Distribution-Independent PAC Learning of Halfspaces with Massart Noise

Ilias Diakonikolas (UW Madison)

Themis Gouleakis (MPI)

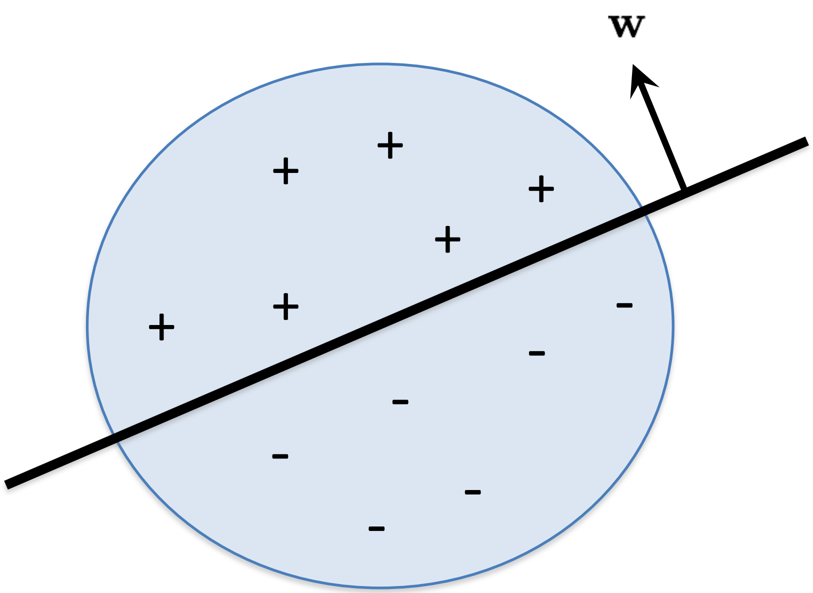
Christos Tzamos (UW Madison)

Explaining the Title

Main Result:

First computationally efficient algorithm for learning halfspaces in the distribution-independent PAC model with Massart noise.

HALFSPACES



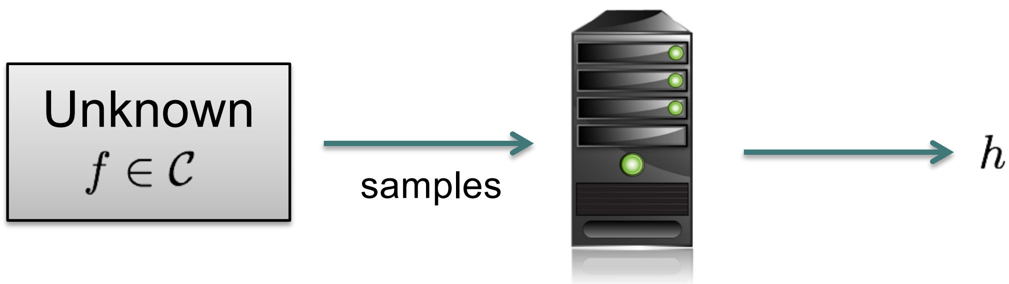
Class of functions $f : \mathbb{R}^d \to \{\pm 1\}$ such that

$$f(\mathbf{x}) = \mathrm{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle - \theta)$$

where $\mathbf{w} \in \mathbb{R}^d, \theta \in \mathbb{R}$

- Also known as: Linear Threshold Functions, Perceptrons, Linear Separators, Threshold Gates, Weighted Voting Games, …
- Extensively studied in ML since [Rosenblatt'58]

(Distribution-Independent) PAC Learning



$\mathcal{C}$: known class of functions $f : \mathbb{R}^d \to \{\pm 1\}$

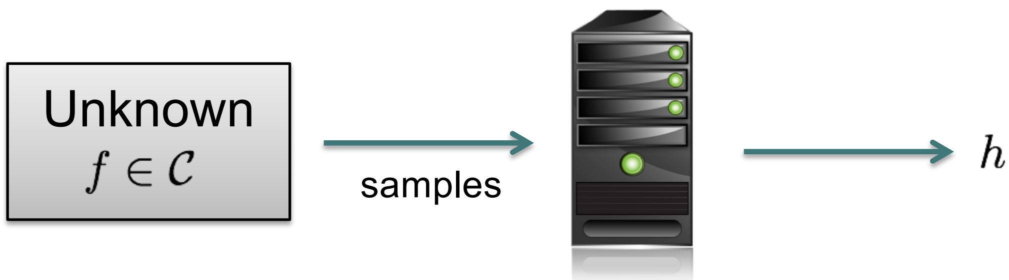
- **Input**: multiset of IID labeled examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ from distribution $\mathcal{D}$ such that: $\mathbf{x}^{(i)} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is **fixed but arbitrary**, and

$$y^{(i)} = f(\mathbf{x}^{(i)})$$

for some fixed unknown target concept $f \in \mathcal{C}$.

- **Goal**: find hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ minimizing $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$

(Distribution-Independent) PAC Learning With Massart Noise



$\mathcal{C}$: known class of functions $f : \mathbb{R}^d \to \{\pm 1\}$

- **Input**: multiset of IID labeled examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ from distribution $\mathcal{D}$ such that: $\mathbf{x}^{(i)} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is **fixed but arbitrary**, and

  $$y^{(i)} = \begin{cases} f(\mathbf{x}^{(i)}), & \text{with probability } 1 - \eta(\mathbf{x}^{(i)}) \\ -f(\mathbf{x}^{(i)}), & \text{with probability } \eta(\mathbf{x}^{(i)}) \end{cases}$$ where $\eta(\mathbf{x}) : \mathbb{R}^d \to [0, \eta], \eta < 1/2$

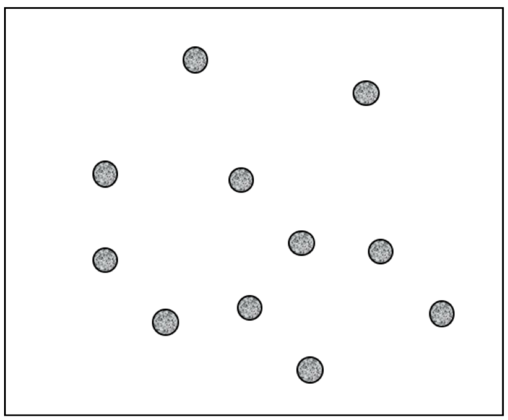
  for some fixed unknown target concept $f \in \mathcal{C}$.

- **Goal**: find hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ minimizing $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$

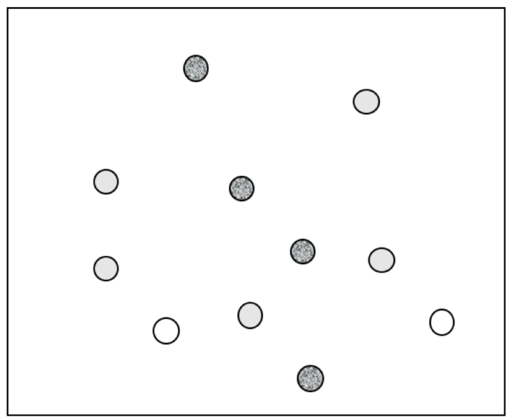
PAC Learning with *Other* Noise

Massart Noise “in between” Random Classification Noise and Agnostic Model:

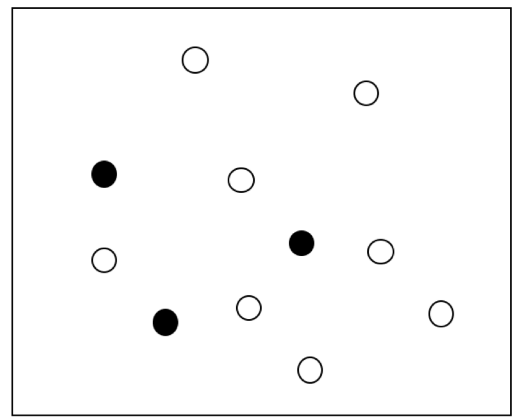
- **Random Classification Noise** (**RCN**) [Angluin-Laird’88]:
  - Special case of Massart noise: For all $\mathbf{x}$, we have that $\eta(\mathbf{x}) = \eta < 1/2$
- **Agnostic Model** [Haussler’92, Kearns-Shapire-Sellie’94]:
  - Adversary can flip *arbitrary* η fraction of the labels: $\inf_{f \in \mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] = \eta$



RCN
Noise Rate **exactly** η



Massart
Noise Rate **at most** η



Agnostic
Arbitrary η fraction

LEARNING HALFSPACES WITH NOISE: PRIOR WORK

Sample Complexity Well-Understood for Learning Halfspaces in all these models.

Fact: $\mathrm{poly}(d, 1/\epsilon)$ samples suffice to achieve misclassification error $\mathrm{OPT} + \epsilon$.

Computational Complexity

- Halfspaces efficiently learnable in realizable PAC model
  - [e.g., Maass-Turan'94].
- Polynomial-time algorithm for learning halfspaces with RCN
  - [Blum-Frieze-Kannan-Vempala'96]
- Learning Halfspaces with Massart Noise
- Weak agnostic learning of LTFs is computationally intractable
  - [Guruswami-Raghevendra'06, Feldman et al.'06, Daniely'16]

LEARNING HALFSPACES WITH MASSART NOISE: OPEN

Malicious misclassification noise [Sloan'88, Rivest-Sloan'94] (equivalent to Massart).

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

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

(Or even for more restricted concept classes?)

[A. Blum, FOCS'03 Tutorial]:

"*Given labeled examples from an unknown Boolean disjunction, 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

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

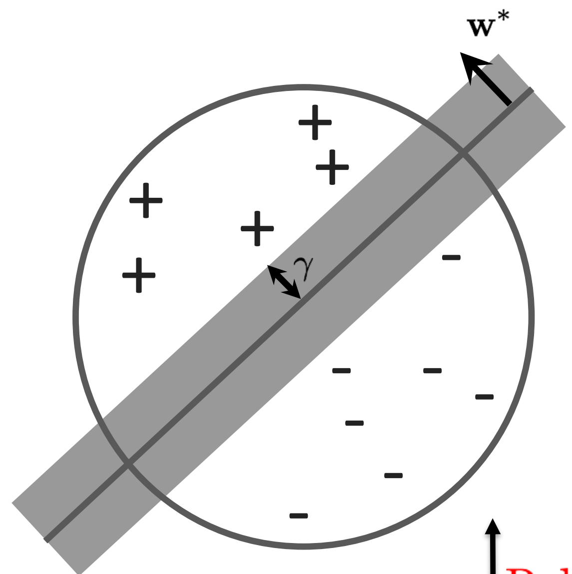
where η is the upper bound on the Massart noise rate, and runs in time $\mathrm{poly}(d, b, 1/\epsilon)$.

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

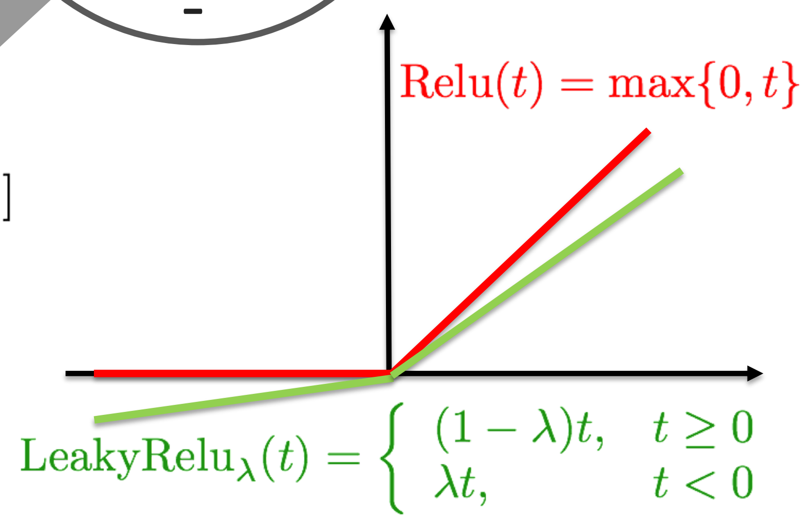
Target vector $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 = 1$
Marginal $\mathcal{D}_\mathbf{x}$ satisfies $|\langle \mathbf{w}^*, \mathbf{x} \rangle| \geq \gamma$



- **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…

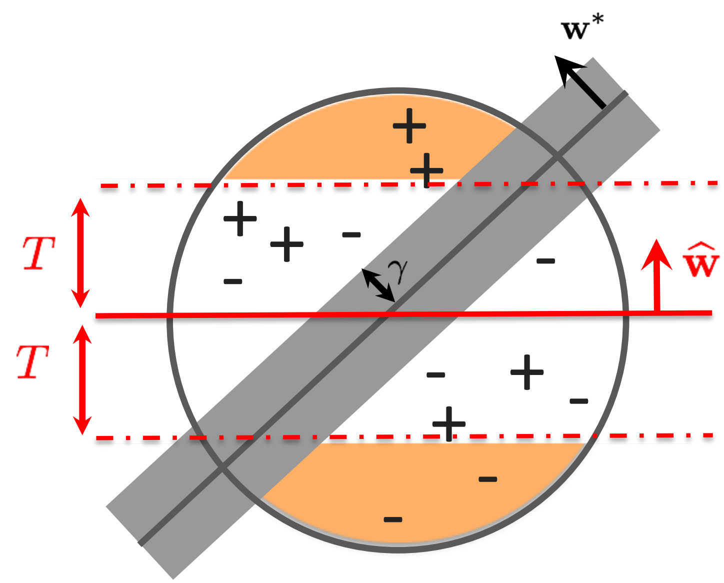
Lemma 2: Let $\widehat{\mathbf{w}}$ be the minimizer of

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

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



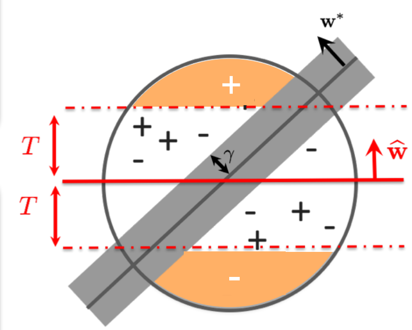
Summary of Approach

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

Open Questions:

- Error $\mathrm{OPT} + \epsilon$?
- Other models of robustness?

Thank you!
Questions?

Poster #226: 5-7 PM Today
East Exhibition Hall B + C