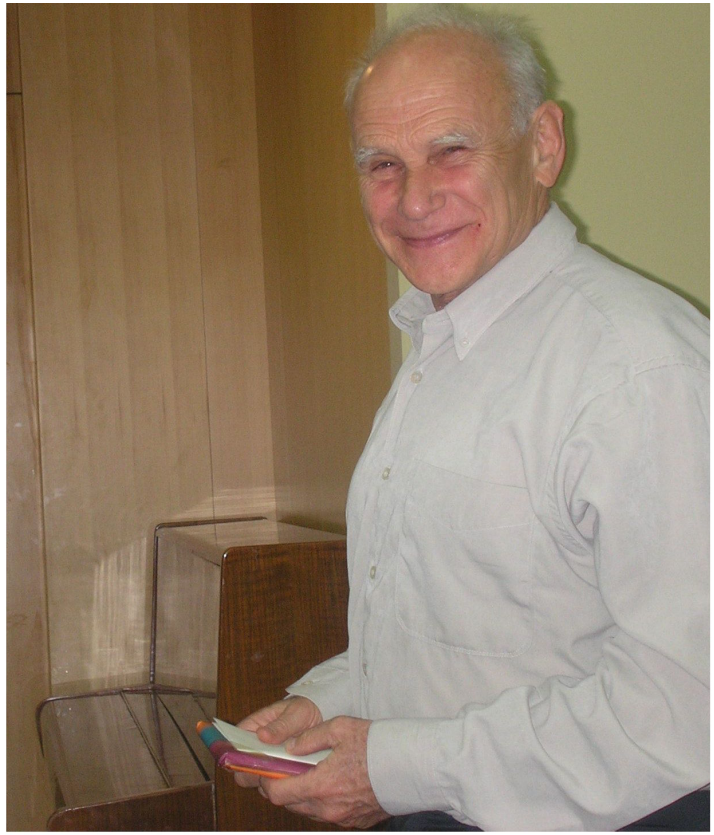
# Gadi Moran (16.05.38 – 01.01.16)



# Sign rank, machine learning, and communication complexity

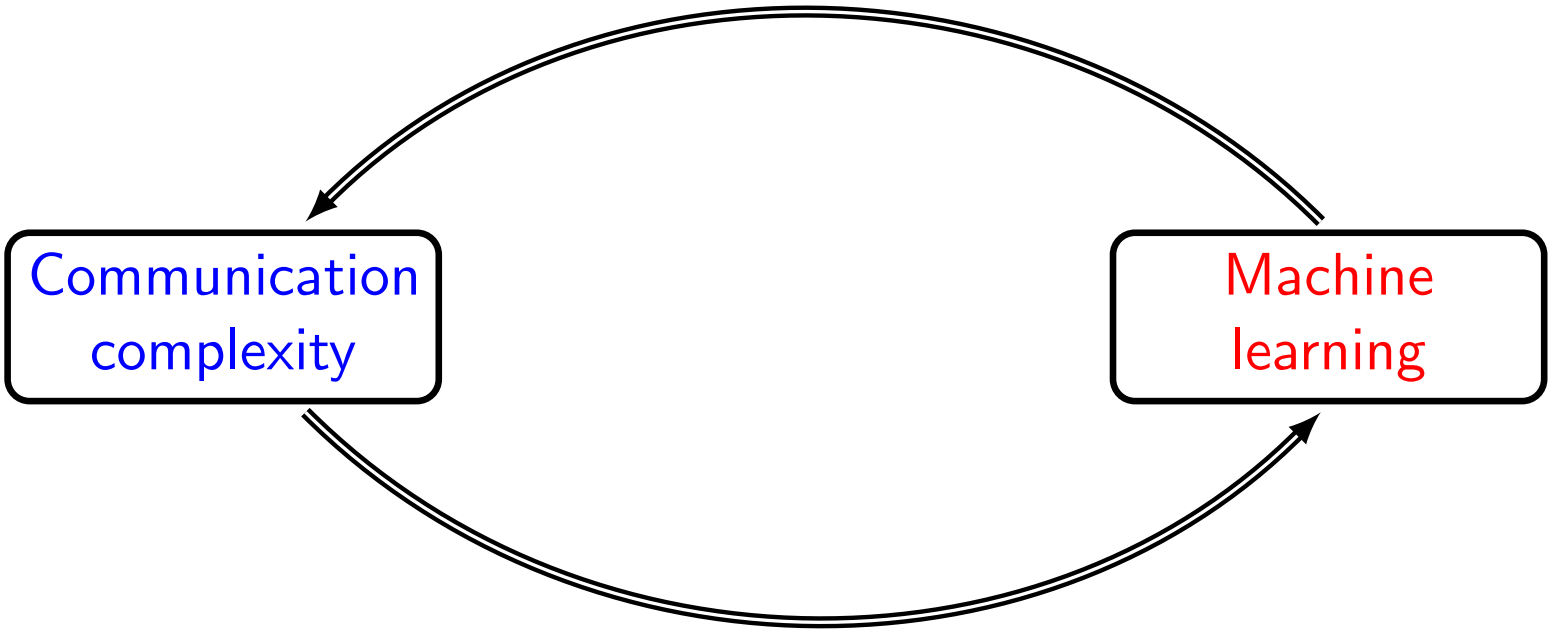
Shay Moran

Technion

January 18, 2016

Partially based on joint work with Noga Alon and Amir Yehudayoff.

# Correspondences between machine learning and communication complexity quantities




1. Unbounded error comm. complexity vs dimension complexity

[Paturi and Simon '86, Ben-David, Eiron, and Simon '03]

2. One-way complexity under product distributions vs VC dimension

[Kremer, Nissan, and Ron '94]

3. Discrepancy vs margin complexity

[Linial and Shraibman '08]

# Plan

- Sign rank in communication complexity
- Sign rank in machine learning
- A "paradox"
  - Resolution 1
  - Resolution 2
- An open problem
- Summary

# Deterministic, randomized, and unbounded error communication complexity

$f$ – a boolean function

$\mathcal{D}(f)$ - deterministic communication complexity

$\mathcal{R}_\epsilon(f)$ - randomized (private coin) communication complexity

$\mathcal{R}_\infty(f) = \min\{\mathcal{R}_\epsilon(f) : \epsilon < \frac{1}{2}\}$

# Rank, approximate rank, and sign rank

$M$ – a boolean matrix

$\mathrm{rank}(M)$

$\mathrm{rank}_\epsilon(M) = \min\{\mathrm{rank}(R) : |R_{i,j} - M_{i,j}| \leq \epsilon\}$

$\mathrm{signrank}(M) = \min\{\mathrm{rank}_\epsilon(M) : \epsilon < \frac{1}{2}\}$

# Rank, approximate rank, and sign rank

$M$ – a boolean matrix

$\text{rank}(M)$

$\text{rank}_\epsilon(M) = \min\{\text{rank}(R) : |R_{i,j} - M_{i,j}| \leq \epsilon\}$

$\text{signrank}(M) = \min\{\text{rank}_\epsilon(M) : \epsilon < \frac{1}{2}\}$

Equivalently, for a sign matrix $S$, the sign rank is defined as $\min\{\text{rank}(R) : \text{sign}(R) = S\}$

# The logarithms of ranks lower bound the communication complexities

$f$ – boolean function
$M_f$ – matrix representing $f$

$\log \operatorname{rank}(M_f) \leq \mathcal{D}(f)$ [Mehlhorn and Schmidt '82]

$\log \operatorname{rank}_\epsilon(M) \leq \mathcal{R}_\epsilon(f)$ [Krause '96]

$\log \operatorname{signrank}(M) \leq \mathcal{R}_\infty(f)$ [Paturi and Simon '86]

# Are the log-ranks lower bounds tight?

$f$ – boolean function
$M_f$ – matrix representing $f$

Log rank conjecture: $\mathcal{D}(f) \leq \text{poly}\log \text{rank}(M_f)$?
[Lovász and Saks '88]

Log approx. rank conjecture: $\mathcal{R}_\epsilon(f) \leq \text{poly}\log \text{rank}_\epsilon(M)$?
[Lee and Shraibman '09]

Log sign rank theorem: $\mathcal{R}_\infty(f) \leq \log \text{signrank}(M) + 2$!!!
[Paturi and Simon '86]

# Recapitulation

1. Sign rank captures unbounded error communication complexity

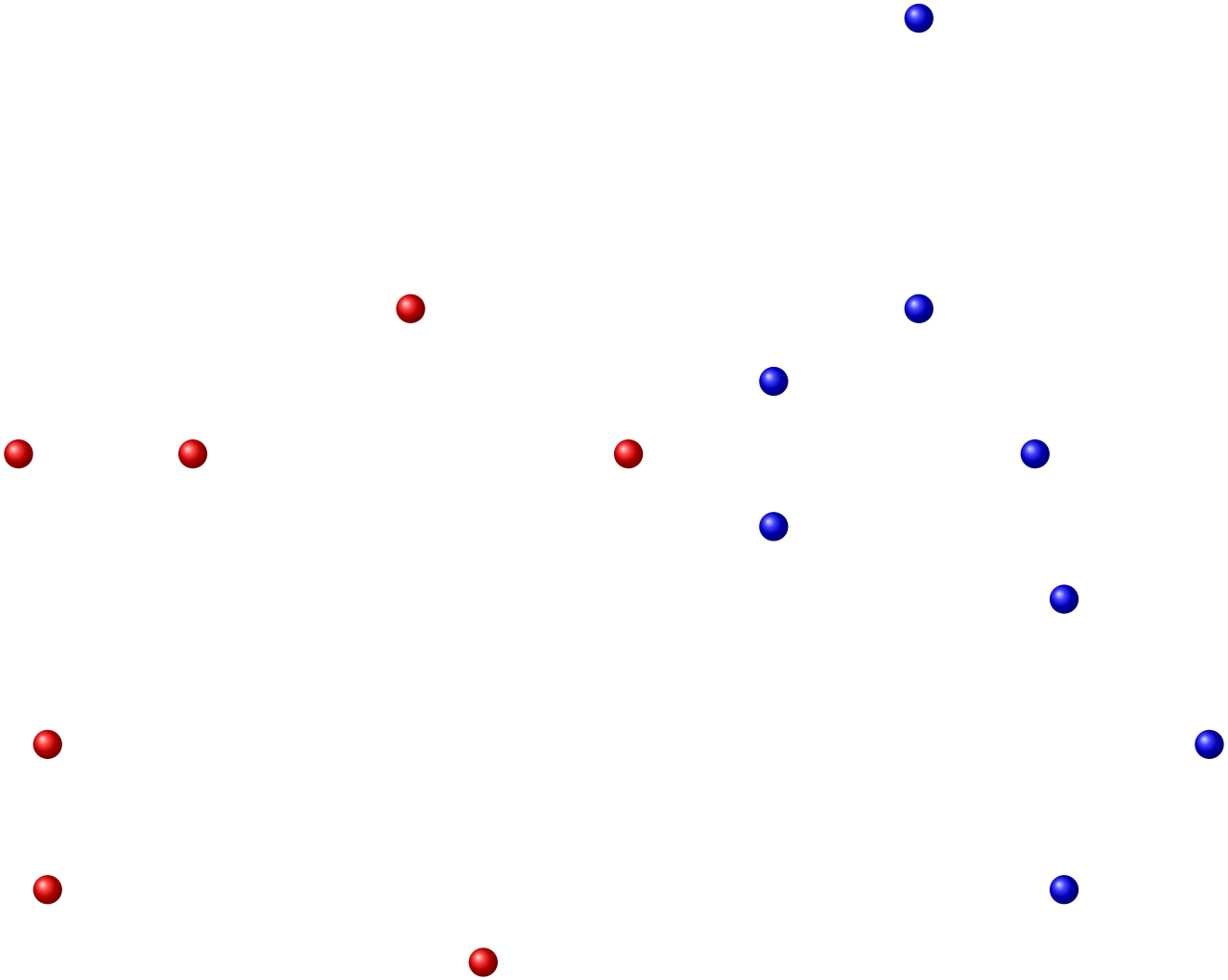
# Plan

# The support vector machine algorithm

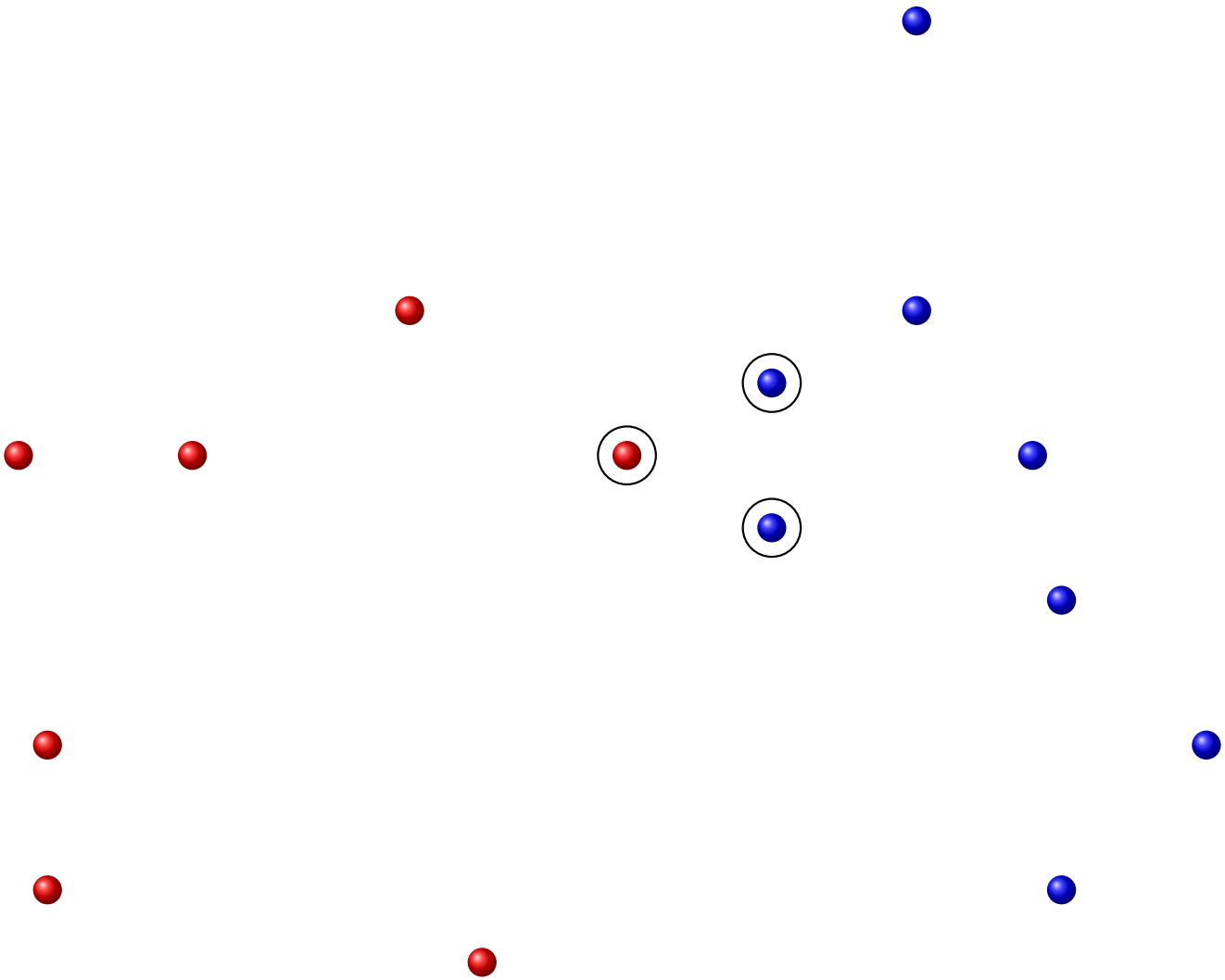
Input: two linearly separable sets $R, B \subseteq \mathbb{R}^d$

Output: hyperplane of maximum margin which separates $R$ from $B$

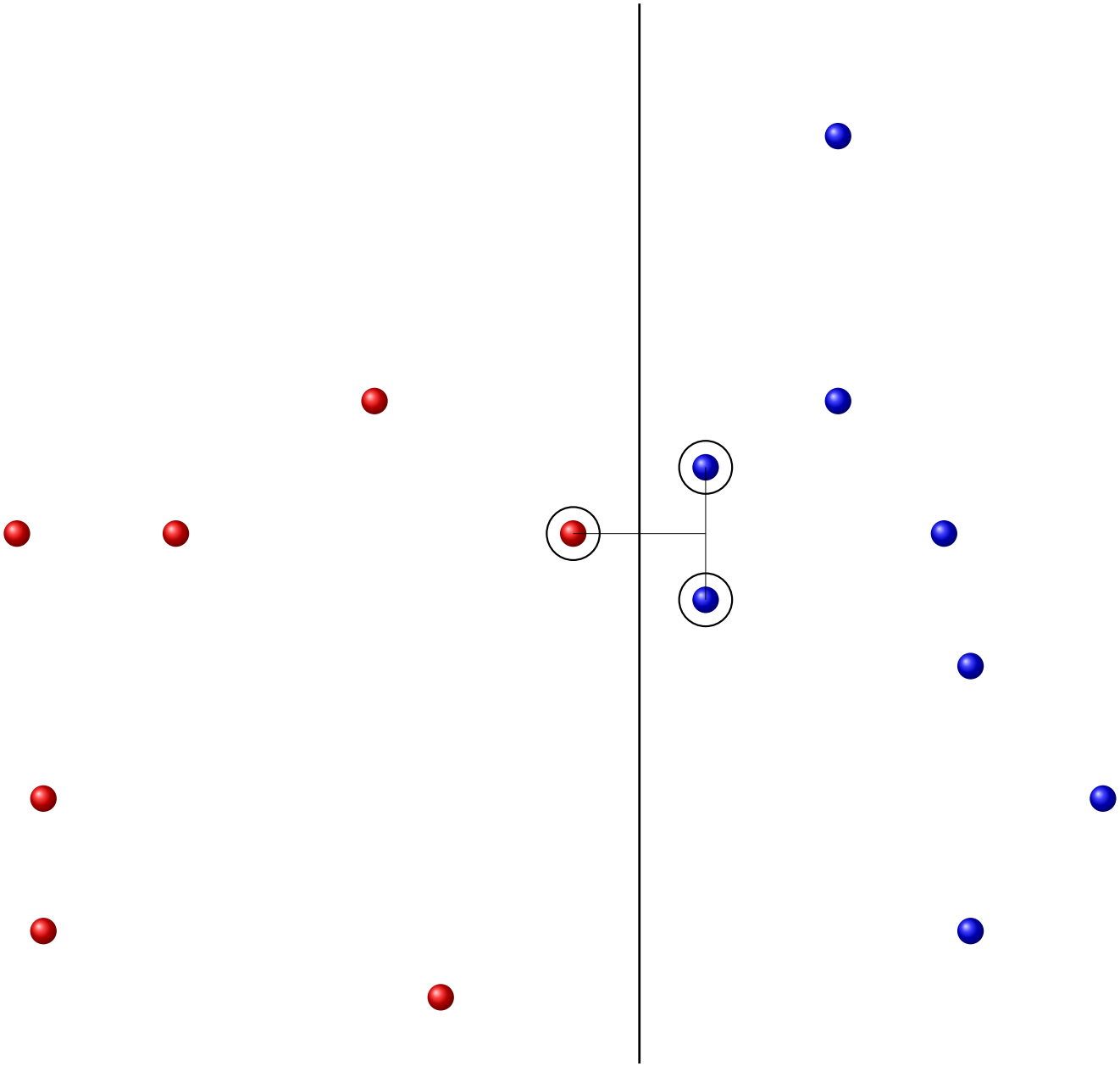
# Support vector machines: illustration



# Support vector machines: illustration



# Support vector machines: illustration



# Extending the applicability of SVM

$X$ – a set
$C \subseteq \{\pm 1\}^X$ – a concept class

SVM can be applied when $X = \mathbb{R}^d$ and $C$ contains half-spaces

**Q:** How to use SVM when $C$ is arbitrary?
**A:** Reduce $C$ to to half spaces:

# Extending the applicability of SVM

$X$ – a set
$C \subseteq \{\pm 1\}^X$ – a concept class

SVM can be applied when $X = \mathbb{R}^d$ and $C$ contains half-spaces

**Q:** How to use SVM when $C$ is arbitrary?
**A:** Reduce $C$ to to half spaces:

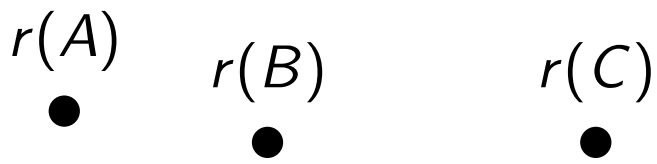
$r : X \to \mathbb{R}^d$ separates $C$ if

$\forall c \in C$, $r(c^{-1}(+1))$ is linearly separable from $r(c^{-1}(-1))$.

e.g. kernel functions

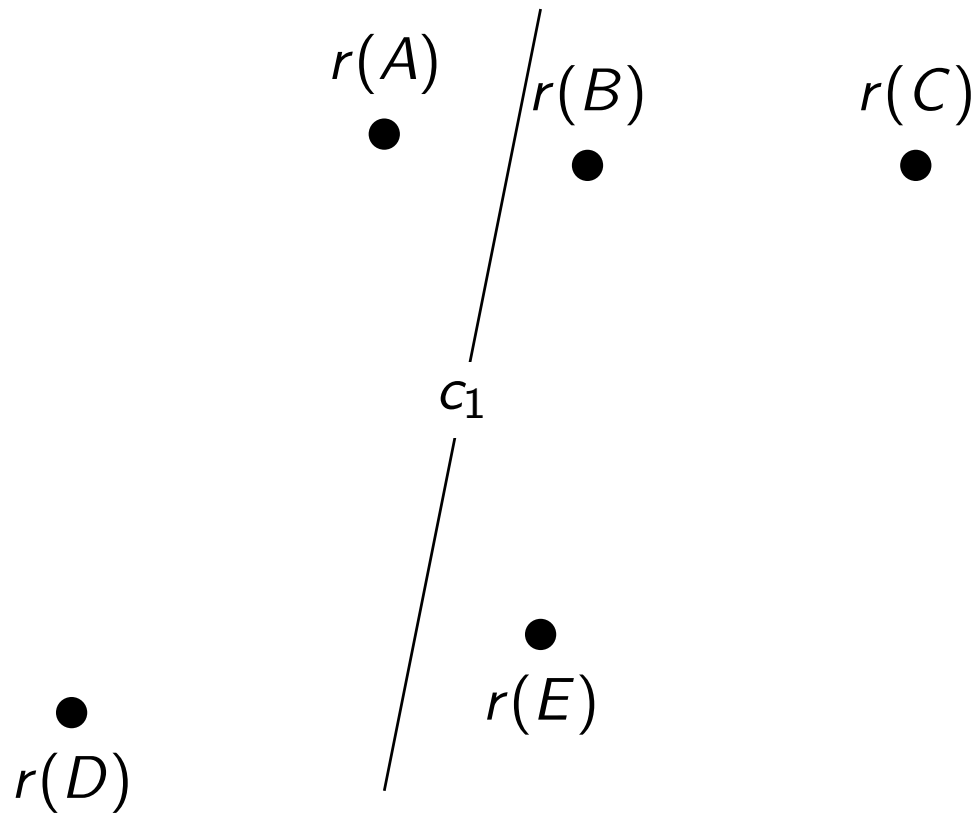
# Example

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $c_1$ | $+$ | $-$ | $-$ | $+$ | $-$ |
| $c_2$ | $+$ | $+$ | $+$ | $-$ | $-$ |
| $c_3$ | $+$ | $+$ | $-$ | $-$ | $-$ |



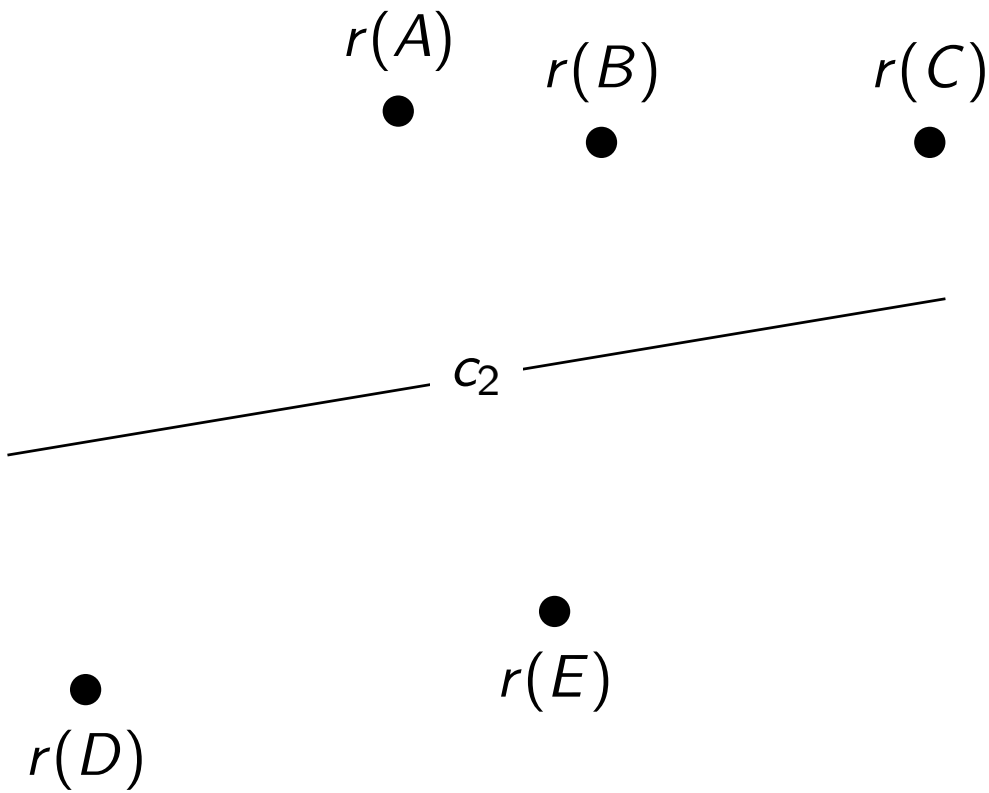

# Example

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $c_1$ | $+$ | $-$ | $-$ | $+$ | $-$ |
| $c_2$ | $+$ | $+$ | $+$ | $-$ | $-$ |
| $c_3$ | $+$ | $+$ | $-$ | $-$ | $-$ |



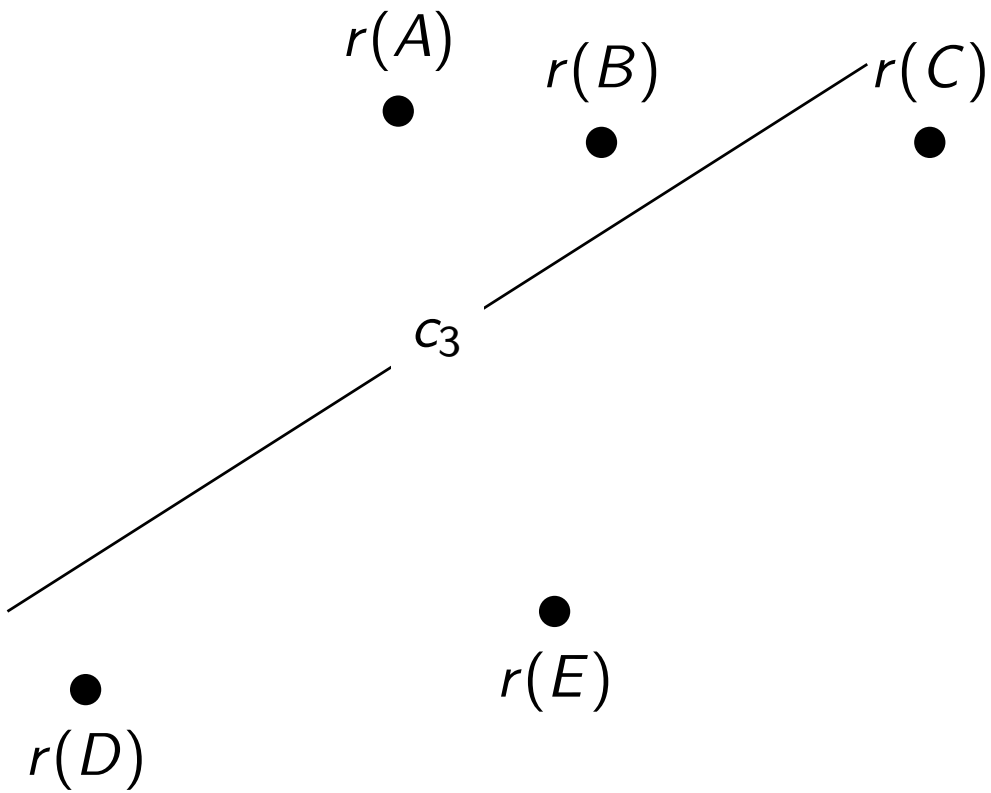
# Example

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $c_1$ | $+$ | $-$ | $-$ | $+$ | $-$ |
| $c_2$ | $+$ | $+$ | $+$ | $-$ | $-$ |
| $c_3$ | $+$ | $+$ | $-$ | $-$ | $-$ |



# Example

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $c_1$ | $+$ | $-$ | $-$ | $+$ | $-$ |
| $c_2$ | $+$ | $+$ | $+$ | $-$ | $-$ |
| $c_3$ | $+$ | $+$ | $-$ | $-$ | $-$ |



# Dimension complexity

$C \subseteq \{\pm 1\}^X$ – a concept class

$r$ – a $C$-separating map to $\mathbb{R}^d$
the *dimension* of $r$ is $d$

**Definition:** The dimension complexity of $C$ is the minimum dimension of a separating map for it.

# Low dimension complexity implies succesful learning

$C \subseteq \{\pm 1\}^X$ – a concept class with dimension complexity $d$

$r$ – a $C$-separating map to $\mathbb{R}^d$

$L$ – a learning algorithm that applies $r$ and then uses SVM (e.g. kernel machines)

The sample complexity of $L$ is $O(d)$

# Margin

$C \subseteq \{\pm 1\}^X$ – a concept class

$r$ – a $C$-separating map to $\mathcal{B}^d \subseteq \mathbb{R}^d$
the margin of $r$ is the minimum distance between $\text{conv}(c^{-1}(+1))$ and $\text{conv}(c^{-1}(1))$ over all $c \in C$

**Definition:** The margin complexity of $C$ is the maximum margin of a separating map for it.

# Large margin complexity implies succesfull learning

$C \subseteq \{\pm 1\}^X$ – a concept class with margin complexity $\gamma$

$r$ – a $C$-separating map with margin $\gamma$

$L$ – a learning algorithm that applies $r$ and then uses SVM (e.g. kernel machines)

The sample complexity of $L$ is $O\left(\frac{1}{\gamma^2}\right)$

# Large margin complexity implies low dimension complexity

If there exists a $C$-separating map with large margin then there exists a $C$-separating map with low dimension

- apply a random projection

# Large margin complexity implies low dimension complexity

If there exists a $C$-separating map with large margin then there exists a $C$-separating map with low dimension

- apply a random projection

**"Corollary":** If $C$ is efficiently learned by a kernel machine then its dimension complexity is low.

# Sign rank and dimension complexity are equivalent

$C \subseteq \{\pm 1\}^X$ – a concept class with dimension complexity $d$

$M$ – a matrix whose rows are the concepts of $C$

$$d \leq \text{signrank}(M) \leq d + 1$$

# Recapitulation

1. Sign rank captures unbounded error communication complexity

2. Good performance of kernel machines on $C$ implies it has a low sign rank

# Plan

# "In theory": most learnable classes have large sign rank

**Theorem** [BES '02, AMY '15]
For any fixed $d \geq 2$, most concept classes $C \subseteq \{\pm 1\}^N$ of VC dimension $d$ have sign rank $N^{\Omega(1)}$.

Thus, a random concept class with a constant sample complexity can not be learned by first embedding the data to a point set with
(i) a constant dimension, or
(ii) a constant margin.

# "In practice": many learning tasks are performed by kernel machines

Many practical learning problems are efficiently learned by kernel machines.

listed among the top classifiers to try first [e.g. by stackexchange.com]

handwriting recognition, image classification, medical science, bioinformatics, and more...

# Recapitulation

1. Sign rank captures unbounded error communication complexity

2. Good performance of kernel machines on $C$ implies it has a low sign rank

3. "A paradox":
   - In practice kernel machines perform many learning tasks
   - Most learnable classes have a large sign rank

# Plan

# Resolution 1: A sublinear upper bound on the sign rank of learnable classes

**Theorem**[Alon,M,Yehudayoff]
For any fixed $d$, every class $C \subseteq \{\pm 1\}^N$ with VC dimension $d$ has sign rank $o(N)$.

- (almost) matches the lower bound

# Resolution 1: A sublinear upper bound on the sign rank of learnable classes

**Theorem**[Alon,M,Yehudayoff]
For any fixed $d$, every class $C \subseteq \{\pm 1\}^N$ with VC dimension $d$ has sign rank $o(N)$.

- (almost) matches the lower bound

How can this be used to bridge the gap?

# Resolution 1: A sublinear upper bound on the sign rank of learnable classes

**Theorem**[Alon,M,Yehudayoff]
For any fixed $d$, every class $C \subseteq \{\pm 1\}^N$ with VC dimension $d$ has sign rank $o(N)$.

# Resolution 1: A sublinear upper bound on the sign rank of learnable classes

**Theorem**[Alon,M,Yehudayoff]
For any fixed $d$, every class $C \subseteq \{\pm 1\}^N$ with VC dimension $d$ has sign rank $o(N)$.

**Definition** A class $C \subseteq \{\pm 1\}^X$ is weakly separable if for every $\{x_1, \ldots, x_m\} \subseteq X$, $C|_{\{x_1,\ldots,x_m\}}$ has sign rank at most $k = o(m)$.

# Resolution 1: A sublinear upper bound on the sign rank of learnable classes

**Theorem**[Alon,M,Yehudayoff]
For any fixed $d$, every class $C \subseteq \{\pm 1\}^N$ with VC dimension $d$ has sign rank $o(N)$.

**Definition** A class $C \subseteq \{\pm 1\}^X$ is weakly separable if for every $\{x_1, \ldots, x_m\} \subseteq X$, $C|_{\{x_1,\ldots,x_m\}}$ has sign rank at most $k = o(m)$.

**Corollary**
Every learnable class is weakly separable

# Weak separability is useful for learning

**Definition.** A class $C \subseteq \{\pm 1\}^X$ is weakly separable if for every $\{x_1, \ldots, x_m\} \subseteq X$, $C|_{\{x_1,\ldots,x_m\}}$ has sign rank at most $k = o(m)$.

A recipe for learning weakly separable classes:

$(x_1, y_1), \ldots, (x_m, y_m)$ - input sample

1. Embed $\{x_1, \ldots, x_m\}$ in $\mathbb{R}^k$.

2. Output $c \in C$ that agrees withimum margin separating hyperplane.

# Weak separability is useful for learning

**Definition.** A class $C \subseteq \{\pm 1\}^X$ is weakly separable if for every $\{x_1, \ldots, x_m\} \subseteq X$, $C|_{\{x_1,\ldots,x_m\}}$ has sign rank at most $k = o(m)$.

A recipe for learning weakly separable classes:

$(x_1, y_1), \ldots, (x_m, y_m)$ - input sample

1. Embed $\{x_1, \ldots, x_m\}$ in $\mathbb{R}^k$.

2. Output $c \in C$ that agrees withimum margin separating hyperplane.

**A generalization bound.** As $m$ grows, The error decays like $\epsilon \log \frac{1}{\epsilon}$, where $\epsilon = k/m = o(1)$.

# Recapitulation

1. Sign rank captures unbounded error communication complexity

2. Good performance of kernel machines on $C$ implies it has a low sign rank

3. "A paradox":
   - In practice kernel machines perform many learning tasks
   - Most learnable classes have a large sign rank

4. "Resolution" 1: Every learnable class has a sublinear sign rank

# Plan

# Resolution 2: practical learning problems have structure

"A paradox":

- In practice, kernel machines perform many learning tasks
- Most learnable classes have a large sign rank

Perhaps concept classes that appear in practical applications typically have a low sign rank.

# Resolution 2: practical learning problems have structure

"A paradox":

- In practice, kernel machines perform many learning tasks
- Most learnable classes have a large sign rank

Perhaps concept classes that appear in practical applications typically have a low sign rank.

**Goal.** Study the structure of concept classes/matrices with low sign rank.

# Plan

# How do standard operations affect the sign rank?

A basic type of "structural" questions concerns variability under standard operations.

# How do standard operations affect the sign rank?

A basic type of "structural" questions concerns variability under standard operations.

$C_1, C_2$ – two concept classes of sign rank at most $r$.

Consider a class obtained by some natural operation on $C_1, C_2$:

1. $\{\neg c_1 : c_1 \in C_1\}$
2. $\{c_1 \oplus c_2 : c_1 \in C_1, c_2 \in C_2\}$
3. $\{c_1 \wedge c_2 : c_1 \in C_1, c_2 \in C_2\}$

Is the sign rank of these classes bounded in terms of $r$?

# How do standard operations affect the sign rank?

A basic type of "structural" questions concerns variability under standard operations.

$C_1, C_2$ – two concept classes of sign rank at most $r$.

Consider a class obtained by some natural operation on $C_1, C_2$:

1. $\{\neg c_1 : c_1 \in C_1\}$
2. $\{c_1 \oplus c_2 : c_1 \in C_1, c_2 \in C_2\}$
3. $\{c_1 \wedge c_2 : c_1 \in C_1, c_2 \in C_2\}$

Is the sign rank of these classes bounded in terms of $r$?

1. the sign rank is at most $r$
2. the sign rank is at most $r^2$ [Derzinsky and Warmuth]
3. ???

# A question

$C_1, C_2$ – two concept classes of sign rank at most $r$.

**Question.**
Is the sign rank of $\{c_1 \wedge c_2 : c_1 \in C_1, c_2 \in C_2\}$ bounded in terms of $r$?

# Interpretation in machine learning

$C_1, C_2$ – two concept classes of sign rank at most $r$.

**Question.**
Is the sign rank of $\{c_1 \wedge c_2 : c_1 \in C_1, c_2 \in C_2\}$ bounded in terms of $r$?

Given efficient kernel machines for $C_1$ and $C_2$, can we construct an efficient kernel machine for $\{c_1 \wedge c_2 : c_1 \in C_1, c_2 \in C_2\}$?

# Interpretation in communication complexity

$f_1, f_2$ – two function with unbounded complexity at most $c$.

define $f_1 \wedge f_2$ as follows:
Alice's input is $x_1, x_2$
Bob's input is $y_1, y_2$
Their goal is to compute $f_1(x_1, y_1) \wedge f_2(x_2, y_2)$

**Question.**
Is the unbounded complexity of $f_1 \wedge f_2$ bounded in terms of $r$?

# Interpretation in communication complexity

$f_1, f_2$ – two function with unbounded complexity at most $c$.

define $f_1 \wedge f_2$ as follows:
Alice's input is $x_1, x_2$
Bob's input is $y_1, y_2$
Their goal is to compute $f_1(x_1, y_1) \wedge f_2(x_2, y_2)$

**Question.**
Is the unbounded complexity of $f_1 \wedge f_2$ bounded in terms of $r$?

*Are repetitions necessary for computing two decision problems in a randomized fashion?*

# Plan

# Recapitulation

1. Sign rank captures unbounded error communication complexity

2. Good performance of kernel machines on $C$ implies it has a low sign rank

3. "A paradox":
   - In practice kernel machines perform many learning tasks
   - Most learnable classes have a large sign rank

4. "Resolution" 1: Every learnable class has a sublinear sign rank

5. "Resolution" 2: Practical classes have low sign rank

6. Goal: Study the structure of classes/matrices with low sign rank
   - How do the sign-rank changes under standard operations?
   - Interpretation in machine learning
   - Interpretation in communication complexity