Theoretically Principled Trade-off for Stateful Defenses against Query-Based Black-Box Attacks

Ashish Hooda [*] [1] Neal Mangaokar [*] [2] Ryan Feng [2] Kassem Fawaz [1] Somesh Jha [1] Atul Prakash [2]

Abstract

Adversarial examples threaten the integrity of machine learning systems with alarming success rates even under constrained black-box conditions. Stateful defenses have emerged as an effective countermeasure, detecting potential attacks by maintaining a buffer of recent queries and detecting new queries that are too similar. However, these defenses fundamentally pose a trade-off between attack detection and false positive rates, and this trade-off is typically optimized by hand-picking feature extractors and similarity thresholds that empirically work well. There is little current understanding as to the formal limits of this trade-off and the exact properties of the feature extractors/underlying problem domain that influence it. This work aims to address this gap by offering a theoretical characterization of the trade-off between detection and false positive rates for stateful defenses. We provide upper bounds for detection rates of a general class of feature extractors and analyze the impact of this trade-off on the convergence of black-box attacks. We then support our theoretical findings with empirical evaluations across multiple datasets and stateful defenses.

1. Introduction

Adversarial examples pose a significant threat to the security and integrity of machine learning systems (Eykholt et al., 2018; Sayles et al., 2021). These examples are subtly manipulated inputs that deceive the models and cause misclassifications (Szegedy et al., 2014; Carlini & Wagner, 2017; Hooda et al., 2022). Even in the challenging black-box setting, where the attacker has limited information access, adversarial examples have been remarkably successful (Ilyas et al., 2018; Chen et al., 2020a; Feng et al., 2022; Maho et al., 2021; Andriushchenko et al., 2020; Li et al., 2020).

Recent research has shown that stateful defenses offer a promising approach to mitigate the impact of such attacks (Li et al., 2022; Choi et al., 2023; Chen et al., 2020b). These defenses leverage the observation that black-box attackers often submit numerous highly similar queries, e.g., querying nearby points for gradient estimation. To counter this, stateful defenses maintain a buffer of recent queries and compare incoming queries in some feature space to identify potential attacks. If the similarity between queries exceeds a predefined threshold, defensive action is taken, e.g., banning the user's account (Chen et al., 2020b) or rejecting queries (Li et al., 2022).

The success of stateful defenses hinges on their ability to detect and flag attack queries without flagging benign ones. This suggests the existence of a trade-off between the detection and false positive rates of a stateful defense (much like the trade-off between robustness and accuracy for existing white-box defenses (Tsipras et al., 2018; Yang et al., 2020; Raghunathan et al., 2020)). In light of this, existing defenses typically tune their similarity threshold to manipulate the trade-off, i.e., such that the defense only permits an empirically computed false positive rate. However, this does not provide any guarantees for the detection rate, and little is currently known about the exact properties of the feature spaces and problem domains that influence this trade-off. This work aims to address this gap by theoretically characterizing the trade-off between the detection rate and the false positive rate of stateful defenses. Specifically, we provide upper bounds for the detection rate for a general class of feature extractors. We then empirically validate that the takeaways from these bounds hold for multiple datasets and defenses and also analyze how this trade-off affects the convergence of black-box attacks.

[*]Equal contribution [1]University of Wisconsin-Madison [2]University of Michigan. Correspondence to: Ashish Hooda <ahooda@wisc.edu>, Neal Mangaokar <nealmgkr@umich.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Background

2.1. Black-box Attacks

Adversarial Examples are perturbed inputs that intentionally mislead or deceive machine learning models. Specifically, given an image $\mathbf{x}$ with label y and a classifier f, such attacks aim to construct an adversarial example $\mathbf{x}_{adv}$ such that:

$$f(\mathbf{x}_{adv}) \neq y \text{ and } ||\mathbf{x}_{adv} - \mathbf{x}||_p \leq \epsilon \tag{1}$$

where ϵ is the perturbation budget per some ℓ_p norm. In the black-box setting, these attacks only have on query access to the model. One common characteristic of black-box attacks is the use of similar queries to gather information about the model's behavior. Specifically, by making queries with slight perturbations to the input and observing the corresponding model outputs, attackers can gain insights into the model's decision-making process.

Consider the initial stage of many black-box adversarial attacks, which involves estimating the direction to move the input to achieve the desired adversarial effect. For example, the NES (Ilyas et al., 2018), HSJA (Chen et al., 2020a), and QEBA (Li et al., 2020) attacks estimate the gradient by sampling nearby points from a Gaussian (or similar) probability distributions, and computing finite differences over these points. Other attacks such as SurFree (Maho et al., 2021) and Square (Andriushchenko et al., 2020) also sample nearby points to estimate a "random search" direction (not a gradient) in which to move the input. We will often refer to the interplay between such queries made during the direction estimation stage and a stateful defense, particularly because the attack's overall convergence properties are often directly influenced by choice of direction.

2.2. Stateful Defenses

The overall intuition behind stateful defenses is that black-box attackers often submit highly similar queries as part of the optimization procedure for their chosen adversarial task. These highly similar queries can then be detected. Defenses such as Blacklight (Li et al., 2022) have reduced attack success rate (ASR) of state-of-the-art black-box attacks to as low as 0%.

A stateful defense typically comprises a classifier f, feature extractor H (with some associated distance metric), query store q, and threshold τ. The defense then compares an incoming query against all queries stored in q. If similarity with any example in q exceeds τ, the defense deploys preventive measures such as query rejection or account banning.

Different stateful defenses primarily vary in their choices of H. Specifically, some defenses such as Blacklight and PIHA (Li et al., 2022; Choi et al., 2023) leverage discrete-valued metrics such as hamming distance over hashes, e.g., SHA-256 hashes of quantized pixels. Others, such as Stateful Defense (SD) (Chen et al., 2020b), employ real-valued metrics, e.g., ℓ_2 distance between embeddings from neural similarity encoders. In this work, we evaluate Blacklight and PIHA since they are available for both the CIFAR-10 and ImageNet datasets.

Model Stealing Recent work has also proposed stateful defenses against model-stealing attacks. Such attacks aim to steal a local "clone" model f^c such that the behavior of f^c is similar to that of f. Defenses such as SEAT (Zhang et al., 2021) have also been successful here and can force the attacker to create as many as 65 accounts to steal a single model. This success can be similarly explained by the submission of highly similar queries. For example, at iteration t of a Jacobian-based Augmentation (JBA) attack (Papernot et al., 2017), the adversary constructs a "useful" but highly similar query $\mathbf{x}_{t+1}$ by perturbing previous query $\mathbf{x}_t$ so that it maximizes the loss $\mathcal{L}$ of f^c_t:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta * sgn(\nabla_{\mathbf{x}_t}\mathcal{L}(\mathbf{x}_t, f(\mathbf{x}_t))) \tag{2}$$

where η is some step size.

3. Trade-offs between Detection and False Positives

In this section, we demonstrate that there exists an implicit trade-off between detecting attack queries and avoiding false positives in the context of stateful defenses. We begin with a constructive model through which we provide explicit characterizations of the feature extractor and data distributions. We use this toy model to highlight the trade-off, and then relax the assumptions to provide a more general bound that highlights the direct influence of the feature extractor and the problem domain.

3.1. Toy Model

Feature extractor. We begin by considering an explicit class of feature extractors based on simple quantization. The feature extractor is given by $H : \mathbb{R}^d \rightarrow \mathbb{Z}^d$ with a discrete output space. Specifically,

$$H(\mathbf{x}) = \lfloor \mathbf{x} + \mathbf{0.5} \rfloor \tag{3}$$

where the $\lfloor . \rfloor$ operation is element-wise. Many defenses employ quantization to provide perceptual similarity (Li et al., 2022; Choi et al., 2023). In this model, we consider a query to be an attack query if and only if it produces the exact same features as that of a prior query. Later, in Section 3.2 we expand beyond the toy model to consider the case where H is a generic feature extractor, and queries are considered attack queries when their features are within some distance τ of a prior query.

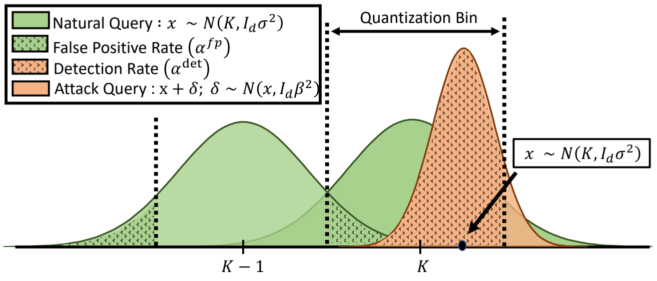
Natural Query : $x \sim N(K, I_d\sigma^2)$
False Positive Rate (α^{fp})
Detection Rate (α^{det})
Attack Query : $x + \delta$; $\delta \sim N(x, I_d\beta^2)$
Quantization Bin
$x \sim N(K, I_d\sigma^2)$
$K-1$
K

Figure 1. Illustration of the Toy Model in 1-D. We assume that any natural query $\mathbf{x}$ is sampled from a distinct Gaussian distribution (green). Two such distributions are shown, centered at $K-1$ and K. For a given natural query $\mathbf{x}$, the attack queries $\mathbf{x}+\boldsymbol{\delta}$ are sampled from another Gaussian distribution (orange) centered around a natural query. The feature extractor is designed to map each natural query to a unique output. Therefore, H maps all values within each quantization bin to the same output. This means that the green shaded area represents α^{fp}, and the orange shaded area represents α^{det} for the attack queries.

Natural Query Distribution. Stateful defenses assume that natural images are sufficiently “spread out”, or dissimilar enough such that they can be distinguished. Therefore, for our model we assume that natural images originate from one of several Gaussian distributions, which are uniformly dispersed across input space $\mathbb{R}^d$ [1]. Each natural image is obtained from a distinct Gaussian distribution. This may be viewed as a “best case” situation for the defense, where natural images are sufficiently spread out across the input space to avoid false positives. For simplicity, we assume isotropic Gaussian distributions: $\mathcal{N}(\mathbf{p}, \mathbf{I_d}\sigma^2)$ where $\mathbf{p} \in \mathbb{Z}^d$. Intuitively, when applying H to a natural image $\mathbf{x} \sim \mathcal{N}(\mathbf{p}, \mathbf{I_d}\sigma^2)$, it should output the discrete feature vector $\mathbf{p}$ with high probability.

Attack Query Distribution. To estimate the gradient at input $\mathbf{x}$, a Monte Carlo simulation approach would require sampling a total of q perturbations $\{\mathbf{x}, \mathbf{x}+\boldsymbol{\delta}_\mathbf{1}, ..., \mathbf{x}+\boldsymbol{\delta}_\mathbf{q}\}$. For our model, we consider the distribution of perturbations for $\mathbf{x}$ to be $\mathcal{N}(0, \mathbf{I_d}\beta^2)$, i.e., the adversary is estimating a gradient using finite differences on a Gaussian basis (Ilyas et al., 2018).

Given the setting described above (also illustrated in Figure 1), we now present the following result, which bounds the detection rate with the false positive rate:

Theorem 3.1. *Let the adversary sample a natural image $\mathbf{x}$ from one of the above distributions $\mathcal{N}(\mathbf{p}, \mathbf{I_d}\sigma^2)$, and perturb it with $\boldsymbol{\delta} \sim \mathcal{N}(0, \mathbf{I_d}\beta^2)$ to estimate a gradient. Given that the stateful defense incurs a false positive rate α^{fp}, the detection rate α^{det} for the perturbed query $\mathbf{x}+\boldsymbol{\delta}$ is then bounded as follows:*

$$\alpha^{det} \leq 1 - \left(2 - 2\Phi\left(0.5\beta^{-1}\right)\right)^d \left(1 - \alpha^{fp}\right) \tag{4}$$

Proof. H fails to detect the attack query $\mathbf{x}+\boldsymbol{\delta}$ if and only if $H(\mathbf{x}+\boldsymbol{\delta}) \neq H(\mathbf{x})$. Therefore,

$$\begin{aligned}
\alpha^{det} &= 1 - \mathbb{P}[H(\mathbf{x}) \neq H(\mathbf{x}+\boldsymbol{\delta})] && (5)\\
&\leq 1 - \mathbb{P}[H(\mathbf{x}+\boldsymbol{\delta}) \neq \mathbf{p}, H(\mathbf{x}) = \mathbf{p}] && (6)\\
&= 1 - \mathbb{P}[H(\mathbf{x}+\boldsymbol{\delta}) \neq \mathbf{p} \mid H(\mathbf{x}) = \mathbf{p}]\mathbb{P}[H(\mathbf{x}) = \mathbf{p}] && (7)\\
&\leq 1 - \mathbb{P}[H\left(\mathbf{p}+\boldsymbol{\delta}\right) \neq \mathbf{p}]\mathbb{P}[H(\mathbf{x}) = \mathbf{p}] && (8)\\
&= 1 - \mathbb{P}[||\boldsymbol{\delta}||_\infty > 0.5]\mathbb{P}[H(\mathbf{x}) = \mathbf{p}] && (9)\\
&= 1 - \left(2 - 2\Phi\left(0.5\beta^{-1}\right)\right)^d \left(1 - \alpha^{fp}\right) && (10)
\end{aligned}$$

where Φ is the cummulative distribution function of $\mathcal{N}(0, 1)$. Note that to go from (7) to the inequality in (8), we assign a specific value $\mathbf{x} = \mathbf{p}$, i.e., placing $\mathbf{x}$ at the center of the quantization bin for H (see Equation 3). By placing it at the center, the probability of evasion when adding $\boldsymbol{\delta}$ is minimized, and the resulting event is also independent of event $H(\mathbf{x}) = \mathbf{p}$. Finally, going from (9) to (10) uses standard results for the CDF of a multivariate Gaussian. □

Takeaway. There exists a trade-off between the detection rate α^{det} and the false positive rate α^{fp}, i.e., decreasing α^{fp} also decreases the upper bound for α^{det}. Furthermore, this trade-off also depends on the standard deviation β of the perturbation distribution, i.e., high values of β lead to a lower detection rate.

3.2. General Analysis

Recall that our toy model assumed a quantization-based feature extractor and a uniform natural image distribution. We now extend our results to a more generic perceptual feature extractor and image distribution. Specifically, consider $H : \mathbb{R}^d \rightarrow \mathbb{R}^y$ where y is the dimensionality of the output feature space. We assume H to be Lipschitz continuous with constants K_L and K_U :

$$K_L||\mathbf{x_1} - \mathbf{x_2}|| \leq ||H(\mathbf{x_1}) - H(\mathbf{x_2})|| \leq K_U||\mathbf{x_1} - \mathbf{x_2}||, \tag{11}$$

$\forall\ (\mathbf{x_1}, \mathbf{x_2}) \in \mathbb{R}^d$. Note that we no longer assume the implementation of H as in the toy model; the continuity assumption here is only needed to ensure that H captures perceptual similarity, i.e., similar images should indeed have similar features. Furthermore, since H is now continuous, we extend to a threshold based detection setting i.e. a query $\mathbf{x}$ is considered an attack query if and only if

[1] For the case where the input space is constrained, for instance to [0,255], the natural images can instead be sampled from truncated Gaussian distributions.

$||H(\mathbf{x}) - H(\mathbf{x_h})|| \leq \tau$ where $\mathbf{x_h}$ is any historical query. Given these changes, we can now re-analyze the detection α^{det} for a perturbed query $\mathbf{x} + \boldsymbol{\delta}$:

Theorem 3.2. *Let the adversary sample natural image $\mathbf{x}$, and perturb it with $\boldsymbol{\delta} \sim \mathcal{N}(0, \mathbf{I_d}\beta^2)$ to estimate a gradient. For a false positive rate α^{fp}, the detection rate α^{det} for perturbed query $\mathbf{x} + \boldsymbol{\delta}$ is then bounded as follows:*

$$\alpha^{det} \leq \frac{1}{\Gamma(\frac{d}{2})} \gamma \left(\frac{d}{2}, \frac{1}{2} \left(\frac{K_U}{K_L} \frac{M_{\mathcal{D}}}{\beta} \frac{1}{1 - \alpha^{fp}} \right)^2 \right) \tag{12}$$

where $M_{\mathcal{D}} = \mathbb{E}[||\mathbf{x_1} - \mathbf{x_2}||]$, i.e., the expected spread of natural queries, and γ and Γ are the monotonic lower incomplete and complete Gamma functions respectively.

Proof. H fails to detect the attack query $\mathbf{x} + \boldsymbol{\delta}$ if and only if $||H(\mathbf{x}) - H(\mathbf{x} + \boldsymbol{\delta})|| > \tau$. Therefore,

$$\alpha^{det} = \mathbb{P}\left[||H(\mathbf{x}) - H(\mathbf{x} + \boldsymbol{\delta})|| \leq \tau\right] \tag{13}$$

Similarly, H produces a false positive for two natural images $\mathbf{x_1}$ and $\mathbf{x_2}$ if and only if $||H(\mathbf{x_1}) - H(\mathbf{x_2})|| \leq \tau$. Therefore,

$$\alpha^{fp} = \mathbb{P}\left[||H(\mathbf{x_1}) - H(\mathbf{x_2})|| \leq \tau\right] \tag{14}$$

Using Equation 11 with 13 and 14:

$$\alpha^{det} \leq \mathbb{P}\left[||\boldsymbol{\delta}|| \leq \frac{\tau}{K_L} \right] \tag{15}$$

$$\alpha^{fp} \geq \mathbb{P}\left[||\mathbf{x_1} - \mathbf{x_2}|| \leq \frac{\tau}{K_U} \right] \tag{16}$$

Finally, using a CDF for the norm of a Gaussian, i.e., a chi-distribution in Equation 15 and Markov's inequality in Equation 16, we get:

$$\alpha^{det} \leq \frac{1}{\Gamma(\frac{d}{2})} \gamma \left(\frac{d}{2}, \frac{1}{2} \left(\frac{K_U}{K_L} \frac{M_{\mathcal{D}}}{\beta} \frac{1}{1 - \alpha^{fp}} \right)^2 \right) \tag{17}$$

where:

$$\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt \tag{18}$$

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt \tag{19}$$

□

Takeaway. The trade-off observed in the toy model also extends to the more general setting, i.e., the detection rate α^{det} and the false positive rate α^{fp} are still at odds with each other. Furthermore, this trade-off depends upon the standard deviation β of the perturbation distribution, the expected spread $M_{\mathcal{D}}$ of natural queries, and the Lipschitz constant ratio K_U/K_L of H.

4. Experiments

Motivated by our analysis in Section 3, we conduct experiments to validate our findings empirically, and thus answer the following questions:

Q1. How does the trade-off empirically depend upon the spread, i.e., variance β of the attack queries?

Q2. How does the trade-off empirically depend upon the Lipschitz constant ratio K_U/K_L of the feature extractor?

Q3. What are the implications of the trade-off for the convergence of black-box attacks?

4.1. Experimental Setup

Feature extractors. We focus our evaluation on feature extractors from two state-of-the-art stateful defenses: Blacklight (Li et al., 2022) and PIHA (Choi et al., 2023). Below we provide detailed descriptions and hyper-parameters for both.

Blacklight operates on an input image with pixel values in the range of [0, 255]. First, it discretizes the pixels into bins of size 50. Second, a sliding window technique is applied to the discretized image, utilizing a window size of 20 for TinyImages (Torralba et al., 2008) and 50 for ImageNet (Russakovsky et al., 2015). During this process, each window is hashed using the SHA-256 algorithm. Finally, the resulting set of hashes obtained from all the windows is considered as the "feature" for the image. For efficiency purposes, Blacklight utilizes only the top 50 hashes. To quantify the distance between two hash sets, Blacklight computes the number of non-common hashes, which can be interpreted as an ℓ_1 distance.

PIHA also operates on input images with the same pixel range. First, it runs a 3x3 low-pass Gaussian filter with standard deviation 1 over the image. Second, the image is converted to the HSV color space with the S and V components discarded. Finally, PIHA runs a sum-pooling operation over 7x7 image blocks, and the "feature" is computed as the output of the local binary pattern algorithm (Ojala et al., 1994) on the sum-pooled image.

Datasets. We evaluate Blacklight and PIHA using two datasets, TinyImages and ImageNet. The TinyImages dataset is a collection of 32x32 images and is the superset collection from which the popular CIFAR-10 dataset is sampled (providing nearly 80 million images as opposed to only 60,000). The ImageNet dataset comprises over 1 million 256x256 images. We sample a random subset of 1 million images from both datasets for our experiments.

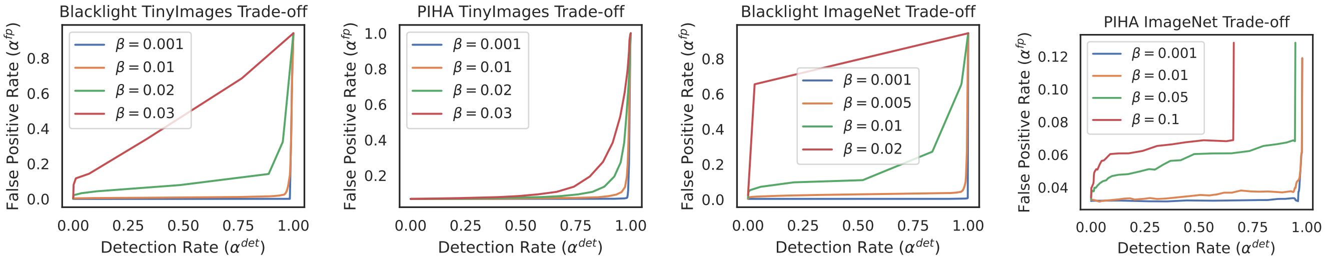

Figure 2. There exists a trade-off between detection rate α^{det} and false positive rate α^{fp} for stateful defenses. This trade-off is worsened for larger β values. Each curve is computed by varying threshold τ for the chosen feature extractor, and each setting presents four curves corresponding to different β values.

4.2. Q1. Variance of Attack Queries

Theorem 3.2 suggests a clear inverse relationship between the (α^{det}, α^{fp}) trade-off and β. We now empirically validate this relationship, i.e., for any given feature extractor and dataset, we plot α^{fp} against α^{det} for a variety of thresholds τ. We compute α^{fp} over 1 million images for all settings except PIHA on ImageNet, for which we compute on 100k and extrapolate due to computational complexity. We compute α^{det} over 100 images by sampling perturbations from Gaussians with different standard deviations β.

Results are presented in Figure 2. Notably, we first observe that for any β, the trade-off between α^{det} and α^{fp} indeed exists across all thresholds. More specifically, to obtain a larger α^{det} always requires an increase in α^{fp} as well. This validates the takeaways from Theorems 3.1 and 3.2. Furthermore, the inverse relationship with β also exists, i.e., achieving the same α^{det} requires a larger α^{fp} when β is increased. Interestingly, PIHA can achieve higher α^{det} on the low-dimensional TinyImages compared to Blacklight, but both suffer on ImageNet when β increases beyond $\beta = 0.01$.

4.3. Q2. Lipschitz Constants of the Feature Extractor

Theorem 3.2 also suggests that the (α^{det}, α^{fp}) trade-off is influenced by Lipschitz constants K_U and K_L of the feature extractor. However, this assumes a continuous feature extractor — although the feature extractors from Blacklight and PIHA are not continuous, they are still designed to approximate the perceptual likeness of images (yielding closer features for similar queries and further features for dissimilar ones). Given the lack of closed-form expressions, we resort to an empirical estimation of K_U and K_L.

We create image pairs $\mathbf{x}$ and $\mathbf{x} + \boldsymbol{\delta}$ where $\mathbf{x}$ is sampled from the dataset (TinyImages/ImageNet), and $\boldsymbol{\delta} \sim \mathcal{N}(0, \mathbf{I_d}\beta^2); \beta = 0.01$. For each pair, we then calculate the ratio between the ℓ_2 distance in the feature space and the input space, i.e., $\frac{||H(\mathbf{x}) - H(\mathbf{x}+\boldsymbol{\delta})||}{||\boldsymbol{\delta}||}$. We construct 10000 such pairs and plot the distribution of these distance ratios.

Figure 3(a) plots these distributions for ImageNet images processed by both Blacklight and PIHA feature extractors. We note a larger distribution spread in the histogram for PIHA compared to Blacklight, hinting at a greater value for $\frac{K_U}{K_L}$ for PIHA. As per Theorem 3.2, this suggests that PIHA possesses the potential for superior detection rates compared to Blacklight. We corroborate this empirically by plotting α^{fp} against α^{det} in a manner akin to that in Section 4.2. As presented in Figure 3(b), PIHA indeed manifests higher detection rates when compared with Blacklight.

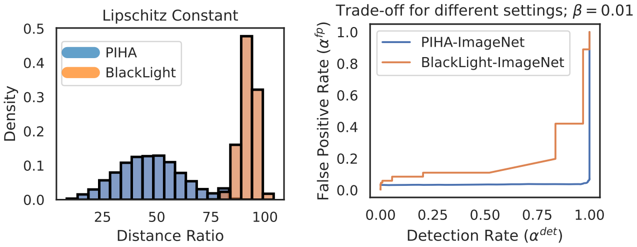

Figure 3. Lipschitz constant ratio of the feature extractors is directly proportional to the quality of the trade-off. On the left, we present the distribution of ratios between pairwise distance in the feature space and pairwise distance in the input space — a larger distribution spread implies a larger Lipschitz ratio for that feature extractor. On the right, we present the corresponding (α^{det}, α^{fp}) trade-off.

4.4. Q2. The Trade-off and Attack Convergence

Given that increasing β worsened the trade-off of the defense (Q1 in Section 4.2), we now question the impact of increasing β on the attack convergence itself. We specifically consider the adversary goal of gradient estimation via finite differences. Formally, it can be shown through the following result that increasing β should worsen the quality of the estimated gradient:

Theorem 4.1. *Let ∇_x be the true gradient of $\mathbf{x}$ for the*

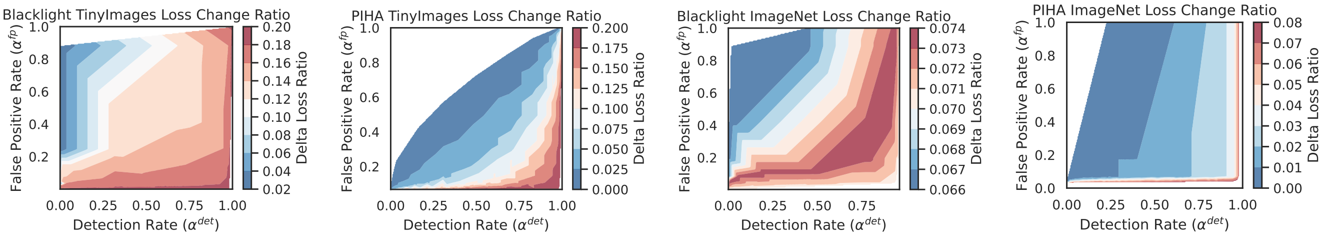

Figure 4. Even though the (α^{det}, α^{fp}) trade-off is worsened for larger β values, the larger β values also produce an inferior direction which does not increase the loss much. Each shaded region corresponds to the change in loss induced by a gradient estimated with different β values. Red regions imply a large increase in loss, and blue imply a small increase.

classifier's loss, and G be a matrix of rows $g_1, \cdots, g_k \sim \mathcal{N}(0, \mathbf{I_d}\beta^2)$. Then, the norm of estimated gradient $G \cdot \nabla_{\mathbf{x}}$ is bounded in probability by:

$$\mathbb{P}[(1-\epsilon)\|\nabla_{\mathbf{x}}\| \le \|G \cdot \nabla_{\mathbf{x}}\| \le (1+\epsilon)\|\nabla_{\mathbf{x}}\|] \ge 1 - 2 \cdot exp\left(-k - \frac{1+\epsilon}{2\beta^2}\right)$$

where $0 \le \epsilon \le 1$ is the estimation error.

A detailed proof of this result can be found in Appendix A.0.1. The left-hand side represents the probability that our estimated gradient is "good", i.e., produces the same increase-in-loss as the true gradient. As β increases, the lower bound on this probability decreases (right-hand side), suggesting that the estimate is less likely to produce the same increase-in-loss.

We empirically validate this impact of increasing β in Figure 4, which plots the increase in loss when following gradients estimated with different β. These figures present a clearer overall picture — for any given α^{fp}, even though larger β decreases the detection rate, a gradient estimated with larger β is also strictly worse for the adversary, i.e., does not increase the loss as much (see the gradation from red to blue). In other words, these findings suggest that the worsening of the (α^{det}, α^{fp}) trade-off at larger β is not without a negative impact on the adversary.

5. Conclusion

In conclusion, our work offers a more formal understanding of how stateful defenses prevent black-box adversarial attacks. We outlined a crucial trade-off between detecting attack detection and false positives, and highlighted its dependence upon the distribution of attack and natural queries, and the properties of the defense's feature extractor. Our analysis can help illuminate why certain defenses perform better against black-box attacks, which can help to refine current strategies and potentially guide the design of future defenses. As the landscape of adversarial attacks and defenses evolves, our findings contribute to the development of more robust and resilient machine learning models under the realistic black-box threat model.

6. Acknowledgements

This material is based upon work supported by DARPA under agreement number 885000, National Science Foundation Grant No. 2039445, and National Science Foundation Graduate Research Fellowship Grant No. DGE 1841052. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of our research sponsors.

References

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pp. 484–501. Springer, 2020.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, Los Alamitos, CA, USA, may 2017. IEEE Computer Society. doi: 10.1109/SP.2017.49. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.49.

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

Chen, S., Carlini, N., and Wagner, D. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pp. 30–39, 2020b.

Choi, S.-H., Shin, J., and Choi, Y.-H. Piha: Detection method using perceptual image hashing against query-based adversarial attacks. *Future Generation Computer Systems*, 2023.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

Feng, R., Mangaokar, N., Chen, J., Fernandes, E., Jha, S., and Prakash, A. Graphite: Generating automatic physical examples for machine-learning attacks on computer vision systems. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 664–683. IEEE, 2022.

Hooda, A., Mangaokar, N., Feng, R., Fawaz, K., Jha, S., and Prakash, A. Towards adversarially robust deepfake detection: An ensemble approach, 2022.

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

Li, H., Xu, X., Zhang, X., Yang, S., and Li, B. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1221–1230, 2020.

Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., and Zhao, B. Y. Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2117–2134, 2022.

Maho, T., Furon, T., and Le Merrer, E. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10430–10439, 2021.

Ojala, T., Pietikainen, M., and Harwood, D. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pp. 582–585. IEEE, 1994.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sayles, A., Hooda, A., Gupta, M., Chatterjee, R., and Fernandes, E. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14666–14675, June 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Zhang, Z., Chen, Y., and Wagner, D. Seat: similarity encoder by adversarial training for detecting model extraction attack queries. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pp. 37–48, 2021.

A. Supplementary Proofs

Lemma A.1. *Let G be a $k \times d$ random matrix with rows $\sigma g^i \sim \mathcal{N}(0, \mathbf{I_d}\sigma^2)$ $\forall 1 \leq i \leq k$. Then, for any unit vector $v \in \mathbb{R}^d$,*

$$P[|\|Gv\|^2 - 1| > \epsilon] \leq 2exp\left(-\left(k + \frac{\epsilon + 1}{2\sigma^2}\right)\right)$$

Proof. Note that by rotational invariance of Gaussians, $Gv \overset{D}{=} Ge^1$, where e^1 is the standard basis vector. This implies that $\|Gv\|^2 \overset{D}{=} \|Ge^1\|^2 \overset{D}{=} \sigma^2\chi_k^2$, where χ_k^2 is a chi-square random variable with k-degrees of freedom. Then, by Chernoff's bounding method:

$$P[|\|Gv\|^2 - 1| > \epsilon] = P[|\sigma^2\chi_k^2 - 1| > \epsilon]$$

$$\begin{aligned}
&\leq 2 \inf_{t>0} e^{-\epsilon t}\mathbb{E}[e^{t(\sigma^2\chi^2 - 1)}] \\
&\leq 2 \inf_{t>0} e^{-\epsilon t - t}\mathbb{E}[e^{t\sigma^2\chi^2}] \\
&\leq 2 \inf_{t>0} e^{-\epsilon t - t}(1 - 2\sigma^2 t)^{\frac{-k}{2}} \\
&\leq 2 \inf_{t>0} e^{-\epsilon t - t - \frac{k}{2}\log(1 - 2\sigma^2 t)} \\
&\leq 2e^{-\frac{\epsilon - k\sigma^2 + 1}{2\sigma^2} - \frac{k}{2}\log(\frac{\sigma^2 k}{\epsilon + 1})} \\
&\leq 2e^{-\frac{(\epsilon + 1 - \sigma^2 k)^2}{2\sigma^2(1+\epsilon)}} \\
&\leq 2e^{-\frac{(\epsilon+1)^2 - 2\sigma^2 k(\epsilon+1)}{2\sigma^2(1+\epsilon)}} \\
&\leq 2e^{-\frac{\epsilon + 1 - 2\sigma^2 k}{2\sigma^2}} \\
&\leq 2e^{-k}e^{-\frac{\epsilon+1}{2\sigma^2}} \\
&\leq 2e^{-k - \frac{\epsilon+1}{2\sigma^2}}
\end{aligned}$$

□

A.0.1. Proof for Theorem 4.1

Let $\nabla_\mathbf{x}$ be the true gradient of $\mathbf{x}$ for the classifier's loss, and G be a matrix of rows $g_1, \cdots, g_k \sim \mathcal{N}(0, \mathbf{I_d}\beta^2)$. Then, the norm of estimated gradient $G \cdot \nabla_\mathbf{x}$ is bounded in probability by:

$$\mathbb{P}[(1-\epsilon)\|\nabla_\mathbf{x}\| \leq \|G \cdot \nabla_\mathbf{x}\| \leq (1+\epsilon)\|\nabla_\mathbf{x}\|] \geq 1 - 2 \cdot exp\left(-k - \frac{1+\epsilon}{2\beta^2}\right)$$

where $0 \leq \epsilon \leq 1$ is the estimation error.

Proof.

$$\begin{aligned}
&P[(1-\epsilon)\|\nabla_\mathbf{x}\| \leq \|G\nabla_\mathbf{x}\| \leq (1+\epsilon)\|\nabla_\mathbf{x}\|] \\
&= P[(1-\epsilon)^2\|\nabla_\mathbf{x}\|^2 \leq \|G\nabla_\mathbf{x}\|^2 \leq (1+\epsilon)^2\|\nabla_\mathbf{x}\|^2] \\
&\geq P\left[1 - \epsilon \leq \frac{\|G\nabla_\mathbf{x}\|^2}{\|\nabla_\mathbf{x}\|^2} \leq 1 + \epsilon\right] \\
&= P\left[\left|\frac{\|G\nabla_\mathbf{x}\|^2}{\|\nabla_\mathbf{x}\|^2} - 1\right| \leq \epsilon\right] \\
&= P\left[\left|\left\|G\frac{\nabla_\mathbf{x}}{\|\nabla_\mathbf{x}\|}\right\|^2 - 1\right| \leq \epsilon\right] \\
&\geq 1 - 2e^{-k - \frac{\epsilon+1}{2\beta^2}}
\end{aligned}$$

Where the last step is by Lemma 1. □