ	0.001
	the binary search's stopping threshold	1×10^{-4}
	number of iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm for the gradient clipping operation	1.0

Sign-OPT and SVM-OPT. Hyperparameters of Sign-OPT and SVM-OPT are listed in Table 4. For fair comparison, we set the hyperparameters of Prior-OPT and Prior-Sign-OPT to be same with Sign-OPT and SVM-OPT, *e.g.,* the same number of vectors for the gradient estimation q and the same α and β.

HSJA, TA and G-TA. Hyperparameters of HSJA, TA and G-TA are listed in Table 5. TA has no additional hyperparameters. G-TA has an additional hyperparameter radius ratio r to control the shape of the virtual semi-ellipsoid, which is set to be different values in CIFAR-10 and ImageNet, as shown in Table 5.

Evolutionary. We follow the official implementation source code of Evolutionary to set its hyperparameters, as shown in Table 6.

GeoDA. GeoDA only supports untargeted attacks, and the convergence of ℓ_2 norm attacks of GeoDA is proved. Thus, we conduct experiments of ℓ_2 norm untargeted attacks of GeoDA, and the hyperparameters of GeoDA are shown in Table 7.

Triangle Attack. We set the hyperparameter of "ratio mask" as 0.1 and 0.3 in the ImageNet dataset and the CIFAR-10 dataset, respectively. All hyperparameters of Triangle Attack are shown in Table 8.

SurFree. SurFree only supports the ℓ_2 norm attacks. We translate the code from SurFree's official version into the PyTorch version for experiments, and the hyperparameters of SurFree are shown in Table 9.

SQBA. SQBA only supports the ℓ_2 norm untargeted attacks, and the hyperparameters of SQBA are shown in Table 10.

Table 4: The hyperparameters of Sign-OPT and SVM-OPT.

Hyperparameter	Value
q, number of queries for estimating an approximate gradient	200
α, the update step size of the direction θ	0.2
β, used for the gradient estimation of θ and determining the stopping threshold of binary search	0.001
the number of iterations	1,000
the binary search's stopping threshold of the CIFAR-10 dataset	$\frac{\beta}{500}$
the binary search's stopping threshold of the ImageNet dataset	1×10^{-4}

Table 5: The hyperparameters of HSJA, TA and G-TA.

Hyperparameter	Value
γ, threshold of the binary search	1.0
B_0, the initial batch size for gradient estimation	100
$B_{\max}$, the maximum batch size for gradient estimation	10,000
the search method for step size	geometric progression
number of iterations	64
radius ratio r for the ImageNet dataset in G-TA	1.1
radius ratio r for the CIFAR-10 dataset in G-TA	1.5

Table 6: The hyperparameters of Evolutionary.

Hyperparameter	Value
c_{cov}, the hyperparameter of updating the diagonal covariance matrix $\mathbf{C}$	0.001
σ, the deviation for bias	0.03
μ, the mean for bias, and it is a critical hyper-parameter controlling the strength of going towards the original image	0.01
maxlen, the maximum length of successful attacks for calculating μ	30

Table 7: The hyperparameters of GeoDA.

Dataset	Hyperparameter	Value
CIFAR-10	subspace dimension, the dimension of 2D DCT basis's subspace	10
	ϵ, the step size of searching the decision boundary	0.5
ImageNet	subspace dimension, the dimension of 2D DCT basis's subspace	75
	ϵ, the step size of searching the decision boundary	5

Table 8: The hyperparameters of Triangle Attack.

Dataset	Hyperparameter	Value
CIFAR-10	d, the number of picked dimensions	3
	ratio mask, the ratio of mask size for obtaining the mask in the low frequency	0.3
	initial θ, the initial angle of a subspace=initial $\theta \times \pi/32$	2
	initial α, the initial angle of alpha	$\frac{\pi}{2}$
	the maximum iteration number of attack algorithm in 2D subspace	2
ImageNet	d, the number of picked dimensions	3
	ratio mask, the ratio of mask size for obtaining the mask in the low frequency	0.1
	initial θ, the initial angle of a subspace=initial $\theta \times \pi/32$	2
	initial α, the initial angle of alpha	$\frac{\pi}{2}$
	the maximum iteration number of attack algorithm in 2D subspace	2

BBA. BBA only supports the ℓ_2 norm attacks, and the hyperparameters of BBA are shown in Table 11. We only use the surrogate model's bias and the hyperparameter `pg_factor` controls the strength of

Table 9: The hyperparameters of SurFree.

Hyperparameter	Value
`BS_gamma`, the stopping threshold in the binary search of α	0.01
`BS_max_iteration`, the maximum iterations in the binary search for α	10
ρ, the parameter for determining $\theta_{\max}$	0.98
`T`, the parameter for determining the range of α and the best θ	3
$\theta_{\max}$, the parameter for determining the range of α	30
`n_ortho`, the parameter for finding the direction of the lowest ϵ in `_get_candidates`	100
the binary search's stopping threshold of the ImageNet dataset	1×10^{-4}
`frequence_range`, the parameter used in constructing `dct_mask`	$0 \sim 0.5$
`with_distance_line_search`, the parameter used in `_get_candidates`	false
`with_interpolation`, the parameter used in `_get_candidates`	false
`with_alpha_line_search`, the parameter used in `_get_best_theta`	true

Table 10: The hyperparameters of SQBA.

Hyperparameter	Value
`threshold`, the stopping threshold in the binary search	0.001
`min_randoms`, the value indirectly determines the number of queries in each gradient estimation	10

Table 11: The hyperparameters of BBA.

Hyperparameter	Value
`use_surrogate_bias`, whether to use a surrogate model as the bias	True
`use_mask_bias`, whether to use regional masks as the bias	False
`use_perlin_bias`, whether to use Perlin Noise as the bias	False
`pg_factor`, the hyperparameter that controls the strength of the bias	0.3

this bias. When the value of `pg_factor` $= 1$, then the orthogonal step would be equivalent to an iteration of PGD attack. Brunner et al. (2019) suggest that `pg_factor` $= 0.3$.

F Additional Experimental Results

In this section, we provide the additional experimental results of the ImageNet dataset and the CIFAR-10 dataset.

F.1 Computational Overhead

The primary additional computational cost of Prior-OPT compared to Sign-OPT arises from the binary search procedure performed on the priors during gradient estimation. Let d represent the dimension of the input image, q the number of vectors used in gradient estimation, and $f(d)$ the inference time of the target model for an input of dimension d. The time complexity of gradient estimation in Sign-OPT is $O(q \cdot f(d))$. In Prior-OPT, s priors are introduced. Each prior requires a binary search procedure, which involves approximately k inference steps. While k may vary slightly depending on the specific prior or the input configuration, its value generally remains bounded and logarithmic in scale, given the nature of binary search. Consequently, the time complexity of the gradient estimation step of Prior-OPT can be expressed as:

$$O((q - s + (s+1) \cdot k) \cdot f(d) + s \cdot \hat{f}(d)). \tag{137}$$

When q is large, s and k is relatively small (*i.e.,* the number of priors is small, and k typically ranges in the tens), $s \cdot \hat{f}(d)$ denotes the time of taking s priors, the additional overhead introduced by Prior-OPT is limited compared to Sign-OPT. While Prior-OPT introduces extra computation due to the binary search procedure on the priors, the increase in time complexity is relatively modest, especially when s remains much smaller than q. This demonstrates that Prior-OPT achieves a balance between computational efficiency and improved gradient estimation.

Table 12 demonstrates the time consumption of Sign-OPT, SVM-OPT, Prior-Sign-OPT and Prior-OPT, which are measured by conducting untargeted attack on the ImageNet dataset. The surrogate model employs the ResNet-50 architecture and the GPU is NVIDIA Tesla V100. The additional time overhead of Prior-Sign-OPT is mainly the time of obtaining priors on surrogate models. Prior-OPT calculates Eq. (13) to estimate gradients, which calls $s+1$ times of binary search, where s is the number of surrogate models. This will result in additional time consumption than Prior-Sign-OPT. Note that the major metrics of black-box attacks are not time consumption, but the number of queries and the attack success rate. Because in real-world systems, the number of queries is the main limitation. We need to use as few queries as possible to achieve the highest possible success rate. Table 13 shows the GPU memory allocations of Sign-OPT, Prior-Sign-OPT and Prior-OPT.

Table 12: The time consumption of attacking one image with $10{,}000$ queries, which are measured by `seconds` on a NVIDIA Tesla V100 GPU.

Method	ResNet-101	SENet-154	ResNeXt-101	GC ViT	Swin Transformer
Sign-OPT Cheng et al. (2020)	112	197	91	131	88
SVM-OPT Cheng et al. (2020)	119	189	102	158	98
Prior-Sign-OPT$_{\text{ResNet50}}$	240	372	195	203	183
Prior-OPT$_{\text{ResNet50}}$	342	476	321	357	203

Table 13: The GPU memory allocations of attacks against different target models, which are measured by `MB` on a NVIDIA Tesla V100 GPU.

Method	ResNet-101	SENet-154	ResNeXt-101	GC ViT	Swin Transformer
Sign-OPT Cheng et al. (2020)	4686	6244	7272	7352	8854
Prior-Sign-OPT$_{\text{ResNet50}}$	5222	6750	7828	7856	9410
Prior-OPT$_{\text{ResNet50}}$	5222	6746	7816	7846	9390

Prior-OPT and Prior-Sign-OPT require the transfer-based priors, and thus the additional memory allocation of our approach is mainly consumed in the feed-forward and back-propagation processes of the surrogate models. After obtaining a prior, the GPU memory will be released in a timely manner to save occupancy. Thus, our approach does not take up too much additional GPU memory.

F.2 Comparison with State-Of-The-Art Methods

Tables 14, 15, 16 show the experimental results of untargeted attacks against ViTs and CNNs on the ImageNet dataset, where AUC indicates area under the curve of the mean ℓ_2 distortions versus the number of queires, and the "Mean ℓ_2" is the average ℓ_2 norm distortion of the final adversarial example, and "ASR" indicates the attack success rate of the final adversarial examples. Here, the final adversarial examples are generated by using the query budget of $10{,}000$. The ASR is defined as the the percentage of samples with distortions below a threshold ϵ, which is set to $\epsilon = \sqrt{0.001 \times d}$ in the ImageNet dataset and $\epsilon = 1.0$ in the CIFAR-10 dataset, where d is the image dimension. Tables 14, 15, 16 demonstrate that the Prior-OPT with two surrogate models performs the best in most cases, and the PGD initialization of θ (*i.e.,* Prior-OPT$_{\theta_0^{\text{PGD}} + \text{X}}$) can effectively decrease the AUC number.

Table 17 shows the experimental results of ℓ_∞ norm attacks on the ImageNet dataset. The results show that Prior-OPT and Prior-Sign-OPT also perform well on ℓ_∞ norm attacks, which outperforms Sign-OPT in terms of average ℓ_∞ norm distortions. These results highlight the effectiveness of our approach across various adversarial attack types.

Fig. 7 shows the experimental results of ℓ_2 norm untargeted attacks on the ImageNet dataset. The results demonstrate that Prior-OPT significantly outperform all compared methods, including SQBA and BBA that also uses surrogate models. The results also show that use of multiple surrogate models can further boost the performance. In addition, the PGD θ initialization ensures the algorithm's initial attack direction θ_0 is already good, which enables it to achieve a better untargeted attack performance even with a small number of queries (*e.g.,* the query budget of $1{,}000$).

Table 14: Untargeted attack results of ViTs on the ImageNet dataset, where AUC denotes the area under the curve of mean ℓ_2 distortions versus the number of queries (the lower is better), and ASR denotes the attack success rate of the final adversarial example.

Method	ViT			GC ViT			Swin Transformer		
	Mean ℓ_2	AUC	ASR	Mean ℓ_2	AUC	ASR	Mean ℓ_2	AUC	ASR
HSJA	5.637	102956.7	96.7%	7.955	163915.1	82.8%	10.635	228806.0	70.3%
TA	5.674	104023.3	96.6%	9.102	176063.8	76.7%	10.513	230351.0	68.4%
G-TA	5.643	102013.4	96.4%	8.671	170511.6	77.6%	9.929	219877.9	72.6%
GeoDA	6.313	83176.7	91.0%	12.998	172173.2	54.3%	19.120	245094.6	31.5%
Evolutionary	6.719	128659.9	89.8%	8.615	174592.0	79.1%	15.738	266695.9	52.6%
Triangle Attack	10.097	99746.4	69.0%	30.119	298578.9	21.2%	29.005	288358.2	23.1%
SurFree	6.303	104053.9	91.6%	10.967	193400.6	65.4%	13.059	200688.5	58.3%
BBA$_{\text{ResNet50}}$	9.567	125221.1	74.0%	11.294	161711.4	67.5%	14.084	185551.7	59.9%
BBA$_{\text{ConViT}}$	8.595	105826.6	79.3%	9.188	128468.5	77.6%	12.375	156081.9	59.5%
SQBA$_{\text{ResNet50}}$	5.201	79423.0	95.7%	6.186	100435.8	89.2%	7.557	115845.7	83.2%
SQBA$_{\text{ConViT}}$	4.452	60295.8	98.3%	5.056	79670.8	94.4%	5.883	82141.3	91.0%
Sign-OPT	4.572	111439.9	98.3%	7.185	166001.9	85.9%	9.899	238907.0	74.7%
SVM-OPT	5.070	120008.9	97.1%	7.325	171869.8	83.9%	10.526	249491.3	72.1%
Prior-Sign-OPT$_{\text{ResNet50}}$	4.850	119961.6	97.4%	6.723	165586.5	87.1%	9.254	234462.4	75.7%
Prior-Sign-OPT$_{\text{ConViT}}$	4.313	105379.9	98.1%	5.972	151725.9	89.0%	7.622	198431.3	84.2%
Prior-Sign-OPT$_{\text{ResNet50\&ConViT}}$	3.967	99940.6	98.6%	5.286	137011.9	92.9%	6.331	177589.1	89.2%
Prior-Sign-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	4.331	88120.1	97.7%	5.243	98749.0	92.6%	8.112	128167.1	80.3%
Prior-OPT$_{\text{ResNet50}}$	5.195	106791.0	97.3%	6.066	134255.9	90.7%	9.625	190534.1	73.0%
Prior-OPT$_{\text{ConViT}}$	3.754	62928.1	**99.2%**	4.453	92662.3	95.7%	5.558	102428.7	91.8%
Prior-OPT$_{\text{ResNet50\&ConViT}}$	**3.609**	**60449.0**	**99.2%**	**3.700**	**76896.1**	**98.3%**	**4.896**	**91211.7**	**94.5%**
Prior-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	5.009	90005.1	96.4%	5.502	98555.8	92.8%	8.552	128766.0	76.4%

Table 15: Untargeted attack results of CNNs on the ImageNet dataset, where AUC denotes the area under the curve of mean ℓ_2 distortions versus the number of queries (the lower is better), and ASR denotes the attack success rate of the final adversarial example.

Method	ResNet-101			ResNeXt-101 (64 × 4d)			SENet-154		
	Mean ℓ_2	AUC	ASR	Mean ℓ_2	AUC	ASR	Mean ℓ_2	AUC	ASR
HSJA	5.158	96234.2	95.8%	5.484	110376.8	95.0%	9.385	177364.9	74.9%
TA	5.239	96858.5	95.9%	5.565	110870.1	95.0%	9.379	172600.0	73.8%
G-TA	5.225	95901.1	96.3%	5.524	109990.5	95.0%	5.430	119281.1	92.9%
GeoDA	6.364	82320.0	91.9%	6.898	88947.7	89.3%	8.209	107267.4	80.9%
Evolutionary	5.406	107841.6	93.2%	6.042	123706.5	91.3%	6.111	130032.0	90.1%
Triangle Attack	12.123	117731.5	61.3%	11.883	116639.5	63.7%	15.019	145508.7	48.9%
SurFree	6.627	104285.4	88.1%	7.550	123394.0	83.7%	8.247	131295.4	79.5%
BBA$_{\text{ResNet50}}$	7.295	83314.7	86.9%	9.393	116579.5	74.9%	8.976	115007.9	76.9%
SQBA$_{\text{ResNet50}}$	3.563	46450.8	98.9%	4.058	59316.7	97.8%	4.332	67106.3	97.9%
Sign-OPT	4.754	101907.7	95.9%	5.108	120545.5	95.4%	5.111	124730.7	93.5%
SVM-OPT	4.842	105778.8	95.8%	5.255	126799.4	95.0%	5.125	127568.9	93.7%
Prior-Sign-OPT$_{\text{ResNet50}}$	3.019	79126.4	99.4%	3.518	100999.4	98.9%	4.223	114089.3	96.7%
Prior-Sign-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	**2.045**	27148.4	99.7%	**2.450**	35290.7	99.4%	**2.958**	48708.7	98.3%
Prior-OPT$_{\text{ResNet50}}$	2.158	37218.3	**100.0%**	2.692	53085.7	**99.7%**	3.394	68609.1	**98.9%**
Prior-OPT$_{\theta_0^{\text{PGD}}+\text{ResNet50}}$	2.107	**25627.6**	99.5%	2.486	**33291.5**	**99.7%**	3.215	**48447.6**	98.6%

Fig. 8 shows the experimental results of ℓ_2 norm targeted attacks on the ImageNet dataset. The results of Fig. 8 show that Prior-Sign-OPT outperforms Prior-OPT under targeted attacks on CNNs models, and Prior-Sign-OPT with multiple surrogate models outperforms that with a single surrogate model.

Figs. 9 and 10 show attack success rates of untargeted attacks and targeted attacks on the ImageNet dataset. In untargeted ℓ_2 norm attacks (Fig. 9), Prior-OPT with two surrogate models performs the best, and both Prior-OPT and Prior-Sign-OPT outperform the baseline Sign-OPT. In targeted attack results (Fig. 10), Prior-Sign-OPT outperforms Prior-OPT, and it achieves state-of-the-art performance. We speculate the reason is that α_i is low in targeted attacks, which leads to a better performance of Prior-Sign-OPT according to Theorem A.4 and Theorem 3.3.

Figs. 11, 12 show the experimental results of the CIFAR-10 dataset. It shows that Prior-OPT outperforms Prior-Sign-OPT, and they also outperforms other baseline methods such as Sign-OPT, SurFree and GeoDA.

Figs. 13a, 13b, 13c show the theoretical comparison result by using the equations of $\mathbb{E}[\gamma^2]$ (Eq. (10) for Sign-OPT, Eq. (12) for Prior-Sign-OPT, and Eq. (17) for Prior-OPT). Fig. 13d illustrates the

Table 16: Untargeted attack results of Inception networks on the ImageNet dataset, where AUC denotes the area under the curve of mean ℓ_2 distortions versus the number of queries (the lower is better), and ASR denotes the attack success rate of the final adversarial example.

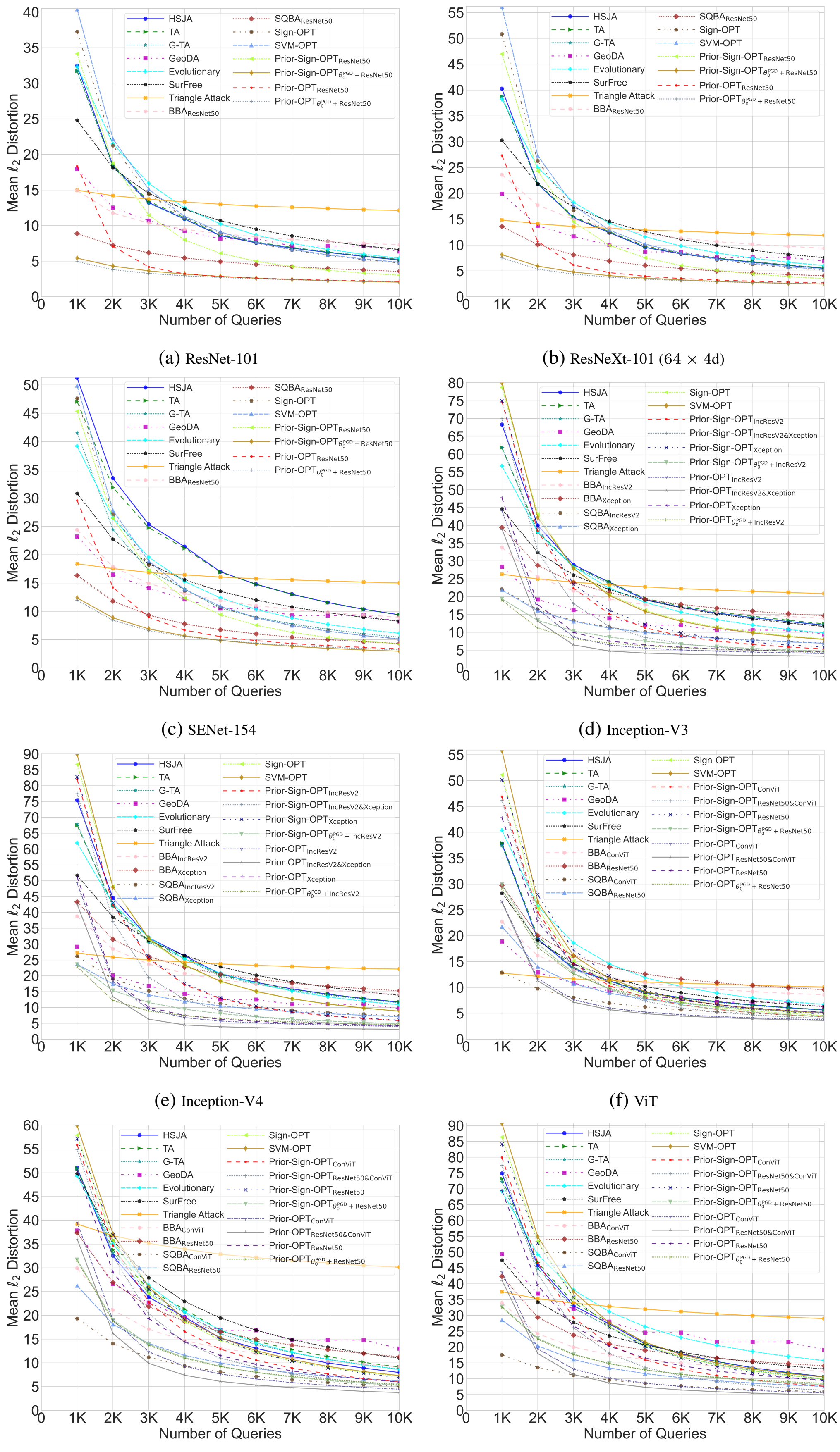
Method	Inception-V3			Inception-V4		
	Mean ℓ_2	AUC	ASR	Mean ℓ_2	AUC	ASR
HSJA	12.014	211938.7	81.1%	11.645	227700.5	82.1%
TA	12.378	208706.8	79.8%	11.694	219707.3	82.0%
G-TA	12.076	205670.7	81.5%	11.448	216797.7	83.3%
GeoDA	9.437	124150.7	87.8%	9.688	128665.4	87.7%
Evolutionary	9.809	192654.1	86.4%	10.839	215405.4	81.6%
Triangle Attack	20.878	205534.2	46.5%	22.132	214723.3	42.5%
SurFree	11.648	186094.3	79.1%	13.818	221197.5	69.7%
$\text{BBA}_{\text{IncResV2}}$	13.952	169881.3	69.0%	14.191	182033.7	68.8%
$\text{BBA}_{\text{Xception}}$	14.657	185798.3	67.2%	15.282	199287.2	63.6%
$\text{SQBA}_{\text{IncResV2}}$	7.020	98767.0	94.3%	7.417	110451.8	93.0%
$\text{SQBA}_{\text{Xception}}$	6.933	97022.6	94.7%	7.115	102939.0	92.2%
Sign-OPT	8.134	195118.7	91.9%	8.786	217576.3	89.8%
SVM-OPT	7.995	193289.7	92.3%	8.839	219673.6	89.0%
$\text{Prior-Sign-OPT}_{\text{IncResV2}}$	5.314	156261.6	97.8%	5.842	174243.9	96.3%
$\text{Prior-Sign-OPT}_{\text{Xception}}$	5.831	163715.5	96.8%	5.958	176950.5	95.7%
$\text{Prior-Sign-OPT}_{\text{IncResV2\&Xception}}$	4.225	130427.5	98.6%	4.199	142032.3	99.0%
$\text{Prior-Sign-OPT}_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	4.713	75088.7	97.1%	4.863	83066.8	96.9%
$\text{Prior-OPT}_{\text{IncResV2}}$	<u>4.067</u>	79685.9	<u>99.3%</u>	<u>4.027</u>	85290.8	99.1%
$\text{Prior-OPT}_{\text{Xception}}$	4.539	88915.0	<u>99.3%</u>	4.261	90492.0	<u>99.3%</u>
$\text{Prior-OPT}_{\text{IncResV2\&Xception}}$	**3.387**	**64461.8**	**99.7%**	**3.167**	**66110.1**	**99.8%**
$\text{Prior-OPT}_{\theta_0^{\text{PGD}}+\text{IncResV2}}$	4.496	<u>65031.0</u>	98.4%	4.548	<u>70165.3</u>	98.1%

performance of Prior-Sign-OPT in ℓ_2 norm untargeted attacks against the Swin Transformer using various values of q on the ImageNet dataset.

Figs. 14, 16, 18 present examples of adversarial images generated using varying numbers of queries from the Sign-OPT, Prior-Sign-OPT, and Prior-OPT methods. Figs. 15, 17, 19 present the corresponding adversarial perturbations of the Sign-OPT, Prior-Sign-OPT, and Prior-OPT. Initially, all methods begin with an image from the target class and iteratively minimize the ℓ_2 norm distance (also referred to as ℓ_2 distortion) to the original image, while maintaining the predicted label as the target class. Prior-Sign-OPT and Prior-OPT achieve a faster reduction in perturbation magnitude compared to Sign-OPT. After approximately $5{,}000$ queries, the visual features of the target-class image become almost imperceptible within the perturbations.

Table 17: Mean ℓ_∞ distortions of untargeted attacks across various query budgets on the ImageNet dataset.

Target Model	Method	Mean ℓ_∞ distortions				
		@1K	@2K	@5K	@8K	@10K
Inception-v3	TA Ma et al. (2021)	**0.397**	0.379	0.359	0.348	0.342
	Sign-OPT Cheng et al. (2020)	0.726	0.403	0.156	0.100	0.084
	SVM-OPT Cheng et al. (2020)	0.723	0.389	0.155	0.102	0.088
	Prior-Sign-OPT$_{\text{IncResV2}}$	0.678	0.365	0.117	0.086	0.080
	Prior-Sign-OPT$_{\text{IncResV2\&Xception}}$	0.640	0.318	**0.102**	**0.083**	**0.080**
	Prior-OPT$_{\text{IncResV2}}$	0.581	0.267	0.138	0.115	0.109
	Prior-OPT$_{\text{IncResV2\&Xception}}$	0.502	**0.208**	0.119	0.111	0.109
Inception-v4	TA Ma et al. (2021)	**0.420**	0.402	0.381	0.370	0.365
	Sign-OPT Cheng et al. (2020)	0.794	0.450	0.175	0.111	0.093
	SVM-OPT Cheng et al. (2020)	0.811	0.446	0.178	0.113	0.096
	Prior-Sign-OPT$_{\text{IncResV2}}$	0.756	0.408	0.136	0.097	0.089
	Prior-Sign-OPT$_{\text{IncResV2\&Xception}}$	0.700	0.348	**0.107**	**0.086**	**0.082**
	Prior-OPT$_{\text{IncResV2}}$	0.645	0.302	0.157	0.131	0.123
	Prior-OPT$_{\text{IncResV2\&Xception}}$	0.558	**0.218**	0.117	0.108	0.106
ResNet-101	TA Ma et al. (2021)	0.301	0.285	0.267	0.258	0.253
	Sign-OPT Cheng et al. (2020)	0.437	0.247	0.101	0.066	0.057
	SVM-OPT Cheng et al. (2020)	0.461	0.254	0.110	0.075	0.066
	Prior-Sign-OPT$_{\text{ResNet50}}$	0.404	0.218	**0.074**	**0.053**	**0.049**
	Prior-OPT$_{\text{ResNet50}}$	**0.289**	**0.138**	0.075	0.064	0.060
ResNeXt-101 (64 $\times$ 4d)	TA Ma et al. (2021)	**0.362**	0.344	0.323	0.313	0.307
	Sign-OPT Cheng et al. (2020)	0.611	0.326	0.131	0.090	0.078
	SVM-OPT Cheng et al. (2020)	0.667	0.336	0.131	0.089	0.078
	Prior-Sign-OPT$_{\text{ResNet50}}$	0.574	0.303	0.104	**0.075**	**0.069**
	Prior-OPT$_{\text{ResNet50}}$	0.428	**0.196**	**0.097**	0.080	0.075
SENet-154	TA Ma et al. (2021)	**0.355**	0.336	0.316	0.306	0.300
	Sign-OPT Cheng et al. (2020)	0.563	0.326	0.132	0.082	0.067
	SVM-OPT Cheng et al. (2020)	0.570	0.325	0.132	0.082	0.068
	Prior-Sign-OPT$_{\text{ResNet50}}$	0.536	0.314	**0.113**	**0.074**	**0.065**
	Prior-OPT$_{\text{ResNet50}}$	0.448	**0.246**	0.129	0.102	0.094
ViT	TA Ma et al. (2021)	**0.399**	0.379	0.358	0.348	0.342
	Sign-OPT Cheng et al. (2020)	0.602	0.302	0.105	0.072	0.064
	SVM-OPT Cheng et al. (2020)	0.651	0.310	0.107	0.075	0.068
	Prior-Sign-OPT$_{\text{ResNet50}}$	0.597	0.334	0.118	0.084	0.077
	Prior-Sign-OPT$_{\text{ResNet50\&ConViT}}$	0.539	0.273	**0.090**	**0.069**	**0.065**
	Prior-OPT$_{\text{ResNet50}}$	0.591	0.352	0.178	0.136	0.123
	Prior-OPT$_{\text{ResNet50\&ConViT}}$	0.429	**0.217**	0.124	0.110	0.106
GC ViT	TA Ma et al. (2021)	**0.380**	0.365	0.348	0.339	0.335
	Sign-OPT Cheng et al. (2020)	0.680	0.434	0.186	0.119	0.098
	SVM-OPT Cheng et al. (2020)	0.678	0.427	0.183	0.116	0.097
	Prior-Sign-OPT$_{\text{ResNet50}}$	0.670	0.445	0.183	0.116	0.097
	Prior-Sign-OPT$_{\text{ResNet50\&ConViT}}$	0.642	0.389	**0.141**	**0.092**	**0.079**
	Prior-OPT$_{\text{ResNet50}}$	0.652	0.455	0.248	0.185	0.163
	Prior-OPT$_{\text{ResNet50\&ConViT}}$	0.538	**0.305**	0.160	0.131	0.122
Swin Transformer	TA Ma et al. (2021)	**0.536**	0.515	0.491	0.479	0.472
	Sign-OPT Cheng et al. (2020)	1.009	0.625	0.258	0.159	0.128
	SVM-OPT Cheng et al. (2020)	1.036	0.622	0.251	0.157	0.131
	Prior-Sign-OPT$_{\text{ResNet50}}$	1.000	0.647	0.262	0.162	0.133
	Prior-Sign-OPT$_{\text{ResNet50\&ConViT}}$	0.909	0.513	0.169	**0.105**	**0.088**
	Prior-OPT$_{\text{ResNet50}}$	0.942	0.619	0.309	0.226	0.198
	Prior-OPT$_{\text{ResNet50\&ConViT}}$	0.662	**0.321**	**0.159**	0.129	0.120



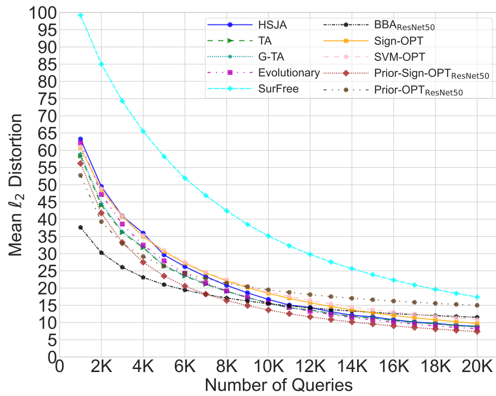
(a) ResNet-101 (b) ResNeXt-101 (64 × 4d)

(c) SENet-154 (d) Inception-V3

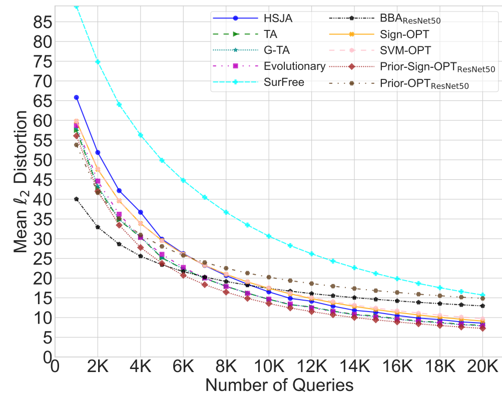
(e) Inception-V4 (f) ViT

(g) GC ViT (h) Swin Transformer

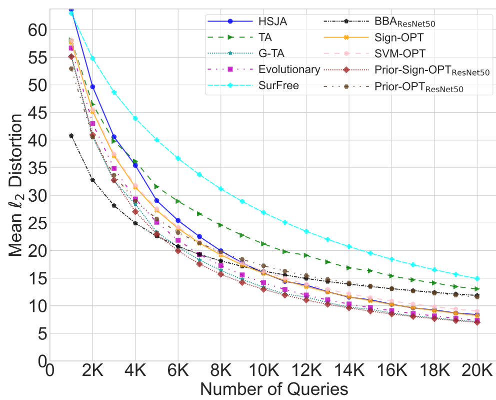
Figure 7: Mean distortions of ℓ_2 norm untargeted attack under different query budgets on the ImageNet dataset.



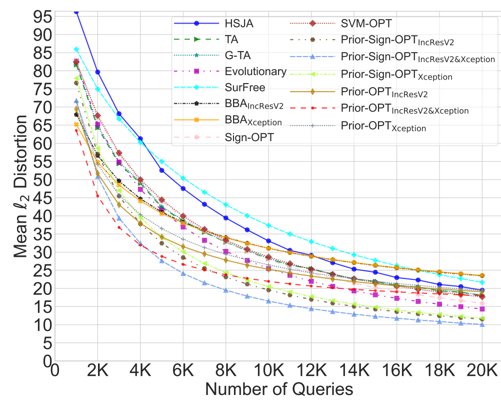
(a) ResNet-101



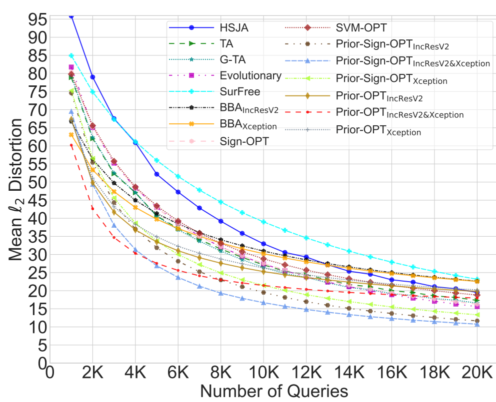
(b) ResNeXt-101 ($64 \times 4d$)



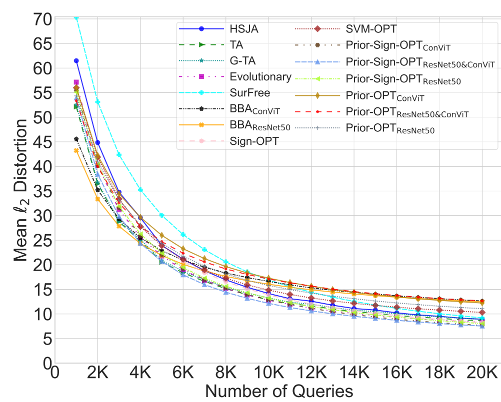
(c) SENet-154



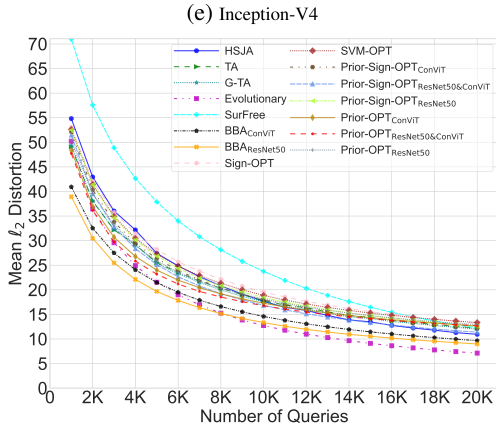
(d) Inception-V3



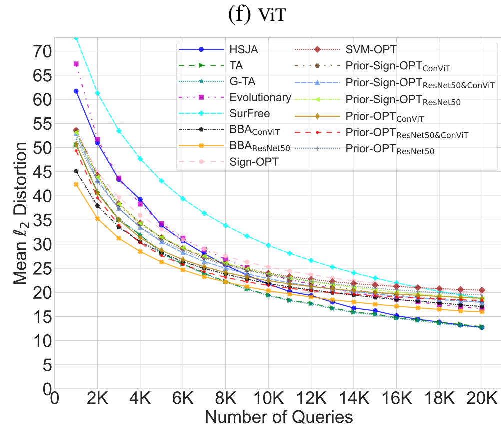
(e) Inception-V4



(f) ViT

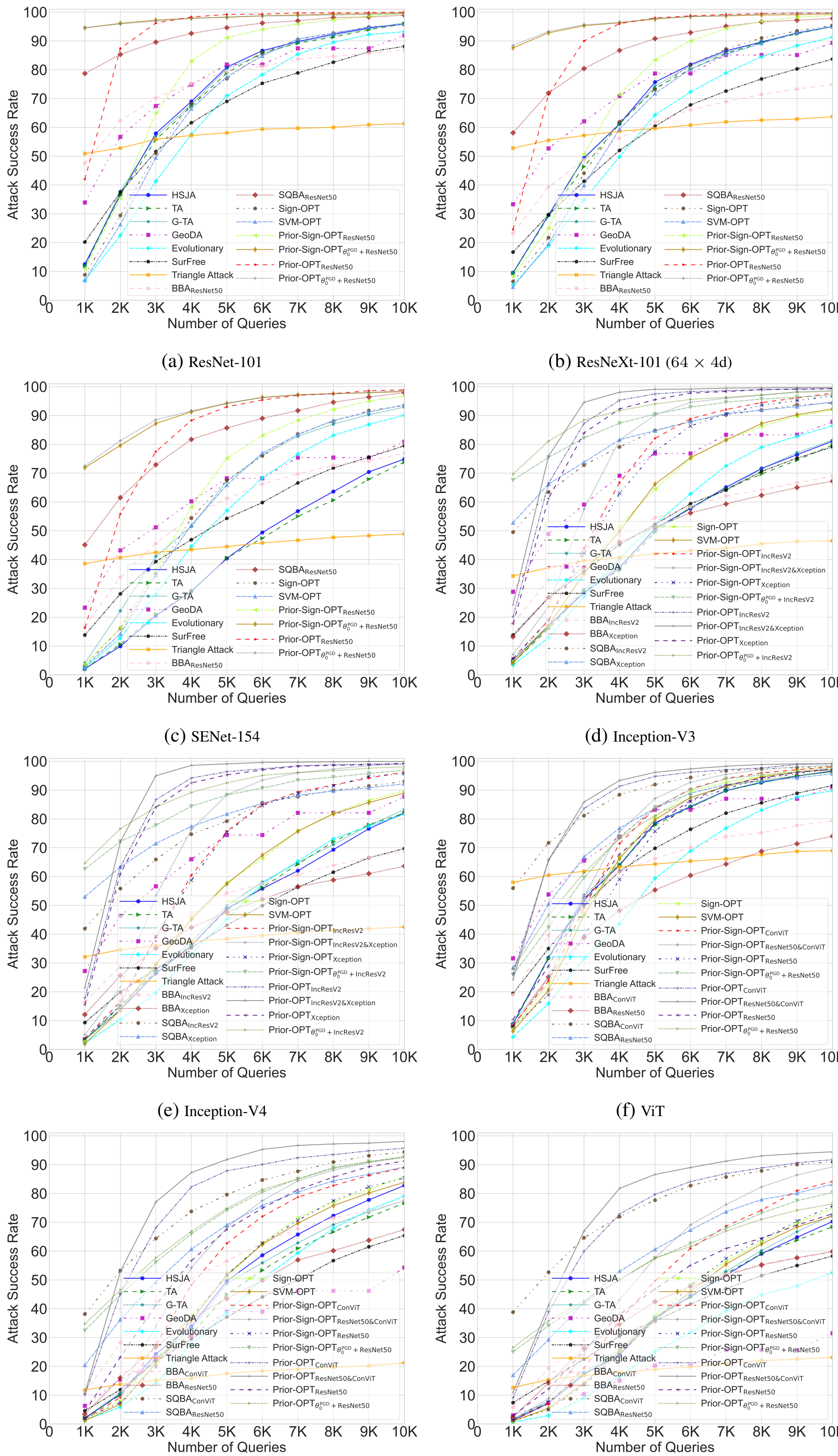


(g) GC ViT



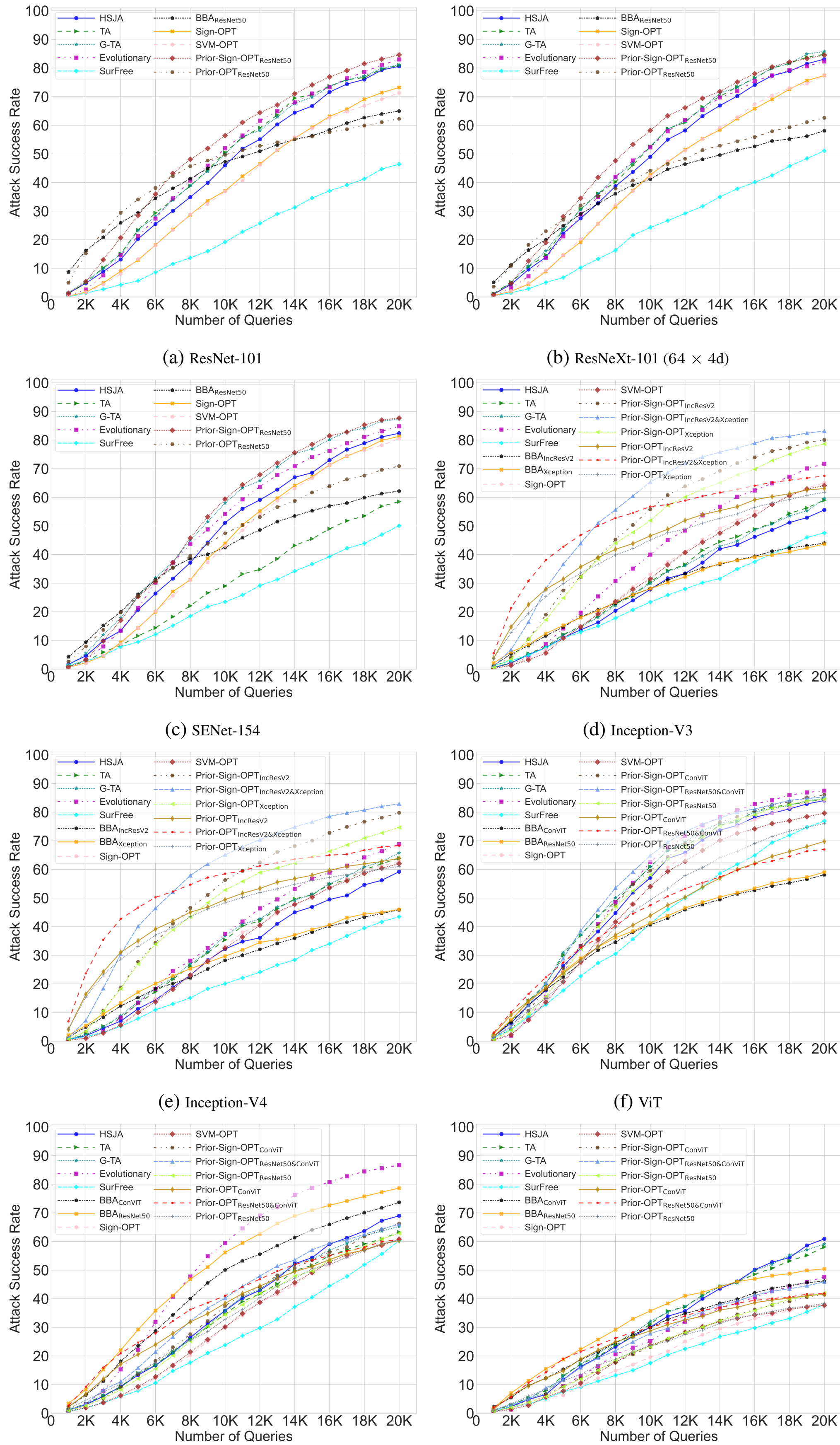
(h) Swin Transformer

Figure 8: Mean distortions of ℓ_2 norm targeted attacks under different query budgets on the ImageNet dataset.



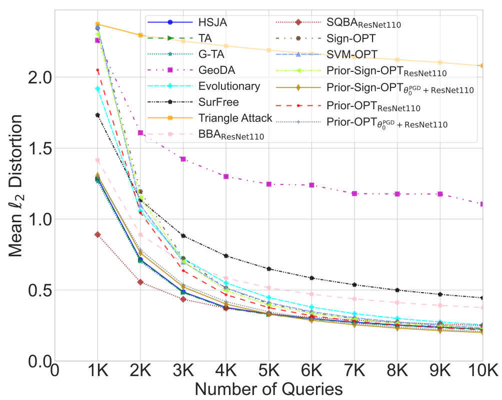
(a) ResNet-101
(b) ResNeXt-101 (64 × 4d)
(c) SENet-154
(d) Inception-V3
(e) Inception-V4
(f) ViT
(g) GC ViT
(h) Swin Transformer

Figure 9: Attack success rates of ℓ_2 norm untargeted attacks under different query budgets on the ImageNet dataset.

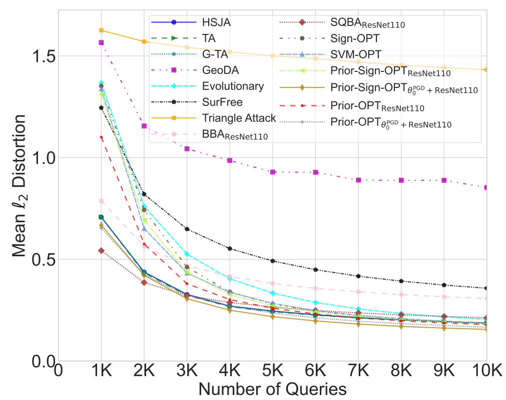


(a) ResNet-101
(b) ResNeXt-101 (64 × 4d)
(c) SENet-154
(d) Inception-V3
(e) Inception-V4
(f) ViT
(g) GC ViT
(h) Swin Transformer

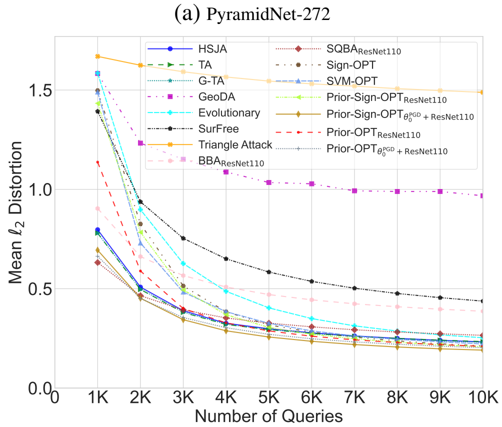
Figure 10: Attack success rates of ℓ_2 norm targeted attacks under different query budgets on the ImageNet dataset.



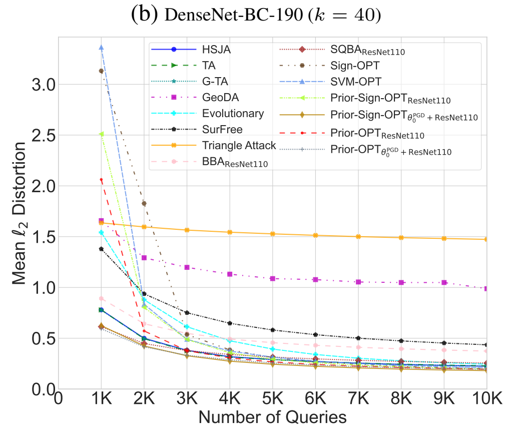
(a) PyramidNet-272



(b) DenseNet-BC-190 ($k = 40$)

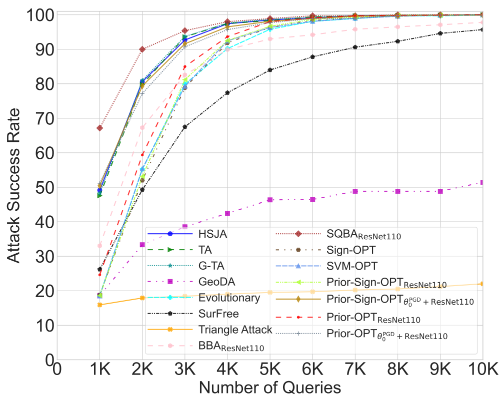


(c) WRN-28

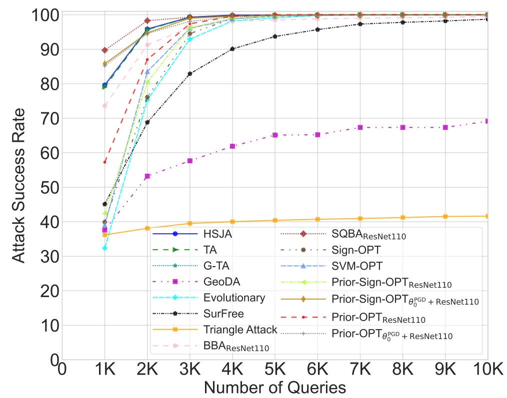


(d) WRN-40

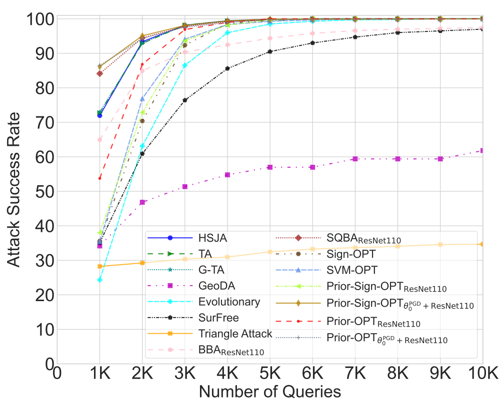
Figure 11: Mean distortions of ℓ_2 norm untargeted attack under different query budgets on the CIFAR-10 dataset.



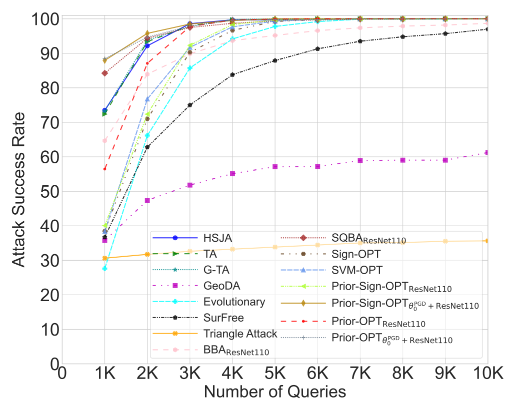
(a) PyramidNet-272



(b) DenseNet-BC-190 ($k = 40$)



(c) WRN-28



(d) WRN-40

Figure 12: Attack success rates of ℓ_2 norm untargeted attacks under different query budgets on the CIFAR-10 dataset.

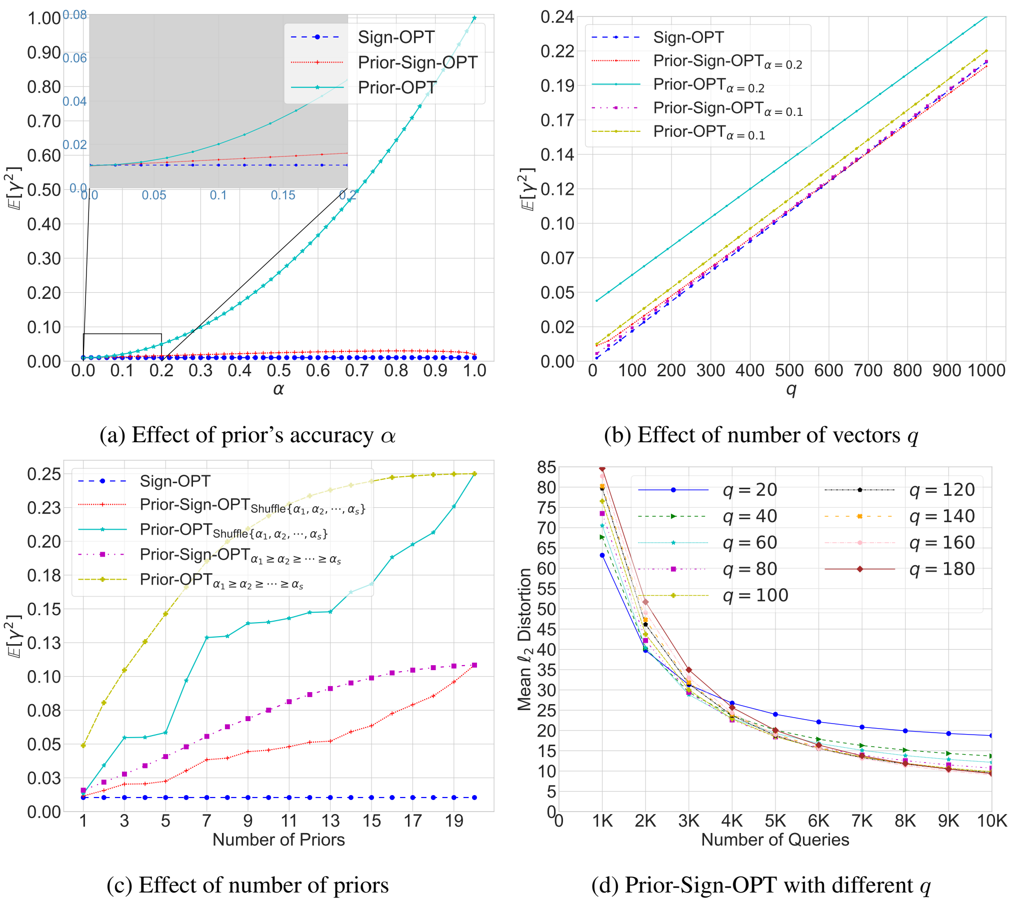
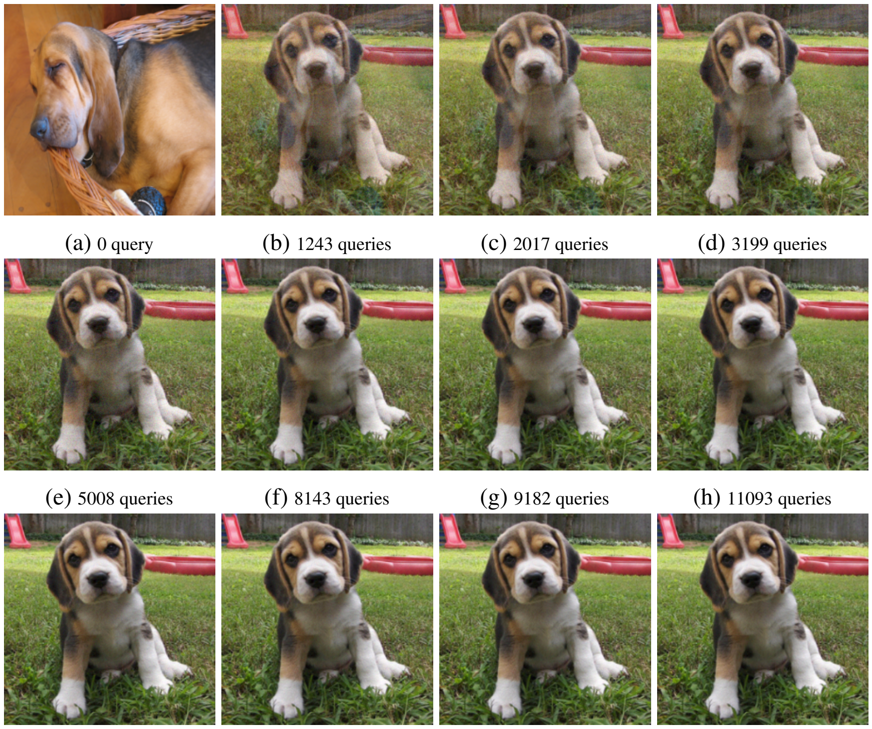


Figure 13: Experimental results of ablation studies of $\mathbb{E}[\gamma^2]$. Figs. 13a, 13b, 13c are based on theoretical results (Eqs. (10), (12) and (17)) with $d = 3072$. Fig. 13d demonstrates the results of attacking against Swin Transformer on the ImageNet dataset using Prior-Sign-OPT with different q.

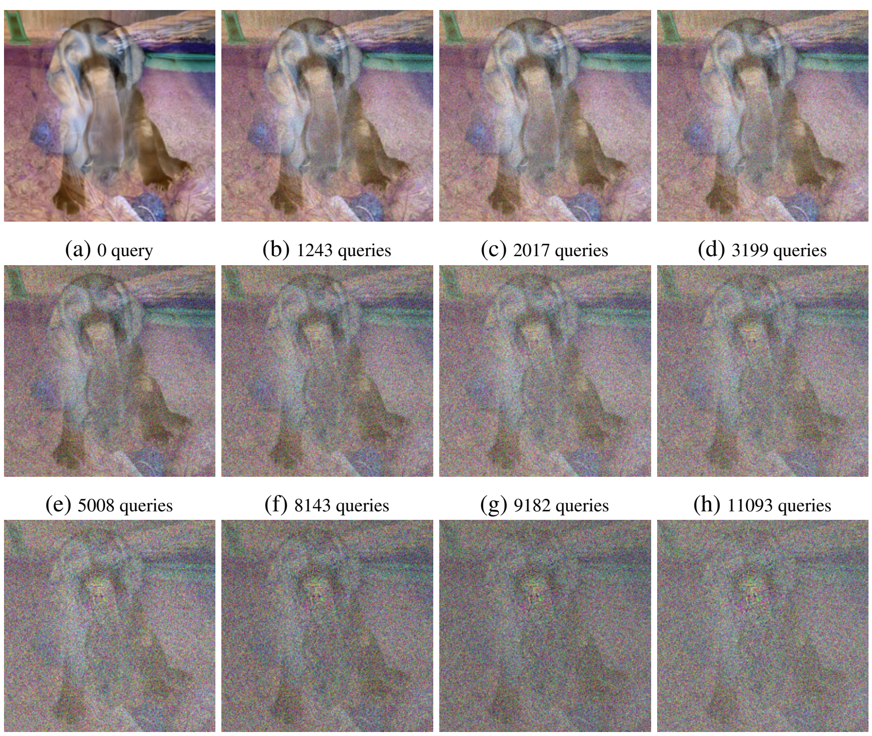


(a) 0 query (b) 1243 queries (c) 2017 queries (d) 3199 queries

(e) 5008 queries (f) 8143 queries (g) 9182 queries (h) 11093 queries

(i) 13051 queries (j) 15227 queries (k) 18101 queries (l) 20003 queries

Figure 14: Adversarial images generated with different queries in targeted Sign-OPT attacks against ResNet-101.



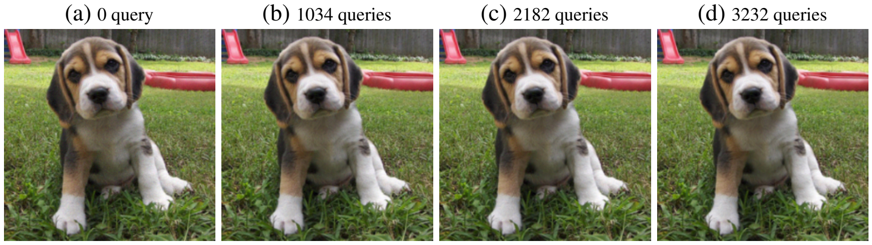
(a) 0 query (b) 1243 queries (c) 2017 queries (d) 3199 queries

(e) 5008 queries (f) 8143 queries (g) 9182 queries (h) 11093 queries

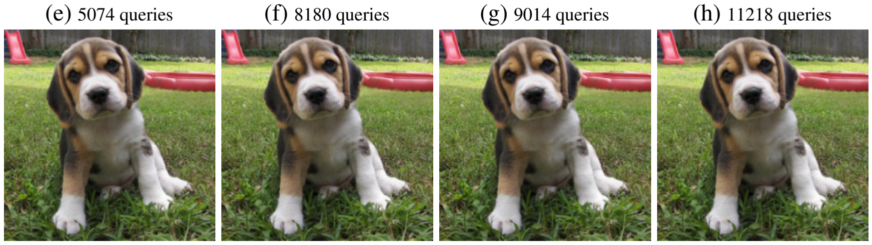
(i) 13051 queries (j) 15227 queries (k) 18101 queries (l) 20003 queries

Figure 15: The corresponding adversarial perturbations generated with different queries in targeted Sign-OPT attacks against ResNet-101.

(a) 0 query (b) 1034 queries (c) 2182 queries (d) 3232 queries



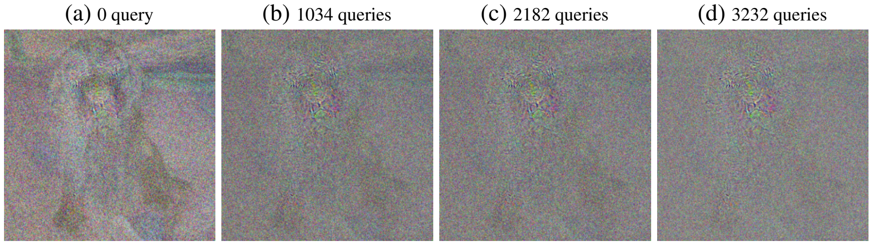
(e) 5074 queries (f) 8180 queries (g) 9014 queries (h) 11218 queries



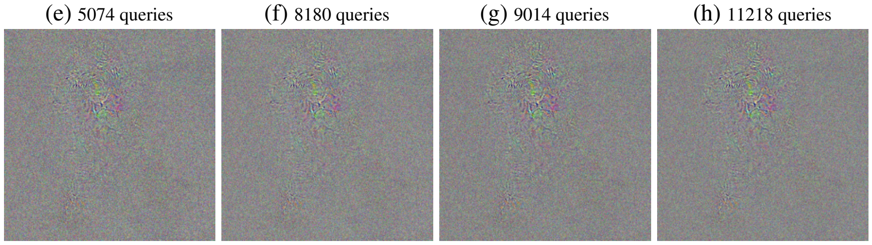
(i) 13117 queries (j) 15034 queries (k) 18101 queries (l) 20181 queries

Figure 16: Adversarial images generated with different queries in targeted Prior-Sign-OPT attacks against ResNet-101.

(a) 0 query (b) 1034 queries (c) 2182 queries (d) 3232 queries



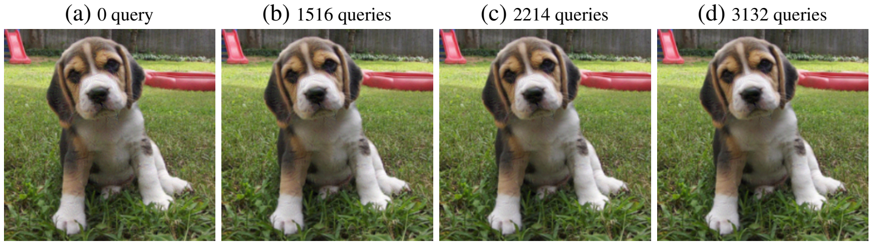
(e) 5074 queries (f) 8180 queries (g) 9014 queries (h) 11218 queries



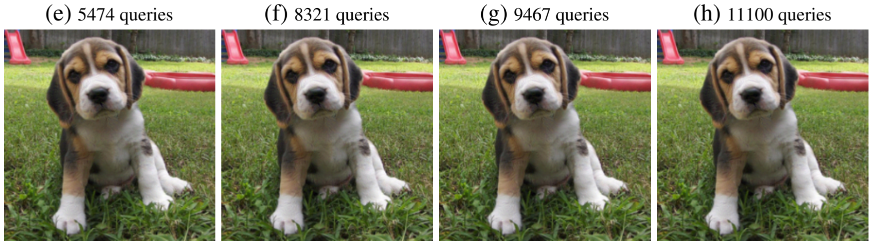
(i) 13117 queries (j) 15034 queries (k) 18101 queries (l) 20181 queries

Figure 17: The corresponding adversarial perturbations generated with different queries in targeted Prior-Sign-OPT attacks against ResNet-101.

(a) 0 query (b) 1516 queries (c) 2214 queries (d) 3132 queries

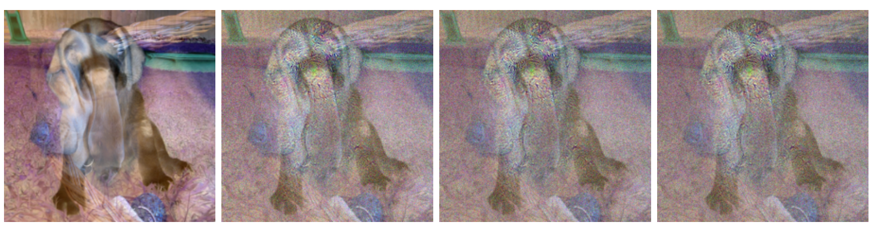


(e) 5474 queries (f) 8321 queries (g) 9467 queries (h) 11100 queries

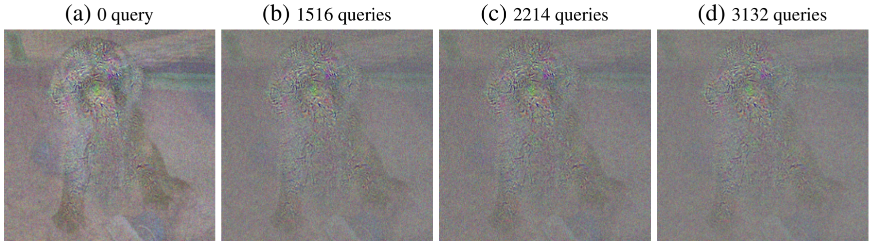


(i) 13227 queries (j) 15259 queries (k) 18376 queries (l) 20291 queries

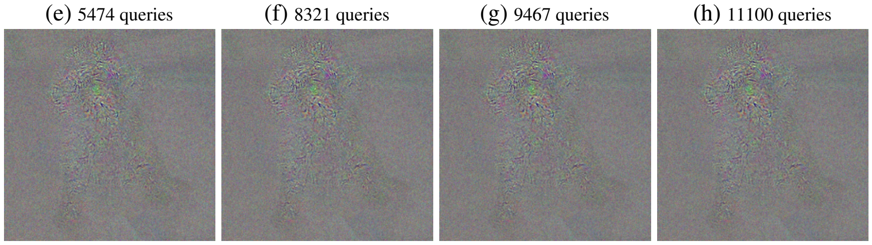
Figure 18: Adversarial images generated with different queries in targeted Prior-OPT attacks against ResNet-101.



(a) 0 query (b) 1516 queries (c) 2214 queries (d) 3132 queries



(e) 5474 queries (f) 8321 queries (g) 9467 queries (h) 11100 queries



(i) 13227 queries (j) 15259 queries (k) 18376 queries (l) 20291 queries

Figure 19: The corresponding adversarial perturbations generated with different queries in targeted Prior-OPT attacks against ResNet-101.