Improved Sample Complexity for Multiclass PAC Learning

Steve Hanneke [1] Shay Moran [2] Qian Zhang [3]

[1]Purdue University, steve.hanneke@gmail.com

[2]Technion and Google Research, smoran@technion.ac.il

[3]Purdue University, zhan3761@purdue.edu

November 2024

- $\mathcal{X}$: feature space.
- $\mathcal{Y}$: label space, a set with $|\mathcal{Y}| > 2$ ($|\mathcal{Y}|$ can be infinite).
- $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$: concept class.
- **Error rate** of a classifier $h : \mathcal{X} \to \mathcal{Y}$ under a probability distribution P over $\mathcal{X} \times \mathcal{Y}$:

$$\mathrm{er}_P(h) := P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : h(x) \neq y\}).$$

- A distribution P is called ($\mathcal{H}$-)**realizable** if

$$\inf_{h \in \mathcal{H}} \mathrm{er}_P(h) = 0.$$

- $\mathrm{RE}(\mathcal{H})$: the set of all $\mathcal{H}$-realizable distributions.

- A **multiclass learner** (or a learner) $\mathcal{A}$ is an algorithm which given a sequence $\mathbf{s} \in \cup_{n=0}^{\infty}(\mathcal{X} \times \mathcal{Y})^n$ and a concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, outputs a classifier $\mathcal{A}(\mathbf{s}, \mathcal{H}) \in \mathcal{Y}^{\mathcal{X}}$.
- The **(PAC) sample complexity** of $\mathcal{A}$ is the function

$$M_{\mathcal{A},\mathcal{H}} : (0,1)^2 \to \mathbb{N},$$
$$(\varepsilon, \delta) \mapsto \inf\{n \in \mathbb{N} : \mathbb{P}_{S\sim P^m}(\mathrm{er}_P(\mathcal{A}(S,\mathcal{H})) > \varepsilon) \leq \delta,\ \forall m \geq n, P \in \mathrm{RE}(\mathcal{H})\}$$

with the convention $\inf \emptyset = \infty$.

- $\mathcal{H}$ is **PAC learnable** by $\mathcal{A}$ if $\mathcal{M}_{\mathcal{A},\mathcal{H}}(\varepsilon, \delta) < \infty$ for all $(\varepsilon, \delta) \in (0,1)^2$. The **(PAC) sample complexity** of $\mathcal{H}$ is defined as $\mathcal{M}_{\mathcal{H}}(\varepsilon, \delta) := \inf_{\mathcal{A}} \mathcal{M}_{\mathcal{A},\mathcal{H}}(\varepsilon, \delta)$, $\forall(\varepsilon, \delta) \in (0,1)^2$.

- **Expected error rate**

$$\varepsilon_{\mathcal{A},\mathcal{H},P} : \mathbb{N} \to [0,1], \quad n \mapsto \mathbb{E}_{S\sim P^n}[\mathrm{er}_P(\mathcal{A}(S,\mathcal{H}))].$$

  Define $\varepsilon_{\mathcal{A},\mathcal{H}} := \sup_{P\in\mathrm{RE}(\mathcal{H})} \varepsilon_{\mathcal{A},\mathcal{H},P}$ and $\varepsilon_{\mathcal{H}} := \inf_{\mathcal{A}} \varepsilon_{\mathcal{A},\mathcal{H}}$.
- **Transductive error rate**

$$\varepsilon_{\mathcal{A},\mathcal{H},\text{trans}} : \mathbb{N} \to [0,1],$$
$$n \mapsto \sup_{\mathsf{s}=((x_1,h(x_1)),\ldots,(x_n,h(x_n)))\in(\mathcal{X}\times\mathcal{Y})^n : h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{h(x_i)\neq\mathcal{A}(\mathsf{s}_{-i},\mathcal{H})(x_i)}.$$

  Define $\varepsilon_{\mathcal{H},\text{trans}} := \inf_{\mathcal{A}} \varepsilon_{\mathcal{A},\mathcal{H},\text{trans}}$.
- By a leave-one-out argument, we have $\varepsilon_{\mathcal{A},\mathcal{H}} \leq \varepsilon_{\mathcal{A},\mathcal{H},\text{trans}}$.
- **Theorem 2.6.** Suppose $\varepsilon_{\mathcal{A},\mathcal{H},P}(n) \leq M_n/n\ \forall n \in \mathbb{N}$ and $P \in \mathrm{RE}(\mathcal{H})$ with M_n nondecreasing in n. Then, there exists a learner $\mathcal{A}'$ such that for any $P \in \mathrm{RE}(\mathcal{H})$, $\delta \in (0,1)$, and $n \geq 4$, sampling $S \sim P^n$, with probability at least $1-\delta$,

$$\mathrm{er}_P(\mathcal{A}'(S,\mathcal{H})) \leq 4.82 \cdot (8.34 M_{\lfloor n/2 \rfloor} + \log(2/\delta))/n.$$

- For $d, k \in \mathbb{N}$, a set $H \subseteq \mathcal{Y}^d$ is called a k-**pseudo-cube** of dimension d if
  - $0 < |H| < \infty$ and
  - For any $h \in H$ and $i \in [d]$, there are at least k i-neighbors of h (g is an i-neighbor of h if $g(i) \neq h(i)$ and $g(j) = h(j)$ for all $j \in [d] \backslash \{i\}$).
- $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}^d$ is k-**DS-shattered** by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if $\mathcal{H}|_{\mathbf{x}} := \{(h(x_1), \ldots, h(x_d)) : h \in \mathcal{H}\}$ contains a d-dimensional k-pseudo-cube.
- The k-**DS dimension** of $\mathcal{H}$ ($\dim_k(\mathcal{H})$) is the maximum size of a k-DS-shattered sequence.
- Pseudo-cube and DS dimension (dim) correspond to 1-pseudo and 1-DS dimension ($\dim_1$).

- Brukhim et al. [2022] proved that
  - a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is PAC learnable if and only if $d := \mathrm{dim}(\mathcal{H}) < \infty$;
  - there exists a multiclass learner $\mathcal{A}$ which for any $P \in \mathrm{RE}(\mathcal{H})$, $\delta \in (0,1)$, $n \in \mathbb{N}$, and $S \sim P^n$, satisfies that with probability at least $1 - \delta$,

$$\mathrm{er}_P(\mathcal{A}(S,\mathcal{H})) = O\left(\frac{(d^{3/2}\log(d) + d\log(\log(n)))\log^2(n) + \log(1/\delta)}{n}\right). \tag{1}$$

- Charikar and Pabbaraju [2023] proved $\varepsilon_{\mathcal{H}}(n) = \Omega\left(d/n\right)$.

- The **one-inclusion graph** (OIG) of $H \subseteq \mathcal{Y}^n$ is a hypergraph $\mathcal{G}(H) = (H, E)$ where H is the vertex-set and E is the edge-set defined as follows.
- For any $i \in [n]$ and $f : [n] \backslash \{i\} \to \mathcal{Y}$, define the set $e_{i,f} := \{h \in H : h(j) = f(j),\ \forall j \in [n] \backslash \{i\}\}$.
- The edge-set is

$$E := \{(e_{i,f}, i) : i \in [n], f : [n] \backslash \{i\} \to \mathcal{Y}, e_{i,f} \neq \emptyset\}.$$

- For any $(e_{i,f}, i) \in E$ and $h \in H$, we say $h \in (e_{i,f}, i)$ if $h \in e_{i,f}$. The size of the edge is $|(e_{i,f}, i)| := |e_{i,f}|$.

- For any hypergraph $G = (V, E)$ and $v \in V$, the **degree** of v in G is $\mathrm{deg}(v; G) := |\{e \in E : v \in e, |e| \geq 2\}|$, written $\mathrm{deg}(v)$ in abbreviation.
- If $|V| < \infty$, the **average degree** and **average out-degree** of G are

$$\mathrm{avgdeg}(G) := \tfrac{1}{|V|} \textstyle\sum_{v \in V} \mathrm{deg}(v; G) = \tfrac{1}{|V|} \textstyle\sum_{e \in E : |e| \geq 2} |e|,$$
$$\mathrm{avgoutdeg}(G) := \tfrac{1}{|V|} \textstyle\sum_{e \in E} (|e| - 1).$$

- For general V, the **maximal average degree** of G is

$$\mathrm{md}(G) := \sup_{U \subseteq V : |U| < \infty} \mathrm{avgdeg}(G[U]),$$

where $G[U] = (U, E[U])$ with $E[U] := \{e \cap U : e \in E, e \cap U \neq \emptyset\}$.
- The **density** of $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is defined as

$$\mu_{\mathcal{H}}(m) := \sup_{\mathbf{x} \in \mathcal{X}^m} \mathrm{md}(\mathcal{G}(\mathcal{H}|_{\mathbf{x}})), \ \forall m \in \mathbb{N}.$$

- $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is **nondegenerate** if there exist $h_1, h_2 \in \mathcal{H}$ and $x_0, x_1 \in \mathcal{X}$ such that $h_1(x_0) = h_2(x_0)$ and $h_1(x_1) \neq h_2(x_1)$.
- $\mathcal{H}$ is **degenerate** if it is not nondegenerate.
- **Theorem 1.9** (Partial summary of Theorem 2.5 and 2.11)**.** For any nondegenerate concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\dim(\mathcal{H}) = d$ and any $(\varepsilon, \delta) \in (0,1)^2$, we have

$$\Omega\left(\frac{d+\log(1/\delta)}{\varepsilon}\right) \leq \mathcal{M}_{\mathcal{H}}(\varepsilon, \delta) \leq O\left(\frac{d^{3/2}\log(d)\log(d/\varepsilon)+\log(1/\delta)}{\varepsilon}\right). \tag{2}$$

Theorem 1.10 (Informal summary of Theorem 2.7 and 2.10)**.** Assume that there exists a list learner which, given a concept class $\mathcal{H}$ with $\dim(\mathcal{H}) = d$ and training sequence of size n, outputs a menu of size $p(\mathcal{H}, n)$ with expected error rate upper bounded by $\beta(\mathcal{H}, n)/n$ for some functions p and β nondecreasing in n. Then, there exists a multiclass learner whose error rate is

$$O\left(\frac{\beta(\mathcal{H},n)+d\log(p(\mathcal{H},n))+\log(1/\delta)}{n}\right) \text{ with probability at least } 1-\delta.$$

Moreover, there exists a list learner satisfying

$$p(\mathcal{H}, n) = O\big((e\sqrt{d})^{\sqrt{d}} \log(n)\big) \text{ and}$$
$$\beta(\mathcal{H}, n) = O\big(d^{3/2} \log(d) \log(n)\big).$$

- A **menu** of size k is a function $\mu : \mathcal{X} \to \{\mathsf{Y} \subseteq \mathcal{Y} : |\mathsf{Y}| \le k\}$. A 1-menu can be viewed as a classifier in $\mathcal{Y}^{\mathcal{X}}$, and vice versa.
- A **list learner** $\mathcal{A}$ of size k is an algorithm which, given a sequence $\mathbf{s} \in \cup_{n=0}^{\infty}(\mathcal{X} \times \mathcal{Y})^n$ and a concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, outputs a k-menu $\mathcal{A}(\mathbf{s}, \mathcal{H})$. A 1-list learner can be viewed as a multiclass learner, and vice versa.
- Charikar and Pabbaraju [2023] proved that $\mathcal{H}$ is k-list learnabale if and only if $d_k := \dim_k(\mathcal{H}) < \infty$, and there exists a k-list learner $\mathcal{A}^k$ which for any $P \in \mathrm{RE}(\mathcal{H})$, $\delta \in (0, 1)$, $n \in \mathbb{N}$, and $S \sim P^n$, satisfies that with probability at least $1 - \delta$,

$$\mathrm{er}_P(\mathcal{A}^k(S, \mathcal{H})) = O\left(\frac{k^6 d_k(\sqrt{d_k}\log(d_k) + \log(k\log(n)))\log^2(n) + \log(1/\delta)}{n}\right). \quad (3)$$

- Charikar and Pabbaraju [2023] proved the lower bound $\varepsilon_{\mathcal{H}}^k(n) = \Omega\left(d_k/(kn)\right)$.

- A concept class $\mathcal{H} \in \mathcal{Y}^{\mathcal{X}}$ is called k-**nondegenerate** for $k \in \mathbb{N}$ if there exist $h_1, \dots, h_{k+1} \in \mathcal{H}$ and $x_0, x_1 \in \mathcal{X}$ such that $|\{h_j(x_0) : j \in [k+1]\}| = 1$ and $|\{h_j(x_1) : j \in [k+1]\}| = k+1$.
- $\mathcal{H}$ is called k-**degenerate** if it is not k-nondegenerate.
- **Theorem 2.5.** For any $k \in \mathbb{N}$, k-nondegenerate concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\dim_k(\mathcal{H}) = d_k \in \mathbb{N}$, $\varepsilon \in \left(0, \frac{1}{8(k+1)}\right)$, and $\delta \in \left(0, \frac{1}{4(k+1)}\right)$, we have

$$\mathcal{M}_{\mathcal{H}}^{k}(\varepsilon, \delta) \geq \frac{(d_k - 1)\log(2) + 4\log(1/\delta)}{16(k+1)\varepsilon}.$$

In particular, when $k = 1$, for any $\varepsilon \in (0, 1/16)$ and $\delta \in (0, 1/8)$, we have

$$\mathcal{M}_{\mathcal{H}}(\varepsilon, \delta) \geq \frac{(\dim(\mathcal{H}) - 1)\log(2) + 4\log(1/\delta)}{32\varepsilon}. \tag{4}$$

Reduction from multiclass learning to list learning

Algorithm 1: Multiclass learner $\mathcal{A}_{\mathrm{red}}$ using a list learner $\mathcal{A}_{\mathrm{list}}$

Input: List learner $\mathcal{A}_{\mathrm{list}}$, concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, training sequence $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ for $n \geq 3$, test feature $x_{n+1} \in \mathcal{X}$.

Output: A label $y \in \mathcal{Y}$ for the feature x_{n+1}.

$n_1 \leftarrow n - 2\lfloor n/3 \rfloor, \;\; n_2 \leftarrow \lfloor n/3 \rfloor$;

$S^1 \leftarrow ((x_i, y_i))_{i \in [n_1]}, \;\; S^2 \leftarrow ((x_i, y_i))_{i=n_1+1}^{n}, \;\; \mathbf{x}' \leftarrow (x_{n_1+1}, \ldots, x_n, x_{n+1})$;

$\widehat{\mu} \leftarrow \mathcal{A}_{\mathrm{list}}(S^1, \mathcal{H}), \;\; N \leftarrow \sum_{(x,y) \in S^2} \mathbb{1}_{y \notin \widehat{\mu}(x)}$;

$\mathcal{H}_{\mathbf{x}'} \leftarrow \{h|_{\mathbf{x}'} : h \in \mathcal{H}, |\{i \in [n+1] \backslash [n_1] : h(x_i) \notin \widehat{\mu}(x_i)\}| \leq N + 1\}$;

Sample $(I_1, \ldots, I_{n_2}) \sim \mathsf{Unif}([2n_2])^{n_2}$;

$\widehat{h} \leftarrow A_G(T, \mathcal{H}_{\mathbf{x}'})$ where $T \leftarrow ((I_j, y_{I_j + n_1}))_{j \in [n_2]}$;

return the label $\widehat{h}(2n_2 + 1)$.

Reduction from multiclass learning to list learning

- In step 6 of Algorithm 1, A_G is a multiclass PAC learner for classes $\mathcal{H}$ of bounded graph dimension ($\dim_G(\mathcal{H})$) [Natarajan and Tadepalli, 1988].
- We prove in Proposition H.5 that for any $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, $n \in \mathbb{N}$, $\delta \in (0,1)$, and $S \sim \mathcal{D}^n$, with probability at least $1-\delta$,

$$\mathrm{er}_{\mathcal{D}}(A_G(S,\mathcal{H})) = O\left(\frac{\dim_G(\mathcal{H}) + \log(1/\delta)}{n}\right)$$

Sampled boosting of list learners

- Brukhim et al. [2022] proposed a list sample compression scheme of size $r = O(d^{3/2}\log(n))$ for concept classes of DS dimension d and sample size n.
- Its error rate is $O\left((r\log(n/r) + \log(1/\delta))/n\right)$ by standard techniques for sample compression schemes [David et al., 2016]. There is an extra log factor $\log(n/r)$.
- da Cunha et al. [2024] proposed stable randomized sample compression schemes and a subsampling-based boosting algorithm for weak learners for binary classification whose generalization does not induce the extra log factor in n.
- For $K \in \mathbb{N}$ menus $\mu_1, \ldots, \mu_K$ each of size p, we define their **majority vote** to be $\mu = \mathrm{Maj}(\mu_1, \ldots, \mu_K)$ with

$$\mathrm{Maj}(\mu_1, \ldots, \mu_K)(x) := \{y \in \mathcal{Y} : |\{k \in [K] : y \in \mu_k(x)\}| > K/2\},\ \forall x \in \mathcal{X}.$$

- μ has size $2p - 1$. For $p = 1$, the above definition recovers the majority vote of classifiers.

Sampled boosting of list learners

Algorithm 2: Sampled boosting $\mathcal{A}_{\text{boost}}$ of a list learner $\mathcal{A}_{\text{list}}$

Input: List learner $\mathcal{A}_{\text{list}}$, concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, training sequence $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, $\gamma \in (0, 1/2)$, $\nu \in (0, \gamma/18]$, $\delta \in (0, 1)$.

Output: Menu μ.

for $i = 1, \ldots, n$ **do**
- $\mathcal{D}_1(\{(x_i, y_i)\}) \leftarrow 1/n$;

$\alpha \leftarrow \frac{1}{2} \log\left((1+\gamma)/(1-\gamma)\right)$, $m \leftarrow \mathcal{M}_{\mathcal{A}_{\text{list}}, \mathcal{H}}(1/2 - \gamma, \nu)$, $K \leftarrow \lceil 4 \log(n/\delta)/\gamma \rceil$;

for $k = 1, \ldots, K$ **do**
- Draw m samples $S^k \sim \mathcal{D}_k^m$;
- $\mu_k \leftarrow \mathcal{A}_{\text{list}}(S^k, \mathcal{H})$;
- **for** $i = 1, \ldots, n$ **do**
  - $\mathcal{D}_{k+1}(\{(x_i, y_i)\}) \leftarrow \mathcal{D}_k(\{(x_i, y_i)\}) \exp\left(-\alpha \left(2\mathbb{1}_{y_i \in \mu_k(x_i)} - 1\right)\right)$;
- $\mathcal{D}_{k+1} \leftarrow \mathcal{D}_{k+1} / \left(\sum_{i=1}^{n} \mathcal{D}_k(\{(x_i, y_i)\}) \exp\left(-\alpha \left(2\mathbb{1}_{y_i \in \mu_k(x_i)} - 1\right)\right)\right)$;

return $\mu \leftarrow \mathrm{Maj}\left((\mu_k)_{k \in [K]}\right)$.

Theorem 2.8. Assume that $\mathcal{A}_{\mathrm{list}}$ is a list learner with $\mathcal{M}_{\mathcal{A}_{\mathrm{list}},\mathcal{H}}(1/2-\gamma,\nu) < \infty$ for some $\gamma \in (0,1/2)$ and $\nu \in (0,\gamma/18]$. Then, for any $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, $n \in \mathbb{N}$, and $\delta > 0$, sampling $S \sim \mathcal{D}^n$, with probability at least $1-\delta$, the menu μ produced by $\mathcal{A}_{\mathrm{boost}}$ using $\mathcal{A}_{\mathrm{list}}$ in Algorithm 2 satisfies that

$$\mathrm{er}_{\mathcal{D}}(\mu) = O\left(\frac{\mathcal{M}_{\mathcal{A}_{\mathrm{list}},\mathcal{H}}(1/2-\gamma,\nu)\log(n/\delta)}{\gamma n}\right).$$

Theorem 2.10. There exists a list learner $\mathcal{A}_L$ which for any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\dim(\mathcal{H}) = d$ and sample size $n \in \mathbb{N}$ outputs a menu of size $O\big((e\sqrt{d})^{\sqrt{d}}\log(n)\big)$ with $\varepsilon_{\mathcal{A}_L,\mathcal{H}}(n) = O\left(\frac{d^{3/2}\log(d)\log(n)}{n}\right)$.

- **Theorem 2.11.** There exists a multiclass learner $\mathcal{A}_{\mathrm{multi}}$ such that for any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of DS dimension d, $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, $\delta \in (0,1)$, $n \geq d+1$, and $S \sim \mathcal{D}^n$, with probability at least $1-\delta$, we have

$$\mathrm{er}_{\mathcal{D}}(\mathcal{A}_{\mathrm{multi}}(S,\mathcal{H})) = O\left(\frac{d^{3/2}\log(d)\log(n)+\log(1/\delta)}{n}\right), \quad (5)$$

which implies that

$$\mathcal{M}_{\mathcal{A}_{\mathrm{multi}},\mathcal{H}}(\varepsilon,\delta) = O\left(\frac{d^{3/2}\log(d)\log(d/\varepsilon)+\log(1/\delta)}{\varepsilon}\right), \ \forall \varepsilon,\delta \in (0,1).$$

- The existing upper bound in Brukhim et al. [2022]: $O\left(\frac{(d^{3/2}\log(d)+d\log(\log(n)))\log^2(n)+\log(1/\delta)}{n}\right)$.

Theorem 2.11 (cont'd). If there exists a list learner $\mathcal{A}_{\text{goodlist}}$ of size $f_1(d)$ and expected error rate $\varepsilon_{\mathcal{A}_{\text{goodlist}},\mathcal{H}}(n) \leq f_2(d)/n$ for some functions $f_1 : \mathbb{N} \to \mathbb{N}$ and $f_2 : \mathbb{N} \to [0, \infty)$, then, there exists a multiclass learner $\mathcal{A}_{\text{lin}}$ such that

$$\mathcal{M}_{\mathcal{A}_{\text{lin}},\mathcal{H}}(\varepsilon, \delta) = O\left(\frac{d\log(f_1(d)) + f_2(d) + \log(1/\delta)}{\varepsilon}\right), \ \forall \varepsilon, \delta \in (0, 1).$$

Open Question 1. Does there exist a list learner such that given a concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, its size is $f_1(\text{dim}(\mathcal{H}))$ and its expected error rate is $\varepsilon_{\mathcal{A}_{\text{list}},\mathcal{H}}(n) = f_2(\text{dim}(\mathcal{H}))/n$ for some functions $f_1 : \mathbb{N} \to \mathbb{N}$ and $f_2 : \mathbb{N} \to [0, \infty)$?

- **Proposition 3.1** (Daniely and Shalev-Shwartz 2014, Charikar and Pabbaraju 2023, Aden-Ali et al. 2023). For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $n \in \mathbb{N}$, we have

$$\mu_{\mathcal{H}}(n)/(2en) \leq \varepsilon_{\mathcal{H}} \leq \varepsilon_{\mathcal{H},\text{trans}} \leq \mu_{\mathcal{H}}(n)/n. \tag{6}$$

  Assume that $\mu_{\mathcal{H}}(n) \leq f(\text{dim}(\mathcal{H}))$ for some function $f : \mathbb{N} \to [0, \infty)$ and all $n \in \mathbb{N}$. Then, there exists a learner $\mathcal{A}$ based on orienting the one-clusion graph of the projected concept class [Aden-Ali et al., 2023, Appendix A] with sample complexity $\mathcal{M}_{\mathcal{A},\mathcal{H}}(\varepsilon, \delta) = O\big(\frac{f(\text{dim}(\mathcal{H}))+\log(1/\delta)}{\varepsilon}\big)$, $\forall \varepsilon, \delta \in (0, 1)$.
- Haussler et al. [1994] proved that $\mu_{\mathcal{H}} \leq 2\text{dim}(\mathcal{H})$ for binary classes, which motivates the conjecture for multiclasses.
- **Theorem 3.2.** For any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\text{dim}(\mathcal{H}) = 1$, we have $\mu_{\mathcal{H}}(n) \leq 2, \ \forall n \in \mathbb{N}$. Thus, $\mathcal{M}_{\mathcal{H}}(\varepsilon, \delta) = \Theta\left(\log(1/\delta)/\varepsilon\right)$ for any positive $\varepsilon, \delta \in O(1)$ and any $\mathcal{H}$ with $\text{dim}(\mathcal{H}) = 1$.

- Motivated by the proof for binary classes [Haussler et al., 1994, Lemma 2.4], we consider upper bounding the density by induction on the size of the sequence the class projects to.
- The analysis for binary classes does not apply to general concept classes.
- The analysis in the induction step proceeds seamlessly for some special concept classes where a common label which we call a "pivot" exists for each edge in the last dimension of size greater than 1 in its one-inclusion graph.

Definition 3.5 (Pivot of finite concept class)**.** For any $n \in \mathbb{N}\backslash\{1\}$ and $V_n \subseteq \mathcal{Y}^n$, we define

$$\mathfrak{P}(V_n) := \cup_{y\in\mathcal{Y}} \cup_{y'\in\mathcal{Y}\backslash\{y\}} \big\{(y_1,\ldots,y_{n-1}) \in \mathcal{Y}^{n-1} : (y_1,\ldots,y_{n-1},y),(y_1,\ldots,y_{n-1},y') \in V_n\big\}.$$

$a \in \mathcal{Y}$ is said to be a **pivot** of V_n if $(y_1,\ldots,y_{n-1},a) \in V_n$ for all $(y_1,\ldots,y_{n-1}) \in \mathfrak{P}(V_n)$. When $\mathfrak{P}(V_n) = \emptyset$, every $a \in \mathcal{Y}$ is a pivot of V_n.

Lemma 3.6. Assume that for some $n \in \mathbb{N}\backslash\{1\}$, any $d \in \mathbb{N}$, any $m \in [n-1]$, and any $H \subseteq \mathcal{Y}^m$ with $\dim(H) \le d$ and $|H| < \infty$, we have $\operatorname{avgoutdeg}(\mathcal{G}(H)) \le d$. Consider an arbitrary set $V_n \subseteq \mathcal{Y}^n$ such that $|V_n| < \infty$ and $\dim(V_n) \le d$. If V_n has a pivot, then we have $\operatorname{avgoutdeg}(\mathcal{G}(V_n)) \le d$.

For any $n \in \mathbb{N}\backslash\{1\}$, $a \in \mathcal{Y}$, and $V_n \subseteq \mathcal{Y}^n$ with $|V_n| < \infty$, we define

$$\begin{aligned}\mathfrak{P}_a(V_n) := \cup_{y\in\mathcal{Y}}\big\{(y_1,\ldots,y_{n-1}) \in \mathcal{Y}^{n-1} :&(y_1,\ldots,y_{n-1},y) \in V_n,\\ &(y_1,\ldots,y_{n-1},a) \notin V_n\big\}.\end{aligned}$$

For $\mathbf{y} = (y_1,\ldots,y_{n-1}) \in \mathfrak{P}_a(V_n)$ and the edge $(e_{n,\mathbf{y}}, n)$ in $\mathcal{G}(V_n)$, define

$$L_{\mathbf{y}} := \{y \in \mathcal{Y} : (y_1,\ldots,y_{n-1},y) \in (e_{n,\mathbf{y}}, n)\}.$$

A mapping $\gamma : \mathfrak{P}_a(V_n) \to \mathcal{Y}$ is called a **pivot shifting** on V_n to a if $\gamma(\mathbf{y}) \in L_{\mathbf{y}}$ for all $\mathbf{y} \in \mathfrak{P}_a(V_n)$.
Let Γ_{a,V_n} denote the set of all pivot shifting on V_n to a. For any $\gamma \in \Gamma_{a,V_n}$, we define

$$V_n^{\gamma} := (V_n \backslash \{(\mathbf{y},\gamma(\mathbf{y})) : \mathbf{y} \in \mathfrak{P}_a(V_n)\}) \cup \{(\mathbf{y},a) : \mathbf{y} \in \mathfrak{P}_a(V_n)\};$$

i.e., $V_{n,\gamma}$ is obtained by replacing the label $\gamma(\mathbf{y})$ in $(\mathbf{y},\gamma(\mathbf{y}))$ with a for all $\mathbf{y} \in \mathfrak{P}_a(V_n)$.

- **Lemma 3.8.** For any $a \in \mathcal{Y}$, $V \subseteq \cup_{n=2}^{\infty} \mathcal{Y}^n$ with $|V| < \infty$, and $\gamma \in \Gamma_{a,V}$, we have

$$\text{avgoutdeg}(\mathcal{G}(V^\gamma)) \geq \text{avgoutdeg}(\mathcal{G}(V)).$$

- **Open Question 2.** For any $d \in \mathbb{N}$ and any $V \subseteq \cup_{n=d+2}^{\infty} \mathcal{Y}^n$ with $|V| < \infty$ and $\dim(V) = d$, are there some $a \in \mathcal{Y}$ and $\gamma \in \Gamma_{a,V}$ such that $\dim(V^\gamma) \leq d$?
- A positive resolution of the above question would lead to the conclusion that $\mu_{\mathcal{H}} \leq 2\dim(\mathcal{H})$.

Thank You!

I. Aden-Ali, Y. Cherapanamjeri, A. Shetty, and N. Zhivotovskiy. Optimal pac bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1203–1223, Los Alamitos, CA, USA, nov 2023. IEEE Computer Society. doi: 10.1109/FOCS57990.2023.00071. URL https://doi.ieeecomputersociety.org/10.1109/FOCS57990.202

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1713–1726, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585190. URL https://doi.org/10.1145/3564246.3585190.

Arthur da Cunha, Kasper Green Larsen, and Martin Ritzert.

Boosting, voting classifiers and randomized sample compression schemes. *arXiv preprint arXiv:2402.02976*, 2024.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/fil

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Balas K Natarajan and Prasad Tadepalli. Two new frameworks for learning. In *Machine Learning Proceedings 1988*, pages 402–415. Elsevier, 1988.