Hard-label Manifolds: Unexpected advantages of query efficiency for finding on-manifold adversarial examples

Washington Garcia (UF), Kevin Butler (UF)

Scott Clouse (AFRL/ACT3)

Pin-Yu Chen (IBM), Somesh Jha (UW)

UF UNIVERSITY of FLORIDA

Duke UNIVERSITY

TEXAS The University of Texas at Austin

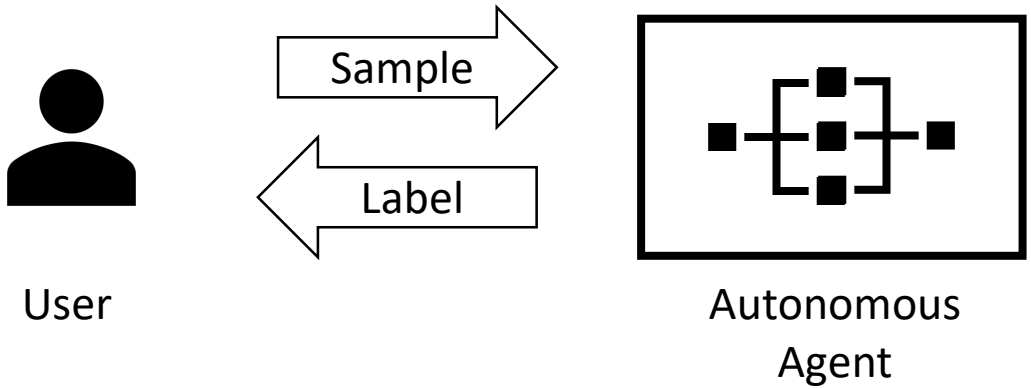
UC SANTA CRUZ

Context

Autonomous agents deployed as a service:

- Perform knowledge extraction and communicate only the extracted knowledge.

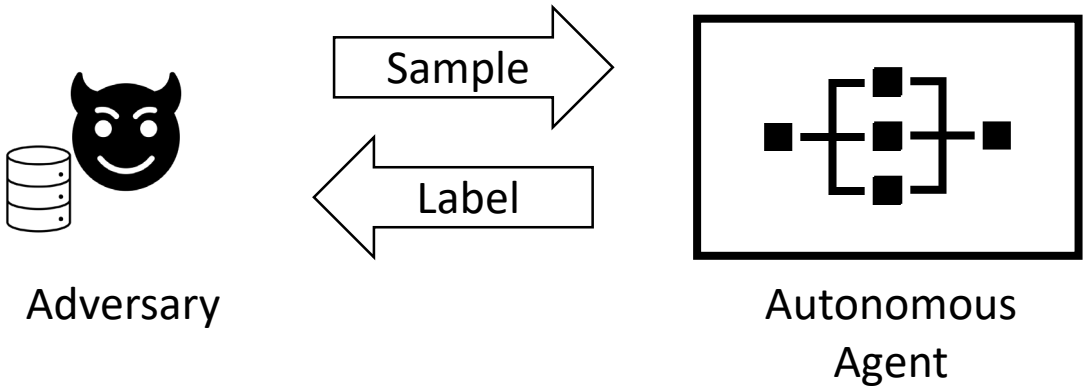
- Communication is often “hard-label” i.e., there is no side information about the system’s reasoning.



Context

Hard-label adversarial machine learning attacks are a "grand-prize":

- Adversary only needs ***query access*** to generate "label-flipped" samples (e.g., through compromised user)
- Hard-label attacks are gaining popularity, but not well characterized apart from convergence guarantees.



Background

Gradient level setting: The goal of adversarial attacks is to generate adversarial sample $\mathbf{x}$ based on original sample $\mathbf{x_o}$ such that

$$\mathbf{x} \text{ is close to } \mathbf{x}_0 \text{ and } f(\mathbf{x}) \neq f(\mathbf{x}_0).$$

Let $f(\mathbf{x}) = \text{argmax}_i(Z(\mathbf{x})_i)$ for logits $Z(\mathbf{x}) \in \mathbb{R}^K$, formally satisfy the above goal by

$$h(\mathbf{x}) := \underset{\mathbf{x}}{\text{argmin}} \left\{ ||\mathbf{x} - \mathbf{x}_0||_p + c\mathscr{L}(Z(\mathbf{x})) \right\},$$

for an adversarial loss function $\mathscr{L}(\cdot)$ and L_p-norm $||\cdot||_p$.
A popular choice is the Carlini-Wagner objective

$$\mathscr{L}(Z(\mathbf{x})) = \max \left\{ \left[Z(\mathbf{x})\right]_{y_0} - \max_{i \neq y_0} \left[Z(\mathbf{x})\right]_i, -\kappa \right\},$$

where y_o is the original label and κ is a parameter for controlling "confidence" of the adversarial sample.

Background

Gradient level setting: The goal of adversarial attacks is to generate adversarial sample $\mathbf{x}$ based on original sample $\mathbf{x_o}$ such that

$$\mathbf{x} \text{ is close to } \mathbf{x}_0 \text{ and } f(\mathbf{x}) \neq f(\mathbf{x}_0).$$

Let $f(\mathbf{x}) = \text{argmax}_i(Z(\mathbf{x})_i)$ for logits $Z(\mathbf{x}) \in \mathbb{R}^K$, formally satisfy the above goal by

$$h(\mathbf{x}) := \underset{\mathbf{x}}{\text{argmin}} \left\{ ||\mathbf{x} - \mathbf{x}_0||_p + c\mathscr{L}(Z(\mathbf{x})) \right\},$$

for an adversarial loss function $\mathscr{L}(\cdot)$ and L_p-norm $||\cdot||_p$.
A popular choice is the Carlini-Wagner objective

$$\mathscr{L}(Z(\mathbf{x})) = \max \left\{ [Z(\mathbf{x})]_{y_0} - \max_{i \neq y_0} [Z(\mathbf{x})]_i, -\kappa \right\},$$

where y_o is the original label and κ is a parameter for controlling "confidence" of the adversarial sample.

Background

Hard-label setting: We get no logits information, only the *step-function* informing us of the label (i.e., no access to $Z(\mathbf{x})$).

Instead, reformulate attack goal as a (continuous) function of **distance to the model decision boundary** $g(\cdot)$, along a search direction **θ,**

$$g(\boldsymbol{\theta}) = \text{argmin}_{\lambda>0} \left(f(\mathbf{x}_0 + \lambda \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||}) \neq y_0 \right),$$

$$\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) .$$

The optimal solution is of the form

$$\mathbf{x}^* = \mathbf{x}_0 + g(\boldsymbol{\theta}^*) \frac{\boldsymbol{\theta}^*}{||\boldsymbol{\theta}^*||} . \qquad \text{(Cheng et al. 2019)}$$

Background

Caveat: The gradient $\nabla g(\boldsymbol{\theta})$ cannot be solved directly, we can only use evaluations of $g(\cdot)$.

Instead use zeroth-order optimization (ZOO) to estimate the gradient, a common choice is Randomized Gradient-Free (RGF) method over $\boldsymbol{q}$ random directions,

$$\hat{\mathbf{g}} = \frac{1}{q}\sum_{i=0}^{q} \frac{g(\boldsymbol{\theta} + \beta \mathbf{u}_i) - g(\boldsymbol{\theta})}{\beta} \cdot \mathbf{u}_i \,. \qquad \text{(Cheng et al. 2019)}$$

where $\mathbf{u}$ is a Gaussian vector and $\beta > 0$ is a small smoothing parameter.

Background

Many recent advances in hard-label attacks:

- Originated as random walk on decision boundary (Brendel et al. 2017)

- First convergence guarantees using boundary-distance zeroth-order formulation (Cheng et al. 2019, **shown in previous slides**)

- Subsample search with sign estimate (Cheng et al. 2020): Sign-OPT
- Subsample search with single-point estimate (Chen et al. 2019): HSJA
- Subsample search with super-pixel grouping (Chen & Gu 2020): RayS

Query efficiency is gained through search subsampling. Yet, ***this reduces search fidelity***.

Central Problem

Query efficiency is achieved by searching over a dimension-reduced (reduced fidelity) version of the original.

Despite this:

- search efficiency increases
- distortion is lowered

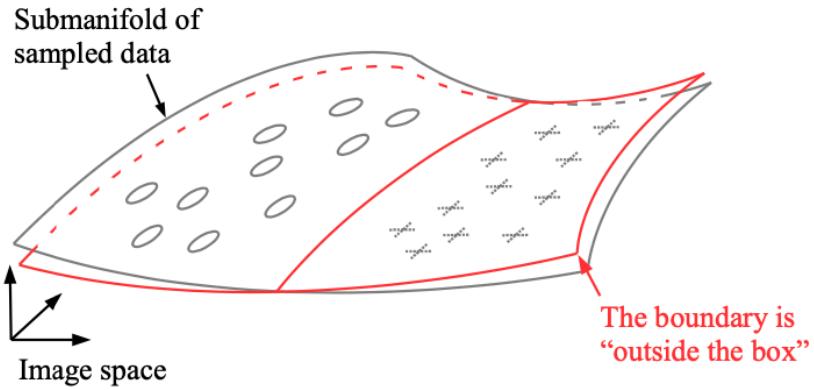
We want to understand this behavior in a controlled manner, thus

- Leverage geometric perspective of adversarial samples for context
- introduce dimension-reduced variant of each attack to measure effect of search fidelity

Approach

Build a perspective of hard-label attacks using recent developments:

- The boundary tilting perspective: ”Data-geometric” view of adversarial sample behavior (Tanay & Griffin 2016)

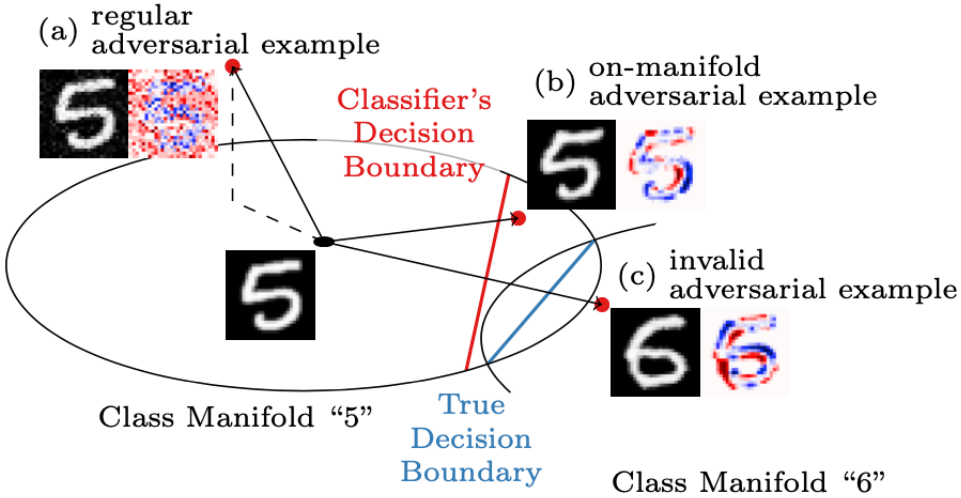


(Tanay & Griffin 2016)

Approach

Build a perspective of hard-label attacks using recent developments:

- Regular adversarial examples leave the data manifold, on-manifold adversarial examples are generalization errors (Stutz et al. 2019)



(Stutz et al. 2019)

Approach

Investigate search fidelity in a **controlled** way, by considering the dimension-reduced search variants,

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

Notably we have **θ'** the dimension-reduced **θ**, and **u'** the dimension-reduced **u**.

How to get θ'?

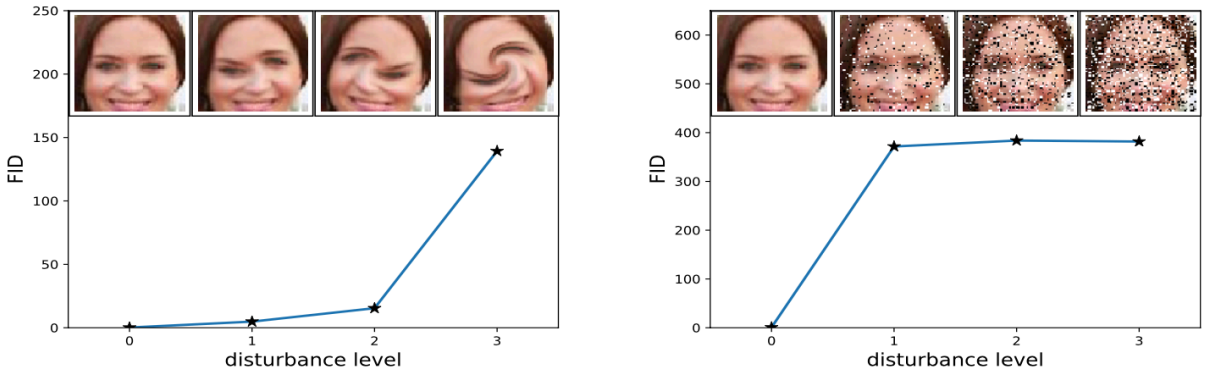
Use either a bilinear resizing function (BiLN) or trained encoder-decoder functions (i.e., auto-encoder (AE)).

Now, how to measure zeroth-order search deviation from reduced fidelity?

Approach

Instead, measure search deviation as a function of distance to manifold

- Leverage Fréchet Inception Distance (FID) proposed by Heusel et al. (2018).
- FID leverages high level coding level of Inception-V3
- Thus FID will correlate with distortion of high-level features, and act as a surrogate for distance to manifold.



(Heusel et al. 2018)

Setup

Compare attack behavior under different subsampling scenarios:

- 3 hard-label attacks, up to 4 variants each (3 BiLN, AE)
  - Sign-OPT, HSJA, RayS
  - 100 adversarial samples each
- Natural and ε-robust models for two image classification datasets:
  - CIFAR-10: Madry et al. (2018) L_∞ adversarial training **(shown today)**
  - ImageNet: Cohen et al. (2020) L_2 randomized smoothing (behavior was similar to CIFAR-10)

Approx. 2 weeks of cluster compute time

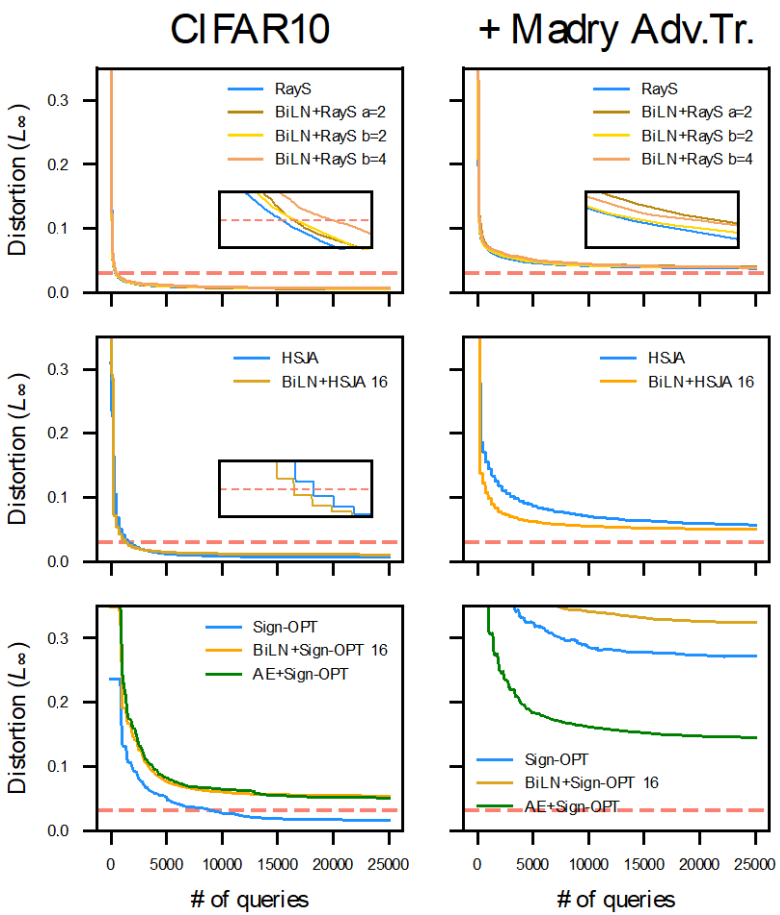
Primarily comparing:

- Distortion vs. number of queries
  - Measures attack efficiency
- FID score
  - Measures approximate distance to the data manifold

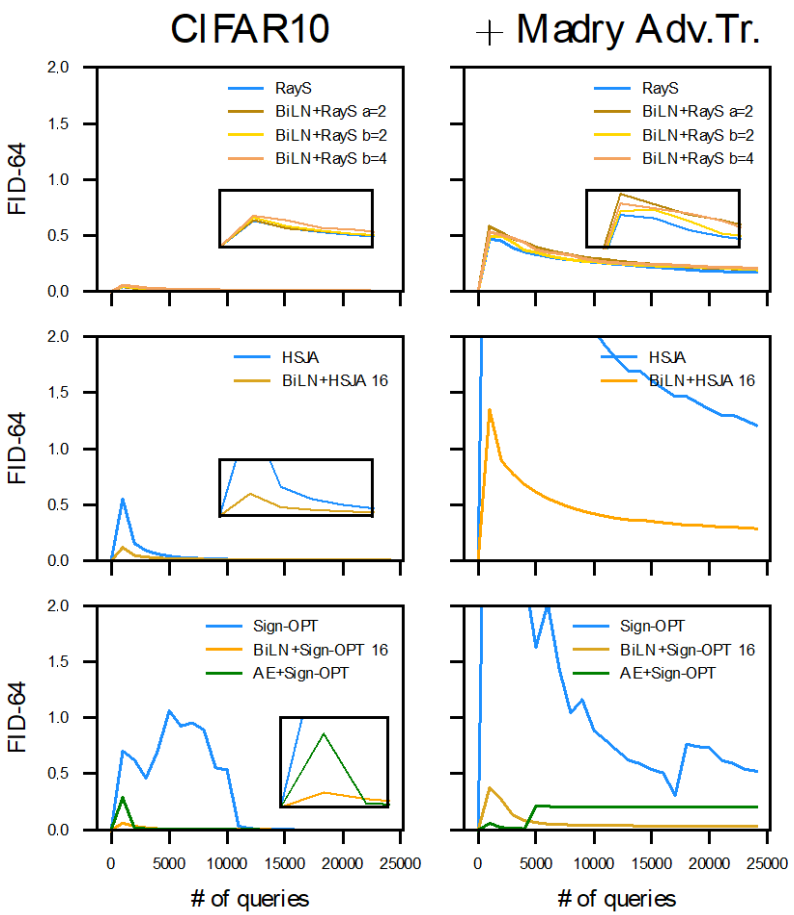
Results

CIFAR-10 (L_∞)

Average distortion by query count



FID (approx. distance to manifold)

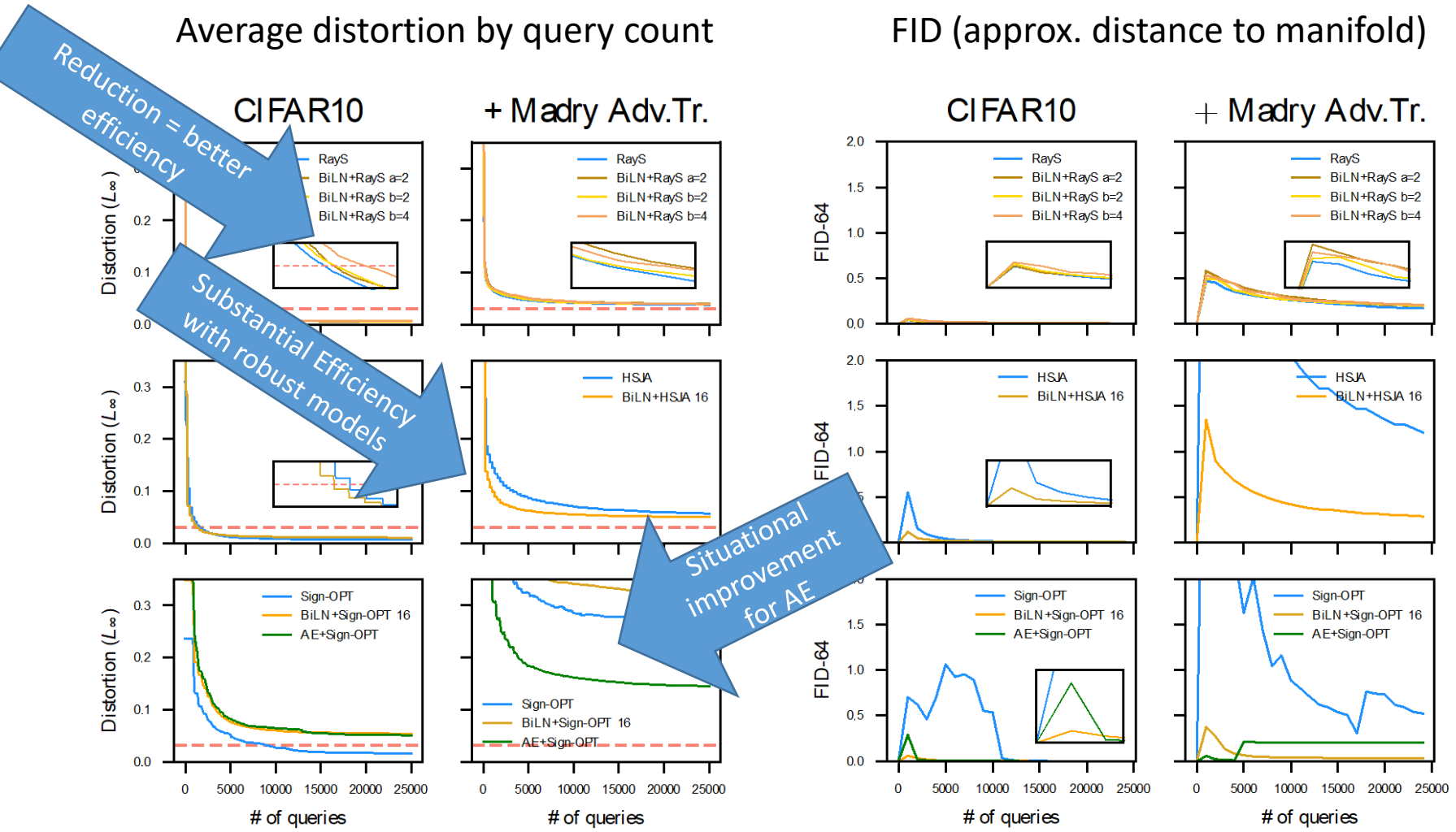


Results

CIFAR-10 (L_∞)

Average distortion by query count

FID (approx. distance to manifold)

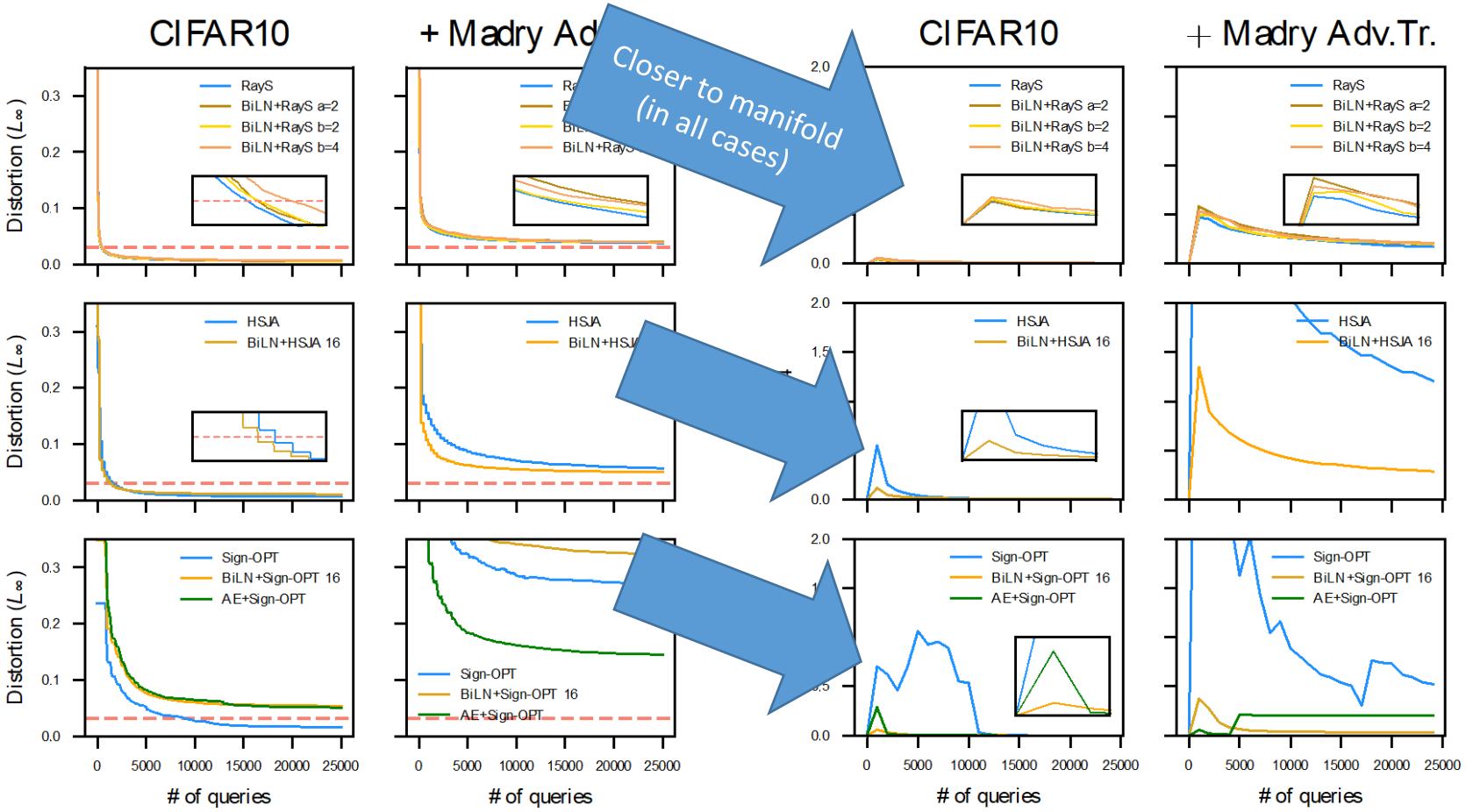


Results

CIFAR-10 (L_∞)

Average distortion by query count

FID (approx. distance to manifold)



UF UNIVERSITY of FLORIDA

Duke UNIVERSITY

TEXAS The University of Texas at Austin

UC SANTA CRUZ

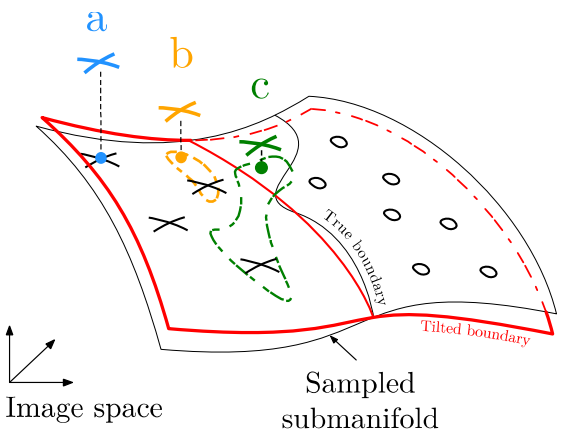
Interpretation

Characterize hard-label attacks as a “data-geometric” hierarchy:

a) Regular attacks: leave the manifold, low similarity

b) Query-efficient attacks: Near the manifold, high similarity

c) On-manifold (e.g., AE-enabled) attacks: low similarity due to on-manifold model and true boundaries



a: off-manifold, low-similarity
b: on-manifold, high-similarity
c: on-manifold, low-similarity
- - - - - - - - : search space

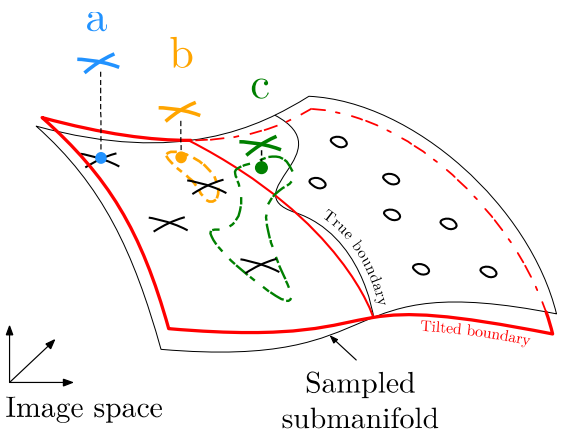
Interpretation

Characterize hard-label attacks as a "data-geometric" hierarchy:

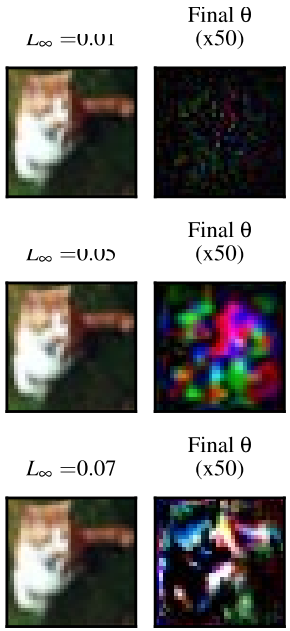
a) Regular attacks: leave the manifold, low similarity

b) Query-efficient attacks: Near the manifold, high similarity

c) On-manifold (e.g., AE-enabled) attacks: low similarity due to on-manifold model and true boundaries



a: off-manifold, low-similarity
b: on-manifold, high-similarity
c: on-manifold, low-similarity
: search space



Takeaways

Key observation: Query efficient attack samples lie closer to the manifold, robustness can hurt.

Our interpretation: Adversary leverages a "noisy manifold distance oracle" to improve query efficiency

Mutual information between model's gradient and data manifold was shown by Engstrom et al. (2019).

Our information theoretic argument (future work):

It can be shown by data processing inequality (Beaudry & Renner 2012) that,

$$I(\mathscr{M}, \ddot{\nabla} f) \quad \text{increases with} \quad I(\mathscr{M}, \nabla f)$$

for data manifold $\mathscr{M}$, model gradient ∇f, and ZO gradient estimate $\ddot{\nabla} f$.

Thank you

w.garcia@ufl.edu