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
$$f(x) = \text{sgn}(\langle w, x \rangle - \theta)$$
where $w \in \mathbb{R}^d$, $\theta \in \mathbb{R}$

- 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

\[ C : \text{known class of functions } f : \mathbb{R}^d \rightarrow \{-1, 1\} \]

- **Input:** multiset of IID labeled examples \( \{(x^{(i)}, y^{(i)})\}_{i=1}^n \) from distribution \( \mathcal{D} \) such that:
  \[ x^{(i)} \sim \mathcal{D}_x, \text{ where } \mathcal{D}_x \text{ is fixed but arbitrary, and} \]
  \[ y^{(i)} = f(x^{(i)}) \]
  for some fixed unknown target concept \( f \in C \).
- **Goal:** find hypothesis \( h : \mathbb{R}^d \rightarrow \{-1, 1\} \) minimizing \( \Pr_{(x,y) \sim \mathcal{D}}[h(x) \neq y] \)
(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 $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ from distribution $\mathcal{D}$ such that:
  
  $x^{(i)} \sim \mathcal{D}_x$, where $\mathcal{D}_x$ is fixed but arbitrary, and
  
  $y^{(i)} = \begin{cases} 
  f(x^{(i)}), & \text{with probability } 1 - \eta(x^{(i)}) \\
  -f(x^{(i)}), & \text{with probability } \eta(x^{(i)})
  \end{cases}$

  where $\eta(x) : \mathbb{R}^d \to [0, \eta]$, $\eta < 1/2$

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

- **Goal**: find hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ minimizing $\Pr_{(x,y) \sim \mathcal{D}}[h(x) \neq y]$
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 \( x \), we have that \( \eta(x) = \eta < 1/2 \)

- **Agnostic Model** [Haussler’92, Kearns-Shapire-Sellie’94]:
  - Adversary can flip arbitrary \( \eta \) fraction of the labels: \( \inf_{f \in C} \Pr_{(x,y) \sim D}[f(x) \neq y] = \eta \)

\[ \text{RCN} \]  
Noise Rate exactly \( \eta \)

\[ \text{Massart} \]  
Noise Rate at most \( \eta \)

\[ \text{Agnostic} \]  
Arbitrary \( \eta \) fraction
**Learning Halfspaces with Noise: Prior Work**

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

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

**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

$$\Pr_{(x,y) \sim D} [h(x) \neq y] \leq \eta + \epsilon$$

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

Remarks:

- Hypothesis is a decision-list of halfspaces.
- Misclassification error is $\eta + \epsilon$, as opposed to $\text{OPT} + \epsilon$.
- First non-trivial guarantee in sub-exponential time.
**INTUITION: LARGE MARGIN CASE**

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

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

- **Random Classification Noise:**
  SGD on $L_\lambda(\mathbf{w}) = \mathbf{E}_{(x,y) \sim \mathcal{D}}[\text{LeakyRelu}_\lambda(-y\langle \mathbf{w}, \mathbf{x} \rangle)]$
  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 $\hat{\mathbf{w}}$ be the minimizer of

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

for $\lambda \approx \eta$.

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

- $\Pr_{(\mathbf{x},y) \sim \mathcal{D}}[R_T \leq \epsilon \gamma] \leq \eta + \epsilon$.

**Summary of Approach**

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

- $Pr_{(x,y) \sim D}[R_T] \geq \epsilon \gamma$, and
- $Pr_{(x,y) \sim D}[h_{\hat{w}}(x) \neq y | R_T] \leq \eta + \epsilon$.

**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 \( \eta \) is an upper bound on the noise rate.

Open Questions:

• Error \( \text{OPT} + \epsilon \) ?
• Other models of robustness?

Thank you!
Questions?

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