CS 521: Homework 1

Jamie Fulford

Due September 19, 2025

Link to files: https://github.com/jfulfo/cs521/tree/main/hw1

Problem 1

For the first part of the problem I used this code:

```
eta = eps * x.grad.sign()
adv_x = x - eta
```

This fools the classifier successfully, as it reports the target class of 0.

For the second part of the problem, the given epsilon of 0.5 did not successfully fool the classifier. The natural solution was to increase the size of the ε ball we used. I iterated through different ε values, increasing by 0.01, and found that 1.14 to 1.35 would fool the classifier. This is the same as the norm of the difference between the adversarial input and the given input.

Problem 2

My implementations largely followed the FGSM example, as well as the hint given for the L_2 case. I computed the multi-norm robust accuracy by ANDing two tensors together for the union-threat model:

```
correct_linf = (pred_linf == y_batch)
correct_l2 = (pred_l2 == y_batch)
tot_acc += (correct_linf & correct_l2).float().sum().item()
tot_test += x_batch.size(0)
```

Model	Standard	L_∞ Robust	L_2 Robust	Multi-norm
L_∞-trained	82.80%	**51.20%**	48.81%	47.04%
L_2-trained	**88.75%**	30.49%	54.39%	30.49%
RAMP	81.19%	49.62%	**59.67%**	**49.62%**

The accuracies of the models show that specifically trained models excel in their respective categories but do not generalize well to other categories. The L_2 model, in particular, fails to protect against L_∞ attacks. The RAMP is more well balanced across the attacks, including having the greatest minimum, so it makes sense that it would perform well in the multi-norm setting. We also can see the trade-off between standard accuracy and robust accuracy: the more robust a model is on average, the worse its standard accuracy tends to be.

Problem 3

The authors argue that current techniques for testing the robustness of neural networks rely on L_p balls. To address this issue, the authors propose several new non L_p attacks, as well as using the new attacks to create a benchmark, which they argue is more realistic than existing ones. To measure the new attacks they created a threat model for "unforeseen robustness"; robustness for an entire distribution of adversaries and attacks.

The main strength of the paper is the proven novelty of the technique. They successfully show that non L_p based techniques have not been done before in literature to the extent they have been done in the paper. Another strength is the benchmark they provide, which is especially useful for further research. The paper is rather succinct, but contains a large appendix, lending to readability.

The most apparent weakness of the paper is the lack of real world application. The paper mainly argues that their research matters in a few sentences at the beginning, but fails to maintain this argument throughout the paper. Along the same vein, another weakness is that they argue their new attacks are more "realistic", but fail to provide ample reasons as to why this is true. How their proposed benchmark compares with existing measures could be expanded. Most of what they say can be simplified down to "unforeseen robustness is largely different". They should include an argument as to why worst case robustness matters.

The natural extension of this to LLMs is to similarly define a "distribution of attacks" measure as they do in the paper. Not sure how exactly this would be done, as in the paper they have a different perturbation sets to define their collection of adversaries. One would need a way to directly quantifies the adversarial attacks against LLMs.