THEORY OF MACHINE LEARNING

LECTURE 25

ROBUSTNESS (PART 2)

ANNOUNCEMENTS

- **Homework 4 – due last on Tuesday Apr 26**
  - Discussion → Issues in analyzing convergence of a non-convex opt problem.
- **Project presentations:** starting next week!
  - Please sign up! → On Tuesday, saw how to structure your pres.
  - 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 / ~~no~~ noisy points.
  - Median instead of mean (low dimensions)
  - Truncated gradients (clustering, mean estimation)
- "Inlier pursuit": key idea is that inliers "reinforce" one another
  - RANSAC algorithm
  - More sophisticated "filtering" algorithms

Promising idea/hypothesis: if you had only inliers a "simpler" model is possible.

- 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]

deep learning models have tons of params

stop sign Confidence: 0.9153 | Adversarial perturbation | flowerpot Confidence: 0.8374

out of dist. input ? so generalization guarantees don't mean anything.

- Obvious consequences
- Why possible? (statistical explanation)

ROBUSTNESS OF TRAINED MODELS

input x, $x+\delta$ $\longrightarrow \boxed{M} \longrightarrow y$.

$M(x+\delta) \simeq M(x) + \langle \delta, \nabla M(x) \rangle$

- How can we "generate" such adversarial examples?

→ ~~...~~ examples that look "normal" but fool the model

- 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

(implemented in standard DL libraries).

assume you have access to model..

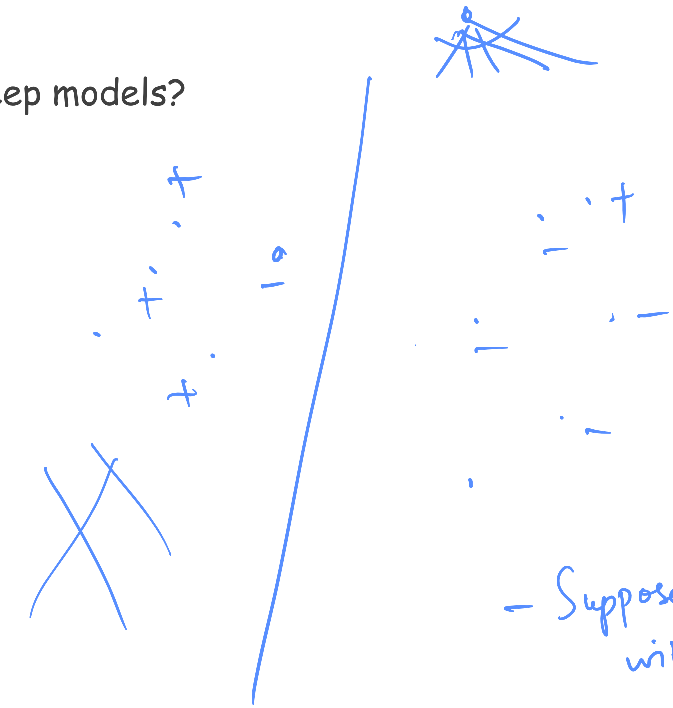
→ Same model arch; ImageNet.

"generalizable" adversarial examples.

WHY DO ADVERSARIAL EXAMPLES EXIST?

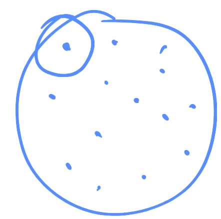
- I.e., why only for deep models?

- Margin theory



100 × 100 → $d = 10^4$.

inputs in $\mathbb{R}^d$.



$\left(\frac{1}{\varepsilon}\right)^d$ points are necessary for an ε-net.

- Suppose I want an input pt that is within ε of every point on the sphere.

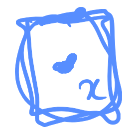
"PROVABLY" DEFENDING?

$\forall$ training points x

$\forall$ "real inputs" x, x & $x+\delta$ must have same classification, as long as $\|\delta\|_\infty \le \varepsilon$.

- Can we show that *no corruption of small magnitude* can hurt the classifier?
- "Adversarial training": minimize $\mathbb{E}_{x \sim \mathcal{D}_{inputs}} \ell(x, y; w)$ — y: label(x), w: parameters.
  - Instead of minimizing empirical risk, minimize empirical "robust risk"

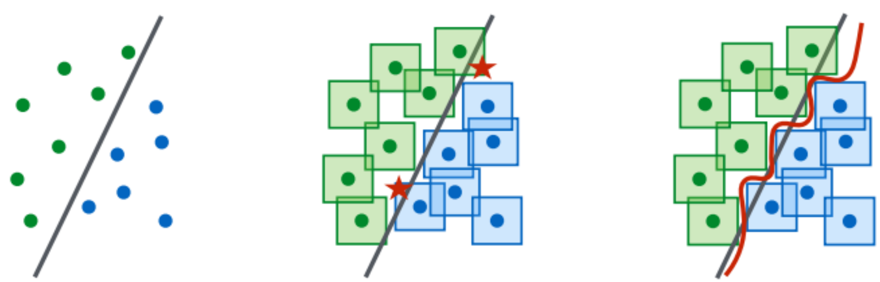
$$\min_{w} \; \mathbb{E}_{x \sim \mathcal{D}_{inputs}} \; \max_{\delta,\ \|\delta\| \le \varepsilon} \ell(x+\delta, \text{label}(x); w).$$



[Wald].

Saddle point opts.

  - How to solve this optimization problem? (uses theorem of Danskin – gives a way of solving min-max opt problems) → [Madry, .. 2018].
  - Do we need "richer models"?



Trade off between robustness & "regular" accuracy.

Tradeoff between robustness & accuracy.

$x_1, x_2, \ldots, x_N$.

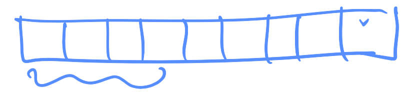
clean training data.

model M → clean accuracy → 80%. → noisy accuracy → 5%.

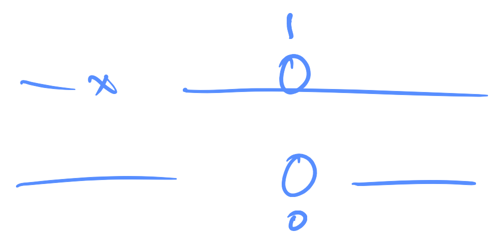
model M' trained ~~on~~ using min max training. → noisy accuracy → 60%.

clean accuracy → 65%.

"PROVABLY" DEFENDING – CONNECTION TO PRIVACY

Differential privacy: dataset $\underline{X}$ $\rightarrow$ 

Clever formulation by [Dwork, Nissim, Naor, Smith.] $\rightarrow$ Suppose my "trusted party" ensures that $\forall$ query that analyst gives, Answer(x) & Answer(x', where you flipped your bit) are indistinguishable.



Answer$(x) \simeq$ Answer$(x+\delta)$.

MORE NUGGETS

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