THEORY OF MACHINE LEARNING

LECTURE 25

ROBUSTNESS (PART 2)

ANNOUNCEMENTS

- **Homework 4 – due last on Tuesday Apr 26**
  - Discussion
- **Project presentations:** starting next week!
  - Please sign up!
  - Optional – submit presentation pdf on canvas for smoother transitions

RECAP: LEARNING IN THE PRESENCE OF ADVERSARIES

- Training time versus test time

- Training time: adversary corrupts small fraction of inputs
- Test time: adversary evaluates model on inputs with "imperceptible error" added (can be viewed as input distribution vs test)

- Former has multiple models – Benign noise, Huber's corruption model, data poisoning
- Field of robust statistics

ALGORITHMS AT HIGH LEVEL

- Use entire data, but limit *influence* of outliers
  - Median instead of mean (low dimensions)
  - Truncated gradients
- "Inlier pursuit": key idea is that inliers "reinforce" one another
  - RANSAC algorithm
  - More sophisticated "filtering" algorithms
- Main problem of study – robust mean estimation
  - Can be used as a subroutine in other algorithms (use robust mean estimation for gradients!)

ROBUSTNESS OF TRAINED MODELS

- "Intriguing" property of deep learning models – models that generalize well are surprisingly brittle! [Szegedy et al. 2013]

- Obvious consequences
- Why possible? (statistical explanation)

ROBUSTNESS OF TRAINED MODELS

- How can we "generate" such adversarial examples?

- Now standard approaches
  - Fast Gradient Sign Method (FGSM) – way to use the gradients to carefully choose direction of movement
  - Projected Gradient Descent (PGD) – iterative procedure to "maximally" affect loss function

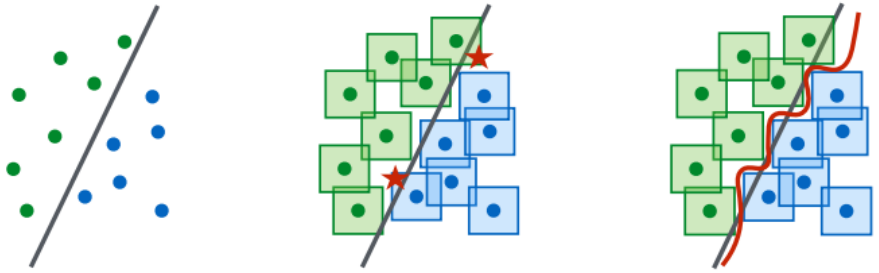
WHY DO ADVERSARIAL EXAMPLES EXIST?

- I.e., why only for deep models?

- Margin theory

"PROVABLY" DEFENDING?

- Can we show that *no corruption of small magnitude* can hurt the classifier?
- "Adversarial training":
  - Instead of minimizing empirical risk, minimize empirical "robust risk"
  - How to solve this optimization problem? (uses theorem of Danskin – gives a way of solving min-max opt problems)
  - Do we need "richer models"?



“PROVABLY” DEFENDING – CONNECTION TO PRIVACY

MORE NUGGETS

- Do we need more data for obtaining robust models?
- State of the art accuracies
- Expressibility vs trainability vs robustness