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

Akhil Nadigatla

28 March 2021

1 Major Changes

One main hurdle that we have encountered so far is the lack of existing packages surrounding adversarial robustness that fit our needs. We have looked into packages like the Adversarial Robustness Toolkit (ART) by IBM and the Advertorch package by Borealis AI. However, there was a need for a lot of modification to done in order to use these in a manner that fits our project needs.

2 Accomplishments

During this milestone, our main accomplishment was successful refactoring and creation of a code base that allows for the training and examination of 9 models to be used in the ensemble.

3 Meeting Fourth Milestone

The scope of our project has significantly diminished from our initial proposal. While we had allocate additional time anticipating such circumstances, scheduling coupled with the sheer magnitude of the existent research work has slowed down progress.

4 Surprises

As mentioned, we did not anticipate how much code needed to be actually written in order to meet our goals. We had hoped existing adversarial robustness packages would have sufficed.

5 Looking Ahead

Our goal moving ahead is going to be trying to see if we can extract concrete theorems from our observations that can explain why the model ensemble will likely be more adversarially robust than existing methods.

6 Revisions to Future Milestones

No revisions made to future milestones.

7 Resources Needed

No additional resources are needed on my end.