Less is More: Dimension Reduction Finds On-Manifold Adversarial Examples in Hard-Label Attacks

Washington Garcia*, Pin-Yu Chen†, Hamilton Scott Clouse‡, Somesh Jha§ and Kevin R.B. Butler*
*Department of Computer and Information Science and Engineering, University of Florida
{w.garcia,butler}@ufl.edu
†IBM Research
pin-yu.chen@ibm.com
‡ACT3, Air Force Research Laboratory
hamilton.clouse.1@afrl.af.mil
§Computer Sciences Department, University of Wisconsin
jha@cs.wisc.edu

***Abstract*—Designing deep networks robust to adversarial examples remains an open problem. Recently, it was shown that adversaries relying on only top-1 feedback (i.e., the hard-label) from an image classification model can arbitrarily shift an image towards an intended target prediction. Likewise, these hard-label adversaries enjoy performance comparable to first-order adversaries relying on the full model gradient. It was also shown in the gradient-level setting that regular adversarial examples leave the data manifold, while their on-manifold counterparts are in fact generalization errors. In this paper, we argue that query efficiency in the hard-label setting is also connected to an adversary's traversal through the data manifold. To explain this behavior, we propose an information-theoretic argument based on a *noisy manifold distance oracle*, which leaks manifold information through the adversary's distribution of gradient estimates. Through numerical experiments of manifold-gradient mutual information, we show this behavior acts as a function of the effective problem dimensionality. On high-dimensional real-world datasets and multiple hard-label attacks using dimension reduction, we observe the same behavior to produce samples closer to the data manifold. This can result in up to 10x decrease in the manifold distance measure, regardless of the model robustness. Our results suggest that our variant of hard-label attack can find a higher concentration of generalization errors than previous techniques, leading to improved worst-case analysis for model designers.**

***Index Terms*—adversarial machine learning, zero knowledge attacks**

I. Introduction

Adversarial examples against deep learning models have become a persistent topic of investigation in recent years, as they offer a principled approach to studying worst-case behavior in machine learning systems [3], [4]. Formal methods for discovering adversarial examples were originally conceived by assuming gradient-level (i.e., first-order) access to machine learning models, and these became the first techniques to reach widespread attention within the deep learning community [5]–[9]. In order to compute the necessary gradient information, gradient-level techniques required access to the sensitive model parameters and a sizeable query budget. These shortcomings were later addressed by the creation of score-level attacks, which only required the confidence values output by the deep learning models [10]–[13]. However, score-level attacks still rely on models to divulge information that would be impractical to receive in real-world systems. By contrast, *hard-label attacks* make no assumptions about available model information, and instead rely only on the top-1 predicted class, thus providing the weakest, yet most realistic adversarial threat model to date. Hard-label attacks have traditionally been formulated by leveraging a framework known as zeroth-order optimization, which allows to estimate the input gradient of

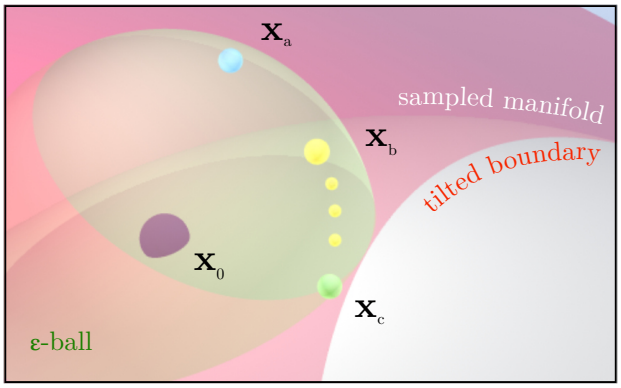


Figure 1. Our geometric interpretation of hard-label attack behavior in the context of boundary tilting [1] (red) around original sample $\mathbf{x}_0$ within a green ϵ-ball: ($\mathbf{x}_a$) zeroth-order attack using any search direction, leaving the manifold, ($\mathbf{x}_b$) an efficient zeroth-order attack (and dimension-reduced variants, denoted by smaller points), which find on-manifold directions and maximize manifold-gradient mutual information through dimension-reduction, and ($\mathbf{x}_c$) Stutz et al. [2]-style attack on the manifold, where adversarial samples are least concentrated, and may not be found within the ϵ radius.

This work was supported by AFOSR FA8650-19-1-0169.

a sample with respect to the adversary's desired label [14]–[16]. These methods have been carefully refined to offer convergence guarantees [17], query efficiency [18], [19], and powerful capabilities in the physical world [20].

Despite the steady improvements of hard-label attacks, open questions persist about their behavior and adversarial machine learning (AML) attacks at large, particularly in the task of image classification (the focus of this work). Adversarial examples were originally assumed to lie in rare pockets of the input space [4], but this conventional wisdom was later challenged by the boundary tilting assumption [1], [21], which adopts a "data-geometric" view that observable inputs concentrate on a lower-dimensional connected region, called a manifold. Through this lens, Stutz et al. [2] suggest that adversarial examples can be categorized based on their distance to the manifold. Adversaries may add unnatural high-frequency distortion (e.g., noise), causing the image to leave the data manifold to cross the model's decision boundary (blue orb in Figure 1). Another option is to only add natural distortions that alter the image without changing the true label (e.g., color shifting or distorting entire pixel groups). This latter option produces on-manifold adversarial examples, which are essentially generalization errors, i.e., examples not captured by the model's approximation of the manifold (green orb in Figure 1). One can also produce arbitrarily near-manifold examples, which can be considered on-manifold for brevity (yellow orbs in Figure 1). As a direct consequence, an adversarial example's distance to the manifold describes the feasibility for such an example to generate naturally from the true data distribution. This makes it advantageous to produce on-manifold adversarial examples, since the adversary can exploit the inherent generalization error of the model, while producing samples that are perceptually similar for humans. Discovery of generalization errors are also motivated from the perspective of adversarial training, since it was observed that models adversarially trained with traditional off-manifold examples can be easily bypassed by perceptual distance attacks in the gradient-level setting [22]. Unfortunately, the *true* data manifold is either difficult or impossible to describe, and relying solely on approximations of the manifold can lead to the creation of crude, high distortion examples [2].

In this paper, we adopt the boundary-tilting assumption and demonstrate an unexpected benefit of dimension-reduced zeroth-order attacks. These attacks are more likely to discover on-manifold examples, which we theoretically demonstrate is the result of manifold-gradient mutual information. Our results suggest that this quantity can *increase* when data dimensionality is reduced, allowing an attacker to learn useful semantic variations of the data from the distribution of gradient estimates alone. With this knowledge, we empirically demonstrate how to improve hard-label worst-case analyses in a generic yet principled way, by enabling the discovery of new generalization errors within "robust" models that may be useful for adversarial training in future work. Due to insights from our experiments, we provide a geometric interpretation of hard-label attack behavior, summarized in Figure 1.

Our specific contributions are as follows, with key contributions and flow of arguments illustrated in Figure 2:

- **Introduction of manifold distance oracle.** To create on-manifold examples, the adversary must leverage manifold information during the attack phase. We thus propose an information-theoretic formulation of the noisy manifold distance (NMD) oracle, which can explain how zeroth-order attacks craft on-manifold examples. We theoretically demonstrate on a Gaussian data model that *manifold-gradient mutual information can increase as a function of data dimensionality*. We empirically show this is true even on large-scale image datasets such as CIFAR-10 and ImageNet. This finding relates to known behavior in the gradient-level setting, where useful variations of the data (e.g., shapes and textures) can be leaked from the model gradient [23].
- **Reveal new insights of manifold feedback during query-efficient zeroth-order search.** In practice, the data manifold is difficult to characterize. We propose the use of two proxies for manifold distance, which all show consistent results in terms of an adversary's ability to search near the manifold. This methodology allows us to empirically demonstrate the connection between dimension reduction, model robustness, and manifold feedback from the model, beyond the known convergence rates tied to dimensionality [24]. Our findings inform how to search closer to the manifold (Table I and Tables II-V), improve query efficiency (Figure 6), and reduce gradient deviation (Table VI) in a simple and generic way for hard-label attacks.
- **Attack-agnostic method for semantic super-pixel construction.** We show that spatial dimension reduction of a decision-based gradient estimate acts as an attack- and knowledge-agnostic method for learning the most important super-pixels of an image. More importantly, this allows an attacker to synthesize the semantic search directions for a sample using *only* the knowledge gathered during hard-label boundary traversal. Our attack formulation leads to a 2x success rate improvement for state-of-the-art hard-label attacks such as HSJA [18], accompanied by up to 10x lower manifold distance, well below the manifold distance of previously-proposed RayS [25], the current state-of-the-art for crafting on-manifold examples.

Code to reproduce experiments is made available online.[1]

II. Related Work

We investigate the scenario where an adversary uses top-1 label feedback to estimate an input input gradient direction with respect to a desired label, showcased by attacks such as Sign-OPT [19] and HopSkipJumpAttack [18]. These contemporary attacks are variants of random gradient-free method (RGF) [24] and aim to approximate the true input gradients $\mathcal{G}$ with respect to the adversarial label through zeroth-order estimates $\tilde{\mathcal{G}}$. The core idea is to convert the top-1 (hard) label, which is a step function, into a continuous real-valued function

[1] https://github.com/FICS/hard_label_manifolds

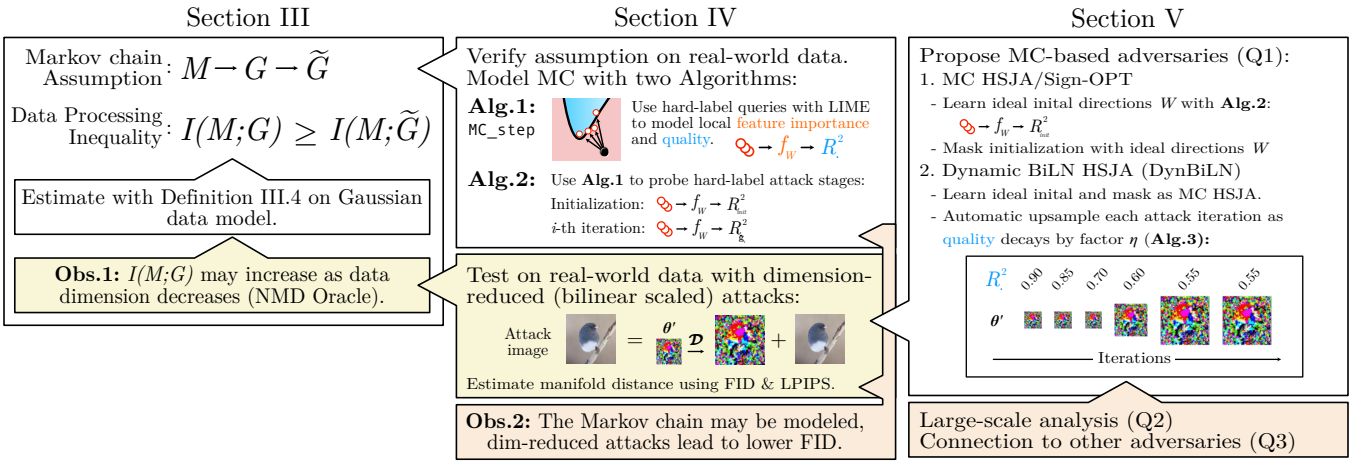


Figure 2. Approach flowchart for our results based on the Markov chain (MC) assumption (left) and subsequent analysis using interpretable modeling techniques (middle). Based on empirical observations on real-world data, we propose MC-based adversaries (right) and show a tendency for dimension-reduced attacks to reduce manifold distance estimates in large-scale datasets.

$g : \mathbb{R}^d \rightarrow \mathbb{R}$, which takes search direction $\boldsymbol{\theta} \in \mathbb{R}^d$ and outputs the distance to the nearest adversarial example [15]. The gradient estimates $\tilde{\mathcal{G}}$ are conceived as a function of the local search direction gradient ∇g, and can be estimated with either two samples of information (Sign-OPT) [19], or a single point (HopSkipJumpAttack) [18]. Details of specific formulations for each attack are provided in Appendix B.

The theory of gradient estimation error and convergence provides some clues to improve efficiency, such as the fact that the estimation cost is polynomial in d, the dimension of the optimized variable, thus motivating standard dimension-reduction techniques based on autoencoding [26]. Alternatively, it is possible to change entire groups of pixels at a time, a technique known as super-pixel grouping, which is effective in attacks that leverage heuristic-based search rather than gradient estimation, such as RayS [25]. However, it is not completely understood how dimensionality-reduction relates to the adversary's probing of the data manifold. In the score-level setting, it was shown that better query efficiency can be achieved by leveraging time- and data-scale dependencies within the distribution of gradient estimates [27]. We examine analagous dependencies in service of generating on-manifold adversarial examples, since to date it is unclear how such dependencies interact with the model's encoded manifold, particularly in the hard-label setting.

III. Noisy Manifold Distance Oracle

Main idea: Let $\mathcal{F}$ be a classifier that returns the top-1 predicted label for some query, $\mathcal{G}$ the distribution of true input gradients with respect to an adversary's desired label, and $\tilde{\mathcal{G}}$ the distribution of respective hard-label gradient estimates. By the Data Processing Inequality (Definition III.1) we posit that their mutual information (I) has the relation $I(\mathcal{F};\mathcal{G}) \geq I(\mathcal{F};\tilde{\mathcal{G}})$. We use $I(\mathcal{F};\tilde{\mathcal{G}})$ as a measure for how much information $\mathcal{F}$ leaks in the gradient estimates $\tilde{\mathcal{G}}$. The main hypothesis is that as data dimension d increases, so does $I(\mathcal{F};\mathcal{G})$, but at a lower rate when scaling for d in the Schmidt et al. [28] data model, giving higher $I(\mathcal{F};\mathcal{G})$ at lower d.

A. Preliminaries

A common observation is that due to the low probability of encountering "interesting" images in an entire input space (e.g., images of numbers or animals), the data of interest highly concentrates along a lower-dimension region of connected points, referred to as a *manifold* and denoted $\mathcal{M}$ [1], [29], [30]. The manifold is considered lower-dimension because there are generally only a few valid variations for interesting inputs, such as translating or rotating objects in images, and connected because traversal along the manifold yields new valid inputs (i.e., rotating the object in an image does not change the object's true label). Likewise, one can consider a subset of the manifold dimensions, which is analagous to considering a subset of useful variations. It is also known that deep learning models encode information about the data manifold [31]. Thus we denote the model's learned manifold as $\mathcal{M}(\mathcal{F})$ and use $\mathcal{M}$ for brevity.

B. Approach

Due to the fact that observable data generates from a random process, any function of the observed data (and corresponding manifold) is random, and thus the model point estimate is also a random variable. As described in Section II, hard-label attacks assume the existence of true input gradients $\mathcal{G}$ (a function of the model's point estimate) which points in the direction of a desired adversarial label. The task of hard-label attacks is to synthesize an estimate of the true input gradients, $\tilde{\mathcal{G}}$, using only the top-1 feedback from the model. Since the victim model's loss cannot be observed, hard-label attacks usually synthesize a continuous surrogate loss for estimation, e.g., distance to nearest adversarial sample, meaning the adversary relies on data-driven feedback. We assume that a Markov chain maps the hard-label attack pipeline originating from the data manifold: $\mathcal{M} \rightarrow \mathcal{G} \rightarrow \tilde{\mathcal{G}}$, where $\mathcal{G}$ ($\tilde{\mathcal{G}}$) describes

the true (estimated) input gradient distributions. From the hard-label attacker's perspective, the Markov chain is only partially observable due to having access to top-1 feedback alone. We posit that the adversary may be able to increase their available information by leveraging a basic result from data processing.

Definition III.1 (Data Processing Inequality (DPI) [32])**.** If three random variables form the Markov chain $X \to Y \to Z$, then their mutual information (I) has the relation $I(X;Z) \leqslant I(X;Y)$.

In the context of Definition III.1 and our hard-label attack pipeline ($\mathcal{M} \to \mathcal{G} \to \tilde{\mathcal{G}}$), $I(\mathcal{M};\mathcal{G})$ acts as an upper bound on $I(\mathcal{M};\tilde{\mathcal{G}})$. In the information-theoretic sense, increasing the upper bound allows an optimal hard-label adversary to further maximize $I(\mathcal{M};\tilde{\mathcal{G}})$, i.e., make the adversarial sample look as if it generated from the true data distribution, which is desireable for exploiting the generalization error of the model as discussed by [2]. Ideally, one would also like to show the lower bound of $I(\mathcal{M};\tilde{\mathcal{G}})$ will also increase with the upper bound if the adversary is stronger. However, to the best of our knowledge such an ideal lower bound (e.g., some constant $c < 1$ such that $c \cdot I(\mathcal{M};\mathcal{G})$ is a lower bound of $I(\mathcal{M};\tilde{\mathcal{G}})$) does not exist in general, unless one makes further assumptions (e.g., the true distribution is a known Gaussian) [33]. Leveraging DPI is a natural interpretation of attack behavior, since maximum likelihood estimation (MLE) can be interpreted as minimizing the dissimilarity between the empirical data distribution (defined by the data manifold) and the learned model distribution (defined by the point estimate). In practice the dissimilarity is quantified by measures of KL-divergence (e.g., binary cross entropy loss between distributions of interest) [29], [33]. Although DPI only offers an upper bound, rather than a guaranteed increase, it allows us to operationalize the approach to hard-label attacks by attempting to increase $I(\mathcal{M};\mathcal{G})$. Since the states are only partially observable by the adversary, how could they increase the right-hand term $I(\mathcal{M};\mathcal{G})$? Recent work by [28] prove that adversarial robustness requires a significantly larger number of data samples as a function of data dimensionality (e.g., $\sqrt{d}$ larger sample complexity for Gaussian data models), which can be argued as an effect of the curse of dimensionality. In a similar twist, [21] showed that the volume of an error region along a subset of "adversarial directions" becomes exceedingly small with high data dimensionality. Under the data model of [28], we expect the data categories to become more separated with higher dimensionality, effectively concentrating in smaller regions. This leads us to our main hypothesis.

Hypothesis 1. *Assume the manifold-gradient Markov chain $\mathcal{M} \to \mathcal{G} \to \tilde{\mathcal{G}}$ exists. Then if we decrease the data dimension d by considering only a subset data dimensions, or equivalently, considering lower-dimension versions of the manifold distribution, we can yield higher mutual information $I(\mathcal{M};\mathcal{G})$ in the Schmidt et al. [28] data model.*

In order to empirically verify Hypothesis 1, we perform synthetic experiments quantifying $I(\mathcal{M};\mathcal{G})$ under different dimensionality values. In particular, we leverage the Gaussian data model and results from [28] to derive an analytical solution for $I(\mathcal{M};\mathcal{G})$.

Definition III.2 (Data model and optimal weights [28])**.** Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be the per-class centers (means) and let $\sigma > 0$ be the variance parameter. Then the $(\boldsymbol{\mu}, \sigma I)$-Gaussian model is defined by the following distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$: First, draw a label $y \in \{\pm 1\}$ uniformly at random. Then sample the data point $\mathbf{x} \in \mathbb{R}^d$ from $\mathcal{N}(y \cdot \boldsymbol{\mu}, \sigma I)$.

Definition III.3 (Optimal classification weight [28])**.** Fix $\sigma \leq c_1 d^{\frac{1}{4}}$ for the universal constant c_1, and samples $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ drawn $i.i.d$ from the $(\boldsymbol{\mu}, \sigma I)$-Gaussian model with $||\boldsymbol{\mu}|| = \sqrt{d}$ (i.e., $\boldsymbol{\mu}_k = 1$ for all dimensions $k \in \{0, \ldots, d\}$). Schmidt et al. [28] prove the weight setting $\widehat{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{x}_i$ yields an l_∞^ϵ-robust classification error of at most 1% for the linear classifier $f_{\widehat{\mathbf{w}}} : \mathbb{R}^d \to \{\pm 1\}$ instantiated as $f_{\widehat{\mathbf{w}}}(x) = \text{sgn}(\widehat{\mathbf{w}}^T \mathbf{x})$ if

$$n \geq \begin{cases} 1, & \text{for } \epsilon \leq \frac{1}{4} d^{-\frac{1}{4}} \\ c_2 \epsilon^2 \sqrt{d}, & \text{for } \frac{1}{4} d^{-\frac{1}{4}} \leq \epsilon \leq \frac{1}{4} \end{cases}, \tag{1}$$

for a universal constant c_2.

Note that the instantiation of $\widehat{\mathbf{w}}$ must change with choice of ϵ and d. We can leverage the weight settings as a function of n and d to give a definition of manifold-gradient mutual information.

C. Manifold-Gradient Mutual Information

Notice the classifier $\text{sgn}(\cdot)$ in Definition III.3 is discontinuous at $\mathbf{x}_k = 0$ for any dimension k. Instead we consider the sub-gradient of the classifier at $\mathbf{x}_k < 0$ and $\mathbf{x}_k > 0$. In either case (non-robust or robust), the input sub-gradient for $f_{\widehat{w}}(\mathbf{x}'_k)$ is defined dimension-wise for our isotropic Gaussian as $\nabla_{\mathbf{x}'_k} f_{\widehat{\mathbf{w}_k}} = \text{sgn}(\mathbf{w}_k)$. Since the weight of each dimension is Gaussian distributed with $\widehat{\mathbf{w}_k} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2)$, we can define the distribution of gradients as $\mathcal{G} \sim \text{Rademacher}\left(\mathbb{P}_{\widehat{\mathbf{w}_k} \sim \mathcal{N}}\left[\widehat{\mathbf{w}_k} \geq 0\right]\right)$. Using this fact, we define manifold-gradient mutual information in three parts: (1) defining the manifold-gradient point-wise joint probabilities between $\mathbf{g}_k$ and $\mathbf{x}_k$ at each dimension k for the sub-gradient cases where $\mathbf{x}_k > 0$ and $\mathbf{x}_k < 0$, (2) defining the manifold-gradient marginal probability under the gradient, and (3) defining the marginal probability under the manifold. The complete derivation of the joint and marginal probabilities can be found in Appendix A. The three parts are used in the standard definition of mutual information [33].

Notation. Fix $\sigma = c_1 d^{\frac{1}{4}}$ for both cases. We denote the sub-manifold sampled from the positive ($y = 1$) and negative ($y = -1$) classes as $\mathcal{M}^+$ and $\mathcal{M}^-$, respectively. For brevity we label $\mathbf{x}_k > 0$ as $\mathbf{x}^+$ and $\mathbf{x}_k < 0$ as $\mathbf{x}^-$.

Definition III.4 (Manifold-Gradient Mutual Information)**.** We define the manifold-gradient mutual information, based on the

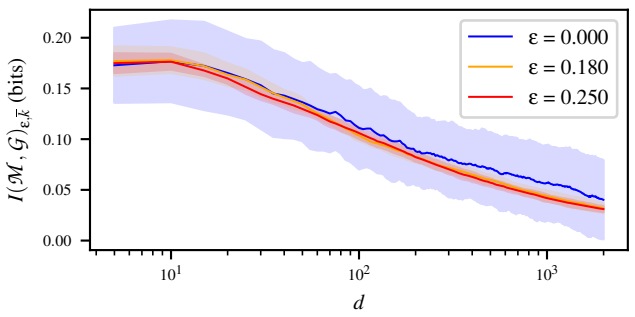


Figure 3. Average per-dimension mutual information (I) over dimension d for values of ϵ in Equation 2, log-scale d-axis with $d \in [5, 2000)$, average over ten seeds. The approximate mutual information is higher for robust and standard models at lower d regardless of ϵ (all nearly overlapping).

standard definition of mutual information from information theory [33], as

$$I(\mathcal{M};\mathcal{G})_{\epsilon,k} = 2\int_{\mathcal{M}^+} p(1,\mathbf{x}^+)\log(\frac{p(1,\mathbf{x}^+)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}^+)})\,d\mathbf{x}^+ + 2\int_{\mathcal{M}^+} p(-1,\mathbf{x}^+)\log(\frac{p(-1,\mathbf{x}^+)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}^+)})\,d\mathbf{x}^+. \quad (2)$$

with the total unnormalized mutual information defined as the summation over dimensions (due to dimension co-independence) $I(\mathcal{M};\mathcal{G})_\epsilon = \sum_{k=1}^{d} I(\mathcal{M};\mathcal{G})_{\epsilon,k}$.

D. Mutual information as a function of dimensionality

To provide numerical support for Hypothesis 1, we run experiments using the Riemann approximation of Equation 2, provided in the Appendix as Equation 15. We estimate the average per-dimension mutual information, $I(\mathcal{M};\mathcal{G})_{\epsilon,\overline{k}} = \frac{I(\mathcal{M};\mathcal{G})_\epsilon}{d}$, for the case where $\mathbf{x} \in \mathbb{R}^d$ while varying the dimensionality term $d \in [5, 2000)$ with $\epsilon \in \{0.000, 0.180, 0.250\}$. We target an error within 10^{-1} (i.e., $0.9 \leq p_{\mathcal{G}}(1) + p_{\mathcal{G}}(-1) \leq 1.0$) by setting the multiplicative factor $c_2 = 10^2$, and multiplying each branch of Equation 1 by a large constant (10^4). We run the approximation over ten different random seeds and show the average with standard error shaded.

The estimation result is shown in Figure 3 with log-scale x-axis. Regardless of ϵ, lower values of the dimensionality evidence a higher mutual information. Since the curve is normalized with respect to d, it shows that although mutual information scales with d, the rate of increase for mutual information is slower than the rate of increase for d.

Observation 1. *Given reduced data dimensionality, an adversary could increase $I(\mathcal{M};\mathcal{G})_{\epsilon,\overline{k}}$ and lead to leaking better search direction through the gradient. This supports Hypothesis 1.*

An implication of Observation 1 is that better search directions come from a better encoding of the data manifold. This same behavior was observed empirically by [23] and [30], who demonstrated that the input gradient from a "robustified" model leads to synthesizing images with higher visual alignment to the training data, which are essentially on-manifold synthetic images. In this capacity, it is useful to describe any sample's distance to the true manifold w.l.o.g as *manifold distance*, often quantified by the L_p-norm of the distance between a synthetic sample (e.g., machine-generated or adversarial sample) and the original when projected onto an approximate manifold representation [2], [31]. Given Observation 1 and the previous empirical results by Santurkar et al. [30], we denote the true input gradient distribution $\mathcal{G}$ as a *manifold distance oracle*, because it offers useful variations of the input data for decreasing manifold distance. In the hard-label setting, the data manifold, true gradient, and model parameters are not accessible, but they may be connected due to the Markov chain assumption of Hypothesis 1. Thus we propose the *noisy* manifold distance (NMD) oracle, an abstraction of the information captured by the distribution of gradient estimates $\tilde{\mathcal{G}}$. From the security perspective, the NMD oracle acts as a side channel leaking information (e.g., useful semantic directions) as a factor of the data dimensionality.

IV. Zeroth-order Search Through the Manifold Distance Oracle

According to Observation 1, dimensionality reduction could increase the upper bound on $I(\mathcal{M}, \tilde{\mathcal{G}})$. Under our Markov chain assumption, this means the distribution of gradient estimates from an attack algorithm can act as a noisy manifold distance oracle. However, Observation 1 mainly applies to a Gaussian data model with a simple manifold, thus it does not necessarily lift into real-world datasets. Likewise, to our best knowledge, previous work have yet to decompose the hard-label attack process into the Markov chain described in Section III. Since Observation 1 relies on its existence, we proceed by first formulating dimension-reduced versions of existing hard-label attacks, then demonstrating through an empirical study that in practice, the Markov chain assumption is reasonable due to the learnability of semantic directions in the neighborhood around a hard-label sample. We subsequently show that lower search dimensions lead to a higher quality model of the adversarial sample's semantic directions, which in turn enables a lower manifold distance measured by standard metrics such as LPIPS and FID, introduced later.

A. Dimension-reduced zeroth-order search

We modify existing hard-label attacks to produce dimension-reduced variants. In practice we modify attacks so they generate dimension-reduced gradient estimates to update their candidate search direction $\boldsymbol{\theta}' \in \mathbb{R}^{d'}$ for reduced image dimension d', where $d' < d$ and d is the original data dimension. To update an adversarial example and query the victim model, the candidate direction is upsampled using a decoding map $\mathcal{D} : \mathbb{R}^{d'} \to \mathbb{R}^d$. In general the adversarial sample is created by $\mathbf{x} = \mathbf{x}_0 + g\left(\mathcal{D}(\boldsymbol{\theta}')\right)\frac{\mathcal{D}(\boldsymbol{\theta}')}{||\mathcal{D}(\boldsymbol{\theta}')||}$, where $g : \mathbb{R}^d \to \mathbb{R}$ is a regularizing function dependent on the attack formulation. For Sign-OPT attack, g is the distance to the decision boundary in direction $\mathcal{D}(\boldsymbol{\theta}')$, while for HSJA, g represents the optimal step-size in direction $\mathcal{D}(\boldsymbol{\theta}')$. For our purposes, the decoder function $\mathcal{D}$ is initialized with a

Algorithm 1: Local Markov chain step (`MC_step`)

Input: Hard-label Gaussian process (`GP`), LIME kernel width k

Output: Sample feature coefficients $W \in \mathbb{R}^d$ and their quality score $R^2 \in \mathbb{R}$, `GP` result (`res`)

1 initialize LIME Ridge regression trainer (`LIME`) [34]
2 `/* Execute GP to collect samples */`
3 $X, Y, \texttt{res} \leftarrow \texttt{GP}()$
4 $f_W \leftarrow \texttt{LIME}(X, Y, k)$
5 $R^2 \leftarrow f_W(X)$
6 **return** W, R^2, res

bilinear upsampler (henceforth referred to as BiLN followed by d') which simply scales the height and width dimensions of candidate directions (shaded middle section of Figure 2). It is possible to instantiate more complex mapping functions, for example the use of autoencoders [2], [26]. We forego analysis of autoencoder variants, since hard-label adversaries may not have access to the full training or test set necessary to create such an approximation in the first place.

B. Attack-agnostic model of the manifold-gradient Markov chain

Due to the hard-label threat model, an adversary does not have access to the true manifold distribution (e.g., training samples) $\mathcal{M}$ or the distribution of true input gradients $\mathcal{G}$. The distribution of gradient estimates $\tilde{\mathcal{G}}$ is updated over the course of a hard-label attack, and in the score-level setting, is known to possess useful time-scale dependencies [27]. Unfortunately, current hard-label attacks immediately discard the distribution after each attack iteration [18], [19]. Due to the complexity of real-world data, even if each of the global distributions was known, it would be intractable for designers, much less adversaries, to analytically model every possible relationship between manifold points and gradient estimates. Instead, we can study the *local Markov chain* at a single point on the manifold (e.g., starting from a singular clean image). If an adversary can learn the semantic features of an image using *only* previous adversarial attempts, it may validate the Markov chain assumption.

It was previously shown that a designer can find a sample's semantic directions using only a linear model of their classifier's local decision boundary, e.g., the LIME technique proposed by Ribeiro et al. [34]. The linear model (standard Ridge regression) is trained using uniform-random perturbations around a sample. The linear model coefficients are interpreted as per-pixel feature importance, the quality of which is measured by the linear model's R^2 score. Rather than use uniform-randomly perturbed samples, we propose to leverage samples collected during hard-label attack initialization and gradient estimation to train the LIME model. If the hard-label adversary's LIME model yields a high quality (i.e., high R^2) set of coefficients, it shows that the local manifold-gradient Markov chain can be directly modeled over the course of an attack. In fact, the R^2 score is a weighted estimate based on an exponential kernel, which in our implementation is chosen to correspond with a local measure of mutual information (i.e,. binary cross entropy between original sample and hard-label attempts). We can later use the R^2 score to measure the quality of the linear model after applying different dimension-reduction schemes, providing additional empirical validation for Observation 1.

In practice, LIME was shown to have unstable behavior due to the difficulty in selecting a kernel width for the weighted estimate of coefficients [35]. Instead we leverage OptiLIME, which factorizes individual training runs of LIME as Gaussian processes, and maximizes R^2 by trying different kernel width selections at each run through Bayesian optimization [35]. Similarly, we decompose the typical hard-label attack into independent Gaussian processes corresponding to common parameter-independent attack stages, e.g., sample initialization, gradient approximation, and binary search for nearest boundary [15], [18], [19]. A stage can be called multiple times to collect independent sets of samples, allowing to train OptiLIME's Bayesian classifier and obtain optimal coefficients for that stage. We implement this logic for whole-attack Markov chain (MC) analysis across two algorithms, illustrated visually in Figure 2. The first algorithm, described in Algorithm 1 (`MC_step`), takes as input an attack stage Gaussian process (denoted `GP`) and acts as the vehicle for collecting a neighborhood of hard-label samples around the attacker's current sample. LIME is used in this algorithm alongside a *candidate* kernel width k to train a linear model f, parameterized by model coefficients $W \in \mathbb{R}^d$, which outputs an R^2 score that acts as an objective function of the candidate kernel width. Thus `MC_step` is used as the input to OptiLIME to conduct independent trials for kernel width optimization at each attack stage.

At each attack stage, `MC_step` is instantiated with the stage's Gaussian process and given to OptiLIME in Algorithm 2 (Lines 2 and 9), which handles the collection of feature coefficients and linear model scores at different attack stages for a generic hard-label attack. For simplicity, we only consider attack initialization (Line 1-3) and subsequent gradient estimation attempts (Lines 7-9). Although the OptiLIME Bayesian optimization requires several trials to choose an optimal kernel width, in practice OptiLIME can obtain reasonable R^2 scores from only twenty LIME training runs (each requiring 10-200 hard-label queries, depending on the attack and particular stage). The final output of Algorithm 2 is the feature coefficients W and corresponding R^2 score at each attack stage from initialization (`Init`) through to the i-th gradient estimation step ($\hat{g}_i$). We emphasize that Algorithm 2 does not interfere with the initialization or gradient approximation of the attack, and merely serves to model the manifold-gradient Markov-chain in generic hard-label attacks. An example of this process is illustrated in Figure 4 on a single sample from CIFAR-10. Each column after the input represents the activated feature coefficients at the specified attack stage, visualized by white pixels (positive coefficients)

Algorithm 2: Markov chain probing of hard-label attack

Input: Benign sample $\mathbf{x}_0$, Hard-label attack initialization (`init`) and gradient approximation (`approximate_gradient`), OptiLIME bayesian optimization routine for kernel width (`OptiLIME`), num. iterations n

Output: Hard-label attack sample $\mathbf{x}$, whole-attack feature coefficients $\{W_{\texttt{init}}, W_{\hat{\mathbf{g}}_1}, \ldots, W_{\hat{\mathbf{g}}_n}\}$ and their quality scores $\{R^2_{\texttt{init}}, R^2_{\hat{\mathbf{g}}_1}, \ldots, R^2_{\hat{\mathbf{g}}_n}\}$

```
GP := (init, x_0)
kernel width k ← OptiLIME(MC_step, GP)
/* Initialize through MC_step */
W_init, R^2, x ← MC_step(GP, k)
for i := 1 to n do  Hard-label attack loop
    GP ← (approximate_gradient, x)
    k ← OptiLIME(MC_step, GP, x')
    /* Approximate through MC_step */
    W_ĝi, R^2, θ ← MC_step(GP, k)
    Update x from θ using attack formulation
end
return x, {W_init, W_ĝ1, ..., W_ĝn},
  {R^2_init, R^2_ĝ1, ..., R^2_ĝn}
```

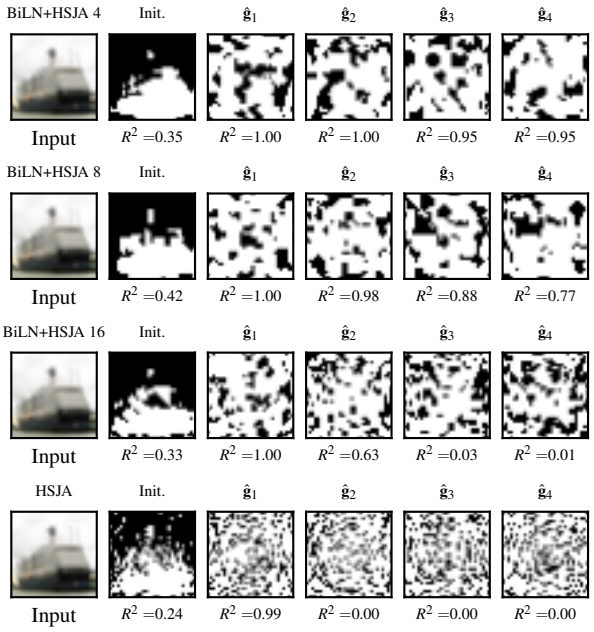


Figure 4. Comparison of learned feature importance coefficients from hard-label boundary traversal alone, each column representing different stages of attack. Dimension-reduced attack variants (first two rows) can isolate the ship in the single input image with better clarity during initialization (quantified by higher R^2 score), and enable larger pixel grouping during subsequent estimation stages.

and black pixels (negative coefficients). Notably, the hard-label adversary's initialization (`Init.`) is enough to learn a coarse spatial representation of the subject in the image, seen as the rough outlines of the cruise ship in the second column. Counter-intuitively, the subject can be isolated with better detail using dimension-reduced variants of HSJA (e.g., BiLN 4 in row one and BiLN 8 in row two) compared to regular HSJA in the final row. The increased semantic quality of the dimension-reduced variants is reflected in the higher R^2 score of the learned feature coefficients (e.g., $R^2 = 0.42$ of BiLN+HSJA 8 compared to $R^2 = 0.24$ of regular HSJA). Likewise, we see that the feature importance of subsequent gradient estimation steps is scattered throughout the spatial dimension of the image, although larger patches of the image are affected in lower-dimension variants, which retain a higher R^2 score for more iterations.

To gain a holistic view of the phenomenon observed in Figure 4, we attack 500 samples on a standard CIFAR-10 model with variants of HSJA in the l_∞ setting and obtain the mean R^2 score at each stage using Algorithm 2, shown in Figure 5. As before, we observe that on average, each HSJA variant can learn a reasonable representation of the semantic features in the image, evidenced by high R^2 scores up to the second gradient estimation stage, thereafter decreasing quickly depending on the use of dimension reduction. Surprisingly, dimension-reduced variants can retain high semantic quality of the learned feature coefficients for more iterations, as evidenced by the ability for BiLN+HSJA 4 to retain an $R^2 \geq 0.8$ on average through every stage.

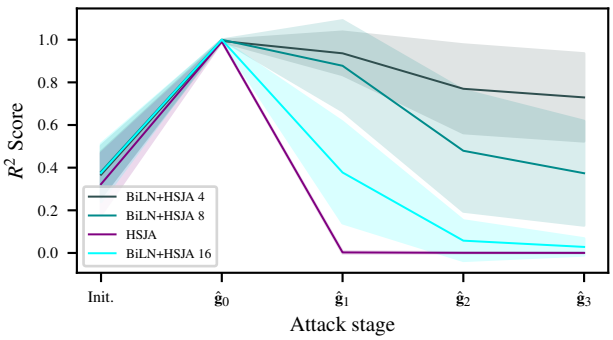


Figure 5. Comparison of the learned semantic coefficients quality (R^2 score) on CIFAR-10 at different HSJA attack stages and dimension-reduced variants, each averaged over 500 samples.

C. Estimating manifold distance

Given the ability for the hard-label adversary to model coarse semantic features of the input sample, this may indicate that the manifold-gradient Markov chain is a reasonable assumption. However, the result in Observation 1 implies that the dimension-reduced attack variants should lead to adversarial samples lying closer to the original concentration of training points (i.e., closer to the manifold). To connect Observation 1 to the findings in Figures 4 and 5, we leverage an estimate of manifold distance to determine if higher R^2 score does in fact lead to samples lying closer to the data manifold. The Learned

Table I
MARKOV CHAIN MODELING STATISTICS FOR HSJA VARIANTS.

	Attack Variant	$\bar{R}^2$	FID	SR@40k (ϵ=0.031)	LPIPS
Madry CIFAR-10	HSJA	0.259	0.244	0.272	0.676 ± 0.275
	$\rightarrow$ BiLN 16	0.363	0.074	**0.298**	0.654 ± 0.277
	$\rightarrow$ BiLN 8	0.624	0.026	0.224	0.668 ± 0.304
	$\rightarrow$ BiLN 4	**0.779**	**0.026**	0.130	0.709 ± 0.345
Natural CIFAR-10	HSJA	0.263	0.240	**1.000**	0.496 ± 0.211
	$\rightarrow$ BiLN 16	0.368	0.085	0.984	0.543 ± 0.227
	$\rightarrow$ BiLN 8	0.622	0.028	0.826	0.624 ± 0.253
	$\rightarrow$ BiLN 4	**0.759**	**0.012**	0.472	0.651 ± 0.297

Perceptual Image Patch Similarity (LPIPS) acts as such a measure for manifold distance, since it was previously shown to correlate with human visual perception [22], [31]. We use the same LPIPS code and checkpoint provided by the authors. In addition to LPIPS, we rely on the popular Fréchet Inception Distance (FID) to measure the semantic drift between sets of original samples and their adversarial counterparts [36]. Although FID lacks human studies, Heusel et al. showed that the internal representations of Inception network capture the visual semantic quality of synthetically generated images [36]. Our attack variants do not target the Inception network directly, meaning the results should not be affected by the addition of low-magnitude adversarial noise. When measuring either LPIPS or FID, we use the first successful adversarial sample which was below the noise threshold $\epsilon = 0.031$ studied by previous hard-label attacks in the l_∞ setting [18], [25]. In this way, all comparison samples for the purposes of measuring FID and LPIPS are "clipped" at the same distortion threshold, so as to not bias the computation of either score.

We measure FID and the average LPIPS of the same 500 samples generated during the Markov-chain attack probe described in Algorithm 2. As observed in Table I, the dimension-reduced variants create samples with consistently lower FID, which is strongly negatively correlated with the average whole-attack R^2 score from these same variants (denoted $\bar{R}^2$). Shown in bold, the BiLN 4 variant of HSJA achieves over ten times lower FID score on an adversarially trained model [37] (denoted Madry CIFAR-10) compared to the baseline attack. The same trend holds on the natural model without adversarial training, where the FID is instead twenty times lower. The trade-off in the natural case is the success rate; as is observed in the fourth column, success rate after forty thousand queries (`SR@40k`) is consistently lower as the dimension is reduced, but the opposite is true for the adversarially trained model. Instead, we see that the BiLN 16 variant (bold) outperforms the standard HSJA attack on adversarially trained models. From the perspective of human alignment, we see that the average LPIPS score for each attack variant is within the standard deviation of the baseline attack's LPIPS score, meaning these variants are likely not discernible by the human eye. We summarize our findings so far.

Observation 2. *Motivated by Hypothesis 1 and Observation 1, we proposed a probing algorithm to investigate if the manifold-gradient Markov chain assumption is reasonable on real-world datasets. Based on the holistic comparison in Figure 5 and Table I, we show it is possible to model the Markov chain (i.e., the noisy manifold distance oracle) from boundary traversals alone, leading to discovery of the semantic directions which correspond to manifold traversal, as illustrated by the feature importance masks of Figure 4. The lower FID in Table I for dimension-reduced variants offer further empirical support for Hypothesis 1.*

V. INFORMING PRACTICE

Our findings so far offer empirical support for Hypothesis 1 through the existence of a usable manifold-gradient Markov chain, which could serve as a passive noisy manifold distance (NMD) oracle in dimension-reduced attacks. In this sense, we are left with key research questions concerning the practical application of the increased manifold-gradient mutual information that the NMD oracle provides, beyond the contemporary interest of query efficiency. We outline our key research questions as follows.

Q1. Although dimension-reduction offers lower manifold distance, it may not be clear how many dimensions to choose from the outset of an attack. From the active learning sense, how can Algorithm 2 be modified so that an adversary *actively maximizes* the manifold-gradient mutual information (by way of automatically maximizing their whole-attack R^2 score)?

Q2. Does the phenomenon of manifold-gradient mutual information hold across different attacks, datasets, and levels of model robustness?

Q3. Is there a connection between gradient estimation attacks (Sign-OPT [19], HSJA [18]) and combinatorial search attacks (RayS [25]) in terms of the hard-label adversary's manifold traversal?

We subsequently answer each research question through our experimental design, enabling the following experimental highlights.

A1. We propose two variants of existing gradient estimation-based hard-label attacks by modifying our Markov chian probe described in Algorithm 2: (1) Markov chain (MC) variant and (2) Dynamic R^2-based upsampling attack variant (denoted DynBiLN), based on the MC variant, which can automatically upsample the attack search dimension based on the quality of the learned feature importance W (described fully in Algorithm 3).

A2. By conducting experiments over three state-of-the-art attacks, two large-scale image datasets, and both natural and adversarially trained models, we show a consistent trend for dimension-reduced attacks to achieve lower FID, while remaining visually imperceptible to humans according to LPIPS standard deviation.

A3. Through dimension-reduction techniques and attack variants proposed in A1, we show it is possible for theoretically-grounded gradient estimation attacks such

as Sign-OPT and HSJA to achieve up to 39x lower FID than the state-of-the-art combinatorial search-based RayS attack.

Experiment setup. To provide widespread empirical evidence for the NMD oracle and address the research questions, we compare between three state-of-the-art attacks which adopt unique approaches to the hard-label attack problem.

1) Sign-OPT [19] is a variant of random gradient-free (RGF) method [24] and enjoys both query efficiency (due to its sample update using the sign of the gradient estimate) and convergence guarantees grounded in the theory of zeroth-order optimization [38].
2) HSJA [18] is another variant of RGF, which like Sign-OPT, only uses the sign of the gradient estimate to update the attack sample. HSJA enjoys a bounded estimation error and design optimized for l_∞ setting, which enables better query efficiency than Sign-OPT, but lacks the respective convergence guarantees.
3) RayS [25] does not employ gradient estimation, and instead performs a combinatorial ray search in image space using progressively larger search dimensions. RayS can create semantically similar adversarial samples in the least amount of queries, but lacks the theoretical grounding of Sign-OPT and HSJA.

We perform experiments with the above attacks (and our variants) using CIFAR-10 [39] and ImageNet [40] as input image data. The natural CIFAR-10 network is the same implementation open-sourced by Cheng et al. [19]. The architecture (and accompanying pre-trained weights) for natural ImageNet are taken from the ResNet-50 network implementation in the PyTorch Torchvision library.[2] In addition, we leverage the representative adversarial training technique proposed by Madry et al. [37] (and their $\epsilon = \frac{8}{255} = 0.031$ checkpoints for l_∞ setting) as the robust models for CIFAR-10 and ImageNet. In all experiments, our BiLN variants rescale the attack search directions to the dimension shown next to their denomination (e.g., `BiLN 4` scales attack spatial dimensions for $\boldsymbol{\theta}'$ to 4×4). We use l_∞-norm versions of attacks for all experiments, and the same $\epsilon = 0.031$ distortion threshold to measure success for both natural and robust models (hereafter referred to as Madry CIFAR-10 and Madry ImageNet). All attacks run for 25k queries without early stopping on correctly classified samples. For brevity, we only show results for the untargeted case. Additional implementation details, such as hyperparameters and hardware used, can be found in Appendix C. Code for experiments will be released publicly to encourage reproducibility.

A. Actively exploiting the NMD oracle (Q1)

Motivated by the results in Section IV, we propose an attack algorithm which can leverage the learned feature importance coefficients W from the LIME Ridge regression model to inform the initial attack search directions. This leads to our Markov chain (MC) variant, which uses the learned W from trials of attack initialization to mask the search directions for a "final" attack initialization step (i.e., only perform Line 7 in Algorithm 3). This MC variant acts as a baseline for the case where W informs only the attack sample initialization. Further motivated by the negative correlation between FID and R^2 score in Table I, we propose a dynamic attack variant which in addition to using W for attack initialization, also maximizes the average whole-attack R^2 score. Our proposed dynamic R^2-based upsampling attack variant, denoted DynBiLN and described in Algorithm 3, automatically upsamples the attack search dimension d' based on the quality of the learned feature importance coefficients. This is achieved by recording the best R^2 score at the current reduced attack dimension d' (starting on Line 5, denoted $R^2_{d'}$) and updating each time the upsampling occurs (Line 13). Upsampling is performed when the current linear model R^2 score drops below a small differential factor η of $R^2_{d'}$ (Lines 12-14). In this way, we can gain the query efficiency of low search dimension, without incurring the damaging estimation bias during later stages (which lead to low SR AUC in Table I). To avoid the query overhead of Bayesian optimization for kernel width at every attack iteration, the adversary re-uses the kernel width obtained from initialization (Line 2), as in practice, this choice offers reasonable linear model quality for subsequent iterations. Since DynBiLN does not make assumptions about the underlying intitialization or gradient approximation routines, it can be adapted to any gradient estimate-based hard-label attack that uses sample initialization [15], [18], [19]. Although

Algorithm 3: Dynamic R^2-based upsampling attack variant (DynBiLN)

Input: Benign sample $\mathbf{x}_0$, original attack dimension d, intial reduced attack dimension $d' << d$, DynBiLN differential coefficient η
Output: Hard-label attack sample $\mathbf{x}$

1 $\texttt{GP} := (\texttt{init}, \mathbf{x}_0)$
2 kernel width $k \leftarrow \texttt{OptiLIME}(\texttt{MC\_step}, \texttt{GP})$
3 $W, R^2, \mathbf{x} \leftarrow \texttt{MC\_step}(\texttt{GP}, k)$
4 /* Set best score at current d' */
5 $R^2_{d'} \leftarrow R^2$
6 /* Use W to mask initialization */
7 $\mathbf{x} \leftarrow \texttt{init}(\mathbf{x}_0, W)$
8 **for** $i := 1$ *to* n **do** *Hard-label attack loop*
9 $\quad \texttt{GP} \leftarrow (\texttt{approximate\_gradient}, \mathbf{x})$
10 $\quad$ /* Re-use initial kernel width */
11 $\quad W, R^2, \boldsymbol{\theta}' \leftarrow \texttt{MC\_step}(\texttt{GP}, k)$
12 $\quad$ **if** $R^2_{d'} - R^2 > \eta\, R^2_{d'}$ **then**
13 $\quad\quad d' \leftarrow \texttt{min}(2d', d)$
14 $\quad\quad R^2_{d'} \leftarrow R^2$
15 $\quad$ **end**
16 $\quad$ Update $\mathbf{x}$ from $\boldsymbol{\theta}'$ using attack formulation
17 **end**
18 **return** $\mathbf{x}$

[2]https://pytorch.org/docs/stable/torchvision/models.html

Table II
NATURAL CIFAR-10 COMPARISON OF MANIFOLD DISTANCES AND SUCCESS RATE FOR OUR ATTACK VARIANTS (ITALICIZED).

Attack Variant	FID	SR AUC (ϵ=0.031)	LPIPS
HSJA	0.238	0.967	0.506±0.212
→ *BiLN 4*	0.016 ↓	0.485	0.642±0.297
→ *BiLN 8*	0.028 ↓	0.804	0.614±0.263
→ *BiLN 16*	0.088 ↓	0.950	0.538±0.225
MC HSJA	0.192 ↓	0.967	0.489±0.226
→ *BiLN 4*	0.018 ↓	0.521	0.623±0.303
→ *BiLN 8*	0.028 ↓	0.808	0.610±0.286
→ *BiLN 16*	0.071 ↓	0.948	0.521±0.234
→ *DynBiLN*	0.030 ↓	0.907	0.581±0.250
Sign-OPT	0.023	0.507	0.212±0.089
→ *BiLN 4*	0.252	0.132	0.512±0.830
→ *BiLN 8*	0.076	0.171	0.288±0.504
→ *BiLN 16*	**0.001** ↓	0.230	0.153±0.065
MC Sign-OPT	0.248	0.469	0.224±0.223
→ *BiLN 4*	0.051	0.122	0.348±0.625
→ *BiLN 8*	0.032	0.155	0.270±0.453
→ *BiLN 16*	0.020 ↓	0.206	0.188±0.300
RayS	0.039	**1.000**	0.706±0.269

Table III
NATURAL IMAGENET COMPARISON OF MANIFOLD DISTANCES AND SUCCESS RATE FOR OUR ATTACK VARIANTS (ITALICIZED).

Attack Variant	FID	SR AUC (ϵ=0.031)	LPIPS
HSJA	1.989	0.832	0.513±0.223
→ *BiLN 16*	0.411 ↓	0.890 ↑	0.765±0.275
→ *BiLN 32*	1.176 ↓	0.950 ↑	0.816±0.274
→ *BiLN 64*	2.871	0.947 ↑	0.795±0.271
MC HSJA	1.813 ↓	0.839 ↑	0.511±0.218
→ *BiLN 16*	0.390 ↓	0.858 ↑	0.756±0.269
→ *BiLN 32*	1.101 ↓	0.948 ↑	0.795±0.279
→ *BiLN 64*	2.502	0.931 ↑	0.774±0.247
→ *DynBiLN*	0.716 ↓	0.905 ↑	0.772±0.276
Sign-OPT	0.087	0.142	0.148±0.109
→ *BiLN 8*	0.002 ↓	0.108	0.123±0.072
→ *BiLN 16*	0.002 ↓	0.133	0.179±0.234
→ *BiLN 32*	0.001 ↓	0.143 ↑	0.147±0.075
MC Sign-OPT	0.220	0.127	0.179±0.220
→ *BiLN 8*	**0.001** ↓	0.088	0.114±0.056
→ *BiLN 16*	0.001 ↓	0.127	0.160±0.237
→ *BiLN 32*	0.005 ↓	0.135	0.179±0.244
RayS	0.445	**1.000**	0.851±0.296

DynBiLN depends on the differential parameter η, we found through grid search experiments any choice of $\eta \in [0.2, 0.5]$ provides adequate scaling performance for both CIFAR-10 and ImageNet. For the sake of comparison, we choose $\eta = 0.3$ for CIFAR-10 and $\eta = 0.5$ for ImageNet. The full grid search for η on CIFAR-10 is made available in Appendix D.

We demonstrate the effectiveness of our MC and DynBiLN variants by implementing MC HSJA, MC Sign-OPT, and DynBiLN HSJA variants to compare against baseline versions. The average success rate across 200 samples is plotted against 25k queries in Figure 6 (and the corresponding distortion plots with error regions are made available in Appendix D). The MC HSJA (dashed blue lines) and MC Sign-OPT (dashed green lines) are able to closely match the SR curve of their baseline versions, but can subsequently lower the FID as shown in MC HSJA and MC Sign-OPT rows of Tables II-V. The same is true for DynBiLN HSJA (cyan lines), which can automatically strike a balance between the query efficiency of down-sampling attacks (orange lines) and simplicity of the baseline HSJA attack (blue lines), meanwhile achieving lower FID compared to HSJA. In fact, on large scale data such as ImageNet (Figure 6b), DynBiLN enables a gradient estimate-based attack, such as HSJA, to gain enough query efficiency to be competitive against RayS (red lines), the current state-of-the-art which does not use gradient estimation (and likewise cannot enjoy the theoretical guarantees from zeroth-order optimization).

B. NMD oracle across attacks, datasets, and robustness (Q2)

We summarize results on key attack implementations in Figure 6, and provide quantitative analysis in Tables II-V through FID, LPIPS, and normalized area-under-curve (AUC) of the average success rate curves for all attacks across 25k queries and 200 samples. On natural models (Tables II and III), we

Table IV
MADRY CIFAR-10 COMPARISON OF MANIFOLD DISTANCES AND SUCCESS RATE FOR OUR ATTACK VARIANTS (ITALICIZED).

Attack Variant	FID	SR AUC (ϵ=0.031)	LPIPS
HSJA	0.253	0.537	0.683±0.284
→ *BiLN 4*	0.026 ↓	0.342	0.700±0.351
→ *BiLN 8*	0.023 ↓	0.574 ↑	0.646±0.306
→ *BiLN 16*	0.074 ↓	0.720 ↑	0.623±0.266
MC HSJA	0.213 ↓	0.545 ↑	0.645±0.244
→ *BiLN 4*	**0.022** ↓	0.356	0.695±0.316
→ *BiLN 8*	0.026 ↓	0.577 ↑	0.616±0.240
→ *BiLN 16*	0.068 ↓	0.705 ↑	0.636±0.239
→ *DynBiLN*	0.030 ↓	0.607 ↑	0.651±0.256
RayS	0.057	**1.000**	0.827±0.310

observe the consistent trend for BiLN HSJA variants to reduce the FID compared to baseline (downward green arrows). This trend is only present for Sign-OPT on large-scale ImageNet data, although in both cases of CIFAR-10 and ImageNet, BiLN Sign-OPT variants enjoy the lowest FID score across all attacks (FID of 0.001 for Sign-OPT BiLN 16 and MC Sign-OPT BiLN 8, respectively). Although RayS retains the best query efficiency (SR AUC of 1.0), BiLN variants have an inverse effect on the query efficiency depending on original data dimension. On natural CIFAR-10 (Table II), SR AUC is lowered alongside BiLN dimension, whereas for natural ImageNet (Table III), SR AUC is consistently higher for lower BiLN dimension on HSJA (upward green arrows), and similar or higher on Sign-OPT. In both cases, LPIPS of attack variants remains within the margin of error for their baseline versions.

To summarize, the effectiveness of dimension reduction may depend on the attack difficulty. We can track this idea by investigating the HSJA attack performance on adversarially-trained models (Tables IV and V). We omit Sign-OPT due

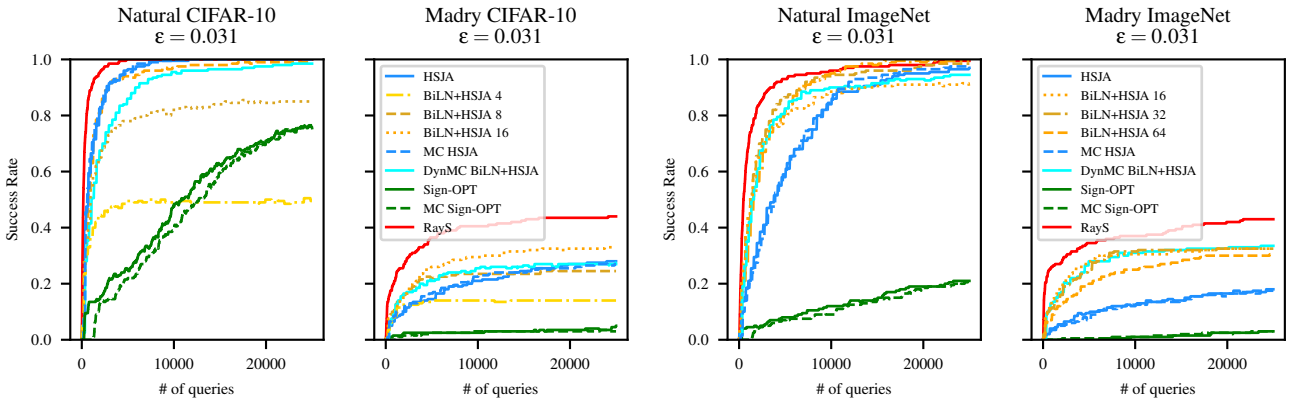


Figure 6. Success rates across key attack implementations over 200 samples on CIFAR-10 (a) and ImageNet (b). Corresponding distortion plots are available in Appendix D.

Table V
MADRY IMAGENET COMPARISON OF MANIFOLD DISTANCES AND SUCCESS RATE FOR OUR ATTACK VARIANTS (ITALICIZED).

Attack Variant	FID	SR AUC (ϵ=0.031)	LPIPS
HSJA	1.541	0.344	0.480±0.211
→ *BiLN 16*	0.312 ↓	0.777 ↑	0.696±0.288
→ *BiLN 32*	1.085 ↓	0.771 ↑	0.785±0.317
→ *BiLN 64*	2.567	0.655 ↑	0.777±0.350
MC HSJA	1.591	0.331	0.492±0.202
→ *BiLN 16*	**0.271** ↓	0.772 ↑	0.654±0.269
→ *BiLN 32*	1.079 ↓	0.771 ↑	0.797±0.293
→ *BiLN 64*	2.287	0.615 ↑	0.753±0.316
→ *DynBiLN*	0.657 ↓	0.774 ↑	0.725±0.293
RayS	0.302	**1.000**	0.676±0.301

to achieving insufficient samples within the adversarial radius $\epsilon = 0.031$ for FID and LPIPS calculation. As in the case of Natural ImageNet, regardless of dataset or baseline attack choice, dimension-reduced variants showcase the consistent trend of lower FID (downward green arrows) and increased SR AUC (upward green arrows). We can likewise observe that the DynBiLN HSJA variant from Q1 offers a balance between low FID and higher SR AUC, evidenced by the performance on Madry ImageNet (doubling SR AUC and halving FID compared to HSJA in Table V).

C. Connection to combinatorial search attacks (Q3)

Dimension-reduced attacks have so far exhibited the ability to achieve lower FID compared to their baseline versions, in some cases reducing FID by 10x on HSJA (e.g., MC HSJA+BiLN 4 in Table V) and 87x on Sign-OPT (MC Sign-OPT+BiLN 8 in Table III). However, there is a persistent need for query efficiency in the design of hard-label attacks. As evidenced in Tables II-V, RayS achieves the best query efficiency, but is often out-performed on FID compared to dimension-reduced variants of gradient estimation attacks (and in some cases, regular Sign-OPT as well). In this sense, our results show that the ability to generate adversarial samples closer to the original model observations (i.e., the data manifold) stem from the ability to abstract away unnecessary dimensions in the attack search space, rather than the specific attack formulation or design (e.g., using gradient estimation or combinatorial search techniques). RayS upsamples the attack search space dynamically, thereby eliminating unnecessary search directions early on, similar to our DynBiLN variant, but certain BiLN variants show there is untapped potential in simply searching a single attack dimensionality over the entire attack. In the case of Madry CIFAR-10 (Table IV), the MC HSJA+BiLN 4 variant achieves *less than half* the FID that was possible using RayS. Dynamic subsampling is not unwarranted, since in comparison, DynBiLN on the same model achieves a similar FID (0.030) with double the SR AUC (0.607). Our empirical results highlight a limitation of Observation 1; although higher manifold-gradient mutual information may lower the manifold distance, the search directions near the manifold yield a lower concentration of adversarial samples on average. This presents an interesting trade-off which builds on previous work investigating the data geometric properties of deep learning models. Stutz et al. [2] used autoencoders to synthesize on-manifold examples which are far from a sample's nearest decision boundary (giving the lowest success rate), while Chen et al. [25] showcased the ability for RayS to find the nearest decision boundary (giving the highest success rate). In this sense, we find that on-manifold generalization errors are more common than originally captured by Stutz et al. [2]; some can be found near the decision boundary, showcased by the low FID of RayS, but they also become more common as we drift away from the boundary, as we obtain the lowest FID using the lowest-dimension BiLN attacks, at the cost of lower SR AUC.

VI. DISCUSSION

Through our theoretical results in Section III, we evidenced the ability for manifold-gradient mutual information to increase with lower data dimension. This motivated empirical

experiments on real-world data to support our Markov chain assumption. By borrowing techniques from the interpretable ML literature, we showed the gradient estimate distribution can be leveraged to model the Markov chain, by way of learning the feature importance coefficients in image space that correspond to manifold traversal. By reducing the attack dimensionality, we showed that an attacker increases the quality of these coefficients (represented by R^2 score), and counter-intuitively, better inform the possible search directions during the course of an attack. As a result, we propose a novel attack-agnostic variant (Algorithm 3) which can initialize attack samples in the most semantically plausible direction. By tracking the decrease of R^2 during each attack iteration, an attacker can also automatically increase the attack search space dimensionality to ensure better search resolution in later attack stages. In the end, the reduced-dimension attacker can achieve lower FID than was previously possible with RayS, and an attack variant based on our proposed DynBiLN can automatically select the ideal search dimension with minimal tuning.

Geometric interpretation of hard-label attacks. Based on our results, we can view zeroth-order attacks as following a geometric hierarchy that reveals the inherent concentration of adversarial samples. Our interpretation is illustrated in Figure 1. Each technique offers a traversal direction which is either away from the manifold (towards $\mathbf{x}_a$), arbitrarily near the manifold ($\mathbf{x}_b$), or along the manifold ($\mathbf{x}_c$). Efficient attacks create samples such as $\mathbf{x}_b$, by finding the direction of the nearest decision boundary which also points to the data manifold. This is representative of RayS, which can find adversarial samples closer to the manifold through elimination of off-manifold directions. The smaller points below $\mathbf{x}_b$ are representative of BiLN and DynBiLN variants, which can get closer to the manifold, despite this region having a lower concentration of adversarial samples. In contrast, traversing close to an approximate manifold description leads to $\mathbf{x}_c$, which may not find adversarial samples inside the ϵ radius. As shown by Stutz et al. [2], strict manifold traversal leads to a lower concentration of adversarial samples and high average distortion. To this end, the R^2 score of local semantic features can inform the geometric behavior of hard-label attacks which resemble the smaller yellow points near $\mathbf{x}_b$, rather than $\mathbf{x}_a$ or $\mathbf{x}_c$. This ultimately enables a better evaluation of model robustness in future work, as we are able to find errors which are closer to the manifold.

Implications for robust models. As discussed in Section V-C, our results highlight specific nuances for designing dimension-reduced hard-label attacks against robust models. When measuring FID as a proxy of manifold distance in Tables II-V, we observed that dimensionality-reduction favored "harder" learning problems such as Madry CIFAR-10 and ImageNet. We know from empirical results by Santurkar et al. [30] that gradients of adversarially trained models contain better semantic alignment with the original image. From the gradient estimation perspective, the gradient estimates could leverage the upper-bounded noisy mutual information during attacks (Hypothesis 1). This would manifest in a lower gradient deviation, or in other words, a lower distance between the true gradient from a robust model and the gradient estimate at the first attack step. Since adversarial training effectively smooths the *sampled* data manifold by augmenting perturbed data samples during training, robust models have a well-defined boundary that aligns with salient input changes [30].

Table VI
PER-PIXEL GRADIENT l_2-DEVIATION MEASUREMENT FOR OUR ATTACK VARIATIONS (ITALICIZED).

Attack Variant	Natural CIFAR-10	Madry CIFAR-10
HSJA	5.42 ± 0.02	5.46 ± 0.06
→ *BiLN 4*	$3.72 \pm 0.35\downarrow$	$3.72 \pm 0.39\downarrow$
→ *BiLN 8*	$3.83 \pm 0.22\downarrow$	$3.84 \pm 0.27\downarrow$
→ *BiLN 16*	$3.83 \pm 0.15\downarrow$	$3.85 \pm 0.21\downarrow$
MC HSJA	$5.42 \pm 0.02\downarrow$	$5.47 \pm 0.05\uparrow$
→ *BiLN 4*	$3.74 \pm 0.33\downarrow$	$3.76 \pm 0.34\downarrow$
→ *BiLN 8*	$3.82 \pm 0.20\downarrow$	$3.89 \pm 0.22\downarrow$
→ *BiLN 16*	$3.83 \pm 0.12\downarrow$	$3.91 \pm 0.16\downarrow$
→ *DynBiLN*	$3.96 \pm 0.33\downarrow$	$3.96 \pm 0.25\downarrow$
Sign-OPT	0.12 ± 0.42	0.84 ± 1.16
→ *BiLN 4*	$0.16 \pm 0.29\uparrow$	$0.41 \pm 0.42\downarrow$
→ *BiLN 8*	$0.15 \pm 0.30\uparrow$	$0.73 \pm 0.44\downarrow$
→ *BiLN 16*	$0.16 \pm 0.24\uparrow$	$0.81 \pm 0.43\downarrow$
MC Sign-OPT	$0.13 \pm 0.32\uparrow$	$0.72 \pm 0.51\downarrow$
→ *BiLN 4*	$0.15 \pm 0.23\uparrow$	$0.45 \pm 0.44\downarrow$
→ *BiLN 8*	$0.18 \pm 0.32\uparrow$	$0.73 \pm 0.40\downarrow$
→ *BiLN 16*	$0.17 \pm 0.30\uparrow$	$0.82 \pm 0.41\downarrow$

We posit that adversarial trained models lower the variance of the gradient estimate compared to natural models. We test this idea by calculating per-pixel gradient deviation $\frac{||\mathbf{g}-\hat{\mathbf{g}}||_2}{H \times W}$ for true gradient $\mathbf{g}$ (in the direction of the adversarial label) from an adversarially trained model, first gradient estimate $\hat{\mathbf{g}}$, estimate height H, and estimate width W. When taking the true input gradient in the direction of the adversarial label, we use the robust model's original criterion to calculate the gradient, which was cross-entropy for all models in our evaluation. The results of this experiment are shown in Table VI, which exhibits a consistent trend for dimension-reduced attacks to reduce the gradient deviation, particularly for HSJA (which is known to have a higher gradient estimation error due to relying on a single point estimate [16]). In this sense, dimension-reduced attacks can turn the strength of adversarial training into a weakness, since the robust model now leaks more semantically meaningful decision boundaries. This reveals an interesting direction for future work, as it can lead to creation of hard-label attacks specifically formulated for adversarial training schemes, which can heal the inherent generalization errors due to adversarial training on narrow threat models [22]. On supplemental experiments in the Appendix, we show that our hard-label attacks are in fact capable of lower manifold distance than strong white-box attacks such as AutoAttack [41]. Although it remains outside the scope of this work, incorporating the hard-label threat model into adversarial training provides a ripe direction for future work.

VII. Conclusion

Despite the recent progress in zeroth-order attack methods, open questions remain about their precise behavior. We develop an information-theoretic analysis that sheds light on their ability to produce on-manifold adversarial examples. Through experiments on real-world datasets, we show an over two-fold increase in attack success rates by leveraging new findings about manifold distance and gradient deviation. With knowledge of the manifold-gradient relationship, it is possible to further refine hard-label attacks, and inform a better evaluation of model robustness. Given the availability of larger datasets in the future, our method may turn the strength of deep learning, which is efficiently extracting patterns in large-scale data, into a weakness.

Acknowledgements

This work was funded by the AFOSR Center of Excellence on Assured Autonomy (AFOSR FA8650-19-1-0169). We thank the numerous anonymous conference reviewers who have positively shaped this work through their feedback and suggestions.

References

[1] T. Tanay and L. Griffin, "A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples," *arXiv:1608.07690 [cs, stat]*, Aug. 2016, arXiv: 1608.07690. [Online]. Available: http://arxiv.org/abs/1608.07690

[2] D. Stutz, M. Hein, and B. Schiele, "Disentangling Adversarial Robustness and Generalization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2019, p. 12.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," pp. 1–10, 2013, arXiv: 1312.6199 ISBN: 1549-9618. [Online]. Available: http://arxiv.org/abs/1312.6199

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," 2014, arXiv: 1412.6572 ISBN: 1412.6572. [Online]. Available: http://arxiv.org/abs/1412.6572

[5] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016*, pp. 372–387, 2016, arXiv: 1511.07528 ISBN: 9781509017515. [Online]. Available: http://arxiv.org/abs/1511.07528

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," 2015, arXiv: 1511.04599 ISBN: 9781467388511. [Online]. Available: http://arxiv.org/abs/1511.04599

[7] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Security and Privacy (SP)*, 2016, pp. 582–597, arXiv: 1608.04644 ISSN: 10816011.

[8] ——, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec '17*. Dallas, Texas, USA: ACM Press, 2017, pp. 3–14. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3128572.3140444

[9] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: elastic-net attacks to deep neural networks via adversarial examples," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[10] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15*, pp. 1322–1333, 2015, iSBN: 9781450338325. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2810103.2813677

[11] F. Tramèr, F. Zhang, F. E. Epfl, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," 2016, iSBN: 978-1-931971-32-4. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

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

[13] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box Adversarial Attacks with Limited Queries and Information," *arXiv:1804.08598 [cs, stat]*, Jul. 2018, arXiv: 1804.08598. [Online]. Available: http://arxiv.org/abs/1804.08598

[14] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models," *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec '17*, pp. 15–26, 2017, arXiv: 1708.03999. [Online]. Available: http://arxiv.org/abs/1708.03999

[15] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-Efficient Hard-label Black-box Attack:An Optimization-based Approach," *arXiv:1807.04457 [cs, stat]*, Jul. 2018, arXiv: 1807.04457. [Online]. Available: http://arxiv.org/abs/1807.04457

[16] S. Liu, P.-Y. Chen, B. Kailkhura, G. Zhang, A. Hero, and P. K. Varshney, "A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning," *arXiv:2006.06224 [cs, eess, stat]*, Jun. 2020, arXiv: 2006.06224. [Online]. Available: http://arxiv.org/abs/2006.06224

[17] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," *International Conference on Learning Representations*, 2019.

[18] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack," *arXiv:1904.02144 [cs, math, stat]*, Apr. 2019, arXiv: 1904.02144. [Online]. Available: http://arxiv.org/abs/1904.02144

[19] M. Cheng, S. Singh, P. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh, "Sign-OPT: A Query-Efficient Hard-label Adversarial Attack," *The International Conference on Learning Representations (ICLR)*, p. 16, 2020. [Online]. Available: https://openreview.net/forum?id=SklTQCNtvS

[20] R. Feng, J. Chen, N. Manohar, E. Fernandes, S. Jha, and A. Prakash, "Query-Efficient Physical Hard-Label Attacks on Deep Learning Visual Classification," *arXiv:2002.07088 [cs]*, Feb. 2020, arXiv: 2002.07088. [Online]. Available: http://arxiv.org/abs/2002.07088

[21] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, "The Relationship Between High-Dimensional Geometry and Adversarial Examples," *arXiv:1801.02774 [cs]*, Sep. 2018, arXiv: 1801.02774. [Online]. Available: http://arxiv.org/abs/1801.02774

[22] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual Adversarial Robustness: Defense Against Unseen Threat Models," *arXiv:2006.12655 [cs, stat]*, Jul. 2021, arXiv: 2006.12655. [Online]. Available: http://arxiv.org/abs/2006.12655

[23] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Learning Perceptually-Aligned Representations via Adversarial Robustness," *arXiv:1906.00945 [cs, stat]*, Jun. 2019, arXiv: 1906.00945. [Online]. Available: http://arxiv.org/abs/1906.00945

[24] Y. Nesterov and V. Spokoiny, "Random Gradient-Free Minimization of Convex Functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, Apr. 2017. [Online]. Available: http://link.springer.com/10.1007/s10208-015-9296-2

[25] J. Chen and Q. Gu, "RayS: A Ray Searching Method for Hard-label Adversarial Attack," *arXiv:2006.12792 [cs, stat]*, Jun. 2020, arXiv: 2006.12792. [Online]. Available: http://arxiv.org/abs/2006.12792

[26] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng, "AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 742–749, Jul. 2019. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/3852

[27] A. Ilyas, L. Engstrom, and A. Madry, "Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors," Mar. 2019, arXiv:1807.07978 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1807.07978

[28] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially Robust Generalization Requires More Data," *arXiv:1804.11285 [cs, stat]*, May 2018, arXiv: 1804.11285. [Online]. Available: http://arxiv.org/abs/1804.11285

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[30] S. Santurkar, D. Tsipras, B. Tran, A. Ilyas, L. Engstrom, and A. Madry, "Image Synthesis with a Single (Robust) Classifier," *arXiv:1906.09453 [cs, stat]*, Jun. 2019, arXiv: 1906.09453. [Online]. Available: http://arxiv.org/abs/1906.09453

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 586–595. [Online]. Available: https://ieeexplore.ieee.org/document/8578166/

[32] N. J. Beaudry and R. Renner, "An intuitive proof of the data processing inequality," *arXiv:1107.0740 [quant-ph]*, Sep. 2012, arXiv: 1107.0740. [Online]. Available: http://arxiv.org/abs/1107.0740

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., ser. Wiley-Interscience. John Wiley & Sons, 2006.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," 2016, arXiv: 1602.04938 ISBN: 9781450321389.

[35] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms," Feb. 2022, arXiv:2006.05714 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.05714

[36] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *arXiv:1706.08500 [cs, stat]*, Jan. 2018, arXiv: 1706.08500. [Online]. Available: http://arxiv.org/abs/1706.08500

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv:1706.06083 [cs, stat]*, Jun. 2017, arXiv: 1706.06083. [Online]. Available: http://arxiv.org/abs/1706.06083

[38] S. Liu, J. Chen, P.-Y. Chen, and A. O. Hero, "Zeroth-Order Online Alternating Direction Method of Multipliers: Convergence Analysis and Applications," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018*, vol. 84, Lanzarote, Spain, 2018, p. 10.

[39] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *University of Toronto*, p. 60, 2009.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[41] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[42] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," *arXiv:1712.04248 [cs, stat]*, Dec. 2017, arXiv: 1712.04248. [Online]. Available: http://arxiv.org/abs/1712.04248

[43] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On Evaluating Adversarial Robustness," *arXiv:1902.06705 [cs, stat]*, Feb. 2019, arXiv: 1902.06705. [Online]. Available: http://arxiv.org/abs/1902.06705

APPENDIX

A. Derivation of Manifold-Gradient Mutual Information (MI)

We define the manifold-gradient point-wise joint probability in a case-wise manner, for the respective values under $\mathbf{g} \in \{-1, 1\}^d$ and $\mathbf{x} \in \mathbb{R}^d$. We are concerned with the sub-gradient cases where $\mathbf{x} > 0$ (denoted $\mathbf{x}^+$) and $\mathbf{x} < 0$ (denoted $\mathbf{x}^-$) which correspond to fixed values of $\mathbf{g}$ based on class means $y \cdot \boldsymbol{\mu}$ with $y \in \{-1, 1\}$. This gives for each dimension k,

$$\begin{aligned} p(\mathbf{g}_k = 1, \mathbf{x}_k^+) &= \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ - \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \\ p(\mathbf{g}_k = 1, \mathbf{x}_k^-) &= \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ + \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \end{aligned} \tag{3}$$

$$\begin{aligned} p(\mathbf{g}_k = -1, \mathbf{x}_k^+) &= \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ + \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \\ p(\mathbf{g}_k = -1, \mathbf{x}_k^-) &= \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ - \boldsymbol{\mu}_k}{\sigma}\right)^2\right). \end{aligned} \tag{4}$$

Since the Schmidt et al. Gaussian mixture is created symmetrically (the probability mass is evenly split between the two classes i.e., the mixture comprises one Gaussian offset by $\boldsymbol{\mu}_k$ and mirrored at $\mathbf{x}_k = 0$) we can simplify to

$$p(\mathbf{g}_k = 1, \mathbf{x}_k^+) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ - \boldsymbol{\mu}_k}{\sigma}\right)^2\right), \tag{5}$$

$$p(\mathbf{g}_k = -1, \mathbf{x}_k^+) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ + \boldsymbol{\mu}_k}{\sigma}\right)^2\right), \tag{6}$$

where $\mathbf{x} \sim \mathcal{N}(y \cdot \boldsymbol{\mu}, \sigma I)$. In words, Equation 6 is the symmetrical tail of the Gaussian mixture marginal while Equation 5 is the remainder of the mixture.

Similarly, a point-wise gradient is given as the Rademacher outcome $\mathbf{g}_k \in \{\pm 1\}$. The choice of ϵ directly influences the marginal probability over the manifold. The marginal probability over the manifold can be given as the Riemann approximations

$$p_{\mathcal{G}}(\mathbf{g}_k = 1)_\epsilon = \frac{1}{2\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_{i,k}^* - \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \Delta_i \tag{7}$$

and

$$p_{\mathcal{G}}(\mathbf{g}_k = -1)_\epsilon = \frac{1}{2\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_{i,k}^* + \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \Delta_i, \tag{8}$$

with $\Delta_i = \mathbf{x}_{i,k}^+ - \mathbf{x}_{i-1,k}^+$ for arbitrary $\mathbf{x}_{i,k}^* \in [\mathbf{x}_{i-1,k}^+, \mathbf{x}_{i,k}^+]$, and n is controlled by the hyper-parameter ϵ.

The marginal for the manifold under the gradient is given similarly as

$$\begin{aligned} p_{\mathcal{M}}(\mathbf{x}_k) &= p(\mathbf{g}_k = 1, \mathbf{x}_k^+) + p(\mathbf{g}_k = -1, \mathbf{x}_k^+) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ - \boldsymbol{\mu}_k}{\sigma}\right)^2\right) \\ &+ \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_k^+ + \boldsymbol{\mu}_k}{\sigma}\right)^2\right), \end{aligned} \tag{9}$$

where $\mathbf{x}_k^+ > 0$ for all dimensions k. Denote the sub-manifold sampled from the positive ($y = 1$) and negative ($y = -1$) classes as $\mathcal{M}^+$ and $\mathcal{M}^-$, respectively. Our definition for manifold-gradient mutual information is based on the standard definition of mutual information from information theory [33],

$$I(\mathcal{M};\mathcal{G})_{\epsilon,k} = \int_{\mathcal{M}}\int_{\mathcal{G}} p(\mathbf{g}_k,\mathbf{x}_k)\log(\frac{p(\mathbf{g}_k,\mathbf{x}_k)}{p_{\mathcal{G}}(\mathbf{g}_k)p_{\mathcal{M}}(\mathbf{x}_k)})\,d\mathbf{g}_k\,d\mathbf{x}_k, \tag{10}$$

where ϵ is treated as a hyper-parameter controlling the value of n in $p_{\mathcal{G}}(\mathbf{g}_k)$. By substitution into Equation 10 we have

$$\begin{aligned} I(\mathcal{M};\mathcal{G})_{\epsilon,k} = &\int_{\mathcal{M}} p(1,\mathbf{x}_k)\log(\frac{p(1,\mathbf{x}_k)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_k)})\,d\mathbf{x}_k \\ &+ \int_{\mathcal{M}} p(-1,\mathbf{x}_k)\log(\frac{p(-1,\mathbf{x}_k)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_k)})\,d\mathbf{x}_k. \end{aligned} \tag{11}$$

This is split further similar to true positive, true negative, false positive, and false negative, as

$$\begin{aligned} I(\mathcal{M};\mathcal{G})_{\epsilon,k} = &\int_{\mathcal{M}^+} p(1,\mathbf{x}_k^+)\log(\frac{p(1,\mathbf{x}_k^+)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_k^+)})\,d\mathbf{x}_k^+ \\ &+ \int_{\mathcal{M}^-} p(1,\mathbf{x}_k^-)\log(\frac{p(1,\mathbf{x}_k^-)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_k^-)})\,d\mathbf{x}_k^- \\ &+ \int_{\mathcal{M}^+} p(-1,\mathbf{x}_k^+)\log(\frac{p(-1,\mathbf{x}_k^+)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_k^+)})\,d\mathbf{x}_k^+ \\ &+ \int_{\mathcal{M}^-} p(-1,\mathbf{x}_k^-)\log(\frac{p(-1,\mathbf{x}_k^-)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_k^-)})\,d\mathbf{x}_k^-, \end{aligned} \tag{12}$$

and simplified due to symmetry at 0 as

$$\begin{aligned} I(\mathcal{M};\mathcal{G})_{\epsilon,k} = &2\int_{\mathcal{M}^+} p(1,\mathbf{x}_k^+)\log(\frac{p(1,\mathbf{x}_k^+)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_k^+)})\,d\mathbf{x}_k^+ \\ &+ 2\int_{\mathcal{M}^+} p(-1,\mathbf{x}_k^+)\log(\frac{p(-1,\mathbf{x}_k^+)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_k^+)})\,d\mathbf{x}_k^+. \end{aligned} \tag{13}$$

The total un-normalized mutual information is given by the summation over dimensions $I(\mathcal{M};\mathcal{G})_\epsilon = \sum_{k=1}^d I(\mathcal{M};\mathcal{G})_{\epsilon,k}$. Notably the cases for each possible scenario under detection theory are represented. Each case is bounded by the results of [28]. By substitution from each marginal and joint probability in Equations 7, 3, and 4 respectively, we have the closed form solution for mutual information.

This leads to the Riemann approximation of Equation 2,

$$\begin{aligned} I(\mathcal{M};\mathcal{G})_{\epsilon,k} = &2\sum_{i=1}^{n} p(1,\mathbf{x}_{i,k}^*)\log(\frac{p(1,\mathbf{x}_{i,k}^*)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_{i,k}^*)})\,\Delta_i \\ &+ 2\sum_{i=1}^{n} p(-1,\mathbf{x}_{i,k}^*)\log(\frac{p(-1,\mathbf{x}_{i,k}^*)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_{i,k}^*)})\,\Delta_i. \end{aligned} \tag{14}$$

with $\Delta_i = \mathbf{x}_{i,k}^+ - \mathbf{x}_{i-1,k}^+$ for arbitrary positive $\mathbf{x}_{i,k}^* \in [\mathbf{x}_{i-1,k}^+, \mathbf{x}_{i,k}^+]$. Since $\mathbf{x}^+$ is a standard multi-variate Gaussian [33], the final mutual information is the summation over each dimension,

$$\begin{aligned} I(\mathcal{M};\mathcal{G})_{\epsilon} = &2\sum_{k=1}^{d}\sum_{i=1}^{n} p(1,\mathbf{x}_{i,k}^*)\log(\frac{p(1,\mathbf{x}_{i,k}^*)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}_{i,k}^*)})\,\Delta_i \\ &+ 2\sum_{k=1}^{d}\sum_{i=1}^{n} p(-1,\mathbf{x}_{i,k}^*)\log(\frac{p(-1,\mathbf{x}_{i,k}^*)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}_{i,k}^*)})\,\Delta_i. \end{aligned} \tag{15}$$

B. Hard-label attack formulation

Contemporary hard-label attacks are variants of random gradient-free method (RGF) [24], a gradient estimator which yields the estimate $\hat{\mathbf{g}}$ over q random directions $\{\mathbf{u}_i\}_{i=1}^q$.

OPT-Attack For benign example $\mathbf{x}_0$, true label y_0, and hard-label black-box function $f : \mathbb{R}^d \to \{1,\ldots,K\}$, [17] define the objective function $g : \mathbb{R}^d \to \mathbb{R}$ as a function of search direction $\boldsymbol{\theta}$, where the optimal solution is $g(\boldsymbol{\theta}^*)$, the minimum distance from $\mathbf{x}_0$ to the nearest adversarial example along the direction $\boldsymbol{\theta}^*$. For the untargeted attack, $g(\boldsymbol{\theta})$ is the distance to any decision boundary along direction $\boldsymbol{\theta}$, and allows for estimating the gradient as

$$\hat{\mathbf{g}} = \frac{1}{q}\sum_{i=0}^{q}\frac{g(\boldsymbol{\theta}+\beta\mathbf{u}_i) - g(\boldsymbol{\theta})}{\beta}\cdot\mathbf{u}_i, \tag{16}$$

where β is a small smoothing parameter. Notably, $g(\boldsymbol{\theta})$ is continuous even if f is a non-continuous step function.

Sign-OPT [19] later improved the query efficiency by only considering the sign of the gradient estimate,

$$\hat{\nabla} g(\boldsymbol{\theta}) \approx \hat{\mathbf{g}} := \sum_{i=1}^{q}\operatorname{sgn}\left(g(\boldsymbol{\theta}+\beta\mathbf{u}_i) - g(\boldsymbol{\theta})\right)\mathbf{u}_i. \tag{17}$$

We focus on the Sign-OPT variant, since the findings are more relevant to the current state-of-the-art.

HopSkipJumpAttack Similar to Sign-OPT, HopSkipJumpAttack (HSJA) [18] uses a zeroth-order sign oracle to improve Boundary Attack [42]. HSJA lacks the convergence analysis of Sign-OPT and relies on one-point gradient estimate. Regardless, HSJA is competitive and can excel in the L_∞ setting.

Dimension-reduced Sign-OPT & HSJA. In general, for attacks relying on the [17] formulation, the update in Equation 16 becomes

$$\hat{\mathbf{g}} = \frac{1}{q}\sum_{i=0}^{q}\frac{g(\boldsymbol{\theta}'+\beta\mathbf{u}_i') - g(\boldsymbol{\theta}')}{\beta}\cdot\mathbf{u}_i' \tag{18}$$

for the reduced-dimension Gaussian vectors $\{\mathbf{u}_i' \in \mathbb{R}^{d'}\}_{i=0}^q$ for integer $d' < d$ and direction $\boldsymbol{\theta}' \in \mathbb{R}^{d'}$. The reduced-dimension direction $\boldsymbol{\theta}'$ is initialized randomly with $\boldsymbol{\theta}' \sim \mathcal{N}(0,1)$ for the untargeted case, or for the targeted case as $\boldsymbol{\theta}' = \mathcal{E}(\mathbf{x}_t)$, where $\mathbf{x}_t$ is a test sample correctly classified as target class t by the victim model. This scheme also applies to HSJA, since HSJA performs a single-point sign estimate. As in the normal variants, $\hat{\mathbf{g}}$ is used to update $\boldsymbol{\theta}'$.

Table VII
COMPARISON OF KEY ATTACK VARIANTS FROM THE MAIN TEXT AGAINST STRONG WHITE-BOX ATTACKS FORMULATED FOR ADVERSARIAL TRAINING.

	SR	LPIPS	FID
Benign	0.132	-	-
HSJA	0.240	0.696 ± 0.272	0.262
BiLN+HSJA 8	0.220	0.735 ± 0.315	0.035
DynBiLN+HSJA	0.240	0.666 ± 0.271	0.046
RayS	0.306	0.822 ± 0.297	0.054
l_∞ AutoAttack [41]	0.560	1.096 ± 0.286	0.072
LPA [22]	0.998	0.470 ± 0.038	0.120
PPGD [22]	0.990	0.387 ± 0.084	0.092

Table VIII
ADOPTING THE NEURAL PERCEPTUAL THREAT MODEL (NPTM) [22] TO COMPARE "UNCONSTRAINED" HARD-LABEL ATTACKS AGAINST NPTM-FORMULATED ATTACKS WHICH DIRECTLY OPTIMIZE LPIPS.

NPTM (no ϵ-ball)	SR	l_∞ distance	LPIPS	FID
HSJA	0.748	0.058 ± 0.041	1.407 ± 0.588	1.715
BiLN+HSJA 8	0.886	0.061 ± 0.044	1.458 ± 0.506	0.305
DynBiLN+HSJA	0.736	0.041 ± 0.038	1.222 ± 0.475	0.397
RayS	0.832	0.041 ± 0.027	1.292 ± 0.478	0.355
LPA [22]	0.998	0.277 ± 0.155	0.470 ± 0.038	0.120
PPGD [22]	0.990	0.223 ± 0.136	0.387 ± 0.084	0.092
l_∞ AutoAttack [41]	0.560	0.013 ± 0.016	1.096 ± 0.286	0.072

C. Implementation details

1) Hardware and Attack Hyperparameters: All experiments in the main paper were performed on an internal high-performance compute cluster equipped with NVIDIA Tesla V100 Tensor Core GPUs and high-speed non-volatile flash storage. In total 16 GPUs, 1TB main system memory, and 40 Intel Xeon CPU cores were used to run experiments completely.

Depending on dataset dimension, HSJA requires tuning of parameter γ for best performance. On CIFAR-10 we used $\gamma = 10.0$. For ImageNet, it was necessary to set $\gamma \geq 1000.0$ to re-create the published results of the regular variant [18]. Due to similar performance we use $\gamma = 1000.0$ for regular and dimension-reduced variants. We note that the dimension-reduced variants like HSJA+BiLN were less sensitive to γ, performing similarly regardless of the setting.

2) Data sampling: Original samples are chosen from the test set using a technique similar to Chen et al. [18]: on CIFAR-10, twenty random samples are taken from each of the ten chosen classes (200 total samples). On the ImageNet dataset, twenty classes are uniform-randomly chosen and ten random samples taken from each (200 total samples).

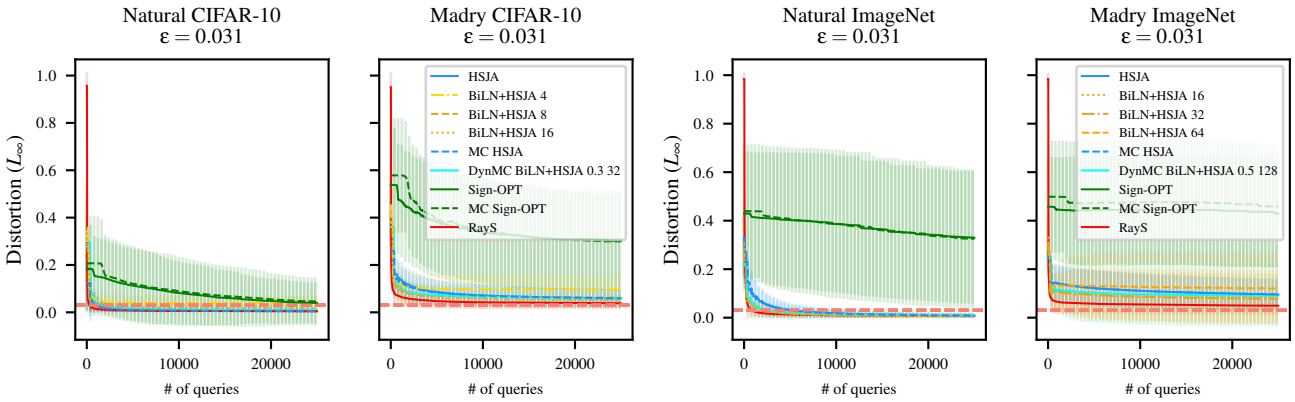
D. Supplemental Results

1) Query vs. Distortion Plots: We show the model queries against attack distortion measurement in Figure 7 to accompany the results in the main paper.

2) Grid search for DynBiLN Differential Factor: We performed a simple grid search to find an ideal differential factor η for our proposed DynBiLN attack variant. The result on CIFAR-10 is shown in Figure 8, averaged over 40 samples.

3) Comparison to strong white-box attacks: Recent work found that traditional adversarial training relies on narrow threat models which do not accurately capture the generalization error of a model, or equivalently, the distinction between off- and on-manifold examples [22]. Laidlaw et al. proposed to instead formulate LPA and PPGD within the *neural perceptual threat model* (NPTM), directly optimizing for lower LPIPS, since it acts as a reliable estimator of perceptual distance (i.e., manifold distance) from adversarial sample to original. In this way LPA and PPGD act as a best-case scenario for the creation of on-manifold adversarial examples in the gradient-level setting. We examine the relationship of these idealized gradient-level attacks to hard-label variants discussed in the main text, by attacking 500 samples with each attack on an adversarially trained model (equivalent to Madry CIFAR-10 in the main text). We show in Table VII that on successful samples within the ϵ-ball, our hard-label attack variants can achieve a similar decrease in LPIPS and FID, undercutting the narrow threat model and gradient-level baseline of l_∞ AutoAttack. Notably our attacks can generate adversarial examples closer to the manifold, without having to directly optimize LPIPS (e.g., LPA and PPGD).

The success rate gap between hard-label and gradient-level attacks in Table VII evidences the need for further improvements in hard-label attack efficiency. To this end, it is notable that LPA and PPGD generate adversarial samples outside the ϵ-ball, which would be considered invalid in our threat model. However, Laidlaw et al. demonstrate that incorporating these "invalid" samples into adversarial training leads to models that are holistically more robust against unseen adversaries compared to previous adversarial training schemes. If we relax the requirement for hard-label attacks to land inside the traditional l_∞ ϵ-ball, and instead adopt the NPTM, can hard-label attacks still find on-manifold examples? We perform this comparison in Table VIII by examining the "unconstrained" performance of the same hard-label attack variants. We provide results within the radius using AutoAttack as a comparison in the last row. Under the NPTM, we can see that hard-label success rate is much higher due to finding samples arbitrarily close to the $\epsilon = 0.031$ radius (a known behavior of adversarially trained models [7], [43]), at the cost of higher LPIPS and FID, since our hard-label variants still rely on l_∞ distance for their objectives (e.g., formulation of g for gradient estimation is optimal l_∞ step size in HSJA). LPA and PPGD, which are gradient-level attacks, have the advantage of directly optimizing LPIPS instead of l_∞, resulting in lower FID and LPIPS. It can be observed that FID is still consistently lower for dimension-reduced variants (e.g., 0.305/1.715 for BiLN+HSJA/HSJA). This motivates an exciting new research direction for hard-label attacks which formulate their objective surrogates under NPTM, rather than traditional l_p-based



(a) (b)

Figure 7. Query vs. distortion plots for a) CIFAR-10 and b) ImageNet, corresponding to the success rate plots in the main text. Horizontal dashed lines denote the value of $\epsilon = 0.031$.

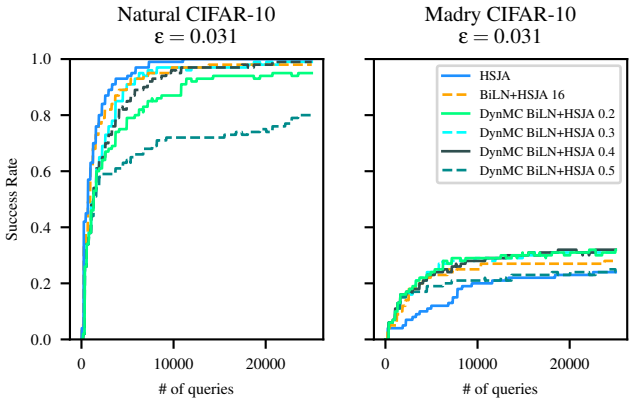


Figure 8. Result of grid search for DynBiLN differential parameter η on CIFAR-10 over 40 samples.

metrics.