Nonlinear Projection Based Gradient Estimation for Query Efficient Blackbox Attacks

Huichen Li[1*] **Linyi Li**[1*] **Xiaojun Xu**[1] **Xiaolu Zhang**[2] **Shuang Yang**[3] **Bo Li**[1]
[1] University of Illinois at Urbana-Champaign [2] Ant Financial [3] Alibaba Group US
{huichen3, linyi2, xiaojun3, lbo}@illinois.edu, yueyin.zxl@antfin.com, shuang.yang@alibaba-inc.com
* *The first two authors contributed equally.*

Abstract

Gradient estimation and *vector space projection* have been studied as two distinct topics. We aim to bridge the gap between the two by investigating how to efficiently estimate gradient based on a projected low-dimensional space. We first provide lower and upper bounds for gradient estimation under both linear and nonlinear gradient projections, and outline checkable sufficient conditions under which one is better than the other. Moreover, we analyze the query complexity for the projection-based gradient estimation and present a sufficient condition for query-efficient estimators. Built upon our theoretic analysis, we propose a novel query-efficient Nonlinear Gradient Projection-based Boundary Blackbox Attack (**NonLinear-BA**). We conduct extensive experiments on four datasets: ImageNet, CelebA, CIFAR-10, and MNIST, and show the superiority of the proposed methods compared with the state-of-the-art baselines. In particular, we show that the projection-based boundary blackbox attacks are able to achieve much smaller magnitude of perturbations with 100% attack success rate based on efficient queries. Both linear and nonlinear projections demonstrate their advantages under different conditions. We also evaluate NonLinear-BA against the commercial online API MEGVII Face++, and demonstrate the high blackbox attack performance both quantitatively and qualitatively. The code is publicly available at `https://github.com/AI-secure/NonLinear-BA`.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Gradient estimation and vector space projection have both been extensively studied in machine learning, but largely for different purposes. Gradient estimation is used when gradient-based optimization such as back-propagation is employed but the exact gradients are not directly accessible, for example, in the case of blackbox adversarial attacks (Chen et al., 2020; Li et al., 2020). Vector space projection, especially gradient projection (or sparsification), on the other hand, has been used to speedup training, for instance, by reducing the complexity of communication and/or storage when performing model update in distributed training (Wangni et al., 2018). In this paper, we aim to bridge the gap between the two and attempt to answer the following questions: *Can we estimate gradients from a projected low-dimensional subspace? How do different projections affect the gradient estimation quality?*

Our investigation is motivated in particular by the challenging problem of blackbox adversarial attacks (Bhagoji et al., 2017; Ilyas et al., 2018). Adversarial attacks have the ability to mislead machine learning models with potentially catastrophic consequences while staying imperceptible to humen. While extensive progresses have been made in white-box attacks (Carlini and Wagner, 2017; Eykholt et al., 2018; Xu et al., 2018) where attackers have complete knowledge about the target model, the more realistic scenario of blackbox attacks where the attacker only has query access to the target model remains challenging. One major challenge is the excessive query complexity. For example, boundary-based blackbox attacks (BA) (Brendel et al., 2017) have shown promising attack effectiveness, but the required query number is too large to be practically feasible (e.g., many approaches require 10^5 or more queries per attack, which could take hours or even days given the rate limit of public machine learning APIs). This inefficiency stems partially from the high dimensionality of the gradient since the Monte Carlo gradient estimation relies on sampling perturbations

from the gradient space.

In this work, we study the properties of a general vector space projection $\mathbf{f}$, that transforms vectors from low-dimensional subspace $\mathbb{R}^n$ to the original gradient space $\mathbb{R}^m$ for gradient estimation. We theoretically provide the lower and upper bounds of cosine similarity between the estimated and true gradients, based on sampling distribution analysis and Taylor expansion. These bounds imply that it is possible to estimate the gradient effectively under checkable sufficient condition. Intuitively, the condition measures how well the estimated and true gradients of the target model align with each other. Furthermore, we compare linear and nonlinear gradient projections in terms of the cosine similarity between the estimated and true gradients, and prove the existence of nonlinear projection that is able to achieve strictly higher cosine similarity lower bound. We finally analyze the query complexity of gradient estimation and present a sufficient condition for query-efficient projection-based gradient estimation. The analysis provides theoretic answers to the aforementioned questions. *Our theoretic analysis on the projection-based gradient estimation is not specific to adversarial attacks, but can shed light on a broader range of applications such as gradient sparsification and distributed training.*

Based on our analysis for query-efficient gradient estimation, we propose **NonLinear-BA**, which applies deep generative models such as AEs, VAEs, and GANs as the nonlinear projections to perform blackbox attack, and therefore evaluate the power of projection-based gradient estimation empirically. Once trained, these generative models are used to project the sampled low-dimensional vectors back to high-dimensional gradient space and query the target model to estimate the gradient. We experimentally evaluate NonLinear-BA with three proposed nonlinear projections on four image datasets: ImageNet (Deng et al., 2009), CelebA (Liu et al., 2015), CIFAR-10 (Krizhevsky et al., 2009) and MNIST (LeCun et al., 1998). We show that NonLinear-BA can achieve 100% attack success rate more efficiently with smaller magnitude of perturbation compared with baselines. We also evaluate the NonLinear-BA against a commercial online API MEGVII Face++ (MEGVII, 2021c). Both quantitative and qualitative results are shown to demonstrate its attack effectiveness.

Contributions: (1) We provide the first general theoretical analysis framework for the projection-based gradient estimation, analyzing the cosine similarity between estimated and true gradients under different vector space projections. (2) We prove and compare the lower bounds of gradient cosine similarities for linear and nonlinear projections. We also analyze the query complexity of the projection based gradient estimators. (3) We propose a novel nonlinear gradient projection-based blackbox attack (NonLinear-BA) which exploits the power of nonlinear-projection based gradient estimation. (4) We conduct extensive experiments on both offline ML models and commercial online APIs with high-dimensional image datasets to demonstrate the high attack performance of NonLinear-BA. The empirical results verify our theoretical findings that the projection-based gradient estimation via sampling is query efficient, and some projections outperform others under certain conditions.

Related Work: The vulnerability of ML to adversarial attacks has been demonstrated by recent studies (Szegedy et al., 2014; Goodfellow et al., 2014). To better understand such threat, new attacks have been consistently proposed over years, which lie in two major branches: *whitebox attacks* and *blackbox attacks*. The whitebox attacks, e.g., (Goodfellow et al., 2014; Carlini and Wagner, 2017; Kurakin et al., 2016; Madry et al., 2018; Athalye et al., 2018), assume full knowledge of the victim model for an attacker; while blackbox attacks only require limited access to the victim model, which is more applicable in practice.

The *blackbox attacks* can be divided into two categories: transfer-based and query-based attacks. The transfer-based attacks rely on adversarial transferability (Papernot et al., 2016; Tramèr et al., 2017), where the adversarial examples generated against one ML model can also attack another model. Different approaches including ensemble methods have been explored to enhance the adversarial transferability (Liu et al., 2017). The query-based attacks utilize the zeroth-order information, i.e., the prediction confidence score, to estimate the gradient of the blackbox model via queries. In query-based attacks, a series of work (Chen et al., 2017; Bhagoji et al., 2018; Ilyas et al., 2018; Tu et al., 2019; Cheng et al., 2019b) has been proposed to improve the efficiency of gradient estimation and reduce the number of queries.

Boundary-based blackbox attack (BA) (Brendel et al., 2017) is another type of query-based attack where the blackbox model only provides the final prediction instead of the prediction confidence scores for each query. Some work has been conducted to improve the query efficiency for BA. For instance, Cheng et al. (2019a) perform gradient sign estimation to improve attack efficiency, Chen et al. (2020) apply the Monte-Carlo sampling strategy to perform gradient estimation for BA, and Li et al. (2020) improve it by sampling from representative low-dimensional orthonormal subspace. Our work, on the other hand, aims to explore more general projection based gradient estimators with a unified theoretical analysis framework..

2 Problem Definition

In this section, we will first introduce the framework of *boundary-based blackbox attack*, and then focus on tackling the challenge of query-based gradient estimation.

Boundary-Based Blackbox Attack (BA). Given an instance x drawn from certain distribution $\mathcal{D}$: $x \sim \mathcal{D}$, where $x \in \mathbb{R}^m$, a C-way classification model $G : \mathbb{R}^m \mapsto \mathbb{R}^C$ is trained to output the confidence score for each class. The final prediction of the model is obtained by selecting the class with the highest confidence score $y = \text{argmax}_{i \in [C]} G(x)_i$ ($[C] = \{1, \ldots, C\}$). The model G is referred to as 'target model' throughout our discussion as it is the target of the adversarial attack. In this work we focus on the scenario where the adversaries do not have access to the details of model G (i.e. blackbox attack) and can only query the model to obtain the final prediction label y instead of the confidence scores.

The general framework of a BA is as follows: given a target-image $x_{tgt} \in \mathbb{R}^m$ whose true label is $y_{ben} \in [C]$, the attacker's goal is to craft an adversarial image x_{adv} that is predicted as a maliciously chosen label $y_{mal} \in [C]$, while the distance $D(x_{tgt}, x_{adv})$ between the two images is as small as possible. Here D is a L_p-norm based distance function which aims to restrict the perturbation added to the target-image in order to make it less noticeable. In this paper we only consider targeted attack with an intentionally chosen y_{mal} since untargeted attack is a trivial extension of the targeted case (by randomly sampling a y_{mal}).

Definition 1 ((G, y_{mal}))-Difference Function)**.** Given a model G, and malicious target y_{mal}, the *difference function* $S : \mathbb{R}^m \to \mathbb{R}$ is defined as $S(x) = G(x)_{y_{mal}} - G(x)_{y_{ben}}$, where y_{ben} denotes the ground truth label.

The difference function S is an important indicator of whether the image is successfully perturbed from being predicted as y_{ben} to y_{mal}. A *boundary-image* is an image x that lies on the decision boundary between y_{ben} and y_{mal} , i.e., $S(x) = 0$.

Projection-Based Gradient Estimation. There are three main steps to perform the BA: (1) gradient estimation at G's decision boundary, (2) move the boundary-image along the estimated gradient direction, and (3) project the image back to the decision boundary. Typically, the first step requires to estimate the gradient based on the sign of *difference function* defined in Definition 1 given multiple queries. It is very computationally expensive as the high-dimensional gradient estimation requires a large number of queries (Chen et al., 2020). Based on recent advances in efficient communication and gradient sparsification (Wangni et al., 2018), we hypothesize that there exist lower dimensional *supports* for gradient vectors and we aim to project the gradient to these lower dimensional supports and perform the estimation efficiently. In particular, we theoretically analyze the impacts of linear and nonlinear gradient projections on gradient estimation.

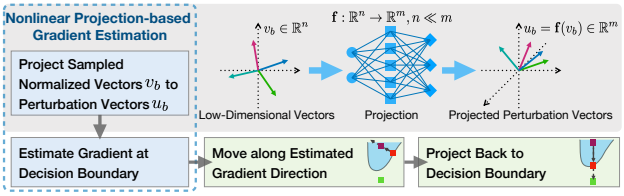

Figure 1: Algorithm illustration for NonLinear-BA.

3 NonLinear-BA: Nonlinear Gradient Projection-based Boundary Blackbox Attack

In this section we introduce the proposed nonlinear gradient projection-based boundary blackbox attack (NonLinear-BA) as illustrated in Figure 1, followed by the detailed theoretical analysis and guarantees in Section 4.

In standard BA, the way to estimate the gradient given the query results is done by Monte Carlo sampling method (Chen et al., 2020):

$$\widetilde{\nabla S}(x_{adv}^{(t)}) = \frac{1}{B} \sum_{b=1}^{B} \text{sgn}\left(S\left(x_{adv}^{(t)} + \delta u_b\right)\right) u_b, \qquad (1)$$

where $x_{adv}^{(t)}$ is the boundary-image at iteration t obtained by binary search with precision threshold θ following Equation 4 (to be shown later). The u_b's are B perturbation vectors uniformly sampled from the unit sphere in $\mathbb{R}^m$. The size of random perturbation δ is chosen as a function of image size and the binary search threshold (Chen et al., 2020) to control the gradient estimation error caused by the boundary-image's offset from the exact decision boundary due to binary search precision. The function $\text{sgn}(S(\cdot))$ denotes the sign of the difference function (Definition 1). Its value is acquired by querying the victim model and comparing the output label with y_{mal}. It is clear that the query cost is very high when the input dimension m is large. A typical 3-channel 224×224 image gradient vector has a dimension m of over 150k. It is challenging to perform accurate estimation in such a high-dimensional space with limited queries. To reduce the query complexity, Li et al. (2020) propose to search for a *representative subspace* with orthonormal mappings $\mathbf{W} = [w_1, \ldots, w_n] \in \mathbb{R}^{m \times n}, n \ll m$ and $\mathbf{W}^\intercal \mathbf{W} = I$. The perturbation vectors are generated by first sampling n-dimensional unit vectors v_b and project them with $u_b = \mathbf{W} v_b$.

Nonlinear Projection-Based Gradient Estimation. To search for the gradient representative subspaces more efficiently, we propose to perform the nonlinear projection-based gradient estimation. In particular, we propose to leverage generative models given their expressive power. Here we mainly consider AE, VAE and GAN as examples. There are two phases in NonLinear-BA: training and attacking. The detailed model structure and the training phase are described in Section F.1. Note that all these models typically have two components: an 'encoder' and a 'decoder' for AE and VAE, and a 'generator' and a 'discriminator' for the GAN. The 'decoder' or 'generator' projects a latent representation or random vector to sample space. The latent dimension is usually much lower than the sample space and this property is exactly desired. We unify the notations and denote both the 'decoder' of AE and VAE and the 'generator' part of GAN as 'projection-based gradient estimator' in our following discussion. The gradient estimator is then used as the projection $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ in the attacking phase. We first randomly sample unit latent vectors v_b's in $\mathbb{R}^n$, then the perturbation vectors generated as $u_b = \mathbf{f}(v_b) \in \mathbb{R}^m$ are used in the gradient estimation, yielding our gradient estimator as

$$\widetilde{\nabla S}(x_{adv}^{(t)}) = \frac{1}{B}\sum_{b=1}^{B} \operatorname{sgn}\left(S\left(x_{adv}^{(t)} + \delta \mathbf{f}(v_b)\right)\right)\mathbf{f}(v_b). \quad (2)$$

Move along Estimated Gradient Direction. After getting the estimated gradient $\widetilde{\nabla S}$, the boundary-image $x_{adv}^{(t)}$ is moved along that direction by:

$$\hat{x}_{t+1} = x_{adv}^{(t)} + \xi_t \cdot \frac{\widetilde{\nabla S}}{\|\widetilde{\nabla S}\|_2}, \quad (3)$$

where ξ_t is a step size chosen by searching with queries similar with HSJA (Chen et al., 2020).

Project Back to Decision Boundary. In order to move closer to the target-image and enable the gradient estimation in the next iteration, we map the new adversarial image $x_{adv}^{(t)}$ back to the decision boundary. This is achieved via binary search assisted by queries to find a suitable weight α_t:

$$x_{adv}^{(t+1)} = \alpha_t \cdot x_{tgt} + (1 - \alpha_t) \cdot \hat{x}_{t+1}. \quad (4)$$

4 Projection-Based Gradient Estimation Analysis

To study the effectiveness of our projection-based gradient estimator in Equation (2) in terms of improving the estimation accuracy and reducing the number of queries, in this section, we theoretically analyze the expected cosine similarity between the estimated gradient $\widetilde{\nabla S}(x_{adv}^{(t)})$ and the true gradient $\nabla S(x_{adv}^{(t)})$ for the boundary-image $x_{adv}^{(t)}$ at step t.

4.1 Generalized Gradient Estimator

We first formally define the gradient projection function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, which maps from the low-dimensional representative space $\mathbb{R}^n$ to the original high-dimensional space $\mathbb{R}^m$, where $n \le m$. Note that the projection function here could be *nonlinear*, which is different from projections (linear transformations) defined in standard linear algebra[1].

Definition 2 (Generalized Projection-Based Gradient Estimator)**.** Suppose $\mathbf{f}(x_0)$ is a boundary-image, i.e., $S(\mathbf{f}(x_0)) = 0$, let $u_1, u_2, \ldots, u_B$ be a subset of orthonormal basis of space $\mathbb{R}^n$ sampled uniformly ($B \le n$), we define

$$\widetilde{\nabla \mathbf{f}^\intercal \nabla} S := \frac{1}{B}\sum_{i=1}^{B} \operatorname{sgn}\left(S\left(\mathbf{f}(x_0 + \delta u_i)\right)\right) u_i. \quad (5)$$

Then, the *generalized gradient estimator* for $\nabla S(\mathbf{f}(x_0))$ is defined as

$$\widetilde{\nabla S}(\mathbf{f}(x_0)) := \nabla \mathbf{f}(x_0)\widetilde{\nabla \mathbf{f}^\intercal \nabla} S. \quad (6)$$

We abbreviate $\widetilde{\nabla S}(\mathbf{f}(x_0))$ as $\widetilde{\nabla S}$ when there is no ambiguity.

All the aforementioned gradient estimators are concretization of this generalized gradient estimator with different projections $\mathbf{f}$'s, including HSJA (Chen et al., 2020), QEBA (Li et al., 2020) and our proposed NonLinear-BA. We defer the instantiations to Appendix A.

We now impose local Lipschitz and local smoothness conditions on the projection $\mathbf{f}$ and the difference function S.

Definition 3 (Local L-Lipschitz)**.** A (scalar or vector) function f is called local L-Lipschitz around x_0 with radius r, if for any two inputs $x, x' \in \{x_0 + \delta : \|\delta\|_2 \le r\}$,

$$\frac{\|f(x) - f(x')\|_2}{\|x - x'\|_2} \le L.$$

Definition 4 (Local β-Smoothness)**.** A (scalar or vector) function f is called local β-smooth around x_0 with radius r, if (1) f is differentiable everywhere in region $\{x_0 + \delta : \|\delta\|_2 \le r\}$; and (2) for any two inputs $x, x' \in \{x_0 + \delta : \|\delta\|_2 \le r\}$,

$$\frac{\lambda_{\max}(\nabla f(x) - \nabla f(x'))}{\|x - x'\|_2} \le \beta,$$

[1] https://en.wikipedia.org/wiki/Projection_(linear_algebra)

where $\lambda_{\max}(\mathbf{M})$ denotes the maximum eigenvalue of the matrix $\mathbf{M}$. Specifically, if M is a vector, $\lambda_{\max}(M) = \|M\|_2$.

The Lipschitz and smoothness definitions follow the general definitions in the literature (Boyd et al., 2004; Bubeck, 2015; Hardt et al., 2016). Specifically, for a general function f (e.g. S or $\mathbf{f}$), when $\beta = 0$, the gradient ∇f is a constant in the region thus f is locally linear. If f is a neural network, there exists a local Lipschitz constant L (Zhang et al., 2019), and under generalized differential operator, there also exists a local smoothness constant β (Nesterov, 2013; Clarke et al., 2008).

Assumptions. Throughout the section, we assume the projection $\mathbf{f}$ is $L_{\mathbf{f}}$-Lipschitz and $\beta_{\mathbf{f}}$-smooth around x_0 with radius δ, and the difference function S is L_S-Lipschitz and β_S-smooth around $\mathbf{f}(x_0)$ with radius $L_{\mathbf{f}}\delta$.

For the convenience of our analysis, we define the constant ω as such:

Definition 5 (Gradient Cosine Similarity Indicator ω)**.**

$$\omega := \delta\left(\frac{1}{2}\beta_{\mathbf{f}}L_S + \frac{1}{2}\beta_S L_{\mathbf{f}}^2 + \frac{1}{2}\delta\beta_{\mathbf{f}}\beta_S L_{\mathbf{f}} + \frac{1}{8}\delta^2\beta_{\mathbf{f}}^2\beta_S\right). \quad (7)$$

The gradient cosine similarity indicator ω is an important quantity appearing in the cosine similarity lower bound. The δ in definition denotes the step size used in gradient estimation which is chosen according to HSJA (Chen et al., 2020).

Theorem 1 (General Bound for Gradient Estimator)**.** *Let $\mathbf{f}(x_0)$ be a boundary-image, i.e., $S(\mathbf{f}(x_0)) = 0$. The projection $\mathbf{f}$ and the difference function S satisfy the assumptions in Section 4.1. Over the randomness of the sampling of orthogonal basis subset $u_1, u_2, \ldots, u_B$ in $\mathbb{R}^n$ space, the expectation of cosine similarity between $\widetilde{\nabla S}(\mathbf{f}(x_0))$ ($\widetilde{\nabla S}$ for short) and $\nabla S(\mathbf{f}(x_0))$ (∇S for short) satisfies*

$$\left(2\left(1 - \frac{\omega^2}{\|\nabla\mathbf{f}^\intercal\nabla S\|_2^2}\right)^{(n-1)/2} - 1\right)\frac{\|\nabla\mathbf{f}^\intercal\nabla S\|_2}{L_{\mathbf{f}}\|\nabla S\|_2}\sqrt{\frac{B}{n}}c_n \leq \mathbb{E}\cos\langle\widetilde{\nabla S}, \nabla S\rangle \leq \frac{\|\nabla\mathbf{f}^\intercal\nabla S\|_2}{l_{\mathbf{f}}\|\nabla S\|_2}\sqrt{\frac{B}{n}}c_n, \quad (8)$$

where ω is defined in Definition 5, and we assume $\omega \leq \|\nabla\mathbf{f}^\intercal\nabla S\|_2$; $c_n \in (2/\pi, 1)$ is a constant depended on n; $l_{\mathbf{f}} := \lambda_{\min}(\nabla\mathbf{f}(x_0))$.

Proof sketch. Based on Taylor expansion, the projected length $\langle u_i, \nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))\rangle$ is correlated with $\operatorname{sgn}(S(\mathbf{f}(x_0 + \delta u_i)))$, where u_i is a sampled base vector. Concretely, when the projection smoothness is bounded, we show that when $|\langle u_i, \nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))\rangle|$ is larger than some threshold, it always has the same sign as $\operatorname{sgn}(S(\mathbf{f}(x_0 + \delta u_i)))$. On the other hand, we study the distribution of $\langle u_i, \frac{\nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))}{\|\nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))\|_2}\rangle$, and derive the closed-form PDF for the distribution. The cosine similarity between $\widetilde{\nabla\mathbf{f}^\intercal\nabla S}$ and $\nabla\mathbf{f}^\intercal\nabla S$ can be expressed as the sum of the products of these two terms over the basis: $\operatorname{sgn}(S(\mathbf{f}(x_0 + \delta u_i))) \cdot \langle u_i, \frac{\nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))}{\|\nabla\mathbf{f}(x_0)^\intercal\nabla S(\mathbf{f}(x_0))\|_2}\rangle$. Thus, we can derive the bounds for cosine similarity between $\widetilde{\nabla\mathbf{f}^\intercal\nabla S}$ and $\nabla\mathbf{f}^\intercal\nabla S$. From these bounds, we obtain the bounds for cosine similarity between $\widetilde{\nabla S}$ and ∇S. We defer the detailed proof to Appendix B. □

Remark. This theorem provides the lower and upper bounds of the cosine similarity between our generalized gradient estimator $\widetilde{\nabla S}$ and the true gradient ∇S for different models. As long as the Lipschitz and smoothness conditions in Section 4.1 are satisfied, this bound is valid regardless of the concrete form of the projection $\mathbf{f}$ or how well the projection $\mathbf{f}$ is aligned with S, so we call it a 'general bound'. We remark that smaller ω implies larger lower bound for the cosine similarity, which induces a tighter and improved gradient estimation. Detailed discussions of these bounds are presented in Section 4.2.

4.2 Gradient Estimation Based on Different Gradient Projections

From Theorem 1, one may think that linear projection is better than nonlinear one since when the $\nabla\mathbf{f}$ is the same, linear projection implies $\beta_{\mathbf{f}} = 0$, which leads to smaller ω and higher lower bound of the gradient cosine similarity. However, this lower bound is applied to all models satisfying the Lipschitz and smoothness condition. In fact, there exists nonlinear projection $\mathbf{f}$ leading to higher cosine similarity lower bound. (We will focus on the discussion of lower bound below, since the upper bound is irrelevant with $\beta_{\mathbf{f}}$ from Theorem 1, meaning linear and nonlinear projections would share the same upper bound.)

Linear Projection. First, let us consider the linear projection $\mathbf{f}$. Throughout the text, we use $\lambda_{\max}(\mathbf{M})$ to denote the largest eigenvalue of matrix $\mathbf{M}$, and $\lambda_{\min}(\mathbf{M})$ the smallest eigenvalue of matrix $\mathbf{M}$.

Corollary 1 (Linear projection Bound, informal)**.** *Under the same setting of Theorem 1 with additional condition that projection $\mathbf{f}$ is locally linear around x_0 with radius δ and $L_{\mathbf{f}} := \lambda_{\max}(\nabla\mathbf{f}(x_0))$, the expectation of cosine similarity satisfies Equation* (8) *with*

$$\omega := \frac{1}{2}\delta\beta_S L_{\mathbf{f}}^2. \quad (9)$$

We assume that $\omega \leq \|\nabla \mathbf{f}^\intercal \nabla S\|_2$. $c_n \in (2/\pi, 1)$ *is a constant depended on* n.

Remark. We defer the formal statement to Appendix D.1. This is a direct application of Theorem 1 with $\beta_{\mathbf{f}} = 0$ due to linearity. The main difference between the corollary and Theorem 1 is in ω, where the general ω in Equation (7) is altered by Equation (9). Furthermore, if S is also locally linear, then $\beta_S = 0$ and hence $\omega = 0$, which closes the gap between lower bound and upper bound and implies that the gradient estimation is pretty precise (cosine similarity between estimated and true gradient is $c_n \in (2/\pi, 1)$). In addition, Li et al. (2020) provide a cosine similarity bound based on projection taken the form of orthogonal transformation, which can be recovered from Corollary 1 by setting $L_{\mathbf{f}} = l_{\mathbf{f}} = 1$, $\beta_{\mathbf{f}} = 0$, and replacing $\|\nabla S^\intercal \nabla \mathbf{f}\|_2$ with $\|\nabla S\|_2$ in Equation (8) due to the randomness of projection $\mathbf{f}$.

Nonlinear Projection. For nonlinear projection, we have the following theorem.

Theorem 2 (Existence of Better Nonlinear Projection, informal)**.** *Under the same setting of Corollary 1, there exists a* nonlinear *projection* $\mathbf{f}'$ *satisfying the assumptions in Section 4.1, with* $\mathbf{f}'(x_0) = \mathbf{f}(x_0)$ *and* $\nabla \mathbf{f}'(x_0) = \nabla \mathbf{f}(x_0)$*, such that the expectation of cosine similarity between* $\widetilde{\nabla S}(\mathbf{f}'(x_0))$ *(*$\widetilde{\nabla S}$ *for short) and* $\nabla S(\mathbf{f}'(x_0))$ *(*∇S *for short) satisfies Equation* (8) *with*

$$\omega := \frac{1}{2}\delta\beta_S L_{\mathbf{f}}^2 - \frac{1}{5}\beta_{\mathbf{f}}\beta_S \delta^2 L_{\mathbf{f}} < \frac{1}{2}\delta\beta_S L_{\mathbf{f}}^2. \tag{10}$$

We assume that $\omega \leq \|\nabla \mathbf{f}^\intercal \nabla S\|_2$. $c_n \in (2/\pi, 1)$ *is a constant depended on* n.

Proof Sketch. We prove by construction—we construct the nonlinear projection $\mathbf{f}'$ explicitly from $\mathbf{f}(x_0)$, $\nabla \mathbf{f}(x_0)$ and the difference function S. Comparing with the linear projection $\mathbf{f}$, the $\mathbf{f}'$ is allowed to have curvature since $\beta_{\mathbf{f}} > 0$. The constructed $\mathbf{f}'$ exploits this curvature to cancel out the impreciseness caused by large $\nabla \mathbf{f}$ and thus reduces the actual $L_{\mathbf{f}}$. After showing that $\mathbf{f}'$ satisfies the assumptions in Section 4.1, we derive its cosine similarity bound with corresponding ω. □

Remark. This theorem shows that if the difference function S is nonlinear (i.e., $\beta_S > 0$), for any linear projection $\mathbf{f}$, we can define a particular nonlinear projection $\mathbf{f}'$ that aligns with $\mathbf{f}$ in both zeroth order and first order. Compared with ω of $\mathbf{f}$ (Equation (9)), ω of $\mathbf{f}'$ (Equation (10)) is thus reduced. Then, Equation (8) implies that using $\mathbf{f}'$, the cosine similarity between the estimated gradient and the true gradient can be improved.

The formal statement and the full proof are deferred to Appendix C.

Based on the above results, we aim to further analyze two research questions.

Can we estimate gradients from a projected low-dimension subspace?
The answer is yes. According to Theorems 1 and 2, the cosine similarity between true gradient and estimated gradient depends on the ratio B/n, rather than only the subspace dimension n. Now we assume the number of queries B is equal to the subspace dimensionality n. We can observe a *sufficient condition* for good cosine similarity: $\|\nabla \mathbf{f}^\intercal \nabla S\|_2 / \|\nabla S\|_2$ is large, i.e., *the gradient of projection function* $\nabla \mathbf{f}$ *and the gradient of difference function* ∇S *align well.* We remark that the condition is independent with subspace dimensionality n, and is checkable when the gradient of the victim model is known. Specifically, when $\|\nabla \mathbf{f}^\intercal \nabla S\|_2 / \|\nabla S\|_2$ achieves its maximum $L_{\mathbf{f}}$, the cosine similarity lower bound becomes

$$\left(2\left(1 - \frac{\omega^2}{\|\nabla \mathbf{f}^\intercal \nabla S\|_2^2}\right)^{(n-1)/2} - 1\right) c_n, \tag{11}$$

where ω could be either defined by Definition 5 for general projection or defined by Theorem 2 for good nonlinear projection.

We can clearly observe that smaller ω leads to better lower bound for cosine similarity, and when $\omega = 0$ the cosine similarity becomes $c_n \in (2/\pi, 1) \approx (.637, 1)$ which is high. To verify this negative correlation between the ω values and the cosine similarity measurements, we conduct empirical experiments in Appendix G.3.

To better analyze the lower bound, we further lower bound Equation (11) as such:

$$\left(2\left(1 - \frac{\omega^2}{\|\nabla \mathbf{f}^\intercal \nabla S\|_2^2}\right)^{(n-1)/2} - 1\right) c_n \tag{11}$$

$$\geq \left(1 - (n-1)\frac{\omega^2}{\|\nabla \mathbf{f}^\intercal \nabla S\|_2^2}\right) c_n. \tag{12}$$

Since the dimensionality n of the sampling space is small, and $\omega = \Theta(\delta)$ where δ is the step size and is also small, we can observe that with the projection, the similarity lower bound is non-trivial, i.e., *we can estimate gradients from a projected low-dimension space.*

When step size δ in ω approaches 0, the ω approaches 0. If S and $\mathbf{f}$ are both locally linear, with $\beta_S = \beta_{\mathbf{f}} = 0$ we again have $\omega = 0$ and the cosine similarity also becomes c_n, which implies that we can achieve high cosine similarity using just linear projection if the difference function S is locally linear.

On the other hand, we inspect the relation between cosine similarity bound and the number of queries

B. As shown in Equation (8), both the lower and upper bound are in $\Theta(\sqrt{B})$ with respect to number of queries. In other words, to achieve a cosine similarity s, one need to perform $\Theta(s^2)$ number of queries. As a result, moderate cosine similarity requires a small number of queries but high cosine similarity requires much more, and it is indeed better to leverage the reduced subspace dimension n. We formalize the query complexity analysis as below and defer the proof and detail discussion to Appendix B.

Corollary 2 (Query Complexity)**.** *Given the projection* $\mathbf{f}$ *and the difference function* S*, to achieve expected cosine similarity* $\mathbb{E}\langle \nabla S(\mathbf{f}(x_0)), \widetilde{\nabla S}(\mathbf{f}(x_0))\rangle = s$*, the required query number* B *is in* $\Theta(s^2)$*.*

How do different projections affect the gradient estimation quality?
The above analysis allows us to compare projection-based gradient estimators in different boundary attacks directly. We instantiate the general bound in Theorem 1 for HSJA and QEBA respectively, which shows that QEBA is significantly better than HSJA as it achieves the same cosine similarity with much fewer queries. For NonLinear-BA, Theorem 2 points out the possibility and a checkable sufficient condition where NonLinear-BA could be better than corresponding linear projection including HSJA and QEBA, in terms of providing higher lower bound of cosine similarity. We further present another sufficient condition in Appendix D.1 under which NonLinear-BA achieves higher lower bound. *In a nutshell, the nonlinear projection which outperforms linear projection is not rare, however, the efficient search algorithm for it with theoretical guarantees is still open.* Thus, in NonLinear-BA we heuristically train popular neural network structures (e.g. AE and GANs) to function as nonlinear projections proxies, so as to analyze their ability of reducing query complexity and achieving precise gradient estimation. The detailed results and discussions can be found in Appendix D.1. We further discuss potential ways of improving the gradient estimation inspired from the theoretical analysis in Appendix D.2.

5 Experiments

In this section, we conduct extensive experiments to evaluate the performance of different boundary blackbox attacks, and show that (1) with the nonlinear and linear projection-based gradient estimation methods, blackbox attacks can achieve better performance compared with the state-of-the-art baselines; (2) both the nonlinear and linear projection-based gradient estimation methods demonstrate their own advantages under certain conditions. In addition, we also show the high blackbox attack performance against commercial face recognition APIs for NonLinear-BA.

5.1 Experimental Setup

Target Models. We use both offline models on ImageNet, CelebA, CIFAR10 and MNIST datasets, and commercial online APIs, as target models following Li et al. (2020). For offline models, on ImageNet, we use a pretrained ResNet-18 as the target model. On CelebA, a pretrained ResNet-18 is fine-tuned to perform classification on attributes as the target model. The most balanced attribute (e.g., 'Mouth_Slightly_Open') is chosen to enhance benign model performance. On CIFAR10 and MNIST datasets, we scale up the input images to 224×224 with linear interpolation to demonstrate the query reduction for high-dimensional input space. Fine-tuned ResNet-18 models are used as target models. The benign target model performance is shown in Appendix E.2. For commercial online APIs, we use the 'Compare' API from MEGVII Face++ (2021a). Given two images, the API returns a confidence score of whether they are of the same person. Following Li et al. (2020), we convert the confidence score to a discrete prediction by taking scores greater than or equal to 50% as 'same person', and vice versa. The implementation details are discussed in Appendix E.1.

Nonlinear Projection Models. The nonlinear projection models are trained on image gradient dataset. The goal is to train the projection models to project from low-dimensional random vectors to higher-dimensional gradient space so that the projected vectors mimic the distribution of the gradient of the target model. The more aligned the projected vectors are with the ground truth gradient vectors, the more effective the Monte Carlo estimation would be. Image gradients are generated using PyTorch's (Paszke et al., 2019) automatic differentiation functions on five reference models for each dataset. The details including model architectures and training parameters are described in Appendix F.1. The benign accuracy for the reference models are shown in Appendix F.3. Note that although in the offline experiments, we use data from the same distribution to train the reference models with different architectures from the target model, this is not a necessary condition: for the attacks on commercial online APIs, we do not have any information about the training dataset or the model structure, and we use the same projection models trained on ImageNet dataset to attack the face recognition APIs. The experimental results in Section 5.3 and Appendix H.3 show that the projection-based boundary blackbox attack works well despite the mismatch of training data distributions between the target model and the projection model.

Evaluation Metrics. We evaluate NonLinear-BA and compare with the baseline methods based on two standard evaluation metrics: (1) the average magnitude

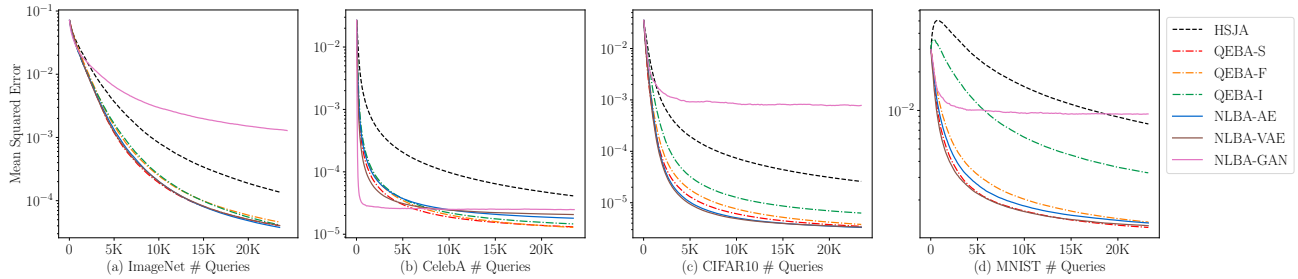

Figure 2: The perturbation magnitude based on different queries for attacks on diverse datasets.

of perturbation at each step, as indicated by the mean squared error (MSE) between the optimized adversarial example and target-image; (2) the attack success rate defined as reaching a specified MSE threshold.

5.2 Blackbox Attack Performance Against Offline Models

Figure 2 shows the attack performance of different approaches in terms of the perturbation magnitude (MSE) between the generated sample and the target-image. The attack success rates are shown in Figure 7 in appendix. The 'NonLinear-BA' is denoted as 'NLBA' in figures. All the results are averaged over 50 randomly sampled pairs of correctly classified images from the corresponding datasets.

The NonLinear-BA with three projection methods exhibit different patterns on the four datasets. NonLinear-BA-AE and NonLinear-BA-VAE are the most consistent across various datasets. They achieve significantly better performance compared with baseline HSJA, and outperform QEBA in many cases. The NonLinear-BA-GAN method, on the other hand, is less stable. For attribute classification model on CelebA, where the model's ground truth gradients have a simpler pattern, it is significant better than the other methods with very few queries (Fig 2(b)). For example, on CelebA dataset, the NonLinear-BA-GAN method takes only a few hundred of queries to get to a distance smaller than 10^{-4} while other projection-based methods take more than one thousand, and the HSJA baseline takes over 10 thousand queries to get to the same distance magnitude. On the other hand, when the gradient patterns are more complex, the NonLinear-BA-GAN method fails to keep reducing the MSE after some relatively small number of queries and converges to a bad local optima. We conjecture this is due to the instability of GAN training, and it would be interesting future work to develop in-depth understanding about the properties of nonlinear GAN-based projection.

Our theoretical analysis and conclusions are well supported by the fact that nonlinear and linear projection models have advantages over each other under various different scenarios, and that except for the unstable NonLinear-BA-GAN case, they both outperform the HSJA baseline which does not have a dimension reduction module via projection.

Verification for Gradient Cosine Similarity and Attack Performance. To verify that the cosine similarity of gradients indeed reflects the blackbox attack performance, we plot the gradient cosine similarity corresponding to different queries in Figure 3. It is clear that the blackbox attack performance highly correlates with the cosine similarity positively: when the cosine similarity is high, the attack performance is better and can converge to a smaller MSE faster.

Theorems 1 and 2 suggest that smaller ω (Definition 5) leads to higher cosine similarity between the estimated and true gradients. To verify it, we use an alternative method to evaluate the effects of ω approximately, and show that cosine similarity is strongly correlated with ω as proved. Details can be found in Appendix G.3.

Case Study: Attack Performance at an Early Stage for CelebA Dataset. We perform a qualitative case study to show the effectiveness of NonLinear-BA. The source-image and target-image are shown in Figure 4. The goal is to generate an adv-image that has a small distance from the target-image with an open mouth, to be mis-recognized as 'Mouth_Slightly_Open=False' by the victim model. The attack results at early attack stages with fewer than 500 queries are shown in Figure 5. We present the results for the baseline HSJA, as well as the baseline QEBA and the proposed NonLinear-BA with the best performance among their variations (e.g., QEBA-S and NonLinear-BA-GAN). The full case study results for all the attack methods are shown in Figure 11 in Appendix H.1. It is obvious that with less than 150 queries, the NonLinear-BA-GAN method already achieves similar or even better performance compared with QEBA-S and HSJA with about 500 queries; with about 250 queries, the quality of the adv-image produced by NonLinear-BA-GAN is so high that visually

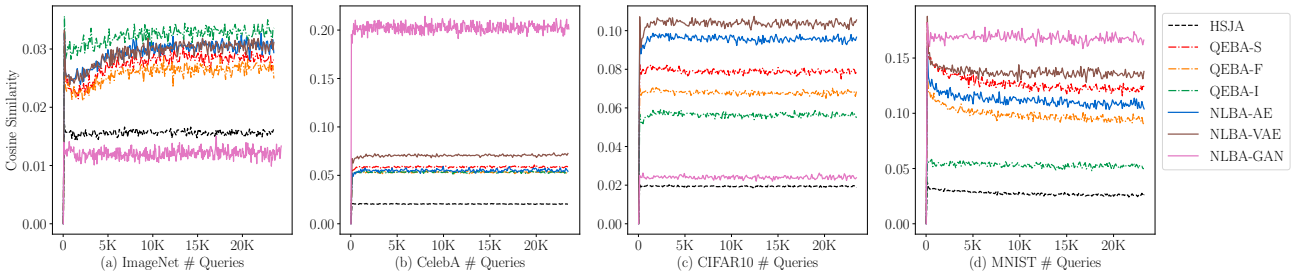

Figure 3: The cosine similarity between the estimated and true gradients with respect to query numbers for attacks on diverse datasets.

(a) Source-image (b) Target-image

Figure 4: The source and target images for the qualitative case study on CelebA dataset. The *attribute-of-interest* is 'Mouth_Slightly_Open', which is labeled as 'False' for the source-image while 'True' for the target-image.

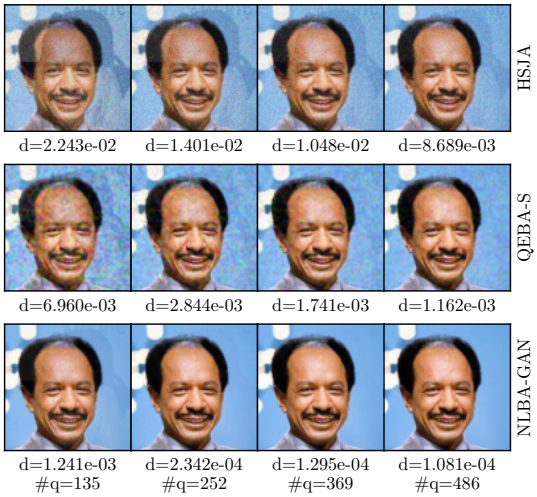

Figure 5: The attack process of different methods on CelebA dataset under different query numbers. d denotes the perturbation magnitude of the generated adversarial example with respect to the target-image. $\#q$ denotes the number of queries.

it is almost indistinguishable compared with the target-image. More qualitative case studies of the attacks can be found in Appendix H.2.

5.3 Blackbox Attack Performance Against Commercial APIs

To demonstrate the practicality of the proposed NonLinear-BA, we perform the blackbox attack against real-world online commercial APIs. Figure 6 shows the MSE between the adv-image and the target-image with different numbers of queries. The attack success rate is always 100% during the whole process. The results are averaged over 40 randomly sampled CelebA face image pairs. The image pairs are the same for each of the seven methods for fair comparison. From Figure 6 it is clear that all the six gradient projection-based methods including both linear and non-linear projections are better than the baseline HSJA in terms of the MSE under the same number of queries, and the nonlinear projection converges faster while observes slightly higher perturbation magnitude. The qualitative results of case studies are shown in Appendix H.3.

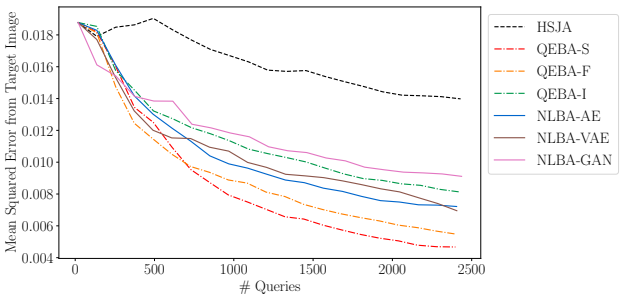

Figure 6: The perturbation magnitude based on different queries against Face++ 'Compare' API.

6 Conclusion

We provide the first theoretic analysis framework for projection-based gradient estimation. We then propose NonLinear-BA, a nonlinear projection-based gradient estimation approach for query-efficient boundary black-box attack. We theoretically show nontrivial cosine similarity bounds for a group of projection based gradient estimation approaches and analyze the properties of different projections. We evaluate the efficiency of NonLinear-BA with extensive experiments against both offline ML models and commercial online APIs.

Acknowledgement This work is partially supported by Amazon research award.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Exploring the space of black-box attacks on deep neural networks. *arXiv preprint arXiv:1712.09491*, 2017.

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *European Conference on Computer Vision*, pages 158–174. Springer, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy (SP)*, pages 668–685. IEEE, 2020.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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*, pages 15–26, 2017.

Minhao Cheng, Simranjit Singh, Patrick H Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019a.

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 32*, pages 10932–10942, 2019b.

Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

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

Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

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*, 2018.

George Marsaglia et al. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646, 1972.

MEGVII. Facial recognition 'compare' api. `https://console.faceplusplus.com/documents/5679308`, 2021a.

MEGVII. Facial recognition 'compare' api query url. `https://api-us.faceplusplus.com/facepp/v3/compare`, 2021b.

MEGVII. Face++. `https://www.faceplusplus.com/`, 2021c.

Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019.

PyTorch. Torchvision.models. `https://pytorch.org/docs/stable/torchvision/models.html`, 2021.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems 31*, pages 1299–1309, 2018.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Huan Zhang, Pengchuan Zhang, and Cho-Jui Hsieh. Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5757–5764, 2019.