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

Akhil Nadigatla

14 March 2021

1 Major Changes

At the moment, while no major changes have been made concrete, we are planning to shift the direction of the project towards a major general question. Rather than focusing on specific adapted game theoretic algorithms against contemporary adversarial training algorithms, we are likely to investigate the effects of switching from attempting to find a pure strategy Nash equilibrium (as is seen in Madry et al.) to looking for a mixed strategy Nash equilibrium instead, and the effects of this change in objective.

2 Accomplishments

Since the second milestone, I have dedicated most of time studying game theory concepts like Nash equilibria, pure and mixed strategies, as well as their application in optimization research thus far. We also felt like a more general question (as proposed above) is more feasible in our current schedule. Moreover, the results from this modified question will likely be more valuable for future research.

3 Meeting Third Milestone

The third milestone was designed for me to get more acquainted with game theory at a higher level and play around with some of its algorithms. However, in the process, we found a more fundamental question to be of more interest and found very little research conducted on this question (pure strategy versus mixed strategy optimization).

4 Surprises

No notable surprises were encountered.

5 Looking Ahead

As mentioned, this is likely to be a small pivot in our direction of investigation. While there will be no change in the actual research question of interest, there will be a shift in how we hoped to harness game theory to improve adversarial robustness. Rather than presenting a laundry list of game theory algorithms and their results on appropriately trained models, we are staring at a more general question that remains under-investigated, at least to our knowledge.

6 Revisions to Future Milestones

Once our changes have been made official, we will be adjusting the milestones accordingly. The fourth milestone, for example, will now focus on the effects of randomization on finding a robust classifier for adversarial perturbed versus benign data points.

7 Resources Needed

No additional resources are needed on my end.