Noisy PAC Learning of Halfspaces

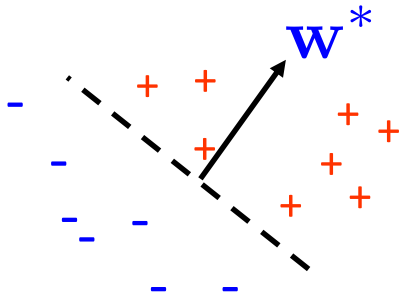
PRANJAL AWASTHI (RUTGERS)

Plan

- Survey of techniques used in robust PAC learning of halfspaces.
- Recent developments and open problems.

PAC learning of halfspaces

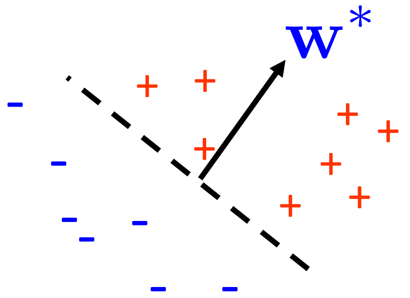
Given measurements $(x, y = sign(w^* \cdot x))$ approximately recover w^*.



- How is x generated?
  - $x \sim D$, where D is an arbitrary distribution over $\mathbb{R}^d$

PAC learning of halfspaces

Given measurements $(x, y = sign(w^* \cdot x))$ approximately recover w^*.



- What is approximate recovery?
  - Input: $(x_1, y_1), (x_2, y_2,), \dots (x_n, y_n)$ generated i.i.d. from D.
  - Output: $h: \mathbb{R}^d \to \{\pm 1\}$ such that $\Pr_{x \sim D}\left(h(x) \neq sign(w^* \cdot x)\right) \leq \epsilon$
  - Want runtime $poly(n)$ and $\mathrm{n} = \tilde{O}(\frac{d}{\epsilon})$

PAC learning of halfspaces

Constraints

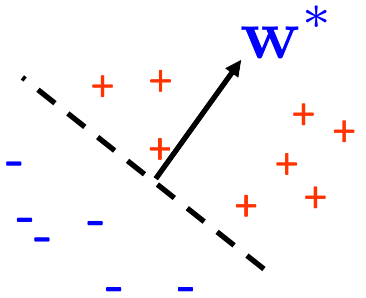
$(x_1, +)$	$w \cdot x_1 > 0$
$(x_2, -)$	$w \cdot x_2 < 0$
….	….
$(x_n, +)$	$w \cdot x_n > 0$

Theorem[VC]

If $\mathrm{n} = \tilde{O}(\frac{d}{\epsilon})$, then w.h.p. $err(w) \leq \epsilon$

PAC learning with malicious noise

Given corrupted measurements $(x, sign(w^* \cdot x))$ approximately recover w^*.



- In each draw
  - $\big(x, sign(w^* \cdot x)\big)$, where $x \sim D$ w.p. $1 - \eta$
  - Arbitrary (x, y) w.p. η

- How much noise $\eta(\epsilon)$ can be tolerated?

PAC learning with malicious noise

- Cannot tolerate $\eta > \frac{\epsilon}{1+\epsilon}$ [Kearns, Li’88]
- Can tolerate $\eta = O(\epsilon)$ (in principle)

Theorem[VC]:

If $\mathrm{n} = \tilde{O}\left(\frac{\mathrm{d}}{\epsilon^2}\right)$ and w has $\frac{\epsilon}{2}$ error on data then w.h.p. $\mathrm{err}(\mathrm{w}) \leq \epsilon$

Minimizing error on noisy data is NP-hard!

PAC learning with malicious noise

- Cannot tolerate $\eta > \frac{\epsilon}{1+\epsilon}$ [Kearns, Li'88]
- Efficiently: Can tolerate $\eta = O(\frac{\epsilon}{d})$
- Proof:
  - Need $\sim \frac{d}{\epsilon}\log(1/\epsilon)$ examples to learn in the noise free case.
  - $\mathbb{P}(no\ example\ is\ corrupted) = (1-\eta)^{\frac{d}{\epsilon}\log(1/\epsilon)} \geq poly(\epsilon)$

PAC learning with malicious noise

- Cannot tolerate $\eta > \frac{\epsilon}{1+\epsilon}$ [Kearns, Li'88]
- Efficiently: Can tolerate $\eta = O(\frac{\epsilon}{d})$

[Daniely'16]: Complexity theoretic evidence that can't get $\eta = O(\epsilon)$ in poly time for a general distribution.

Open: Improve to $\eta = \frac{\epsilon}{d^{0.99}}$

PAC learning with malicious noise (D=Uniform Distribution over S_{d-1})

- Efficiently: Can tolerate
  - $\eta = O(\frac{\epsilon}{d^{\frac{1}{4}}})$ [KKMS'05]
  - $\eta = O(\frac{\epsilon^2}{\log(\frac{d}{\epsilon})})$ [KLS'09]
  - $\eta = O(\epsilon)$ [ABL'14]

PAC learning with malicious noise (D=Isotropic log-concave distribution)

- Efficiently: Can tolerate
  - $\eta = O(\frac{\epsilon^3}{\log^2\left(\frac{d}{\epsilon}\right)})$ [KLS'09]
  - $\eta = O(\epsilon)$ [ABL'14]

Rest of the Talk

- Survey of techniques from [KLS’09, ABL’14]
  - Introduce a margin based technique
- Non-malicious noise models
- Recent developments

PAC Learning with malicious noise

- Cannot tolerate $\eta > \frac{\epsilon}{1+\epsilon}$ [Kearns, Li'88]
- Can tolerate $\eta = O(\epsilon)$ (in principle)

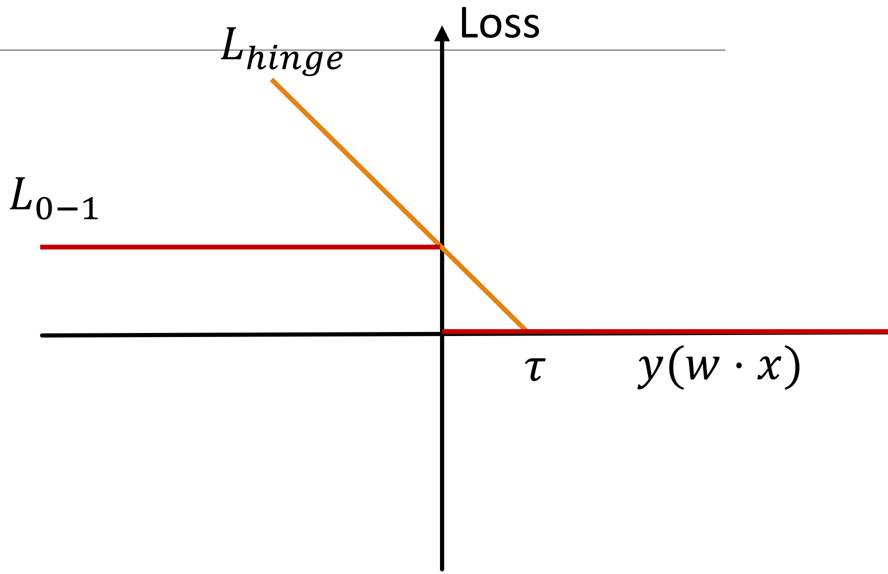
Theorem[VC]:

If $\mathrm{m} = \tilde{O}\left(\frac{\mathrm{d}}{\epsilon^2}\right)$ and w has $\frac{\epsilon}{2}$ error on data then w.h.p. $\mathrm{err}(\mathrm{w}) \leq \epsilon$

Minimizing error on noisy data is NP-hard!

Hinge Loss

$$L_{hinge}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$



Candidate Algorithm

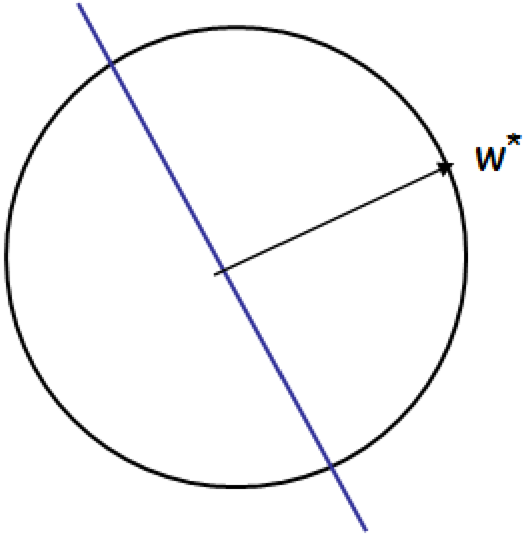
- Sample a set S of noisy examples.
- Output w of small hinge loss over S.

Uniform Distribution over S_{d-1}

$$L_{hinge}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$

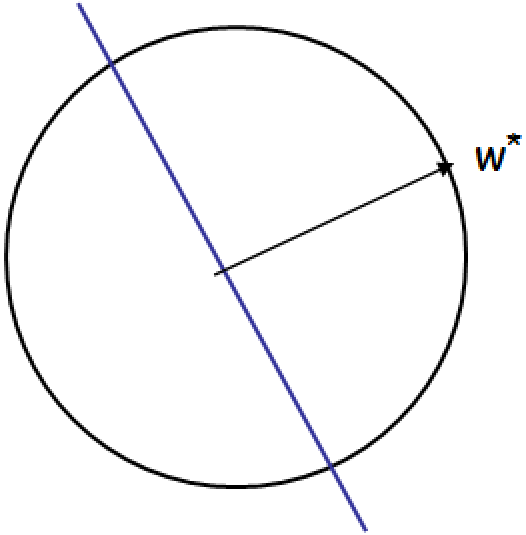
$E\left[L_{hinge}\left(w \quad , x, y\right)\right]$: Expected hinge loss over clean dist.

$\tilde{E}\left[L_{hinge}\left(w \quad , x, y\right)\right]$: Empirical hinge loss over noisy samples.



Uniform Distribution over S_{d-1}

$$L_{hinge}(w,x,y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$



With probability $\geq 1 - \delta$,

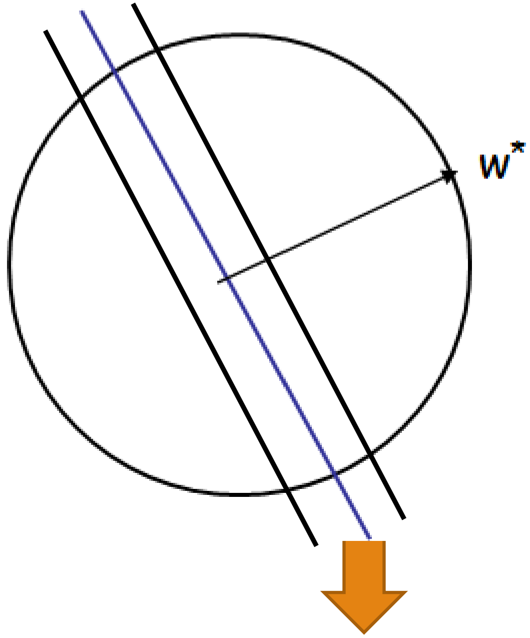
$$\forall w, \left|E\left[L_{hinge}(w,x,y)\right] - \tilde{E}\left[L_{hinge}(w,x,y)\right]\right| \leq O\left(\frac{1}{\tau}\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}\right) + \eta\left(1 + \frac{1}{\tau}\right)$$

Uniform Distribution over S_{d-1}

$$L_{hinge}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$

$$\Pr[|w^* \cdot x| \leq \tau] \leq O(\tau\sqrt{d})$$

$$E\left[L_{hinge}(w^*, x, y)\right] \leq O(\tau\sqrt{d})$$



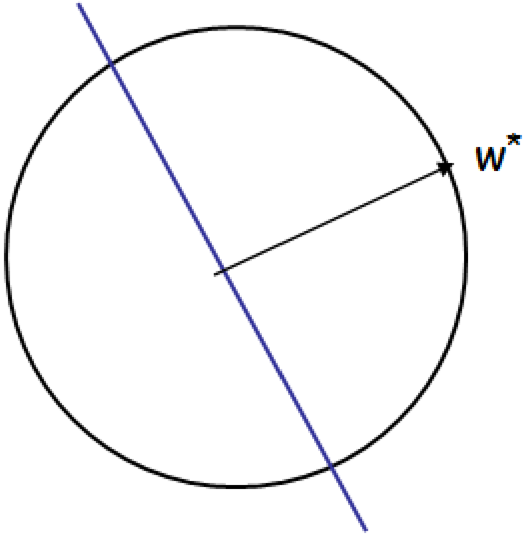
$$|w^* \cdot x| \leq \tau$$

Uniform Distribution over S_{d-1}

If w is the minimizer of hinge loss over noisy data, then w.h.p.

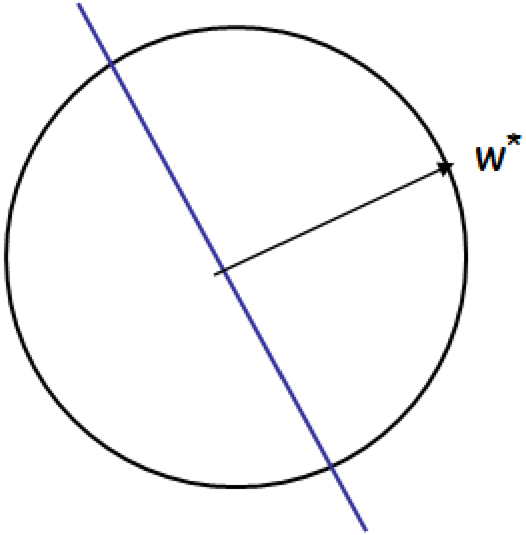
$$E\left[L_{hinge}(w,x,y)\right] \leq O(\tau\sqrt{d} + \epsilon + \eta(1 + \frac{1}{\tau}))$$

Set $\tau = \frac{\epsilon}{c\sqrt{d}}$ ➡ $\eta = \frac{\epsilon^2}{\sqrt{d}}$



Uniform Distribution over S_{d-1}

$$L_{hinge}(w,x,y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$



With probability $\geq 1 - \delta$,

$$\forall w, \left|E\left[L_{hinge}(w,x,y)\right] - \tilde{E}\left[L_{hinge}(w,x,y)\right]\right| \leq O\left(\frac{1}{\tau}\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}\right) + \eta\left(1+\frac{1}{\tau}\right)$$

[KKMS'05]

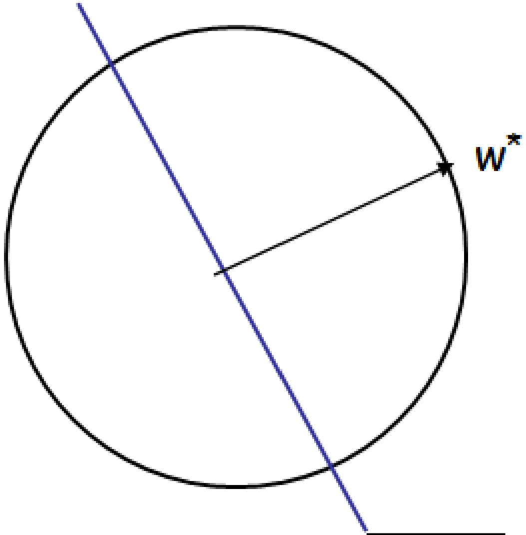
- Sample a set S of noisy examples.
- Remove any pair that is too close to each other
  - Distance less than $\sqrt{2 - c\frac{\log n}{d}}$
- Output w of small hinge loss over S.

[KLS'09]

- Sample a set S of noisy examples.
- While there exists a direction u such that $\mathbb{E}_S[(u.x)^2] > \frac{10 \log n}{d}$
  - Remove any $x \in S$ such that $(u.x)^2 > \frac{10 \log n}{d}$
- Output w of small hinge loss over S.

[KLS'09]

$$L_{hinge}(w,x,y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$



With probability $\geq 1 - \delta$,

$$\forall w, \left|E\left[L_{hinge}(w,x,y)\right] - \tilde{E}\left[L_{hinge}(w,x,y)\right]\right| \leq O\left(\frac{1}{\tau}\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}\right) + \cancel{\eta\left(1+\frac{1}{\tau}\right)} + \eta + \frac{1}{\tau}\sqrt{\frac{\eta \log n}{d}}$$

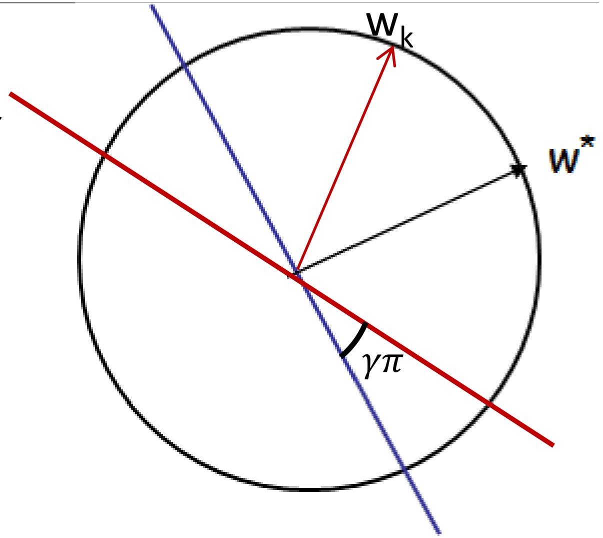
[ABL'14]

- Sample a set S of noisy examples.
- Iteratively do (Outlier removal + hinge loss minimization).

Idea inspired from the active learning literature.

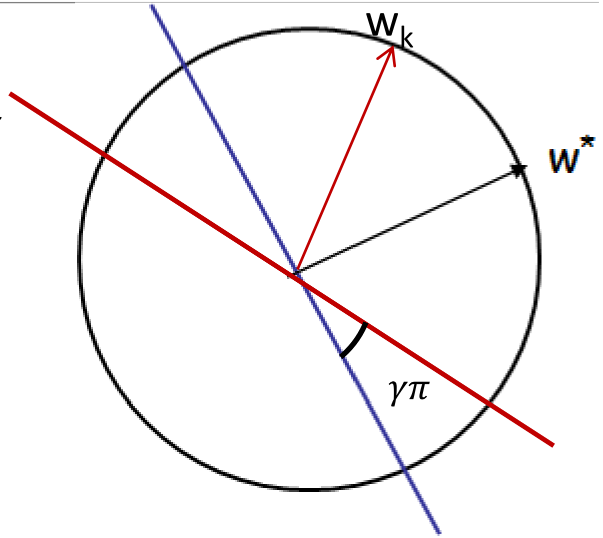
[ABL'14]

- Suppose we have w_k, s.t. $err(w_k) \leq \gamma$
- $\Rightarrow \ \theta(w_k, w^*) \leq \gamma\pi$



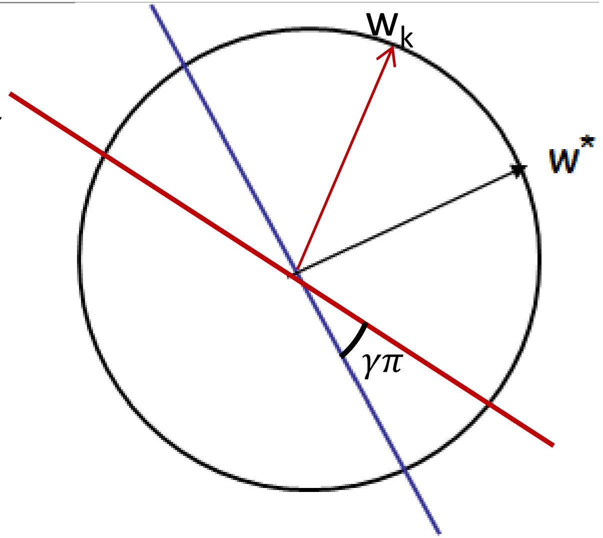
[ABL'14]

- Suppose we have w_k, s.t. $err(w_k) \leq \gamma$
- $\Rightarrow \ \theta(w_k, w^*) \leq \gamma\pi$
- $\mathrm{w}^* \cdot \mathrm{x} = \mathrm{w_k} \cdot \mathrm{x} + \underbrace{(\mathrm{w}^* - \mathrm{w_k})}_{\leq \gamma\pi} \cdot \mathrm{x}$



[ABL'14]

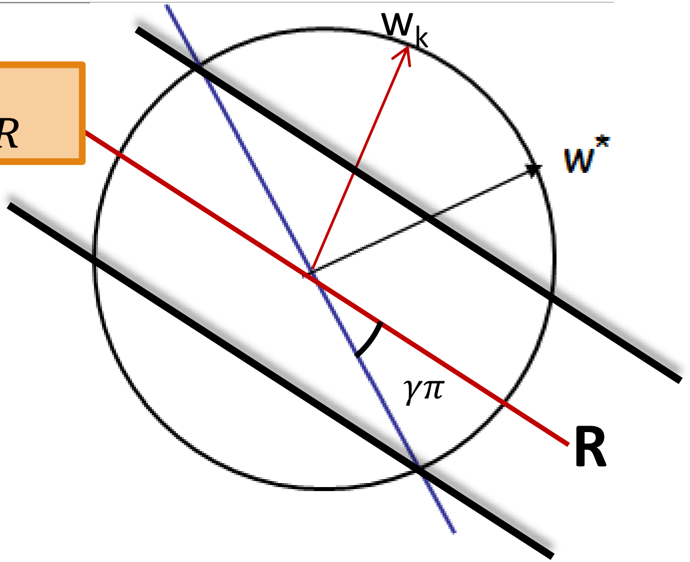
- Suppose we have w_k, s.t. $err(w_k) \leq \gamma$
- $\Rightarrow \ \theta(w_k, w^*) \leq \gamma\pi$
- $$\mathrm{w}^* \cdot \mathrm{x} = \mathrm{w_k} \cdot \mathrm{x} + \underbrace{(\mathrm{w}^* - \mathrm{w_k}) \cdot \mathrm{x}}_{\leq \gamma\pi}$$



- If $|w_k \cdot x| > \pi\gamma$, then w_k and w^* agree on the label of x.

[ABL'14]

Find $w \in B(w_k, \gamma\pi)$ of low error w.r.t. $D_{x \in R}$



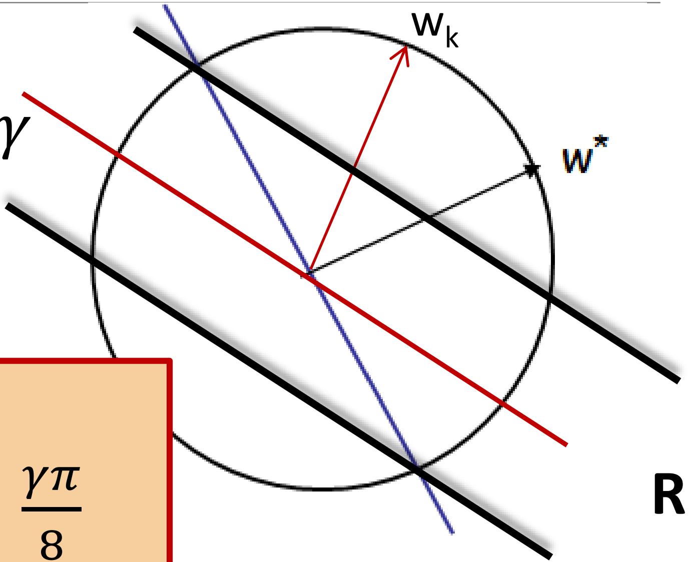
$$R = \{x\colon |w_k \cdot x| \le \pi\gamma\}$$

[ABL'14]

- Suppose we have w_k, s.t. $err(w_k) \leq \gamma$
- $\Rightarrow \ \theta(w_k, w^*) \leq \gamma\pi$

Lemma: If $\theta(w, w^*) = \gamma\pi$,

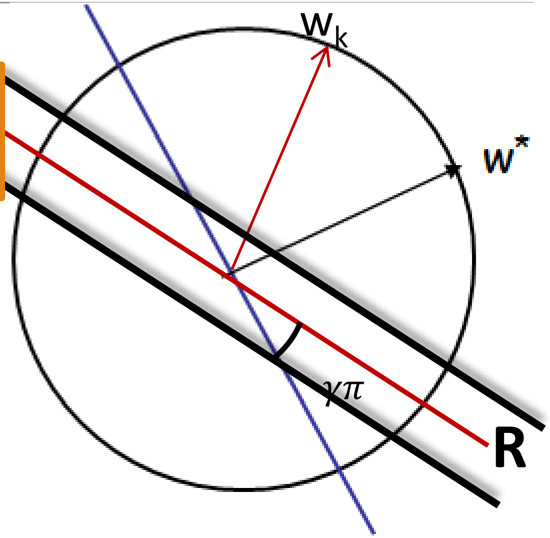
$$\Pr\left[(w^* \cdot x)(w \cdot x) < 0, |w \cdot x| > \frac{2\pi\gamma}{\sqrt{d}}\right] \leq \frac{\gamma\pi}{8}$$



$$R = \{x \colon |w_k \cdot x| \leq \frac{2\pi\gamma}{\sqrt{d}}\}$$

[ABL'14]

Find $w \in B(w_k, \frac{2\gamma\pi}{\sqrt{d}})$ of low error w.r.t. $D_{x \in R}$



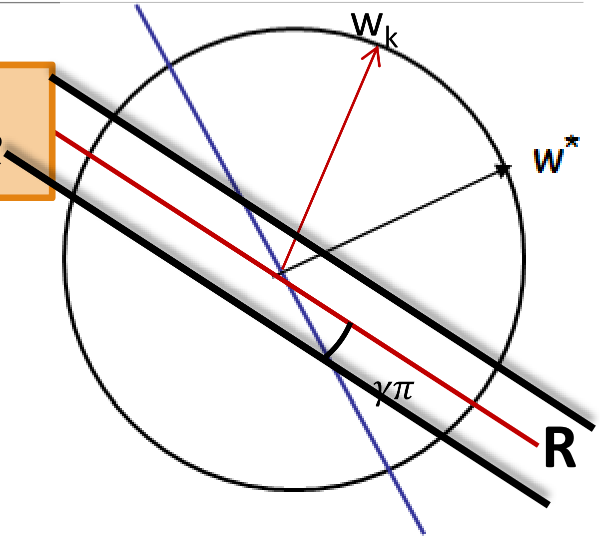
- To improve the error of current w_k from γ to $\frac{\gamma}{2}$, enough to get error $O(1)$ w.r.t. $D_{x \in R}$.
- At each step solve the above subproblem robustly.

$$R = \{x : |w_k \cdot x| \leq \frac{2\pi\gamma}{\sqrt{d}}\}$$

[ABL'14]

Find $w \in B(w_k, \frac{2\gamma\pi}{\sqrt{d}})$ of low error w.r.t. $D_{x \in R}$

- Noise cannot hurt the hinge loss by a lot: $|w \cdot x| \leq 2\pi\gamma$.



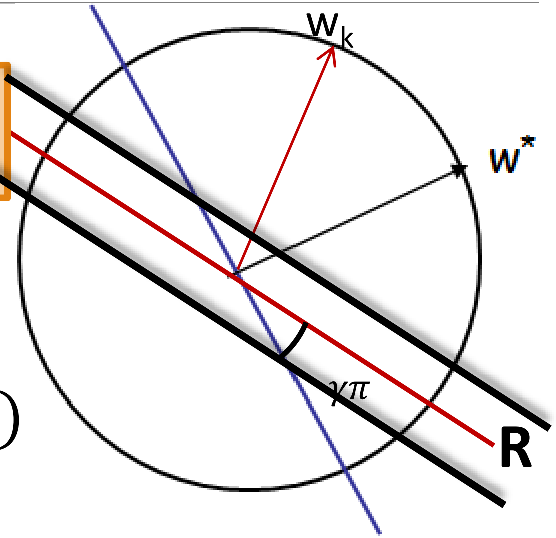
$$R = \{x: |w_k \cdot x| \leq \frac{2\pi\gamma}{\sqrt{d}}\}$$

[ABL'14]

Find $w \in B(w_k, \frac{2\gamma\pi}{\sqrt{d}})$ of low error w.r.t. $D_{x \in R}$

- Can do better outlier removal

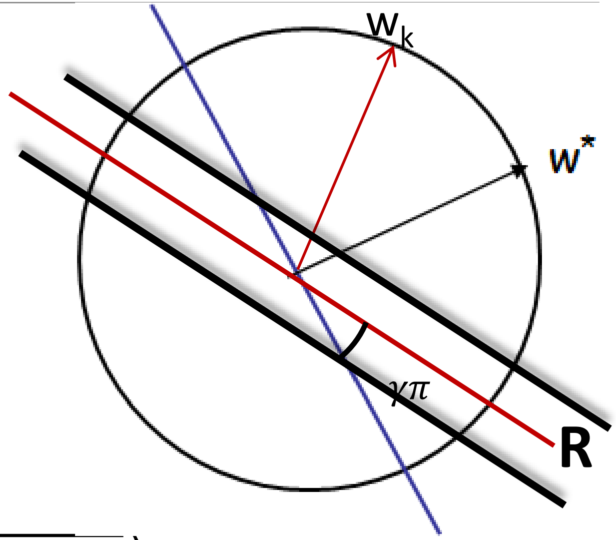
$$\forall w \in B\left(w_k, \frac{2\gamma\pi}{\sqrt{d}}\right), E_{D_{x\in R}}[(w \cdot x)^2] \le O(\frac{\gamma^2}{d})$$



$$R = \{x: |w_k \cdot x| \le \frac{2\pi\gamma}{\sqrt{d}}\}$$

[ABL'14]

$$L_{hinge}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right)$$



With probability $\geq 1 - \delta$,

$$\forall w, \left|E\left[L_{hinge}(w, x, y)\right] - \tilde{E}\left[L_{hinge}(w, x, y)\right]\right| \leq O\left(\frac{1}{\tau}\sqrt{\frac{d + \log\left(\frac{1}{\delta}\right)}{n}}\right) + \cancel{\eta\left(1 + \frac{1}{\tau}\right)} + \eta + \frac{1}{\tau}\sqrt{\frac{\eta\gamma}{d}}$$

[ABL'14]

Initialize w_1 randomly.

Iterate $\mathrm{k} = 2,3, \dots, \log(\frac{1}{\epsilon})$

- Sample m_k examples x satisfying $|w_{k-1} \cdot x| \leq \frac{2\pi\gamma_{k-1}}{\sqrt{d}}$

- Need constant error in each round, hence $O(d)$labeled examples
- Total # labeled examples = $O(dlog(\frac{1}{\epsilon}))$

- Find w_k in $B(w_{k-1}, 2\pi\gamma_{k-1})$ of small hinge loss.
- Clear working set S.

end iterate

PAC learning with malicious noise

- $\eta = O(\frac{\epsilon}{d^{\frac{1}{4}}})$ [KKMS'05]
- $\eta = O(\frac{\epsilon^2}{\log\left(\frac{d}{\epsilon}\right)})$ [KLS'09]
- $\eta = O(\epsilon)$ [ABL'14]
- $\eta = \frac{\epsilon}{2+\delta}$ if only noise in labels [Daniely'15]
  - by combining margin based technique with polynomial regression.
- Extend margin based learning to robustly learn a broader class of distributions [BZ'17].
- Robust learning of non-linear models [DKS'18].

PAC learning with (non)malicious noise

Random Classification Noise

Malicious Noise

Intermediate Noise models?

Each label flipped w.p. $\frac{1}{2} - \beta$

Can learn halfspaces efficiently with $O\left(\frac{d}{\epsilon^2 \beta^2}\right)$ samples[BFKV'98].

Bounded (Massart) Noise

$$p_x = P(y \neq sign(w^* \cdot x)|x) \leq \frac{1}{2} - \beta,$$

$$0 \leq \beta \leq \frac{1}{2}$$

- Can learn with $O(\frac{d}{\epsilon^2 \beta^2})$ samples in exponential time.
  - Current complexity theoretic reductions do not work in this model.
- [ABHZ'16]: Can learn under isotropic log-concave distributions in polynomial time for any constant β.

Bounded (Massart) Noise

$$p_x = P(y \neq sign(w^* \cdot x)|x) \leq \frac{1}{2} - \beta,$$
$$0 \leq \beta \leq \frac{1}{2}$$

- [CLZ’17]: Practical algorithms based on SGD.
- [YZ’17, Zhang’18]: Practical algorithms based on (margin + perceptron). Also label efficient.
- Guarantees hold for uniform distribution. Can handle β arbitrarily close to zero.
- Open: fast algorithms for isotropic log-concave distributions?
- Open: polynomial time algorithms for Massart noise beyond log-concave?

Tsybakov Noise

$$\forall t > 0, \mathbb{P}_X\left(\left|p_x - \frac{1}{2}\right| < t\right) \le Bt^{\alpha}$$

- Can achieve rates $\sim \left(\frac{d \log n + \log\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$
- Open: a polynomial time PAC learning algorithm?

Recap of Open Questions

- PAC learn halfspaces under malicious noise with $\eta = \frac{\epsilon}{d^{0.99}}$?
- Show that SGD based algorithms work for Massart noise beyond the uniform distribution?
- Design polynomial time learning algorithms for Massart noise for a broad class of distributions?
- Design polynomial time algorithm for Tsybakov noise in any non-trivial setting?
- Explore intermediate noise models for PAC learning?

THANK YOU