Learning Halfspaces under Massart Noise

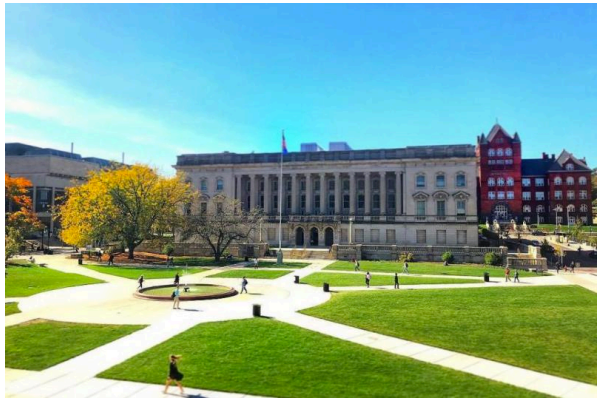
Christos Tzamos (UW Madison)

Based on joint work with **Themis Gouleakis (MPI)** and **Ilias Diakonikolas (UW Madison)**

NeurIPS 2019 (Outstanding Paper Award)

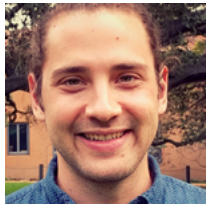
A Few Words About University of Wisconsin-Madison

- Top CS department in US
- Great place to be:
  - Surrounded by lakes
  - University Town



- Many Greek faculty






Learning Theory / Machine Learning Databases Graphics

- And Greek students





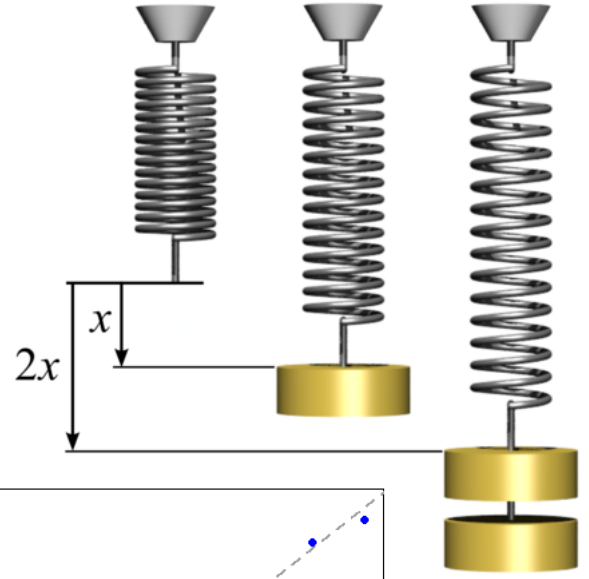
BACKGROUND ON MACHINE LEARNING

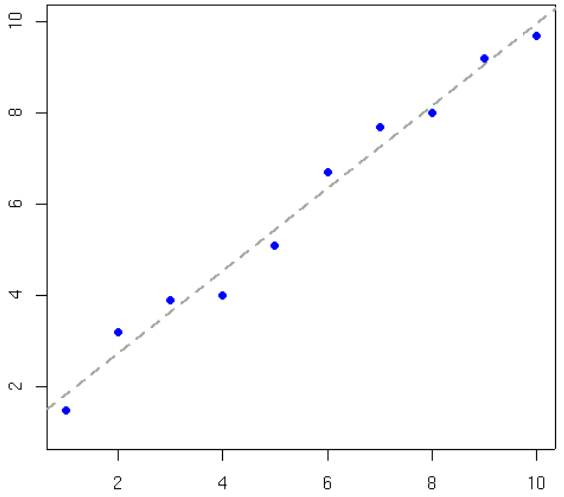
Goal of Machine Learning: Generalization

- Predicting the spring deformation
- Past observations
  - 2 kg → 3.2 cm
  - 6 kg → 6.1 cm
  - 10 kg → 9.6 cm
  - …

- What will the deformation be at 4 kg?
- A linear equation is predictive of the data
  - Hooke's law

- Linear Regression: Fit best line



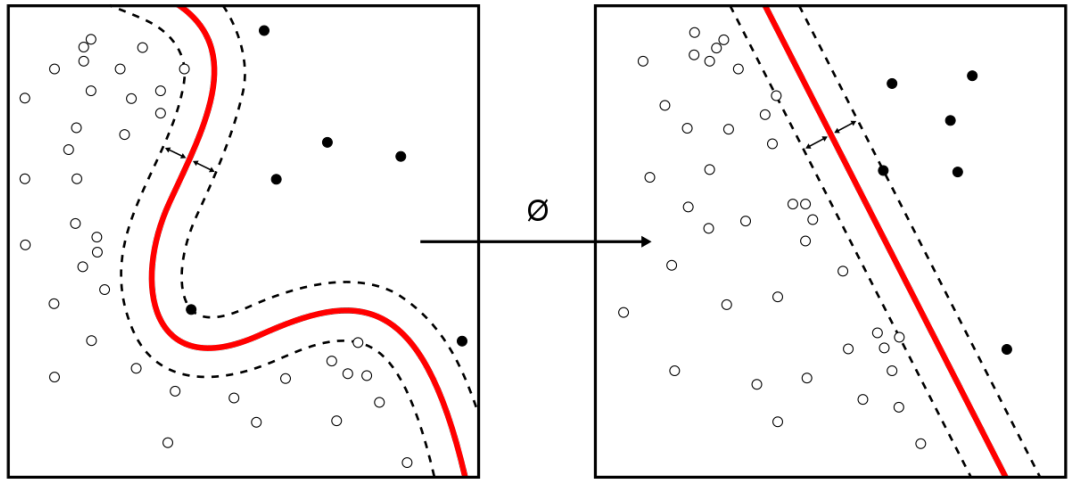


Beyond Linear Dependencies

- How to capture non-linear dependencies?
  - e.g. time it takes for an apple to reach the ground vs height
  - Equation still linear as a function of $\sqrt{\text{height}}$
- Can linearize many different problems
- What about dependence in more than one variable?
  - Multi-variate linear regression
  - $y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$
- More broadly: the goal is to fit the best curve

CLASSIFICATION PROBLEMS

- Goal is to classify data in two or more categories, e.g. sick vs healthy, cats vs dogs
- Find the curve that best separates the data



- Focus on Linear Separators also known as Halfspaces

A Classification Problem

- Train a program to classify a person's hair as blonde vs brown:



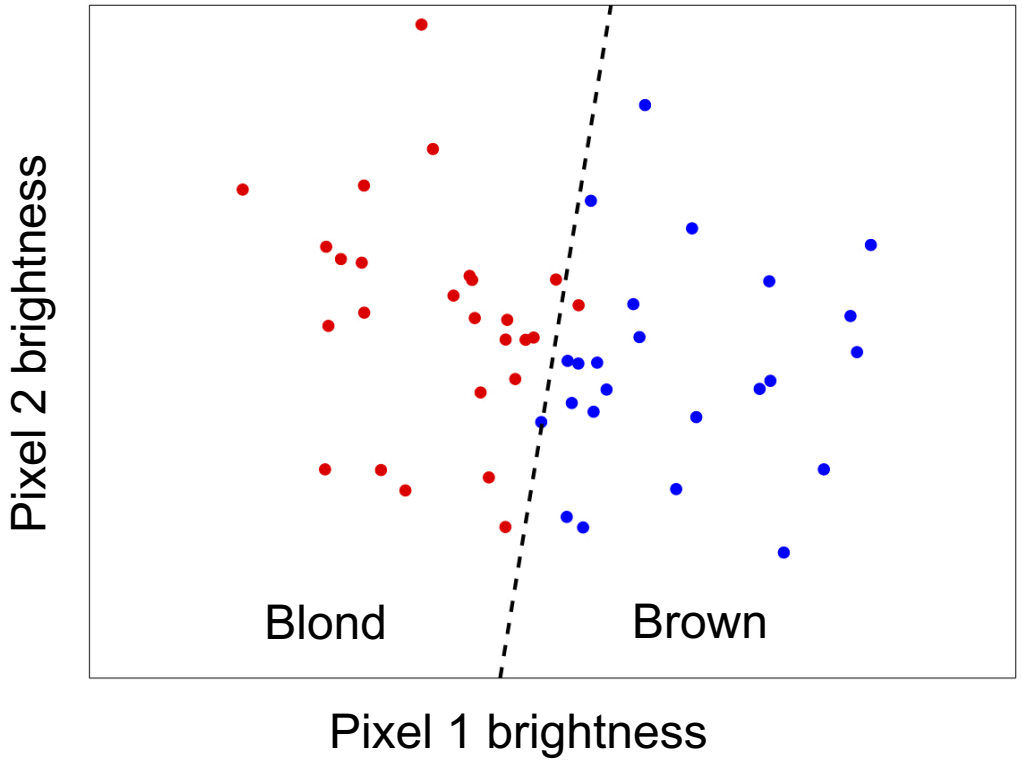
- How to write a program to do it?
  - Simple solution: Consider a pixel in the image and look at how dark it is
  - A simple threshold rule: If **x > threshold** then **blond** else **brown**
  - Consider all thresholds and pick the one that works best.

A More Complex Rule: Look at more pixels

- Look at the brightness of 2 pixels in an image, let (x_1, x_2) be their brightness values
- Plot these values as points in 2d.
- Find a line that best separates the two sets.

- Or consider all pixels: Find the hyperplane that best separates all the points.

- Learning a **Linear separator**:
  - $w_1 x_1 + w_2 x_2 + \cdots + w_d x_d > \theta$

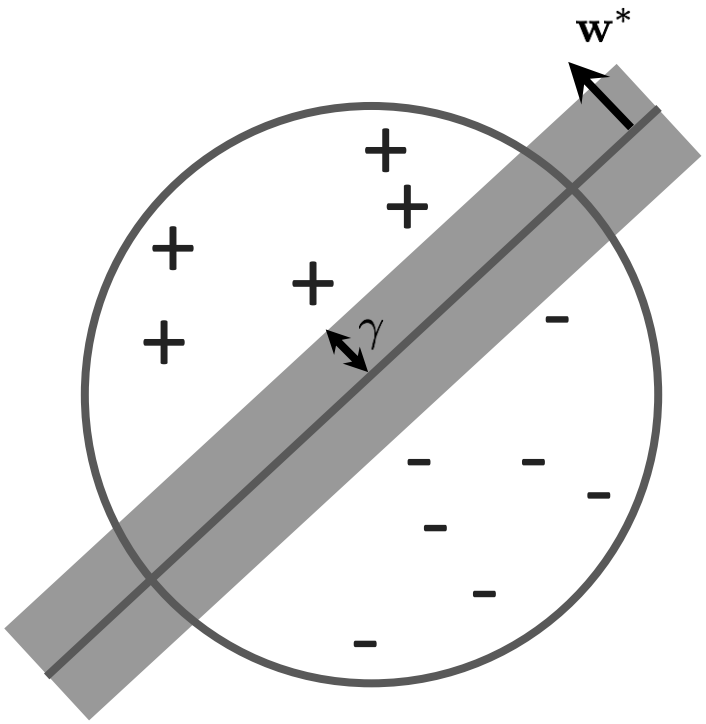


CLASSIFICATION IN MACHINE LEARNING

- Central problem in machine learning
- A dataset with different examples $x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(n)} \in \mathrm{R}^d$ and their labels
  $$y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(n)} \in \{-1, +1\}$$
  - (The examples are assumed to be i.i.d. draws from an underlying distribution D, PAC model)
- A parameterized family C of functions $f : \mathrm{R}^d \rightarrow \{-1, +1\}$
  - Find the function in the family that best separates the data (Minimizes $\Pr_{(x,y)\sim D}[f(\mathrm{x}) \neq y]$)
  - Example: Linear separators: $f(x) = \mathrm{sign}(\langle w, x \rangle - \theta)$
    - $(d+1)$ parameters $w = (w_1, \dots, w_d)$ and θ
- Two challenges
  - How many examples are needed to identify a good function?
    - Well understood in ML, typically proportional to the number of parameters
  - How to algorithmically find a good function that minimizes (misclassification) error?
    - **Focus of this talk**

Linear Separators with Separable Data

- Learning linear separators has been extensively studied in ML since [Rosenblatt'58]
- If the given points can be perfectly separated by some hyperplane: **linearly separable**
- When data are linearly separable, the **Perceptron** algorithm [Rosenblatt'58] finds a perfect linear separator.
  - Number of iterations depend on the margin, i.e. distance of closest point to the linear separator
  - An optimal linear separator can also be found through linear programming in time polynomial in the dimension (no dependence on margin)



Non-Separable Data: Agnostic Model

- How to minimize the misclassification error when data are non-separable?

- Suppose there exists a linear separator with 1% misclassification error.
- This can be thought of as all data are linear separable but 1% are corrupted.

- **Agnostic Model** [Haussler'92, Kearns-Shapire-Sellie'94]:
  - Adversary can flip *arbitrary* 1% of the labels

- Strong noise model!

- **Negative result**: It is even computationally intractable to compute a classifier with 49% error.
  - [Guruswami-Raghevendra'06, Feldman et al.'06, Daniely'16]

Non-Separable Data: Random Classification Noise

- A weaker noise model imposes that noisy examples are not adversarially chosen but randomly

- **Random Classification Noise** (**RCN**) [Angluin-Laird'88]:
  - Every label is randomly flipped with probability 1%

- Polynomial-time algorithm for learning halfspaces with RCN
  - [Blum-Frieze-Kannan-Vempala'96]

- Intuitively even though noise exists, it cancels out because it is uniformly distributed

NON-SEPARABLE DATA: MASSART NOISE

- A noise model in-between RCN and Agnostic,

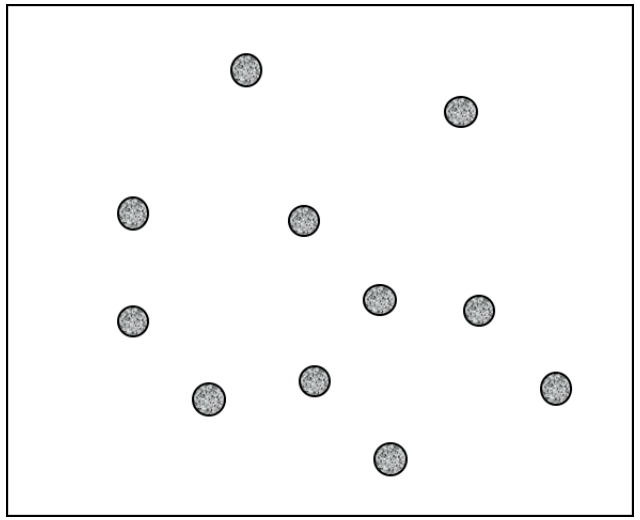
- **Massart Noise**, also known as Malicious misclassification noise [Sloan'88, Rivest-Sloan'94]:
  - Every label is randomly flipped with probability **at most** 1% but the exact probabilities are adversarially chosen

- If $f(x)$ is the true label of example x, then

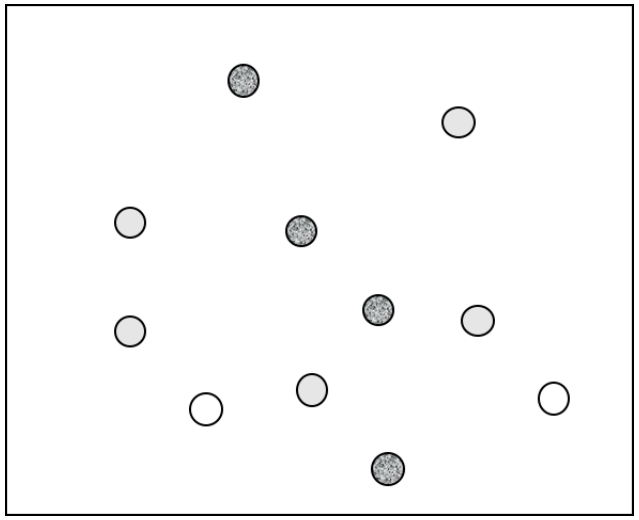
$$y^{(i)} = \begin{cases} f(\mathbf{x}^{(i)}), & \text{with probability } 1 - \eta(\mathbf{x}^{(i)}) \\ -f(\mathbf{x}^{(i)}), & \text{with probability } \eta(\mathbf{x}^{(i)}) \end{cases}$$

where $\eta(x) \leq 1\%$

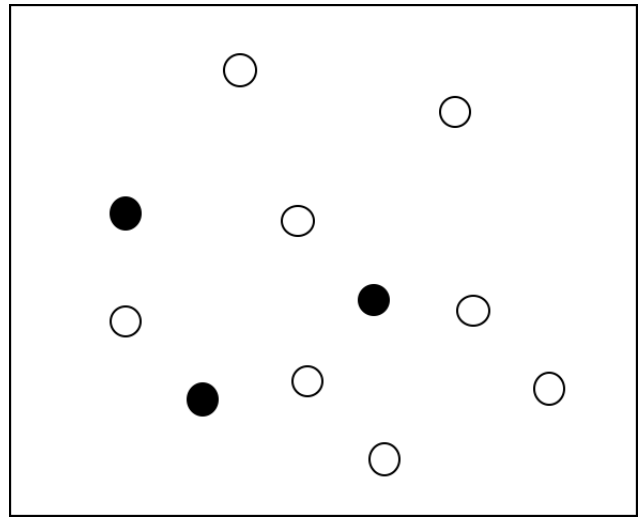
Summary of Noise Models and Computational Results



RCN
Noise Rate **exactly** 1%



Massart
Noise Rate **at most** 1%



Agnostic
Arbitrary 1% fraction

- Linear Separators efficiently learnable without noise
  - [e.g., Maass-Turan'94].
- Efficient algorithm for learning linear separators with RCN
  - [Blum-Frieze-Kannan-Vempala'96]
- Learning Linear Separators with Massart Noise?

OPEN

- Weak agnostic learning of LTFs is computationally intractable
  - [Guruswami-Raghevendra'06, Feldman et al.'06, Daniely'16]

LEARNING WITH MASSART NOISE: OPEN

Open Problem [Sloan'88, Cohen'97, Blum'03]

Is there a polynomial-time algorithm with non-trivial error for linear separators?

(Or even for more restricted concept classes?)

[A. Blum, FOCS'03 Tutorial]:

"*Given labeled linearly separable examples corrupted with 1% Massart noise, can we efficiently find a hypothesis that achieves misclassification error 49%?*"

No progress in distribution-free setting.

Efficient algorithms when marginal is *uniform on unit sphere*

(line of work started by [Awasthi-Balcan-Haghtalab-Urner'15])

Main Algorithmic Result

First efficient algorithm for learning halfspaces with Massart noise.

Main Theorem: There is an efficient algorithm that learns halfspaces on $\mathbb{R}^d$ in the distribution-independent PAC model with Massart noise. Specifically, the algorithm outputs a hypothesis h with misclassification error

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \eta + \epsilon$$

where η is the upper bound on the Massart noise rate, and runs in time $\mathrm{poly}(d, b, 1/\epsilon)$.

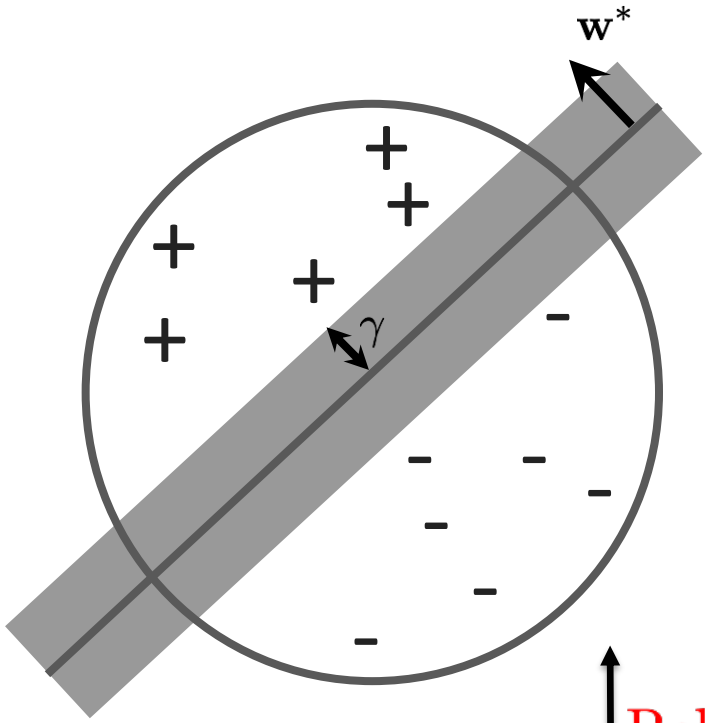
Remarks:

- Hypothesis is a decision-list of halfspaces.
- Misclassification error is $\eta + \epsilon$, as opposed to $\mathrm{OPT} + \epsilon$.
- First non-trivial guarantee in sub-exponential time.

Intuition: Large Margin Case

Target vector $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 = 1$

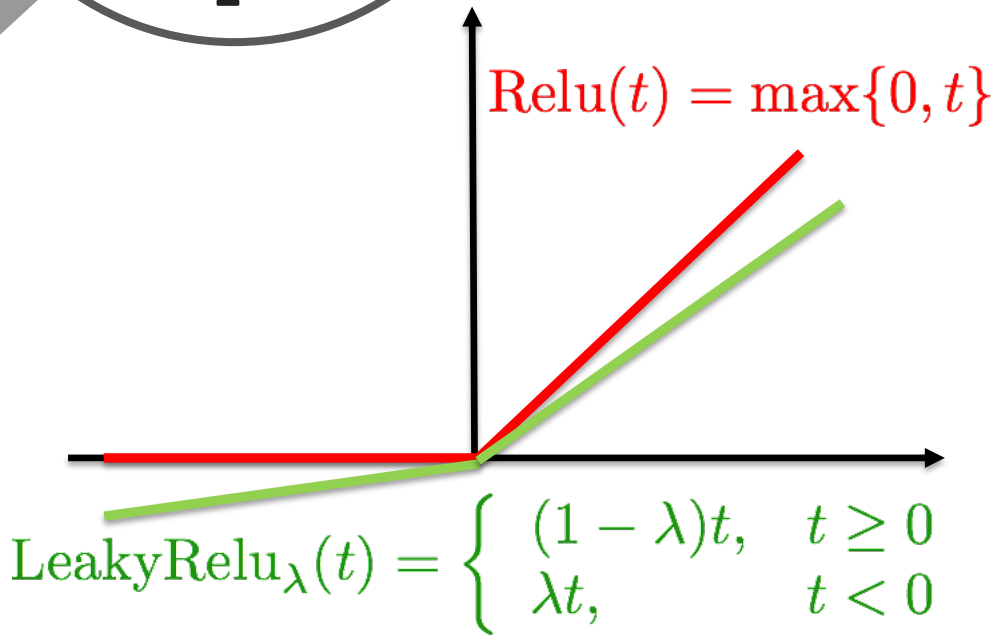
Marginal $\mathcal{D}_{\mathbf{x}}$ satisfies $|\langle \mathbf{w}^*, \mathbf{x} \rangle| \geq \gamma$



- **Realizable Case:**
  (Perceptron =) SGD on
  $$L_0(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{Relu}(-y\langle \mathbf{w}, \mathbf{x} \rangle)]$$

- **Random Classification Noise:**
  SGD on $L_\lambda(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{LeakyRelu}_\lambda(-y\langle \mathbf{w}, \mathbf{x} \rangle)]$
  for $\lambda \approx \eta$

In both cases: $L(\mathbf{w}) \geq 0$ and $L(\mathbf{w}^*) = 0$



LARGE MARGIN CASE: MASSART NOISE

Lemma 1: No convex surrogate works.

But…

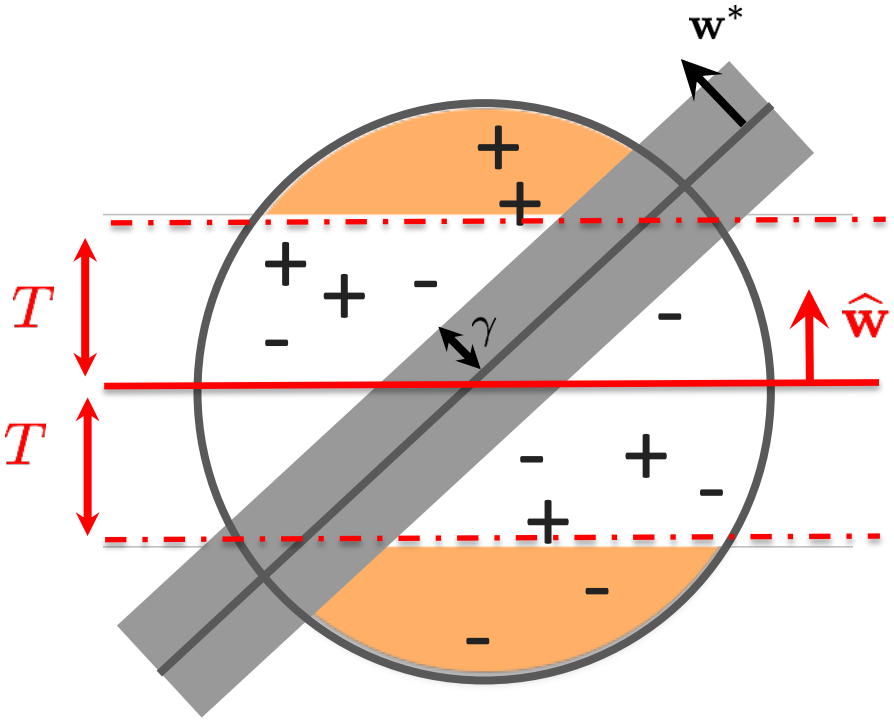
Lemma 2: Let $\widehat{\mathbf{w}}$ be the minimizer of

$$L_\lambda(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{LeakyRelu}_\lambda(-y\langle \mathbf{w}, \mathbf{x}\rangle)]$$

for $\lambda \approx \eta$.

There exists $T > 0$ such that $R_T = \{\mathbf{x} : |\langle \widehat{\mathbf{w}}, \mathbf{x}\rangle| \geq T\}$ has:

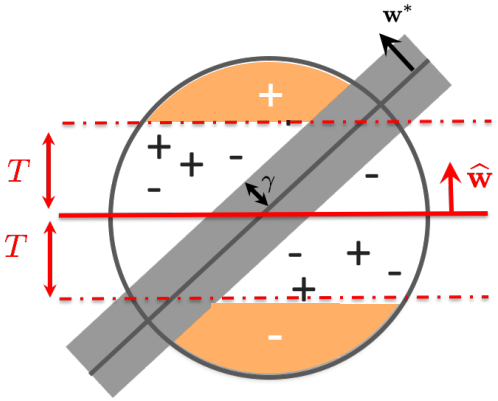
- $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[R_T] \geq \epsilon\gamma$, and
- $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\widehat{\mathbf{w}}}(\mathbf{x}) \neq y \mid R_T] \leq \eta + \epsilon$.



SUMMARY OF APPROACH

Lemma 2: Let $\widehat{\mathbf{w}}$ minimizer of $L_\lambda(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{LeakyRelu}_\lambda(-y\langle\mathbf{w},\mathbf{x}\rangle)]$ for $\lambda \approx \eta$. *There exists* $T > 0$ such that $R_T = \{\mathbf{x} : |\langle\widehat{\mathbf{w}},\mathbf{x}\rangle| \geq T\}$ has:

- $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[R_T] \geq \epsilon\gamma$, and
- $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\widehat{\mathbf{w}}}(\mathbf{x}) \neq y \mid R_T] \leq \eta + \epsilon$.



Large-Margin Case:

- There exists convex surrogate with non-trivial error on *unknown* subset S.
- Can algorithmically identify S using samples.
- Use convex surrogate hypothesis on S.
- Iterate on complement.

General Case:

Reduce to Large Margin Case

Conclusions And Open Problems

- First efficient algorithm with non-trivial error guarantees for
  for distribution-independent PAC learning of halfspaces with Massart noise.
- Misclassification error $\eta + \epsilon$ where η is an *upper bound* on the noise rate.

"Distribution-Independent PAC Learning of Halfspaces with Massart Noise"

I. Diakonikolas, T. Gouleakis and C. Tzamos - NeurIPS 2019

Open Questions:

- Error $\mathrm{OPT} + \epsilon$?
  - In recent work with Diakonikolas, Kontonis and Zarifis, we obtain efficient algorithms for data drawn from log-concave distributions
- Other models of robustness?

Thank you! Questions?