PAC Learning Linear Thresholds from Label Proportions

Anand Brahmbhatt, Rishi Saket and Aravindan Raghuveer
Google Research India

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^{d} \rightarrow$ **{0, 1}**, given **m** samples **(x, f(x))** where **x ~ $\mathcal{D}$**, find a hypothesis h s.t. **$\Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$** w.p. **$1 - \delta$**. Efficient if **$m \leq O(\mathrm{poly}(d, 1/\varepsilon, \log(1/\delta)))$**.

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^d \rightarrow$ **{0, 1}**, given **m** samples **(x, f(x))** where **x ~ $\mathcal{D}$**, find a hypothesis h s.t. $\mathbf{Pr}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$ w.p. **1 - δ**. Efficient if **m ≤ O(poly(d,1/ε, log(1/δ)))**.

- **Learning from Label Proportions (LLP)**: Samples are "bags" of the form $(\{x_1, \ldots, x_q\}, k)$ where $\Sigma_i f(x_i) = k$. Our study: $f \leftarrow$ Linear Threshold function (LTF) $[\ \mathbb{1}\{r^T x + c > 0\}\ ]$

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^d \rightarrow$ **{0, 1}**, given **m** samples **(x, f(x))** where **x ~ $\mathcal{D}$**, find a hypothesis h s.t. $\mathbf{Pr}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$ w.p. **1 - δ**. Efficient if **m ≤ O(poly(d,1/ε, log(1/δ)))**.

- **Learning from Label Proportions (LLP)**: Samples are "bags" of the form $(\{x_1, \ldots, x_q\}, k)$ where $\Sigma_i f(x_i) = k$. Our study: $f \leftarrow$ Linear Threshold function (LTF) $[\ \mathbb{1}\{r^T x + c > 0\}\ ]$
  - LTFs are efficiently PAC learnable to arbitrary accuracy

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^d \rightarrow \{0, 1\}$, given **m** samples **(x, f(x))** where **x ~ $\mathcal{D}$**, find a hypothesis h s.t. $\mathbf{Pr}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$ w.p. **1 - δ**. Efficient if **m ≤ O(poly(d,1/ε, log(1/δ)))**.

- **Learning from Label Proportions (LLP)**: Samples are "bags" of the form $(\{x_1, \ldots, x_q\}, k)$ where $\Sigma_i f(x_i) = k$. Our study: f ← Linear Threshold function (LTF) $[\ \mathbb{1}\{r^T x + c > 0\}\ ]$
  - LTFs are efficiently PAC learnable to arbitrary accuracy

- **LLP Hardness [Saket 21, 22]**: Given a set of bags of size ≤ **q**, s.t. ∃ LTF consistent with all bags, NP-hard to find **<u>any</u>** LTF consistent with ≥ **(1/q + o(1))**-fraction of the bags.

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^d \rightarrow \{0, 1\}$, given **m** samples **(x, f(x))** where **x ~ $\mathcal{D}$**, find a hypothesis h s.t. **$\Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$** w.p. **$1 - \delta$**. Efficient if **$m \leq O(\text{poly}(d, 1/\varepsilon, \log(1/\delta)))$**.

- **Learning from Label Proportions (LLP)**: Samples are "bags" of the form **$(\{x_1, \ldots, x_q\}, k)$** where **$\Sigma_i f(x_i) = k$**. Our study: $f \leftarrow$ Linear Threshold function (LTF) **$[\ \mathbb{1}\{r^T x + c > 0\}\ ]$**
  - ➢ LTFs are efficiently PAC learnable to arbitrary accuracy

- **LLP Hardness [Saket 21, 22]**: Given a set of bags of size ≤ **q**, s.t. ∃ LTF consistent with all bags, NP-hard to find **<u>any</u>** LTF consistent with ≥ **(1/q + o(1))**-fraction of the bags.

Question: *What happens for natural/well-behaved distributions?*

Problem Statement

- **PAC Learning**: For a function f: $\mathbb{R}^d \rightarrow \{0, 1\}$, given **m** samples $(x, f(x))$ where $x \sim \mathcal{D}$, find a hypothesis h s.t. $\Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$ w.p. $1 - \delta$. Efficient if $m \leq O(\mathrm{poly}(d, 1/\varepsilon, \log(1/\delta)))$.

- **Learning from Label Proportions (LLP)**: Samples are "bags" of the form $(\{x_1, \ldots, x_q\}, k)$ where $\Sigma_i f(x_i) = k$. Our study: $f \leftarrow$ Linear Threshold function (LTF) $[\ \mathbb{1}\{r^T x + c > 0\}\ ]$
  - LTFs are efficiently PAC learnable to arbitrary accuracy

- **LLP Hardness [Saket 21, 22]**: Given a set of bags of size $\leq$ **q**, s.t. $\exists$ LTF consistent with all bags, NP-hard to find **any** LTF consistent with $\geq (1/q + o(1))$-fraction of the bags.

Question: *What happens for natural/well-behaved distributions?*

- Bag Oracle for LTF f, $\mathcal{D} = N(\mu, \Sigma)$ and fixed **q**, **k** :- $\mathrm{Ex}(f, \mathcal{D}, q, k)$
  - Samples bag with **k** feature-vecs. from $\mathcal{D}|f(x)=1$ and **q-k** from $\mathcal{D}|f(x)=0$.

Our Results

Homogeneous LTF Standard Gaussian	**Homogeneous LTF Centered Gaussian**	**Non-Homogeneous LTF General Gaussian**
$k \neq q/2$	$\forall\, k \in \{1, \dots, q-1\}$	$\forall\, k \in \{1, \dots, q-1\}$
$\mathcal{D} := N(\mathbf{0}, \mathbf{I})$	$\mathcal{D} := N(\mathbf{0}, \mathbf{\Sigma})$	$\mathcal{D} := N(\mu, \mathbf{\Sigma})$
$f(\mathbf{x}) := \mathbf{1}\{\mathbf{r}_*^\top \mathbf{x} > 0\}$	$f(\mathbf{x}) := \mathbf{1}\{\mathbf{r}_*^\top \mathbf{x} > 0\}$	$f(\mathbf{x}) := \mathbf{1}\{\mathbf{r}_*^\top \mathbf{x} + c_* > 0\}$
$\hat{f}(\mathbf{x}) := \mathbf{1}\{\hat{\mathbf{r}}^\top \mathbf{x} > 0\}$	$\hat{f}(\mathbf{x}) := \mathbf{1}\{\hat{\mathbf{r}}^\top \mathbf{x} > 0\}$	$\hat{f}(\mathbf{x}) := \mathbf{1}\{\hat{\mathbf{r}}^\top \mathbf{x} + \hat{c} > 0\}$
$O\left(\frac{d}{\varepsilon^2} \log\left(\frac{d}{\delta}\right)\right)$	$O\left(\frac{d}{\varepsilon^4} \log\left(\frac{d}{\delta}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^6 q^8\right)$	$O\left(\frac{d}{\varepsilon^4} \log\left(\frac{d}{\delta}\right) \frac{O(\ell^2)}{\Phi(\ell)(1-\Phi(\ell))} \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^4 \left(\frac{\sqrt{\lambda_{\max}} + \lVert\mu\rVert_2}{\sqrt{\lambda_{\min}}}\right)^4 q^8\right)$

$d \leftarrow$ dimension of the feature-vectors, $\quad \ell := -(c_* + \mathbf{r}_*^\top \mu)/\lVert \mathbf{\Sigma}^{1/2} \mathbf{r}_* \rVert_2$

$\lambda_{\max} \leftarrow$ maximum eigenvalue of $\mathbf{\Sigma}$, $\quad \lambda_{\min} \leftarrow$ minimum eigenvalue of $\mathbf{\Sigma}$

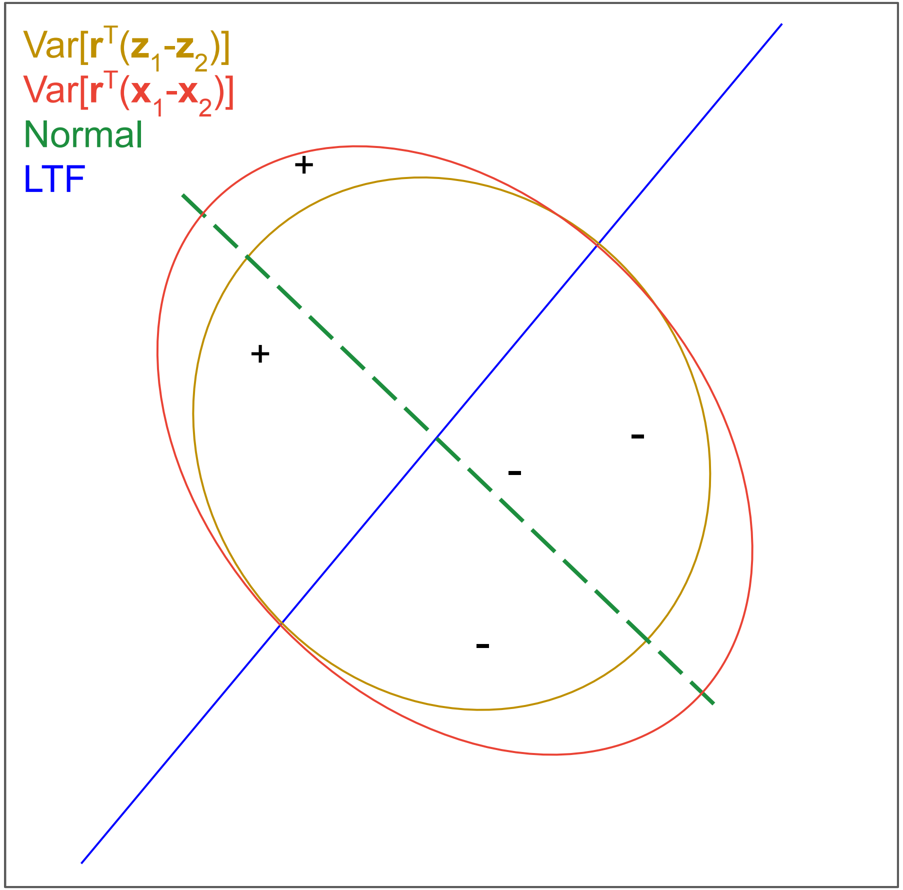
Normal Estimation

- **Observation**: Sampling a pair of feature vectors from a bag
  - ($\mathbf{x}_1$, $\mathbf{x}_2$) independently u.a.r: $\Pr[f(\mathbf{x}_1) \neq f(\mathbf{x}_1)] = 2k(q-k)/q^2$
  - ($\mathbf{z}_1$, $\mathbf{z}_2$) pair u.a.r w/o replacement: $\Pr[f(\mathbf{z}_1) \neq f(\mathbf{z}_2)] = 2k(q-k)/q(q-1)$

Theorem: $\rho(\mathbf{r}) := \mathrm{Var}\left[\mathbf{r}^\top(\mathbf{z}_1 - \mathbf{z}_2)\right] / \mathrm{Var}\left[\mathbf{r}^\top(\mathbf{x}_1 - \mathbf{x}_2)\right]$ is maximized when $\mathbf{r} = \pm\mathbf{r}_*$.

Proof Sketch: Case $\mathcal{D} = N(\mathbf{0}, \mathbf{I})$ and $k = q/2$. For any $\mathbf{r} \in \mathbb{S}^{d-1}$

$$\rho(\mathbf{r}) = \begin{cases} 1 & \text{if } \mathbf{r}^\top\mathbf{r}_* = 0, \text{ i.e. } \mathbf{r} \text{ lies on the LTF,} \\ 1 + \frac{1}{q-1}\left(\frac{2}{\pi}\right) & \text{if } \mathbf{r} = \pm\mathbf{r}_*, \text{ i.e. } \mathbf{r} \text{ is aligned} \\ & \text{with the normal to the LTF,} \\ 1 + \frac{1}{q-1}\left(\frac{2}{\pi}\right)\cos^2\theta & \text{if } \mathbf{r}^\top\mathbf{r}_* = \cos\theta, \text{ i.e. the angle} \\ & \text{between } \mathbf{r} \text{ and } \mathbf{r}_* \text{ is } \theta. \end{cases}$$



Homogeneous LTF with $\mathbf{N(0, \Sigma)}$

- Let $\boldsymbol{\Sigma}_B = \mathbf{E}[(\mathbf{x_1}-\mathbf{x_2})(\mathbf{x_1}-\mathbf{x_2})^T]$ and $\boldsymbol{\Sigma}_D = \mathbf{E}[(\mathbf{z_1}-\mathbf{z_2})(\mathbf{z_1}-\mathbf{z_2})^T]$
- **Objective**: $\text{argmax}_{||\mathbf{r}||=1} \mathbf{r}^T\boldsymbol{\Sigma}_B\mathbf{r}/\mathbf{r}^T\boldsymbol{\Sigma}_D\mathbf{r} = \boldsymbol{\Sigma}_B^{-1/2}\mathbf{PrincipalEigenVector}(\boldsymbol{\Sigma}_B^{-1/2}\boldsymbol{\Sigma}_D\boldsymbol{\Sigma}_B^{-1/2})$

Homogeneous LTF with $\mathbf{N(0, \Sigma)}$

- Let $\boldsymbol{\Sigma}_B = \mathbf{E}[(\mathbf{x_1}-\mathbf{x_2})(\mathbf{x_1}-\mathbf{x_2})^T]$ and $\boldsymbol{\Sigma}_D = \mathbf{E}[(\mathbf{z_1}-\mathbf{z_2})(\mathbf{z_1}-\mathbf{z_2})^T]$
- **Objective**: $\text{argmax}_{||\mathbf{r}||=1} \mathbf{r}^T\boldsymbol{\Sigma}_B\mathbf{r}/\mathbf{r}^T\boldsymbol{\Sigma}_D\mathbf{r} \;=\; \boldsymbol{\Sigma}_B^{-\frac{1}{2}}\mathbf{PrincipalEigenVector}(\boldsymbol{\Sigma}_B^{-\frac{1}{2}}\boldsymbol{\Sigma}_D\boldsymbol{\Sigma}_B^{-\frac{1}{2}})$
- **Stability Theorem**: The ratio maximization computed with high probability approximations of $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_D$ is close to the normal with high probability.

Homogeneous LTF with $\mathbf{N(0, \Sigma)}$

- Let $\mathbf{\Sigma}_B = \mathbf{E}[(\mathbf{x_1}-\mathbf{x_2})(\mathbf{x_1}-\mathbf{x_2})^T]$ and $\mathbf{\Sigma}_D = \mathbf{E}[(\mathbf{z_1}-\mathbf{z_2})(\mathbf{z_1}-\mathbf{z_2})^T]$
- **Objective**: $\text{argmax}_{||\mathbf{r}||=1} \mathbf{r}^T\mathbf{\Sigma}_B\mathbf{r}/\mathbf{r}^T\mathbf{\Sigma}_D\mathbf{r} \quad = \quad \mathbf{\Sigma}_B^{-1/2}\mathbf{PrincipalEigenVector}(\mathbf{\Sigma}_B^{-1/2}\mathbf{\Sigma}_D\mathbf{\Sigma}_B^{-1/2})$
- **Stability Theorem**: The ratio maximization computed with high probability approximations of $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_D$ is close to the normal with high probability.
- Geometric bound ⟶ Bound on sample error
  - An algorithm to find a high probability estimator of c_* given a high probability estimator of $\mathbf{r}_*$

Homogeneous LTF with $\mathbf{N(0, \Sigma)}$

- Let $\boldsymbol{\Sigma}_B = \mathbf{E}[(\mathbf{x}_1-\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)^T]$ and $\boldsymbol{\Sigma}_D = \mathbf{E}[(\mathbf{z}_1-\mathbf{z}_2)(\mathbf{z}_1-\mathbf{z}_2)^T]$
- **Objective**: $\text{argmax}_{||\mathbf{r}||=1}\mathbf{r}^T\boldsymbol{\Sigma}_B\mathbf{r}/\mathbf{r}^T\boldsymbol{\Sigma}_D\mathbf{r} = \boldsymbol{\Sigma}_B^{-½}\mathbf{PrincipalEigenVector}(\boldsymbol{\Sigma}_B^{-½}\boldsymbol{\Sigma}_D\boldsymbol{\Sigma}_B^{-½})$
- **Stability Theorem**: The ratio maximization computed with high probability approximations of $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_D$ is close to the normal with high probability.
- Geometric bound → Bound on sample error
  - An algorithm to find a high probability estimator of c_* given a high probability estimator of $\mathbf{r}_*$
- Bound on sample error → Bound on Generalization error
  - **Generalization Error Bound**

Homogeneous LTF with **N(0, Σ)**

- Let $\mathbf{\Sigma}_B = \mathbf{E}[(\mathbf{x_1}-\mathbf{x_2})(\mathbf{x_1}-\mathbf{x_2})^T]$ and $\mathbf{\Sigma}_D = \mathbf{E}[(\mathbf{z_1}-\mathbf{z_2})(\mathbf{z_1}-\mathbf{z_2})^T]$
- **Objective**: $\text{argmax}_{||\mathbf{r}||=1}\mathbf{r}^T\mathbf{\Sigma}_B\mathbf{r}/\mathbf{r}^T\mathbf{\Sigma}_D\mathbf{r} \;=\; \mathbf{\Sigma}_B^{-½}\mathbf{PrincipalEigenVector}(\mathbf{\Sigma}_B^{-½}\mathbf{\Sigma}_D\mathbf{\Sigma}_B^{-½})$
- **Stability Theorem**: The ratio maximization computed with high probability approximations of $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_D$ is close to the normal with high probability.
- Geometric bound → Bound on sample error
  - An algorithm to find a high probability estimator of c_* given a high probability estimator of $\mathbf{r}_*$
- Bound on sample error → Bound on Generalization error
  - **Generalization Error Bound**
- Sub-gaussian concentration bounds for thresholded Gaussians.

Thank You