Communication Complexity of Classification Problems

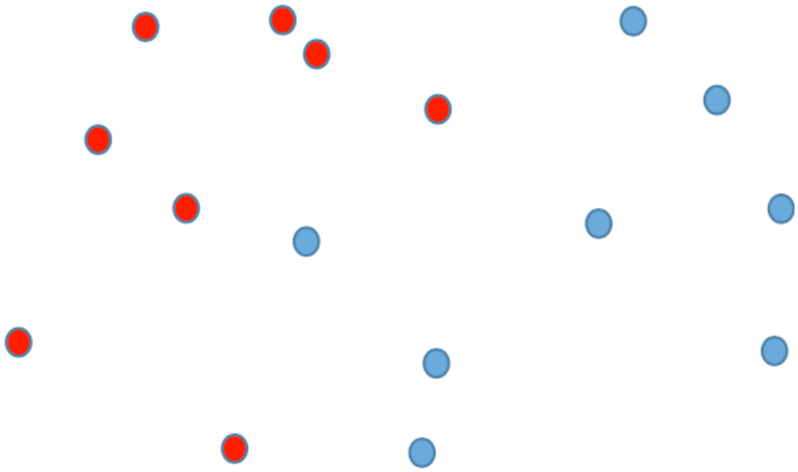
Lecturer: Matt Lawhon

Introduction

In this lecture, we explore the paper "On the Communication Complexity of Classification Problems" by Daniel Kane, Roi Livni, Shay Moran, Amir Yehudayoff in 2018. In this paper, they introduce a new communication model motivated by Yao's model, distributed learning, and interesting real would problems. They then proceed to derive a number of general results on proper/improper, and agnostic/realizable learning, along with some specific results for the interesting motivating problems of the model. As of 2022, the paper has been cited 13 times. Thus, though the communication model has yet to become widespread, it has been the source of inspiration in some subsequent work in learning theory.

Example: Convex Set Disjointness

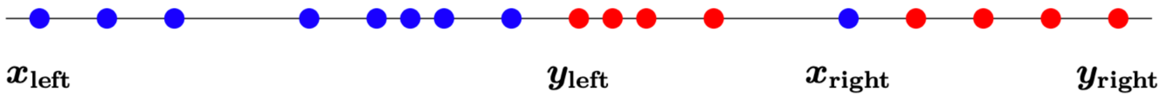
- Setup: Alice and Bob are each given n points in $\mathbb{R}^d$.
- Problem: Do the convex hulls of their inputs intersect?
- Problem with solving this in Yao's model…
  - Transmitting bits doesn't work because they have infinite domains.



An Extension to Yao's Model

The idea here is to extend Yao's model to allow Alice and Bob to *send points in their input* or bits as one unit of communication. Note that it is important to specify that they cannot send points *not in* their input? In the case of Convex Set Disjointness, this would admit a 2-bit solution independent of d, n using an $\mathbb{R}^{nd} \to \mathbb{R}$ bijection.

Example: Convex Set Disjointness
Consider $d = 1$. What is the communication complexity?



- $CC = 3$. Alice communicate 2 endpoints, Bob publishes the output

Suppose Alice and Bob are each given a subset of $U \subset \mathbb{R}^d$ where $d \geq n - 1 \geq 1,\ \ |U| = n$

- It is possible that points in U are affinely independent.
- Which implies that $A \cap B = \emptyset \iff conv(A) \cap conv(B) = \emptyset$
- $CC = \Theta(n)$ because this is just set disjointness.

"In this way, Convex Set Disjointness can be seen as a geometric interpolation between Set Disjointness (when d ≥ n − 1), and Greater-Than (when d = 1)" – Shay Moran

The Model

The setup is usually defined over a given domain and hypothesis space.

- Let X be the domain, and $Z = X \times \{-1,1\}$ be the examples domain.
- Alice and Bob are given $S_a,\ S_b \subset Z$ and can communicate elements of their inputs or bits, for one unit of communication.
- Typically learning something about a hypothesis $h: X \to \{-1,1\}$ for a given hypothesis class H.

Problems we can study

Decision (about a property – we focus on Realizability):

- Given $S_a,\ S_b$ decide if there exists some $h \in H$ that is consistent with $S_a,\ S_b$
- Eg. CSD: S_a labelled 1, S_b labelled 0, H the hypothesis class of half-spaces

Search:

- Given $S_a,\ S_b$ output a hypothesis that makes no-more (or perhaps ϵ more) mistakes than the best $h \in H$
  - If we are confined to output some $h \in H$ as our hypothesis, then we are *properly* learning
  - If the best $h \in H$ is consistent with $S_a,\ S_b$ then we are learning in the *realizable* case, as opposed to the agnostic case

Related work

This communication model can be viewed as yielding distributed sample compression schemes in the search problem setting. A *Sample Compression Scheme* of size $k < n$ is defined by:

- Compressor: $c: (x \times y)^n \to (x \times y)^k$
- Reconstructor: $r: (x \times y)^k \to H$, s.t. $\forall (x,\ y),\ \ r\big(c(x,y)\big) = y$

Sample Compression Schemes were introduced by Littlestone and Warmuth who showed that compression implies (PAC-)learnability and asked whether learnability implies compression. In fact, it is still an open whether every H has a SCS of size $O(VC\text{–}dim(H))$, and the best result is exponential (Yehudayoff et al, 2015).

We note there is a strict difference in sample complexity for distributed and non-distributed. For half planes:

- A trivial SCS of size $d + 1$ using the data as support vectors.
- A lower bound of $\Omega(d \log(n/d))$ is known in the two-party setting for randomized and/or improper learning (Braverman et al, 2019).

More generally, there are SCS of size only depending on VC-Dimension, but this is not the case for DSCS.

Main Results

Theorem 1. Let H be the class of half-spaces in $\mathbb{R}^d$, $d \geq 2$, $\epsilon \leq 1/3$**.** Any Protocol that learns H in the realizable case has sample complexity $\widetilde{\Omega}(d + \log(1/\epsilon))$.
Proof:

- First, we prove $\widetilde{\Omega}(d)$ samples are required.
  - We know that ϵ-approximate non-distributed sample compression schemes for any fixed ϵ (say 1/3), require $\widetilde{\Omega}(d)$ samples.
- Next, we prove $\widetilde{\Omega}(\log\,(1/\epsilon))$ samples are required.
  - Not actually – this proof is 6 pages long.

Some Definitions

- We define analogous complexity classes of P, NP and coNP.
  - H is in P if there is an efficient protocol (in terms of sample complexity) for the realizability problem.
  - H is in NP if there is a poly-log proof that certifies realizability.
  - H is in coNP if there is a poly-log proof that certifies non-realizability.
- VC-Dimension(H)
  - The size of the largest $S \subset X$ s.t. $\forall T \subset S$, $\exists h \in H$ s.t. $\{s | s \in S \wedge h(s) = 1\} = T$
- CoVC-Dimension(H) (aka dual Helly number)
  - The smallest k s.t. every non-realizable sample has a non-realizable subsample of size at most k
  - Also - separator between proper and improper learning of linear separators (SVMS) with optimal sample complexity (finite = proper) (Bousquet et al, 2020)

CoVC/VC-dimension Examples

Consider the following hypothesis classes:

$$H = \{s \subset [n] : |s| = 1\},\ X = [n]$$

- VC-Dim(H) = 1
- CoVC-Dim(H) = n.
    - The sample labeled -1 everywhere of size n is unrealizable and has no non-realizable subsample.

$$H = \{[n] \to \{-1,1\} :\ \forall i \geq \frac{n}{2},\ h(i) = -1\}$$

- VC-Dim(H) = n/2
- CoVC-Dim(H) = 1.
    - Every non-realizable sample must contain an example $(i,\ 1)$ with $i\ \geq\ n/2$, which is not realizable.

$$H = \{f : \mathbb{R}^d \to \{-1,1\} : f(x) = sign(w \cdot (x,1)), w \in R^{d+1}\}$$

- VC-Dimension = $d + 1$.
    - Radon's Theorem – Any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two sets whose convex hulls intersect.
    - $\{v \in \mathbb{R}^d : L_0(v) = L_1(v) = 1\} \cup \{\mathbf{0}\}$
- CoVC-Dimension $\leq 2d + 2$
    - Carathéodory's Theorem – If $x \in \mathbb{R}^d$ lies in the convex hull of a set $P,$ then x can be written as the convex combination of at most $d + 1$ points in P.
    - Let S be a non-realizable set and denote the positive and negatively labelled points $S_+,\ S_-$. The convex hulls of S_+ and S_- intersect.
    - x lies in the convex hull of some $d + 1$ points in S_- and of some $d + 1$ points in S_+.
    - The union of these (of size $2d + 2$) is non-realizable.

Main Result

Theorem 5. The following statements are equivalent for a hypothesis class H

i. H in in P
ii. H is in NP ∩ coNP
iii. H has finite VC dimension and finite coVC dimension
iv. There exists a protocol for the realizability problem for H with sample complexity $\widetilde{O}(dk^2 \log(|S|))$ for $d = VC\ dim(H)$ and $k = coVC\ dim(H)$

We will prove this by showing $i \Rightarrow ii \Rightarrow iii \Rightarrow iv \Rightarrow i$. Note that showing $iv \Rightarrow i \Rightarrow ii$ is trivial.

$\boldsymbol{H}$ in NP ∩ coNP ⇒ $\boldsymbol{H}$ has finite VC dimension and coVC dimension

Theorem 6. For every class H with VC dimension $d \in \mathbb{N} \cup \{\infty\}$,

$$N_H^{np}(n) = \widetilde{\Omega}(\min(d, n))$$

Proof:

- **Lemma 4.** Let H be a hypothesis class and let $R \subseteq X$ be a subset of size n that is shattered by H. $\exists F_a, F_b$ that map n bit-strings to labelled examples from R such that for every $x, y \in \{0,1\}^n$, $x \cap y = \emptyset$ iff the joint sample $S = \langle F_a(x); F_b(y) \rangle$ is realizable by H.
  Proof:
  - Since R is shattered by H, it follows that a sample S with examples from R is realizable by H if and only if it contains no point with two opposite labels. Set $F_a(x) = \{(i, 1): x_i = 1\}$ and set $F_b(y) = \{(i, -1): y_i = 1\}$.
  - If $i \in x \cap y$ then having $(i,\ 1) \in F_a(x)$ and $(i,\ -1) \in F_b(y)$ implies that the joint sample S is not realizable. On the other hand, since R is shattered, we have that if $x \cap y = \emptyset$, then S is realizable.
- Let R be a shattered set of size d. Since all $x \in$ R can be encoded by $O(log(d))$ bits, it follows that every NP-proof of sample complexity T for the realizability problem for H implies an NP-proof for DISJ$_d$ with bit-complexity $O(T\ log(d))$ in Yao's model.
- $\mathrm{NP}^{\mathrm{CC}}$(DISJ$_d$) = $\Omega(d)$ so $\mathrm{T} = \Omega(d)$ (or all of R can be sent). QED.

Theorem 7. For every class H with coVC dimension $k \in \mathbb{N} \cup \{\infty\}$,

$$N_H^{conp}(n) = \widetilde{\Theta}(\min(k, n))$$

Proof of upper bound (lower bound similar to 6):

- assume that the coVC dimension is $k < \infty$.
- If $S = \langle S_a; S_b \rangle$ is not realizable then it contains a non-realizable sample S of size at most coVC-dim$(H) = k$ that serves as a proof that S is not realizable. If $k = \infty$, then the whole sample S serves as a proof of size n that it is not realizable.

$\boldsymbol{H}$ has finite VC and coVC dimension ⇒ $\boldsymbol{H}$ has a protocol with bounded sample complexity.

For every class H with VC-dim$(H) = d$ and coVC-dim$(H) = k$ there exists a protocol for the realizability problem over H with sample complexity $O(dk^2 logk\ log|S|)$.

Proof idea:

- We give a protocol derived as a tailored version of adaboost s.t. at iteration t Alice and Bob agree on a hypothesis h_t which is an α-weak hypothesis for $\alpha = \frac{1}{2} - \frac{1}{5k}$ for Alice's distribution p_t^a on S_a and Bob's distribution p_t^b on S_b.
- This protocol will output "non-realizable" if at any iteration such a protocol doesn't exist. If it never outputs this after $O(k \log|S|)$ rounds, it outputs "realizable."

How many samples are enough at iteration t to achieve the desired $\boldsymbol{\alpha}$?

- **$\boldsymbol{\epsilon}$-net theorem.** Let H be a class of VC dimension d and let S be a realizable sample. For every distribution p over S there exists a subsample S' of S of size $O(d\log(1/\epsilon)/\epsilon)$ s.t.
  $$\forall h \in H : L_{S'}(h) = 0 \ \Rightarrow L_p(h) \le \epsilon.$$
- Thus, at each iteration a sample of size $O(dk\log(k))$ suffices.

We now observe two claims that will all but suffice to show the protocol works as desired.

Claim 6.1. Let H be a class with coVC dimension $k > 0$. For any unrealizable sample S and for any $h_1, \dots, h_T \in H$, there is $(x, y) \in S$ so

$$\frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] \geq \frac{1}{k}$$

Proof: Pick some unrealizable subsample $S'$of S s. t. $|S'| \leq k.$ Note that $L_{S'}(h) \geq 1/k$

$$\max_{(x,y)\in S'} \frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] \geq \frac{1}{|S'|}\sum_{(x,y)\in S'} \frac{1}{T}\cdot \sum_{t=1}^{T} 1[h_t(x) \neq y]$$

$$\geq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|S'|}\sum_{(x,y)\in S'} 1[h_t(x) \neq y] \geq \frac{1}{k}$$

Lemma 2. Set η in Adaboost to be $ln2$. Let $T \ \geq\ 2k \log|S|$ for k > 0, and have $h_1, \ . \ . \ . \ , \ h_T$ denote the weak hypotheses returned by an arbitrary α-weak learner with $\alpha = \frac{1}{2} - \frac{1}{5k}$ during the execution of Adaboost. Then, for every $(x,\ y) \in S$:

$$\frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] \leq \frac{1}{k}$$

NTS: If the protocol terminates then the sample is realizable.

- If $T \geq 4(k+1)\log|S|$ (by lemma 2)

$$\forall (x,y) \in \langle S_a,\ S_b\rangle : \sum_{t=1}^{T} 1[h_t(x) \neq y] < \frac{1}{2(k+1)}$$

- Thus, by claim 6.1, the sample is realizable.
- Further, this guarantees that if the given sample is not realizable, then this protocol will find such a sample.
- Thus, we have shown $i \Rightarrow ii \Rightarrow iii \Rightarrow iv \Rightarrow i$

Conclusions

- The sample complexity of the realizability problem over H is either $O(\log n)$ or $\widetilde{\Omega}(n)$.
- Convex Set Disjointness can be decided by $\widetilde{\mathrm{O}}(d^3 \log\ n)$ points.

$\mathbf{P} = \mathbf{NP} \cap \mathbf{coNP}$ for Realizability Problems

This is analogous to the classic result in standard communication complexity. The derivation of the result, however, comes about in an entirely different manner, in that the algorithm showing this is a boosting protocol parameterized by VC-Dimension and coVC-Dimension.

It is also of note that while in standard communication complexity, deterministic communication complexity is symmetrically upper bounded by nondeterministic and co-nondeterministic communication complexity, $O(N_0 \cdot N_1)$, this is not the case here. The authors speculate this may be because while decision problems are closed under negation, realizability problems are not.

Open Questions, Further Research

- Closing the gap for the Convex Set Disjointness problem.
$$\widetilde{\Omega}(d + \log n), \Omega(d \log(n/d)), \tilde{O}(d^3 \log n)$$
- Provide a combinatorial upper bound on $N_H^{np}(n)$
    - This would be implied by another open problem – the existence of *proper sample compression schemes* of polylogarithmic sample size.
- Distributed Sample Compression for more than two parties
- Multiclass categorization.

Bibliography

- Kane, Daniel, et al. "On communication complexity of classification problems." *Conference on Learning Theory*. PMLR, 2019.
- Littlestone, Nick, and Manfred Warmuth. "Relating data compression and learnability." (1986).
- Moran, Shay, and Amir Yehudayoff. "Sample compression schemes for VC classes." *Journal of the ACM (JACM)* 63.3 (2016): 1-10.
- Braverman, Mark, et al. "Convex set disjointness, distributed learning of halfspaces, and LP feasibility." *arXiv preprint arXiv:1909.03547* (2019).
- Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *icml*. Vol. 96. 1996.