On communication complexity of classification

Shay Moran (Princeton)

Joint with Daniel Kane (UCSD), Roi Livni (TAU), and Amir Yehudayoff (Technion)
Warmup (i): convex set disjointness

• Alice’s input: \( n \) points in \( \mathbb{R}^d \)
• Bob’s input: \( n \) points in \( \mathbb{R}^d \)
• Decide: do the convex hulls of their inputs intersect?

• Extension of sparse set disjointness
Convex Set Disjointness

• Can we model this problem in Yao’s model?
  • obstacle: input domain is infinite…
    • possible solution: discretization

• We extend Yao’s model by allowing the parties to send input points
  • Each input point costs one sample complexity unit

• 1-dim CSD (next slide)
Convex Set Disjointness in 1D

Alice’s input: $n$ points
Bob’s input: $n$ points

1. Alice sends the two endpoints of her input

2. Bob checks whether the intervals intersect and publishes the output

Output: not disjoint

- Communication complexity: $3 = 2$ points + 1 bit (for the output)
- more interesting in higher dimensions (more details later)
Warmup (ii): distributed sample compression schemes

**Goal:**
output $h$ that is consistent with all examples

$X$ - a domain

$H$ – a class of “$X \rightarrow \{0,1\}$” functions //hypothesis class

e.g. Halfspaces

Alice and Bob are given examples labeled by an unknown $h \in H$

- Alice’s input: $S_a = (x_1, y_1), \ldots, (x_n, y_n)$

- Bob’s input: $S_b = (u_1, v_1), \ldots, (u_n, v_n)$

**Goal:** output $h : X \rightarrow \{0,1\}$ that is consistent with all examples

One-way protocols are Sample Compression Schemes
One-way protocols =
Sample Compression Schemes

Input = \((x_1, y_1)\ldots(x_n, y_n)\)

Output = \(h\)

\((x_{i_1}, y_{i_1})\ldots(x_{i_d}, y_{i_d})\)

Consistency: \(h(x_i) = y_i\) for \(i=1,\ldots,n\)
(including the examples that were not sent)
Example: Sample Compression Scheme of size d+1 for Halfspaces
Sample Compression Schemes (SCS)

- A (universal) method for deriving generalization bounds in statistical learning theory

- Open question (Littlestone & Warmuth ’86):
  
  Does every $H$ have SCS of size $\text{VCdim}(H)$
  
  (best upper bound is exponential in $\text{VCdim}$ [M-Yehudayoff ‘15])

- This work: a distributed extension of sample compression schemes

  - **Question:** Can the SCS of size $d+1$ for Halfspaces be extended to the distributed case? (i.e. input sample is distributed between Alice and Bob)
Overview

• Model

• Decision problems
  • Convex set disjointness
  • The realizability problem

• Search problems
  • Realizable/Agnostic learning
  • Proper/non-proper learning

• Summary
Communication model
Communication Model

\( \mathcal{X} \) – domain
\{0,1\} – labels

Two party communication protocols

• Parties inputs are **samples**
  • Alice’s input: \( S_a = (x_1, y_1), \ldots, (x_n, y_n) \)
  • Bob’s input: \( S_b = (u_1, v_1), \ldots, (u_n, v_n) \)

• Parties may send examples from their inputs (and/or bits)
  • but **not** arbitrary examples

• Sample complexity = total number of examples/bits sent
Problems we study: decision problems

- \( h: \mathcal{X} \rightarrow \mathcal{Y} \): a hypothesis
- \( \mathcal{H} \): a hypothesis class

The realizability problem
Given input samples \( S_a, S_b \) decide if there is \( h \in \mathcal{H} \) that is consistent with the joint sample \( S = (S_a, S_b) \)

- Focus on infinite classes \( \mathcal{H} \) (uniform model like e.g. Turing machines)
- Convex set disjointness - \( \mathcal{H} \) is half-spaces
Problems we study: search/learning problems

- **H**: hypothesis class

**Learning problems**

Given input samples $S_a, S_b$ output a hypothesis that makes no more mistakes on $S = (S_a, S_b)$ than the best $h \in H$

- Proper/non-proper
- Realizable/agnostic
- Allow for $\epsilon$ slackness (may make $\epsilon$ more errors than the best)
Decision problems

Convex set disjointness
Convex set disjointness

**Upper bound:**

Theorem. There is a protocol for convex set disjointness with sample complexity $\tilde{O}(d^3 \log n)$

**Lower bound:**

Theorem.
1. A lower bound of $\tilde{\Omega}(d)$ holds always
2. A lower bound of $\tilde{\Omega}(\log n)$ holds for $d \geq 2$

- In particular the d+1 size sample compression scheme for halfspaces can not be extended to the distributed case

**Open:** find tight bounds
Convex set disjointness: upper bound

Lemma [Center-subset].
Let $X$ be a set of $n$ points in $\mathbb{R}^d$. Then there is $Y \subseteq X$ of size $\tilde{O}(d / \epsilon)$ such that every half-space containing $Y$ contains $(1 - \epsilon)n$ points from $X$.

• Also extends for arbitrary weights on the points in $X$
Convex set disjointness: upper bound

Alice’s input: \( n \) points
Bob’s input: \( n \) points
Convex set disjointness: upper bound

Alice’s input: \( n \) points
Bob’s input: \( n \) points
Initialize the weight of each point to 1
Convex set disjointness: upper bound

Alice’s input: $n$ points
Bob’s input: $n$ points

Initialize the weight of each point to 1

1. Alice & Bob publish an $1/(8d)$-center-subset w.r.t current weights (of size $O(d^2)$)
Convex set disjointness: upper bound

Alice’s input: $n$ points
Bob’s input: $n$ points

Initialize the weight of each point to 1

1. Alice & Bob publish $1/(8d)$-center-sets w.r.t current weights
2. If the hulls of the center-sets intersect output “intersect”
Convex set disjointness: upper bound

Alice’s input: \( n \) points
Bob’s input: \( n \) points

Initialize the weight of each point to 1

1. Alice & Bob publish \( 1/(8d) \)-center-sets w.r.t current weights

2. If the hulls of the center-sets intersect
   output “intersect”

3. Else, find a separator and double the weights of misclassified points
Convex set disjointness: upper bound

Alice’s input: \textit{n points}
Bob’s input: \textit{n points}

Initialize the weight of each point to 1

1. Alice & Bob publish \(1/(8d)\)-center-sets w.r.t current weights
2. If the hulls of the center-sets intersect output “intersect”
3. Else, find a separator and double the weights of misclassified points
Convex set disjointness: upper bound

Alice’s input: \( n \) points  
Bob’s input: \( n \) points

Initialize the weight of each point to 1

1. Alice & Bob publish \( 1/(8d) \)-center-sets w.r.t current weights
2. If the hulls of the center-sets intersect output “intersect”
3. Else, find a separator and double the weights of misclassified points
4. Repeat 1-3 for \( 4d \log n \) times
5. Output “disjoint”
Convex set disjointness: upper bound

- If the hulls are **disjoint** then the output will be "**disjoint**"

- If the hulls intersect, then (by Caratheodory) there is a subset $A$ of at most $2d$ points whose hulls intersect:

- At every round $T$ we have $W_T(A) \leq W_T$ (all input points)

- At least one point from $A$ **doubles** in every round, so: $W_T(A) \geq 2d \cdot 2^{T/2d}$

- Since Alice&Bob send epsilon=1/8d **center subsets**:

$$W_T \text{(all input points)} \leq 2n \cdot (1+1/8d)^T \leq 2n \cdot 2^{T/4d}$$

- So, the protocol proceeds only for as long as $2d \cdot 2^{T/2d} \leq 2n \cdot 2^{T/4d}$ which implies that $T \leq 4d \log n$
Convex set disjointness

**Question:** When the hulls are disjoint, can Alice and Bob find a separating hyperplane?

- Yes, but requires a more subtle analysis
- Related to *properly learning* halfspaces (more details later)
Decision problems

The realizability problem
The realizability problem: the class $P$

Goal: decide whether the input samples are simultaneously consistent with some $h$ in $H$.

$P$ – all classes $H$ for which the realizability problem can be decided with sample complexity $\text{polylog}(n)$, where $n$ is the joint input sample size.

- $P$ is a set of hypothesis classes
- Half-spaces are in $P$ (by convex set disjointness upper bound)
Understanding the class P

• Goal: characterize the classes $H$ for which the realizability problem can be decided fast

• We will give two characterizations
  
  • combinatorial
  
  • complexity-theoretical (using analogs of NP and coNP)
The realizability problem: \( NP \) and \( coNP \)

The class \( NP \)

\( H \) is in \( NP \) if there are two predicates \( A,B \) such that for every \( S = (S_a, S_b) \):

- If \( S \) is realizable then there is a \( \text{polylog}(n) \) size proof \( p \) such that \( A(S_a, p) = B(S_b, p) = \text{True} \),

- If \( S \) is not realizable then \( A(S_a, p) = \text{false} \) or \( B(S_b, p) = \text{false} \) for any proof

- The proof must consist only of examples from \( S \) (and/or bits).
- Define \( coNP \) similarly.
NP and coNP sample complexities of half-spaces

**Claim.** The NP sample complexity of half-spaces in $d$ dimensions is at most $d+1$.

S is realizable $\rightarrow$ there is separating hyperplane $\rightarrow$ the maximum margin hyperplane can be encoded using at most $d+1$ examples (support vectors)
NP and coNP sample complexities of half-spaces

**Claim.** The coNP sample complexity of half-spaces in $d$ dimensions is at most $2d$.

S is not realizable $\rightarrow$ the positive and negative convex hulls intersect $\rightarrow$ can be certified with $2d$ examples:
A characterization of P

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

1. $H$ is in $P$.
2. $H$ is in $NP \land coNP$.
3. $H$ has finite $VC$ dimension and $coVC$ dimension.

Moreover, if either the VC or coVC are unbounded then the sample complexity is $\Omega(n)$.

- The coVC dimension of $H$ is the min $k$ such that every non-realizable sample contains a non-realizable subsample of size at most $k$.

- $1 \rightarrow 2$: trivial
- $2 \rightarrow 3$: via reductions to classical results in Yao’s model (set-disjointness)
- $3 \rightarrow 1$: similar to the convex set-disjointness upper bound
A characterization of $\text{coNP}$

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

1. $H$ is in $\text{coNP}$.
2. $H$ has a finite *coVC dimension*.

- The coVC dimension of $H$ is the min $k$ such that every **non-realizable** sample contains a **non-realizable** subsample of size $k$.

- $1 \rightarrow 2$: via reductions to classical results in Yao’s model (set-disjointness)
- $2 \rightarrow 1$: the non-realizable subsample of size $k$ serves as a proof
Open: a characterization of NP

**Conjecture.** The following statements are equivalent for a hypothesis class $H$.

1. $H$ is in NP.
2. $H$ has a finite VC *dimension*.

We show $1 \rightarrow 2$ via reductions to known statements in Yao’s model (set-disjointness).

Would follow from establishing the existence of proper sample compression schemes for every VC class (which would be a breakthrough).
One more result concerning the realizability problem

• A natural possibility of studying realizability problems in Yao’s model is via discretization:

  • Assume a prespecified arbitrary finite \( R \subseteq \mathcal{X} \)

  • Inputs samples come from \( R \) (and labelled by unknown \( h \) in \( H \))

  • \( R \) is known to both Alice and Bob

  • Decide the realizability problem using \( poly \log(| R |) \) bits

  • Do the class \( P \) change under this (nonuniform) definition?
Theorem. The following statements are equivalent for a hypothesis class $H$.

1. $H$ is in $P$.

2. For any finite $R \subseteq \mathcal{X}$ there is a protocol for the realizability problem of $H|_R$ in Yao’s model of sample complexity $O(\log^2 |R|)$. 

A compactness result + connection to Yao’s model
Learning problems
Learnability - definitions

Definition.
$H$ is learnable with sample complexity $T(\epsilon)$ if there is a protocol that for every input sample $S=(S_a,S_b)$ outputs a hypothesis with error at most $\epsilon$ more than the best $h \in H$

- Realizable – input samples restricted to be realizable
- Proper – output hypothesis must reside in $H$
Realizable case

**Upper bound:**

**Theorem** [Balcan Blum Fine Mansour ‘12]. Every $\mathcal{H}$ is learnable in the realizable case with sample complexity $O(d \log(1/\epsilon))$, where $d$ is the VC dimension of $\mathcal{H}$.

**Lower bound:**

**Theorem.**
1. A lower bound of $\Omega(d)$ holds for any $\mathcal{H}$ ($\epsilon = 1/4$).
2. A lower bound of $\Omega(\log(1/\epsilon))$ holds for half-planes.
Proper learning in the realizable-case

**Theorem.**
1. If $H$ is in $P$ then* it can be properly learned with $O(\log(1/\epsilon))$ sample complexity (the $O()$ hides polynomial dependency on VC and coVC)

2. If $H$ is not in $P$ then any protocol that properly learns it must use $\Omega(1/\epsilon)$ sample complexity.

* - our proof assumes that $H$ is “nice”: either the domain $X$ is countable, or $H$ is closed in the product topology on $\{0,1\}^X$

- Analysis is more subtle than in the realizability problem
Agnostic case

**Lower bound:**

Theorem. Let $H$ be the class of singletons over $\mathbb{N}$. Then every protocol that learns $H$ agnostically has sample complexity $\tilde{\Omega}(1/\epsilon)$.

**Upper bound:**

Theorem. Every $H$ is learnable in the agnostic case with sample complexity $O\left(1/\epsilon^{2d+2}\right) = o(1/\epsilon^2)$ (e.g., halfspaces).

Open: what is the correct bound? halfspaces?
Summary

• An extension of Yao’s model
• Decision problems
  • Convex set disjointness
    • Upper and lower bounds (open: not tight)
  • The realizability problem
    • P,NP,coNP
    • Open: NP = finite VC dim?
• Search problems
  • Agnostic/Realizable/Proper/Non-proper
  • Separations
  • Open: sample complexity of agnostic learning
Misc. slides
Connection with $P=\mathsf{NP} \land \mathsf{coNP}$ in Yao’s model?

**Corollary.** $P=\mathsf{NP} \land \mathsf{coNP}$ for **realizability** problems.

- Proof gives a protocol with sample complexity $\tilde{O}(s_1 \cdot s_0^2 \log |S|)$ where $s_1, s_0$ are the $\mathsf{NP}$, $\mathsf{coNP}$ samp. comp.

**Theorem**[Aho Ullman Yannakakis ‘83].

$P=\mathsf{NP} \land \mathsf{coNP}$ in Yao’s model for any decision problem

- Proof gives a protocol with sample complexity $n_1 \cdot n_0$ where $n_1, n_0$ are the $\mathsf{NP}$, $\mathsf{coNP}$ bit comp.

How do these two results relate to each other?
Connection with $P=NP \land coNP$ in Yao’s model?

- Different proofs: in Yao’s model combinatorial, our proof is more analytical (fractional combinatorics)

- **Conjecture.** $P \neq NP \cap coNP$ in our model for general decision problems. (have explicit candidate)

- Different bounds: $s_1 \cdot s_0^2 \log |S|$ vs $n_1 \cdot n_0$
  - ours is non-symmetric
  - ours has an extra $\log |S|$ factor
    - necessary (convex set disjointness)