# Supplementary material: Exploiting the local parabolic landscapes of adversarial losses to accelerate black-box adversarial attack

Hoang Tran[1], Dan Lu[1], and Guannan Zhang[1]*

Oak Ridge National Laboratory, Oak Ridge TN 37830, USA
{tranha,lud1,zhangg}@ornl.gov

In this supplementary section, we include the following materials:

1. Additional experimental results, where we compare BABIES with two baseline methods, PPBA [2] and PRGF [1] (Section 1),
2. Ablation study showing the robustness of our method to the random seed (Section 2) and step size $\varepsilon$ (Section 3),
3. Scatter plots showing the accuracy of the loss values approximated by parabolas for additional image classifiers (Section 4).

## 1 Comparison with PPBA and PRGF

We compare our method with two recent approaches, i.e., Projection & Probability-driven Black-box Attack (PPBA) and Prior-guided Random Gradient-free (PRGF), using the same setting (e.g., maximum perturbation and maximum queries) and dataset as in our main paper. The code for PPBA and PRGF were acquired from `github.com/theFool32/PPBA` and `github.com/thu-ml/Prior-Guided-RGF`, respectively. We use the hyperparameters suggested by the authors of the methods. In particular, for PRGF, ResNet-v2-152 is used as the surrogate model to provide the transfer gradient. We evaluate two variants of PRGF, i.e., with biased sampling (PRGF-BS) and with gradient averaging (PRGF-GA), incorporated with data-dependent prior. Since the Github repositories of PPBA and PRGF only provide codes and hyperparameters for ImageNet untargeted attacks, we only evaluate them in those cases. The performance of PPBA and PRGF can be compared directly with other baselines in our main evaluation for the untargeted tests (Tables 3 and 5 in the paper). We reproduce those here for reader's convenience.

The results are shown **Table 1**. We observe that on successful attacks, PPBA and PRGF require fewer queries than BABIES. On standard models, the average and median queries of PRGF are the best, with PPBA and Square-attack slightly trailing behind, while PPBA performs better than PRGF in robust models. However, the advantage of both baselines in the number of queries metric is somehow offset by their low success rates. Here, our BABIES algorithm leads by a remarkable margin: 10% on two standard models, and 20%-30% on robust

* Corresponding author

models. Our experiment suggests that by exploiting the successful past steps or gradients from surrogate models to guide the attacks, PPBA and PRGF can save queries in many examples. On the other hand, they may struggle in many others, perhaps because the searching directions are narrowed down, and prior information is not always useful (e.g., when the surrogate model behaves very different from target model), thus degrading the overall success rates.

| | Inception v3 **(untargeted)** | | | ResNet50 **(untargeted)** | | | ResNet18, eps=3 **(untargeted)** | | | ResNet50, eps=3 **(untargeted)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Attack** | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR |
| Bandits | 2200 | 1118 | 66.6% | 1238 | 672 | 96.6% | 2745 | 1660 | 54.3% | 3952 | 3558 | 43.6% |
| Square-attack | 1543 | 777 | 87.8% | 1130 | 563 | 97.3% | **1027** | **209** | 48.9% | **1614** | **758** | 33.4% |
| SimBA-DCT | 1897 | 1183 | 78% | 1673 | 1115 | 94.4% | 4563 | 3839 | 50.7% | 5554 | 5160 | 43.2% |
| PPBA | 1546 | 676 | 75.7% | 1384 | 683 | 89.9% | 1839 | 1062 | 51.1% | 2560 | 1791 | 38.4% |
| PRGF-BS | **1124** | 402 | 76.2% | 1125 | **338** | 84.2% | 1916 | 1140 | 47.6% | 2725 | 2114 | 32% |
| PRGF-GA | 1159 | **364** | 79.6% | **961** | 346 | 86.4% | 1923 | 1336 | 50.4% | 2699 | 1996 | 35% |
| BABIES-DCT | 1907 | 908 | **87.9%** | 1276 | 742 | **98%** | 2553 | 1594 | **71.8%** | 3348 | 2898 | **63.2%** |

**Table 1.** Comparison of PPBA and PRGF with BABIES on untargeted attacks against four standard and robust models for ImageNet. PPBA and PRGF require fewer queries than BABIES on successful attacks, but also have significantly lower success rates. (The results of other baselines are reproduced from Table 3 and 5 in the paper for reference).

## 2 Influence of the random seed on our algorithm

Since our algorithm is essentially a random search algorithm, it is necessary to demonstrate that the performance of our method does not vary dramatically with the change of the random seed. To this end, we use the case of *untargeted attack on standard ImageNet classifiers* (Inception_v3 and ResNet50) to test the robustness of our algorithm with respect to the random seed. We attack a set of 200 images from the ImageNetV2 and run our algorithm with 20 randomly generated random seeds. All other settings are set the same as in the Experimental Evaluation in the main manuscript. The testing results are given in **Table 2**. We can see that our algorithm performs stably when changing the random seed. The success rate varies within 0.3% for ResNet50 and 2.5% for Inception_v3. The maximum and minimum numbers for both Avg.QY and Med.QY vary around $5\% \sim 7\%$ of the mean values, where the standard deviation of those quantities are smaller than 10%. Thus, our idea of exploiting the parabolic landscape of loss to accelerate random search is a statistically effective approach.

## 3 Influence of step size $\varepsilon$ on our algorithm

To emphasize that the performance of BABIES also does not change dramatically with $\varepsilon$, we provide results for BABIES with additional values of $\varepsilon$. **Table 3** here extends Table 2 in the main manuscript (Comparison on attacks against standard models for CIFAR-10), where results of BABIES with $\varepsilon = 1$ and $\varepsilon = 1.4$ are added. **Bold** numbers denote the best overall performance and *italic*

| **Inception_v3** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Avg.QY | | | Med.QY | | | SR | | |
| Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| 2084 | 1999 | 2196 | 968 | 907 | 1074 | 91.1% | 90.2% | 92.5% |
| **ResNet50** | | | | | | | | |
| Avg.QY | | | Med.QY | | | SR | | |
| Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| 1240 | 1182 | 1293 | 671 | 634 | 709 | 99.5% | 99.3% | 99.6% |

**Table 2.** Results on influence of the random seed (untargeted attacks on ImageNet)

numbers denote a better performance of BABIES against baseline methods. For the untargeted attacks, our method still shows the significantly lower median queries when reducing $\varepsilon$ from 2 to 1 with competitive success rate. For the targeted attacks, BABIES achieves the lowest number of queries for $\varepsilon = 1.4$ and $\varepsilon = 2$ and the best success rate in all cases.

In **Table 4**, we show the comparison on attacks against the $\ell_2$-robust models on a set of 100 correctly labeled images from the ImageNetV2. We perform BABIES with $\varepsilon = 4, 8$ and 10. We see that changing $\varepsilon$ does not affect the comparative performance of BABIES to other methods. Our method consistently leads in success rate by a large margin in three out of four cases. In the remaining case (targeted ResNet50), it is comparable to Bandits but requires much fewer queries for all considered $\varepsilon$. The performance of all the approaches here also agrees with Table 5 in the paper (same test on a larger sample set).

| | Inception v3 **(untargeted)** | | | Inception v3 **(targeted)** | | | VGG13 **(untargeted)** | | | VGG13 **(targeted)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Attack** | Avg. | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR |
| Bandits | 409 | 46 | 94.5% | 817 | 314 | 82.9% | 199 | 46 | 99.6% | 601 | 252 | 97.7% |
| Square-attack | **170** | 41 | 97.5% | 345 | 91 | 89.4% | 96 | 28 | **99.9%** | **130** | 51 | 99.3% |
| SimBA | 203 | 177 | 65.9% | 603 | 486 | 46% | 184 | 140 | 83.2% | 492 | 356 | 86.9% |
| BABIES ($\varepsilon = 2$) | 329 | *17* | **97.9%** | **239** | **62** | **97.1%** | *94* | **5** | 99.2% | 159 | **47** | *99.3%* |
| BABIES ($\varepsilon = 1.4$) | 213 | **15** | 96.1% | *280* | *66* | *95.5%* | **79** | *6* | 99.3% | 177 | 79 | *99.3%* |
| BABIES ($\varepsilon = 1$) | 203 | *19* | 95.3% | 357 | 132 | *92.1%* | 115 | *18* | 99.4% | 246 | 107 | **99.5%** |

**Table 3.** Performance of BABIES with difference choices of step size $\varepsilon$ compared to other baselines on attacks against the standard models for CIFAR-10.

## 4 Additional illustration on the accuracy of approximated loss values

We provide in Figure 1 scatter plots showing the correlation between true and approximated loss values given by parabolas for three additional classifiers: Small-CNN on MNIST, Inception_v3 on CIFAR and ResNet50 on ImageNet. Each plot

| | ResNet18, eps=3 **(untargeted)** | | | ResNet18, eps=3 **(targeted)** | | | ResNet50, eps=3 **(untargeted)** | | | ResNet50, eps=3 **(targeted)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Attack** | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR | Avg. QY | Med. QY | SR |
| Bandits | 2875 | 2028 | 51% | 14971 | 14135 | 36% | 4094 | 3256 | 43% | 17101 | 12966 | 86% |
| Square-attack | **1387** | **594** | 51% | 14279 | 8280 | 55% | **2033** | **1156** | 37% | 11798 | 7940 | 48% |
| SimBA-DCT | 3028 | 2954 | 54% | 17474 | 17585 | 40% | 3871 | 3708 | 44% | 25397 | 27635 | 28% |
| BABIES ($\varepsilon = 8$) | 3107 | 2069 | **73%** | **7291** | **4597** | *82%* | 3537 | 2763 | **63%** | **8967** | **5795** | *86%* |
| BABIES ($\varepsilon = 4$) | 3115 | 2073 | **73%** | *9251* | *5774* | **83%** | 3431 | 2667 | *62%* | *9435* | *6221* | **88%** |
| BABIES ($\varepsilon = 10$) | 2757 | 1784 | *68%* | *8530* | *5492* | *81%* | 3200 | 2458 | *57%* | *9589* | *6032* | *86%* |

**Table 4.** Performance of BABIES with difference choices of $\varepsilon$ compared to other baselines on attacks against the $\ell_2$-robust models for ImageNet.

is generated using 5000 random points in the neighborhood of 50 images (described in detail in the main manuscript, Section 3). Our observation here is consistent with that in the main paper, that the correlation between the true and approximated loss values is strong in DCT setting, yielding that the adversarial losses can be well-approximated by parabolas in the frequency directions, but much less so in the pixel directions.

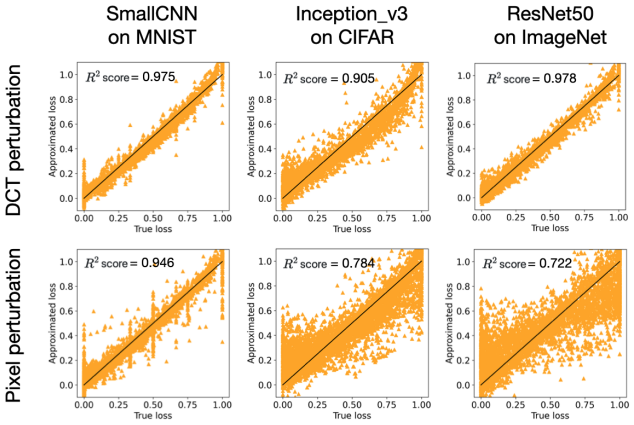


**Fig. 1.** Scatter plot displays the correlation between true and approximated loss values on 5000 random points, sampled from 5000 segments along DCT directions **(top)** and pixel directions **(bottom)**. This plot extends our illustration in Figure 3 (main paper) for three additional image classifiers.

## References

1. Dong, Y., Cheng, S., Pang, T., Su, H., Zhu, J.: Query-efficient black-box adversarial attacks guided by a transfer-based prior. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3126733
2. Li, J., Ji, R., Liu, H., Liu, J., Zhong, B., Deng, C., Tian, Q.: Projection and probability-driven black-box attack. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)