Refining Adversarial Training Methods Using Game Theory
15-400 Project Milestone Report

Akhil Nadigatla

28 March 2021

1 Major Changes

As previously mentioned, we have made a slight pivot towards out approach to the research project. Our two main interests in this with respect to Alexander Madry's saddle-point problem formulation are:

1. Are there better ways of solving the inner maximization problem rather than using PGD, a heuristic approach?
2. Can we approach the outer minimization to produce an ensemble of classifiers rather than single one?

2 Accomplishments

For the two question outline above, here is our progress thus far:

1. Literature review on this topic has showed us that all proposed improvements to the initial proposal made by Madry Lab for the inner maximization seem to be rooted in PGD or simply give a slightly modified version of the same procedure. This is surprising given that PGD is provably sub-optimal. Our question of interest is: what is the reason for continuing down the same path? Is it because of the simplicity provided by this method?
2. Before getting deep into the process of weighting the different classifiers in the particular ensemble, we want to first see if using an ensemble adds an value. We have tried to find attacks that are different from the standard black-box PGD-based ones used in Madry's paper, but very few exist. The reason why we want to choose non-PGD models as part of our ensemble

3 Meeting Fourth Milestone

Our question has evolved into a fascinating meta discussion about why adversarial robustness research has stagnated when it comes to the inner maximization problem (maximizing the expected loss across all possible adversarial perturbations possible for a particular data distribution) and continued to focus on PGD rather than experimenting with other methods. Arriving at this question took a substantial amount of literature review. The surprising lack of models for initial experimentation using an ensemble around the outer minimization also slowed down our progress. Nevertheless, we are still engaging in interesting discussions using the few models we found.

4 Surprises

No notable surprises were encountered.

5 Looking Ahead

Looking ahead, we hope to continue our goals as planned. While the end of the semester is edging closer, we hope to (at least for the purpose of 15-400) provide some discussion comparing the (potential) robustness gains with and without using an ensemble of classifiers. We also hope to dig deeper into understanding why PGD remains a 'standard' method while it is clearly sub-optimal.

6 Revisions to Future Milestones

No revisions made to future milestones.

7 Resources Needed

No additional resources are needed on my end.