Learning in High Dimensions with Asymmetric Label Noise

Ilias Diakonikolas (UW Madison)

HALG 2020
Can we develop *supervised* learning algorithms that are *robust* to a *constant* fraction of *corruptions*?
MOTIVATION

- Model Misspecification/Robust Statistics
  [Fisher 1920s, Tukey 1960s, Huber 1960s]

- Adversarial/Secure ML
DATA POISONING

Fake Reviews [Mayzlin et al. ‘14]

Recommender Systems [Li et al. ‘16]
Crowdsourcing [Wang et al. ‘14]
Malware/spam [Nelson et al. ‘08]

So Many Misleading, “Fake” Reviews
**DISTRIBUTION-INDEPENDENT) PAC LEARNING**

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

- **Input**: multiset of IID labeled examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ from distribution $\mathcal{D}$ such that:
  \[ x^{(i)} \sim \mathcal{D}_x, \text{ where } \mathcal{D}_x \text{ is fixed but arbitrary, and} \]
  \[ y^{(i)} = f(x^{(i)}) \]
  for some fixed unknown target concept $f \in \mathcal{C}$.

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

---

Diakonikolas, HALG’20
Learning in High Dimensions with Asymmetric Label Noise
(Distribution-Independent) PAC Learning With Massart Noise

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

- **Input**: multiset of IID labeled examples \( \{(x^{(i)}, y^{(i)})\}_{i=1}^n \) from distribution \( \mathcal{D} \) such that:
  \[ x^{(i)} \sim \mathcal{D}_x, \text{ where } \mathcal{D}_x \text{ is fixed but arbitrary, and} \]
  \[ y^{(i)} = \begin{cases} 
    f(x^{(i)}), & \text{with probability } 1 - \eta(x^{(i)}) \\
    -f(x^{(i)}), & \text{with probability } \eta(x^{(i)})
  \end{cases} \]
  where \( \eta(x) : \mathbb{R}^d \rightarrow [0, \eta], \eta < 1/2 \)

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

- **Goal**: find hypothesis \( h : \mathbb{R}^d \rightarrow \{\pm 1\} \) 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 OPT fraction of the labels: 
    \[
    \inf_{f \in C} \Pr_{(x, y) \sim D}[f(x) \neq y] = \text{OPT}
    \]

---

**RCN**

Noise Rate **exactly** \( \eta \)

**Massart**

Noise Rate **at most** \( \eta \)

**Agnostic**

Arbitrary **OPT** fraction
OUTLINE

• Part I:
  – Distribution-Independent PAC Learning with Massart Noise

• Part II:
  – Distribution-Specific PAC Learning with Massart (and Other) Noise
Are there realistic noise models that allow for efficient algorithms without distributional assumptions?
Main Result of Part I

Main Result [D-Gouleakis-Tzamos’19]:
First computationally efficient algorithm for learning halfspaces in the distribution-independent PAC model with Massart noise.
Class of functions $f : \mathbb{R}^d \rightarrow \{\pm 1\}$ 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]
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.
Main Algorithmic Result

First efficient algorithm for learning halfspaces with Massart noise.

Theorem [D-Gouleakis-Tzamos’19]
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 $h$ such that

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

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.
- Optimal misclassification error is $\text{OPT} + \epsilon$, where $\text{OPT} = \mathbb{E}_{x \sim \mathcal{D}}[\eta(x)]$.
- 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
  \[ L_0(\mathbf{w}) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\text{Relu}(-y\langle \mathbf{w}, \mathbf{x} \rangle)] \]

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

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

\[ \text{Relu}(t) = \max\{0, t\} \]

\[ \text{LeakyRelu}_\lambda(t) = \begin{cases} 
(1 - \lambda)t, & t \geq 0 \\
\lambda t, & t < 0 
\end{cases} \]
**Large Margin Case: Massart Noise**

**Lemma 1**: No convex surrogate works.

But…

**Lemma 2**: Let $\hat{w}$ be the minimizer of

$$L_\lambda(w) = E_{(x,y) \sim D}[\text{LeakyRelu}_\lambda(-y<w,x>)]$$

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

• $Pr_{(x,y) \sim D}[h_{\hat{w}}(x) \neq y \mid R_T] \leq \eta + \epsilon$.
SUMMARY OF APPROACH: LARGE MARGIN CASE

Lemma 2: Let $\hat{w}$ minimizer of $L_\lambda(w) = \mathbb{E}_{(x,y) \sim D}[\text{LeakyRelu}_\lambda(-y\langle w, x \rangle)]$ 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
- $\Pr_{(x,y) \sim D}[h_{\hat{w}}(x) \neq y \mid R_T] \leq \eta + \epsilon$.

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: Reduction to Large Margin Case**

**Lemma [Dunagan-Vempala’04]**
Using $m = \tilde{O}(d^2 b)$ samples from $D_x$, we can efficiently find an ellipsoid $E$ such that $\Pr_{x \sim D_x}[x \in E] \geq 1/2$ and every point $x$ in $D_x|_E$ satisfies $\langle w, x \rangle^2 \leq \tilde{O}(db) E_{x \sim D_x}[\langle w, x \rangle^2]$ for all $w \in \mathbb{R}^d$.

Leads to sample complexity $\text{poly}(d, b, 1/\epsilon)$

[D-Kane-Tzamos’20]
Different reduction leads to sample complexity $\text{poly}(d, 1/\epsilon)$
**Subsequent Work**

**Theorem [Chen-Koehler-Moitra-Yau’20]**
There exists a polynomial time *proper* learner with same error guarantee.

**Theorem [Chen-Koehler-Moitra-Yau’20]**
Achieving error $\text{OPT} + \epsilon$ requires time $d^{\Omega(\log(1/\epsilon))}$ in the Statistical Query model.
SUMMARY AND MAIN OPEN QUESTION

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

Main Open Question:
Is there a polynomial time learner with misclassification error $O(\text{OPT}) + \epsilon$? If not, can we achieve error $g(\text{OPT}) + \epsilon$?
Question: How about more general/other concept classes?
**Boosting in the Presence of Massart Noise**

**Boosting**: Technique to improve the accuracy of any given “weak” learner.

- **Weak learner**: Algorithm that achieves small advantage $\Pr_{(x,y) \sim D}[h(x) \neq y] \leq 1/2 - \gamma$.

- **Extensively studied in TCS and ML**
  [Schapire’90, Freund’95, Freund-Schapire’97, Mansour-McAllester’02,…]

- **Challenge**: Boosting in the presence of noise
  - RCN [Kalai-Servedio’03]
  - Agnostic setting [Kalai-Mansour-Verbin’08, Feldman’10]

**Question**: Can we design efficient boosting algorithms in the presence of Massart noise?
**Boosting in the Presence of Massart Noise**

**Question:** Can we design boosting algorithms in the presence of Massart noise?

- **Weak learner:** Algorithm that achieves \( \Pr_{(x,y) \sim D}[h(x) \neq y] \leq 1/2 - \gamma \).

**Theorem [D-Impagliazzo-Kane-Lei-Sorrell-Tzamos’20]**

Let \( \mathcal{C} \) be any concept class on \( \mathbb{R}^d \). Suppose there exists a \( \text{poly}(d) \) time weak learner for \( \mathcal{C} \) with advantage \( \gamma \) in the distribution-independent Massart PAC model. There exists a boosting algorithm that learns \( \mathcal{C} \) in the distribution-independent Massart PAC model. The algorithm runs in \( \text{poly}(d, 1/\gamma, 1/\epsilon) \) time and outputs a hypothesis \( h \) such that

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

where \( \eta \) is the upper bound on the Massart noise rate.

- **Remark:** Upper bound above is optimal for black-box boosting.
OUTLINE

• Part I:
  – Distribution-Independent PAC Learning with Massart Noise

• Part II:
  – Distribution-Specific PAC Learning with Massart (and Other) Noise
Can we obtain *near-optimal* error guarantees for broad classes of *structured distributions*?
PRIOR WORK ON DISTRIBUTION-SPECIFIC MASSART LEARNING

Goal: Achieve near-optimal error, i.e., $\Pr_{(x,y) \sim D}[h((x)) \neq y] \leq \text{OPT} + \epsilon$

Equivalently: Approximate the true classifier to any accuracy, i.e., $\Pr_{x \sim D_x}[h((x)) \neq f(x)] \leq \epsilon$.

Uniform Distribution on Unit Sphere:
[Awasthi-Balcan-Haghtalab-Urner’15, Yan-Zhang’17, Zhang-Liang-Charikar’17, Mangoubi, Vishnoi’19]

Log-Concave Distributions:
[Awasthi-Balcan-Haghtalab-Zhang’16]

poly($d, 1/\epsilon, 1/(1 - 2\eta)$)

$d^{2\text{poly}(\frac{1}{\epsilon})}/\text{poly}(\epsilon)$

Open: Is there a polynomial time algorithm for more general distributions?
Theorem [D-Kontonis-Tzamos-Zarifis’20]

There is an efficient algorithm that learns halfspaces in the presence of Massart noise, assuming the distribution on examples is “well-behaved”. The algorithm has sample complexity $N = O(d/\epsilon^4)$, runs in poly($N$) time, and outputs a hypothesis $h$ such that

$$\Pr_{x \sim D_x}[h(x) \neq f(x)] \leq \epsilon$$

where $f$ is the Bayes optimal classifier.

Distribution is well-behaved if its 2-d projections have good concentration and (anti-)anti-concentration.

Corollary: First polynomial-time algorithm for log-concave distributions.

See also concurrent work [Zhang-Shen-Awasthi’20].
**INTUITION: CONVEX VERSUS NON-CONVEX RELAXATION**

- **Population Risk:** Minimize $\mathcal{L}_{0/1}(w) = \mathbf{E}_{(x,y) \sim D}[1\{ -y\langle w, x \rangle > 0 \}]$

- **Convex Relaxation:**

Minimize $\mathcal{L}_G(w) = \mathbf{E}_{(x,y) \sim D}[G(-y\langle w, x \rangle)]$ for some convex $G$.

**Lemma:** No convex surrogate works, even for Gaussian data.

- **Idea:** How about *non*-convex relaxations?

Minimize $\mathcal{L}_{\text{logistic}_\sigma}(w) = \mathbf{E}_{(x,y) \sim D}[\text{logistic}_\sigma(-y\langle w, x \rangle)]$

where $\text{logistic}_\sigma(t) = \frac{1}{1 + e^{-t/\sigma}}$
STRUCTURAL RESULT: STATIONARY POINTS SUFFICE

Non-convex landscape is well-behaved.

Lemma:
For $\sigma \lesssim \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$. Then we have that
\[
\|\nabla L_{\text{logistic}}(w)\|_2 \gtrsim 1 - 2\eta.
\]

Corollary:
Stochastic Gradient Descent (SGD) efficiently converges to a near-optimal solution.
**STRUCTURAL RESULT: INTUITION**

**Lemma:**
For $\sigma \leq \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$. Then we have that

$$\|\nabla L_{\text{logistic}}(w)\|_2 \gtrsim 1 - 2\eta.$$
**Structural Result: Intuition**

**Lemma:**
For \( \sigma \leq \epsilon \sqrt{1 - 2\eta} \) the following holds: Let \( w \) be any halfspace such that \( \theta(w, w^*) \geq \epsilon \). Then we have that
\[
\| \nabla L_{\text{logistic}}(w) \|_2 \gtrsim 1 - 2\eta.
\]
**Structural Result: Intuition**

**Lemma:**
For $\sigma \leq \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$. Then we have that

$$\|\nabla \mathcal{L}_{\text{logistic}_{\sigma}}(w)\|_2 \gtrsim 1 - 2\eta.$$
**Structural Result: Intuition**

**Lemma:**
For $\sigma \leq \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$. Then we have that

$$\|\nabla \mathcal{L}_{\text{logistic}}(w)\|_2 \gtrsim 1 - 2\eta.$$
Structural Result: Intuition

Lemma:
For $\sigma \leq \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$.
Then we have that

$$\| \nabla L_{\text{logistic}}(w) \|_2 \gtrsim 1 - 2\eta.$$
**Structural Result: Intuition**

**Lemma:**
For $\sigma \leq \epsilon \sqrt{1 - 2\eta}$ the following holds: Let $w$ be any halfspace such that $\theta(w, w^*) \geq \epsilon$. Then we have that

$$\|\nabla L_{\text{logistic}, \sigma}(w)\|_2 \geq 1 - 2\eta.$$
**Stronger than Massart? Tsypakov Noise Model**

**Definition (Tsypakov noise with parameters- \((\alpha, A)\))**
The label of each \(x\) is independently flipped with probability \(\eta(x)\), where \(\eta(x)\) unknown and satisfies

\[
\Pr_{x \sim D_x}[\eta(x) \geq 1/2 - t] \leq A t^{1-\alpha}
\]

for all \(t \in (0, 1/2]\).

- Extensively studied
  
  [Mammen-Tsypakov'99, Boucheron-Bouquet-Lugosi'06, Bartlett-Jordan-Mcauliffe’07, Balcan-Broder-Zhang’07, …]

- Sample complexity well-understood.
- No efficient algorithm, for any non-trivial setting.
LEARNING HALFSPACES WITH TSYBAKOV NOISE

First algorithmic progress on this problem.

**Theorem [D-Kontonis-Tzamos-Zarifis’20]**
There exists an algorithm that learns halfspaces to optimal accuracy in the presence of Tsybakov noise, assuming the distribution on examples is “well-behaved”. The algorithm has sample complexity and running time $d^{O(\log^2(1/\epsilon))}$ and outputs a halfspace hypothesis $h$ such that with high probability

$$\Pr_{x \sim D_x}[h(x) \neq f(x)] \leq \epsilon$$

where $f$ is the Bayes optimal halfspace.

No previous bound upper beyond agnostic learning.
**INTUITION: LEARNING VIA CERTIFYING (NON)-OPTIMALITY**

**Easier Problem:** Given candidate $\mathbf{w}$, certify if it is (sub)-optimal.

**Fact:** Let $\mathbf{w}$ be such that $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Then there exists $T : \mathbb{R}^d \to \mathbb{R}^d_+$ such that

$$\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [T(\mathbf{x}) \langle \mathbf{w}, \mathbf{x} \rangle y] < 0.$$

Given an efficient certificate, can find a near-optimal $\mathbf{w}$ via **online convex optimization**.
**STRUCTURAL RESULT: EFFICIENT CERTIFICATE**

**Lemma**
Suppose the distribution on examples is well-behaved. Let \( w \) by any halfspace such that \( \theta(w, w^*) \geq \epsilon \). There exists a degree-\( k \) polynomial \( p : \mathbb{R}^d \to \mathbb{R} \), for \( k = O(\log^2(1/\epsilon)) \), satisfying \( \|p\|_2 = d^{O(k)} \) such that
\[
\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ p(x)^2 1\left\{ 0 \leq \langle w, x \rangle \lesssim \epsilon \right\} y \langle w, x \rangle \right] \lesssim -\epsilon.
\]
Moreover, such a polynomial can be computed with sample complexity and runtime \( d^{O(k)} \).

- Explicit construction via Chebyshev polynomials.
- Efficient computation via SDP.
POLYNOMIAL TIME ALGORITHM?

[D-Kane-Kontonis-Tzamos-Zarifis’20]

More sophisticated algorithm for certificate computation.
CONCLUSIONS AND FUTURE DIRECTIONS

Summary:
• First algorithmic results for distribution-independent learning with Massart noise. Noise-tolerant learning under arbitrary distributions is algorithmically possible!
• Optimal learning with Massart/Tsybakov noise under structured distributions.

Future Directions:
• More general concept classes?
• Other natural semi-random models?
• Applications in data poisoning?