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

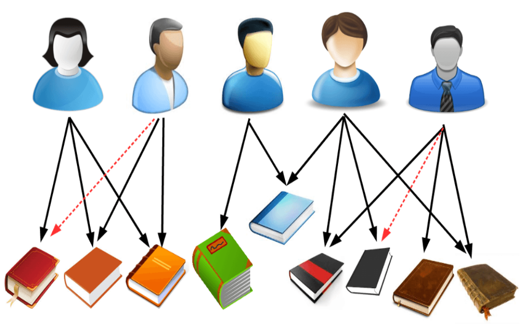
- **Model Misspecification/Robust Statistics**
  [Fisher 1920s, Tukey 1960s, Huber 1960s]

- **Adversarial/Secure ML**

DATA POISONING

Fake Reviews [Mayzlin et al. ‘14]

Recommender Systems



[Li et al. ‘16]

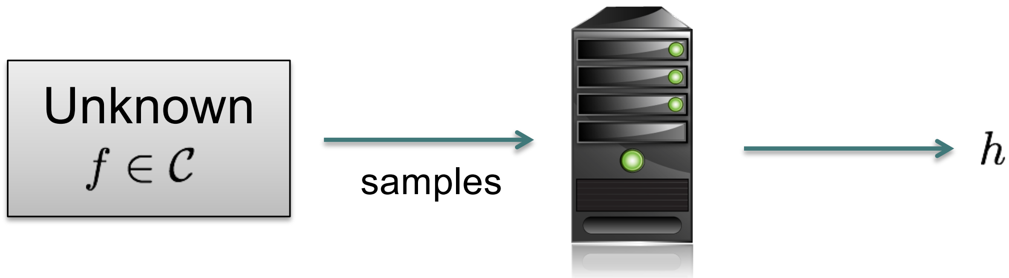
Crowdsourcing

[Wang et al. ‘14]

Malware/spam

[Nelson et al. ‘08]

(Distribution-Independent) PAC Learning



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

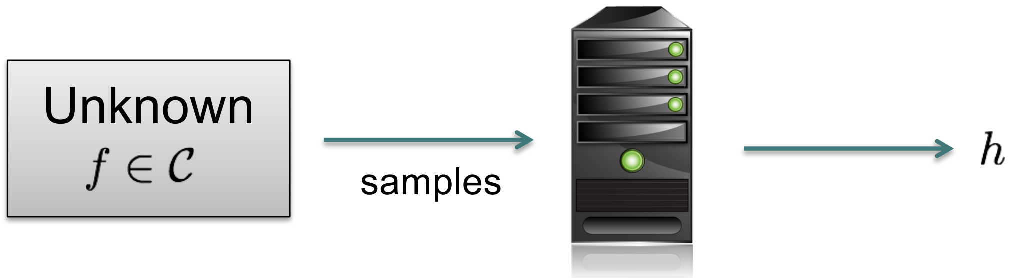
- **Input**: multiset of IID labeled examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ from distribution $\mathcal{D}$ such that: $\mathbf{x}^{(i)} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is **fixed but arbitrary**, and

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

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

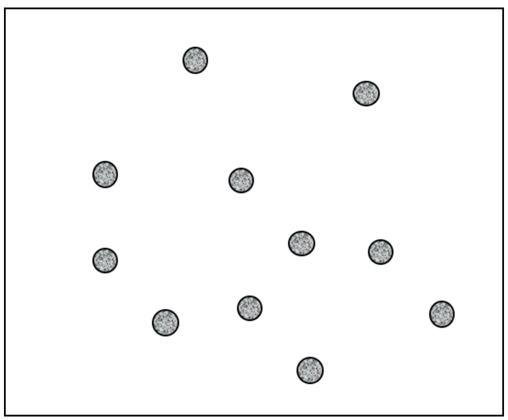
for some fixed unknown target concept $f \in \mathcal{C}$.

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

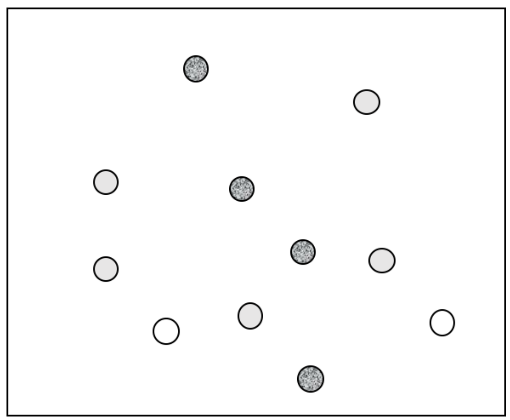
PAC LEARNING WITH *OTHER* NOISE

Massart Noise "in between" Random Classification Noise and Agnostic Model:

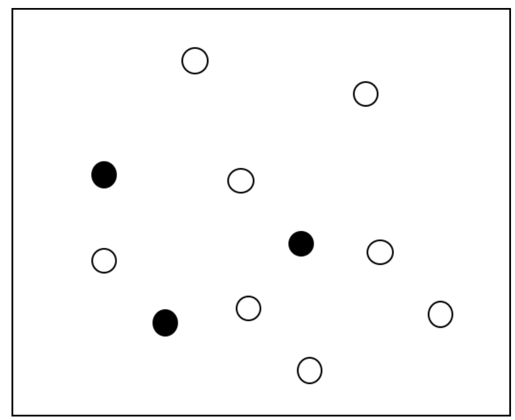
- **Random Classification Noise** (**RCN**) [Angluin-Laird'88]:
  – Special case of Massart noise: For all $\mathbf{x}$, we have that $\eta(\mathbf{x}) = \eta < 1/2$
- **Agnostic Model** [Haussler'92, Kearns-Shapire-Sellie'94]:
  – Adversary can flip *arbitrary* OPT fraction of the labels: $\inf_{f \in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



RCN
Noise Rate **exactly** η



Massart
Noise Rate **at most** η



Agnostic
Arbitrary OPT fraction

OUTLINE

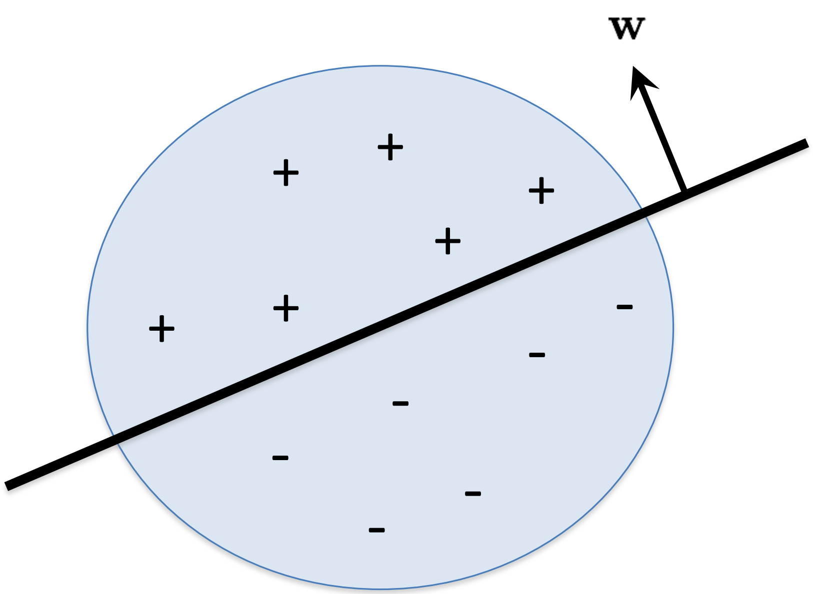
- **Part I:**
  - Distribution-Independent PAC Learning with Massart Noise
- **Part II:**
  - Distribution-Specific PAC Learning with Massart (and Other) Noise

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.

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]

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.

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

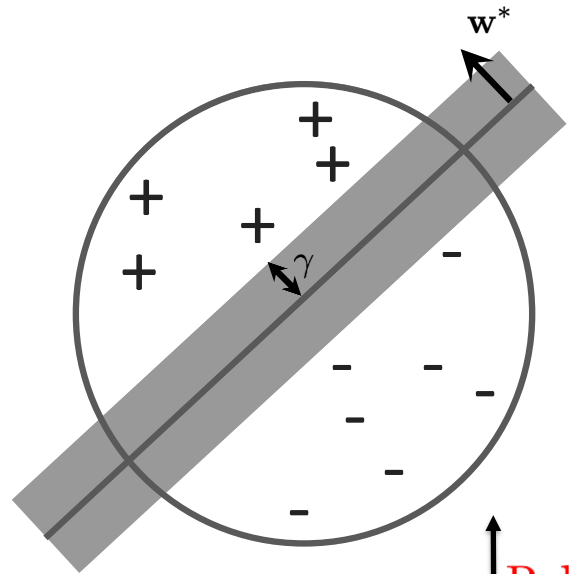
$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \eta + \epsilon$$

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.
- Optimal misclassification error is $\mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x})]$.
- First non-trivial guarantee in sub-exponential time.

INTUITION: LARGE MARGIN CASE

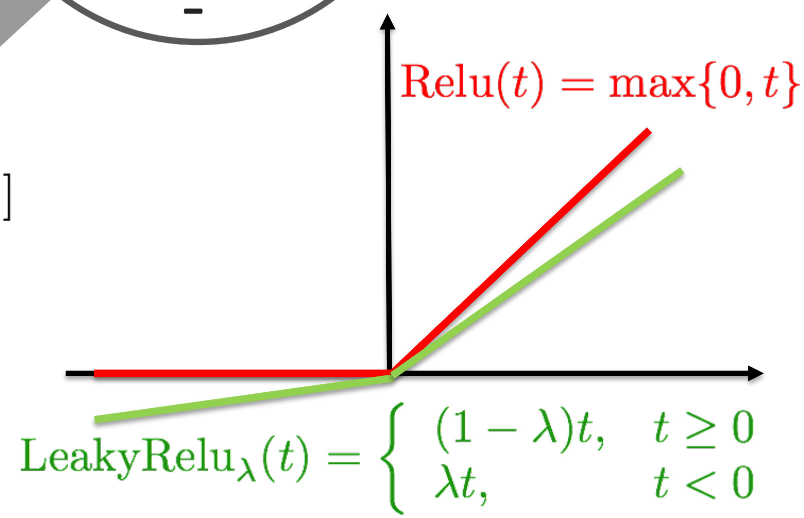


Target vector $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 = 1$
Marginal $\mathcal{D}_{\mathbf{x}}$ satisfies $|\langle \mathbf{w}^*, \mathbf{x} \rangle| \geq \gamma$

- **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…

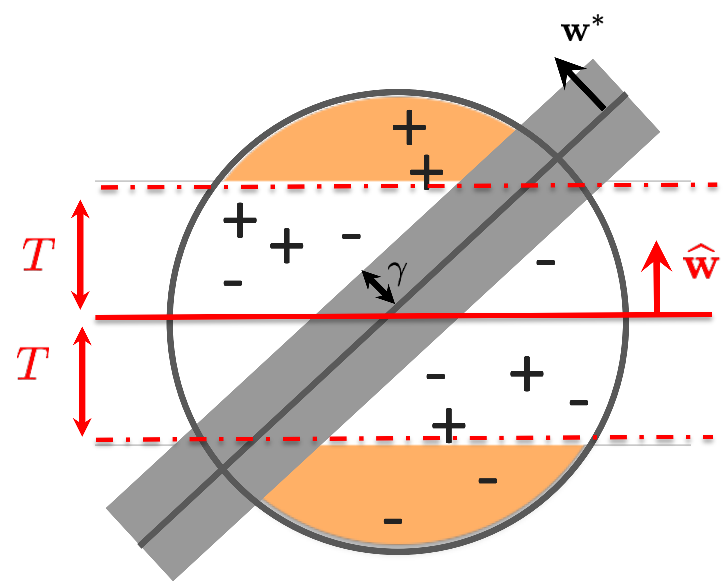
Lemma 2: Let $\widehat{\mathbf{w}}$ be the minimizer of

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

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



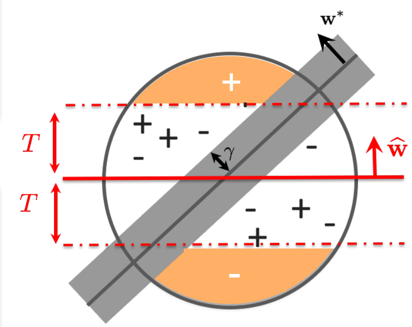
SUMMARY OF APPROACH: LARGE MARGIN CASE

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: REDUCTION TO LARGE MARGIN CASE

Lemma [Dunagan-Vempala'04]
Using $m = \tilde{O}(d^2 b)$ samples from $\mathcal{D}_{\mathbf{x}}$, we can efficiently find an ellipsoid E such that $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\mathbf{x} \in E] \geq 1/2$ and every point $\mathbf{x}$ in $\mathcal{D}_{\mathbf{x}}|_E$ satisfies

$$\langle \mathbf{w}, \mathbf{x} \rangle^2 \leq \tilde{O}(db)\, \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\langle \mathbf{w}, \mathbf{x} \rangle^2]$$

for all $\mathbf{w} \in \mathbb{R}^d$.

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

[D-Kane-Tzamos'20]
Different reduction leads to sample complexity $\mathrm{poly}(d, 1/\epsilon)$

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 $\mathrm{OPT} + \epsilon$ 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 η is an *upper bound* on the noise rate.

Main Open Question:
Is there a polynomial time learner with misclassification error $O(\mathrm{OPT}) + \epsilon$?
If not, can we achieve error $g(\mathrm{OPT}) + \epsilon$?

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 $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq 1/2 - \gamma$.

- **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 $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq 1/2 - \gamma$.

Theorem [D-Impagliazzo-Kane-Lei-Sorrell-Tzamos'20]
Let $\mathcal{C}$ be any concept class on $\mathbb{R}^d$. Suppose there exists a $\mathrm{poly}(d)$ time weak learner for $\mathcal{C}$ with advantage γ 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 $\mathrm{poly}(d, 1/\gamma, 1/\epsilon)$ time and outputs a hypothesis h such that

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

where η is the upper bound on the Massart noise rate.

- **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., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h((x)) \neq y] \leq \mathrm{OPT} + \epsilon$

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

Uniform Distribution on Unit Sphere: $\mathrm{poly}(d, 1/\epsilon, 1/(1-2\eta))$

[Awasthi-Balcan-Haghtalab-Urner'15, Yan-Zhang'17, Zhang-Liang-Charikar'17, Mangoubi, Vishnoi'19]

Log-Concave Distributions: $d^{2^{\mathrm{poly}\left(\frac{1}{1-2\eta}\right)}}/\mathrm{poly}(\epsilon)$

[Awasthi-Balcan-Haghtalab-Zhang'16]

Open: Is there a polynomial time algorithm for more general distributions?

DISTRIBUTION-SPECIFIC MASSART LEARNING OF HALFSPACES

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 $\mathrm{poly}(N)$ time, and outputs a hypothesis h such that

$$\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

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

- **Convex Relaxation**:

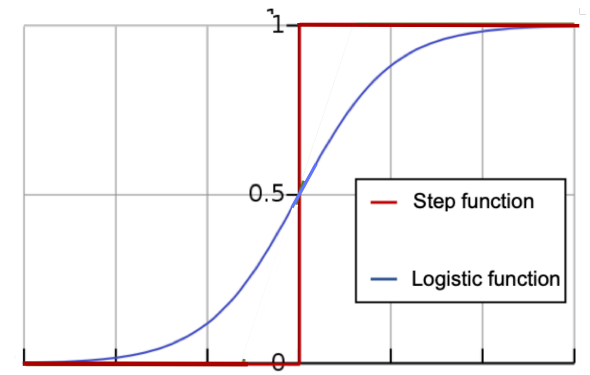
Minimize $\mathcal{L}_G(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[G(-y\langle \mathbf{w}, \mathbf{x}\rangle)\}]$ for some convex G.

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

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

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

where $\mathrm{logistic}_\sigma(t) = \dfrac{1}{1+e^{-t/\sigma}}$



STRUCTURAL RESULT: STATIONARY POINTS SUFFICE

Non-convex landscape is well-behaved.

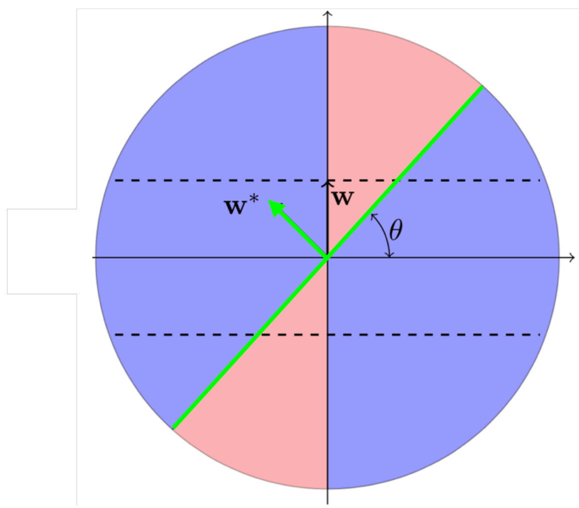
Lemma:
For $\sigma \lesssim \epsilon\sqrt{1-2\eta}$ the following holds: Let w be any halfspace such that $\theta(\mathbf{w}, \mathbf{w}^*) \geq \epsilon$.
Then we have that
$$\|\nabla \mathcal{L}_{\mathrm{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1 - 2\eta \ .$$

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

STRUCTURAL RESULT: INTUITION

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

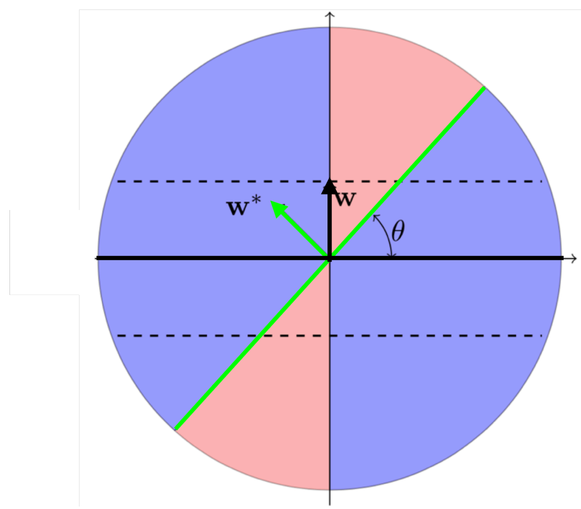
$$\|\nabla \mathcal{L}_{\mathrm{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1-2\eta \ .$$



STRUCTURAL RESULT: INTUITION

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

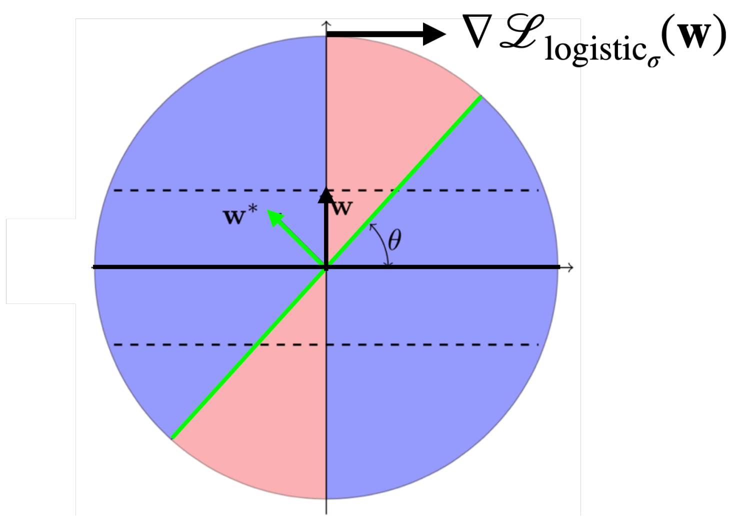
$$\|\nabla \mathcal{L}_{\mathrm{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1 - 2\eta \ .$$



Structural Result: Intuition

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

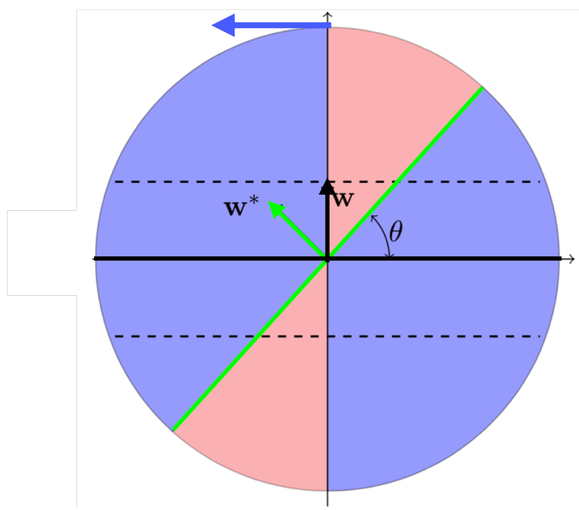
$$\|\nabla \mathcal{L}_{\text{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1 - 2\eta \ .$$



Structural Result: Intuition

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

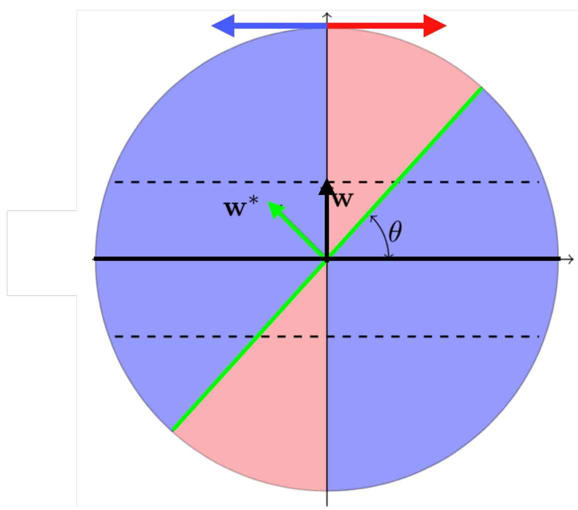
$$\|\nabla \mathcal{L}_{\mathrm{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1 - 2\eta \ .$$



Structural Result: Intuition

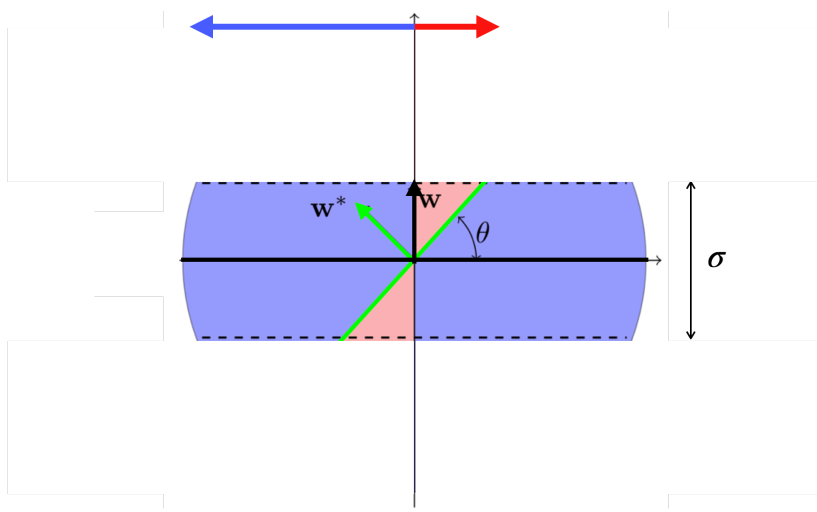
Lemma:
For $\sigma \lesssim \epsilon\sqrt{1-2\eta}$ the following holds: Let w be any halfspace such that $\theta(\mathbf{w}, \mathbf{w}^*) \geq \epsilon$.
Then we have that

$$\|\nabla \mathcal{L}_{\mathrm{logistic}_\sigma}(\mathbf{w})\|_2 \gtrsim 1 - 2\eta \ .$$



Structural Result: Intuition

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



Stronger than Massart? Tsybakov Noise Model

Definition (**Tsybakov noise** with parameters- (α, A))
The label of each $\mathbf{x}$ is independently flipped with probability $\eta(\mathbf{x})$, where $\eta(\mathbf{x})$ unknown and satisfies

$$\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x}) \geq 1/2 - t] \leq A t^{\frac{\alpha}{1-\alpha}}$$

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

- Extensively studied
  [Mammen-Tsybakov'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

$$\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon$$

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 $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \not\equiv \mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. Then there exists $T : \mathbb{R}^d \to \mathbb{R}^d_+$ such that

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

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 $\boldsymbol{w}$ by any halfspace such that $\theta(\mathbf{w}, \mathbf{w}^*) \geq \epsilon$. 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

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\underbrace{p(\mathbf{x})^2\,\mathbf{1}\{0 \leq \langle \mathbf{w}, \mathbf{x}\rangle \lesssim \epsilon\}}_{T(\mathbf{x})}\, y\,\langle \mathbf{w}, \mathbf{x}\rangle\Big] \lesssim -\epsilon\ .$$

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?