Supplementary Material: On the Effectiveness of Small Input Noise for Defending Against Query-based Black-Box Attacks

Junyoung Byun, Hyojun Go, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{bjyoung, gohyojun15, changick}@kaist.ac.kr

1. Illustration of Defense Against Decision-based Attack with Small Input Noise

We illustrate the working principle of SND against decision-based attacks in Fig. 1.

2. Detailed Experimental Settings

2.1. Settings of Attack Methods

BA: We adopt BA provided by Adversarial Robustness Toolbox (ART) library [2] with default parameters.

Sign-OPT: We take the code[1] provided by the authors without changing the special parameters of the attack.

HSJA: We adopt HSJA provided by the ART library with default parameters except for increasing the maximum number of iterations to 64 to follow the authors' code[2].

GeoDA: We take the code[3] provided by the authors without changing the special parameters of the attack.

SimBA & SimBA-DCT: We take the code[4] provided by the authors. Following the authors, we use `freq_dims`=28, `order`=strided, and `stride`=7 for SimBA-DCT.

Bandit-TD: We take the code[5] provided by the authors with default parameters except for `batch_size`=1 and `epsilon`=4.9.

Subspace Attack: We take the code[6] provided by the authors with the original settings for the ℓ_∞ norm untargeted attacks for the ImageNet. We use the pre-trained ResNet-18 and ResNet-34 trained on the `imagenetv2-val` dataset as reference models that the authors provide.

2.2. Settings of Defense Methods

Baseline: We trained a ResNet-20 model on the CIFAR-10 dataset for 200 epochs and used this model for our experiments. For the ImageNet dataset, we used the pre-trained ResNet-50 model provided by TorchVision library[7].

[1]https://github.com/cmhcbb/attackbox
[2]https://github.com/Jianbo-Lab/HSJA
[3]https://github.com/thisisalirah/GeoDA
[4]https://github.com/cg563/simple-blackbox-attack
[5]https://github.com/MadryLab/blackbox-bandits
[6]https://github.com/ZiangYan/subspace-attack.pytorch

Dataset	CIFAR-10			
# of queries	2K×T	5K×T	10K×T	P_{mis}
BA (T=10)	20.3%	28.5%	32.9%	0.413
SO (T=10)	20.5%	21.1%	21.9%	0.354
HSJA (T=10)	21.0%	27.3%	34.9%	0.376
GeoDA (T=10)	12.1%	12.2%	12.2%	0.413
Dataset	ImageNet			
# of queries		5K×T	10K×T	20K×T
SimBA-DCT (T=10)		8.8%	9.6%	11.2%
Bandit-TD (T=10)		8.8%	8.8%	9.2%

Table 1: Attack success rates of the adaptive attacks against SND ($\sigma = 0.01$) with T=10.

PNI: We used the pre-trained ResNet-20 model trained on the CIFAR-10 dataset with PNI-W (channel-wise) provided by the authors.

PGD-AT: We used the adversarilly trained ResNet-50 model for the ℓ_2 norm with $\epsilon_{train} = 3$ provided from *robustness* library [1] with PGD on the ImageNet dataset for comparisons.

RSE: We trained a RSE-based ResNet-20 with $\sigma_{\text{init}} = 0.2$ and $\sigma_{\text{inner}} = 0.1$. Considering computational efficiency, we used 5 ensembles for each prediction of RSE.

R&P: R&P applies random resizing and random padding to its input sequentially. It first rescales an input image of size $W \times H \times 3$ with a scale factor s which is sampled from $[s_{min}, s_{max}]$, and places it in a random position within an empty image of size $s_{max}W \times s_{max}Y \times 3$. Following the authors, we set s_{min} and s_{max} as $\frac{310}{299}$ and $\frac{331}{299}$ respectively.

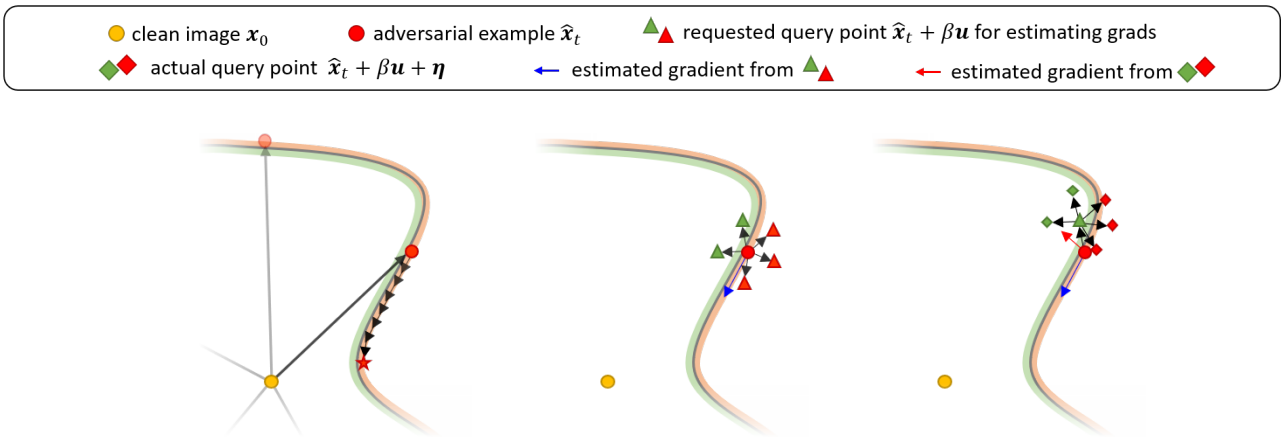

Figure 1: Illustrations of how small noise can defend against decision-based attacks (left) An adversary wants to reach the optimal adversarial example from an initial misclassified image. (middle) To find the next update's direction, it evaluates $\hat{x}_t + \beta u$ (right) Small noise can disturb this gradient estimation. The illustration shows that the prediction for each query image can have different values because of the small input noise.

3. Evaluation of Adaptive Attacks Against SND

We apply the expectation-based adaptive attack to BA, SO, and GeoDA on CIFAR-10 and two score-based attacks (SimBA-DCT and Bandit-TD) on ImageNet for comprehensive comparisons. The experimental results are shown in Table 1.

4. Evaluation of Varying σ for Each Inference

Detailed experimental results are shown in Table 2.

5. Evaluation of Attack Success Rates of Subspace Attack

Detailed experimental results are shown in Table 3.

References

[1] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[2] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

[7] https://github.com/pytorch/vision

	Clean Acc. (%)	$\mathbb{E}\|\|\boldsymbol{\eta}\|\|_2$	Sign-OPT			HSJA		
# of queries	-	-	2K	5K	10K	2K	5K	10K
σ=0.001	91.33 ± 0.02	0.055	20.6%	22.2%	23.4%	48.1%	67.6%	81.9%
σ=0.01	90.57 ± 0.09	0.550	21.7%	22.3%	22.8%	16.5%	19.9%	22.7%
σ=0.01, α=β=1	91.04 ± 0.12	0.276	21.6%	22.4%	22.9%	18.1%	23.9%	30.0%
σ=0.01, α=β=2	91.15 ± 0.04	0.275	20.3%	20.8%	21.7%	19.7%	23.5%	28.5%
σ=0.01, α=β=0.5	91.06 ± 0.06	0.275	20.9%	21.4%	22.4%	19.8%	26.3%	32.6%
σ=0.02	87.56 ± 0.18	1.098	26.4%	26.5%	26.6%	24.2%	26.7%	30.1%
σ=0.02, α=β=1	90.17 ± 0.09	0.552	22.0%	22.0%	22.1%	19.6%	23.2%	25.4%
σ=0.02, α=β=2	90.44 ± 0.08	0.550	21.6%	21.8%	22.1%	18.0%	22.1%	25.0%
σ=0.02, α=β=0.5	89.99 ± 0.24	0.549	22.5%	22.5%	23.3%	20.3%	24.4%	27.5%

Table 2: Experimental results of varying σ with the CIFAR-10 dataset. We evaluate the mean and standard deviation of clean accuracy in 5 repetitive experiments on the original test dataset.

Attack method	Subspace Attack		
# of queries	5K	10K	20K
Baseline	99.6% (99.6%)	100.0% (100.0%)	100.0% (100.0%)
SND ($\sigma = 0.01$)	61.2% (59.6%)	61.2% (59.6%)	61.2% (59.6%)
SND ($\sigma = 0.001$)	64.4% (64.4%)	66.4% (66.4%)	68.4% (68.4%)
PGD-AT	71.6% (45.6%)	78.4% (52.4%)	81.6% (55.6%)
PGD-AT + SND ($\sigma = 0.01$)	40.8% (14.8%)	42.0% (16.0%)	42.4% (16.4%)
PGD-AT + SND ($\sigma = 0.001$)	62.0% (36.0%)	62.4% (36.4%)	63.2% (37.2%)
R&P	73.2% (68.4%)	74.0% (69.2%)	74.0% (69.2%)

Table 3: Evaluation of attack success rates of Subspace Attack against defenses on the ImageNet dataset. We also calculate the attack success rate without initially misclassified images and denote it in the parenthesis.