Trustworthy Machine Learning - EX1

Ishay Yemini

April 16, 2025

1 White-box vs. query-based black-box attack

1. The benign accuracy of the model is 87.5%.

2. Untargeted success rate: 98.5%.
   Targeted success rate: 94.5%.

3. The success rates for the Black-Box untargeted and targeted attacks are presented in table 1. We can clearly see that these attacks are not as successful as the White-Box attack from the last subsection. This aligns with our expectation, as the White-Box attack is able to use the real gradient, while the Black-Box attack can only use an estimate of it. Additionally, momentum improves the success rate and lowers the number of queries needed, for both untargeted and targeted attacks. This maybe happens because the momentum causes the algorithm to use the gradients from the previous iterations, thereby making the estimation less noisy than without momentum. This allows for better convergence, which results in higher success rates and less queries, as our results show.
   The box plots from our experiment are present in figures 1 and 2.

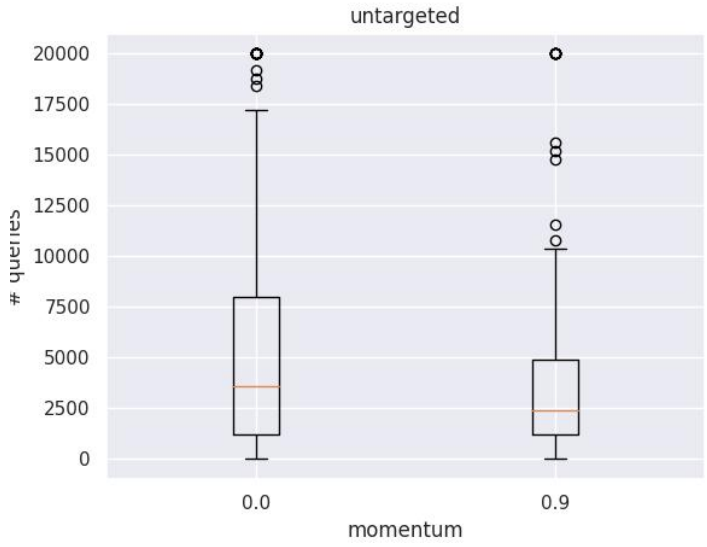

Figure 1: Black-box untargeted number of queries

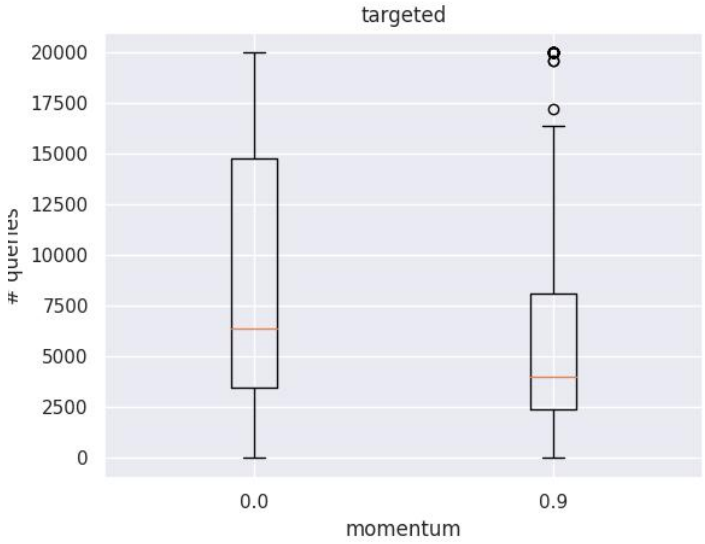

Figure 2: Black-box targeted number of queries

Type	Momentum	Success Rate	Queries
Untargeted	0	93%	3,600
Targeted	0	81%	6,400
Untergeted	0.9	96.5%	2,400
Targeted	0.9	88%	4,000

Table 1: Black-box attacks success rates and number of queries

2 Transferability-based black-box attack.

We can see that untargeted attacks transfer much better than targeted attacks, as presented in tables 2 and 3.
When using an ensemble attack, trained on models 1 and 2, we get much better results: for the untargeted attack we get a success rate of 74%, much better than the 55.5%-54% we got with only one of 1 or 2. Similarly, we for the targeted attack we get a success rate of 50% - almost two times better than the 25.5% we got with only a single model!
This can be explained by the fact that now, the attack is less susceptible to specificities in different models, and by training on two models, this smoothes out the PGD process, thus finding features that generalize better.

		Attacked		
		0	1	2
Trained	0	98.5%	55.5%	54%
	1	68.5%	96.5%	58.5%
	2	60%	54.5%	95.5%

Table 2: Untargeted attacks' transferability.

		Attacked		
		0	1	2
Trained	0	92.5%	25.5%	25.5%
	1	37%	89.5%	27%
	2	32.5%	23.5%	84%

Table 3: Targeted attacks' transferability.

3 Bit-flip attacks

1. The maximum RAD is 74.55%.
2. The fraction of bits that lead to $> 15\%$ RAD when flipped is 0.0213 bits.
3. We can see in figure 3 that bit 1 has the highest median RAD. This makes sense: bit 0 is the sign bit, so bit 1 is the MSB, so flipping it is expected to have the most effect on the parsed float.

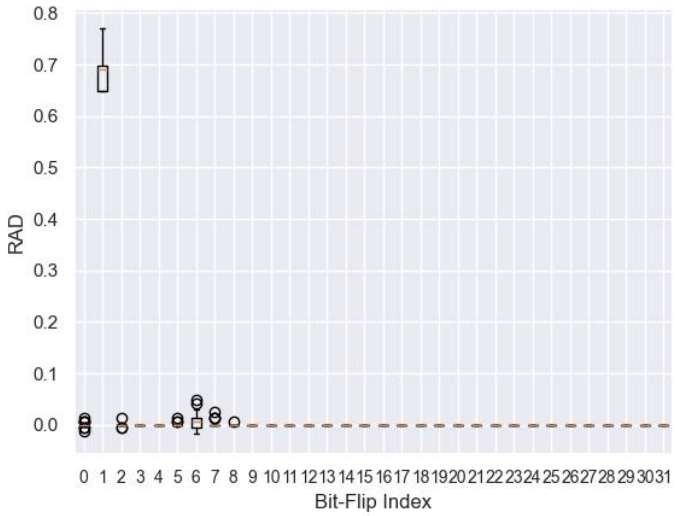

Figure 3: Bit-Flip index vs. RAD