DEPENDABLE AND SECURE AI-ML (AI60006)

Assignment 1

Abhinav Bohra **18CS30049** horizontal line

# 1. Constructing adversarial examples

**Models used:**  ResNet-50 **Dataset used:** CIFAR

|  |  |
| --- | --- |
| **Before Adding Noise** | **After Adding Noise** |
|  |  |
| **Prediction:** Pomegranate | **Prediction:** Candle |

# 2. Adversarial Training

# Models used: We used two different architectures to train our model: a 2-layer fully connected MLP, (referred to as **DNN** in the report), and a 6-layer ConvNet, (referred to as **CNN** in the report)

**Dataset used:** MNIST

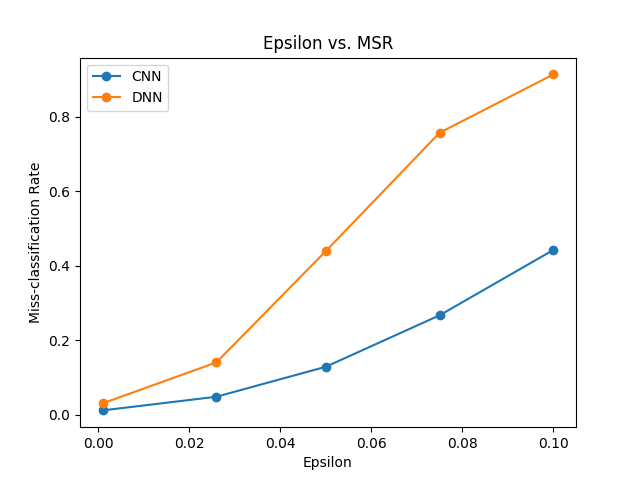
# **2.1 Original model outputs without perturbations**

|  |  |  |
| --- | --- | --- |
| **Model** | **Miss-classification Rate** | **Output Screenshot** |
| DNN | 0.0279 | original_preds_dnn |
| CNN | 0.0106 | original_preds_cnn |

**2.2 Fast Gradient Sign Method (FGSM)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Epsilon** | **Model** | **Miss-classification Rate** | **Output\_Screenshot** |
| 0.1 | CNN | 0.4423 |  |
| 0.075 | CNN | 0.2667 |  |
| 0.05 | CNN | 0.1284 |  |
| 0.026 | CNN | 0.048 |  |
| 0.001 | CNN | 0.0113 |  |
| 0.1 | DNN | 0.9137 |  |
| 0.075 | DNN | 0.757 |  |
| 0.05 | DNN | 0.4388 |  |
| 0.026 | DNN | 0.1403 |  |
| 0.001 | DNN | 0.0301 |  |

**Analysis of impact of *epsilon* on miss-classification rate**

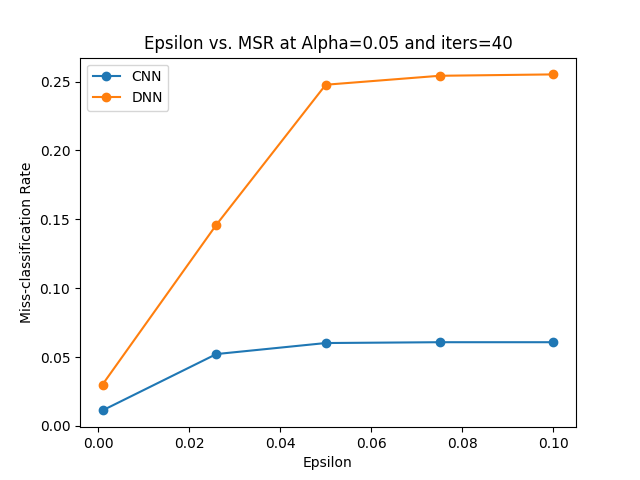
***Observation:*** It is observed that miss classification rate (MSR) increases with increase in epsilon value for both the models. The rate of change of MSR with respect to epsilon value (slope) is higher for DNN as compared to CNN. Thus we infer that FGSM attack method impacts DNN more than CNN.

**2.3 Projected gradient descent(PGD)**

**2.3.1 Analysis of impact of *epsilon* on miss-classification rate**

First, we analyzed the impact of epsilon on miss-classification rate (MSR) by keeping alpha constant (0.05) and the number of iterations at 40.

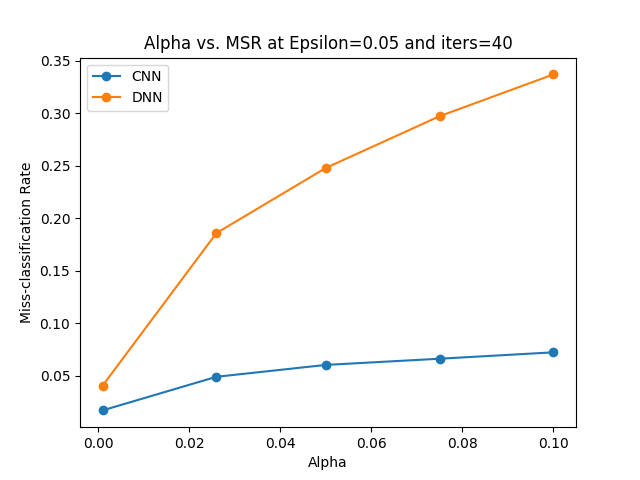
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.1 | 0.05 | 40 | CNN | 0.0608 |  |
| 0.075 | 0.05 | 40 | CNN | 0.0608 |  |
| 0.05 | 0.05 | 40 | CNN | 0.0602 |  |
| 0.026 | 0.05 | 40 | CNN | 0.0522 |  |
| 0.001 | 0.05 | 40 | CNN | 0.0113 |  |
| 0.1 | 0.05 | 40 | DNN | 0.2552 |  |
| 0.075 | 0.05 | 40 | DNN | 0.2542 |  |
| 0.05 | 0.05 | 40 | DNN | 0.2477 |  |
| 0.026 | 0.05 | 40 | DNN | 0.1461 |  |
| 0.001 | 0.05 | 40 | DNN | 0.0301 |  |

***Observation:*** It is observed that miss classification rate (MSR) increases with increase in epsilon value for both the models. The rate of change of MSR with respect to epsilon value (slope) is higher for DNN as compared to CNN. Thus we infer that PGD attack method impacts DNN more than CNN.

**2.3.2 Analysis of impact of *alpha* on miss-classification rate**

Then, we analyzed the impact of alpha on miss-classification rate by keeping epsilon constant (0.05) and the number of iterations at 40.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.05 | 0.1 | 40 | CNN | 0.0722 |  |
| 0.05 | 0.075 | 40 | CNN | 0.0661 |  |
| 0.05 | 0.05 | 40 | CNN | 0.0602 |  |
| 0.05 | 0.026 | 40 | CNN | 0.0489 |  |
| 0.05 | 0.001 | 40 | CNN | 0.0169 |  |
| 0.05 | 0.1 | 40 | DNN | 0.3368 |  |
| 0.05 | 0.075 | 40 | DNN | 0.297 |  |
| 0.05 | 0.05 | 40 | DNN | 0.2477 |  |
| 0.05 | 0.026 | 40 | DNN | 0.1857 |  |
| 0.05 | 0.001 | 40 | DNN | 0.0404 |  |

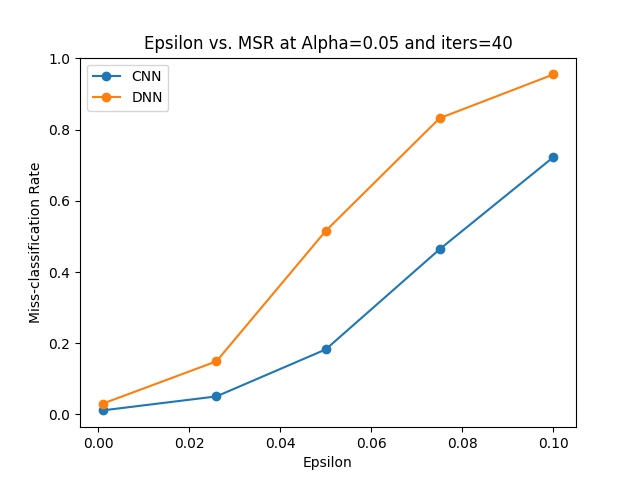
***Observation:*** It is observed that miss classification rate (MSR) increases with increase in alpha value for both the models. The rate of change of MSR with respect to alpha value (slope) is higher for DNN as compared to CNN. Thus we infer that PGD attack method impacts DNN more than CNN.

**2.4 Projected gradient descent with infinity norm**

**2.4.1 Analysis of impact of *epsilon* on miss-classification rate**

First, we analyzed the impact of epsilon on miss-classification rate by keeping alpha constant (0.05) and the number of iterations at 40.

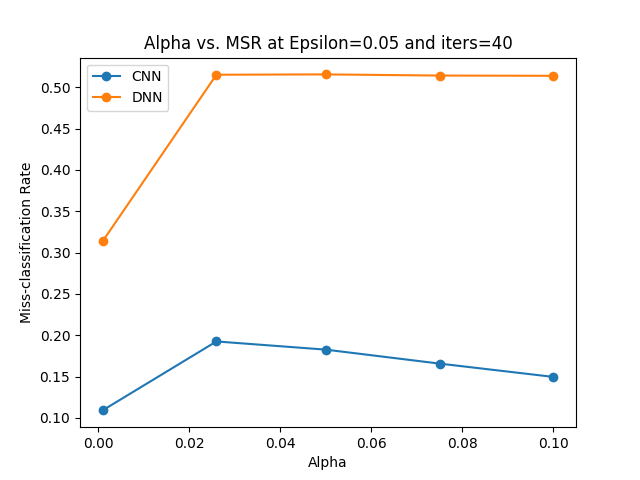
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.1 | 0.05 | 40 | CNN | 0.723 |  |
| 0.075 | 0.05 | 40 | CNN | 0.4639 |  |
| 0.05 | 0.05 | 40 | CNN | 0.1826 |  |
| 0.026 | 0.05 | 40 | CNN | 0.0506 |  |
| 0.001 | 0.05 | 40 | CNN | 0.0113 |  |
| 0.1 | 0.05 | 40 | DNN | 0.9552 |  |
| 0.075 | 0.05 | 40 | DNN | 0.8321 |  |
| 0.05 | 0.05 | 40 | DNN | 0.5155 |  |
| 0.026 | 0.05 | 40 | DNN | 0.1491 |  |
| 0.001 | 0.05 | 40 | DNN | 0.0301 |  |

***Observation:*** It is observed that miss classification rate (MSR) increases with increase in epsilon value for both the models. The rate of change of MSR with respect to epsilon value (slope) is higher for DNN as compared to CNN. Thus we infer that PGD attack method impacts DNN more than CNN.

**2.4.2 Analysis of impact of *alpha* on miss-classification rate**

Then, we analyzed the impact of alpha on miss-classification rate by keeping epsilon constant (0.05) and the number of iterations at 40.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.05 | 0.1 | 40 | CNN | 0.1496 |  |
| 0.05 | 0.075 | 40 | CNN | 0.1657 |  |
| 0.05 | 0.05 | 40 | CNN | 0.1826 |  |
| 0.05 | 0.026 | 40 | CNN | 0.1925 |  |
| 0.05 | 0.001 | 40 | CNN | 0.1091 |  |
| 0.05 | 0.1 | 40 | DNN | 0.5138 |  |
| 0.05 | 0.075 | 40 | DNN | 0.5141 |  |
| 0.05 | 0.05 | 40 | DNN | 0.5155 |  |
| 0.05 | 0.026 | 40 | DNN | 0.5151 |  |
| 0.05 | 0.001 | 40 | DNN | 0.3135 |  |

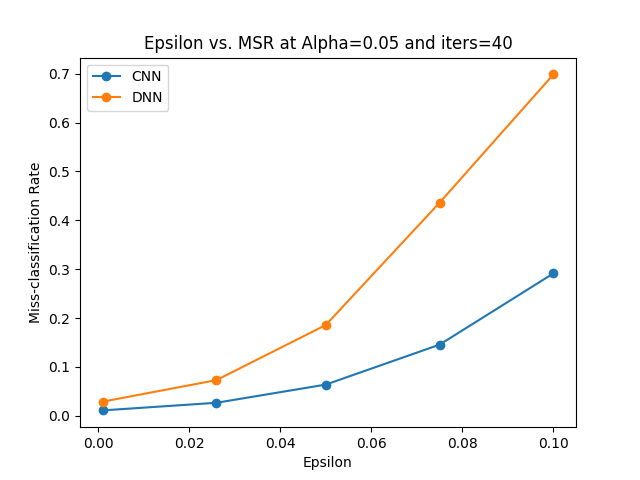
 ***Observation:*** It is observed that miss classification rate (MSR) first increases and then plateaus (after alpha = 0.26) with increase in alpha value for DNN, on the other hand it starts decreasing (after alpha = 0.26) for CNN.

**2.5 Projected gradient descent infinity norm with targeted attack**

**2.5.1 Analysis of impact of *epsilon* on miss-classification rate**

First, we analyzed the impact of epsilon on miss-classification rate by keeping alpha constant (0.05) and the number of iterations at 40. For analysis we used targ2 function, since its formulation was found to be better. The analysis presented below is for y-targ=0

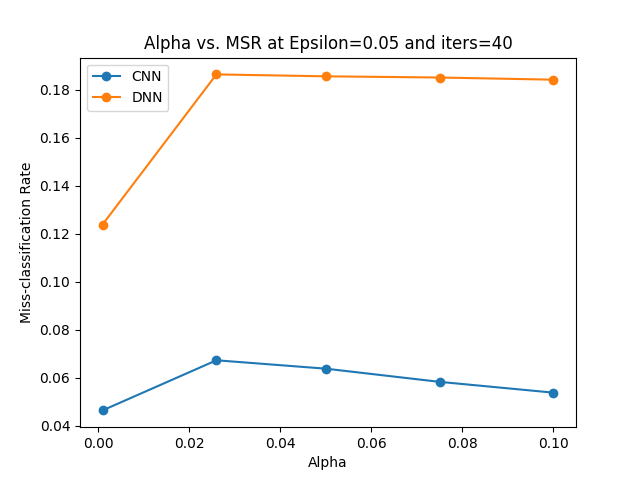
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Y-Targ** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.1 | 0.05 | 40 | 0 | CNN | 0.2914 |  |
| 0.075 | 0.05 | 40 | 0 | CNN | 0.1455 |  |
| 0.05 | 0.05 | 40 | 0 | CNN | 0.0637 |  |
| 0.026 | 0.05 | 40 | 0 | CNN | 0.0265 |  |
| 0.001 | 0.05 | 40 | 0 | CNN | 0.0109 |  |
| 0.1 | 0.05 | 40 | 0 | DNN | 0.6987 |  |
| 0.075 | 0.05 | 40 | 0 | DNN | 0.4361 |  |
| 0.05 | 0.05 | 40 | 0 | DNN | 0.1856 |  |
| 0.026 | 0.05 | 40 | 0 | DNN | 0.0729 |  |
| 0.001 | 0.05 | 40 | 0 | DNN | 0.0287 |  |

***Observation:*** It is observed that miss classification rate (MSR) increases with increase in eplsion value for both the models. However rate of change is observed to be higher for DNN model

**2.5.2 Analysis of impact of *alpha* on miss-classification rate**

Then, we analyzed the impact of alpha on miss-classification rate by keeping epsilon constant (0.05) and the number of iterations at 40. For analysis we used targ2 function, since its formulation was found to be better. The analysis presented below is for y-targ=0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Y-Targ** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.05 | 0.1 | 40 | 0 | CNN | 0.0537 |  |
| 0.05 | 0.075 | 40 | 0 | CNN | 0.0582 |  |
| 0.05 | 0.05 | 40 | 0 | CNN | 0.0637 |  |
| 0.05 | 0.026 | 40 | 0 | CNN | 0.0672 |  |
| 0.05 | 0.001 | 40 | 0 | CNN | 0.0463 |  |
| 0.05 | 0.1 | 40 | 0 | DNN | 0.1842 |  |
| 0.05 | 0.075 | 40 | 0 | DNN | 0.1851 |  |
| 0.05 | 0.05 | 40 | 0 | DNN | 0.1856 |  |
| 0.05 | 0.026 | 40 | 0 | DNN | 0.1864 |  |
| 0.05 | 0.001 | 40 | 0 | DNN | 0.1238 |  |

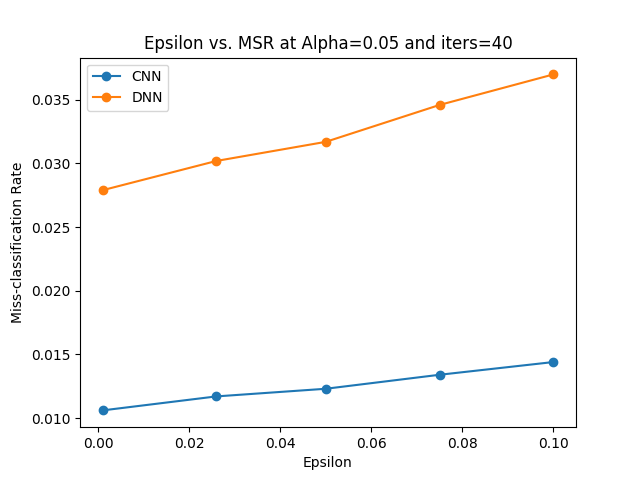
  
***Observation:*** It is observed that miss classification rate (MSR) first increases and then plateaus (after alpha = 0.26) with increase in alpha value for DNN, on the other hand it starts decreasing (after alpha = 0.26) for CNN.

**2.6 Projected gradient descent with l2 norm**

**2.6.1 Analysis of impact of *epsilon* on miss-classification rate**

First, we analyzed the impact of epsilon on miss-classification rate by keeping alpha constant (0.05) and the number of iterations at 40.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.1 | 0.05 | 40 | CNN | 0.0144 |  |
| 0.075 | 0.05 | 40 | CNN | 0.0134 |  |
| 0.05 | 0.05 | 40 | CNN | 0.0123 |  |
| 0.026 | 0.05 | 40 | CNN | 0.0117 |  |
| 0.001 | 0.05 | 40 | CNN | 0.0106 |  |
| 0.1 | 0.05 | 40 | DNN | 0.037 |  |
| 0.075 | 0.05 | 40 | DNN | 0.0346 |  |
| 0.05 | 0.05 | 40 | DNN | 0.0317 |  |
| 0.026 | 0.05 | 40 | DNN | 0.0302 |  |
| 0.001 | 0.05 | 40 | DNN | 0.0279 |  |

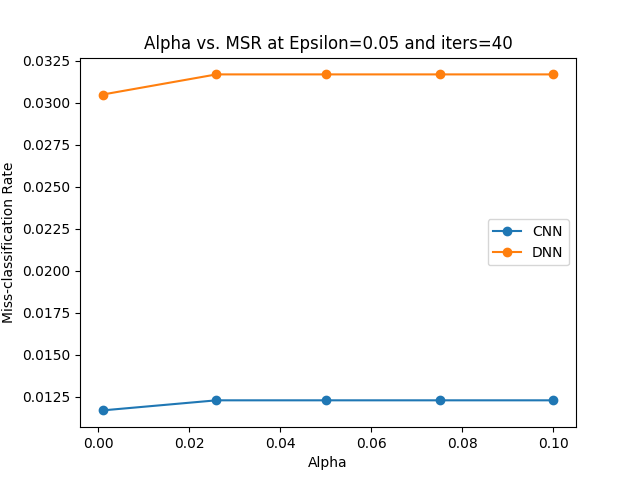


***Observation:*** It is observed that miss classification rate (MSR) increases with increase in epsilon value for both the models. However, the rate of change of MSR with respect to epsilon value (slope) is smaller as compared to that of other attacks for both the models.

**2.6.2 Analysis of impact of *alpha* on miss-classification rate**

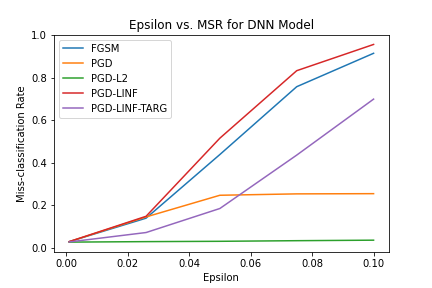
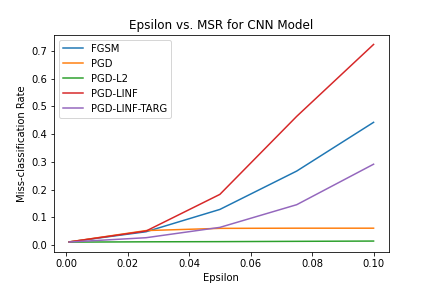
Then, we analyzed the impact of alpha on miss-classification rate by keeping epsilon constant (0.05) and the number of iterations at 40.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Epsilon** | **Alpha** | **Iters** | **Model** | **MSR** | **Output\_Screenshot** |
| 0.05 | 0.1 | 40 | CNN | 0.0123 |  |
| 0.05 | 0.075 | 40 | CNN | 0.0123 |  |
| 0.05 | 0.05 | 40 | CNN | 0.0123 |  |
| 0.05 | 0.026 | 40 | CNN | 0.0123 |  |
| 0.05 | 0.001 | 40 | CNN | 0.0117 |  |
| 0.05 | 0.1 | 40 | DNN | 0.0317 |  |
| 0.05 | 0.075 | 40 | DNN | 0.0317 |  |
| 0.05 | 0.05 | 40 | DNN | 0.0317 |  |
| 0.05 | 0.026 | 40 | DNN | 0.0317 |  |
| 0.05 | 0.001 | 40 | DNN | 0.0305 |  |

 ***Observation:*** It is observed that miss classification rate (MSR) initially increases with increase in alpha value but plateaus (after alpha=0.26) for both the models. The magnitude of MSR is smaller in CNN as compared to that in DNN.

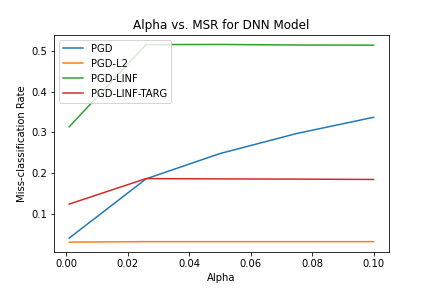
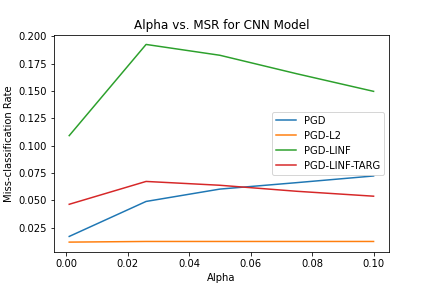
**2.7 Comparison of miss-classification rate relative to all methods**

**2.7.1 Epsilon vs MSR Comparison of all attack methods at fixed alpha (0.05) & iters (40)**



***Observation:*** For both the models, the attack methods follow similar trend with PGD-infinity norm with highest miss-classification rate at any given value of epsilion.

**2.7.2 Alpha vs MSR Comparison of all attack methods at fixed epsilon (0.05) & iters (40)**



***Observation:*** For both the models, the attack methods follow similar trend with PGD-infinity norm with highest miss-classification rate for any given value of alpha. It is noted that all the attack methods except vanilla PGD observe a decreasing trend for alpha > 0.26