THEORY OF MACHINE LEARNING

LECTURE 24

ROBUSTNESS

ANNOUNCEMENTS

- **Homework 4 – out soon, due ~ 2 weeks**
- **Project presentations:** starting next week! (~18 projects)
  - Dates: April 19, 22, 26 (5 projects /class), couple online
- This week and next: representation learning, robustness

LAST WEEK

- NNs and "representation learning"
  - Intermediate layers of NN
  - NN transforms inputs -> "feature space embeddings"
  - Supervised vs unsupervised representations (when is a rep "good")
  - Self-supervision
    (using supervised learning to do unsupervised learning)
- Representations in NLP
  - Embeddings for words (Firth's hypothesis, n-grams)
  - Embeddings for nodes in a graph

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)

NOISE IN TRAINING DATA

- "Benign" noise:
  - Very common – standard regression analysis, clustering, label noise in supervised learning, … (iid mean zero noise)
  - Less common – few (random subset) of points are "badly" corrupted (Huber's contamination model)
- Adversarial noise (data poisoning)
  - Carefully chosen subset of points is corrupted
  - Even basic problems are hard! (robust mean estimation for Gaussian data, robust PCA)
  - Lot of work on robust mean estimation (why?)

CLASSIC ALGORITHMS

- Mean estimation in low dimensions – median vs mean

- "Inlier pursuit": key idea is that inliers "reinforce" one another
  - RANSAC algorithm

PROBLEM OF DIMENSIONALITY

- High dimensional mean estimation
  - Clean data = n iid samples from Gaussian in d dimensions (mean μ, covariance matrix Σ)
  - Corrupted data = ϵn points from clean data are replaced with some adversarially chosen points (in R^d)
- Can you recover the parameters μ' and Σ' so that $N(\mu, \Sigma) \approx N(\mu', \Sigma')$ [Can show that if you allow exponential time, this can be done to $O(\epsilon)$, if n is big enough (polynomial in d)]
- Simpler problem: assume $\Sigma = I$

ROLE OF DIMENSION

- In 1-D, problem fairly easy
- What about d dimensions?

- Main result [Diakonikolas et al. 2016, Lai et al. 2016]: there exists an algorithm that can efficiently recover the mean to error ~ $\epsilon\sqrt{\log 1/\epsilon}$
- Key idea: "filtering"
- Can also be extended to *arbitrary* distributions (not just Gaussian, as long as variance is bounded)