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 \((x, y = \text{sign}(w^* \cdot x))\) 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 \((x, y = \text{sign}(w^* \cdot x))\) approximately recover \(w^*\).

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

Constraints

\begin{align*}
(x_1, +) & \quad w \cdot x_1 > 0 \\
(x_2, -) & \quad w \cdot x_2 < 0 \\
\ldots & \quad \ldots \\
(x_n, +) & \quad w \cdot x_n > 0
\end{align*}

**Theorem [VC]**

If \( n = \tilde{O}(\frac{d}{\epsilon}) \), then w.h.p. \( \text{err}(w) \leq \epsilon \).
PAC learning with malicious noise

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

- In each draw
  - \((x, \text{sign}(w^* \cdot x))\), where \(x \sim D\) w.p. \(1 - \eta\)
  - Arbitrary \((x, y)\) w.p. \(\eta\)

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

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 $\eta = O\left(\frac{\epsilon}{d}\right)$
- Proof:
  - Need $\sim \frac{d}{\epsilon} \log(1/\epsilon)$ examples to learn in the noise free case.
  - $\mathbb{P}(\text{no example is corrupted}) = (1 - \eta)^{\frac{d}{\epsilon} \log(1/\epsilon)} \geq \text{poly}(\epsilon)$
PAC learning with malicious noise

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

[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
  - $\eta = O\left(\frac{\epsilon}{d^4}\right)$ [KKMS’05]
  - $\eta = O\left(\frac{\epsilon^2}{\log\left(\frac{d}{\epsilon}\right)}\right)$ [KLS’09]
  - $\eta = O(\epsilon)$ [ABL’14]
PAC learning with malicious noise (D=Isotropic log-concave distribution)

- **Efficiently: Can tolerate**
  - \( \eta = O\left(\frac{\epsilon^3}{\log^2\left(\frac{d}{\epsilon}\right)}\right) \) [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 $m = \tilde{O} \left( \frac{d}{\epsilon^2} \right)$ and $w$ has $\frac{\epsilon}{2}$ error on data
then w.h.p. $\text{err}(w) \leq \epsilon$

Minimizing error on noisy data is NP-hard!
Hinge Loss

\[ L_{\text{hinge}}(w, x, y) = \max \left( 0, 1 - \frac{y(w \cdot x)}{\tau} \right) \]
Candidate Algorithm

- Sample a set $S$ of noisy examples.
- Output $w$ of small hinge loss over $S$. 
Uniform Distribution over $S_{d-1}$

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

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

$\hat{E}[L_{hinge}(w, x, y)]$ : 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 $\geq 1 - \delta$,

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

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

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

\[
E[L_{\text{hinge}}(w^*, x, y)] \leq O(\tau \sqrt{d})
\]
Uniform Distribution over $S_{d-1}$

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

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

Set $\tau = \frac{\epsilon}{c\sqrt{d}}$  \quad \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 $\geq 1 - \delta$,

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

[KKMS’05]
Sample a set $S$ of noisy examples.

While there exists a direction $u$ such that $\mathbb{E}_S [(u \cdot x)^2] > \frac{10 \log n}{d}$

- Remove any $x \in S$ such that $(u \cdot x)^2 > \frac{10 \log n}{d}$

- Output $w$ of small hinge loss over $S$. 
\[ L_{\text{hinge}}(w, x, y) = \max \left( 0, 1 - \frac{y(w \cdot x)}{\tau} \right) \]

With probability \( \geq 1 - \delta \),

\[ \forall w, |E[L_{\text{hinge}}(w, x, y)] - \tilde{E}[L_{\text{hinge}}(w, x, y)]| \leq O \left( \frac{1}{\tau} \sqrt{\frac{d + \log \left( \frac{1}{\delta} \right)}{n}} \right) + \eta \left( 1 + \frac{1}{\tau} \right) + \eta + \frac{1}{\tau} \sqrt{\frac{\eta \log n}{d}} \]
Sample a set $S$ of noisy examples.
Iteratively do (Outlier removal + hinge loss minimization).

Idea inspired from the active learning literature.
Suppose we have $w_k$, s.t. $err(w_k) \leq \gamma$
$\Rightarrow \theta(w_k, w^*) \leq \gamma \pi$
Suppose we have $w_k$, s.t. $\text{err}(w_k) \leq \gamma$

$\Rightarrow \theta(w_k, w^*) \leq \gamma \pi$

$w^* \cdot x = w_k \cdot x + (w^* - w_k) \cdot x$

$\leq \gamma \pi$
Suppose we have $w_k$, s.t. $err(w_k) \leq \gamma$

$\Rightarrow \theta(w_k, w^*) \leq \gamma \pi$

$w^* \cdot x = w_k \cdot x + (w^* - w_k) \cdot x \leq \gamma \pi$

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

[ABL’14]
[ABL’14]

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

$$R = \{ x : |w_k \cdot x| \leq \pi \gamma \}$$
[ABL’14]

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

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

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

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

\[
R = \{ x : |w_k \cdot x| \leq \frac{2\pi \gamma}{\sqrt{d}} \}
\]
Find $w \in B(w_k, \frac{2\gamma \pi}{\sqrt{d}})$ of low error w.r.t. $D_{x \in \mathbb{R}}$

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

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

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

\[
R = \{ x : |w_k \cdot x| \leq \frac{2\pi \gamma}{\sqrt{d}} \}
\]
\[ L_{hinge}(w, x, y) = \max\left(0, 1 - \frac{y(w \cdot x)}{\tau}\right) \]

With probability \( \geq 1 - \delta \),

\[ \forall w, |E[L_{hinge}(w, x, y)] - \tilde{E}[L_{hinge}(w, x, y)]| \leq O\left(\frac{1}{\tau} \sqrt{\frac{d + \log\left(\frac{1}{\delta}\right)}{n}}\right) + \eta(1 + \frac{1}{\tau}) + \eta + \frac{1}{\tau} \sqrt{\frac{\eta \gamma}{d}} \]
[ABL’14]

Initialize $w_1$ randomly.

Iterate $k = 2, 3, \ldots, \log\left(\frac{1}{\epsilon}\right)$

- 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 = $O(d \log\left(\frac{1}{\epsilon}\right))$

- 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

- $\eta = O\left(\frac{\varepsilon}{d^4}\right)$ [KKMS’05]
- $\eta = O\left(\frac{\varepsilon^2}{\log\left(\frac{d}{\varepsilon}\right)}\right)$ [KLS’09]
- $\eta = O(\varepsilon)$ [ABL’14]
- $\eta = \frac{\varepsilon}{2+\delta}$ 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
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].

Intermediate Noise models?

Malicious Noise
Bounded (Massart) Noise

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

- Can learn with \( O\left(\frac{d}{\epsilon^2 \beta^2}\right) \) 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 \( \beta \).
Bounded (Massart) Noise

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

- [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 \( \beta \) arbitrarily close to zero.
- Open: fast algorithms for isotropic log-concave distributions?
- Open: polynomial time algorithms for Massart noise beyond log-concave?
Tsybakov Noise

\[ \forall t > 0, \mathbb{P}_X \left( \left| p_x - \frac{1}{2} \right| < t \right) \leq B t^\alpha \]

- Can achieve rates \( \sim \left( \frac{d \log n + \log \left( \frac{1}{\delta} \right)}{n} \right)^{\frac{1}{2+\alpha}} \)

- 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