Distribution Independent PAC Learning of Halfspaces w/ Massart Noise

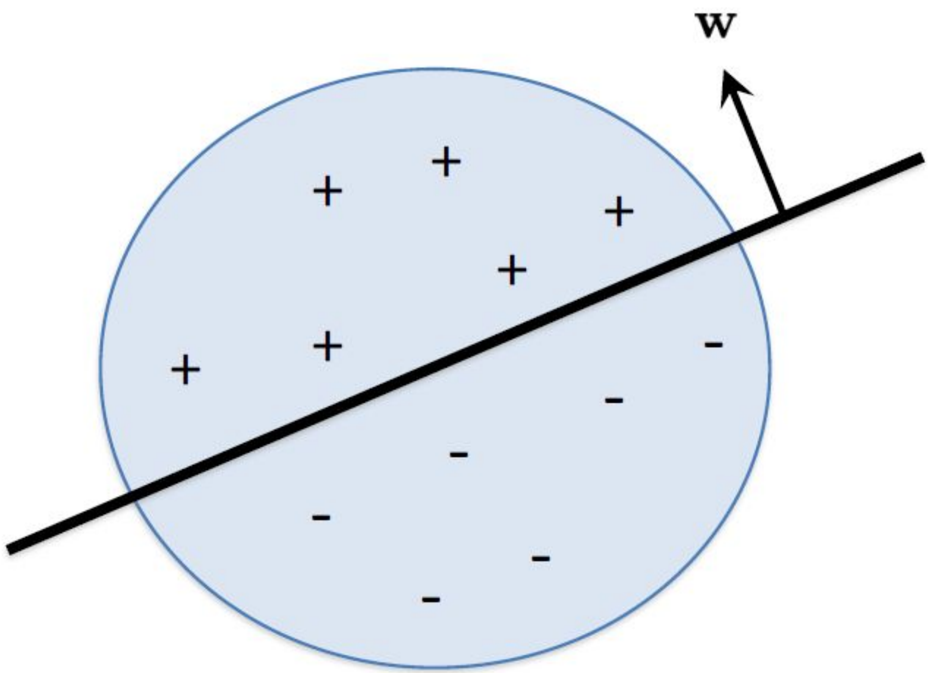
Ilias Diakonikolas, Themis Gouleakis, Christos Tzamos

NeurIPS2019 Outstanding Paper Award

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]

ry

Massart noise

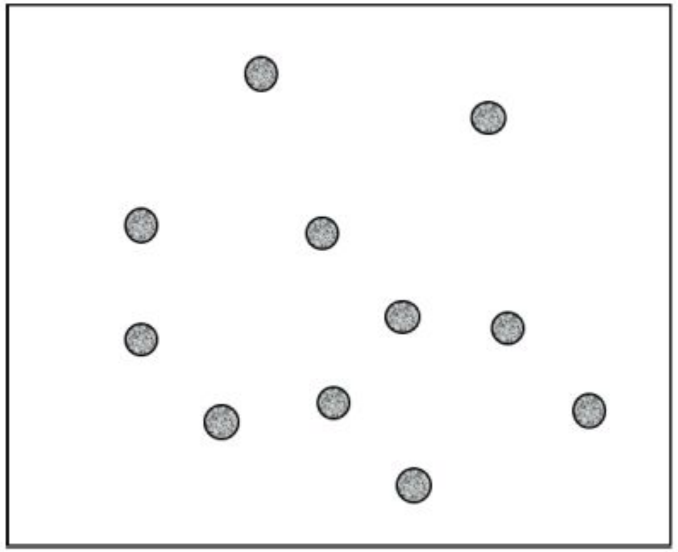
➔ Perturbation of sample label

Given target function $f : R \to \{\pm 1\}$

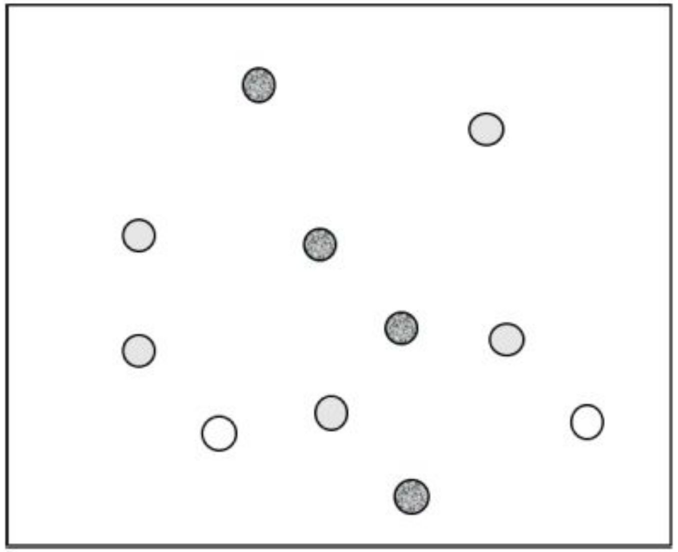
$$y^{(i)} = \begin{cases} f(\mathbf{x}^{(i)}), & \text{with probability } 1 - \eta(\mathbf{x}^{(i)}) \\ -f(\mathbf{x}^{(i)}), & \text{with probability } \eta(\mathbf{x}^{(i)}) \end{cases}$$ where $\eta(\mathbf{x}) : \mathbb{R}^d \to [0, \eta], \eta < 1/2$

Broader Context

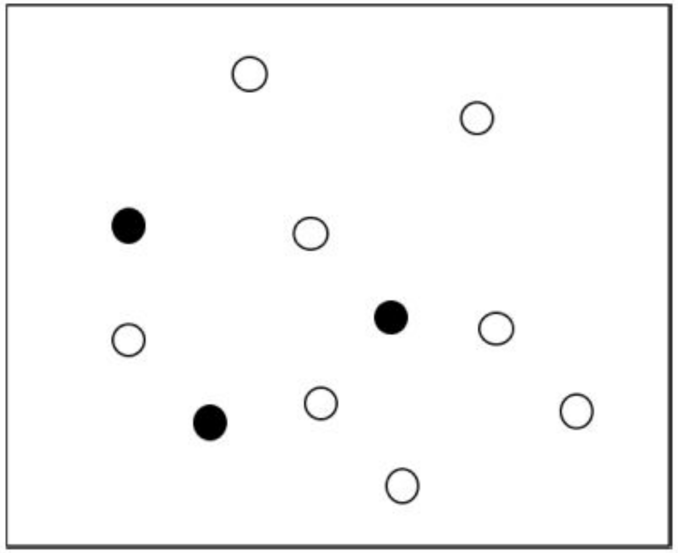
Designing robust estimators w.r.t. natural noise models



RCN
Noise Rate **exactly** η



Massart
Noise Rate **at most** η



Agnostic
Arbitrary η fraction

Broader Context

✔️ Halfspaces efficiently learnable in realizable PAC model [Maass & al, 94]

✔️ Polynomial-time algorithm for learning halfspaces with RCN [Blum & al, 96]

☐ Learning Halfspaces with Massart Noise

❌ Weak agnostic learning of LTFs computationally intractable

Broader Context

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]

Diakonikolas, Gouleakis, Tzamos   Learning Halfspaces with Massart Noise

Broader Context

"Given labeled examples from an unknown Boolean disjunction, corrupted with 1% Massart noise, can we efficiently find a hypothesis that achieves misclassification error 49%"

Problem Setting

$\mathcal{C}$: 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 $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$

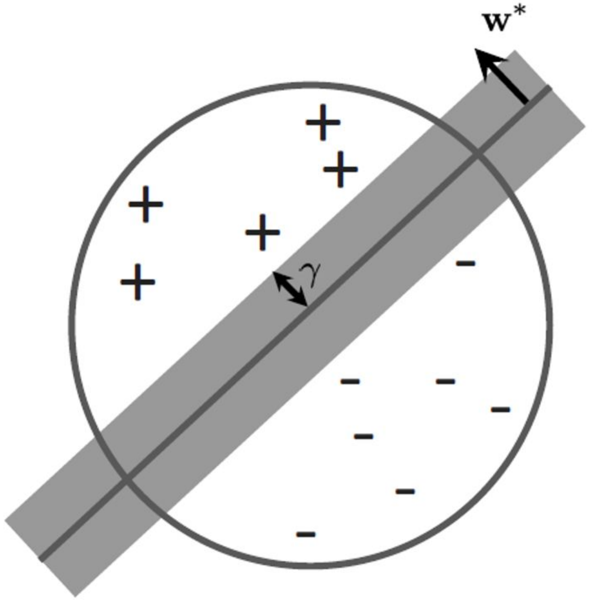
Learning large margin halfspaces

Theorem 2.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{B}_d \times \{\pm 1\}$ such that $\mathcal{D}_{\mathbf{x}}$ satisfies the γ-margin property with respect to $\mathbf{w}^*$ and y is generated by $\mathrm{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ corrupted with Massart noise at rate $\eta < 1/2$. Algorithm 1 uses $\tilde{O}(1/(\gamma^3 \epsilon^5))$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, 1/\epsilon, 1/\gamma)$ time, and returns, with probability $2/3$, a classifier h with misclassification error $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq \eta + \epsilon$.*

<u>Large margin:</u>

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

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



Limitation on loss function

Theorem 3.1: No single convex surrogate can lead to even a weak learner

The problem we're solving is non-convex

$\longrightarrow$ *Theorem 3.1 and proof are not covered in this presentation*

Learning large margin halfspaces

Algorithm 1 Main Algorithm (with margin)

1: Set $S^{(1)} = \mathbb{R}^d$, $\lambda = \eta + \epsilon$, $m = \tilde{O}(\frac{1}{\gamma^2 \epsilon^4})$.

2: Set $i \leftarrow 1$.

3: Draw $O\left((1/\epsilon^2)\log(1/(\epsilon\gamma))\right)$ samples from $\mathcal{D}_{\mathbf{x}}$ to form an empirical distribution $\tilde{\mathcal{D}}_{\mathbf{x}}$.

4: **while** $\mathbf{Pr}_{\mathbf{x}\sim\tilde{\mathcal{D}}_{\mathbf{x}}}\left[\mathbf{x} \in S^{(i)}\right] \geq \epsilon$ **do**

5: Set $\mathcal{D}^{(i)} = \mathcal{D}|_{S^{(i)}}$, the distribution conditional on the unclassified points.

6: Let $L^{(i)}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[\mathrm{LeakyRelu}_\lambda(-y\langle\mathbf{w},\mathbf{x}\rangle)]$

7: Run SGD on $L^{(i)}(\mathbf{w})$ for $\tilde{O}(1/(\gamma^2\epsilon^2))$ iterations to get $\mathbf{w}^{(i)}$ with $\|\mathbf{w}^{(i)}\|_2 = 1$ such that $L^{(i)}(\mathbf{w}^{(i)}) \leq \min_{\mathbf{w}:\|\mathbf{w}\|_2\leq 1} L^{(i)}(\mathbf{w}) + \gamma\epsilon/2$.

8: Draw m samples from $\mathcal{D}^{(i)}$ to form an empirical distribution $\mathcal{D}_m^{(i)}$.

9: Find a threshold $T^{(i)}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_m^{(i)}}[|\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}] \geq \gamma\epsilon$ and the empirical misclassification error, $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_m^{(i)}}[h_{\mathbf{w}^{(i)}}(\mathbf{x}) \neq y \mid |\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}]$, is minimized.

10: Update the unclassified region $S^{(i+1)} \leftarrow S^{(i)} \setminus \{\mathbf{x} : |\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}\}$ and set $i \leftarrow i+1$.

11: Return the classifier $[(\mathbf{w}^{(1)}, T^{(1)}), (\mathbf{w}^{(2)}, T^{(2)}), \cdots]$

Lemmas

Lemma 2.3. *If $\lambda \geq \eta$, then $L(\mathbf{w}^*) \leq -\gamma(\lambda - \mathrm{OPT})$.*

Lemma 2.4 (see, e.g., Theorem 3.4.11 in [Duc16])**.** *Let L be any convex function. Consider the (projected) SGD iteration that is initialized at $\mathbf{w}^{(0)} = \mathbf{0}$ and for every step computes*

$$\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \rho \mathbf{v}^{(t)} \quad \text{and} \quad \mathbf{w}^{(t+1)} = \arg \min_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} \left\| \mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})} \right\|_2 ,$$

where $\mathbf{v}^{(t)}$ is a stochastic gradient such that for all steps $\mathbf{E}[\mathbf{v}^{(t)} | \mathbf{w}^{(t)}] \in \partial L(\mathbf{w}^{(t)})$ and $\left\| \mathbf{v}^{(t)} \right\|_2 \leq 1$. Assume that SGD is run for T iterations with step size $\rho = \frac{1}{\sqrt{T}}$ and let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$. Then, for any $\epsilon, \delta > 0$, after $T = \Omega(\log(1/\delta)/\epsilon^2)$ iterations with probability with probability at least $1 - \delta$ we have that $L(\bar{\mathbf{w}}) \leq \min_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} L(\mathbf{w}) + \epsilon$.

Lemma 2.5. *Consider a vector $\mathbf{w}$ with $L(\mathbf{w}) < 0$. There exists a threshold $T \geq 0$ such that (i) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle \mathbf{w}, \mathbf{x} \rangle| \geq T] \geq \frac{|L(\mathbf{w})|}{2\lambda}$, and (ii) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,|\, |\langle \mathbf{w}, \mathbf{x} \rangle| \geq T] \leq \lambda - \frac{|L(\mathbf{w})|}{2}$.*

Proof of Theorem 2.2 (# Iterations)

We start by noting that with high probability the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$. This can be seen as follows: The empirical probability mass under $\mathcal{D}_m^{(i)}$ of the region $\{\mathbf{x} : |\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| \geq T^{(i)}\}$ removed from $S^{(i)}$ to obtain $S^{(i+1)}$ is at least $\gamma\epsilon$ (Step 9). Since $m = \tilde{O}(1/(\gamma^2\epsilon^4))$, the DKW inequality [DKW56] implies that the true probability mass of this region is at least $\gamma\epsilon/2$ with high probability. By a union bound over $i \leq K = \Theta(\log(1/\epsilon)/(\epsilon\gamma))$, it follows that with high probability we have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq (1-\gamma\epsilon/2)^i$ for all $i \in [K]$. After K iterations, we will have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq \epsilon/3$. Step 3 guarantees that the mass of $S^{(i)}$ under $\tilde{\mathcal{D}}_\mathbf{x}$ is within an additive $\epsilon/3$ of its mass under $\mathcal{D}_\mathbf{x}$, for $i \in [K]$. This implies that the loop terminates after at most K iterations.

Proof of Theorem 2.2 (# Iterations)

We start by noting that with high probability the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$. This can be seen as follows: The empirical probability mass under $\mathcal{D}_m^{(i)}$ of the region $\{\mathbf{x} : |\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| \geq T^{(i)}\}$ removed from $S^{(i)}$ to obtain $S^{(i+1)}$ is at least $\gamma\epsilon$ (Step 9). Since $m = \tilde{O}(1/(\gamma^2\epsilon^4))$, the DKW inequality [DKW56] implies that the true probability mass of this region is at least $\gamma\epsilon/2$ with high probability. By a union bound over $i \leq K = \Theta(\log(1/\epsilon)/(\epsilon\gamma))$, it follows that with high probability we have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq (1-\gamma\epsilon/2)^i$ for all $i \in [K]$. After K iterations, we will have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq \epsilon/3$. Step 3 guarantees that the mass of $S^{(i)}$ under $\tilde{\mathcal{D}}_\mathbf{x}$ is within an additive $\epsilon/3$ of its mass under $\mathcal{D}_\mathbf{x}$, for $i \in [K]$. This implies that the loop terminates after at most K iterations.

2: Set $i \leftarrow 1$.
3: Draw $O\left((1/\epsilon^2)\log(1/(\epsilon\gamma))\right)$ samples from $\mathcal{D}_\mathbf{x}$ to form an empirical distribution $\tilde{\mathcal{D}}_\mathbf{x}$.
4: **while** $\mathbf{Pr}_{\mathbf{x}\sim\tilde{\mathcal{D}}_\mathbf{x}}\left[\mathbf{x} \in S^{(i)}\right] \geq \epsilon$ **do**
5: Set $\mathcal{D}^{(i)} = \mathcal{D}|_{S^{(i)}}$, the distribution conditional on the unclassified points.
6: Let $L^{(i)}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[\mathrm{LeakyRelu}_\lambda(-y\langle\mathbf{w},\mathbf{x}\rangle)]$

Proof of Theorem 2.2 (# Iterations)

We start by noting that with high probability the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$. This can be seen as follows: The empirical probability mass under $\mathcal{D}_m^{(i)}$ of the region $\{\mathbf{x} : |\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| \geq T^{(i)}\}$ removed from $S^{(i)}$ to obtain $S^{(i+1)}$ is at least $\gamma\epsilon$ (Step 9). Since $m = \tilde{O}(1/(\gamma^2\epsilon^4))$, the DKW inequality [DKW56] implies that the true probability mass of this region is at least $\gamma\epsilon/2$ with high probability. By a union bound over $i \leq K = \Theta(\log(1/\epsilon)/(\epsilon\gamma))$, it follows that with high probability we have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq (1-\gamma\epsilon/2)^i$ for all $i \in [K]$. After K iterations, we will have that $\Pr_{\mathcal{D}_\mathbf{x}}[S^{(i+1)}] \leq \epsilon/3$. Step 3 guarantees that the mass of $S^{(i)}$ under $\tilde{\mathcal{D}}_\mathbf{x}$ is within an additive $\epsilon/3$ of its mass under $\mathcal{D}_\mathbf{x}$, for $i \in [K]$. This implies that the loop terminates after at most K iterations.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \qquad x \in \mathbb{R}.$$

$$\Pr\Big(\sup_{x\in\mathbb{R}} (F_n(x) - F(x)) > \varepsilon\Big) \leq e^{-2n\varepsilon^2} \qquad \text{for every } \varepsilon \geq \sqrt{\tfrac{1}{2n}\ln 2}$$

Proof of Theorem 2.2 (# Iterations)

We start by noting that with high probability the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$. This can be seen as follows: The empirical probability mass under $\mathcal{D}_m^{(i)}$ of the region $\{\mathbf{x} : |\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| \geq T^{(i)}\}$ removed from $S^{(i)}$ to obtain $S^{(i+1)}$ is at least $\gamma\epsilon$ (Step 9). Since $m = \tilde{O}(1/(\gamma^2\epsilon^4))$, the DKW inequality [DKW56] implies that the true probability mass of this region is at least $\gamma\epsilon/2$ with high probability. By a union bound over $i \leq K = \Theta(\log(1/\epsilon)/(\epsilon\gamma))$, it follows that with high probability we have that $\Pr_{\mathcal{D}_{\mathbf{x}}}[S^{(i+1)}] \leq (1-\gamma\epsilon/2)^i$ for all $i \in [K]$. After K iterations, we will have that $\Pr_{\mathcal{D}_{\mathbf{x}}}[S^{(i+1)}] \leq \epsilon/3$. Step 3 guarantees that the mass of $S^{(i)}$ under $\tilde{\mathcal{D}}_{\mathbf{x}}$ is within an additive $\epsilon/3$ of its mass under $\mathcal{D}_{\mathbf{x}}$, for $i \in [K]$. This implies that the loop terminates after at most K iterations.

2: Set $i \leftarrow 1$.
3: Draw $O\left((1/\epsilon^2)\log(1/(\epsilon\gamma))\right)$ samples from $\mathcal{D}_{\mathbf{x}}$ to form an empirical distribution $\tilde{\mathcal{D}}_{\mathbf{x}}$.
4: **while** $\mathbf{Pr}_{\mathbf{x}\sim\tilde{\mathcal{D}}_{\mathbf{x}}}\left[\mathbf{x} \in S^{(i)}\right] \geq \epsilon$ **do**
5: Set $\mathcal{D}^{(i)} = \mathcal{D}|_{S^{(i)}}$, the distribution conditional on the unclassified points.
6: Let $L^{(i)}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[\mathrm{LeakyRelu}_\lambda(-y\langle\mathbf{w},\mathbf{x}\rangle)]$

Proof of Theorem 2.2 (SGD)

By Lemma 2.3 and the fact that every $\mathcal{D}^{(i)}$ has margin γ, it follows that the minimizer of the loss $L^{(i)}$ has value less than $-\gamma(\lambda - \mathrm{OPT}^{(i)}) \le -\gamma\epsilon$, as $\mathrm{OPT}^{(i)} \le \eta$ and $\lambda = \eta + \epsilon$. By the guarantees of Lemma 2.4, running SGD in line 7 on $L^{(i)}(\cdot)$ with projection to the unit ℓ_2-ball for $O\left(\log(1/\delta)/(\gamma^2\epsilon^2)\right)$ steps, we obtain a $\mathbf{w}^{(i)}$ such that, with probability at least $1-\delta$, it holds $L^{(i)}(\mathbf{w}^{(i)}) \le -\gamma\epsilon/2$ and $\|\mathbf{w}^{(i)}\|_2 = 1$. Here $\delta > 0$ is a parameter that is selected so that the following claim holds: With probability at least $9/10$, for all iterations i of the while loop we have that $L^{(i)}(\mathbf{w}^{(i)}) \le -\gamma\epsilon/2$. Since the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$, setting δ to $\tilde{\Omega}(\epsilon\gamma)$ and applying a union bound over all iterations gives the previous claim. Therefore, the total number of SGD steps per iteration is $\tilde{O}(1/(\gamma^2\epsilon^2))$. For a given iteration of the while loop, running SGD requires $\tilde{O}(1/(\gamma^2\epsilon^2))$ samples from $\mathcal{D}^{(i)}$ which translate to at most $\tilde{O}\left(1/(\gamma^2\epsilon^3)\right)$ samples from $\mathcal{D}$, as $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x} \in S^{(i)}\right] \ge 2\epsilon/3$.

Proof of Theorem 2.2 (SGD)

By Lemma 2.3 and the fact that every $\mathcal{D}^{(i)}$ has margin γ, it follows that the minimizer of the loss $L^{(i)}$ has value less than $-\gamma(\lambda - \mathrm{OPT}^{(i)}) \leq -\gamma\epsilon$, as $\mathrm{OPT}^{(i)} \leq \eta$ and $\lambda = \eta + \epsilon$. By the guarantees of Lemma 2.4, running SGD in line 7 on $L^{(i)}(\cdot)$ with projection to the unit ℓ_2-ball for $O\left(\log(1/\delta)/(\gamma^2\epsilon^2)\right)$ steps, we obtain a $\mathbf{w}^{(i)}$ such that, with probability at least $1-\delta$, it holds $L^{(i)}(\mathbf{w}^{(i)}) \leq -\gamma\epsilon/2$ and $\|\mathbf{w}^{(i)}\|_2 = 1$. Here $\delta > 0$ is a parameter that is selected so that the following claim holds: With probability at least $9/10$, for all iterations i of the while loop we have that $L^{(i)}(\mathbf{w}^{(i)}) \leq -\gamma\epsilon/2$. Since the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$, setting δ to $\tilde{\Omega}(\epsilon\gamma)$ and applying a union bound over all iterations gives the previous claim. Therefore, the total number of SGD steps per iteration is $\tilde{O}(1/(\gamma^2\epsilon^2))$. For a given iteration of the while loop, running SGD requires $\tilde{O}(1/(\gamma^2\epsilon^2))$ samples from $\mathcal{D}^{(i)}$ which translate to at most $\tilde{O}\left(1/(\gamma^2\epsilon^3)\right)$ samples from $\mathcal{D}$, as $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x} \in S^{(i)}\right] \geq 2\epsilon/3$.

$$T = \Omega(log(1/\delta)/\epsilon^2)$$ SGD Steps

Proof of Theorem 2.2 (SGD)

By Lemma 2.3 and the fact that every $\mathcal{D}^{(i)}$ has margin γ, it follows that the minimizer of the loss $L^{(i)}$ has value less than $-\gamma(\lambda - \mathrm{OPT}^{(i)}) \leq -\gamma\epsilon$, as $\mathrm{OPT}^{(i)} \leq \eta$ and $\lambda = \eta + \epsilon$. By the guarantees of Lemma 2.4, running SGD in line 7 on $L^{(i)}(\cdot)$ with projection to the unit ℓ_2-ball for $O\left(\log(1/\delta)/(\gamma^2\epsilon^2)\right)$ steps, we obtain a $\mathbf{w}^{(i)}$ such that, with probability at least $1-\delta$, it holds $L^{(i)}(\mathbf{w}^{(i)}) \leq -\gamma\epsilon/2$ and $\|\mathbf{w}^{(i)}\|_2 = 1$. Here $\delta > 0$ is a parameter that is selected so that the following claim holds: With probability at least $9/10$, for all iterations i of the while loop we have that $L^{(i)}(\mathbf{w}^{(i)}) \leq -\gamma\epsilon/2$. Since the total number of iterations is $\tilde{O}(1/(\gamma\epsilon))$, setting δ to $\tilde{\Omega}(\epsilon\gamma)$ and applying a union bound over all iterations gives the previous claim. Therefore, the total number of SGD steps per iteration is $\tilde{O}(1/(\gamma^2\epsilon^2))$. For a given iteration of the while loop, running SGD requires $\tilde{O}(1/(\gamma^2\epsilon^2))$ samples from $\mathcal{D}^{(i)}$ which translate to at most $\tilde{O}\left(1/(\gamma^2\epsilon^3)\right)$ samples from $\mathcal{D}$, as $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x} \in S^{(i)}\right] \geq 2\epsilon/3$.

Proof of Theorem 2.2 (Threshold)

Lemma 2.5 implies that there exists $T \geq 0$ such that: (a) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \geq \gamma\epsilon$, and (b)) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,\big|\, |\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \leq \eta + \epsilon$. Line 9 of Algorithm 1 estimates the threshold using samples. By the DKW inequality [DKW56], we know that with $m = \tilde{O}(1/(\gamma^2\epsilon^4))$ samples we can estimate the CDF within error $\gamma\epsilon^2$ with probability $1 - \mathrm{poly}(\epsilon, \gamma)$. This suffices to estimate the probability mass of the region within additive $\gamma\epsilon^2$ and the misclassification error within $\epsilon/3$. This is satisfied for all iterations with constant probability.

Lemma 2.5. *Consider a vector* $\mathbf{w}$ *with* $L(\mathbf{w}) < 0$. *There exists a threshold* $T \geq 0$ *such that (i)* $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \geq \frac{|L(\mathbf{w})|}{2\lambda}$, *and (ii)* $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,\big|\, |\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \leq \lambda - \frac{|L(\mathbf{w})|}{2}$.

Proof of Theorem 2.2 (Threshold)

Lemma 2.5 implies that there exists $T \geq 0$ such that: (a) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \geq \gamma\epsilon$, and (b)) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,\big|\, |\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \leq \eta + \epsilon$. Line 9 of Algorithm 1 estimates the threshold using samples. By the DKW inequality [DKW56], we know that with $m = \tilde{O}(1/(\gamma^2\epsilon^4))$ samples we can estimate the CDF within error $\gamma\epsilon^2$ with probability $1 - \mathrm{poly}(\epsilon, \gamma)$. This suffices to estimate the probability mass of the region within additive $\gamma\epsilon^2$ and the misclassification error within $\epsilon/3$. This is satisfied for all iterations with constant probability.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \qquad x \in \mathbb{R}.$$

$$\Pr\left(\sup_{x\in\mathbb{R}} (F_n(x) - F(x)) > \varepsilon\right) \leq e^{-2n\varepsilon^2} \qquad \text{for every } \varepsilon \geq \sqrt{\tfrac{1}{2n}\ln 2}$$

Proof of Theorem 2.2 (total sample complexity)

In summary, with high constant success probability, Algorithm 1 runs for $\tilde{O}(1/(\gamma\epsilon))$ iterations and draws $\tilde{O}(1/(\gamma^2\epsilon^4))$ samples per round for a total of $\tilde{O}(1/(\gamma^3\epsilon^5))$ samples. As each iteration runs in polynomial time, the total running time follows.

When the while loop terminates, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x} \in S^{(i)}] \leq 4\epsilon/3$, i.e., we will have accounted for at least a $(1-4\epsilon/3)$-fraction of the total probability mass. Since our algorithm achieves misclassification error at most $\eta + 4\epsilon/3$ in all the regions we accounted for, its total misclassification error is at most $\eta + 8\epsilon/3$. Rescaling ϵ by a constant factor gives Theorem 2.2. □

Proof of Theorem 2.2 (misclassification error)

In summary, with high constant success probability, Algorithm 1 runs for $\tilde{O}(1/(\gamma\epsilon))$ iterations and draws $\tilde{O}(1/(\gamma^2\epsilon^4))$ samples per round for a total of $\tilde{O}(1/(\gamma^3\epsilon^5))$ samples. As each iteration runs in polynomial time, the total running time follows.

When the while loop terminates, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x} \in S^{(i)}] \leq 4\epsilon/3$, i.e., we will have accounted for at least a $(1-4\epsilon/3)$-fraction of the total probability mass. Since our algorithm achieves misclassification error at most $\eta + 4\epsilon/3$ in all the regions we accounted for, its total misclassification error is at most $\eta + 8\epsilon/3$. Rescaling ϵ by a constant factor gives Theorem 2.2. $\square$

Proof of Lemma 2.3

Lemma 2.3. *If* $\lambda \geq \eta$, *then* $L(\mathbf{w}^*) \leq -\gamma(\lambda - \mathrm{OPT})$.

Proof. For any fixed $\mathbf{x}$, using Claim 2.1, we have that

$$\ell(\mathbf{w}^*, \mathbf{x}) = (\mathrm{err}(\mathbf{w}^*, \mathbf{x}) - \lambda)|\langle \mathbf{w}^*, \mathbf{x}\rangle| = (\eta(\mathbf{x}) - \lambda)|\langle \mathbf{w}^*, \mathbf{x}\rangle| \leq -\gamma(\lambda - \eta(\mathbf{x})) ,$$

since $|\langle \mathbf{w}^*, \mathbf{x}\rangle| \geq \gamma$ and $\eta(\mathbf{x}) - \lambda \leq 0$. Taking expectation over $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, the statement follows. $\square$

Claim 2.1. *For any* $\mathbf{w}, \mathbf{x}$, *we have that* $\ell(\mathbf{w}, \mathbf{x}) = (\mathrm{err}(\mathbf{w}, \mathbf{x}) - \lambda)|\langle \mathbf{w}, \mathbf{x}\rangle|$.

<u>Proxy loss</u>

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

<u>Optimal misclassification error</u>

$$\mathrm{OPT} = \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\mathrm{err}(\mathbf{w}^*, \mathbf{x})] = \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\eta(\mathbf{x})]$$

Proof of Lemma 2.5

Lemma 2.5. *Consider a vector* $\mathbf{w}$ *with* $L(\mathbf{w}) < 0$. *There exists a threshold* $T \geq 0$ *such that (i)* $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \geq \frac{|L(\mathbf{w})|}{2\lambda}$, *and (ii)* $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,\big|\, |\langle \mathbf{w}, \mathbf{x}\rangle| \geq T] \leq \lambda - \frac{|L(\mathbf{w})|}{2}$.

Claim 2.1. *For any* $\mathbf{w}, \mathbf{x}$, *we have that* $\ell(\mathbf{w}, \mathbf{x}) = (\mathrm{err}(\mathbf{w}, \mathbf{x}) - \lambda)|\langle \mathbf{w}, \mathbf{x}\rangle|$.

Proof of Lemma 2.5

For a T drawn uniformly at random in $[0,1]$, we have that:

1.

$$\int_0^1 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda+\zeta)1_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq T}]dT = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda)|\langle\mathbf{w},\mathbf{x}\rangle|] + \zeta\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\langle\mathbf{w},\mathbf{x}\rangle|]$$
$$\leq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\ell(\mathbf{w},\mathbf{x})] + \zeta = L(\mathbf{w}) + \zeta = L(\mathbf{w})/2 < 0\,.$$

Thus, there exists a $\bar{T}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda+\zeta)\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq\bar{T}}] \leq 0$. Consider the minimum such $\bar{T}$. Then we have

2.

$$\int_{\bar{T}}^1 \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda+\zeta)\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq T}]dT \geq -\lambda\cdot\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle\mathbf{w},\mathbf{x}\rangle|\geq\bar{T}]\,.$$

By definition of $\bar{T}$, it must be the case that $\int_0^{\bar{T}} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda+\zeta)\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq T}]dT \geq 0$. Therefore,

3.

$$\frac{L(\mathbf{w})}{2} \geq \int_{\bar{T}}^1 \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathrm{err}(\mathbf{w},\mathbf{x})-\lambda+\zeta)\mathbb{1}_{|\langle\mathbf{w},\mathbf{x}\rangle|\geq T}]dT \geq -\lambda\cdot\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle\mathbf{w},\mathbf{x}\rangle|\geq\bar{T}]\,,$$

which implies that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\langle\mathbf{w},\mathbf{x}\rangle|\geq\bar{T}] \geq \frac{|L(\mathbf{w})|}{2\lambda}$. This completes the proof of Lemma 2.5.

$\square$

Generalization

Definition 2.7 ([DV04a]). We call a point $\mathbf{x}$ in the support of a distribution $\mathcal{D}_\mathbf{x}$ a β-outlier, if there exists a vector $\mathbf{w} \in \mathbb{R}^d$ such that $\langle \mathbf{w}, \mathbf{x} \rangle^2 \leq \beta \, \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\langle \mathbf{w}, \mathbf{x} \rangle^2]$.

Lemma 2.8 (Rephrasing of Theorem 3 of [DV04a]). *Using $m = \tilde{O}(d^2 b)$ samples from $\mathcal{D}_\mathbf{x}$, one can identify with high probability an ellipsoid E such that $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbf{x} \in E] \geq \frac{1}{2}$ and $\mathcal{D}_\mathbf{x}|_E$ has no $\Gamma^{-1} = \tilde{O}(db)$-outliers.*

Theorem 2.9. *Let $\mathcal{D}$ be a distribution over $(d+1)$-dimensional labeled examples with bit-complexity b, generated by an unknown halfspace corrupted by Massart noise at rate $\eta < 1/2$. Algorithm 2 uses $\tilde{O}(d^3 b^3 / \epsilon^5)$ samples, runs in $\mathrm{poly}(d, 1/\epsilon, b)$ time, and returns, with probability $2/3$, a classifier h with misclassification error $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq \eta + \epsilon$.*

Generalization

Algorithm 2 Main Algorithm (general case)

Set $S^{(1)} = \mathbb{R}^d$, $\lambda = \eta + \epsilon$, $\Gamma^{-1} = \tilde{O}(db)$, $m = \tilde{O}(\frac{1}{\Gamma^2 \epsilon^4})$.
Set $i \leftarrow 1$.
Draw $O\left((1/\epsilon^2)\log(1/(\epsilon\Gamma))\right)$ samples from $\mathcal{D}_{\mathbf{x}}$ to form an empirical distribution $\tilde{\mathcal{D}}_{\mathbf{x}}$.
while $\mathbf{Pr}_{\mathbf{x}\sim\tilde{\mathcal{D}}_{\mathbf{x}}}\left[\mathbf{x} \in S^{(i)}\right] \geq \epsilon$ **do**
    Run the algorithm of Lemma 2.8 to remove Γ^{-1}-outliers from the distribution $\mathcal{D}_{S^{(i)}}$ by filtering points outside the ellipsoid $E^{(i)}$.
    Let $\Sigma^{(i)} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}|_{S^{(i)}}}[\mathbf{x}\mathbf{x}^T]$ and set $\mathcal{D}^{(i)} = \Gamma\Sigma^{(i)-1/2} \cdot \mathcal{D}|_{S^{(i)}\cap E^{(i)}}$ be the distribution $\mathcal{D}|_{S^{(i)}\cap E^{(i)}}$ brought in isotropic position and rescaled by Γ so that all vectors have ℓ_2-norm at most 1.
    Let $L^{(i)}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}^{(i)}}[\mathrm{LeakyRelu}_\lambda(-y\langle\mathbf{w},\mathbf{x}\rangle)]$
    Run SGD on $L^{(i)}(\mathbf{w})$ for $\tilde{O}(1/(\Gamma^2\epsilon^2))$ iterations, to get $\mathbf{w}^{(i)}$ with $\|\mathbf{w}^{(i)}\|_2 = 1$ such that $L^{(i)}(\mathbf{w}^{(i)}) \leq \min_{\mathbf{w}:\|\mathbf{w}\|_2\leq 1} L^{(i)}(\mathbf{w}) + \Gamma\epsilon/2$.
    Draw m samples from $\mathcal{D}^{(i)}$ to form an empirical distribution $\mathcal{D}_m^{(i)}$.
    Find a threshold $T^{(i)}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_m^{(i)}}[|\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}] \geq \Gamma\epsilon$ and the empirical misclassification error, $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_m^{(i)}}[h_{\mathbf{w}}(\mathbf{x}) \neq y \,\big|\, |\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}]$, is minimized.
    Revert the linear transformation by setting $\mathbf{w}^{(i)} \leftarrow \Gamma\Sigma^{(i)-1/2} \cdot \mathbf{w}^{(i)}$.
    Update the unclassified region $S^{(i+1)} \leftarrow S^{(i)} \setminus \{\mathbf{x} : \mathbf{x} \in E^{(i)} \wedge |\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq T^{(i)}\}$ and set $i \leftarrow i+1$.
Return the classifier $[(\mathbf{w}^{(1)}, T^{(1)}, E^{(1)}), (\mathbf{w}^{(2)}, T^{(2)}, E^{(2)}), \cdots]$

Conclusions

- Contributions
  - first non-trivial learning algorithm for the class of halfspaces in the distribution-free PAC model with Massart noise
- Future work
  - Proper learner?
  - Different noise model (closer to agnostic setting?)

Quiz Questions

- Which of the following noises lead to an intractable classification problem?
  - Massart noise
  - Random Classification Noise
  - Agnostic Noise
- Given an example (x,y), a response corrupted with \eta Massart noise is
  - y with probability \eta, -y otherwise
  - y with probability <= \eta(x) <\ \eta, -y otherwise (\eta \in [0,1])
  - y with probability <= \eta(x) <\ \eta, -y otherwise (\eta \in [0,½[)
  - -y if example (x,y) is within adversarially chosen set of samples representing \eta fraction of all samples, and y if not
- Minimizing a single convex surrogate over a space with Massart noise can lead to a weak learner. [True/False]
- Choose the options describing a gamma-margin halfspace (w^* describes the true hyperplane)
  - |\langle w^*, x\rangle| [\geq/\leq] \gamma for [all/some] x in the support
- Given a vector w representing a half-space, how do you predict the label of a given sample x. [write answer]
  - sign(\langle w,x\rangle)