# On Teaching and Learning Intersection-Closed Concept Classes

Farnam Mansouri,
Advisor: Dr. Adish Singla

Sharif University of Technology

*fmansouri@ce.sharif.edu*

November 22, 2019

# Overview

1. Basic Definitions

2. VC-Dimension and Self-Directed Learning
   - Preliminary Results
   - Self-Directed learning of Intersection Closed classes and VC-Dimension
   - The structure of intersection-closed concept classes with high learning complexity

3. Teacher-directed learning
   - Average case
   - Best case

# Basic Definitions

## Definition (intersections of $z$ with concept class $C$)

$\Pi_c(z) = \{c \cap z : c \in C\}$

## Definition (VC dimension (VCD))

v is shattered by c if $\Pi_c(z) = 2^z$

$\Pi_c(d) = max_{|z|=d}\Pi_c(z)$

$VD(C) = maxd : \Pi_c(d) = 2^d$

**Saur's lemma:** If VCD(C) = d, and n is number of instances, we'll define $\Phi_d(n) = \sum_{i=0}^{d} \begin{pmatrix} n \\ i \end{pmatrix}$, then $|C| \leq \Phi_d(n)$

# Basic Definitions (cont.)

## Definition (labeled sequence)

$z|f = \{(x, f(x)) : x \in z\}$

## Definition (version space ($C_{z|f}$))

$c$ is consistent with $z|f$ if $z|f = z|c$

$C_{z|f} = \{c \in C : \text{c is consistent with } z|f\}$

## Definition (most specific and most general)

concept with minimum (maximum, resp.) size of C: S(C) (G(C), resp.)

c is most specific (most general, resp.) concept with respect to $z|f$ if $c \in S(C_{z|f})$ ($c \in G(C_{z|f})$, resp.)

# Basic Definitions (cont.)

## Definition (incidence matrix of C)

$a^i_j = 1$ iff $x_i \in c_j$

## Definition (index number of x and C.)

$Ind(C, x) = |\{c \in C : x \in c\}|$
$Ind(C) = min_{x \in X} Ind(C, x)$

# Self-Directed model

- Learner selects an instance and predicts the label of it.
- number of wrong predictions according to target concept is number of queries.
- Self-Directed learning complexity is defined as $M_{sd}(C) = min_L max_{c_t \in C}(M_{sd}(L, c_t))$, which $M_{sd}(L, c_t)$ is number of wrong predictions algorithm L made.

# Teacher-Directed model

- Teaching-Directed complexity for C with respect to $c_t$ is $M_{td}(C, c_t)$, which is defined as $min_{|z|}|C_{z|t}| = 1$
- We call z minimum sequence for $c_t$
- Teaching complexity of C:
  - $M_{td-worst}(c) = max_{c_t \in C}(M_{td}(C, c_t))$
  - $M_{td-best}(c) = min_{c_t \in C}(M_{td}(C, c_t))$
  - $M_{td-average}(c) = \sum_{c \in C} P(c) M_{td}(C, c_t)$ (with respect to distribution P)

# Preliminary Results

- $D_d = (\lambda_j^i)_{j=0\ldots d}^{i=1\ldots d}$ is a $d * (d+1)$ matrix which $\lambda_j^i = 0$ iff $i = j$.
- **Lemma1:** $c \in IC$ then, C contains a sub matrix $D_d$ iff VCD(C) is at-least d.

  **Outline of the Proof:** $\forall f : c_\cap = \cap_{x_i : f(x_i)=0} c_i \cap c_0$ is in C, so $x_1, ..., x_d | f = x1, ..., x_d | c_\cap$.

  **Example:** By lemma1 we know VCD of 1 is 2, since intersection of $\{x_0, x_1\}$ and $\{c_0, c_1, c_2\}$ is $D_2$, and $C \in IC$, also we can see that $\{x_0, x_1\}$ is shattered.

# Preliminary Results(cont.)

| **Instances** | $c_0$ | $c_1$ | $c_2$ | $c_2$ |
|---|---|---|---|---|
| $x_0$ | 0 | 1 | 1 | 0 |
| $x_1$ | 1 | 0 | 1 | 0 |
| $x_2$ | 1 | 1 | 0 | 0 |

Table: Example of Matrix with VCD 2

# Preliminary Results(cont.)

- **Lemma2:** $VCD(C) = 1$ then, $Ind(C) \leq 1$ or $Ind(\bar{C}) \leq 1$.
  **Proof:**
  **step1**: flipping a row of matrix doesn't affect the VC-Dimension, since if some rows are shattered, by flipping one of them they will still be shattered, and by if they are not shattered they will not be shattered by flipping one of them.
  **step2**: so if we XOR all columns with a specific sequence, it wouldn't affect the VC-dimension, i.e., if we flip some of the rows this will not affect the VC-dimension, so we'll XOR all columns with concept $c_0$, then $c_0$ will be 0.
  **step3**: we call the new matrix $\tilde{C}$. Assume $Ind(\tilde{C}) \geq 2$, we'll prove by induction on k that for any k we can find $x_1, \tilde{c}_1, ..., x_k, \tilde{c}_k$ so that, $\forall j > i : x_1, ..., x_i \in \tilde{c}_i, x_j \notin \tilde{c}_i$. proof: Induction this is true for $k \leq n$ then, $x_n$ is only in $c_n$, so there must
  exist a $c_{n+1}$, which $x_n \in c_{n+1}$, since $VCD(C) = 1$ and $c_0 = 0$,
  $\forall i < n : x_i \in \tilde{c}_{n+1}$, so on $x_1, ..., x_n$, $\tilde{c}_n$, and $\tilde{c}_{n+1}$ are the same so there

# Preliminary Results(cont.)

**step3 (cont.)**: must exist a $x_{n+1}$ which is in one of them and isn't in other one, without loss of generality assume $x_{n+1} \in \tilde{c}_n, x_{n+1} \notin \tilde{c}_{n+1}$, since $VCD(C) = 1$ and $c_0 = 0$, $\forall i < n : x_{n+1} \in \tilde{c}_i$. now since number of instances is finite, this is not possible, so $Ind(\tilde{C}) \leq 1$
**step4**: $\exists x : Ind(\tilde{C}, x) \leq 1$, if $x \notin c_0$ we'll drive that $Ind(C, x) \leq 1$, and if $x \in c_0$ we'll drive that $Ind(\bar{C}, x) \leq 1$, so, $Ind(C) \leq 1$ or $Ind(\bar{C}) \leq 1$.

# Preliminary Results(cont.)

- **Lemma3:** $c \in IC$ and, $VCD(C) = 1$, then $Ind(C) \leq 1$.

  **Outline of the Proof:** if $Ind(C) \geq 2, C \in IC$ then, then the step 3 can be droved without need of a zero row.

## Theorem (Theorem 1)

$VCD(C) = 1$, *then* $M_{sd}(C) = 1$

**Proof:** *if* $Ind(C) = 0$ *or* $Ind(\bar{C}) = 0$, *then there is an instance which predicts all instances the same, we can put that away since the VCD will never rise by this action and it can never get zero because that would mean all concepts are the same, at some point* $Ind(C) = 1$ *or* $Ind(\bar{C}) = 1$. *use lemma2 and each time choose the instances which is in one concept or all except one, and we'll give a prediction of majority of labels, any time that we'll mistake the target concept will be found.*

# Self-Directed learning of Intersection Closed classes and VC-Dimension

When VC-dimension isn't 1, we can't bound the self-directed learning complexity.

**Theorem (Theorem 2)**

$\forall m \geq 1, d \geq 2, \exists C_m^d \in IC, VCD(C_m^d) = d, M_{sd}(C_m^d) \geq m$

**Outline of the Proof:** we create E[Q] from $n * r$ incidence matrix Q (see Fig 16), where $A_{nr}^i$ is a $n * r$ matrix which row i is 1, we'll call row i, and column j of $E_1(Q)$ resp., $I^i$ and $I_j$. Then row j of $I^i$ is $I^{i,j}$, and the intersection of $I^i$ and $I_j$ is $I_j^i$, and row k of $I_j^i$ is $I_j^{i,k}$.

# Self-Directed learning of Intersection Closed classes and VC-Dimension (cont.)

Claims:

- Claim1: $Q_{nr} \in IC \Longrightarrow E(Q_{nr}) \in IC$
- Claim2: $Q_{nr} \in IC, VCD(Q_{nr}) = d \geq 2 \Longrightarrow VCD(E(Q_{nr})) = d$
  **Outline of the Proofs:** we will prove it for all different cases.
- Claim3: For all labeled instances $(x, l) \in X \times \{0, 1\}, Q_{nr} \in E(Q_{nr})_{x|l}$
  **Proof:** imagine $x = I^{i,j}$, then if $c_t$ has x in it, since $I^{i}_{j+1}$ is $A^{j}_{nr}$, so $I^{i,j}_{j+1}$ is all one, then $I_{j+1}$ must be in version space, so $I^{j+1}_{j+1}$ is in version space too, so Q is in version space, else for all $k \neq j, j+1$, $I^{i,j}_{k}$ is all zero so $I_k$ and specially $I^{k}_{k}$ which is Q is in version space.

from Claim3 we'll conclude that adversary can make $E[Q_{nr}]$ make a mistake and still have $Q_{nr}$ in it, so if we repeatedly create $E(Q_{nr})$ from $Q_{nr}$ we can force $M_{sd}(C^{d}_{m})$ to rise 1 at each step.

# Self-Directed learning of Intersection Closed classes and VC-Dimension(cont.)

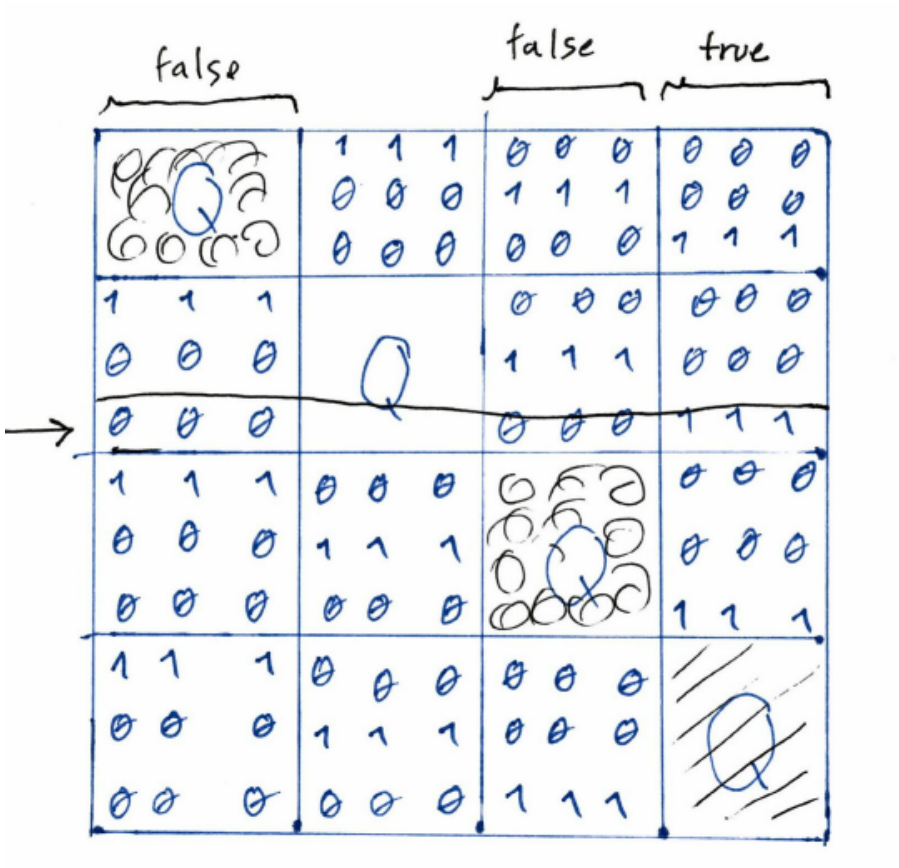


Figure: Illustration of Claim3

# Self-Directed learning of Intersection Closed classes and VC-Dimension(cont.)

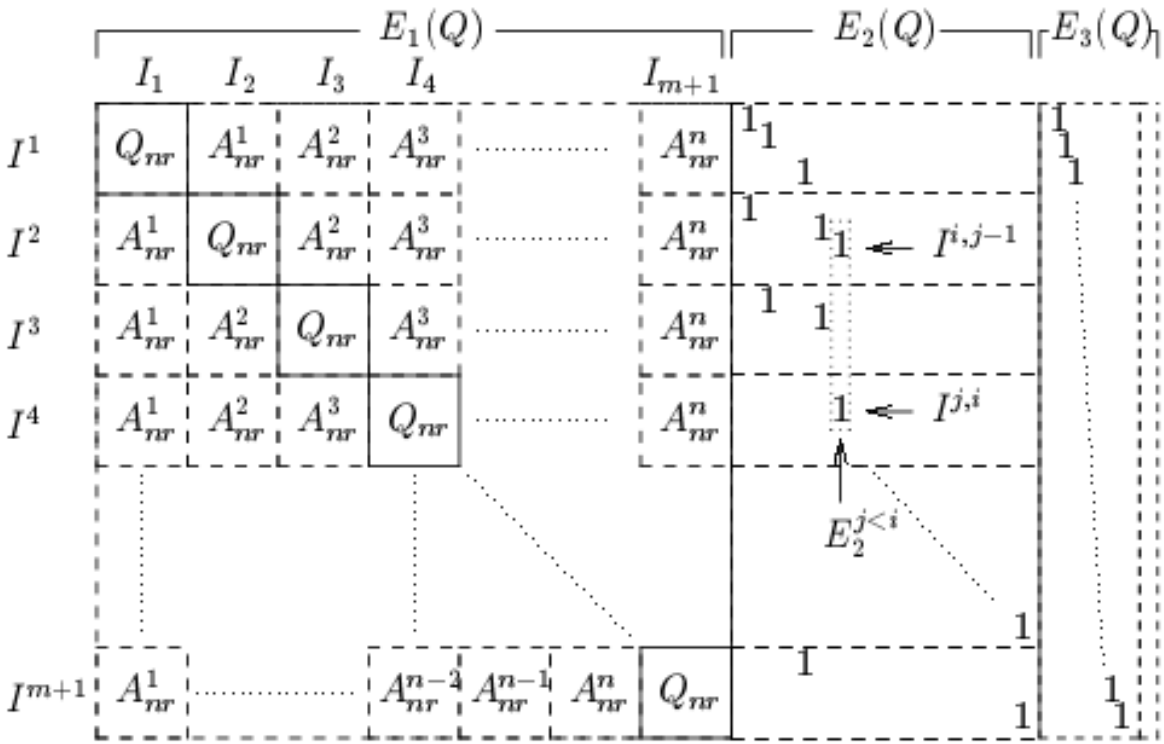


**Fig. 1.** The invariant extension of an intersection-closed incidence matrix $Q$

# The structure of intersection-closed concept classes with high learning complexity

## Theorem (Theorem 3)

*$C \in IC, VCD(C) = 2, M_{sd}(C) = k$. Then incidence matrix of C contains a sub-matrix $A_k$ (which has been shown in Fig. 2).[no proof]*

## Definition (maximal matrix)

A matrix which has equality in Saur's Lemma, i.e., $|C| = \Phi_d(n)$

**Open problems in this area:** can we extend this structure we found for higher VC-Dimensions.

# The structure of intersection-closed concept classes with high learning complexity(cont.)

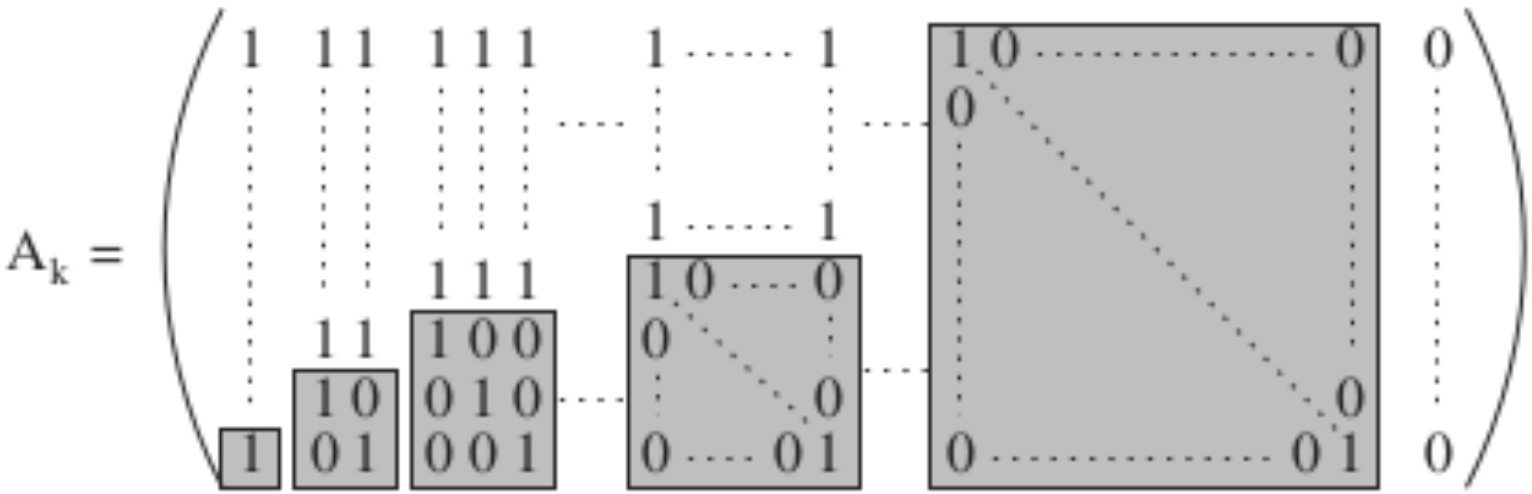


**Fig. 2.** The maximum matrix $A_k$ which is included as a submatrix in the incidence matrix

# Teacher-directed learning - average case

## Definition (k ball with center $z|f_0$)

$B_z^k(f_0) = \{z|f : d_z(f_0, f) \leq k\}$ , and we call $B^k(c_0) = B_X^k(c_0)$ k ball with center $c_0$

## Definition

C contains a k-ball with center $z|c_0$, iff $B_z^k(c_0)$ is a subset of $C|z$.

Since the VC-dimension of a concept class with k-ball is k, a concept class containing a k-ball has VC-dimension of at-least k.

# Teacher-directed learning - average case

**lemma4:** Let $c_t \in C$. For any minimum sequence z for $c_t$, C contains a 1-ball with center $z|c_t$. **Outline of the Proof:** $|C_{z|c_t}| = 1$, and $\forall i : |C_{z-x_i|c_t}| \geq 2$. Therefore, $\forall i : |C_{z-x_i|c_t;x_i|\bar{c}_i}| \geq 1$.

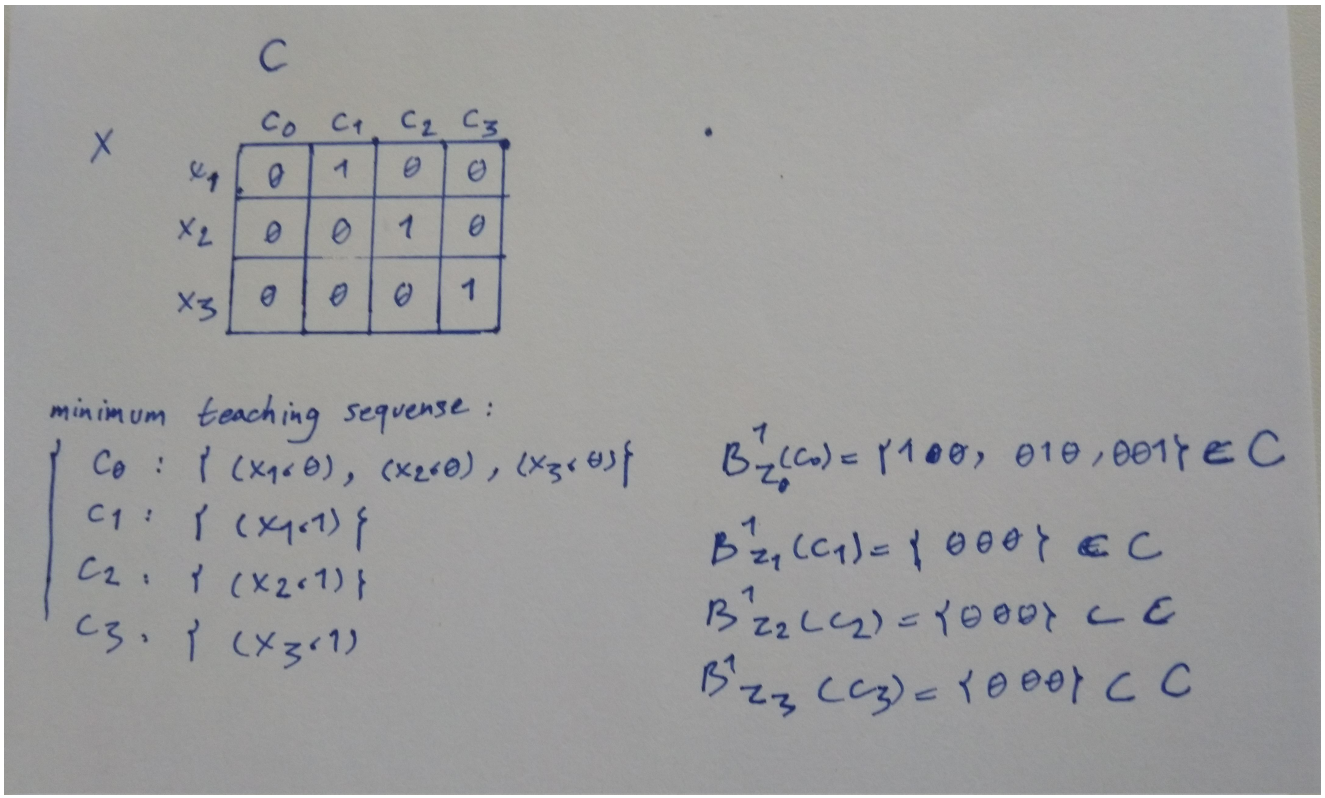


Figure: Example of lemma 4

# Teacher-directed learning - average case (cont.)

**Theorem (Theorem 4)**

*$VCD(C) = 1$, Then $\exists k_0, ..., k_m \in \mathbb{N}_0$ with $\sum_{i=0}^{m} k_i \leq 2m$ such that, $M_{td-average}(C, P) = \sum_{i=0}^{m} k_i P(c_i)$.*
*In particular for uniform distribution $U$, $M_{td-average}(C, U) < 2$.*

**Outline of the Proof:** It is easily proved that every concept class with VCD, 1 can be embedded to a maximum class, so we'll assume we have m instances. we'll prove the theorem by induction on instance space, when we restrict our incidence matrix to m-1 samples the VC-dimension cannot change and by Saur's lemma we know $|C| \leq m - 1 + 1$, so the number of concepts must decrease.

# Teacher-directed learning - average case (cont.)

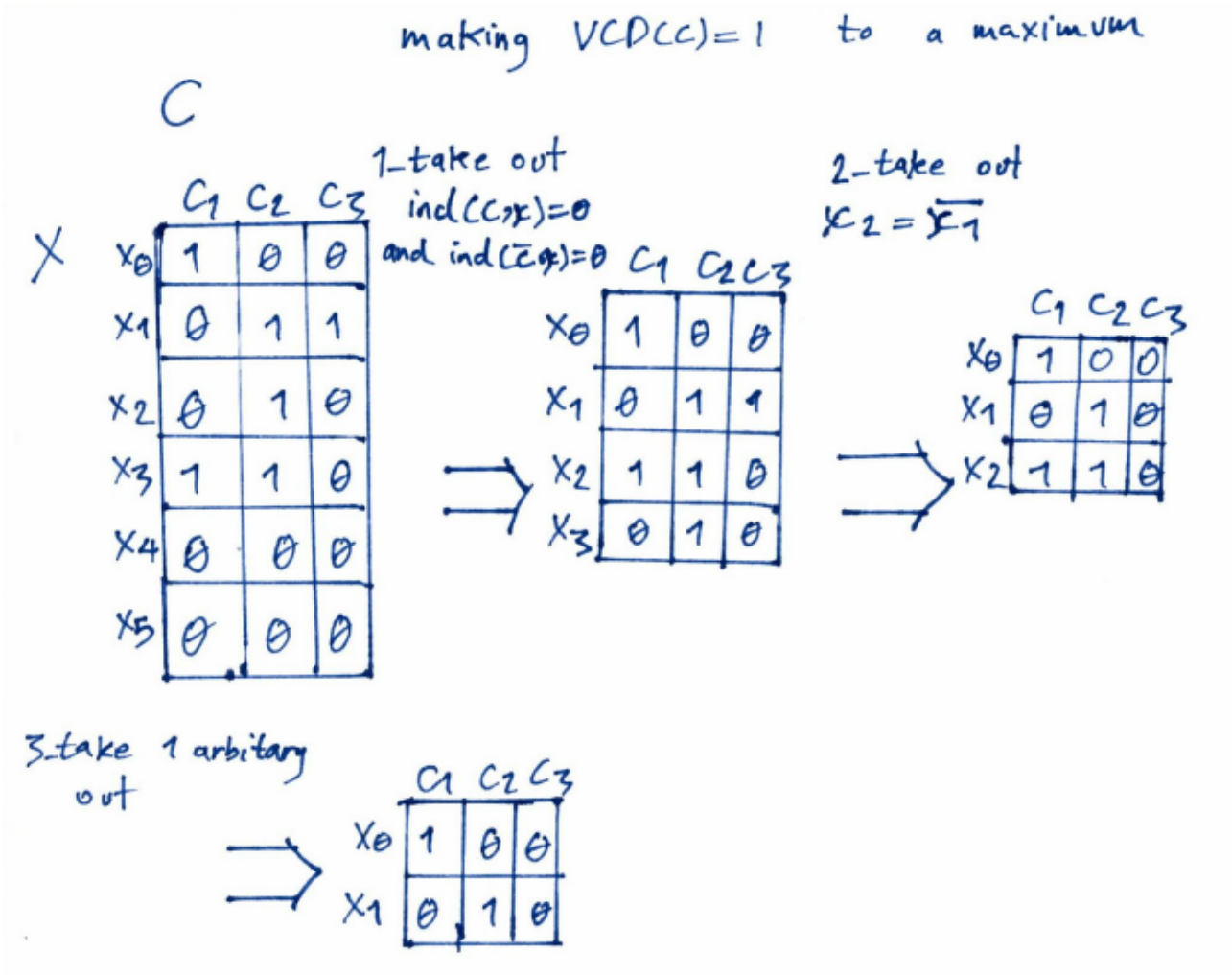


Figure: Procedure of making matrix with VCD 1, maximum

# Teacher-directed learning - average case (cont.)

## Definition (A pair of x)

$(c_1, c_2)_x$ are a pair of x, if $x \notin c_1$ and $c_1 \cup \{x\} = c_2$.

## Definition (Independant)

c is independent if it does not belong to any pair.

**Outline of the Proof(cont.):** Let $\tilde{k}_c = M_{td}(C, c)$,
$k_c = M_{td}(C|(X - x), c)$, by definition $M_{td-average}(C, P) = \sum_{c \in C} k_c P(c)$.

- **Claim1:** There exist at most one pair of x in C (straight forward).
  **Outline of the proof:** if two pairs of x exists in C that would mean $VCD(C) \geq 2$.

# Teacher-directed learning - average case (cont.)

- **Claim2:** Let $(c_1, c_2)_z$ be a pair in C. Then, up to a permutation, $k_{c_1} = \tilde{k}_{c_1} + 1$ and $k_{c_2} = 1$.
  **Proof:** in x, $c_2$ and $c_1$ are different and we already assuming that $Ind(C, x) = 1$, or $Ind(\bar{(C)}, x) = 1$, so all others concept are the same at x, so we can decide one of them, with choosing x, and teaching dimension for the other one will increase at most once.
- **Claim3:** For all independent concepts c, $k_c = \tilde{k}_c$.
  **Proof:** again $Ind(C, x) = 1$, or $Ind(\bar{(C)}, x) = 1$, and since c is independent it's easy to see that x has different label only in one other concept, we'll name it $c_t$, since c is independent, c and $c_t$ must have more than one difference, so we can take out $c_t$ with other x too, and x isn't useful for c.

With claim 1, 2 and 3 and by induction we'll conclude that $\sum_{c \in C} k_c \leq \sum_{c \in C} \tilde{k}_c + 2 \leq 2m$.

# Teacher-directed learning - average case (cont.)

**Lemma5:** For $B^k(c)$, $VCD, M_{td}, M_{td-average}$ and $M_{td-best}$ are independent from c.

**Outline of Proof:** for every $c_1$, $c_2$, we can flip instances that $c_1(x) \neq c_2(x)$ this will not affect the 4 measures mentioned above, and we'll transform $c_1$ to $c_2$.

# Teacher-directed learning - average case (cont.)

## Theorem (Theorem 5)

$M_{td-average}(B^d(c), U) \leq 2d$

**Outline of Proof:** *with help of previous lemma we'll assume* $c = \varnothing$*, then we can partition* $B^d(c)$ *to* $C_0, ..., C_d$*, which* $C_i = c : |c| = i$*. clearly* $\forall i < d, c_i \in C_i : M_{td}(C, c_i) \leq n$*, but* $\forall c_d \in C_d : M_{td}(C - d) \leq d$*, since* $C_d$ *are the largest sequences and can be determined by the ones,* $B^d(c)$ *is a maximum class*

$$M_{td-average}(B^d(c), U) \leq \frac{n|\cup_{i<d} C_i| + d|C_d|}{\Phi_d(n)} = \frac{\Phi_{d-1}(n) * n + d \begin{pmatrix} n \\ d \end{pmatrix}}{\Phi_d(n)} \leq$$

$$d + (n-d)\frac{\Phi_{d-1}(n)}{\Phi_d(n)} \leq 2d$$

**Open problems in this area:** we found that, $M_{td-average}(B^d(c), U) \leq 2d$ but can we generalize this to every intersection closed concept class with VC-dimension d?

# Teacher-directed learning - best case

## Definition (c admits unique most specific concept)

$C \in MSC$ iff $\forall z|f, |S(C_{z|f})| \leq 1$

**Example:** 2 is in MSC, since size of all concepts are different, so $\forall z|f, |S(C_{z|f})| \leq 1$.

| **instances** | $c_0$ | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|---|
| $x_0$ | 1 | 0 | 1 | 0 |
| $x_1$ | 1 | 1 | 0 | 0 |
| $x_2$ | 1 | 1 | 0 | 0 |
| $x_3$ | 0 | 0 | 0 | 0 |

Table: Example of Matrix in MSC

# Teacher-directed learning - best case

## Corollary

*If all concepts have different size, then $C \in MSC$.*

## Corollary

*If $C \in IC$, then $C \in MSC$.*

**Proof:** *if two $c \in C$ are consistent with $z|f$, then their intersection is also consistent with $z|f$. So the smallest member in $C_{z|f}$ is unique.*

# Teacher-directed learning - best case (cont.)

## Definition (spanning sets of c)

For $C \in MSC$ we call $\mathcal{I}(c) = \{$ z with minimum size : $S(C_{z|1}) = \{c\}\}$, the spanning set of c, and we define $|\mathcal{I}(c)|$ as the size of instances of the spanning set, and we'll define $\mathcal{I}(C) = max_c|\mathcal{I}(c)|$

**Example:** It is clear that only for $z \in \{\{x_0, x_1, x_2\}, \{x_0, x_2\}, \{x_0, x_1\}\}$, $S(C_{z|1}) = \{c_0\}$, because only z with minimum size is in $\mathcal{I}(c_0)$, we'll conclude: $\mathcal{I}(c_0) = \{\{x_0, x_2\}, \{x_0, x_1\}\}$ (size of $\{x_0, x_1\}$ and $\{\{x_0, x_2\}$ is 2 so $|\mathcal{I}(c_0)| = 2$) , similarly we can drive $\mathcal{I}(c_3) = \varnothing$, $\mathcal{I}(c_2) = \{\{x_0\}\}$, $\mathcal{I}(c_1) = \{\{x_2\}, \{x_1\}\}$.
Finally, we can drive $\mathcal{I}(C) = 2$, note that $x_0, x_1$ can be shattered, so VCD $= 2$.

## Theorem

*Natarajan has proved that* $\forall C \in MSC, \mathcal{I}(C) \leq VCD(C)$.

# Teacher-directed learning - best case(cont.)

**lemma6:** $M_{td-best}(C) \leq M_{sd}(C)$
**Outline of the proof:** consider each of those instances which self-directed algorithm made mistake, those instances should specify the target concept.

**Theorem (Theorem 6)**

*$C \in MSC$, Then $M_{td-best}(C) \leq VDC(C)$.*

**Outline of the proof:** Let $c_t \in G(C)$, and $z_t \in \mathcal{I}(C_t)$ which $S(z_t|c_t) = \{c_t\}$. Since $|S(z_t|c_t)| = 1$, and $c_t$ is the largest concept of C, $C_{z_t|1} = \{c_t\}$. Hence, $M_{td-best}(C) \leq |z| \leq \mathcal{I}(C)$,
and since $\mathcal{I}(C) \leq VCD(C)$, $M_{td-best}(C) \leq VCD(C)$.

# Teacher-directed learning - best case (cont.)

**Theorem (Theorem 7)**

$\exists C^d, VCD(C^d) = 2d, M_{td-best} = 3d.$

**Outline of the proof:** VCD of Matrix shown in Fig. 3 is 2, and it's $M_{td-best}$ is 3, now $\tilde{C} = C * ... * C$ has the favored property.
**Open problems in this area:** can we find concept class with fix VC-Dimension but arbitrary large $M_{td-best}(C)$?

# Teacher-directed learning - best case (cont.)

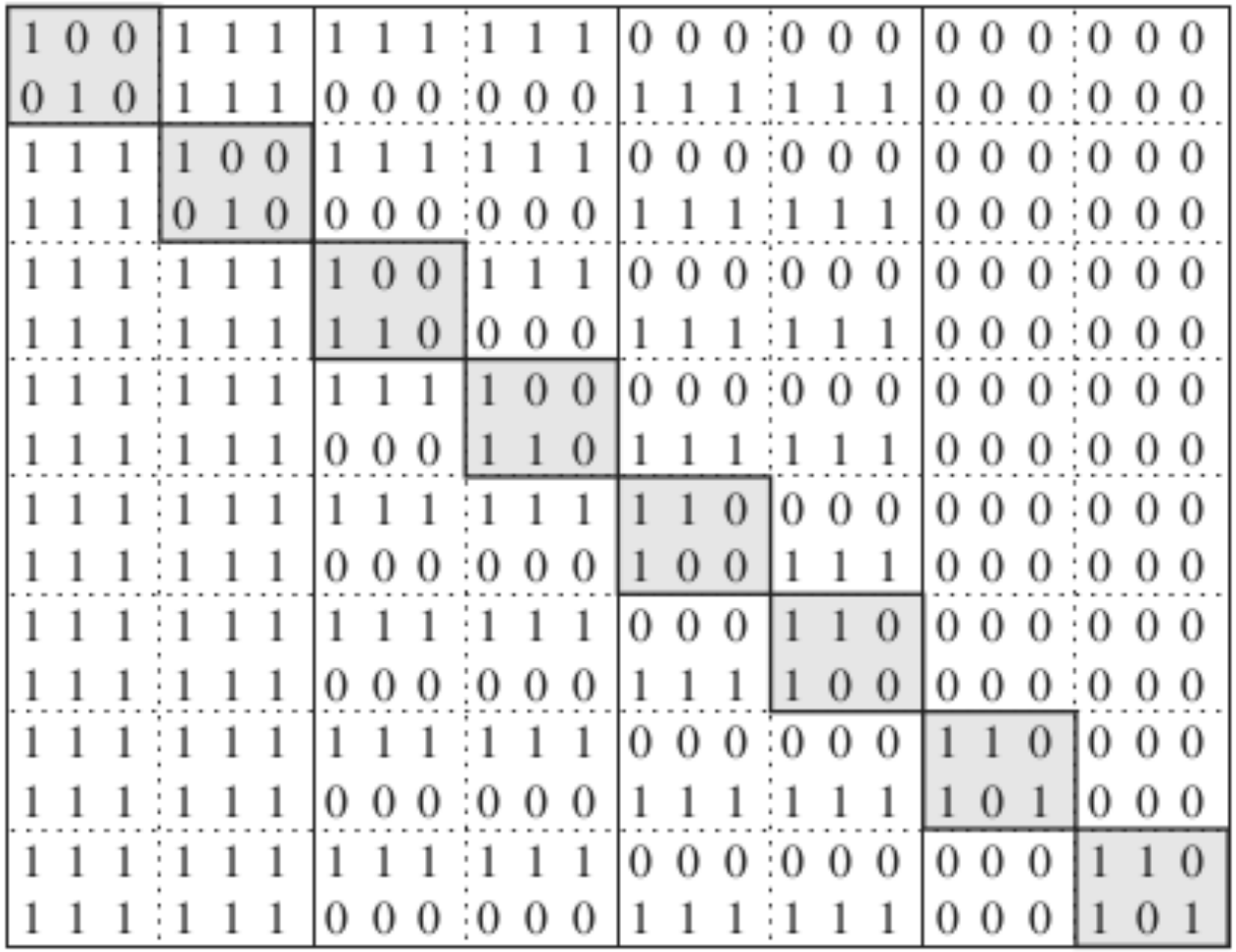

**Fig. 3.** The expansion of powerset $\mathcal{C}$ to a concept class $E(\mathcal{C})$ with $\mathrm{VCD}(E(\mathcal{C})) = 2$ and learning complexity $\mathrm{M}_{td\text{-}best}(E(\mathcal{C})) = 3$

# Open questions all in one slide.

- **Q1:** can we extend this structure we found for higher VC-Dimensions?
- **Q2:** we found that, $_{td-average}(B^d(c), U) \leq 2d$ but can we generalize this to every intersection closed concept class with VC-dimension d?
- **Q3:** can we find concept class with fix VC-Dimension but arbitrary large $M_{td-best}(C)$?

# The End