RUHR-UNIVERSITÄT BOCHUM

SUPERVISED LEARNING AND CO-TRAINING

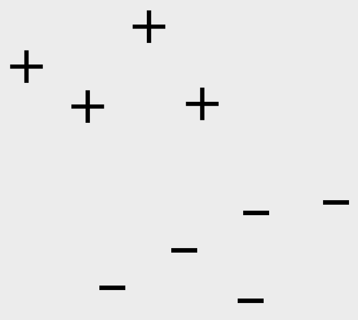
Malte Darnstädt, Hans Ulrich Simon
and Balázs Szörényi
Ruhr-University Bochum and University of Szeged

CONFERENCE ON ALGORITHMIC LEARNING THEORY 2011, ESPOO
October 7, 2011

Outline

1. Motivation and ‘What is Co-Training?’
2. The disagreement coefficient
3. Label complexity bounds
4. Combinatorial bounds
5. Final Remarks

Motivation
Supervised vs. Semi Supervised Learning

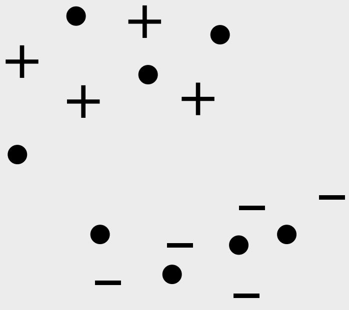


Supervised Learning:

- Concept class $\mathcal{C}$ with target $h^* \in \mathcal{C}$
- Distribution $\mathbb{P}$ over instance space X
- Labeled sample of size m
- Upper bound: $m = O\left(\frac{d \cdot \log 1/\varepsilon + \log 1/\delta}{\varepsilon}\right) = \widetilde{O}\left(\frac{d}{\varepsilon}\right)$
- Lower bound: $m = \Omega\left(\frac{d + \log 1/\delta}{\varepsilon}\right) = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$

Motivation

Supervised vs. Semi Supervised Learning



Semi Supervised Learning:

- Additional unlabeled sample
- Conjecture by Ben-David, Lu and Pál:
  Even for fixed $\mathbb{P}$, only saves a constant factor of labels
  - For special classes and distributions:
    Ben-David et al (2008)
  - For finite classes and 'most' distributions:
    Simon and D. (2011)

Co-Training with conditional independence
by Blum and Mitchell (1998)

- Data points have two ‘views’: $x = (x_1, x_2),\ X = X_1 \times X_2$
- Two target concepts: $h_1^* \in C_1,\ h_2^* \in C_2$
- Distribution $\mathbb{P}$ over $X_1 \times X_2$
    - Perfectly compatible: $h_1^*(x_1) = h_2^*(x_2)$ with probability 1
    - Conditional independence given the label:

$$\mathbb{P}\left(x_1, x_2 | + \right) = \mathbb{P}\left(x_1 | + \right) \cdot \mathbb{P}\left(x_2 | + \right)$$
$$\mathbb{P}\left(x_1, x_2 | - \right) = \mathbb{P}\left(x_1 | - \right) \cdot \mathbb{P}\left(x_2 | - \right)$$

Co-Training with conditional independence
by Blum and Mitchell (1998)

- Balcan and Blum (2010) give a Semi Supervised algorithm that learns with **just one labeled example**!
- Power of unlabeled data?

Co-Training with conditional independence
by Blum and Mitchell (1998)

- Balcan and Blum (2010) give a Semi Supervised algorithm that learns with **just one labeled example**!
- Power of unlabeled data?

$O(1)$ ⟷ $\tilde{\Omega}\left(\frac{d}{\varepsilon}\right)$

Semi Supervised
Co-Training

Supervised

Co-Training with conditional independence
by Blum and Mitchell (1998)

- Balcan and Blum (2010) give a Semi Supervised algorithm that learns with **just one labeled example**!
- Power of unlabeled data?

$$O(1) \longleftrightarrow ?$$

Semi Supervised
Co-Training

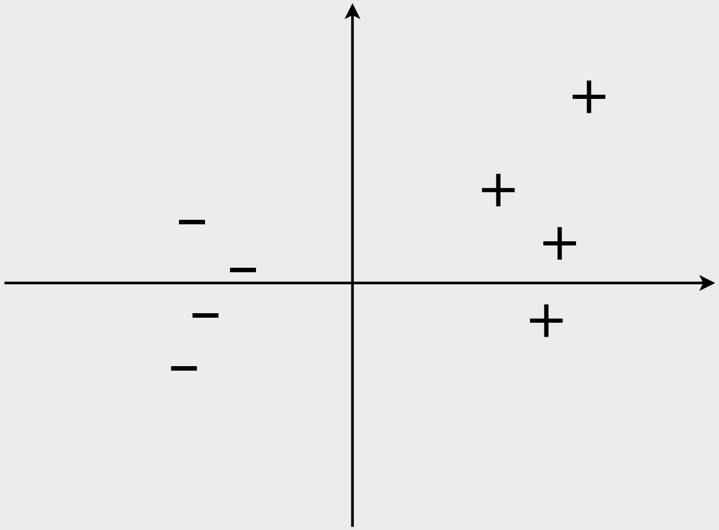
Supervised
Co-Training

Outline

1. Motivation and 'What is Co-Training?'
2. The disagreement coefficient
3. Label complexity bounds
4. Combinatorial bounds
5. Final Remarks

The disagreement region
definition

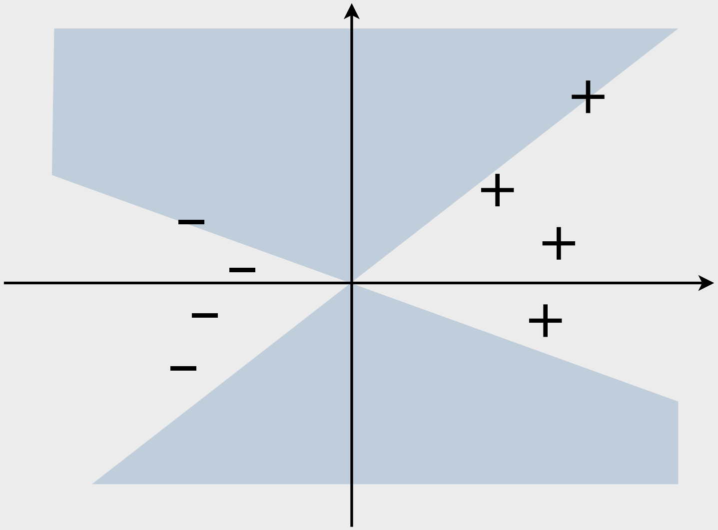
$$\mathrm{DIS}(V) := \{x \in X |\ \exists h, h' \in V : h(x) \neq h'(x)\}$$
for any subset $V \subseteq \mathcal{C}$ (usually a version space)



The disagreement region
definition

$$\mathrm{DIS}(V) := \{x \in X \mid \exists h, h' \in V : h(x) \neq h'(x)\}$$

for any subset $V \subseteq \mathcal{C}$ (usually a version space)



The disagreement coefficient
definition

$$\theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\mathsf{sample}) := \frac{\mathbb{P}\big(\mathrm{DIS}(V_{\mathcal{C}})\big)}{\sup_{h\in V_{\mathcal{H}}} \mathbb{P}\big(h(x)\neq h^*(x)\big)}$$

$$\theta(\mathcal{C},\mathcal{H}) := \sup_{\mathbb{P},h^*,\mathrm{sample}} \theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\mathrm{sample})$$

The disagreement coefficient
definition

size of disagreement region
after seeing the sample

$$\theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\mathsf{sample}) := \frac{\mathbb{P}\big(\mathrm{DIS}(V_{\mathcal{C}})\big)}{\sup_{h\in V_{\mathcal{H}}} \mathbb{P}\big(h(x)\neq h^*(x)\big)}$$

largest error

$$\theta(\mathcal{C},\mathcal{H}) := \sup_{\mathbb{P},h^*,\mathrm{sample}} \theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\mathrm{sample})$$

The disagreement coefficient
definition

size of disagreement region after seeing the sample

$$\theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\text{sample}) := \frac{\mathbb{P}\big(\mathrm{DIS}(V_{\mathcal{C}})\big)}{\sup_{h\in V_{\mathcal{H}}} \mathbb{P}\big(h(x)\neq h^*(x)\big)}$$

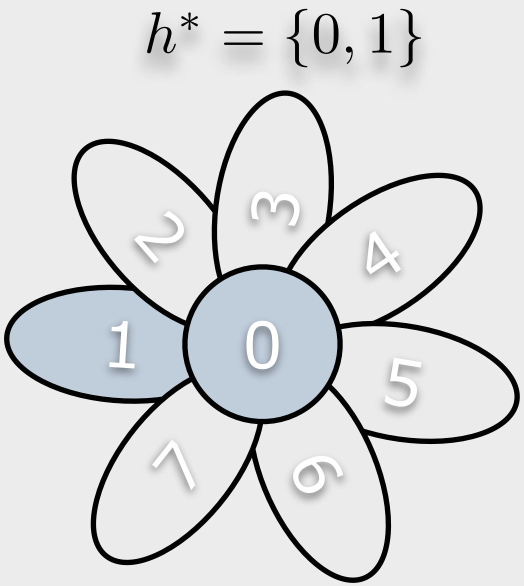
largest error

$$\theta(\mathcal{C},\mathcal{H}) := \sup_{\mathbb{P},h^*,\text{sample}} \theta(\mathcal{C},\mathcal{H}|\mathbb{P},h^*,\text{sample})$$

- A variant of Hanneke's disagreement coefficient (2007) for the realizable case
- Note: error according to hypothesis class $\mathcal{H}$, but the disagreement region is from $\mathcal{C}$

The disagreement coefficient
example

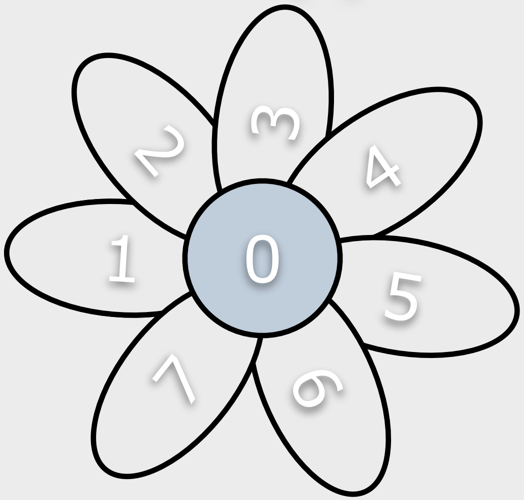
- Simple class that is useful for proving lower bounds
- $\mathrm{SF}_n := \{\{0\},\{0,1\},\{0,2\},\dots,\{0,n\}\}$



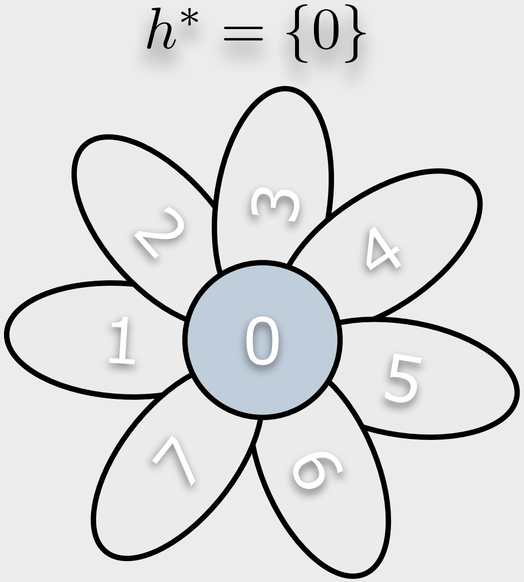
The disagreement coefficient
example

- Simple class that is useful for proving lower bounds
- $\mathrm{SF}_n := \{\{0\},\{0,1\},\{0,2\},\dots,\{0,n\}\}$

$h^* = \{0\}$



The disagreement coefficient
example



- Simple class that is useful for proving lower bounds
- $\mathrm{SF}_n := \{\{0\},\{0,1\},\{0,2\},\ldots,\{0,n\}\}$
- $\theta(\mathrm{SF}_n,\mathrm{SF}_n) = n$
  - Generally: $\theta(\mathcal{C},\mathcal{C}) \leq |\mathcal{C}| - 1$
  - Also $\theta(\mathrm{SF}_n,\mathrm{SF}_n) \geq n$:

    Choose $\mathbb{P}$ as uniform on $\{1,\ldots,n\}$ and $\mathrm{sample} = h^* = \{0\}$
    $\Rightarrow \theta(\mathrm{SF}_n,\mathrm{SF}_n|\ \mathbb{P}, h^*, \mathrm{sample}) = \frac{1}{1/n} = n.$

The disagreement coefficient
application in learning theory

Lemma:
For a sample size of

$$m = \tilde{O}\left(\frac{\theta(\mathcal{C}, \mathcal{H}) \cdot \mathrm{VCdim}(\mathcal{H})}{\varepsilon}\right)$$

it holds with high probability that

$$\mathbb{P}\big(DIS(V_{\mathcal{C}})\big) \leq \varepsilon$$

- Make $\mathcal{H}$ more powerful $\Rightarrow$ θ decreases, but VCdim increases
- Proof: classic PAC bound for class $\mathcal{H}$

Outline

1. Motivation and ‘What is Co-Training?’
2. The disagreement coefficient
3. Label complexity bounds
4. Combinatorial bounds
5. Final Remarks

Back to Co-Training
some notation

- Concept classes $\mathcal{C}_1, \mathcal{C}_2$, hypotheses classes $\mathcal{H}_1, \mathcal{H}_2$
- $d_i := \mathrm{VCdim}(\mathcal{H}_i)$
- $\theta_i := \theta(\mathcal{C}_i, \mathcal{H}_i)$
- $p_+ := \mathbb{P}\big(h^*(x) = \text{``+''}\big)$
- $p_- := \mathbb{P}\big(h^*(x) = \text{``}-\text{''}\big) = 1 - p_+$
- $p_{min} := \min\{p_+, p_-\}$, $d_{max} := \max\{d_1, d_2\}$, ...

Three resolution rules

with upper bounds

- After seeing a labeled sample, the learner has to label a new instance (x_1, x_2):
  - Safe decision, if $x_1 \notin \mathrm{DIS}_1$ or $x_2 \notin \mathrm{DIS}_2$
  - How should we label an instance in $\mathrm{DIS}_1 \times \mathrm{DIS}_2$?

Three resolution rules
with upper bounds

- First fix some consistent $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$

R1: If $h_1(x_1) = h_2(x_2)$ then output this label, otherwise choose the h_i that belongs to the class with higher θ

R2: Same as R1, but when in conflict output the label that occurred less often in the sample

R3: Output the label that occurred less often in the sample

Three resolution rules
with upper bounds

- First fix some consistent $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$

  R1: If $h_1(x_1) = h_2(x_2)$ then output this label, otherwise choose the h_i that belongs to the class with higher θ

  R2: Same as R1, but when in conflict output the label that occurred less often in the sample

  R3: Output the label that occurred less often in the sample

Label complexity

$$\widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right)$$

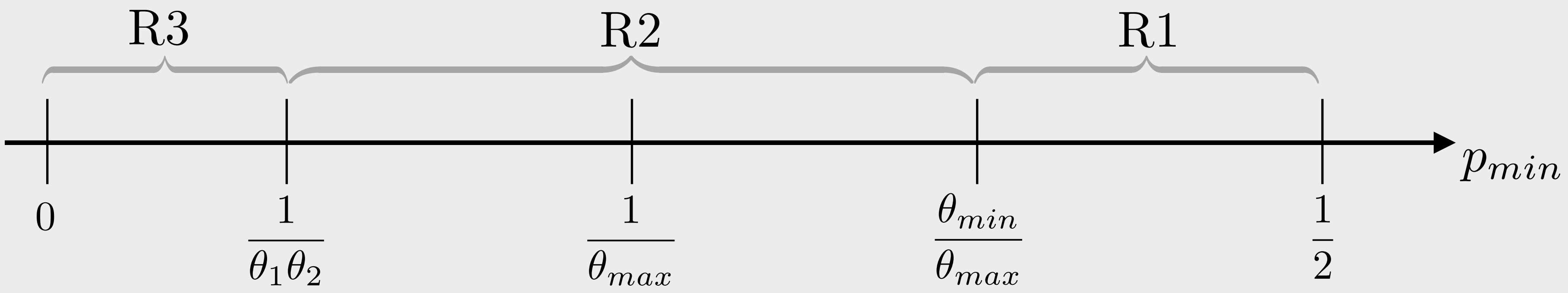
$$\widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \max\left\{\frac{1}{p_{min}}, \theta_{max}\right\}}\right)$$

$$\widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1 \theta_2}\right)$$

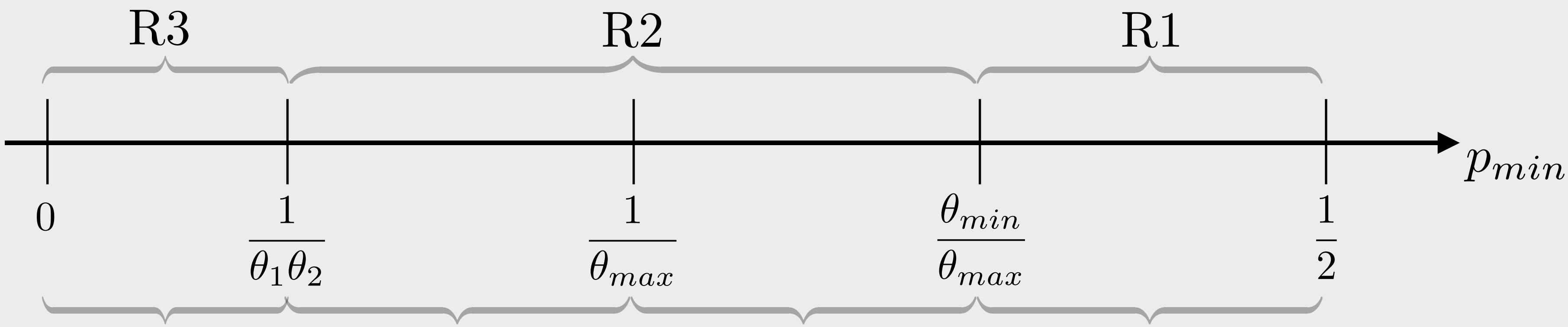
The combined rule

a general upper bound

The combined rule
a general upper bound



The combined rule
a general upper bound

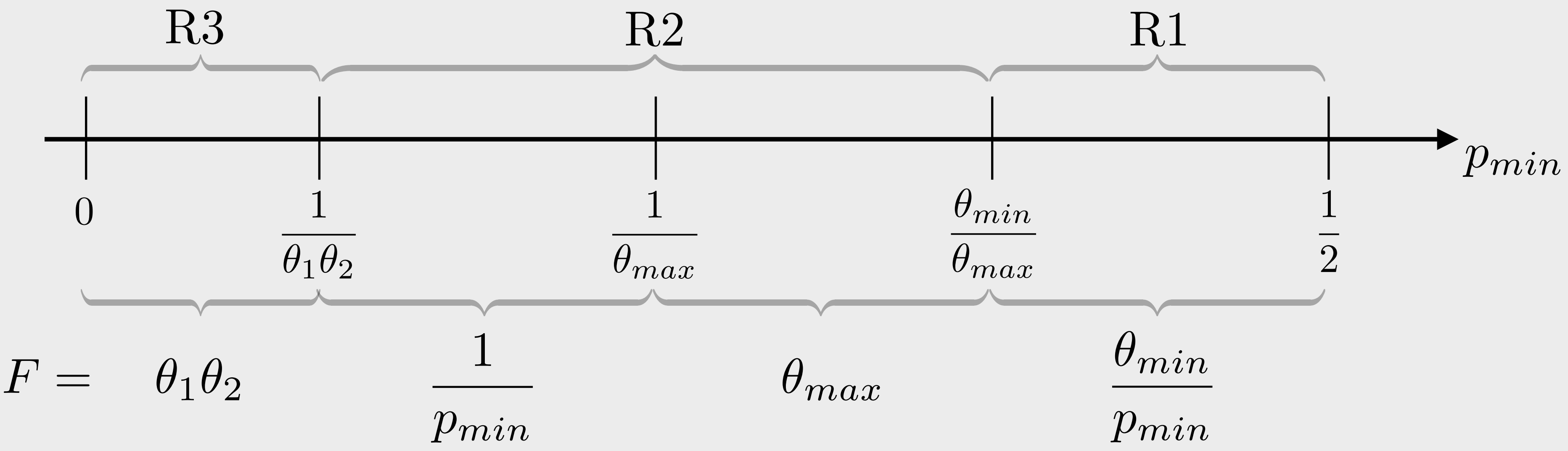


$F =$ $\theta_1\theta_2$ $\quad \frac{1}{p_{min}} \quad \theta_{max} \quad \frac{\theta_{min}}{p_{min}}$

$$m = \widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot F}\right)$$

The combined rule

a general upper bound



R3 R2 R1

0 $\frac{1}{\theta_1\theta_2}$ $\frac{1}{\theta_{max}}$ $\frac{\theta_{min}}{\theta_{max}}$ $\frac{1}{2}$ p_{min}

$F =$ $\theta_1\theta_2$ $\frac{1}{p_{min}}$ θ_{max} $\frac{\theta_{min}}{p_{min}}$

$$m = \widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot F}\right)$$

- Learner has to estimate p_{min} by $\hat{p}_{min}$
- The upper bounds still hold

Proof
only for rule 3

- WLOG $\hat{p}_- \geq 1/2$
- With high probability (after $\widetilde{O}(1)$ examples): $p_- \geq 1/4$
- R3: if $x_1 \in \mathrm{DIS}_1$ and $x_2 \in \mathrm{DIS}_2$ then output '$+$'
- If R3 makes an error on (x_1, x_2), then $x_1 \in \mathrm{DIS}_1$, $x_2 \in \mathrm{DIS}_2$ and the true label is $-$

Proof
only for rule 3

- With high probability:

$$
\begin{aligned}
& \mathbb{P}\big(\text{error on } (x_1, x_2)\big) \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1,\ x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1 | - \big) \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&= \frac{1}{p_-} \cdot \mathbb{P}\big(x_1 \in \mathrm{DIS}_1 | - \big) p_- \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&\leq \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{\mathbb{P}\big(x_1 \in \mathrm{DIS}_1\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_1}{d_2} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \cdot \underbrace{\mathbb{P}\big(x_2 \in \mathrm{DIS}_2\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_2}{d_1} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \leq \epsilon
\end{aligned}
$$

Proof
only for rule 3

- With high probability:

$$\begin{aligned}
&\mathbb{P}\big(\text{error on } (x_1, x_2)\big) \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1,\ x_2 \in \mathrm{DIS}_2 | -\big)p_- \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1|-\big) \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2| -\big)p_- \\
&= \frac{1}{p_-} \cdot \mathbb{P}\big(x_1 \in \mathrm{DIS}_1| -\big)p_- \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2| -\big)p_- \\
&\leq \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{\mathbb{P}\big(x_1 \in \mathrm{DIS}_1\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_1}{d_2} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \cdot \underbrace{\mathbb{P}\big(x_2 \in \mathrm{DIS}_2\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_2}{d_1} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \ \leq\ \epsilon
\end{aligned}$$

conditional independence →

Proof

only for rule 3

- With high probability:

$$
\begin{aligned}
&\mathbb{P}\big(\text{error on } (x_1, x_2)\big) \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1,\ x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&= \mathbb{P}\big(x_1 \in \mathrm{DIS}_1 | -\big) \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&= \frac{1}{p_-} \cdot \mathbb{P}\big(x_1 \in \mathrm{DIS}_1 | - \big) p_- \cdot \mathbb{P}\big(x_2 \in \mathrm{DIS}_2 | - \big) p_- \\
&\leq \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{\mathbb{P}\big(x_1 \in \mathrm{DIS}_1\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_1}{d_2} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \cdot \underbrace{\mathbb{P}\big(x_2 \in \mathrm{DIS}_2\big)}_{\leq \frac{1}{2}\sqrt{\frac{d_2}{d_1} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}} \leq \epsilon
\end{aligned}
$$

conditional independence

$$m = \widetilde{O}\left(\sqrt{\frac{d_1 d_2}{\epsilon}} \cdot \theta_1 \theta_2\right)$$

and the earlier lemma

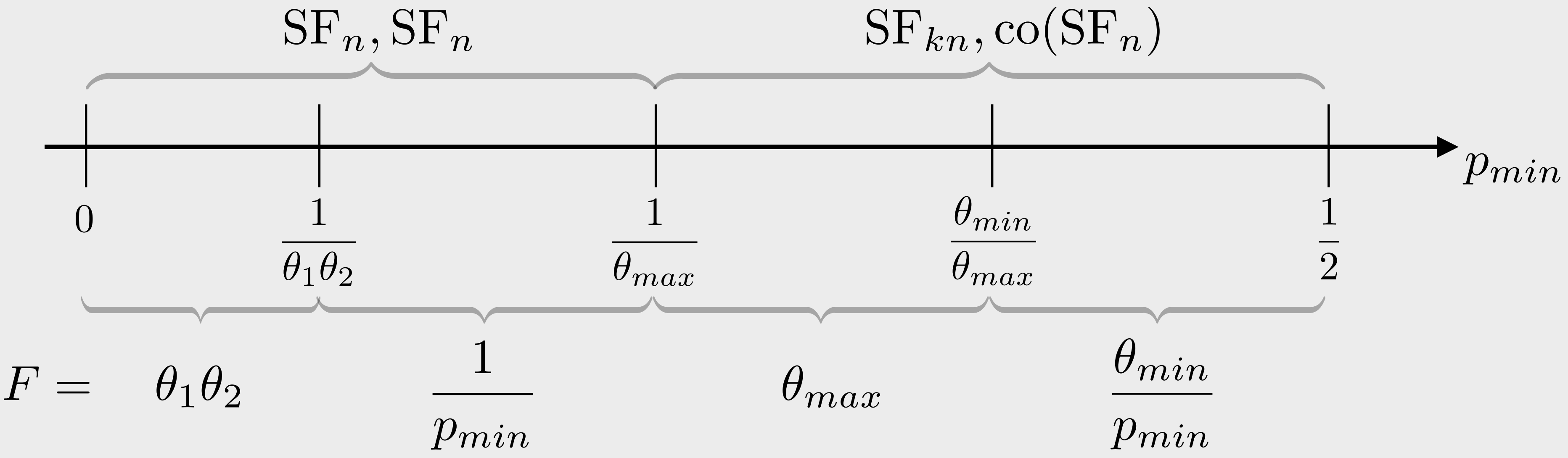
Lower bounds
for special classes

Lower bounds
for special classes

$$F = \quad \theta_1\theta_2 \qquad \frac{1}{p_{min}} \qquad \theta_{max} \qquad \frac{\theta_{min}}{p_{min}}$$

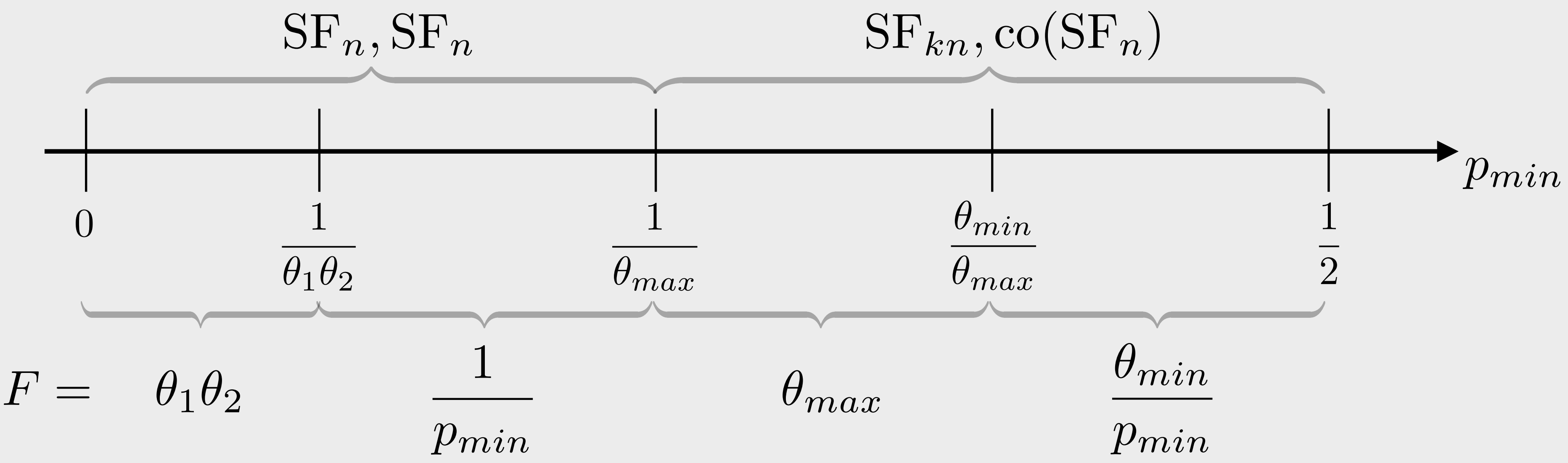
$$m = \widetilde{\Omega}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot F}\right)$$

Lower bounds
for special classes



$$m = \widetilde{\Omega}\left(\sqrt{\frac{d_1 d_2}{\varepsilon}} \cdot F\right)$$

Lower bounds
for special classes

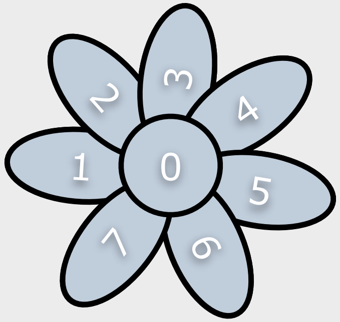


$$m = \widetilde{\Omega}\left(\sqrt{\frac{d_1 d_2}{\varepsilon}} \cdot F\right)$$

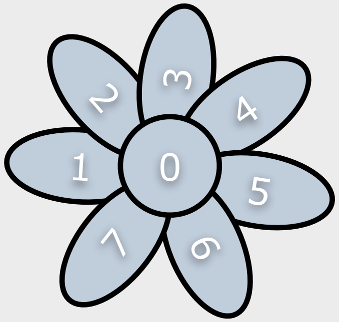
- We have a ‘witness class’ for each upper bound
- The fiendish $\mathbb{P}$ concentrates on the flower head $\{0\}$

Proof technique
padding the VC dimension

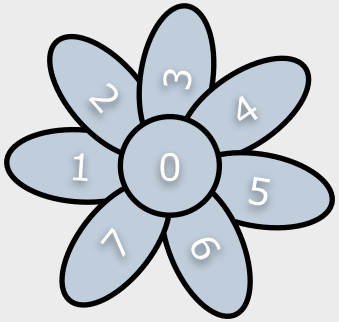
- $\mathrm{VCdim}\big(\mathrm{SF}_n\big) = \mathrm{VCdim}\big(\mathrm{co}(\mathrm{SF}_n)\big) = 1$
- Get classes of arbitrary VC dimension by **padding**:
  - $\mathcal{C}^{[d]} :=$ d-fold “disjoint unions” of $\mathcal{C}$
  - $\mathrm{VCdim}(\mathcal{C}^{[d]}) = d \cdot \mathrm{VCdim}(\mathcal{C})$
  - **Lemma**: $\theta(\mathcal{C}^{[d]}, \mathcal{H}^{[d]}) = \theta(\mathcal{C}, \mathcal{H})$

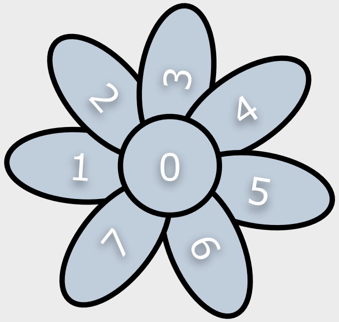


Proof technique
padding the VC dimension



- $\mathrm{VCdim}\big(\mathrm{SF}_n\big) = \mathrm{VCdim}\big(\mathrm{co}(\mathrm{SF}_n)\big) = 1$
- Get classes of arbitrary VC dimension by **padding**:
  - $\mathcal{C}^{[d]} :=$ d-fold "disjoint unions" of $\mathcal{C}$
  - $\mathrm{VCdim}(\mathcal{C}^{[d]}) = d \cdot \mathrm{VCdim}(\mathcal{C})$
  - **Lemma**: $\theta(\mathcal{C}^{[d]}, \mathcal{H}^{[d]}) = \theta(\mathcal{C}, \mathcal{H})$





Outline

1. Motivation and ‘What is Co-Training?’
2. The disagreement coefficient
3. Label complexity bounds
4. Combinatorial bounds
5. Final Remarks

Singleton Size
definition

- $s^+(\mathcal{C}) :=$ size of the largest **singleton** sub-class in $\mathcal{C}$
- $s^-(\mathcal{C}) :=$ size of the largest **co-singleton** sub-class in $\mathcal{C}$
- $\mathcal{C}^+ :=$ all unions of concepts from $\mathcal{C}$
- $\mathcal{C}^- :=$ all intersections of concepts from $\mathcal{C}$

Singleton Size
definition

- $s^+(\mathcal{C})$:= size of the largest **singleton** sub-class in $\mathcal{C}$
- $s^-(\mathcal{C})$:= size of the largest **co-singleton** sub-class in $\mathcal{C}$
- $\mathcal{C}^+$:= all unions of concepts from $\mathcal{C}$
- $\mathcal{C}^-$:= all intersections of concepts from $\mathcal{C}$

$$\text{Singletons}_n := \{\{1\}, \{2\}, \ldots, \{n\}\}$$

$$\text{co-Singletons}_n := \{\{2, 3, \ldots, n\}, \{1, 3, \ldots, n\}, \ldots, \{1, 2, \ldots, n-1\}\}$$

Combinatorial upper bound

Theorem:
For rule R3 and hypothesis classes $\mathcal{H}_{1,2} = \mathcal{C}^+_{1,2} \cup \mathcal{C}^-_{1,2}$
$m = \tilde{O}(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon})$ labeled examples are sufficient.

Combinatorial upper bound

Theorem:
For rule R3 and hypothesis classes $\mathcal{H}_{1,2} = \mathcal{C}^{+}_{1,2} \cup \mathcal{C}^{-}_{1,2}$
$m = \widetilde{O}\left(\sqrt{\max\{s_1^{+} s_2^{+}, s_1^{-} s_2^{-}\}/\varepsilon}\right)$ labeled examples are sufficient.

- Outputs the largest consistent hypothesis in $\mathcal{C}^{+}_{1,2}$ or the smallest consistent hypothesis in $\mathcal{C}^{-}_{1,2}$
- Strong connection to “PAC-learnability from positive examples alone” by Geréb-Graus (1989)

Combinatorial lower bound

Theorem:
Let $3 \leq s^+_{1,2}, s^-_{1,2} < \infty$ and let $\varepsilon > 0$ be sufficiently small, then $m = \widetilde{\Omega}\left(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}\right)$ labeled examples are necessary.

Combinatorial lower bound

Theorem:
Let $3 \leq s^+_{1,2}, s^-_{1,2} < \infty$ and let $\varepsilon > 0$ be sufficiently small, then $m = \widetilde{\Omega}\left(\sqrt{\max\{s^+_1 s^+_2, s^-_1 s^-_2\}/\varepsilon}\right)$ labeled examples are necessary.

- One can drop the restriction $3 \leq s^+_{1,2}, s^-_{1,2}$ and still prove tight bounds
- Valid for $p_{min} = \varepsilon \leq 1/\max\{s^+_1 s^+_2, s^-_1 s^-_2\}$,
  i.e. p_{min} is small

Outline

1. Motivation and 'What is Co-Training?'
2. The disagreement coefficient
3. Label complexity bounds
4. Combinatorial bounds
5. Final Remarks

More results

see the proceedings and upcoming journal version

- Sharper bounds for classes with one-sided errors
- Multiple views $x = (x_1, \ldots, x_k)$ lead to bounds like

$$m = \widetilde{O}\left(\sqrt[k]{\frac{d_1\theta_1 \cdots d_k\theta_k}{\varepsilon}} \right)$$

- Negative result under the α-expanding assumption (weaker than conditional independence)

Open questions and current work

- Bounds for infinite $s^-(\mathcal{C})$, $s^+(\mathcal{C})$?
  - One can find classes with bounds like: $m = \widetilde{\Omega}\left(\sqrt{\frac{d_1 d_2}{\varepsilon}} + \frac{1}{\varepsilon}\right)$
- Can some of our techniques be applied to active learning?

Thank you for your attention!

-- end of talk --