# 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~~ 6 projects /class), couple online

each ~ 10 - 12 minutes.

- ~~This week and next: representation learning, robustness~~

!

# Presentation template :

→ Background : - motivation, why? what the paper is about.

5 mins. - where it fits in with course material.

-

→ Main result(s) → present in 3-4 min.

→ might need to form "informal" versions or special cases..

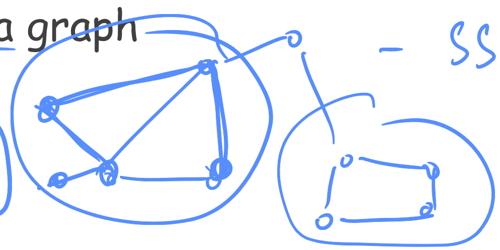
→ Results (experiments / proofs) : 2-3 min.

[be prepared for interruptions if something is not clear..].

## 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 (SSL)
    (using supervised learning to do unsupervised learning)
- Representations in NLP
  - Embeddings for words (Firth's hypothesis, n-grams)
  - Embeddings for nodes in a graph
    (Social networks)

[Finding good "cuts" in graphs.]



- Given lots of text data, find "representations" that capture meanings of words.
- SSL is better than classical methods like parts-o-s tagging, etc.

# LEARNING IN THE PRESENCE OF ADVERSARIES

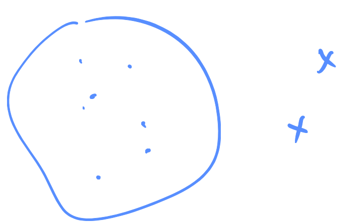
- Training time versus test time

(introducing new data points; mislabeling..)

- 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)

Data ~~> train model ~~> do inference.

Challenge mainly for NN models.

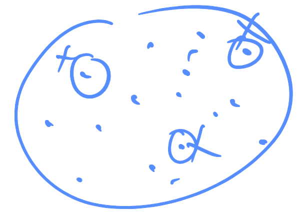


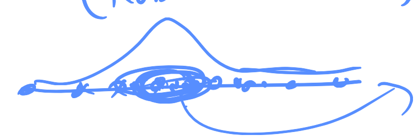
# NOISE IN TRAINING DATA

regression: $\langle a_i, x \rangle + \eta \to$ Gaussian noise.

- "Benign" noise: (common in stats analysis).
  - 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)

(Robust Statistics).





- 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?)

unknown distn with bounded variance (or second moments).

given $x_1, x_2, \dots, x_N \sim \mathcal{D}^{\varepsilon}$, find $\mu$ of $\mathcal{D}$.

$\mathcal{D}^{\varepsilon}$: $\mathcal{D}$, ~~corr~~ with $\varepsilon$ fraction corrupted.

→ Solving some other loss min problem can be <u>reduced</u> (in practice) to mean estimation).

- gradient descent can be viewed as a sequence of mean estimation steps.

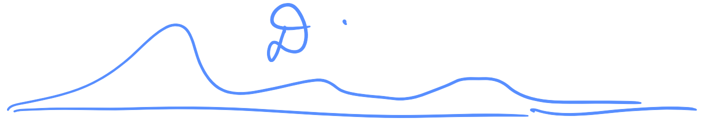
→ (SEVER: Steinhardt, Diakonikolas, et al.)

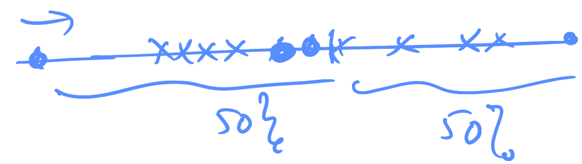
# CLASSIC ALGORITHMS (

$\mathcal{D}^{\varepsilon}$ : (always think of $\varepsilon$ as 1%.)

- Mean estimation in low dimensions – median vs mean

$\mathcal{D}$ :






→ median → within fixed distance of mean.

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

(Random Sampling & Consensus).

→ Say we have a guess for $\mu, \sigma$

→ check for "consistency" → ie, can we remove 1% of pts & get

helps you identify inliers + outliers.

new guess for $\mu, \sigma$.

→ # params is small

→ Error rate is small

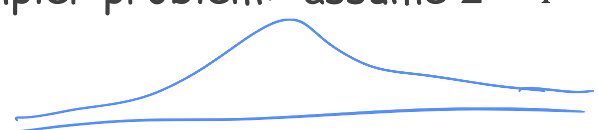
# PROBLEM OF DIMENSIONALITY

[Diakonikolas et al. 2016.
Lai et al. 2016.]

- High dimensional mean estimation
  - Clean data = n iid samples from Gaussian in d dimensions (mean $\mu$, covariance matrix $\Sigma$)
  - Corrupted data = $\epsilon n$ points from clean data are replaced with some adversarially chosen points (in $R^d$ )
- Can you recover the parameters $\mu'$ and $\Sigma'$ 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 $d$)]

Can we recover $\mu$, $\Sigma$ well?

- Simpler problem: assume $\Sigma = I$




find $d/\varepsilon$ points

$|\hat{\mu} - \mu| \sim \sqrt{d}\,\varepsilon.$

estimator

$\rightsquigarrow \frac{\log n}{\delta^2}$

$\frac{1}{\delta} \cdot \varepsilon.$

# ROLE OF DIMENSION

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

used a SDP.

- Main result [Diakonikolas et al. 2016, Lai et al. 2016]: there exists an algorithm that can efficiently recover the mean to error ~ $\epsilon\boxed{\sqrt{\log 1/\epsilon}}$

Jerry Li
Steinhardt.

- Key idea: "filtering"
- Can also be extended to *arbitrary* distributions (not just Gaussian, as long as variance is bounded)