Mind the Gap: Detecting Black-box Adversarial Attacks in the Making through Query Update Analysis (Supplementary Materials)

1. Concentration of Measure

$\mathcal{DS}$ distributions extracted from the sequence of queries from query-based black-box adversarial attacks exhibit unique and distinguishable patterns. To explain this observation, we take an approach that adopts the concentration of measure phenomena in high-dimensional spaces. The properties of high-dimensional spaces often defy the rules based on low-dimensional spaces we are familiar with[2], and there is a set of less appreciate phenomena, especially in data analysis, called the concentration of measure. The concentration of measure phenomena are non-trivial observations and properties of the large number of random variables[17].

Two important and useful geometric metrics of random distributions are the length of a random vector $u_i \in \mathbb{R}^d$, and the angle of two random vectors $u_i, u_j \in \mathbb{R}^d$. In high-dimensional spaces, however, the concentration of measure states these metrics are almost concentrated to a single value in the sense of the measure.

1.1. Concentration of Length

Let d-dimensional vector $u_i \in \mathbb{R}^d$ be a random vector that is sampled from a random distribution $\mathcal{N}(\mu, \sigma^2)$. The law of large numbers states that $\frac{1}{d}\sum_{k=1}^{d} u_i^k$ is almost surely μ as $d \to \infty$ [10], where u_i^k is the k'th element of u_i. If μ and σ^2 is fixed, the length of any vector u_i is expected to converge toward a common value. Especially when $\mu = 0$, the length converges to $\sigma^2\sqrt{d}$.

Suppose a random distribution $\mathcal{N}(\mu, \sigma^2)$ has $\mu = 0$, and the random fluctuation is limited with the unit variance $\sigma^2 = 1$. The expected length of u_i is denoted as[1],

$$E\left[\|u_i\|_2^2\right] = E\left[\sum_{k=1}^{d} |u_i^k|^2\right] = \sum_{k=1}^{d} E\left[|u_i^k|^2\right] = d \quad (1)$$

where $E[u_i^k]$ is the expected value of the random variable u_i^k. Therefore, the Euclidean length of u_i is expected to be approximately $\sqrt{d}$.

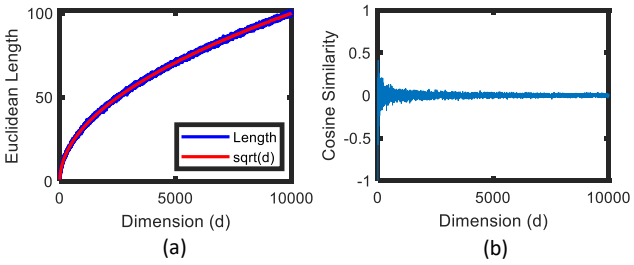

Figure 1. (a) Euclidean length of random vectors with the unit variance $\sigma^2 = 1$ follows $\sqrt{d}$ as the dimension d grows; (b) Cosine Similarity of two random vectors converges toward 0 (90°) as the dimension d grows.

1.2. Concentration of Angle

Let two d-dimensional vectors u_i, $u_j \in \mathbb{R}^d$ be independent random vectors with Rademacher variables as $u_i^k \in \{-1, 1\}$. The angle between two vectors is denoted as $\angle u_i u_j$, and the cosine similarity is denoted as $\cos\angle u_i u_j = \frac{u_i \cdot u_j}{\|u_i\|_2 \|u_j\|_2}$, where $u_i \cdot u_j = \sum_{k=1}^{d} u_i^k u_j^k$ is the sum of independent random variables, hence $E[u_i \cdot u_j] = \sum_{k=1}^{d} E[u_i^k u_j^k] = 0$. Therefore, Hoeffding's inequality[13] can be applied for any given $t > 0$ as

$$P\left(|u_i \cdot u_j| \geq t\right) = P\left(\frac{|u_i \cdot u_j|}{d} \geq \frac{t}{d}\right) \leq 2e^{\left(-\frac{t^2}{2d}\right)} \quad (2)$$

where P is the probability. Now replace t with $\sqrt{2d\log d}$, then the inequality in Equation (2) is rewritten as

$$P\left(\frac{|u_i \cdot u_j|}{d} \geq \sqrt{\frac{2\log d}{d}}\right) \leq 2e^{(-\log d)} \quad (3)$$

From Equation (1), $\|u_i\|_2\|u_j\|_2$ is surely d. Therefore, Equation (3) becomes as follows [1],

$$P\left(|\cos\angle u_i u_j| \geq \sqrt{\frac{2\log d}{d}}\right) \leq \frac{2}{d} \quad (4)$$

As stated in Equation (4), the angles of two independent vectors, sampled from $\mathcal{N}(0, \sigma^2)$, highly likely become narrowly distributed around the mean $\angle u_i u_j = \pi/2$, with a

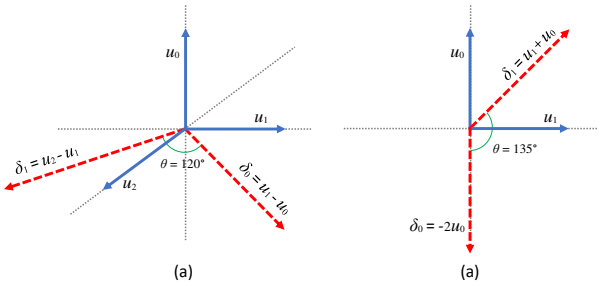

Figure 2. Delta Similarity ($\mathcal{DS}$) of Query-based Black-box Attacks. (a) $\mathcal{DS}$ of HSJA zeroth order optimisation; (b) $\mathcal{DS}$ of NES zeroth order optimistaion.

variance that converges towards zero[19] as the dimension grows.

The concentration of measure phenomena are empirically proven as illustrated in Figure 1.

2. Delta Similarity of Attack methods

We provide $\mathcal{DS}$ analysis of HSJA[6] and NES[15] attack methods. These two attack methods represent hard-label and soft-label based attack strategies respectively.

HSJA : HSJA[6] estimates the direction of gradient via the Monte Carlo algorithm as:

$$\nabla \mathcal{H}(\tilde{x}_t) \approx \frac{1}{n} \sum_{j=1}^{n} \mathcal{H}(\tilde{x}_t + \epsilon u_t^j) u_t^j \tag{5}$$

where $u_t^j \in \mathbb{R}^d$ is i.i.d. random vector. The sequence of queries in Equation (5) is $\{\tilde{x}_t + \epsilon u_t^j,\ \tilde{x}_t + \epsilon u_t^{j+1},\ ...\}$. Therefore, δ becomes a scaled subtraction of two random vectors as $\delta_t^i = \epsilon(u_t^{j+1} - u_t^j)$ and $\delta_t^{i+1} = \epsilon(u_t^{j+2} - u_t^{j+1})$. Three random vectors u_t^j, u_t^{j+1} and u_t^{j+2} are surely orthogonal and have the same length. Hence $\mathcal{DS} = -0.5(120°)$ as illustrated in Figure 2 (a).

NES : Ilyas et al. adopted NES[15] to estimate the gradient. They use search distribution of random Gaussian noise around the intermediate adversarial example in every iteration such as:

$$\nabla F(\tilde{x}_t) \approx \frac{1}{2\epsilon n} \sum_{j=1}^{n} \big(F(\tilde{x}_t + \epsilon u_t^j) - F(\tilde{x}_t - \epsilon u_t^j)\big) u_t^j \tag{6}$$

The sequence of queries in Equation (6) is $\{\tilde{x}_t + \epsilon u_t^j,\ \tilde{x}_t - \epsilon u_t^j,\ \tilde{x}_t + \epsilon u_t^{j+1},\ ...\}$. In this sequence, $\delta_t^i = -2\epsilon u_t^j$, and $\delta_t^{i+1} = \epsilon(u_t^{j+1} + u_t^j)$. Therefore $\mathcal{DS}$ becomes $-0.7071(135°)$. $\mathcal{DS}$ of NES zeroth order optimisation is illustrated in Figure 2 (b).

3. Screener Pre-Processing

GWAD screener pre-process is a cost-effective and light-weight stateful detection to screen out un-suspicious queries

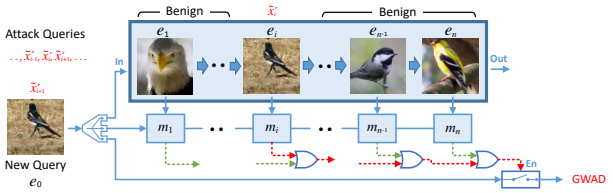

Figure 3. Conceptual diagram of GWAD Screener pre-process. Screener screens out un-suspecious examples deliberately injected by the irregular batch attacks.

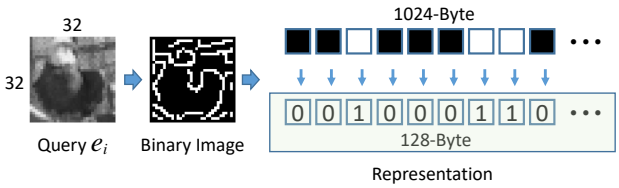

Figure 4. GWAD Screener pre-process 128-Byte representation of an example

Attack	Screener FIFO Depth			
	100	200	600	1K
HSJA	99.50%	99.11%	99.61%	98.86%
NES	99.92%	99.91%	99.92%	99.92%
Sign-Flip	99.76%	99.68%	99.79%	99.78%

Table 1. GWAD$^+$ (Screener + GWAD) detection rates with various Screener FIFO depths over irregular batch attack at $\mathbf{r}_b = 1000\%$ against ResNet-18 trained on CIFAR-10

which are deliberately injected by the irregular batch attacks. The screener represents an image x with the 128-Byte vector. Image x first converted to 32×32 grayscale image, then transformed into a binary image by Canny edge detection algorithm [5]. Each pixel of the binary image is represented with bit 0 or 1, and finally forms a 128-Byte vector. The similarity between two images is scored by the rate of mismatching pixels which is simply calculated with the bit-wise xor of represented vectors and the number of bit 1 in the result, as described in Equation 5 of the main paper. Figure 3 illustrates GWAD screener pre-process and Figure 4 depicts the 128-Byte representation.

***n*-Channel Attack :** n-Channel attack setting denotes where an adversary attacks multiple images in parallel. To cope with the n-Channel attack, screener is extended with a channel-aware detection mechanism that assigns a channel ID (CID) to each query, enabling independent tracking of different query sequences: **i)** Queries are dynamically assigned to distinct channels based on their similarity to prior inputs. **ii)** Each query's $\mathcal{DS}$ score is computed within its respective channel, preventing adversarial alternation. **iii)**

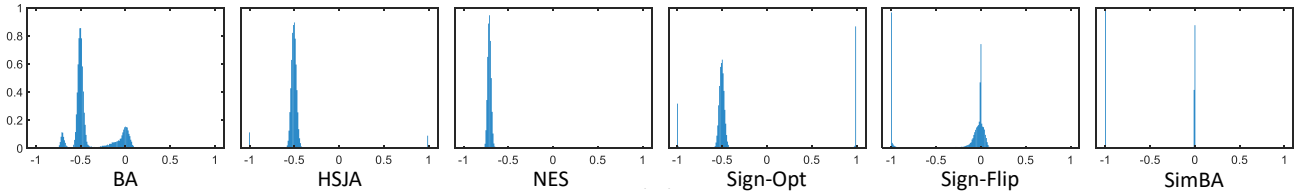

Figure 5. Mean models of CIFAR-10 HoDS training dataset

Layer	Layer Type	Node	Activation
0	Input	201	ReLU
1	Linear	512	ReLU
2	Linear	512	ReLU
3	Linear	256	ReLU
4	Linear	128	ReLU
5	Linear	64	ReLU
5	Linear	7	ReLU
6	LogSoftmax	7	-

Table 2. Architecture of GWAD attack classifier

Attack methods	Query Consumption		
	Zeroth Opt.	Linear Search	Other
BA [4]	90.42%	00.50%	9.08%
HSJA [6]	96.06%	03.93%	0.01%
SimBA [12]	100.00%	00.00%	0.00%
Sign-OPT [8]	71.67%	28.14%	0.19%
Sign-Flip [7]	99.80%	00.19%	0.01%
NES [15]	100.00%	00.00%	0.00%

Table 3. Query consumption of SOTA attack processes. Attacks on MobileNet-V2 trained on CIFAR-10 with 5K query budgets

Screener efficiently maintains separate state information for each attack sequence, ensuring that alternating queries do not evade detection. Our preliminary results confirm the effectiveness of this strategy, achieving over 99% accuracy in detecting a 2-Channel attack.

4. Attack Classifier and Training Feature Set

We train a simple neural network to classify attack queries. As shown in Table 2, the network consists of six full-connected hidden layers with ReLU activations. The classification probabilities are provided by a Log-Softmax output layer.

HoDS Training Feature Set : We first configure all the adversarial attack methods, listed in Table 2 of the main paper, to carry out their attacks with the query budget $q_\epsilon = 5K$ without stopping criteria enabled. To generate a set of HoDS features to train GWAD-CIFAR10, each attack method performs the attack on the pre-trained ResNet-18, with 10 randomly selected examples from the training split. From each attack, 150 HoDS features are extracted at random points in the sequence of queries. Therefore, each attack method generates 1500 HoDS features during its attacks on 10 examples. As a result, 9000 HoDS features are extracted for the six attack classes. Another set of HoDS features to train GWAD-ImageNet is also acquired through the same procedure with the pre-trained VGG-16. We present the visualisation of HoDS features in Figure 5. HoDS feature set to represent the benign class in the training are acquired from the normal distributions instead of the real benign queries. We use four normal distributions, and extract 375 HoDS features from each distribution:

- $\mathcal{N}(0, 0.25)$ and $\mathcal{N}(-0.5, 0.25)$
- $\mathcal{N}(0, 0.14)$ and $\mathcal{N}(-0.5, 0.14)$

The mean and variance of the normal distributions are chosen based on the empirical measure of the $\mathcal{DS}$ distributions. For example, the $\mathcal{DS}$ elements of CIFAR-10 in Figure 2 of the main paper are randomly distributed with $\mu = -0.49447$, and $\sigma^2 = 0.14828$.

5. Dataset and Attack Methods

5.1. Dataset

The experiments of query-based black-box attacks on image classification tasks are conducted over two standard image datasets: CIFAR-10 [16] and ImageNet [9]. All the examples of these datasets are transformed as instructed by Pytorch torchvision library [14]. We present further details of image datasets used in the experiments with benign image queries to simulate the practical use cases as follows.

Tiny-ImageNet [18] is composed of 200 classes of images. Images in the dataset are downsized to 64×64 coloured image space. The dataset is widely used for training and testing various machine learning techniques.

Hollywood Heads [20] is a dataset containing human heads annotated in sequential Hollywood movie frames. As objects are in a single class and extracted from the sequential movie frames, examples tend to exhibit high similarity with neighbouring examples.

FLIR ADAS[11] provides thermal and visible band images for the development of automated systems using modern

Variance Bound	Detection Rate		
	HSJA	NES	Sign-Flip
$0.0 \leq \alpha \leq 1.5$	100.00%	100.00%	99.30%
$0.0 \leq \alpha \leq 2.0$	100.00%	100.00%	99.28%
$0.0 \leq \alpha \leq 2.5$	100.00%	99.98%	99.39%
$0.0 \leq \alpha \leq 3.0$	100.00%	100.00%	99.42%

Table 4. GWAD attack detection performances over the sequence of queries from varying-variance adaptive attacks with HSJA [6], NES [15] and Sign-Flip [7]

DNN models. The dataset was acquired via camera system mounted on a vehicle, and includes images captured in streets and highways in California, USA.

BIRDSAI[3] is an infra-red image dataset specifically designed for Surveillance system with Aerial Intelligence. The dataset was acquired through a long-wave band thermal camera system, and contains night-time images of animals and humans in Southern Africa.

5.2. Attack Methods

We note that all the attack methods used in the experiments are commonly configured to the untargeted setting with l_2 constraint. In the following, we provide the hyper-parameter settings of the methods used in the experiments.

BA[4] : Optimisation steps are initialised with 0.01, and updated every 10 iterations with a learning rate $\epsilon = 1.5$.

HSJA[6] : 100 queries are used to find an initial adversarial example. Binary search threshold θ is set to $0.01/\sqrt{w \times h \times c}$, where $w \times h \times c$ is the area of input space.

SimBA[12] : Attack step size ϵ is set to 0.03 for CIFAR-10, and 0.2 for ImageNet.

Sign-Opt[8] : Gradient search learning rates are initialised to 0.001 for CIFAR-10, and 0.05 for ImageNet. Line search learning rates are set to 2 and 0.25 for further optimisation. Finally, convergence threshold is 255 and 5 for CIFAR-10 and ImageNet respectively.

NES[15] : Number of samples n for the gradient estimate is 50. Learning rate η is set to 0.55 for CIFAR-10 and 2.55 for ImageNet. Search variance σ is 0.1.

Sign-flip[7] : Project step parameter α is initialised to 0.0004 and updatd with a rate of 1.5. Random sign flip step parameter p is initialised to 0 and updated with a step of 0.001.

Query-based attack methods spend the majority of their attack queries on zeroth-order optimisations. Table 3 shows the query consumption profile of six SOTA query-based black-box attack methods based on the settings detailed above.

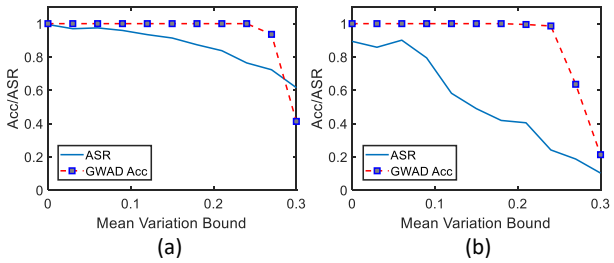

Figure 6. ASR of varying-mean adaptive attacks and detection rates of GWAD as the upper bound $\mathbf{r}_\mu$ of μ variation in random distributions grows: (a) HSJA; (b) NES

6. Additional Experiments

6.1. Moving target attack

In this section, we consider an attacker that implements a "moving-target" strategy to evade detection, while still generating adversarial examples and querying the victim model. The idea is to manipulate the parameters of the noise distribution from which the attack samples the noise.

Noise vectors u_t used in zeroth-order optimisation need to be the zero-mean random distributions with a common variance such as $\mathcal{N}(0, 1)$ to guarantee acceptable quality of attacks (QoA). However, one adaptive attack, with full knowledge of the proposed detection scheme, may introduce a variation in the random distributions by varying μ or σ^2, with no serious consideration of QoA.

Varying-variance attack : We first conduct attacks with varying variance of random distributions as presented in Table 4. In this setting, generated random noises are scaled by a factor α, where $0 \leq \alpha \leq \mathbf{r}_\sigma$. While the attack performance of NES has gradually degraded, GWAD maintains the attack detection performance showing near perfect detection rates against varying variance adaptive attacks, based on HSJA, NES and Sign-Flip method, across all the variations in σ^2.

Varying-mean attack: We now consider the scenario where the adversary varies the mean μ of random vectors. The attack first finds the range of pixel intensities for an input x as $s = \max(x) - \min(x)$, and sets the bounds of variation as a ratio $\mathbf{r}_\mu$ of s. We display the impacts on ASR caused by this adaptive attack setting in Figure 6. The ASR of such attacks are gradually decreased as the rate of variation grows. The adaptive attacks with HSJA and NES show only 61.54% and 10.2% ASRs respectively at $\mathbf{r}_\mu = 0.30$. In contrast, GWAD maintains the robustness in binary classification performance until $\mathbf{r}_\mu = 0.24$ achieving 100% detection rates against both attacks.

	HSJA	NES	SimgBA	Sign-Opt	Sign-Flip	BA
SVM	100.0%	100.0%	100.0%	84.1%	99.7%	94.9%
kNN	89.4%	100.0%	100.0%	93.7%	92.6%	85.6%

Table 5. Non-NN models' classification performance (ImageNet)

Number of	HSJA		Sign-Flip	
Queries	Recogn.	Detect.	Recogn.	Detect.
16	91.7%	100.0%	85.1%	99.4%
32	90.8%	100.0%	87.4%	99.6%
64	98.8%	100.0%	95.8%	99.5%
128	99.9%	100.0%	98.3%	99.3%
256	**100.0%**	100.0%	**99.8%**	99.8%

Table 6. Effect of number of queries to form HoDS for GWAD prediction and detection performance over HSJA and Sign-Flip attacks. "Recogn." corresponds to the attack recognition, and "Detect." corresponds to the binary classification (benign/attack).

6.2. Other Classifier architectures

We explore conventional classifier architectures to evaluate the discriminant power of HoDS features. Among these architectures, the classification performances of two non-DNN classifiers, SVM and kNN, are presented in Table 5. We note that the kNN classifier is directly applied with the HoDS-Mean of CIFAR-10 depicted in Figure 5 at $k = 256$.

6.3. Ablation Studies

GWAD requires a set of queries to predict and detect query-based attack. In Table 6, we report the attack classification and detection performance of GWAD over various number of queries to make HoDS feature. GWAD monitors 10K attack queries from HSJA [6] and Sign-Flip [7]. The classification success rate of GWAD is gradually improved as the number of queries to form HoDS grows reaching almost 100% classification accuracy at 256 queries.

References

[1] A. S. Bandeira. Lecture notes for mathematics of data science (401-4944-201 at eth zurich). *ETH Zurich*, 2020. 1

[2] A. Blum, J. Hopcroft, and R. Kannan. Foundations of data science. *Cambridge University Press*, 2020. 1

[3] E. Bondi, R. Jain, P. Aggrawal, S. Anand, R. Hannaford, A. Kapoor, J. Piavis, S. Shah, L. Joppa, B. Dilkina, and Milind M. Tambe. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 4

[4] W. Brendel, J. Rauber, and M. Bethge. Decision based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations*, 2018. 3, 4

[5] J. Canny. A computational approadh to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, Issue 6.*, 1986. 2

[6] J. Chen, M. I. Jordan, and M. J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. *IEEE Symposium on Security and Privacy*, 2020. 2, 3, 4, 5

[7] W. Chen, Z. Zhang, X. Hu, and B. Wu. Boosting decision-based black-box adversarial attacks with random sign flip. *European Conference on Computer Vision*, 2020. 3, 4, 5

[8] M. Cheng, S. Singh, P. Chen, P.Y. Chen, S. Liu, and C.J. Hsleh. Sign-opt: A query-efficient hard-label adversarial attack. *International Conference on Learning Representations*, 2020. 3, 4

[9] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 3

[10] M. J. Evans, J. S. Rosenthal, and W. H. Freeman. Probability and statistics: The science of uncertainty. *University of Toronto, pp. 204-213*, 2004. 1

[11] FLIR. Free teledyne flir thermal dataset for algorithm training. *https://www.flir.co.uk/oem/adas/adas-dataset-form/*, 2024. 3

[12] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger. Simple black-box adversarial attacks. *International Conference on Learning Representations*, 2019. 3, 4

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association. 58 (301): 13–30*, 1963. 1

[14] https://pytorch.org/vision/stable/models.html. Accessed 10 june 2024. 3

[15] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. *International Conference on Machine Learning*, 2018. 2, 3, 4

[16] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Techinical Report, Citeseer*, 2009. 3

[17] M. Ladoux. The concentration of measure phenomenon. *American Mathematical Society*, 2001. 1

[18] Y. Le and X. S. Yang. Tiny imagenet visual recognition challenge. *https://api.semanticscholar.org/CorpusID:16664790*, 2015. 3

[19] A. Neeman P. Hall, J. S. Marron. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology, Vol. 67, Issue 3, pp. 427-444*, 2005. 2

[20] T. H. Vu, A. Osokin, and I. Laptev. Context-aware cnns for person head detection. *IEEE/CVF International Conference on Computer Vision*, 2015. 3