Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation

Jiawei Zhang [*1] **Linyi Li** [*2] **Huichen Li** [2] **Xiaolu Zhang** [3] **Shuang Yang** [4] **Bo Li** [2]

Abstract

Boundary based blackbox attack has been recognized as practical and effective, given that an attacker only needs to access the final model prediction. However, the query efficiency of it is in general high especially for high dimensional image data. In this paper, we show that such efficiency highly depends on the scale at which the attack is applied, and attacking at the optimal scale significantly improves the efficiency. In particular, we propose a theoretical framework to analyze and show three key characteristics to improve the query efficiency. We prove that there exists an *optimal* scale for projective gradient estimation. Our framework also explains the satisfactory performance achieved by existing boundary blackbox attacks. Based on our theoretical framework, we propose **P**rogressive-**S**cale enabled projective **B**oundary **A**ttack (PSBA) to improve the query efficiency via progressive scaling techniques. In particular, we employ Progressive-GAN to optimize the scale of projections, which we call PSBA-PGAN. We evaluate our approach on both spatial and frequency scales. Extensive experiments on MNIST, CIFAR-10, CelebA, and ImageNet against different models including a real-world face recognition API show that PSBA-PGAN significantly outperforms existing baseline attacks in terms of query efficiency and attack success rate. We also observe relatively stable optimal scales for different models and datasets. The code is publicly available at `https://github.com/AI-secure/PSBA`.

[*]Equal contribution [1]Zhejiang University, China (work done during remote internship at UIUC) [2]UIUC, USA [3]Ant Financial, China [4]Alibaba Group US, USA. Correspondence to: Linyi Li <linyi2@illinois.edu>, Bo Li <lbo@illinois.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

Blackbox attacks against machine learning (ML) models have raised great concerns recently given the wide application of ML (Krizhevsky et al., 2009; He et al., 2016; Vaswani et al., 2017). Among these blackbox attacks, boundary blackbox attack (Brendel et al., 2018; Chen et al., 2020) has shown to be effective. However, it usually requires a large number of queries against the target model given the high-dimensional search space. Recent research shows that it is possible to sample the queries from a lower dimensional sampling space first and project them back to the original space for gradient estimation to reduce the query complexity (Li et al., 2020a; 2021). These observations raise additional questions: *What is the "optimal" projection subspace that we can sample from? How effective such projective attack would be? What is the query complexity for such projective gradient estimation approach?* To answer these questions, in this paper we analyze the general Projective Boundary blackbox Attack framework (PBA), for which only the decision boundary information (i.e. label) is revealed, and propose Progressive-Scale enabled projective Boundary blackbox Attack (PSBA) to gradually search for the optimal sampling space, which can boost the query efficiency of PBA both theoretically and empirically.

The overall pipeline of PSBA and the corresponding analysis framework are shown in Fig. 1, where an attacker starts with a *source image* and manipulate it to be "visually close" with a *target image* while preserving its label and therefore fool a ML model via *progressive-scale* based projective gradient estimation. In particular, an attacker first conducts binary search to find a boundary point based on the source image; then samples several perturbation vectors from a low-dimensional sampling space and project them back to the *projection subspace* via a projection function $\boldsymbol{f}$ to estimate the gradient. Finally, the attacker will move along the estimated gradient direction to construct the adversarial instance. The main goal of PSBA is to search for the optimal *projection subspace* for attack efficiency and effectiveness purpose, and we explore such optimal projection subspace both theoretically and empirically.

Theoretically, we develop a *general framework* to analyze the query efficiency of gradient estimation for PBA. With the

framework, we 1) provide the expectation and concentration bounds for cosine similarity between the estimated and true gradients based on nonlinear projection functions, while previous work only considers identical projection (Chen et al., 2020) or orthogonal sampling (Li et al., 2020a; 2021); 2) discover several key characteristics that contribute to tighter gradient estimation, including small dimensionality of the projection subspace, large projected length of the true gradient onto the projection subspace, and high sensitivity on the projected true gradient direction; 3) analyze the trade-off between small dimensionality and large projected length, and prove the existence of an optimal subspace dimensionality, i.e., an optimal scale; 4) prove that choosing a subspace with large projected length of true gradient as the projection subspace can improve the query efficiency of gradient estimation. This framework not only provides theoretical justification for existing PBAs (Brendel et al., 2018; Chen et al., 2020; Li et al., 2020a; 2021), but also enables the design of more efficient blackbox attacks, where the proposed PSBA is an example.

Inspired by our theoretical analysis, we design PSBA to progressively search for the optimal scale to perform projective gradient estimation for boundary blackbox attacks. We first consider the spatial scale (i.e., resolution of images), and apply PSBA to search over different spatial scales. We then extend PSBA to the frequency scale (i.e., threshold of low-pass filter) and spectrum scale (i.e., dimensionality of PCA). In particular, as a demonstration, we instantiate PSBA with Progressive-GAN (PGAN), and we conduct extensive experiments to 1) justify our theoretical analysis on key characteristics that contribute to tighter gradient estimation and 2) show that PSBA-PGAN *consistently and significantly* outperforms existing boundary attacks such as HSJA (Chen et al., 2020), QEBA (Li et al., 2020a), NonLinear-BA (Li et al., 2021), EA (Dong et al., 2019), and Sign-OPT (Cheng et al., 2019a) on various datasets including MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), CelebA (Liu et al., 2015), and ImageNet (Deng et al., 2009) against different models including a real-world face recognition API.

Technical Contributions. In this paper, we take the *first* step towards exploring the impacts of different projection scales of projection space in boundary blackbox attacks. We make contributions on both theoretical and empirical fronts.

- We propose the first theoretical framework to analyze boundary blackbox attacks with general projection functions. Using this framework, we derive tight expectation and concentration bounds for the cosine similarity between estimated and true gradients.
- We characterize the key characteristics and trade-offs for a good projective gradient estimator. In particular, we theoretically prove the existence of the optimal scale of the projection space.

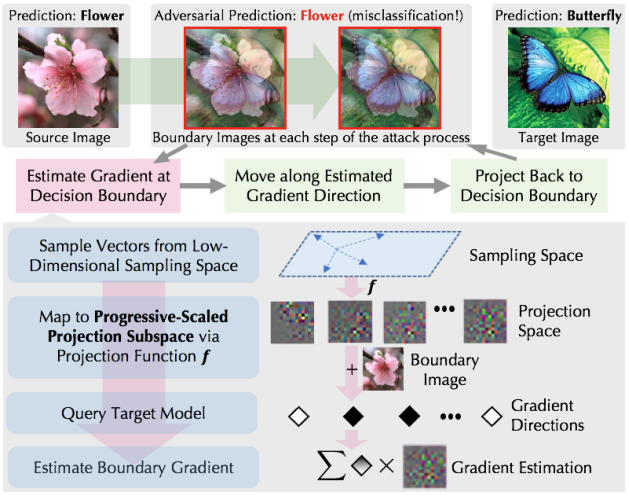

Figure 1. An overview of the progressive-scale boundary blackbox attack (PSBA) via projective gradient estimation.

- We propose Progressive-Scale based projective Boundary Attack (PSBA) via progressively searching for the optimal scale in a self-adaptive way under spatial, frequency, and spectrum scales.
- We instantiate PSBA by PGAN for empirical evaluation. Extensive experiments show that PSBA-PGAN outperforms the state-of-the-art boundary attacks on MNIST, CIFAR-10, CelebA, and ImageNet against different blackbox models and a real-world face recognition API.

Related Work. Adversarial attacks against ML have been conducted to explore the vulnerabilities of learning models and therefore improve their robustness (Szegedy et al., 2013; Kurakin et al., 2016). Most existing attacks (Goodfellow et al., 2014; Madry et al., 2018) assume the white-box setting, where the attacker has full access to the target model including its structure and weights. However, in practice, such as commercial face recognition APIs (MEGVII, 2020c), we cannot access the full model. Thus, several blackbox attacks have been proposed, which mainly fall into three categories: transfer-based, query-based, and hybrid blackbox attacks: 1) The *transfer-based attacks* usually train a surrogate model, attack the surrogate model using white-box attacks, and exploit the adversarial transferability (Papernot et al., 2016; Tramèr et al., 2017) to use the generated adversarial examples to attack the blackbox model. 2) The *query-based attacks* can be further divided into score-based and decision-based. The *score-based attacks* assume that we know the confidence score of the target model's prediction and therefore estimate the gradient based on the prediction scores (Chen et al., 2017; Bhagoji et al., 2018; Tu et al., 2019; Ilyas et al., 2018; Cheng et al., 2019b; Chen et al., 2017; Li et al., 2020b). The *decision-based attacks* assume that we only know the final decision itself which is more practical, such as the boundary attack (Brendel et al., 2018),

EA (Dong et al., 2019) and Sign-OPT (Cheng et al., 2019a). HSJA (Chen et al., 2020) extends the boundary attack by adopting a sampling-based gradient estimation component to guide the search direction. Later on, QEBA (Li et al., 2020a) and NonLinear-BA (Li et al., 2021) are proposed to use a projection function to sample from a lower dimensional space to improve the sampling efficiency. Our work focuses on decision-based attacks. Specifically, our work systematically studies the projective gradient estimation for boundary attacks and reveals that *progressive-scale enabled projection* could improve the query efficiency both theoretically and empirically. 3) The *hybrid attacks* usually train one or multiple surrogate models and leverage the gradient information (Guo et al., 2019; Tashiro et al., 2020) or adversarial examples (Suya et al., 2020) to guide the generation of queries for the target model.

Progressive scaling has long been an effective methodology for different tasks, such as pyramidal-structured objection detection (Lin et al., 2017; Zhang et al., 2020), high-resolution generative neural networks (Karras et al., 2018; Zhang & Khoreva, 2019; Wu et al., 2020), and deep feature extraction (Cai et al., 2016; Ma et al., 2020). In this work, we aim to explore whether it is possible to progressively conduct queries from different subspaces (e.g., spatial and frequency subspaces) against a blackbox machine learning model to perform query efficient blackbox attacks.

2. Preliminaries

In this section, we introduce the related notations for our projective gradient estimation. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For arbitrary two vectors a and b, let $\langle a, b\rangle$ and $\cos\langle a, b\rangle$ denote their dot product and cosine similarity respectively. For a matrix $\boldsymbol{W} \in \mathbb{R}^{n\times m}$ and a vector $v \in \mathbb{R}^n$, we denote its projection on span$(\boldsymbol{W})$ by $\text{proj}_{\boldsymbol{W}} v$.[1] Without loss of generality, we focus on the adversarial attack on an image classifier $G : \mathbb{R}^n \to \mathbb{R}^C$ where C is the number of classes. For given $x \in \mathbb{R}^n$, G predicts the label with highest confidence: $\arg\max_{i\in[C]} G(x)_i$. As for the *threat model*, we consider the practical setting where only the *decision* of the classification model G is accessible for attackers.

Following the literature (Chen et al., 2020; Li et al., 2020a), we define difference function $S(\cdot)$ and sign function $\phi(\cdot)$.

Definition 1. Let label y_0 be model G's prediction for input x^*. For the targeted attack, the adversarial target is $y' \in [C]$. Define the difference function $S_{x^*} : \mathbb{R}^n \to \mathbb{R}$ as below:

$$S_{x^*}(x) := \begin{cases} \max\limits_{y\in[C]:y\neq y_0} G(x)_y - G(x)_{y_0}, & \text{(untargeted attack)} \\ G(x)_{y'} - \max\limits_{y\in[C]:y\neq y'} G(x)_y. & \text{(targeted attack)} \end{cases}$$

[1] From linear algebra, $\text{proj}_{\boldsymbol{W}} v = \boldsymbol{W}(\boldsymbol{W}^\intercal \boldsymbol{W})^{-1}\boldsymbol{W}^\intercal v$.

The function $\phi_{x^*} : \mathbb{R}^n \to \{\pm 1\}$ is the sign function of S_{x^*}:

$$\phi_{x^*}(x) := \begin{cases} +1 & \text{if } S_{x^*}(x) > 0, \\ -1 & \text{otherwise.} \end{cases}$$

When there is no ambiguity, we may abbreviate them as $S(x)$ and $\phi(x)$ respectively. For a target image x^*, the attacker crafts an image x, ensuring that the difference function $S_{x^*}(x) \geq 0$ to perform a success attack while minimizing the distance $\|x - x^*\|_2$.

We call x a *boundary point* if $S_{x^*}(x) = 0$. We assume that S_{x^*} is β_S-smooth. Formally, for any $x, z \in \mathbb{R}^n$[2],

$$\frac{\|\nabla S_{x^*}(x) - \nabla S_{x^*}(z)\|_2}{\|x - z\|_2} \leq \beta_S. \tag{1}$$

Note that we can only query the value of ϕ_{x^*} instead of S_{x^*} according to the threat model. In boundary attack, we estimate the gradient of S by querying ϕ.

We generalize existing gradient estimators for boundary attack in a projective form. Suppose we have a projection $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$, where $m \leq n$. This projection can be obtained in various ways, such as PCA (Li et al., 2020a), VAE (Li et al., 2021), or just an identical projection (Chen et al., 2020). Similar to Eq. (1), we assume $\boldsymbol{f}$ is $\beta_{\boldsymbol{f}}$-smooth[2]. With the projection, given an input x_t that is at or close to the boundary of S_{x^*}, with pre-determined step size δ_t, we can estimate gradient $\nabla S_{x^*}(x_t)$ as such:

Definition 2 (Boundary Gradient Estimator)**.** On $x_t \in \mathbb{R}^n$, with pre-determined step size δ_t, let $\{u_b\}_{b=1}^B$ be B vectors that are uniformly sampled from the m-dimensional unit sphere S^{m-1}. The gradient $\nabla S_{x^*}(x_t)$ is estimated by

$$\widetilde{\nabla S}_{x^*,\delta_t}(x_t) := \frac{1}{B}\sum_{b=1}^{B} \phi_{x^*}\left(x_t + \Delta\boldsymbol{f}(\delta_t u_b)\right)\Delta\boldsymbol{f}(\delta_t u_b), \tag{2}$$

where $\Delta\boldsymbol{f}(x) := \boldsymbol{f}(x) - \boldsymbol{f}(0)$. When there is no ambiguity, we may omit subscript x^* or δ_t. Note that each query of ϕ_{x^*} is a query to the blackbox model, so computing $\widetilde{\nabla S}(x_t)$ requires B queries to the blackbox model.

As previous work shows, the expected cosine similarity between the estimated and true gradients can be bounded. For example, with *identical* projection $\boldsymbol{f}$ (Chen et al., 2020),

$$\cos\langle\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq 1 - \frac{9\beta_S^2\delta_t^2 n^2}{8\|\nabla S(x_t)\|_2^2}. \tag{3}$$

Definition 3 (Sampling Space and Projection Space)**.** For the given projection $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$. We call the domain $\text{dom}(\boldsymbol{f}) = \mathbb{R}^m$ the *sampling space*, and the subspace consisting of projected images the *projection subspace*.

[2] For any non-differentiable point of S_{x^*} or $\boldsymbol{f}$, the assumption generalizes to any Clarke's generalized gradient (Clarke et al., 2008). For simplicity, we assume S_{x^*} and $\boldsymbol{f}$ are differentiable hereinafter. The $\|\cdot\|_2$ operator stands for the ℓ_2 norm for vector ∇S and the spectral norm for Jacobian matrix $\nabla\boldsymbol{f}$.

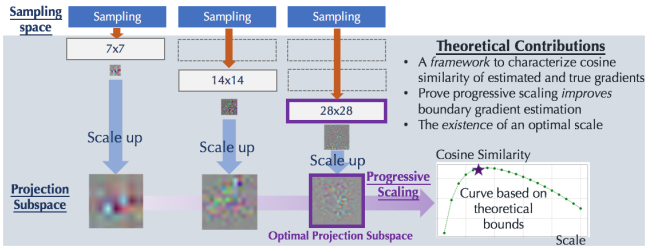

Figure 2. PSBA with progressive scalling on spatial domain.

Remark. The definition reflects the sample-and-project process in gradient estimation. The projection subspace is a subspace of the original input space $\mathbb{R}^n$. Typically, the step size δ_t is small, and $\Delta \boldsymbol{f}(\delta u) \approx \nabla \boldsymbol{f}(0) \cdot (\delta u)$ by linear approximation. Therefore, we can view $\boldsymbol{f}$ as a non-singular linear projection where m is the dimensionality of the projection subspace.

3. Progressive-Scale Blackbox Attack: PSBA

In this section we will introduce the general pipeline of proposed **P**rogressive-**S**cale projective **B**oundary **A**ttack (PSBA), followed by detailed analysis of it in Section 4. On the high level, PSBA progressively increases the scale of *projection subspace* where the perturbation vectors will be chosen from, until it reaches the "optimal" scale as shown in Figure 2. The scales are sampled from domains such as spatial, frequency, and spectral domain. At each scale, PSBA is composed of two stages: *training stage* and *attack stage*.

At the *training stage*, we train a generative model M (e.g., GAN), with gradient images obtained from any model. The input space of M is a low-dimensional space (m). For instance, from the spatial domain, we can sample from different resolutions such as 7×7, 14×14, and 28×28. Then we leverage M to project the sampled vectors back to the original space with interpolation. These projected images form the *projection subspace*.

At the *attack stage*, we adapt the boundary attack pipeline, and first select a source image $\widehat{x_0}$ (drawn from images from the adversarial target class). In each iteration t, the attack reduces the distance from current adversarial sample x_t to the target image x^* via three steps: (1) binary search for the boundary point x_t where $S_{x^*}(x_t) = 0$ on the line connecting $\widehat{x_{t-1}}$ and x^*, with pre-determined precision threshold θ; (2) estimate the gradient at x_t using the boundary gradient estimator (Definition 2) with step size δ_t; (3) normalize the estimated gradient, and perform a step of gradient ascent to get $\widehat{x_t}$. Note that in (1), the binary search for boundary point requires $\mathcal{O}(\log 1/\theta)$ queries to the blackbox model, and in (2), the gradient estimation requries B samples and queries to the blackbox model. In Section 4 we will analyze the relation between B and the quality of estimated gradients in terms of cosine similarity.

We select the optimal scale for projection subspace based on a validation set. If with current scale, after $1,000$ queries, the average distance to the target image x^* is smaller than that of the previous scale, we try a new increased scale. Otherwise, i.e., the average distance is larger than that of the previous one, we select the previous scale as the optimal scale, and use it as the scale of projection subspace. This process can be viewed as climbing to find the maximum point on the curve in Fig. 2 from left to right. Detailed pseudocode can be found in Appendix D.4.

In particular, we mainly consider different scales in the spatial, frequency, and spectrum domains. For spatial domain, M on the progressively grown scale (i.e., resolution) can be effectively trained via Progressive-GAN. The frequency scales correspond to thresholds for the low-pass frequency filter, and the spectrum scales correspond to dimensionalities of PCA. For frequency and spectrum, M can be trained on full-scale and trimmed to fit in the required lower scale as discussed in Section 4.2.

4. Analysis of Gradient Estimation

In this section, we analyze the similarity between the estimated and true gradients for general projective boundary attack frameworks. These attacks all follow Definition 2 to estimate the gradient. Our goal is to improve the gradient similarity while reducing the number of queries (B).

In Section 4.1, we present cosine similarity bounds between the estimated and true gradients for the gradient estimator with general **nonlinear projection function**, and analyze the key characteristics that improve such gradient estimation. In Section 4.2, we analyze the **bounds of cosine similarity** when the output of projection $\boldsymbol{f}$ is constrained on selected projection subspace. We show how constraining on a representative subspace improves gradient estimation compared with performing gradient estimation on the original space, which explains why PSBA outperforms existing methods.

4.1. General Cosine Similarity Bounds

Let $\nabla \boldsymbol{f}(0) \in \mathbb{R}^{n \times m}$ be the Jacobian matrix of the projection $\boldsymbol{f}$ at the origin. Throughout the paper, we assume that $\nabla \boldsymbol{f}(0)$ has full-rank since $\nabla \boldsymbol{f}$ is non-singular in general case. We further assume that there exists a column vector $\nabla \boldsymbol{f}(0)_{:,c}$ that is *aligned with* the projected true gradient $\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)$, and other column vectors are orthogonal to it. Formally, there exists $c \in [m]$, such that $\nabla \boldsymbol{f}(0)_{:,c} = k\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)$ with $k \neq 0$, and for any $i \neq c$, $\langle \nabla \boldsymbol{f}(0)_{:,i}, \nabla \boldsymbol{f}(0)_{:,c} \rangle = 0$. This assumption guarantees that the projection model $\boldsymbol{f}$ produces *no* directional sampling bias for true gradient estimation (Lemma A.3) following the standard setting. We remark that since vectors tend to be orthogonal to each other in high-dimensional

case (Fiers, 2018), this assumption holds with higher confidence in high-dimensional cases. In Section 5.2, we empirically verify this assumption.

Lemma 4.1 ($\nabla \boldsymbol{f}$ Decomposition)**.** *Under the above assumption, there exists a singular value decomposition of* $\nabla \boldsymbol{f}(0) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$ *such that*

$$\boldsymbol{U}_{:,1} = \text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t) / \|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2$$
$$\text{or } \boldsymbol{U}_{:,1} = -\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t) / \|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$, $\boldsymbol{V} \in \mathbb{R}^{m\times m}$ *are orthogonal matrices;* $\boldsymbol{\Sigma} = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_m) \in \mathbb{R}^{n\times m}_{\geq 0}$ *is a rectangular diagonal matrix with* $\alpha_1 > 0$*. Denote* $\max_{i\in[m]} \alpha_i$ *as* $\alpha_{\max}$.

The proof can be found in Appendix A.3. Compared with the standard SVD decomposition, now the first column vector of $\boldsymbol{U}$ can be fixed to the normalized projected true gradient vector or its opposite.

Definition of Sensitivity. In Lemma 4.1, for projection $\boldsymbol{f}$, we can view the resulting α_1 as the *sensitivity* of the projection model on the (projected) true gradient direction; and $\{\alpha_i\}_{i=2}^m$ as the *sensitivity* of the projection model on directions orthogonal to the true-gradient. With higher sensitivity on projected true gradient direction (α_1) and smaller sensitivity on other orthogonal directions ($\{\alpha_i\}_{i=2}^m$), the gradient estimation becomes better as we will show later.

Main Theorems

Here we will present our main theorems for the expectation (Theorem 1) and concentration bound (Theorem 2) of cosine similarity between the estimated and true gradients.

Theorem 1 (Expected cosine similarity)**.** *The difference function* S *and the projection* $\boldsymbol{f}$ *are as defined before. For a point* x_t *that is* θ*-close to the boundary, i.e., there exists* $\theta' \in [-\theta, \theta]$ *such that* $S(x_t + \theta' S(x_t)/\|\nabla S(x_t)\|_2) = 0$*, let estimated gradient* $\widetilde{\nabla S}(x_t)$ *be as computed by Definition 2 with step size* δ *and sampling size* B*. Over the randomness of the sampled vectors* $\{u_b\}_{i=1}^B$*,*

$$\cos\langle \mathbb{E}\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2} \cdot \left(1 - \frac{(m-1)^2\delta^2}{8\alpha_1^2}\left(\frac{\delta\gamma^2}{\alpha_1} + \frac{\gamma}{\alpha_1}\sqrt{\frac{\sum_{i=2}^m \alpha_i^2}{m-1}} + \frac{1.58\beta_{\boldsymbol{f}}}{\sqrt{m-1}} + \frac{\gamma\theta}{\alpha_1\delta} \cdot \frac{\|\nabla S(x_t)\|_2}{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2}\right)^2\right),$$

where

$$\gamma := \beta_{\boldsymbol{f}} + \frac{\beta_S\left(\max_{i\in[m]} \alpha_i + 1/2\delta\beta_{\boldsymbol{f}}\right)^2 + \beta_S\theta^2/\delta^2}{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2}. \quad (4)$$

Proof Sketch. The high-level idea is using Taylor expansion with Lagrange remainder to control both the first-order and higher-order errors, and plug the error terms into the distribution of dot product between $\nabla S(x_t)$ and $\nabla \boldsymbol{f}(0) \cdot u_b$. This dot product follows a linearly transformed Beta distribution (Chen et al., 2020). The error terms are separately controlled for the projected gradient direction (i.e., direction of $\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)$) and other orthogonal directions. Then the controlled directional errors are combined as an ℓ_2-bounded error vector. The complete proof is deferred to Appendix A.6. □

Remark. This bound characterizes the expected cosine similarity of the boundary gradient estimator. For an identical projection $\boldsymbol{f}$, if x_t is exactly the boundary point, i.e., $S(x_t) = 0$, from the theorem we get

$$\cos\langle \mathbb{E}\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq 1 - \frac{(n-1)^2\delta^2\beta_S^2}{2\|\nabla S(x_t)\|_2^2},$$

where we leverage the fact that $\delta = O(1/n)$ is usually small. This bound is of the same order as the previous work (Chen et al., 2020) shown in Eq. (3), while containing a tighter constant $1/2$ instead of $9/8$.

Suppose both the difference function S and projection $\boldsymbol{f}$ are linear, i.e., $\beta_S = \beta_{\boldsymbol{f}} = 0$, then we have $\cos\langle \mathbb{E}\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2}$. From Lemma A.1, this is the optimal cosine similarity obtainable with the projection $\boldsymbol{f}$. Furthermore, for identical projection, $\cos\langle \mathbb{E}\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq 1$, which verifies the optimality of our bound over existing work (Li et al., 2020a; 2021).

Theorem 2 (Concentration of cosine similarity)**.** *Under the same setting as Theorem 1, over the randomness of the sampled vector* $\{u_b\}_{i=1}^B$*, with probability* $1 - p$*,*

$$\cos\langle \widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2} \cdot \left(1 - \frac{(m-1)^2\delta^2}{8\alpha_1^2}\left(\frac{\delta\gamma^2}{\alpha_1} + \frac{\gamma}{\alpha_1}\sqrt{\frac{\sum_{i=2}^m \alpha_i^2}{m-1}} + \frac{1.58\beta_{\boldsymbol{f}}}{\sqrt{m-1}} + \frac{\gamma\theta}{\alpha_1\delta} \cdot \frac{\|\nabla S(x_t)\|_2}{\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2} + \frac{\frac{1}{\delta}\sqrt{\sum_{i=1}^m \alpha_i^2} \cdot \sqrt{\frac{2}{B}\ln(\frac{m}{p})}}{\sqrt{m-1}}\right)^2\right),$$

where γ *is as defined in Eq. (4).*

Proof Sketch. Each u_b is sampled independently, and on each axis the samples are averaged. Therefore, we apply Hoeffding's inequality on each axis, and use the union bound to bound the total ℓ_2 length of the error vector. We propagate this extra error term throughout the proof of Theorem 1. The detail proof is in Appendix A.7. □

Remark. This is the *first* concentration bound for the boundary gradient estimator to our best knowledge. From it we quantitatively learn how increasing the number of queries B increases the precision of the estimator, while the expectation bound cannot reflect it directly.

We note that the above two theorems are general—they provide finer-grained bounds for *all* existing boundary blackbox attacks, e.g., (Chen et al., 2020; Li et al., 2021; 2020a), and the proposed PSBA, thus these bounds provide a principled framework to analyze boundary blackbox attacks. Next, based on this framework we will (1) discover key characteristics that affect the query efficiency of gradient estimation; (2) explain why some existing attacks are more efficient than others; (3) show why PSBA improves upon these attacks.

Key Characteristics of Projective Gradient Estimation

Based on these two main theorems, we draw several observations for key characteristics that would help improve the gradient estimator. For simplification, we will leverage the big-$\mathcal{O}$ notation for the above expectation and concentration bounds (Theorems 1 and 2), as shown in Fig. 3. Compared with the expectation bound, the concentration bound adds a "sampling error term" that makes the bound hold with probability at least $1-p$. In Fig. 3, we label different terms in different colors to represent the key characteristics as optimization goals of a good gradient estimator:

(1) *Reduce the dimensionality m for the projection subspace*: To increase the cosine similarity, we can reduce the dimensionality $\boldsymbol{m}$.

(2) *Increase the projected length of true gradient on the projection subspace*: To increase the cosine similarity, we can increase the brown term $\|\text{proj}_{\nabla f(0)} \nabla \boldsymbol{S}(x_t)\|_2$, which is the projected length of true gradient on the projection subspace of $\boldsymbol{f}$.

(3) *Improve the sensitivity on the true gradient direction*: According to Lemma 4.1, α_1 is the sensitivity for the true gradient direction, and α_i for $i \neq 1$ is the sensitivity for orthogonal directions. To increase the cosine similarity, we can increase the blue term $\boldsymbol{\alpha_1}$, or reduce the green term $\frac{\sum_{i=2}^m \boldsymbol{\alpha_i^2}}{m-1}$ and $\boldsymbol{\alpha_{\max}^4}$. In Section 5.2, we empirically verify that PGAN achieves this, where α_1^2 is consistently and significantly larger than $\frac{\sum_{i=2}^m \alpha_i^2}{m-1}$, and therefore we leverage PGAN in our implementation. Note that for identical projection (Chen et al., 2020) or orthogonal projection (Li et al., 2020a), all α_i's are equal. The performance gain in NonLinearBA (Li et al., 2021) can be explained by this characteristic.

We illustrate these key characteristics in Appendix B. Other factors that can improve the cosine similarity are also revealed, such as smaller step size δ, and larger sampling (i.e., query) numbers B. However, they directly come at the cost of more queries as discussed in (Chen et al., 2020).

To improve the precision and query efficiency of the gradient estimation, next we consider how to optimize the estimator on the above characteristics, especially (1) and (2), given that (3) can be achieved by a PGAN-based projection.

Trade-Off on Dimensionality. From Fig. 3, we observe an apparent trade-off between Key Characteristics (1) and (2): when reducing the dimensionality m of the projection subspace (goal (1)), the preserved gradient information $\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2$ in this m-dimensional linear subspace $\nabla \boldsymbol{f}(0)$ becomes less, which opposes goal (2). To make a tradeoff between (1) and (2), based on above observation, there exists an optimal dimensionality m for the projection, and this dimensionality depends on how much true gradient information the projection subspace can preserve.

4.2. Optimize via Selecting Projection Subspaces

To circumvent the intrinsic trade-off, instead of hoping that the end-to-end trained $\boldsymbol{f}$ can capture much true gradient information in its linear subspace $\nabla \boldsymbol{f}(0)$, we can actively constrain the projection $\boldsymbol{f}$ on a representative subspace.

Here we focus on the linear subspace which can be represented by a linear combination of a set of basis, since the small step size δ of gradient estimator implies that only the local geometry matters and the local geometry of general subspace can be sufficiently approximated in first-order by linear subspace. Concretely, we select an m-dimensional linear subspace $V \subseteq \mathbb{R}^n$. Then, we train the projection $\boldsymbol{f}$ on V, i.e., $\text{im}(\boldsymbol{f}) \subseteq V$. Finally, we estimate the gradient with this $\boldsymbol{f}$ as: $u \in \mathbb{R}^m \overset{\boldsymbol{f}}{\mapsto} \Delta \boldsymbol{f}(\delta u) \in V \subseteq \mathbb{R}^n$, where $\dim(V) = m$. Interchangeably, we call m, the dimensionality of V, as *scale*, since it reflects the scale of the projection subspace V as we will show later.

Now we can analyze the cosine similarity of this new workflow. Since the image of projection $\boldsymbol{f}$ is in V, and $\text{rank}(\nabla \boldsymbol{f}(0)) = \dim(V) = m$, we have $\text{span}(\nabla \boldsymbol{f}(0)) = V$ and $\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2 / \|\text{proj}_V \nabla S(x_t)\|_2 = 1$. Plugging this into the above theorems and deriving from simple geometry (details in Lemma A.1), we find that the cosine similarity bound for subspace-constrained $\boldsymbol{f}$ is of the same form as in Fig. 3, with all $\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2^2$ replaced by $\|\text{proj}_V \nabla S(x_t)\|_2^2$.

Selected Subspace Improves Gradient Estimation. The formulation in Fig. 3 reveals that as long as we select an m-dimensional linear subspace V that preserves more gradient information $\|\text{proj}_V \nabla S(x_t)\|_2$ than the unconstrained projection model $\|\text{proj}_{\nabla \boldsymbol{f}(0)} \nabla S(x_t)\|_2$, the estimated gradient would have higher cosine similarity. We empirically find that the low-frequency subspace[3] satisfies such condition (Section 5.2): for real-world images, the gradient information of classifiers is highly concentrated on low-frequency domain. This is also cross-validated in the literature (Yin et al., 2019). We illustrate this analysis along with curves from numerical experiments in Fig. 4.

[3] Low-frequency subspace in DCT basis is a linear subspace.

$$\cos\langle\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\text{proj}_{\nabla f(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2} \cdot \left(1 - \mathcal{O}\left(m^2 \cdot \frac{\sum_{i=2}^{m} \alpha_i^2}{m-1} \left(\overbrace{\frac{\delta^2 \beta_f^2}{\alpha_1^4} + \frac{\alpha_{\max}^4}{\alpha_1^4} \cdot \frac{\delta^2 \beta_S^2}{\|\text{proj}_{\nabla f(0)} \nabla S(x_t)\|_2^2}}^{\text{expectation}} + \overbrace{\frac{\ln(m/p)}{B\alpha_1^2}}^{\text{sampling error}}\right)\right)\right)$$

Figure 3. Cosine similarity bound in big-$\mathcal{O}$ notation. The "expectation" reflects the expectation bound in Theorem 1. The "sampling error" is the additional term in Theorem 2 that makes the bound hold with probability at least $1-p$. When projection $\boldsymbol{f}$ is constrained in selected linear subspace V (see Section 4.2), the bound has the same form with all $\|\text{proj}_{\nabla f(0)} \nabla S(x_t)\|_2^2$ replaced by $\|\text{proj}_V \nabla S(x_t)\|_2^2$.

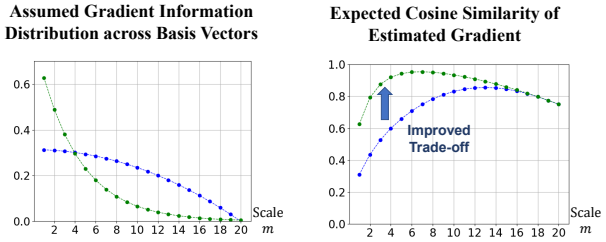

Figure 4. An illustration of why selected subspace improves the gradient estimation. **Left**: we assume a quadratic distribution of gradient information for $\boldsymbol{f}$'s basis (blue curve) and a more concentrated exponential distribution for frequency basis (green curve). **Right**: the corresponding expected cosine similarity is numerically computed with settings $n = 20$, $\beta_S = 0.5$, $\beta_{\boldsymbol{f}} = 0$, $\alpha_i = 1$, $\delta = m^{-1}$. The improved trade-off w.r.t. scale is revealed.

Recall that in PSBA, we train $\boldsymbol{f}$ on a smaller spatial scale and scale up its output, which is equivalent to constraining $\boldsymbol{f}$ on the low-frequency subspace. Since this subspace is more representative than the subspace $\nabla \boldsymbol{f}(0)$ of end-to-end trained $\boldsymbol{f}$, theoretically PSBA can estimate gradient with higher cosine similarity within fewer queries.

Existence of Optimal Scale. From Fig. 3, we find that the trade-off between Key Characteristics (1) and (2) in Section 4.1 still exists for selected subspace in general. Now it transforms to the existence of an optimal scale. This is revealed by the green curve in Fig. 4. For spatial and frequency scales, across different images from the same dataset and model, the coefficients of gradient information on the frequency basis vector tend to be very stable, as Section 5.2 shows. It implies that $\|\text{proj}_V \nabla S(x_t)\|_2$ in Fig. 3 across different images tend to be stable, and the optimal scales for gradient estimation tend to be stable too. Therefore, we can search for the optimal scale with a validation dataset.

5. Experimental Evaluation

With the established general framework of leveraging progressive scaling to improve attack efficiency, in this section, we take PGAN as an instantiation and conduct extensive experiments to 1) verify our theoretical analysis; 2) show that PSBA outperforms existing blackbox attacks by a significantly large margin. We also present some additional interesting findings.

5.1. Experimental Setup

Target Models. We use both offline models and a commercial online API as target models. For offline models, following (Li et al., 2020a; 2021), pretrained ResNet-18 on MNIST, CIFAR-10, CelebA and ImageNet are utilized. We also evaluate model ResNeXt50_32×4d (Xie et al., 2017) to demonstrate the generalization ability. On datasets MNIST and CIFAR-10, we scale up the input images to 224×224 by bilinear interpolation to help explore the influence of different scales. On CelebA, the target model is fine-tuned to perform the binary classification on the attribute 'Mouth_Slightly_Open'. The benign performance of target models is shown in Appendix C.2. For the commercial online API, the 'Compare' API (MEGVII, 2020a) from MEGVII Face++ which determines whether the faces from two images belong to the same person is used as the target model. The compared images are chosen from CelebA. The rationale of selecting these classification tasks and a detailed description of target models are discussed in Appendix C.

Training Procedure of PGAN. PGAN is trained to generate gradient images of reference models with small resolution until reaching convergence, and then new layers will be added to double the output scale. The PGAN training details could be found in Appendix D.2 and reference models' performance are shown in Appendix D.3. For simplicity, we will denote 'PGAN28' as the attack using the output of PGAN with scale 28×28.

Implementation. We follow the description in Section 3 to implement PSBA. Compared to other common attacks, we additionally train PGAN and use an additional validation set of ten images to search for the optimal scale. More implementation details are shown in Appendices D.4 and D.5.

Baselines. We consider six state-of-the-art decision-based attacks as the baselines. Among our baselines, **QEBA** (Li et al., 2020a) and **NLBA** (Li et al., 2021) utilize dimension reduction to sample from low-dimensional space, while the **Sign-OPT** (Cheng et al., 2020) and **HSJA** (Chen et al., 2020) apply direct Monte-Carlo sampling for gradient estimation. **EA** (Dong et al., 2019) adopts evolution algorithm to perform the attack. Note that we directly select the optimal scale for EA to compare under its optimal case. In Appendix E.4 we compare with **RayS** attack (Chen & Gu, 2020).

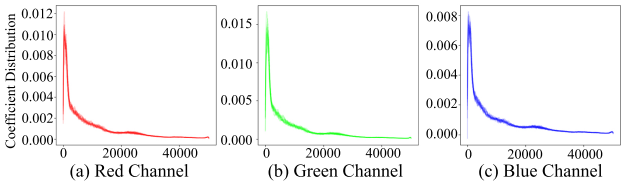

Figure 5. The long tail distribution of the coefficients of gradients generated from 10 images on the validation set of CIFAR-10 and represented on DCT basis for each channel.

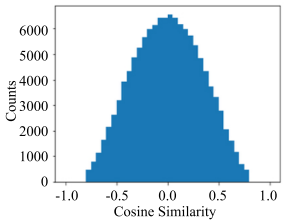

Figure 6. Pairwise cosine similarity of $\nabla \boldsymbol{f}(0)$ column vectors.

Evaluation Metrics. We adopt the standard evaluation metrics: 1) the Mean Squared Error (MSE) between the optimized adversarial examples and target image under different queries (this process will guarantee 100% attack success rate); 2) attack success rate at a specific MSE perturbation bound when the query number is constrained. Note that the conversion between MSE and ℓ_2 metrics are straightforward and order-preserving.

5.2. Verification of Theoretical Findings

Low Frequency Concentration. In order to verify the hypothesis of low frequency concentration, we randomly sample 10 gradient vectors calculated on ResNet-18 model for CIFAR-10 dataset. The gradients are normalized and transformed with DCT basis for each channel. As shown in Fig. 5, the x-axis shows the DCT basis from low to high frequency, and y-axis represents the corresponding coefficients denoised by the Savitzky-Golay filter. The three curves denote the three color channels respectively. The figures show stable long tail distribution across various images, and this implies the benefit of selecting the low-frequency subspace as the projection subspace in gradient estimation, as well as the existence of a stable optimal scale for the same dataset and target model, providing strong evidence for the discussion in Section 4.2. In addition, we draw a graph in a more statistical sense in Appendix E.6.

High Sensitivity on Gradient Direction. As shown in Appendix E.7, for trained PGAN, we compare the sensitivity (Lemma 4.1) for the projected true gradient direction, α_1^2, and the averaged sensitivity for other orthogonal directions, $\sum_{i=2}^{m} \alpha_i^2/(m-1)$. On all datasets, α_1^2 is significantly larger than $\sum_{i=2}^{m} \alpha_i^2/(m-1)$, which implies that trained PGAN achieves higher sensitivity on the true gradient direction. This is exactly the goal (3) in Section 4.1. In constrast, the identical (Chen et al., 2020) and othogonal projection (Li et al., 2020a) have identical directional sensitivity.

To what extent does orthogonality assumption hold. Here, we compute $\nabla \boldsymbol{f}(0)$ of PSBA on ImageNet at the optimal scale and then cluster the similar column vectors (those with cosine similarity > 0.8 or < -0.8) since they contribute to the sensitivity of one direction. Next, we compute the pairwise cosine similarity between clusters. The histogram of clusters based on their cosine similarity is shown in Fig. 6. As we can see, the histogram concentrates at 0, i.e., orthogonal pairs are most frequent. We also remark that recent orthogonal training can also enforce the assumption (Huang et al., 2020).

5.3. Attack Performance Evaluation

In this section, we show that the optimal scale indeed exists and our method PSBA-PGAN outperforms other state-of-the-art baselines in terms of attack effectiveness and efficiency. In addition, by selecting the optimal scale, PSBA-PGAN can also successfully and efficiently attack the online commercial face recognition API. Here, we randomly select 50 pairs of source and target images from validation set that are predicted by the target model as different classes for both offline attack and online attack and the selections of other hyperparameters are shown in Appendix E.1.

Offline Attack. The attack performance of different approaches in terms of the perturbation magnitude (MSE) between the adversarial and target image is shown in Fig. 7 (a)-(c). As we can see from Row 1, PSBA-PGAN effectively decrease the MSE when the number of queries is small and outperforms all baselines. Detailed comparisons on the gradient cosine similarity are in Appendix E.5. From Row 2 we can see that, interestingly, the optimal scale found by PSBA-PGAN across four datasets is consistently 28×28.

In Table 1, we show the attack success rate when the query number is constrained by 2K. This is because we can not easily generate a large number of queries (e.g., exceeding 2K) for attacking one image, and our PSBA is designed to be a query efficient attack with fast convergence speed. We can see that PSBA indeed significantly outperforms other methods when attacking Face++ API under small query budgets. On the other hand, when the query budgets get bigger, most of the methods would converge and attack successfully, then the comparison under this circumstance is not quite useful. We leave the results with large query number constraints in Appendix E.3. Besides, we also compare our method with RayS attack (Chen & Gu, 2020) in Appendix E.4. Detailed discussions on the computation time and resource consumption are in Appendix E.2, and the visualized results for other target models and ImageNet are in Appendix E.3.

Online Attack. To demonstrate the generalization and practicality of PSBA-PGAN, we perform it against a real-world online commercial API as shown in Fig. 7 (d). Although the PGAN used here is trained on ImageNet, PSBA-PGAN

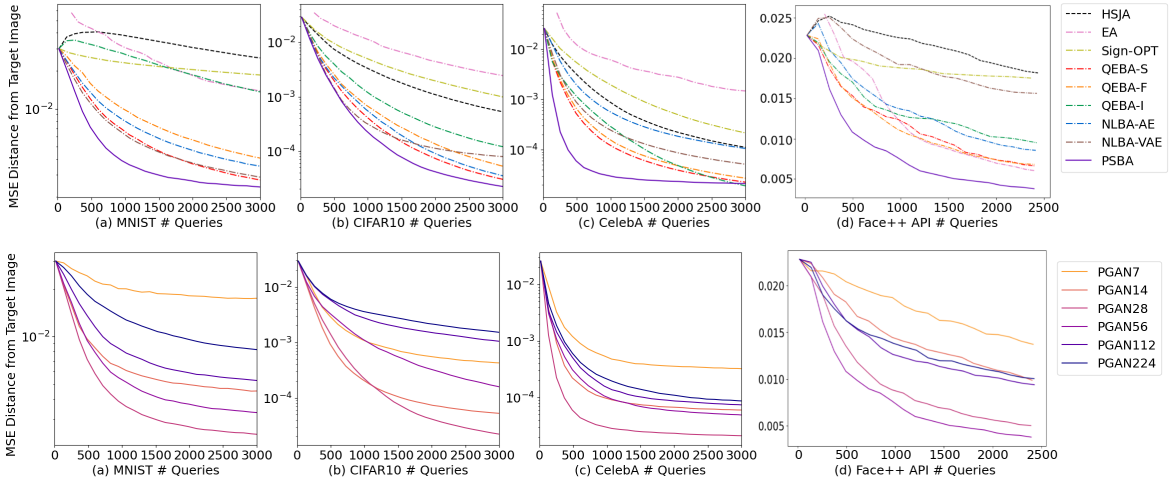

Figure 7. **Row 1**: Perturbation magnitude (MSE) w.r.t. query numbers for attacks on diverse datasets/models. **Row 2**: Perturbation magnitude when choosing different scales as the projection subspaces. The target model in (a)-(c) is ResNet-18, and an online commercial API in (d).

Table 1. Comparison of the attack success rate for different attacks at query number 2K (the perturbation magnitude under MSE for each dataset are: MNIST: $5e-3$; CIFAR10: $5e-4$; CelebA: $1e-4$; ImageNet: $1e-2$).

Data	Model	# Queries = 2K								
		HSJA	EA	Sign-OPT	QEBA-S	QEBA-F	QEBA-I	NLBA-AE	NLBA-VAE	PSBA
MNIST	ResNet	2%	6%	2%	60%	42%	4%	46%	58%	**78%**
	ResNeXt	4%	4%	6%	76%	66%	16%	70%	80%	**88%**
CIFAR10	ResNet	26%	10%	10%	82%	70%	58%	76%	82%	**94%**
	ResNeXt	32%	0%	18%	88%	72%	64%	**90%**	**90%**	**90%**
CelebA	ResNet	20%	8%	2%	80%	70%	72%	20%	46%	**90%**
	ResNeXt	24%	6%	6%	60%	56%	72%	20%	38%	**88%**
ImageNet	ResNet	24%	24%	6%	**54%**	52%	46%	44%	28%	**54%**
	ResNeXt	20%	22%	16%	40%	38%	36%	28%	26%	**42%**

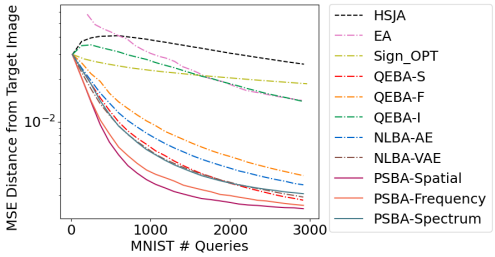

Figure 8. Perturbation magnitude (MSE) under different number of queries on MNIST for PSBA sampling from spatial, frequency, and spectrum domains.

still outperforms other baselines and interestingly, it adopts the optimal scale as 56×56.

Frequency and Spectrum Domains. In addition to the spatial domain, we also evaluate PSBA on frequency and spectrum domains. The results are shown in Fig. 8. PSBA outperforms other baselines. Detailed implementations and other ablation studies are included in Appendix E.8.

Additional Findings. (1) In Appendix E.7, we deliberately adjust the sensitivity on different directions $\{\alpha_i\}_{i=1}^m$ for given projection $\boldsymbol{f}$ to study the correlation between sensitivity and empirical attack performance. The results conform to our theoretical findings (goal (3) in Section 4.1). (2) We empirically study the optimal scale across model structures and show that different models have their own preference, which we believe will lead to interesting future directions. More details can be found in Appendix E.9.

6. Conclusion

In this paper, we propose PSBA, a progressive-scale blackbox attack via projective gradient estimation. We propose a general theoretical framework to analyze existing projective gradient estimators, show key characteristics for improvement, and justify why PSBA outperforms other blackbox attacks. Extensive experiments verify our theoretical findings and show that PSBA outperforms existing blackbox attacks significantly against various target models including a real-world face recognition API.

Acknowledgements

We thank the anonymous reviewers for valuable feedback. This work is partially supported by NSF grant No.1910100, NSF CNS 20-46726 CAR, and Amazon Research Award.

References

Bhagoji, A. N., He, W., Li, B., and Song, D. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *European Conference on Computer Vision*, pp. 158–174. Springer, 2018.

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

Cai, Z., Fan, Q., Feris, R. S., and Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pp. 354–370. Springer, 2016.

Chen, J. and Gu, Q. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1739–1747, 2020.

Chen, J., Jordan, M. I., and Wainwright, M. J. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE symposium on security and privacy (SP)*, pp. 1277–1294. IEEE, 2020.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Cheng, M., Singh, S., Chen, P. H., Chen, P.-Y., Liu, S., and Hsieh, C.-J. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019a.

Cheng, M., Singh, S., Chen, P., Chen, P.-Y., Liu, S., and Hsieh, C.-J. Sign-opt: A query-efficient hard-label adversarial attack, 2020.

Cheng, S., Dong, Y., Pang, T., Su, H., and Zhu, J. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, pp. 10932–10942, 2019b.

Clarke, F. H., Ledyaev, Y. S., Stern, R. J., and Wolenski, P. R. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., and Zhu, J. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722, 2019.

Fiers, T. Why are randomly drawn vectors nearly perpendicular in high dimensions. Mathematics Stack Exchange, 2018. URL https://math.stackexchange.com/q/995678. URL:https://math.stackexchange.com/q/995678 (version: 2018-05-15).

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guo, Y., Yan, Z., and Zhang, C. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015a.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks, 2018.

Huang, L., Liu, L., Zhu, F., Wan, D., Yuan, Z., Li, B., and Shao, L. Controllable orthogonalization in training dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6429–6438, 2020.

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

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation, 2018.

Krizhevsky, A., Hinton, G., et al. *Learning multiple layers of features from tiny images*. Citeseer, 2009.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, H., Xu, X., Zhang, X., Yang, S., and Li, B. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Li, H., Li, L., Xu, X., Zhang, X., Yang, S., and Li, B. Nonlinear gradient estimation for query efficient blackbox attack. In *Proceedings of 24th International Conference on Artificial Intelligence and Statistics*, 2021.

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

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ma, W., Wu, Y., Cen, F., and Wang, G. Mdfn: Multi-scale deep feature learning network for object detection. *Pattern Recognition*, 100:107149, 2020.

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

MEGVII. Face++ facial recognition 'compare' API documentation. https://console.faceplusplus.com/documents/5679308, 2020a.

MEGVII. Face++ facial recognition 'compare' API query URL. https://api-us.faceplusplus.com/facepp/v3/compare, 2020b.

MEGVII. Face++. https://www.faceplusplus.com/, 2020c.

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

PyTorch. Torchvision.models. https://pytorch.org/docs/stable/torchvision/models.html, 2020.

Research, F. Pytorch GAN zoo. https://github.com/facebookresearch/pytorch_GAN_zoo, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Suya, F., Chi, J., Evans, D., and Tian, Y. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1327–1344, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014.

Tashiro, Y., Song, Y., and Ermon, S. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in Neural Information Processing Systems*, 33, 2020.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Wu, R., Zhang, G., Lu, S., and Chen, T. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5021–5030, 2020.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks, 2017.

Yang, G., Duan, T., Hu, E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, 2020.

Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32:13276–13286, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks, 2017.

Zhang, D. and Khoreva, A. Progressive augmentation of gans. In *Advances in Neural Information Processing Systems*, pp. 6249–6259, 2019.

Zhang, J., Xie, Z., Sun, J., Zou, X., and Wang, J. A cascaded r-cnn with multiscale attention and imbalanced samples for traffic sign detection. *IEEE Access*, 8:29742–29754, 2020.