Helper Slides on Abduction-based Minimal Explanations

Supratik Chakraborty
IIT Bombay

Abduction in Logic

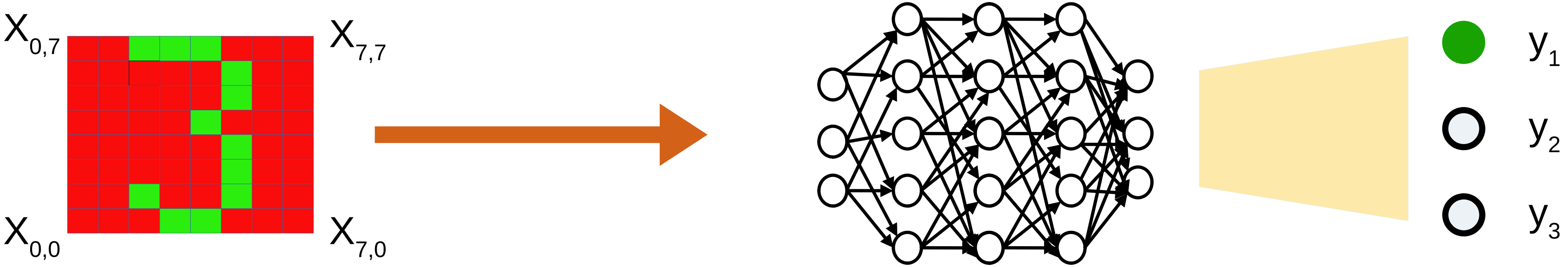
Given a theory (consistent set of sentences) $\mathcal{F}$ and a formula $\mathcal{E}$ in a logic $\mathcal{L}$
Find a formula α such that

- $\alpha \models \mathcal{F} \Rightarrow \mathcal{E}$
- $\mathcal{F} \wedge \alpha$ is consistent

We often want α to be as weak (permissive) as possible.

α is an "explanation" of $\mathcal{E}$ in theory $\mathcal{F}$

Formulating Explanation as Abduction



$$\mathcal{C} =$$

$$(x_{0,7} = R) \wedge (x_{1,7} = R) \wedge (x_{2,7} = G) \wedge (x_{3,7} = G) \wedge \cdots (x_{7,7} = R) \wedge$$

$$\vdots$$

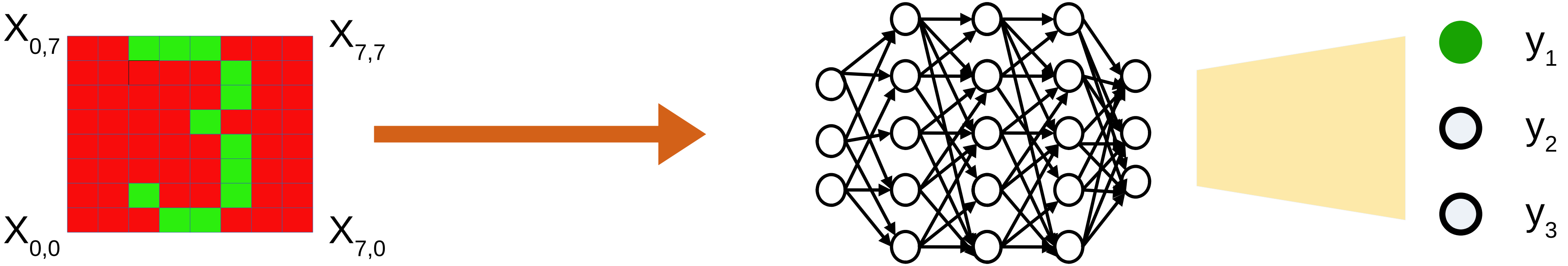
$$(x_{0,0} = R) \wedge (x_{1,0} = R) \wedge (x_{2,0} = R) \wedge (x_{3,0} = G) \wedge \cdots (x_{7,0} = R)$$

$$\mathcal{F}$$

$$\mathcal{E} = (y_1 > y_2) \wedge (y_1 > y_3)$$

Clearly, $\mathcal{C} \wedge \mathcal{F} \wedge \mathcal{E}$ is consistent.

Formulating Explanation as Abduction



$C =$

$$(x_{0,7} = R) \wedge (x_{1,7} = R) \wedge (x_{2,7} = G) \wedge (x_{3,7} = G) \wedge \cdots (x_{7,7} = R) \wedge$$

$$\vdots$$

$$(x_{0,0} = R) \wedge (x_{1,0} = R) \wedge (x_{2,0} = R) \wedge (x_{3,0} = G) \wedge \cdots (x_{7,0} = R)$$

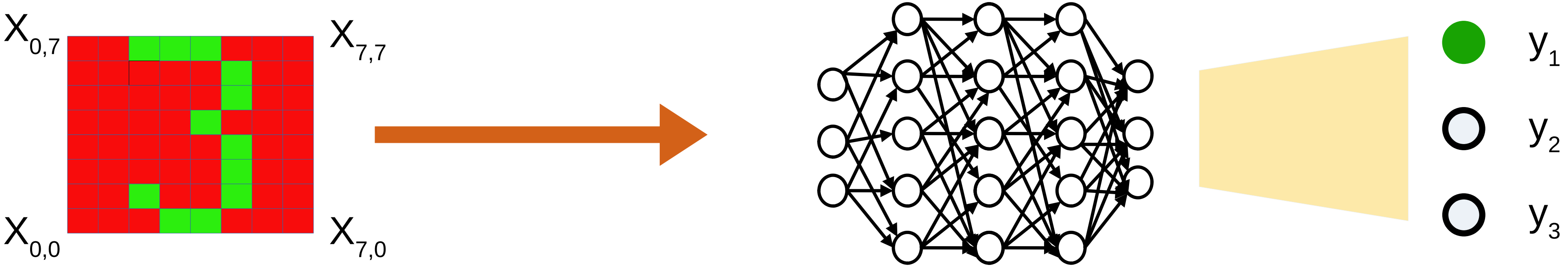
$\mathcal{F}$

$$\mathcal{E} = (y_1 > y_2) \wedge (y_1 > y_3)$$

Find smallest $\mathcal{C}' \subseteq \mathcal{C}$ s.t.

(a) $\mathcal{C}' \wedge \mathcal{F}$ is consistent, and (b) $\mathcal{C}' \models \mathcal{F} \Rightarrow \mathcal{E}$

Building C' Lazily



$C =$

$$(x_{0,7} = R) \wedge (x_{1,7} = R) \wedge (x_{2,7} = G) \wedge (x_{3,7} = G) \wedge \cdots (x_{7,7} = R) \wedge$$

$$\vdots$$

$$(x_{0,0} = R) \wedge (x_{1,0} = R) \wedge (x_{2,0} = R) \wedge (x_{3,0} = G) \wedge \cdots (x_{7,0} = R)$$

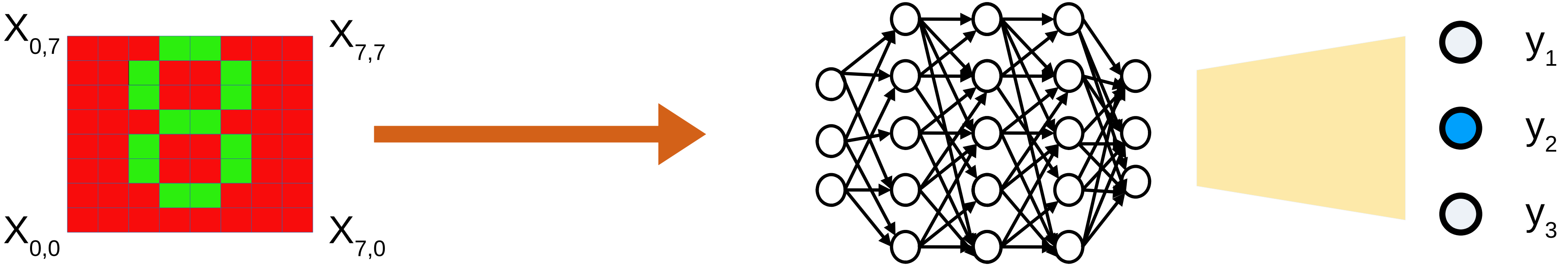
$\mathcal{F}$

$\mathcal{E} = (y_1 > y_2) \wedge (y_1 > y_3)$

Does the empty subset of C suffice?

Does $\models \mathcal{F} \Rightarrow \mathcal{E}$ hold?

Building C' Lazily



$\widehat{C} =$

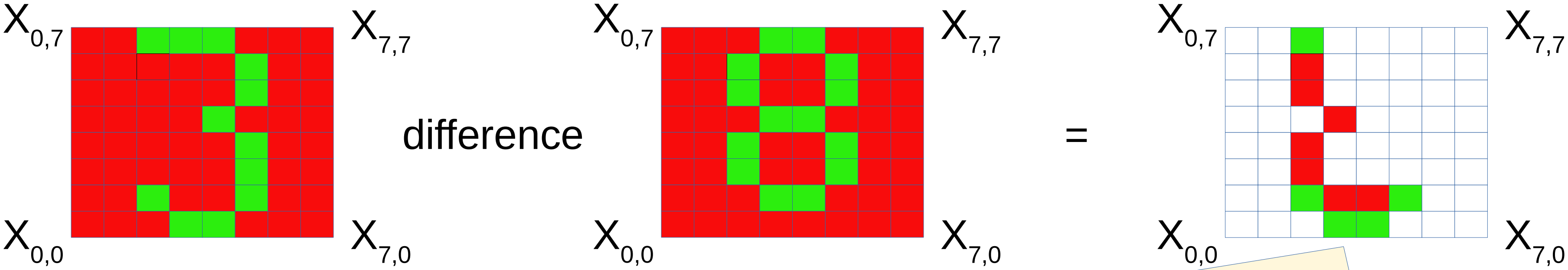
$(x_{0,7} = {\color{red}R}) \wedge (x_{1,7} = {\color{red}R}) \wedge (x_{2,7} = {\color{red}R}) \wedge (x_{3,7} = {\color{green}G}) \wedge \cdots (x_{7,7} = {\color{red}R}) \wedge$ $\mathcal{F}$ $\widehat{\mathcal{E}} = ({\color{blue}y_2} > y_1) \wedge ({\color{blue}y_2} > y_3)$

$$\vdots$$

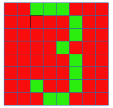
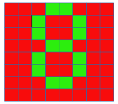
$(x_{0,0} = {\color{red}R}) \wedge (x_{1,0} = {\color{red}R}) \wedge (x_{2,0} = {\color{red}R}) \wedge (x_{3,0} = {\color{red}R}) \wedge \cdots (x_{7,0} = {\color{red}R})$

Certainly $\models \mathcal{F} \Rightarrow \mathcal{E}$ doesn't hold!

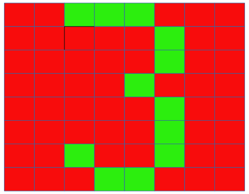
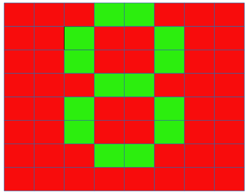
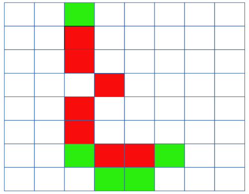
How do the two inputs differ?



$$S_1 = \{(x_{2,7} = G), (x_{2,6} = R), (x_{2,5} = R), (x_{3,4} = R), (x_{2,3} = R), (x_{2,2} = R), (x_{2,1} = G), (x_{3,1} = R), (x_{4,1} = R), (x_{5,1} = G), (x_{3,0} = G), (x_{4,0} = G)\}$$

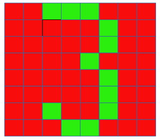
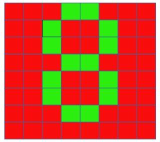
Unless one of the literals in S_1 is included in the explanation C’, we can’t distinguish between  and 

Choosing subset of C

 difference  = 

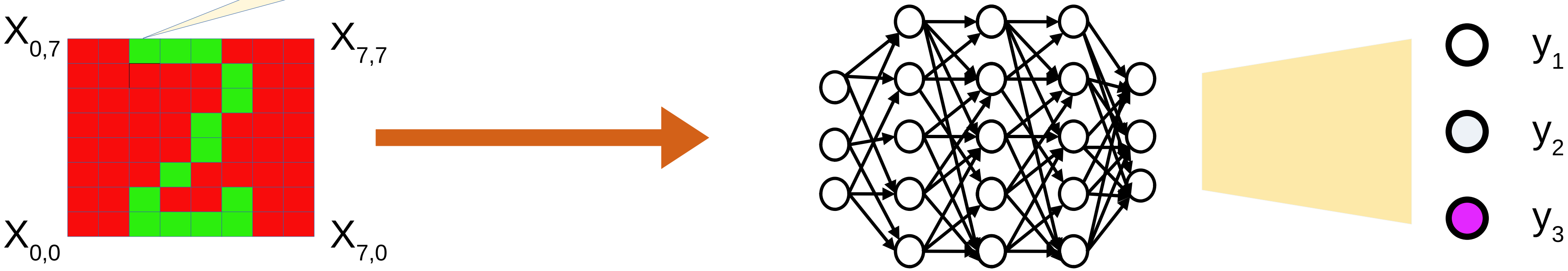
$$S_1 = \{(x_{2,7} = G), (x_{2,6} = R), (x_{2,5} = R), (x_{3,4} = R), (x_{2,3} = R), (x_{2,2} = R), (x_{2,1} = G), (x_{3,1} = R), (x_{4,1} = R), (x_{5,1} = G), (x_{3,0} = G), (x_{4,0} = G)\}$$

Suppose we choose $(x_{2,7} = G)$ for $C' \subseteq C$

Certainly this distinguishes  from 

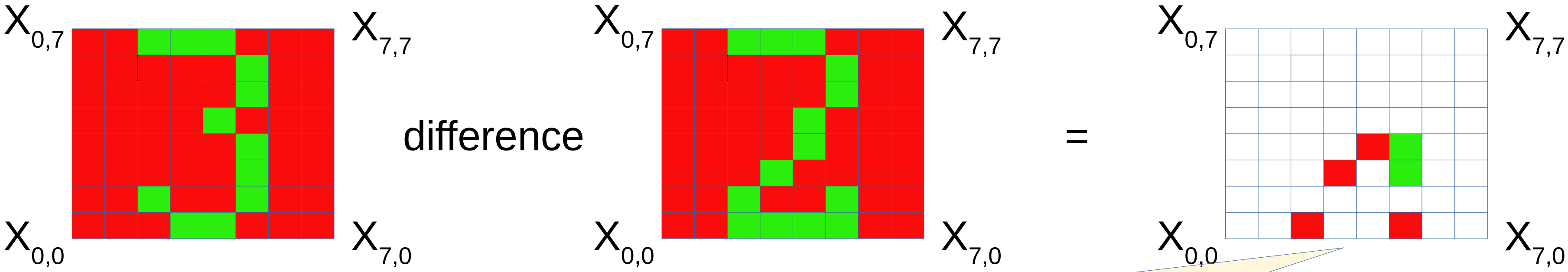
So, have we found the explanation?

Does $(x_{2,7} = G) \models \mathcal{F} \Rightarrow \mathcal{E}$ hold?

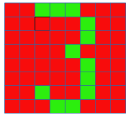
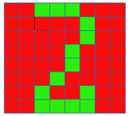


Clearly not!

How do the two inputs differ again?



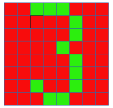
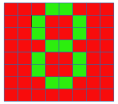
$$S_2 = \{(x_{4,3} = R), (x_{5,3} = G), (x_{3,2} = R), (x_{5,2} = G), (x_{2,0} = R), (x_{5,0} = R)\}$$

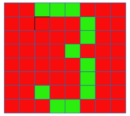
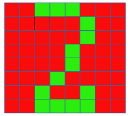
Unless one of the literals in S_2 is included in the explanation C',
we can't distinguish between  and 

Finding updated C’

$$S_1 = \{(x_{2,7} = G), (x_{2,6} = R), (x_{2,5} = R), (x_{3,4} = R), (x_{2,3} = R),$$
$$(x_{2,2} = R), (x_{2,1} = G), (x_{3,1} = R), (x_{4,1} = R),$$
$$(x_{5,1} = G), (x_{3,0} = G), (x_{4,0} = G)\}$$

$$S_2 = \{(x_{4,3} = R), (x_{5,3} = G), (x_{3,2} = R), (x_{5,2} = G),$$
$$(x_{2,0} = R), (x_{5,0} = R)\}$$

Unless one of the literals in S_1 is included in the explanation C’,
we can’t distinguish between  and 

Unless one of the literals in S_2 is included in the explanation C’,
we can’t distinguish between  and 

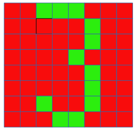
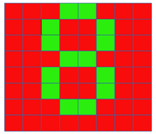
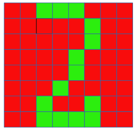
Finding updated C'

$$S_1 = \{(x_{2,7} = G), (x_{2,6} = R), (x_{2,5} = R), (x_{3,4} = R), (x_{2,3} = R), (x_{2,2} = R), (x_{2,1} = G), (x_{3,1} = R), (x_{4,1} = R), (x_{5,1} = G), (x_{3,0} = G), (x_{4,0} = G)\}$$

$$S_2 = \{(x_{4,3} = R), (x_{5,3} = G), (x_{3,2} = R), (x_{5,2} = G), (x_{2,0} = R), (x_{5,0} = R)\}$$

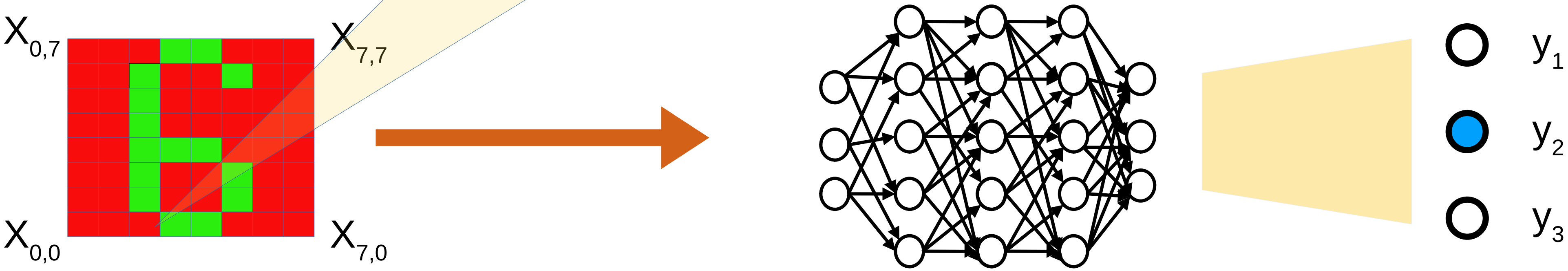
Find a minimum hitting set of S_1 and S_2

$$C' = (x_{3,0} = G) \wedge (x_{2,0} = R)$$

Certainly distinguishes  from both  and 

So, have we found the explanation?

Does $(x_{2,0} = R) \wedge (x_{3,0} = G) \models \mathcal{F} \Rightarrow \mathcal{E}$ hold?



Clearly not!

Continuing the process

- Find difference with current counterexample
- Find another set S_3 from which we must choose a literal
- Find hitting set C' of S_1, S_2, S_3, ...
- Check if C' serves as an abductive explanation
  - $\text{Does } C' \models \mathcal{F} \Rightarrow \mathcal{E} \text{ ?}$
- If not, repeat above steps
- If yes, output C' as minimal explanation