Abstract Interpretation-Based Static Analysis for Machine Learning Verification and Explainability

17th International School on Modeling and Verification of Parallel Processes (MOVEP 2026)

Caterina Urban
Inria & École Normale Supérieure | Université PSL

Static Analysis for Machine Learning

17th International School on Modeling and Verification of Parallel Processes (MOVEP 2026)

Caterina Urban
Inria & École Normale Supérieure | Université PSL

Machine Learning in High-Stakes Systems

Safety-Critical Applications

Socio-Economic Applications

Machine Learning in High-Stakes Systems

HOW CAN WE TRUST?

🔒 CERTIFICATION

📜 Regulatory Standards
(e.g., EU AI Act)

🧪 Benchmarking & Testing

🛡️ Verification
- ROBUSTNESS
- SAFETY
- FAIRNESS
- SECURITY
- PRIVACY

🕵️ ACCOUNTABILITY

🗂️ Traceability
track design decisions, data, updates

🔍 Explainability
understand model decisions

👤 Responsibility
identify who is accountable

⚙️ Governance
oversight & risk management


FORMAL METHODS COMMUNITY
MACHINE LEARNING SOFTWARE
TRADITIONAL SAFETY-CRITICAL SOFTWARE
imgflip.com

Machine Learning in High-Stakes Systems

FORMAL METHODS FOR MACHINE LEARNING

🔒 CERTIFICATION

📜 **Regulatory Standards**
(e.g., EU AI Act)

🧪 **Benchmarking & Testing**

🛡️ **Verification**

- ROBUSTNESS
- SAFETY
- FAIRNESS
- SECURITY
- PRIVACY

🕵️ ACCOUNTABILITY

🗂️ **Traceability**
track design decisions, data, updates

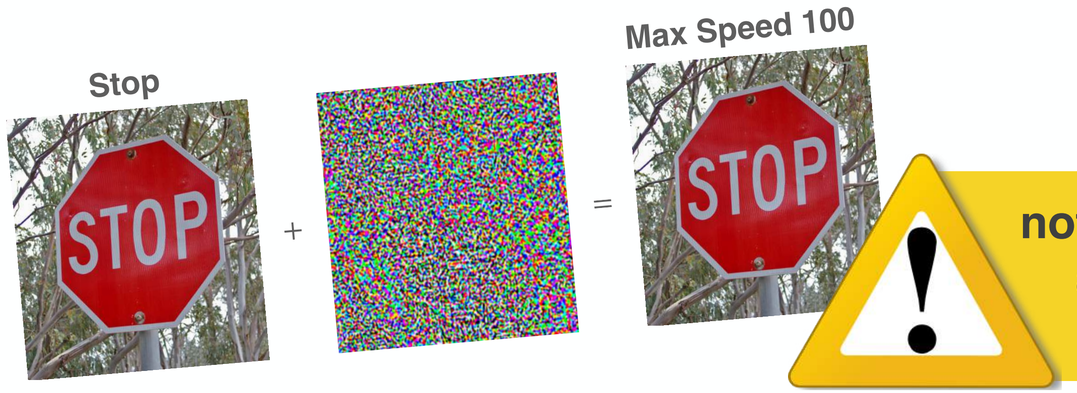
🔍 **Explainability**
understand model decisions

👤 **Responsibility**
identify who is accountable

⚙️ **Governance**
oversight & risk management

Machine Learning Development Pipeline

MODELS ONLY GIVE PROBABILISTIC GUARANTEES



not sufficient for guaranteeing an acceptable failure rate under any circumstance

Static Analysis for Trained Models

Verification

Explainability

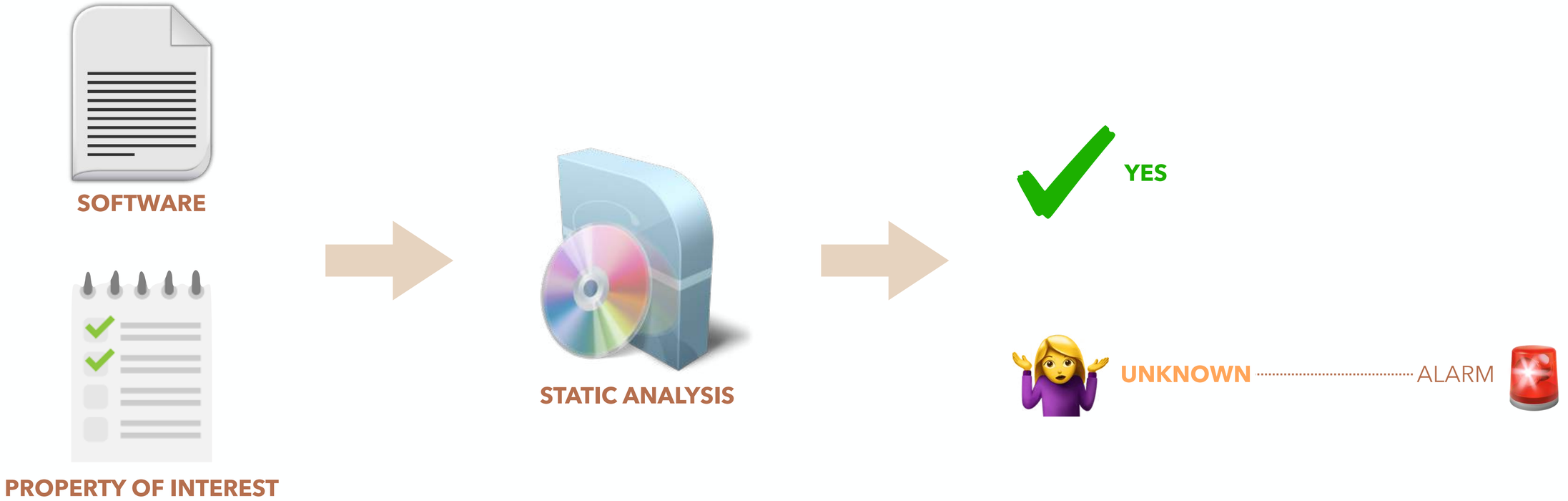
Static Analysis for Trained Models

Verification

Explainability

Static Analysis by Abstract Interpretation

AUTOMATIC COMPUTATION OF FORMAL GUARANTEES



Static Analysis by Abstract Interpretation

INTUITION

ABSTRACTION

$[\![P]\!]^\natural$

SOFTWARE

$[\![P]\!]$

SOUNDNESS

€ 10 +
€ 40 +
€ 30 +
€ 10

€ 90

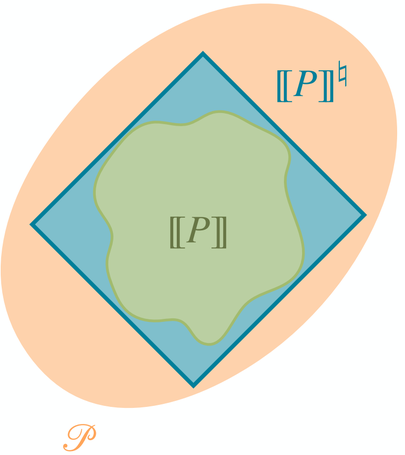
PROPERTY OF INTEREST

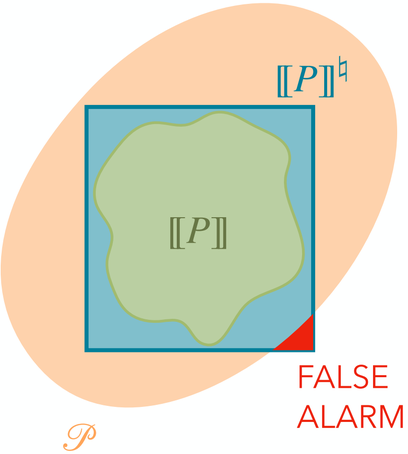
$\mathscr{P}$

€ 9.95 +
€ 35.85 +
€ 27.95 +
€ 4.85

€ 78.60

~~COMPLETENESS~~



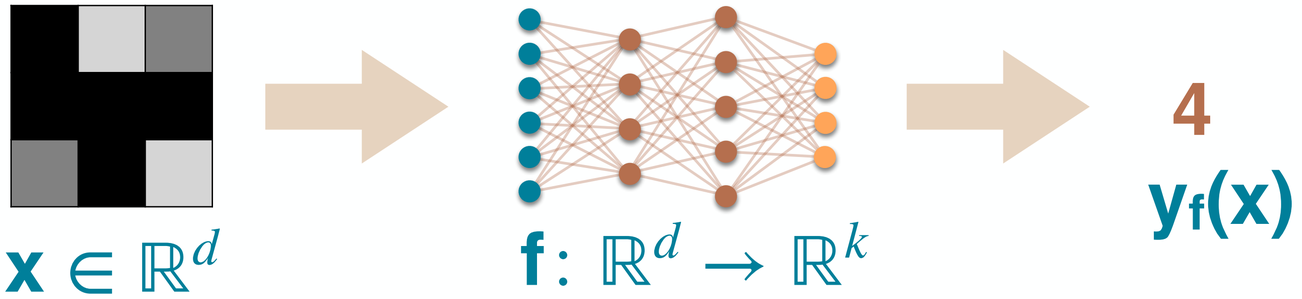


Static Analysis by Abstract Interpretation

SOFTWARE

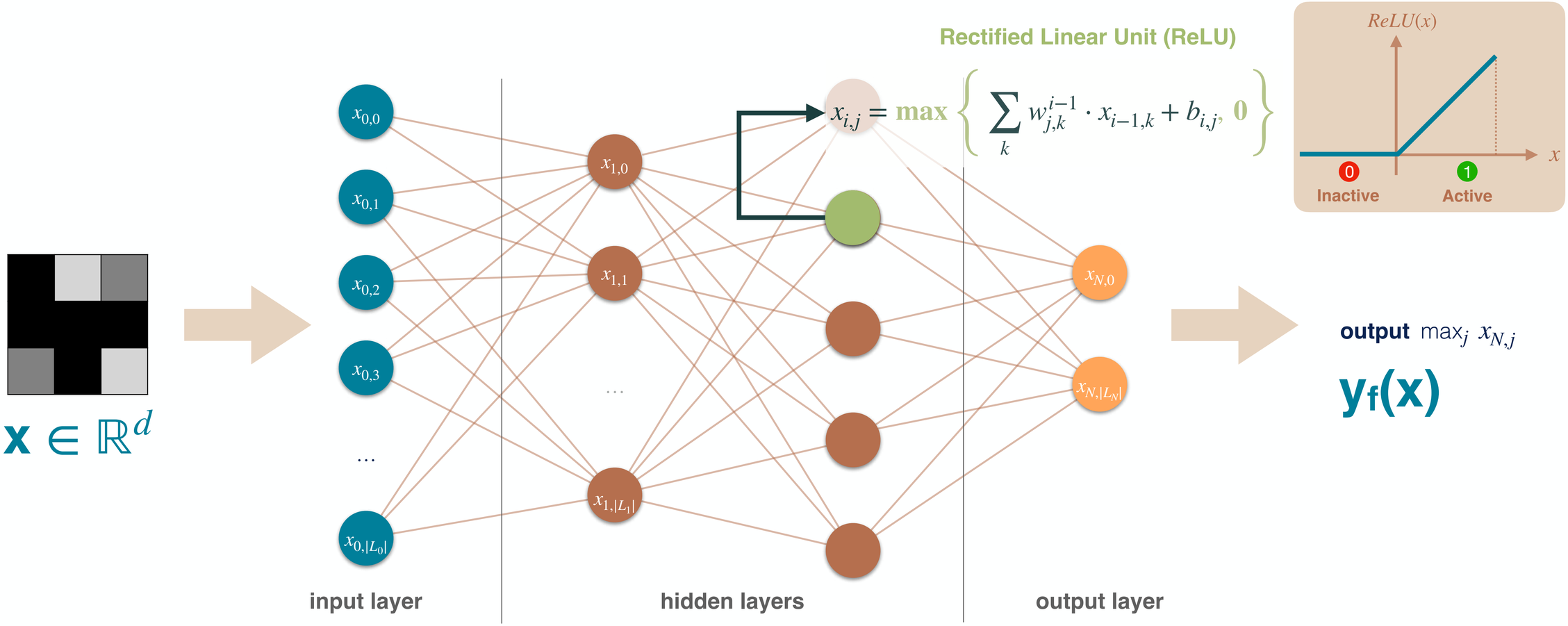
Neural Networks

SINGLE-CLASS CLASSIFIERS



Neural Networks

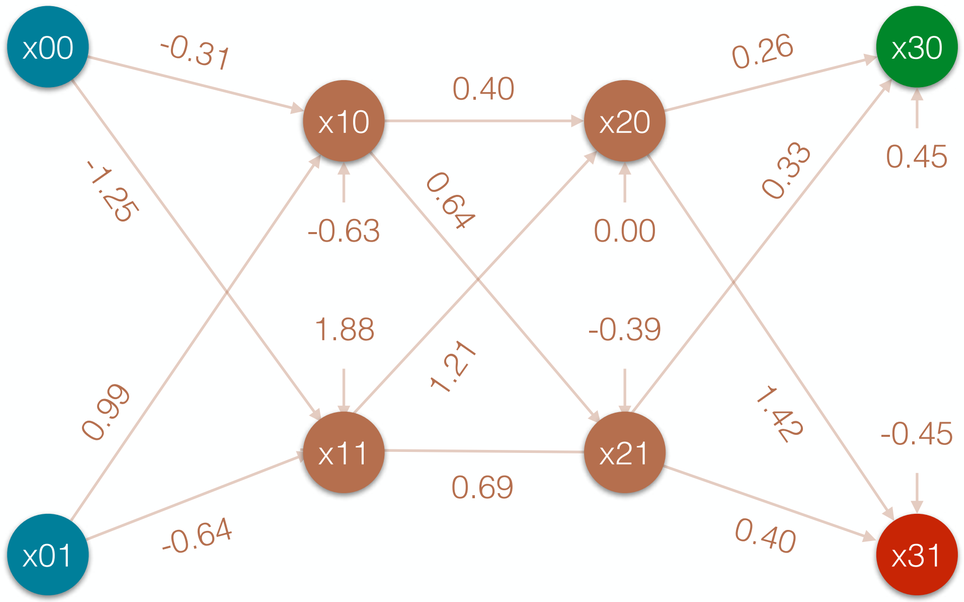
FEED-FORWARD RELU-ACTIVATED NEURAL NETWORKS



$$\mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}^k$$

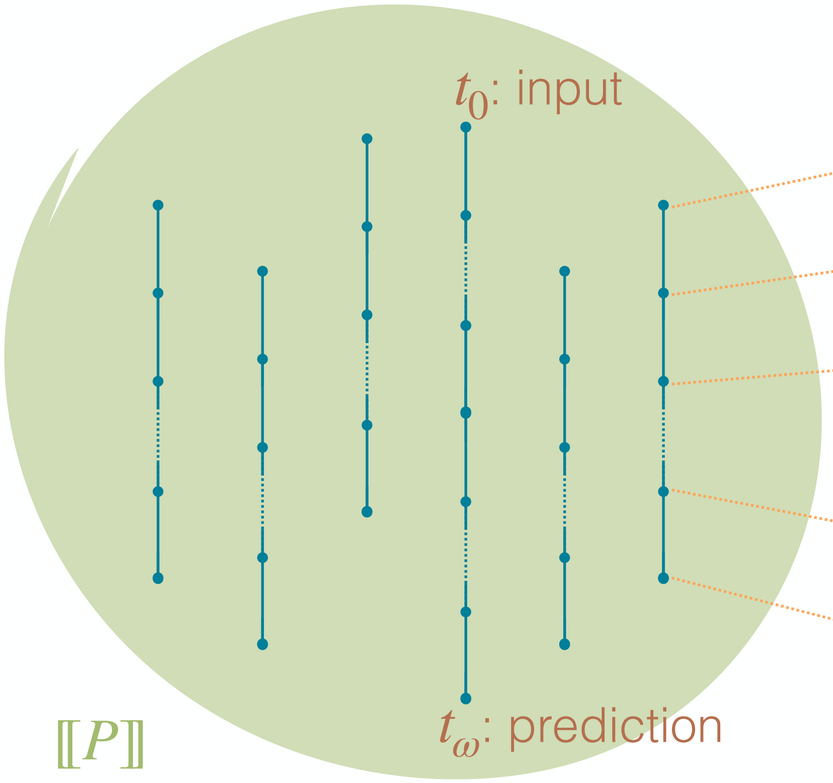
Neural Networks

EXAMPLE




```
x00 = input()
x01 = input()

x10 = -0.31 * x00 + 0.99 * x01 + (-0.63)
x11 = -1.25 * x00 + (-0.64) * x01 + 1.88

x10 = 0 if x10 < 0 else x10
x11 = 0 if x11 < 0 else x11

x20 = 0.40 * x10 + 1.21 * x11 + 0.00
x21 = 0.64 * x10 + 0.69 * x11 + (-0.39)

x20 = 0 if x20 < 0 else x20
x21 = 0 if x21 < 0 else x21

x30 = 0.26 * x20 + 0.33 * x21 + 0.45
X31 = 1.42 * x20 + 0.40 * x21 + (-0.45)

return '  ' if x31 < 30 else '  '
```

Maximal Trace Semantics




```
x00 = input()
x01 = input()

x10 = -0.31 * x00 + 0.99 * x01 + (-0.63)
x11 = -1.25 * x00 + (-0.64) * x01 + 1.88

x10 = 0 if x10 < 0 else x10
x11 = 0 if x11 < 0 else x11

x20 = 0.40 * x10 + 1.21 * x11 + 0.00
x21 = 0.64 * x10 + 0.69 * x11 + (-0.39)

x20 = 0 if x20 < 0 else x20
x21 = 0 if x21 < 0 else x21

x30 = 0.26 * x20 + 0.33 * x21 + 0.45
X31 = 1.42 * x20 + 0.40 * x21 + (-0.45)

return '  ' if x31 < 30 else '  '
```

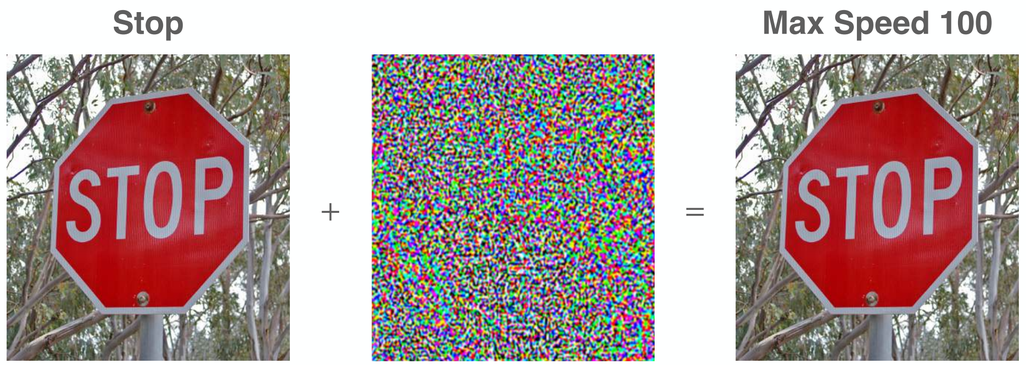
Static Analysis by Abstract Interpretation

PROPERTY OF INTEREST

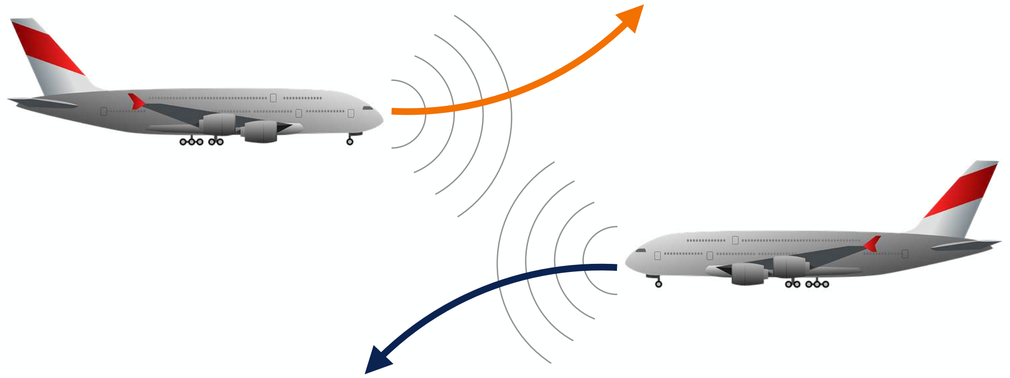
PROPERTY OF INTEREST

$\mathscr{P}$

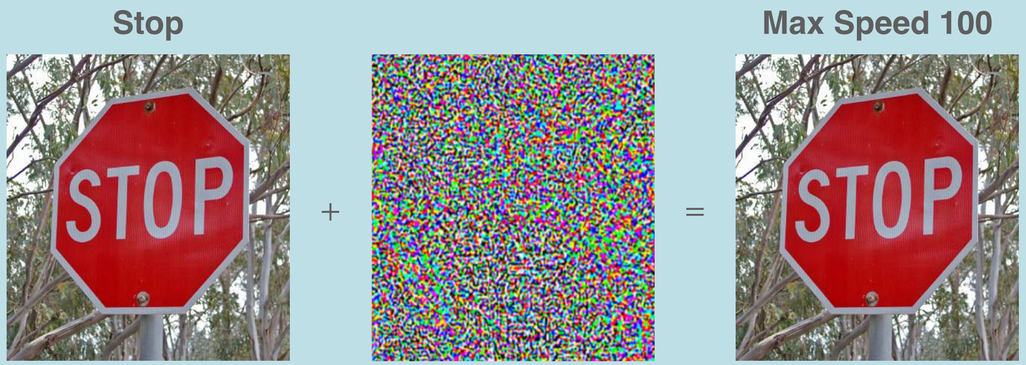
Stability



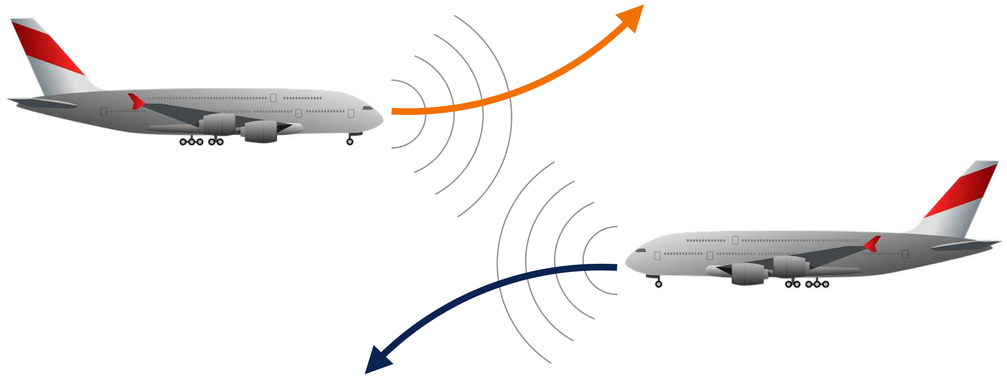
Safety



Stability

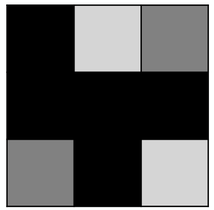


Safety

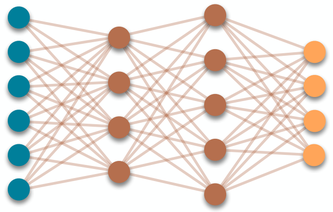


Local Prediction Stability

PREDICTION IS UNAFFECTED BY INPUT PERTURBATIONS



$\mathbf{x} \in \mathbb{R}^d$



f

4

$y_f(\mathbf{x})$

$(\mathbf{x}, \mathscr{A}, \epsilon, \mathbf{f})$

v

1 $\quad \forall x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') = y_f(x)$

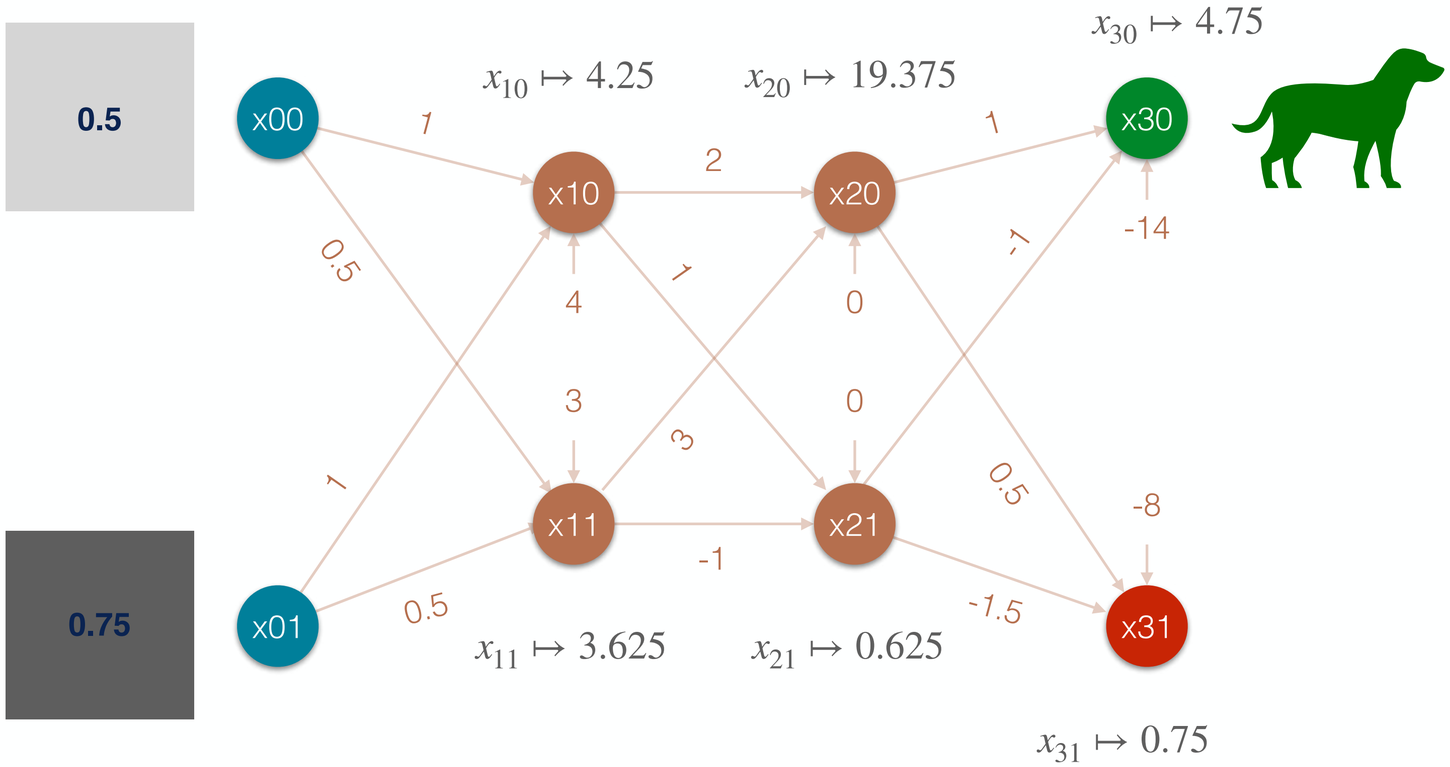
0 🤷‍♀️

-1 $\quad \exists x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') \neq y_f(x)$

$$B^{\epsilon}_{\mathscr{A}}(\mathbf{x}) \stackrel{\text{def}}{=} \left\{ \mathbf{x}' \in \mathbb{R}^d \;\middle|\; \mathbf{x}'_{\overline{\mathscr{A}}} = \mathbf{x}_{\overline{\mathscr{A}}} \wedge \left\| \mathbf{x}'_{\mathscr{A}} - \mathbf{x}_{\mathscr{A}} \right\|_{\infty} \leq \epsilon \right\}$$

Local Prediction Stability

EXAMPLE

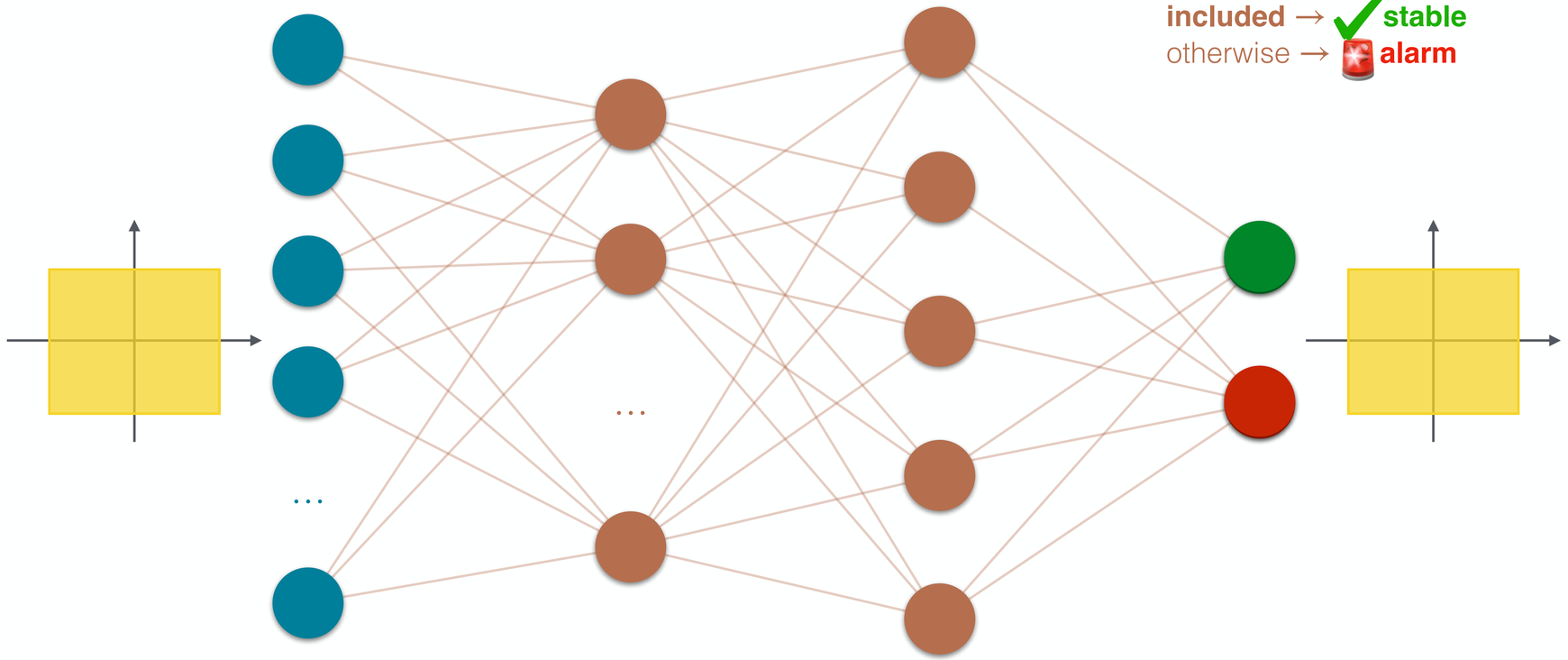


$$B^{\epsilon}_{\mathscr{A}}(\langle \mathbf{0.5}, \mathbf{0.75} \rangle) \stackrel{\text{def}}{=} \left\{ \mathbf{x}' \in \mathbb{R}^2 \;\middle|\; 0 \leq \mathbf{x}'_0 \leq 1 \wedge 0 \leq \mathbf{x}'_1 \leq 1 \right\}$$

Verifying Local Prediction Stability

STATIC FORWARD ANALYSIS

② check output for **inclusion** in **expected output**:
included → ✅ **stable**
otherwise → 🚨 **alarm**



① proceed **forwards from an** **abstraction** of $B^{\epsilon}_{\mathscr{A}}(\mathbf{x})$

Static Analysis by Abstract Interpretation

ABSTRACTION #1: INTERVALS ABSTRACT DOMAIN

ABSTRACTION

$[\![P]\!]^\sharp$

PROPERTY OF INTEREST

$\mathscr{P}$

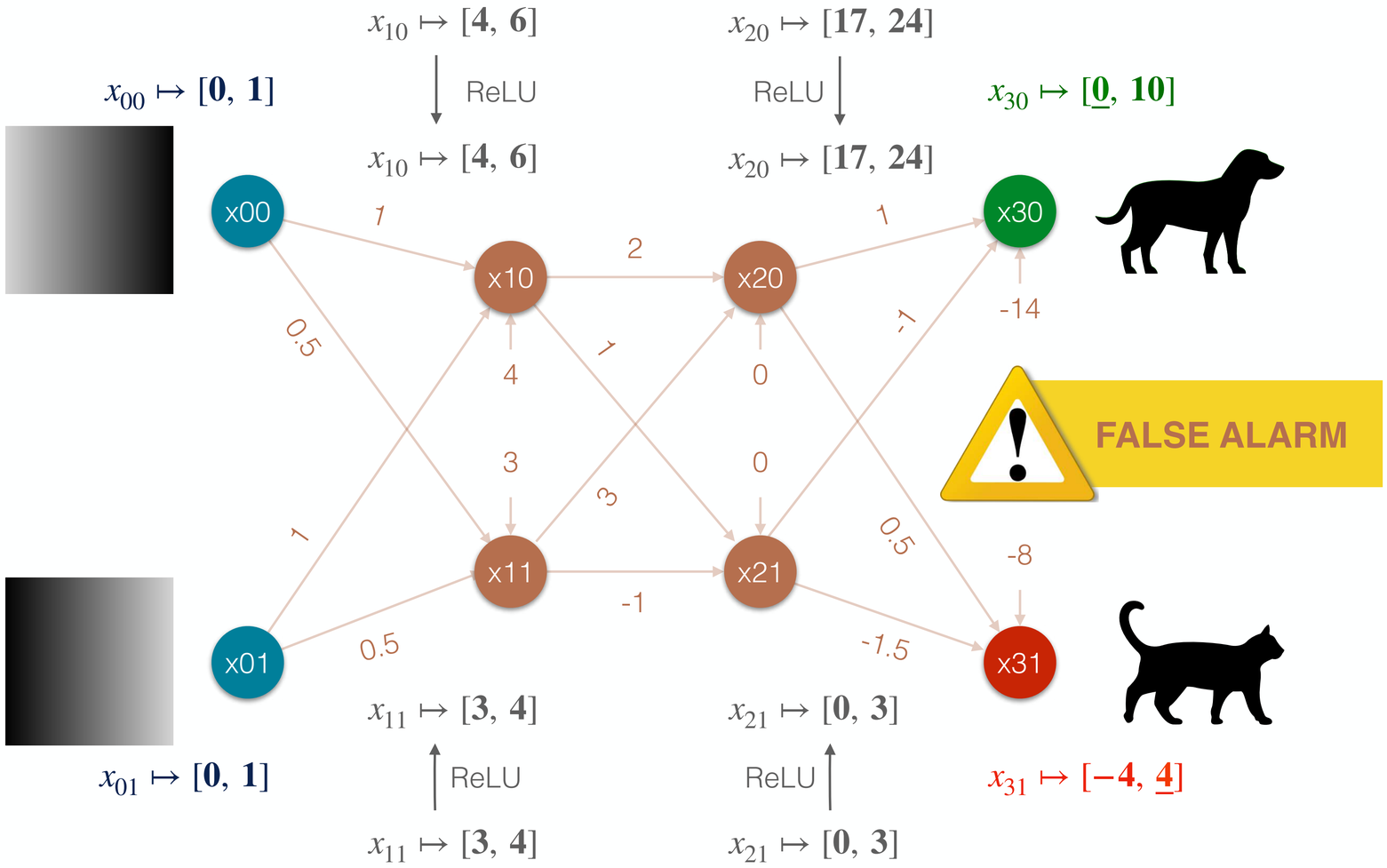
SOFTWARE

$[\![P]\!]$

Intervals Abstract Domain

$$x_{i,j} \mapsto [a, b]$$
$$a, b \in \mathbb{R}$$

EXAMPLE



Static Analysis by Abstract Interpretation

ABSTRACTION #2: SYMBOLIC ABSTRACT DOMAIN

ABSTRACTION

$[\![P]\!]^\sharp$

SOFTWARE

$[\![P]\!]$

PROPERTY OF INTEREST

$\mathscr{P}$

€ 10 +
€ 36 +
€ 28 +
€ 5

€ 79

Symbolic Abstract Domain

represent each neuron as
a linear combination of the inputs
and the previous ReLUs

$$x_{i,j} \mapsto \begin{cases} \sum_{k=0}^{i-1} \mathbf{c}_k \cdot \mathbf{x}_k + \mathbf{c} & \mathbf{c}_k, \mathbf{c} \in \mathbb{R}^{|\mathbf{x}_k|} \\ [a,\ b] & a, b \in \mathbb{R} \end{cases}$$

$$\begin{array}{l} x_{i-1,0} \mapsto \mathbf{E_{i-1,0}} \\ \cdots \\ x_{i-1,j} \mapsto \mathbf{E_{i-1,j}} \\ \cdots \end{array} \xrightarrow{\quad x_{i,j} = \sum_k w_{j,k}^{i-1} \cdot x_{i-1,k} + b_{i,j} \quad} x_{i,j} \mapsto \sum_k w_{j,k}^{i-1} \cdot \mathbf{E_{i-1,k}} + b_{i,j}$$

$$x_{i,j} \mapsto \begin{cases} \mathbf{E_{i,j}} \\ \mathbf{[a,\ b]} \end{cases} \xrightarrow{\quad \text{ReLU} \quad} \begin{array}{ll} x_{i,j} \mapsto \begin{cases} \mathbf{E_{i,j}} \\ \mathbf{[a,\ b]} \end{cases} & 0 \leq a \\ x_{i,j} \mapsto \begin{cases} \mathbf{x_{i,j}} \\ \mathbf{[0,\ b]} \end{cases} & a < 0 \wedge 0 < b \\ x_{i,j} \mapsto \begin{cases} \mathbf{0} \\ \mathbf{[0,\ 0]} \end{cases} & b \leq 0 \end{array}$$

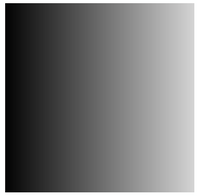
Symbolic Abstract Domain

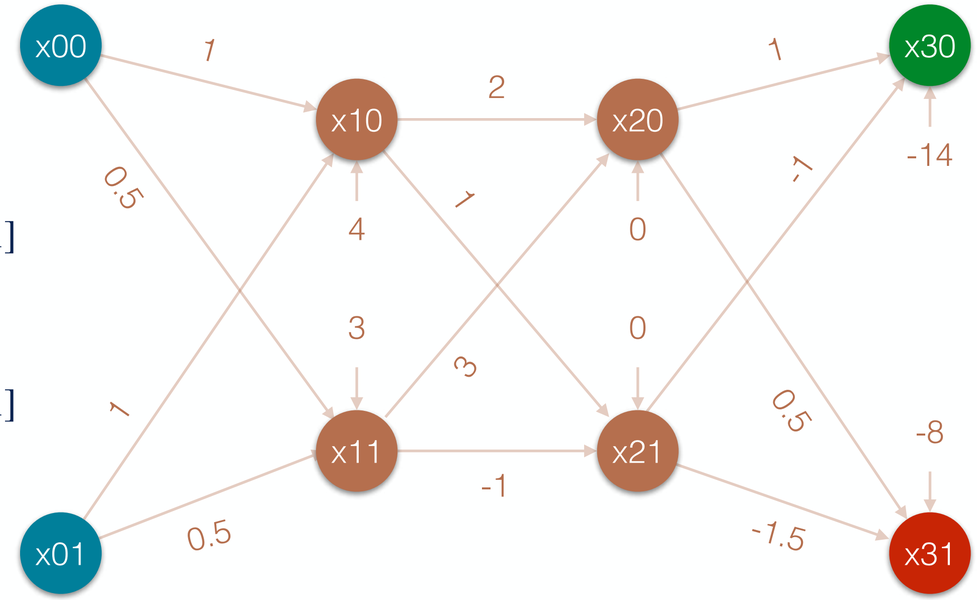
EXAMPLE

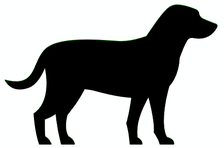
$$x_{10} \mapsto \begin{cases} x_{00} + x_{01} + 4 \\ [\mathbf{4}, \mathbf{6}] \end{cases} \qquad x_{20} \mapsto \begin{cases} 2 \cdot (x_{00} + x_{01} + 4) + 3 \cdot (0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3) \\ [\mathbf{17}, \mathbf{24}] \end{cases}$$

$$x_{00} \mapsto \begin{cases} x_{00} \\ [\mathbf{0}, \mathbf{1}] \end{cases}$$

$$x_{01} \mapsto \begin{cases} x_{01} \\ [\mathbf{0}, \mathbf{1}] \end{cases}$$

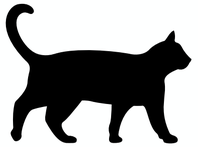






$$x_{30} \mapsto \begin{cases} 3 \cdot x_{00} + 3 \cdot x_{01} + 2 \\ [\underline{\mathbf{2}}, \mathbf{8}] \end{cases}$$

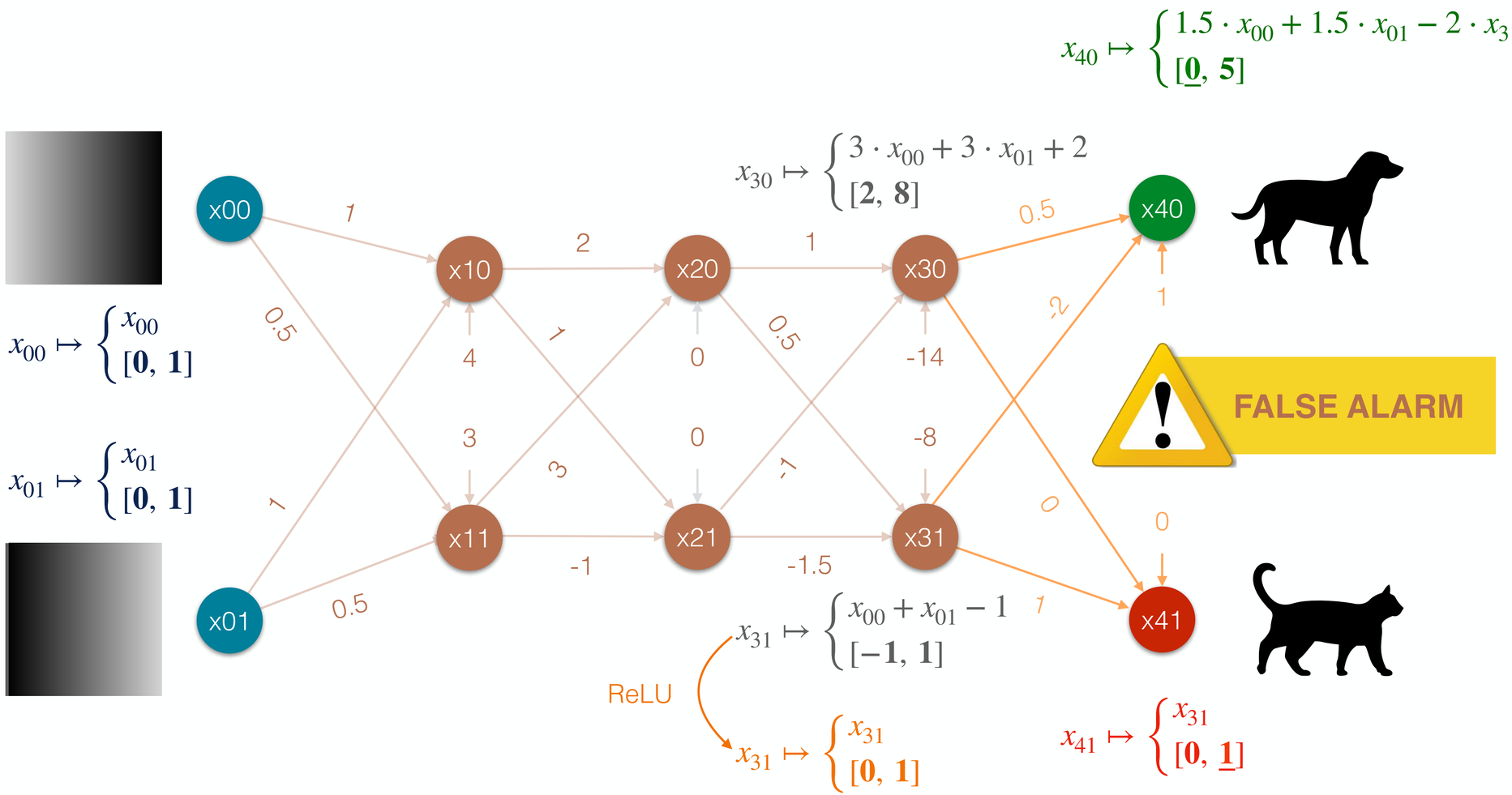
$$x_{31} \mapsto \begin{cases} x_{00} + x_{01} - 1 \\ [-\mathbf{1}, \underline{\mathbf{1}}] \end{cases}$$



$$x_{11} \mapsto \begin{cases} 0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3 \\ [\mathbf{3}, \mathbf{4}] \end{cases} \qquad x_{21} \mapsto \begin{cases} (x_{00} + x_{01} + 4) - 1 \cdot (0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3) \\ [\mathbf{1}, \mathbf{2}] \end{cases}$$

Symbolic Abstract Domain

MODIFIED EXAMPLE



Static Analysis by Abstract Interpretation

ABSTRACTION #3: DEEPPOLY ABSTRACT DOMAIN

PROPERTY OF INTEREST

$\mathcal{P}$

€ 10 +
€ 36 +
€ 28 +
€ 5

€ 79

DeepPoly Abstract Domain

maintain symbolic lower- and upper-bounds for each neuron + convex ReLU approximations

$$x_{i+1,j} \mapsto \begin{cases} [\sum_k c_{i,k} \cdot x_{i,k} + c,\ \sum_k d_{i,k} \cdot x_{i,k} + d] & c_{i,k}, c, d_{i,k}, d \in \mathbb{R} \\ [a,\ b] & a, b \in \mathbb{R} \end{cases}$$

$$x_{i,j} \mapsto \begin{cases} [\mathbf{L_{i,j}}, \mathbf{U_{i,j}}] \\ [\mathbf{a}, \mathbf{b}] \end{cases}$$

ReLU $0 \leq a$

$$x_{i,j} \mapsto \begin{cases} [\mathbf{L_{i,j}}, \mathbf{U_{i,j}}] \\ [\mathbf{a}, \mathbf{b}] \end{cases}$$

ReLU $a < 0 \wedge 0 < b$

$b \leq -a$

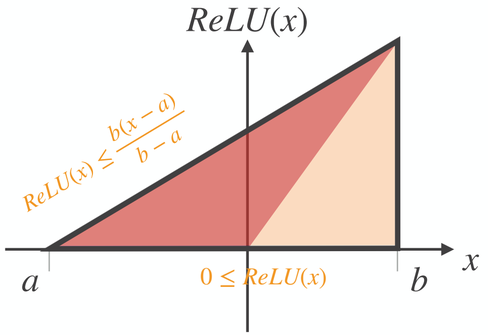
$$x_{i,j} \mapsto \begin{cases} \left[\mathbf{0}, \frac{\mathbf{b}(\mathbf{x_{i,j}} - \mathbf{a})}{\mathbf{b} - \mathbf{a}}\right] \\ [\mathbf{0}, \mathbf{b}] \end{cases}$$

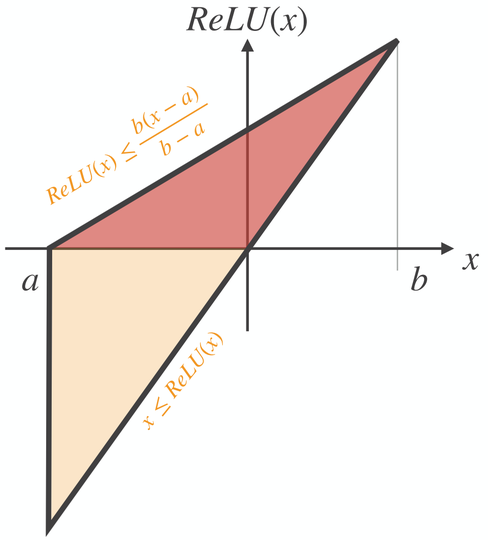
$-a < b$

$$x_{i,j} \mapsto \begin{cases} \left[\mathbf{x_{i,j}}, \frac{\mathbf{b}(\mathbf{x_{i,j}} - \mathbf{a})}{\mathbf{b} - \mathbf{a}}\right] \\ [\mathbf{a}, \mathbf{b}] \end{cases}$$

ReLU $b \leq 0$

$$x_{i,j} \mapsto \begin{cases} [\mathbf{0}, \mathbf{0}] \\ [0, 0] \end{cases}$$

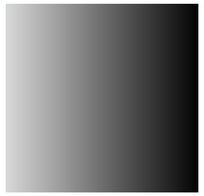




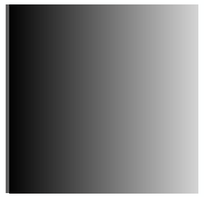
DeepPoly Abstract Domain

EXAMPLE

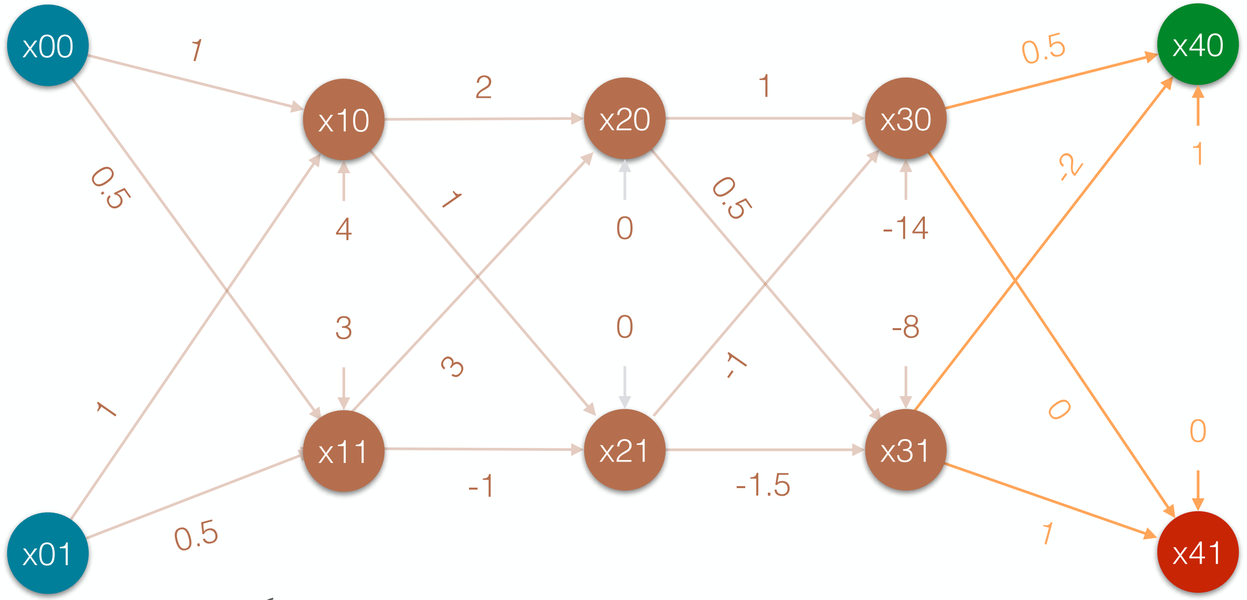
$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$



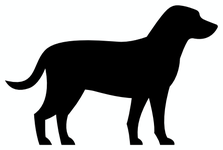
$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

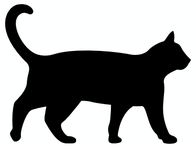


$$x_{10} \mapsto \begin{cases} [x_{00} + x_{01} + 4,\ x_{00} + x_{01} + 4] \\ [\mathbf{4},\ \mathbf{6}] \end{cases}$$



$$x_{11} \mapsto \begin{cases} [0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3,\ 0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3] \\ [\mathbf{3},\ \mathbf{4}] \end{cases}$$

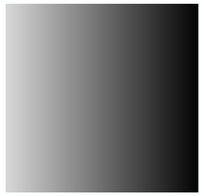




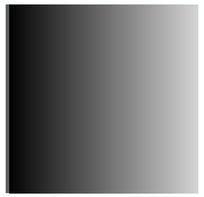
DeepPoly Abstract Domain

EXAMPLE

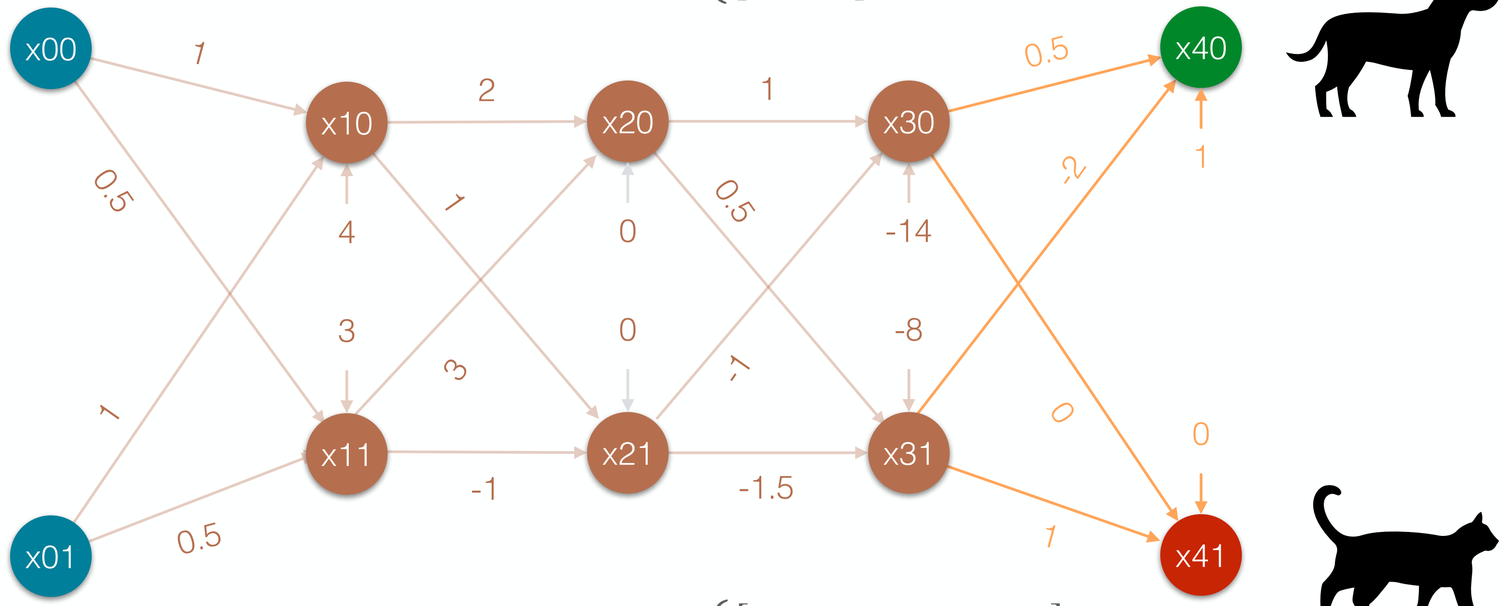
$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$



$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$



$$x_{20} \mapsto \begin{cases} [2 \cdot x_{10} + 3 \cdot x_{11},\ 2 \cdot x_{10} + 3 \cdot x_{11}] \\ [\mathbf{17},\ \mathbf{24}] \end{cases}$$

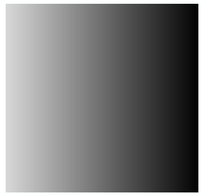


$$x_{21} \mapsto \begin{cases} [x_{10} - x_{11},\ x_{10} - x_{11}] \\ [\mathbf{1},\ \mathbf{2}] \end{cases}$$

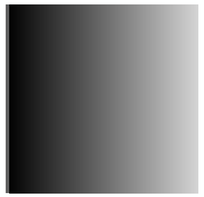
DeepPoly Abstract Domain

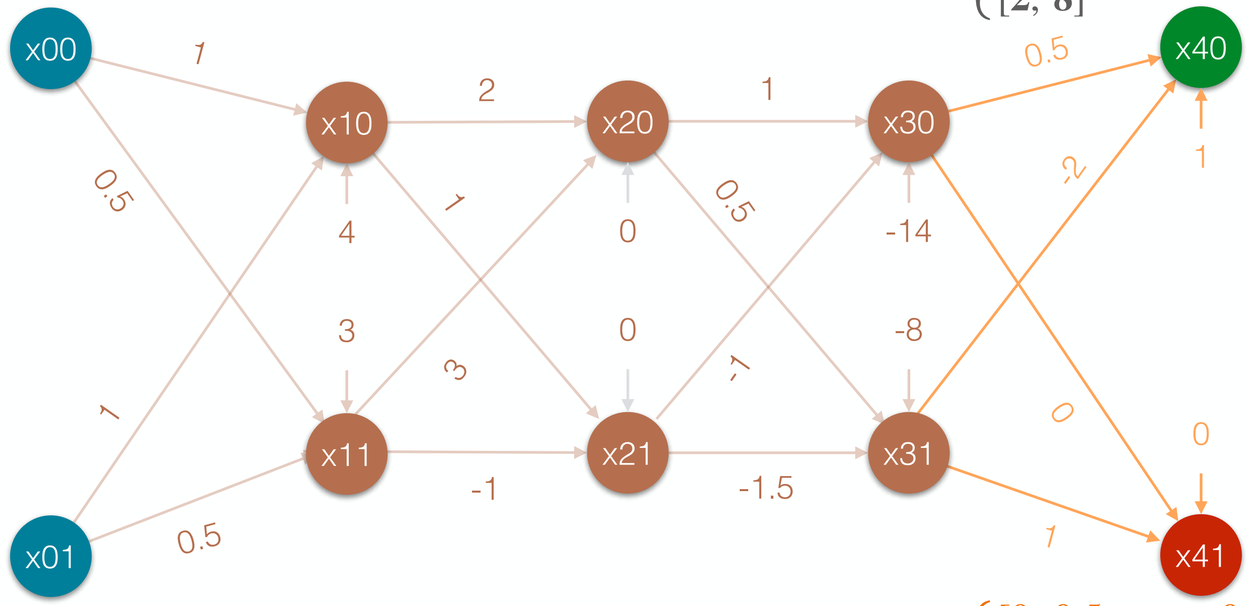
EXAMPLE

$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

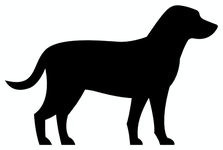


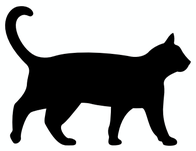
$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

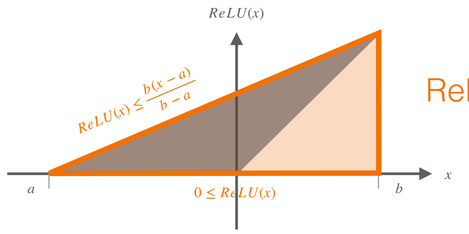




$$x_{30} \mapsto \begin{cases} [x_{20} - x_{21} - 14,\ x_{20} - x_{21} - 14] \\ [\mathbf{2},\ \mathbf{8}] \end{cases}$$







ReLU

$$x_{31} \mapsto \begin{cases} [0,\ 0.5 \cdot x_{31} + 0.5] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

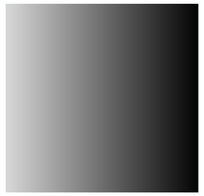
$$x_{31} \mapsto \begin{cases} [0.5 \cdot x_{20} - 1.5 \cdot x_{21} - 8,\ 0.5 \cdot x_{20} - 1.5 \cdot x_{21} - 8] \\ [-\mathbf{1},\ \mathbf{1}] \end{cases}$$

DeepPoly Abstract Domain

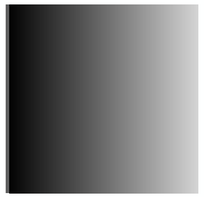
EXAMPLE

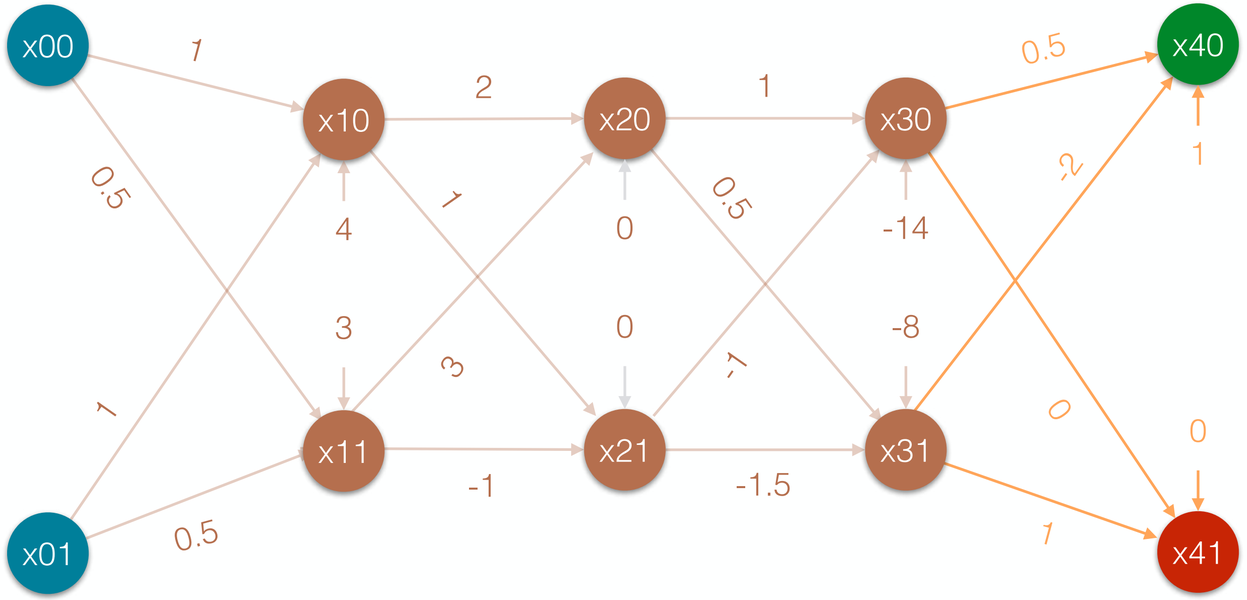
$$x_{40} \mapsto \begin{cases} [0.5 \cdot x_{30} - 2 \cdot x_{31} + 1,\ 0.5 \cdot x_{30} - 2 \cdot x_{31} + 1] \end{cases}$$

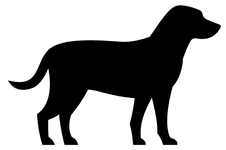
$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

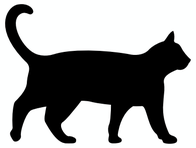


$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$









DeepPoly Abstract Domain

PARTIAL BACK-SUBSTITUTION

$x_{00} \mapsto [\mathbf{0}, \mathbf{1}]$

$x_{01} \mapsto [\mathbf{0}, \mathbf{1}]$

$$x_{10} \mapsto \begin{cases} [x_{00} + x_{01} + 4,\ x_{00} + x_{01} + 4] \\ [\mathbf{4}, \mathbf{6}] \end{cases}$$

$$x_{11} \mapsto \begin{cases} [0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3,\ 0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 3] \\ [\mathbf{3}, \mathbf{4}] \end{cases}$$

$$x_{20} \mapsto \begin{cases} [2 \cdot x_{10} + 3 \cdot x_{11},\ 2 \cdot x_{10} + 3 \cdot x_{11}] \\ [\mathbf{17}, \mathbf{24}] \end{cases}$$

$$x_{21} \mapsto \begin{cases} [x_{10} - x_{11},\ x_{10} - x_{11}] \\ [\mathbf{1}, \mathbf{2}] \end{cases}$$

$$x_{30} \mapsto \begin{cases} [x_{20} - x_{21} - 14,\ x_{20} - x_{21} - 14] \\ [\mathbf{2}, \mathbf{8}] \end{cases}$$

$$x_{31} \mapsto \begin{cases} [0,\ 0.5 \cdot (0.5 \cdot x_{20} - 1.5 \cdot x_{21} - 8) + 0.5] \\ [\mathbf{0}, \mathbf{1}] \end{cases}$$

$$x_{40} \mapsto \begin{cases} [0.5 \cdot x_{30} - 2 \cdot x_{31} + 1,\ 0.5 \cdot x_{30} - 2 \cdot x_{31} + 1] \\ [\underline{\mathbf{0}}, \mathbf{5}] \end{cases}$$

$$\mapsto \begin{cases} [x_{21} + 1,\ 0.5 \cdot x_{20} - 0.5 \cdot x_{21} - 6] \\ [\mathbf{2}, \underline{\mathbf{5.5}}] \end{cases}$$

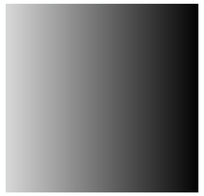
$$\mapsto \begin{cases} [x_{10} - x_{11} + 1,\ 0.5 \cdot x_{10} + 2 \cdot x_{11} - 6] \\ [\underline{\mathbf{1}}, \mathbf{5}] \end{cases}$$

$$\mapsto \begin{cases} [0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 2,\ 1.5 \cdot x_{00} + 1.5 \cdot x_{11} + 2] \\ [\mathbf{2}, \mathbf{5}] \end{cases}$$

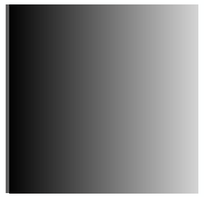
DeepPoly Abstract Domain

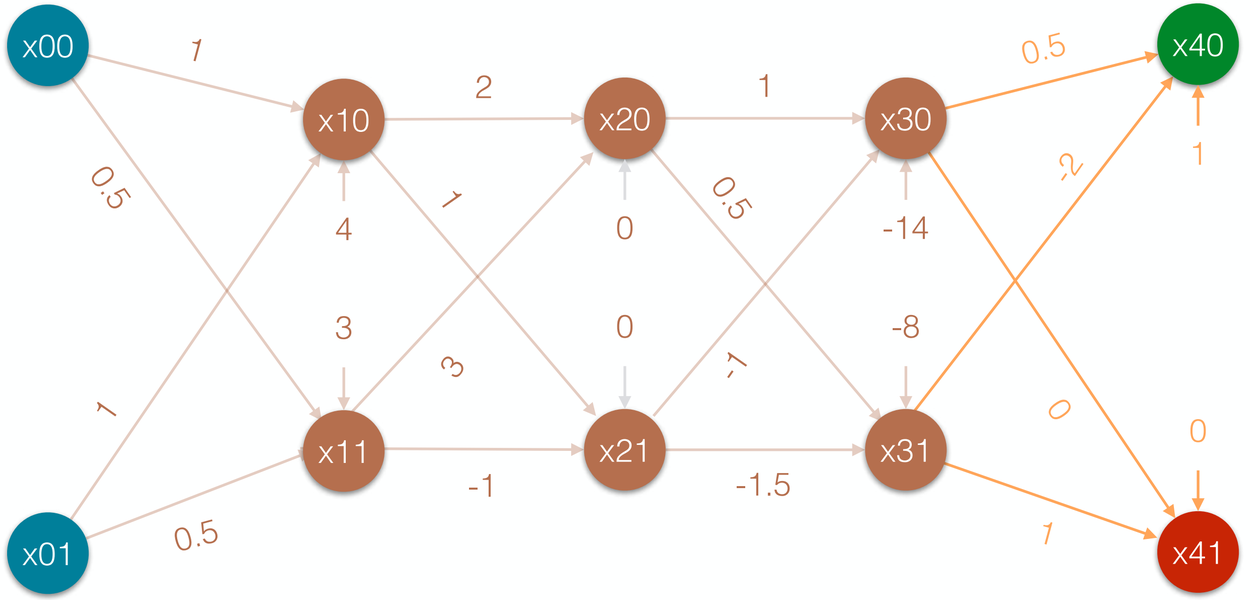
EXAMPLE

$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

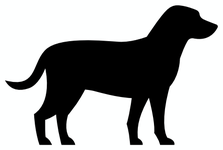


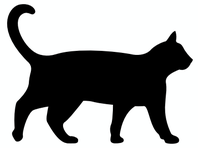
$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$





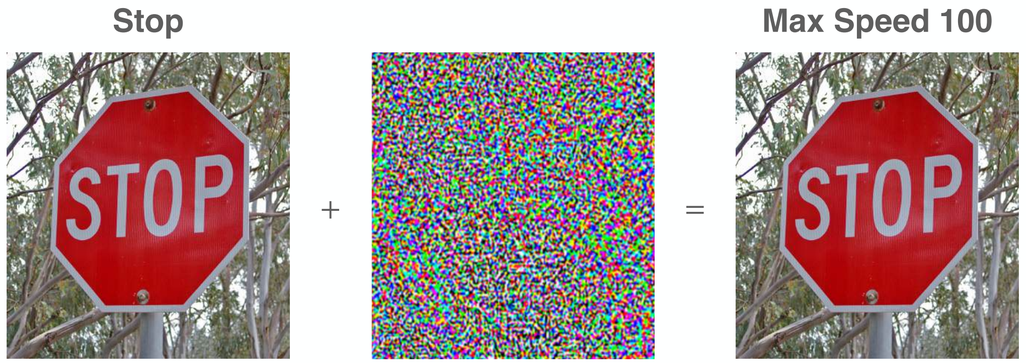
$$x_{40} \mapsto \begin{cases} [0.5 \cdot x_{30} - 2 \cdot x_{31} + 1,\ 0.5 \cdot x_{30} - 2 \cdot x_{31} + 1] \\ [\mathbf{2},\ \mathbf{5}] \end{cases}$$



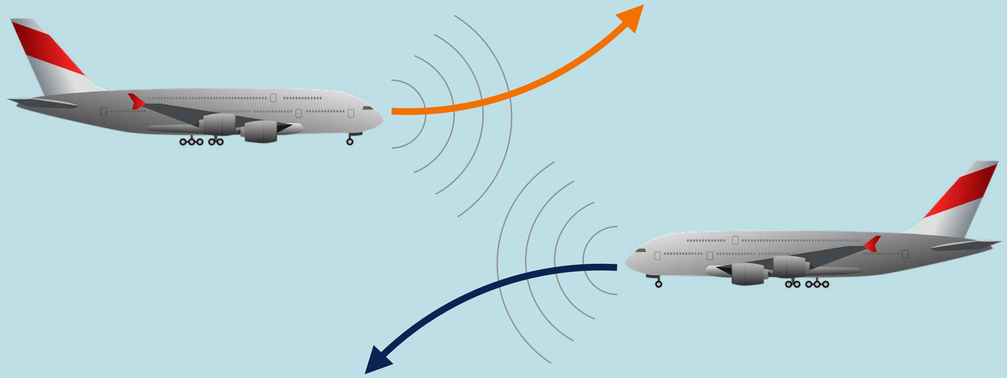


$$x_{41} \mapsto \begin{cases} [x_{31},\ x_{31}] \\ [\mathbf{0},\ \underline{\mathbf{1}}] \end{cases}$$

Stability

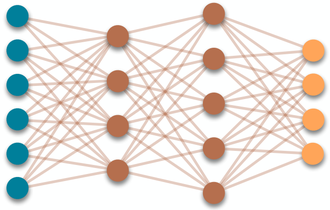


Safety



Safety

INPUT-OUTPUT PROPERTIES



f

(I, f, O)

O: output specification

I: input specification

v

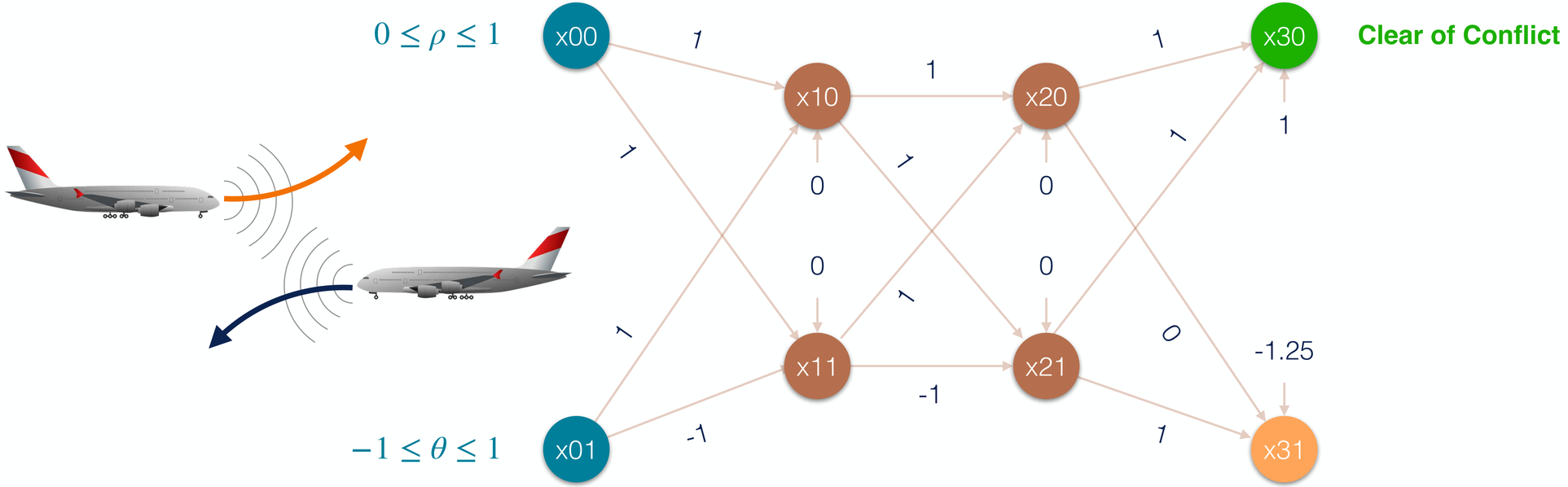
1 $\forall \mathbf{x} \models \mathbf{I} : f(\mathbf{x}) \models \mathbf{O}$

0 🤷‍♀️

-1 $\exists \mathbf{x}' \models \mathbf{I} : f(\mathbf{x}) \not\models \mathbf{O}$

Safety

EXAMPLE

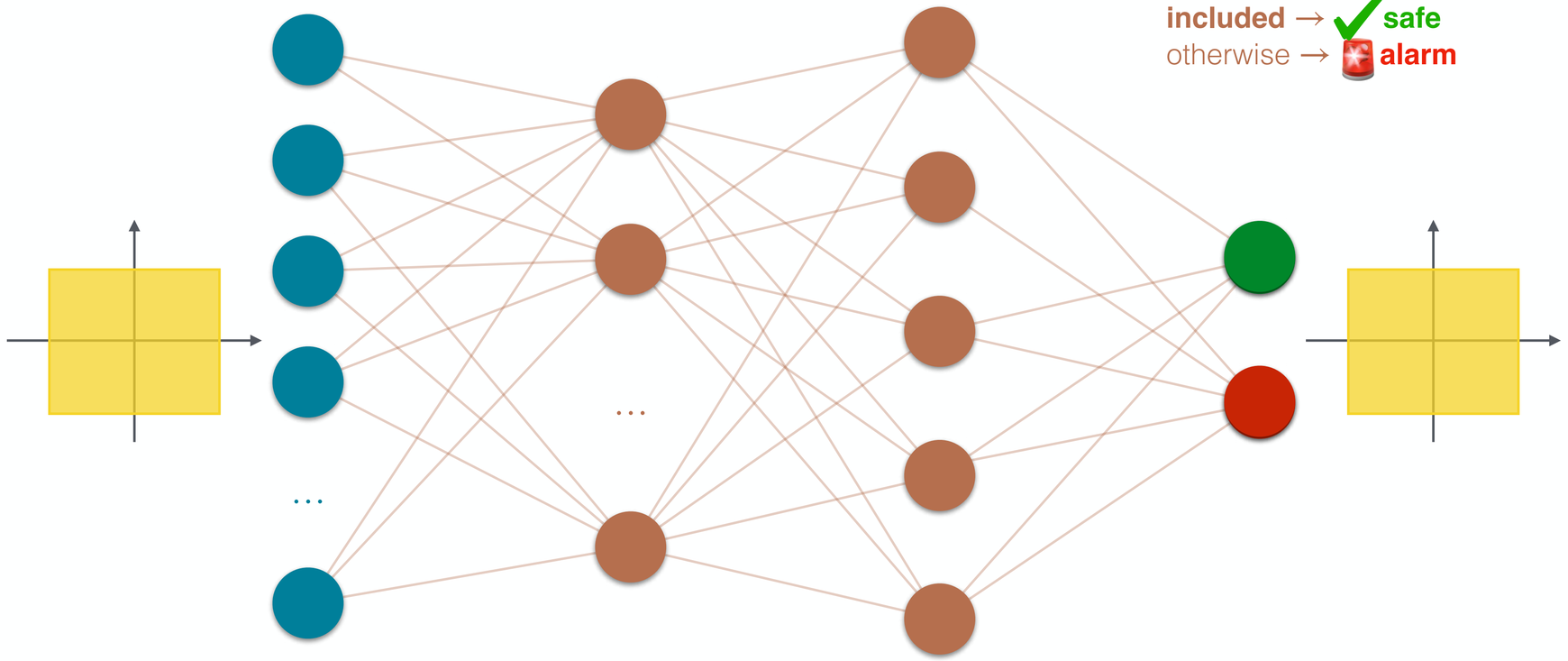


I **O**

Verifying Safety

STATIC FORWARD ANALYSIS

② check output for **inclusion** in **output specification** $\mathbf{O}$:
included → ✅ **safe**
otherwise → 🚨 **alarm**



① proceed **forwards from an abstraction** $\mathbf{I}^{\#}$ of $\mathbf{I}$

Static Analysis by Abstract Interpretation

ABSTRACTION #3: DEEPPOLY ABSTRACT DOMAIN

PROPERTY OF INTEREST

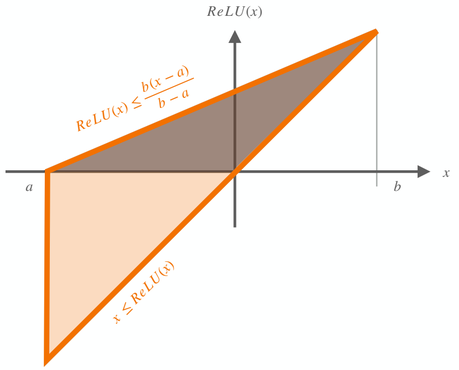
$\mathcal{P}$

€ 10 +
€ 36 +
€ 28 +
€ 5

€ 79

DeepPoly Abstract Domain

EXAMPLE



$$x_{10} \mapsto \begin{cases} [x_{10},\ \frac{2}{3} \cdot x_{10} + \frac{2}{3}] \\ [-\mathbf{1},\ \mathbf{2}] \end{cases}$$

ReLU

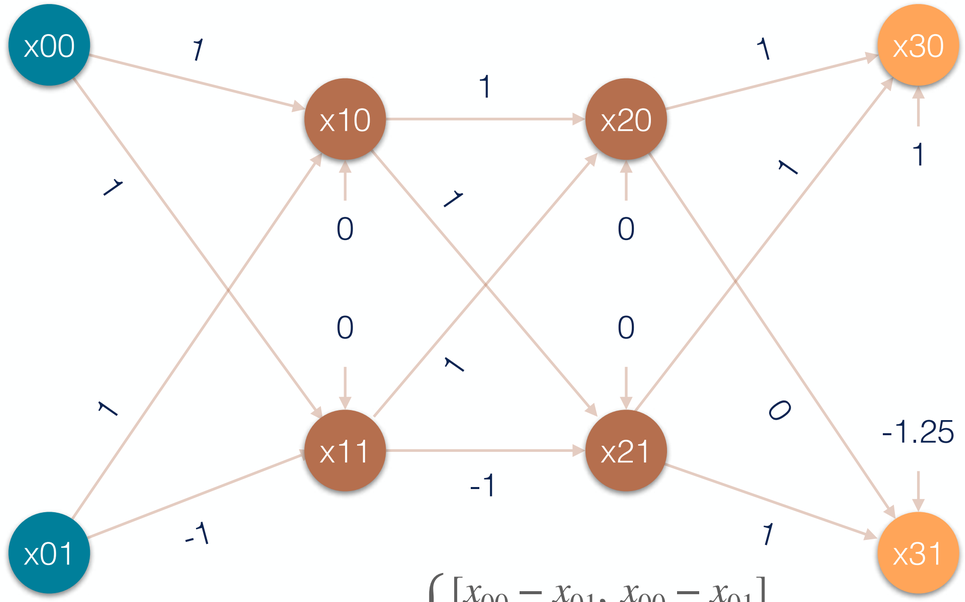
$$x_{10} \mapsto \begin{cases} [x_{00} + x_{01},\ x_{00} + x_{01}] \\ [-\mathbf{1},\ \mathbf{2}] \end{cases}$$

$0 \leq \rho \leq 1$

$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [-\mathbf{1},\ \mathbf{1}] \end{cases}$$

$-1 \leq \theta \leq 1$



Clear of Conflict

Strong Turn

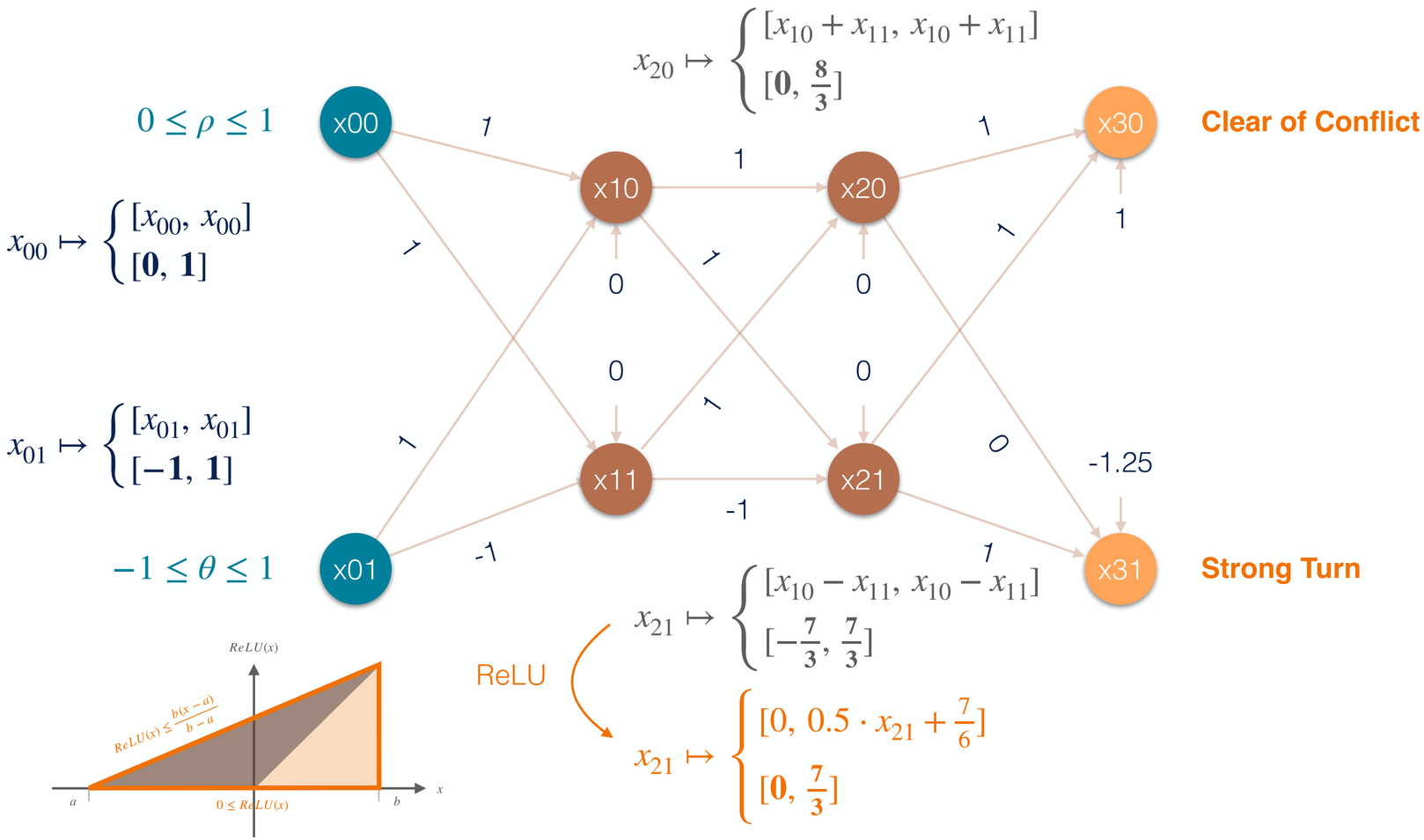
$$x_{11} \mapsto \begin{cases} [x_{00} - x_{01},\ x_{00} - x_{01}] \\ [-\mathbf{1},\ \mathbf{2}] \end{cases}$$

ReLU

$$x_{11} \mapsto \begin{cases} [x_{11},\ \frac{2}{3} \cdot x_{11} + \frac{2}{3}] \\ [-\mathbf{1},\ \mathbf{2}] \end{cases}$$

DeepPoly Abstract Domain

EXAMPLE



DeepPoly Abstract Domain

EXAMPLE

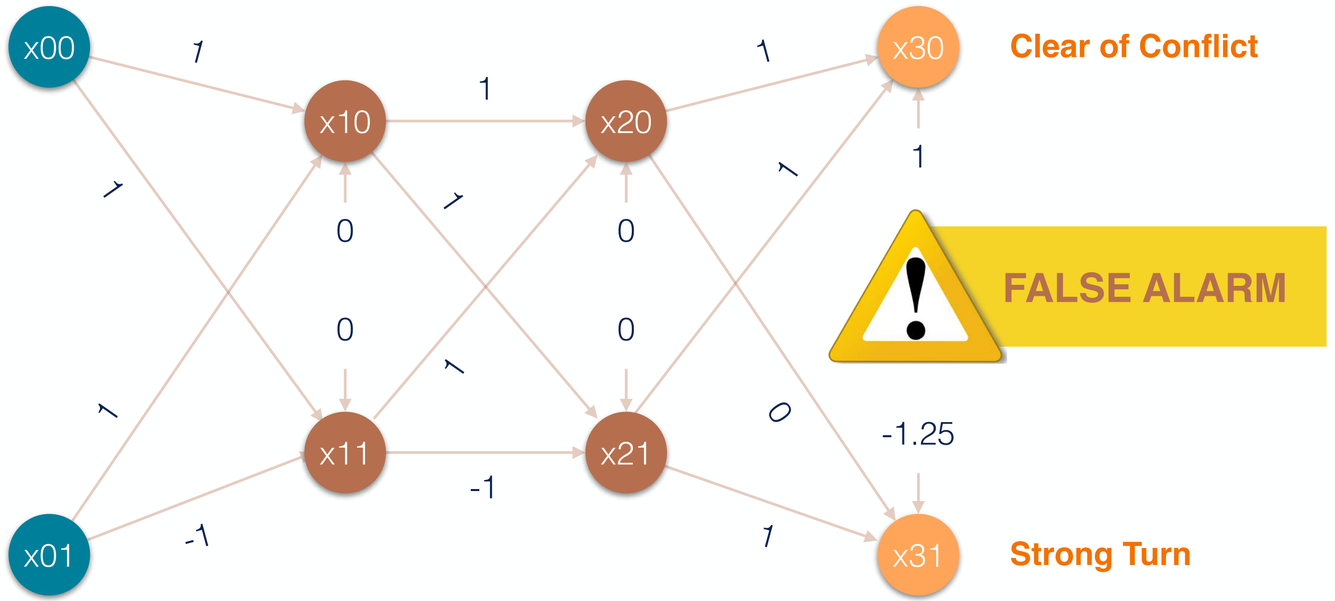
$$x_{30} \mapsto \begin{cases} [x_{20} + x_{21} + 1,\ x_{20} + x_{21} + 1] \\ [\underline{\mathbf{1}},\ \mathbf{5.5}] \end{cases}$$

$0 \leq \rho \leq 1$

$$x_{00} \mapsto \begin{cases} [x_{00},\ x_{00}] \\ [\mathbf{0},\ \mathbf{1}] \end{cases}$$

$$x_{01} \mapsto \begin{cases} [x_{01},\ x_{01}] \\ [\mathbf{-1},\ \mathbf{1}] \end{cases}$$

$-1 \leq \theta \leq 1$



$$x_{31} \mapsto \begin{cases} [x_{21} - 1.25,\ x_{21} - 1.25] \\ [\mathbf{-1.25},\ \underline{\frac{\mathbf{13}}{\mathbf{12}}}] \end{cases}$$

Static Analysis by Abstract Interpretation

ABSTRACTION #2: SYMBOLIC ABSTRACT DOMAIN

ABSTRACTION

$[\![P]\!]^\natural$

SOFTWARE

$[\![P]\!]$

PROPERTY OF INTEREST

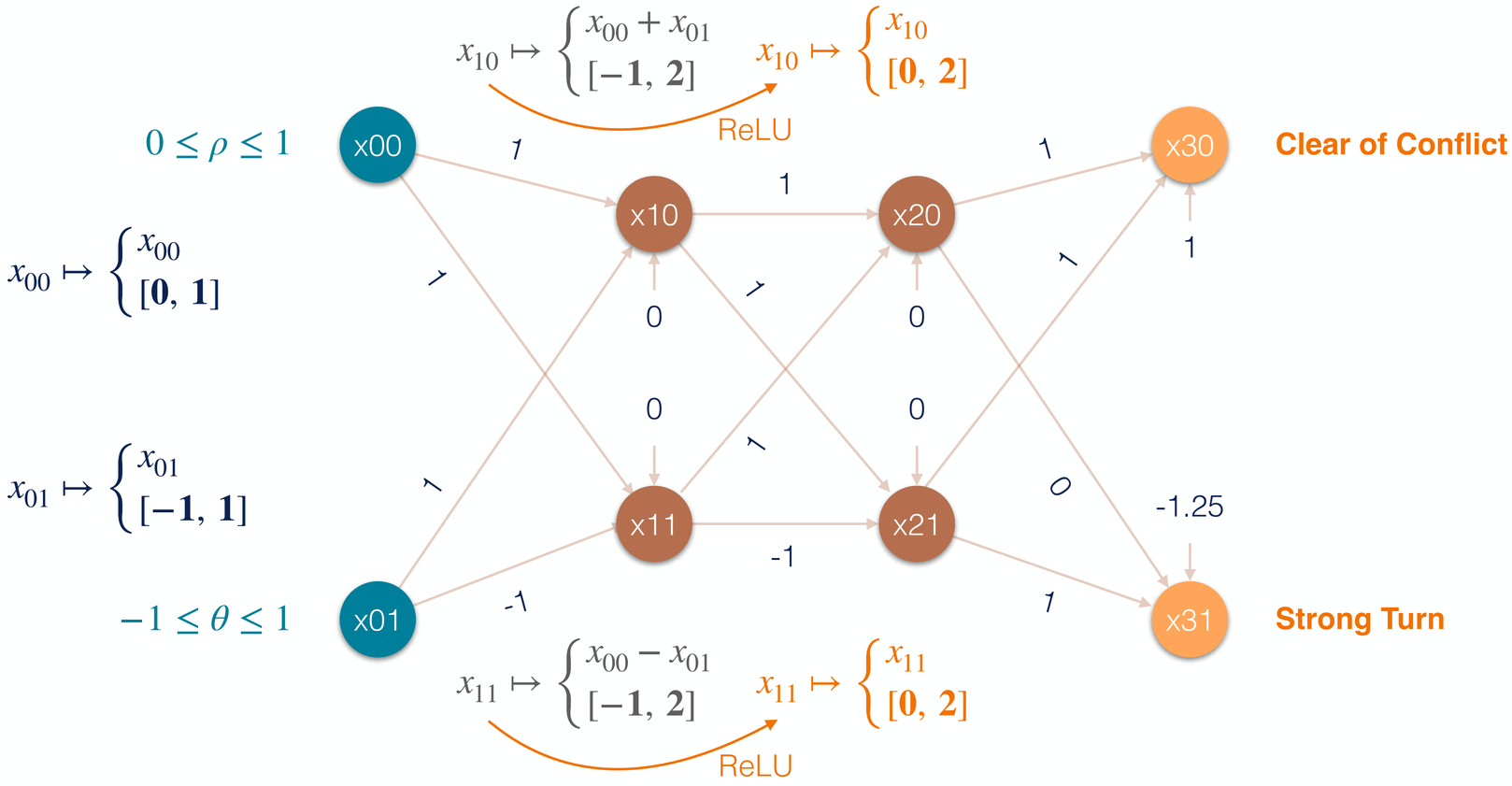
$\mathscr{P}$

€ 10 +
€ 36 +
€ 28 +
€ 5

€ 79

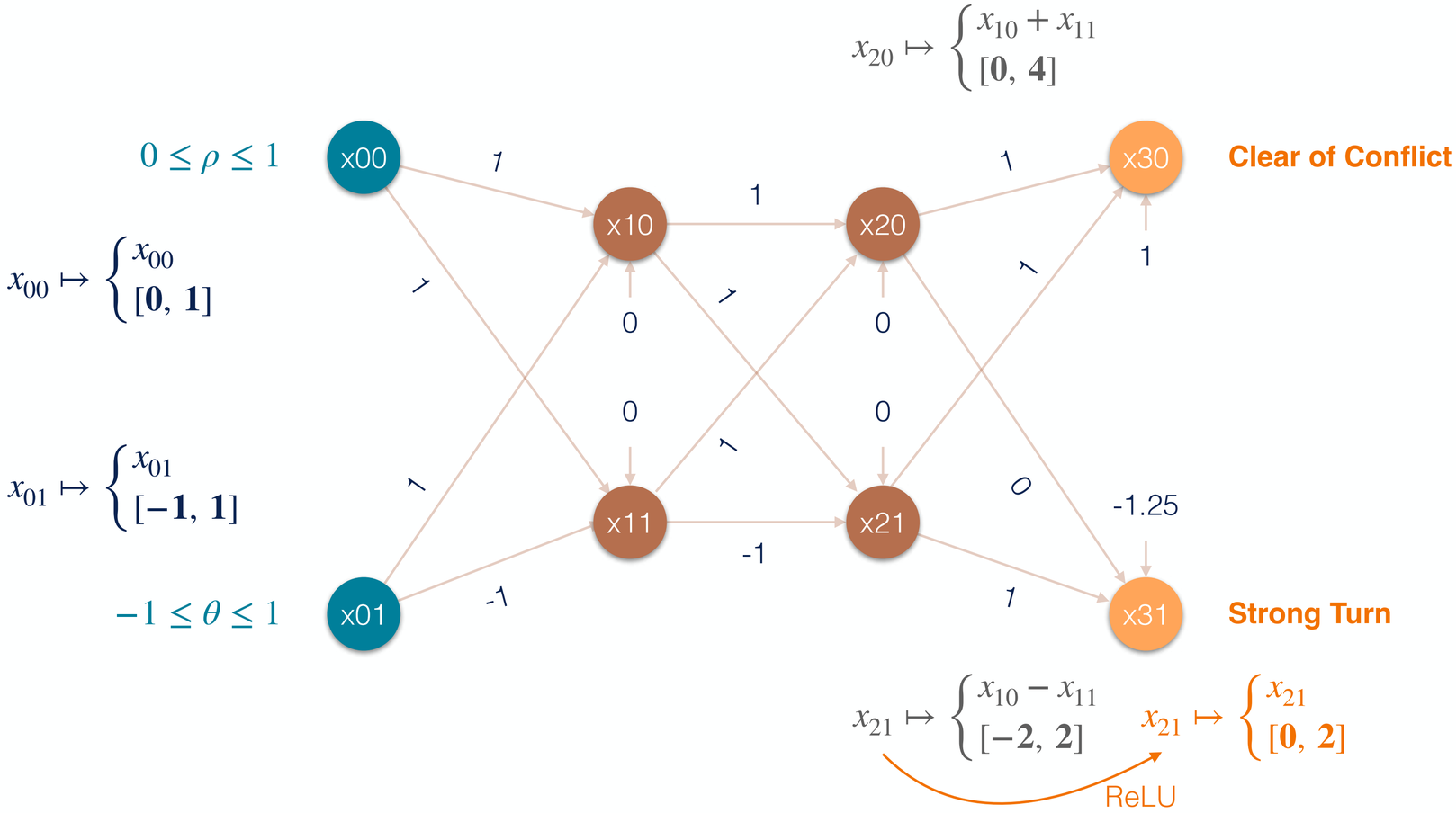
Symbolic Abstract Domain

EXAMPLE



Symbolic Abstract Domain

EXAMPLE



Symbolic Abstract Domain

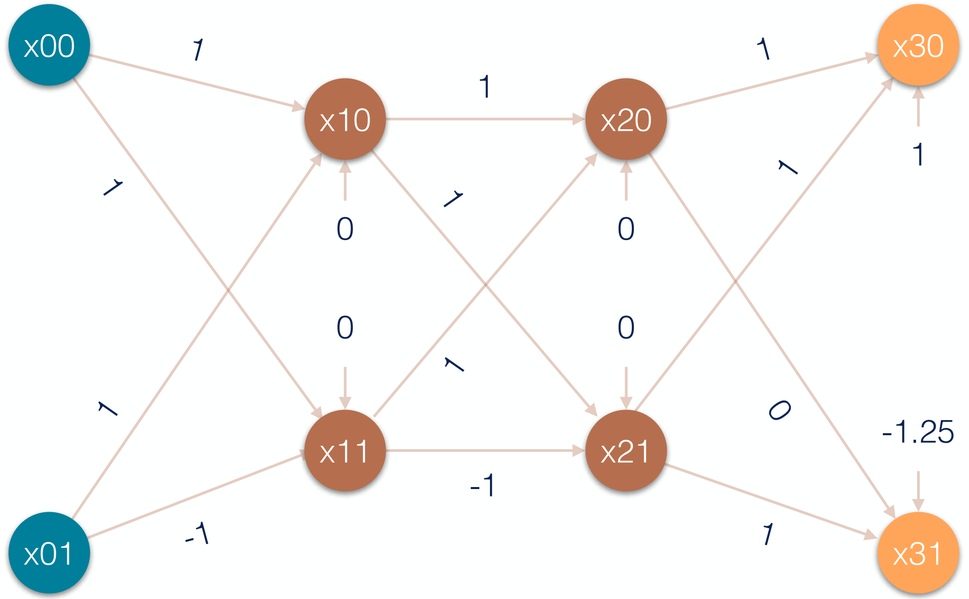
EXAMPLE

$0 \leq \rho \leq 1$

$x_{00} \mapsto \begin{cases} x_{00} \\ [\mathbf{0}, \mathbf{1}] \end{cases}$

$x_{01} \mapsto \begin{cases} x_{01} \\ [-\mathbf{1}, \mathbf{1}] \end{cases}$

$-1 \leq \theta \leq 1$



$x_{30} \mapsto \begin{cases} x_{10} + x_{11} + x_{21} + 1 \\ [\underline{\mathbf{1}}, \mathbf{7}] \end{cases}$ ← $[\underline{\mathbf{1}}, \mathbf{5.5}]$ with DeepPoly

Clear of Conflict

Strong Turn

$x_{31} \mapsto \begin{cases} x_{21} - 1.25 \\ [-\mathbf{1.25}, \underline{\mathbf{0.75}}] \end{cases}$ ← $[-\mathbf{1.25}, \underline{\frac{\mathbf{13}}{\mathbf{12}}}]$ with DeepPoly

Static Analysis by Abstract Interpretation

ABSTRACTION #4: REDUCED PRODUCT DOMAIN

PROPERTY OF INTEREST

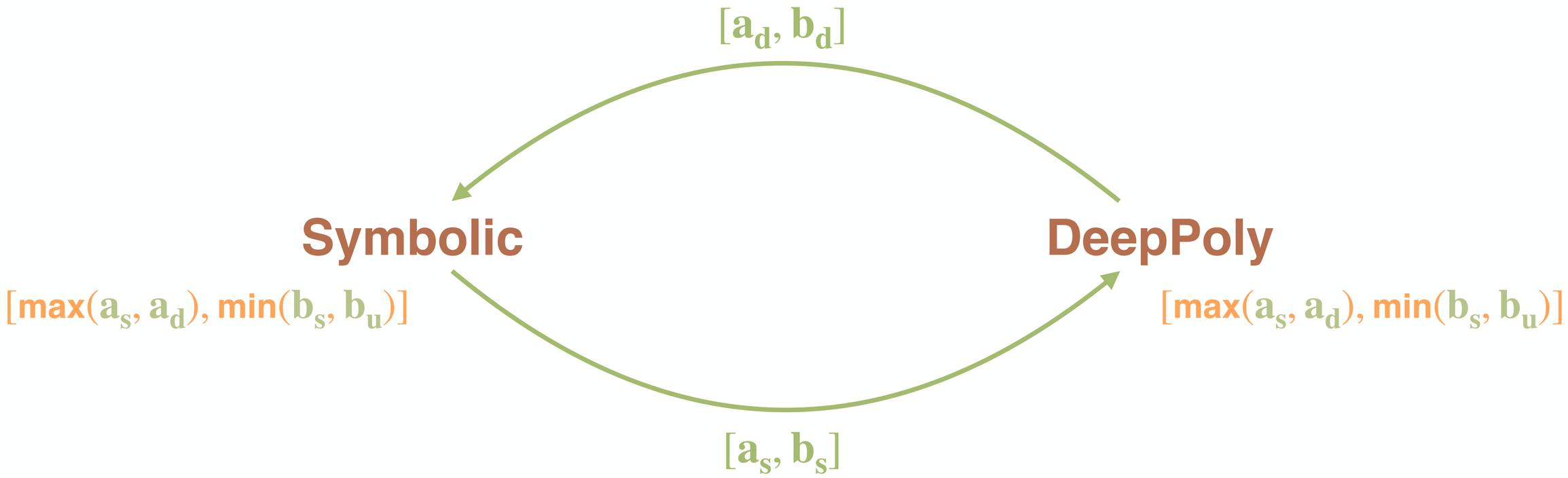
$\mathcal{P}$

€ 10 +
€ 36 +
€ 28 +
€ 5

€ 79

Reduced Product Domain

SYMBOLIC DOMAIN & DEEPPOLY DOMAIN



Reduced Product Domain

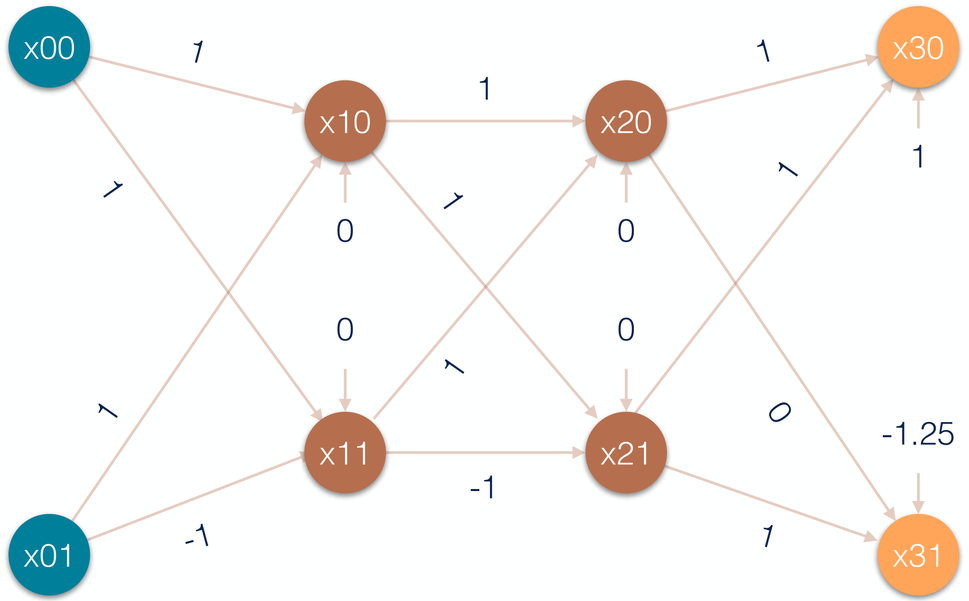
EXAMPLE

$0 \leq \rho \leq 1$

$$x_{00} \mapsto \begin{cases} x_{00} \\ [x_{00}, x_{00}] \\ [\mathbf{0}, \mathbf{1}] \end{cases}$$

$$x_{01} \mapsto \begin{cases} x_{01} \\ [x_{01}, x_{01}] \\ [\mathbf{-1}, \mathbf{1}] \end{cases}$$

$-1 \leq \theta \leq 1$

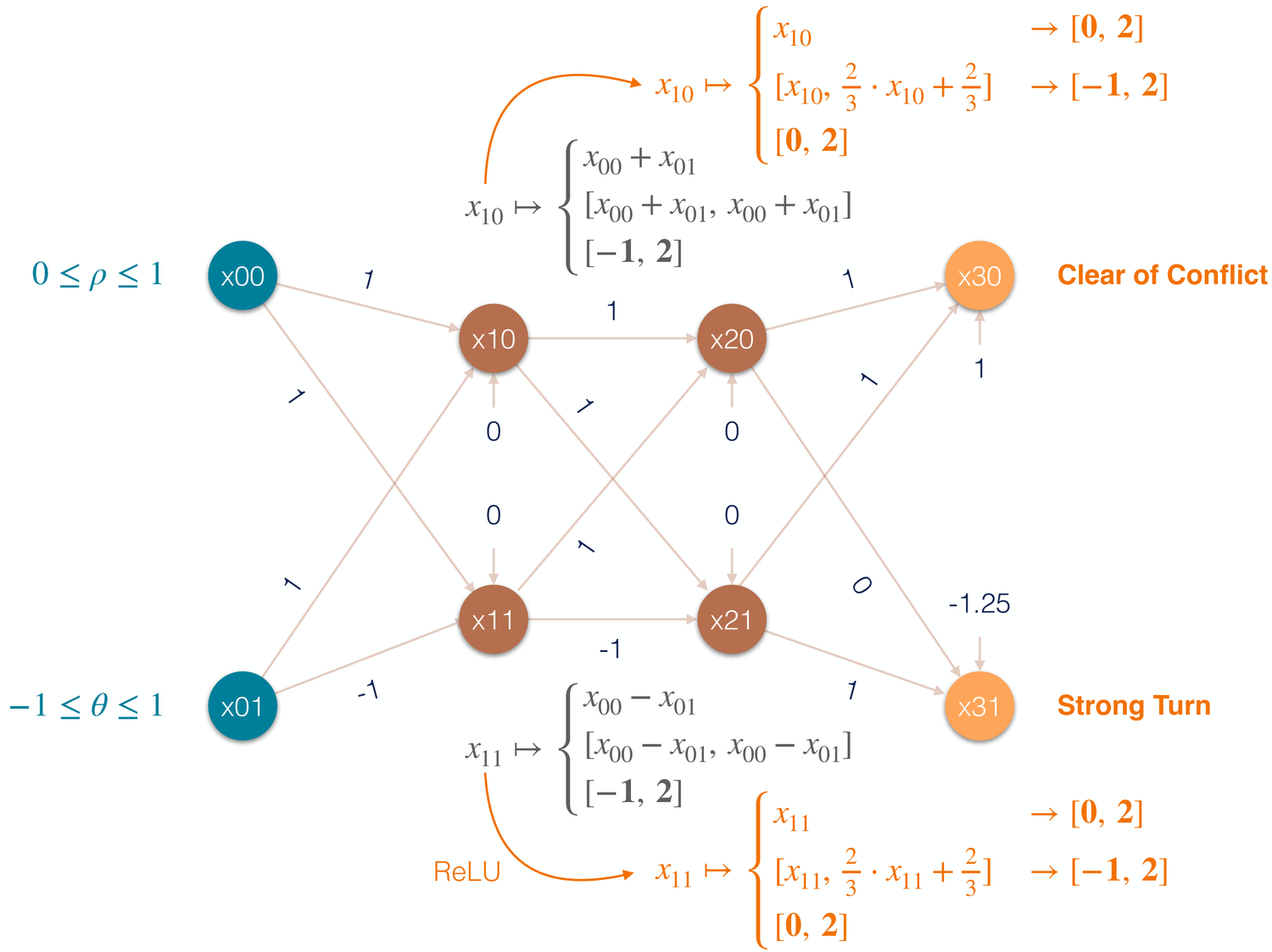


Clear of Conflict

Strong Turn

Reduced Product Domain

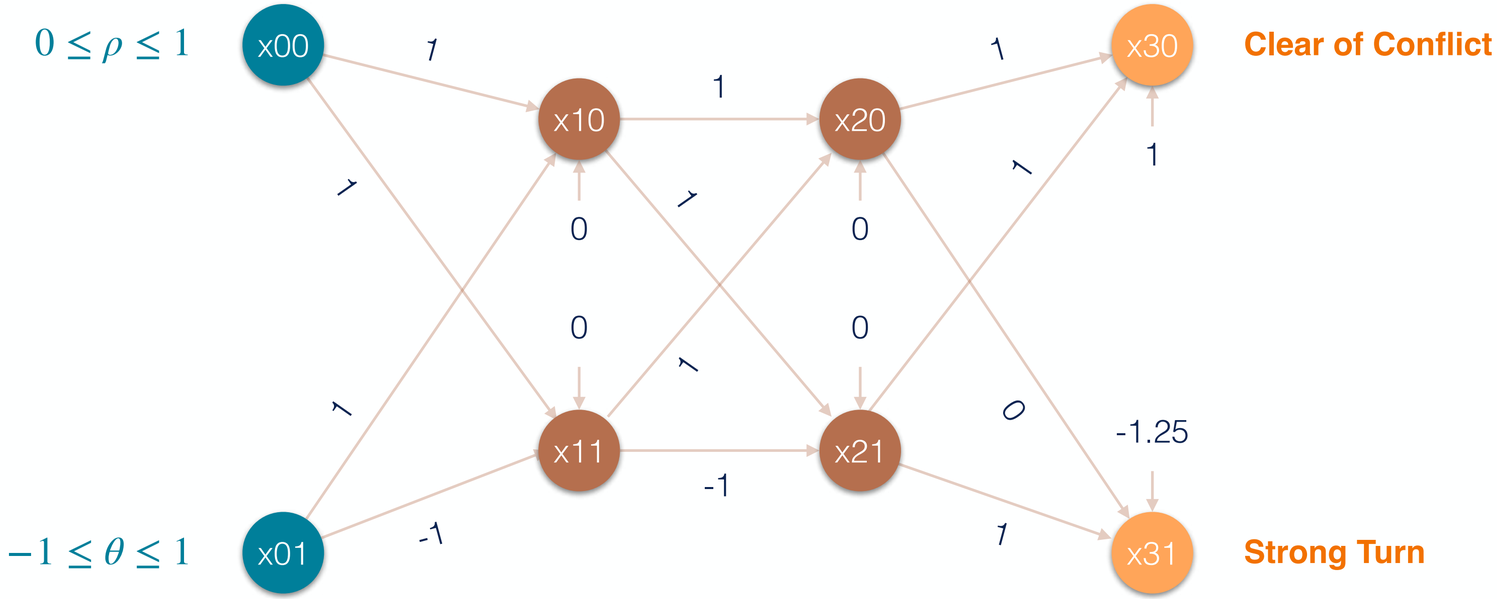
EXAMPLE



Reduced Product Domain

EXAMPLE

$$x_{20} \mapsto \begin{cases} x_{10} + x_{11} & \to [\mathbf{0}, \mathbf{4}] \\ [x_{10} + x_{11},\ x_{10} + x_{11}] & \to [\mathbf{0}, \frac{\mathbf{8}}{\mathbf{3}}] \\ [\mathbf{0}, \frac{\mathbf{8}}{\mathbf{3}}] \end{cases}$$



$$x_{21} \mapsto \begin{cases} x_{10} - x_{11} & \to [-\mathbf{2}, \mathbf{2}] \\ [x_{10} - x_{11},\ x_{10} - x_{11}] & \to [-\frac{\mathbf{7}}{\mathbf{3}}, \frac{\mathbf{7}}{\mathbf{3}}] \\ [-\mathbf{2}, \mathbf{2}] \end{cases}$$

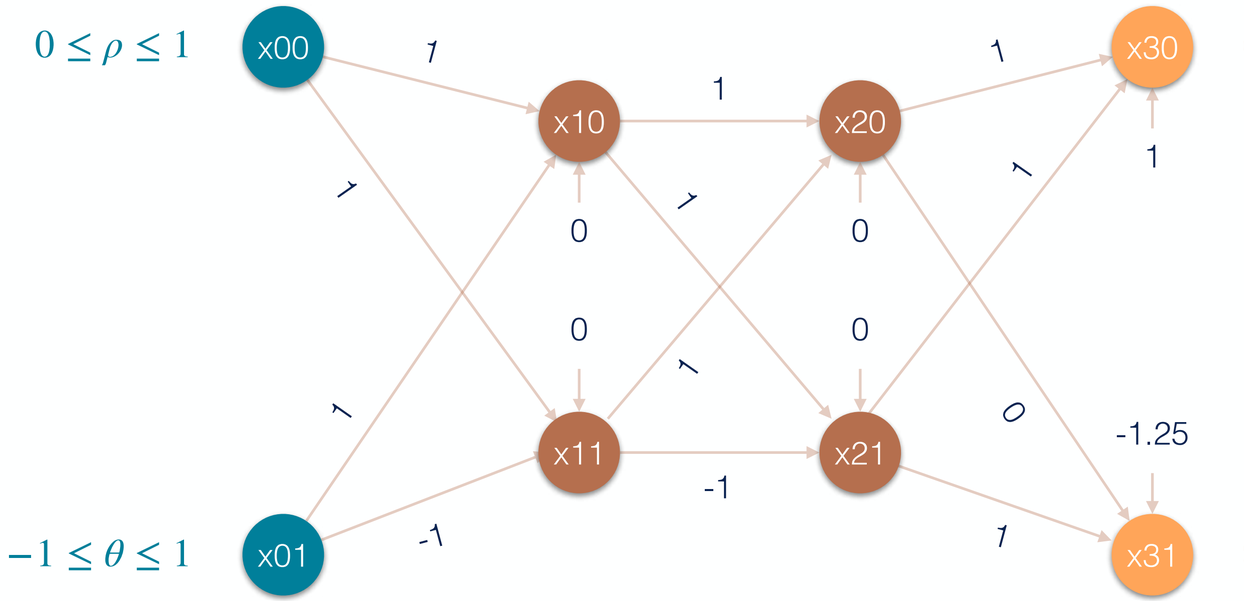
ReLU

$$x_{21} \mapsto \begin{cases} x_{21} & \to [\mathbf{0}, \mathbf{2}] \\ [0,\ 0.5 \cdot x_{21} + 1] & \to [\mathbf{0}, \mathbf{2}] \\ [\mathbf{0}, \mathbf{2}] \end{cases}$$

Reduced Product Domain

EXAMPLE

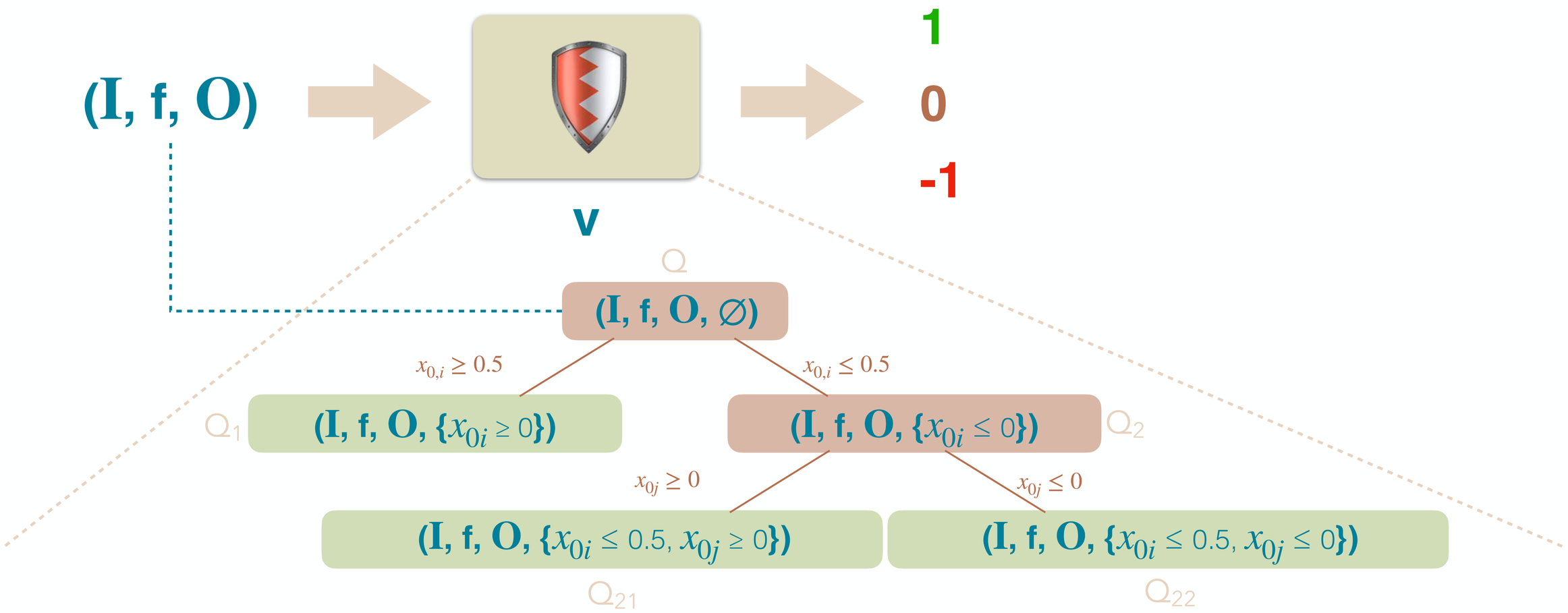
$$x_{30} \mapsto \begin{cases} x_{10} + x_{11} + x_{21} + 1 & \to [\mathbf{1}, \frac{\mathbf{20}}{\mathbf{3}}] \\ [x_{20} + x_{21} + 1,\ x_{20} + x_{21} + 1] & \to [\mathbf{1}, \frac{\mathbf{16}}{\mathbf{3}}] \\ [\underline{\mathbf{1}}, \frac{\mathbf{16}}{\mathbf{3}}] \end{cases}$$



$$x_{31} \mapsto \begin{cases} x_{21} - 1.25 & \to [-\mathbf{1.25}, \mathbf{0.75}] \\ [x_{21} - 1.25,\ x_{21} - 1.25] & \to [-\mathbf{1.25}, \mathbf{0.75}] \\ [-\mathbf{1.25}, \underline{\mathbf{0.75}}] \end{cases}$$

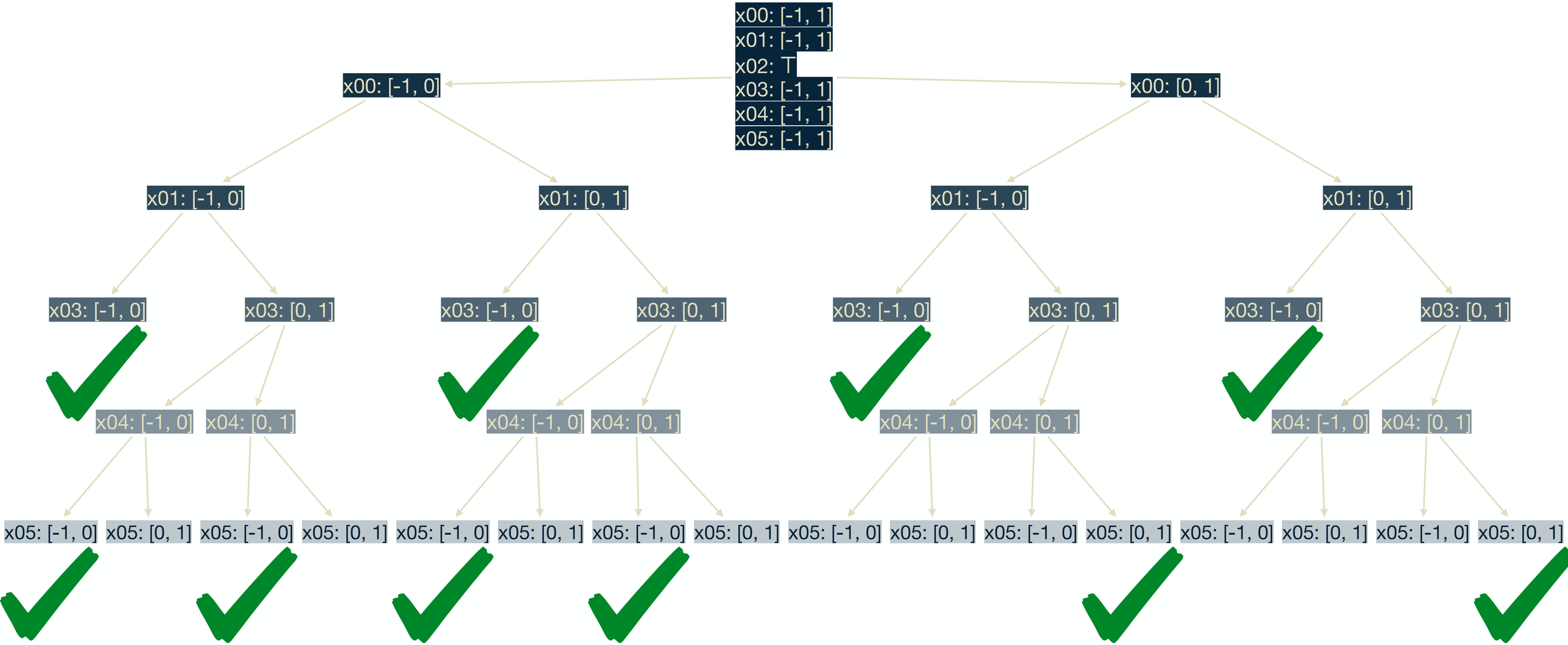
Neural Network Verification

GOING FARTHER: BRANCH-AND-BOUND (BAB) WITH INPUT SPLITTING



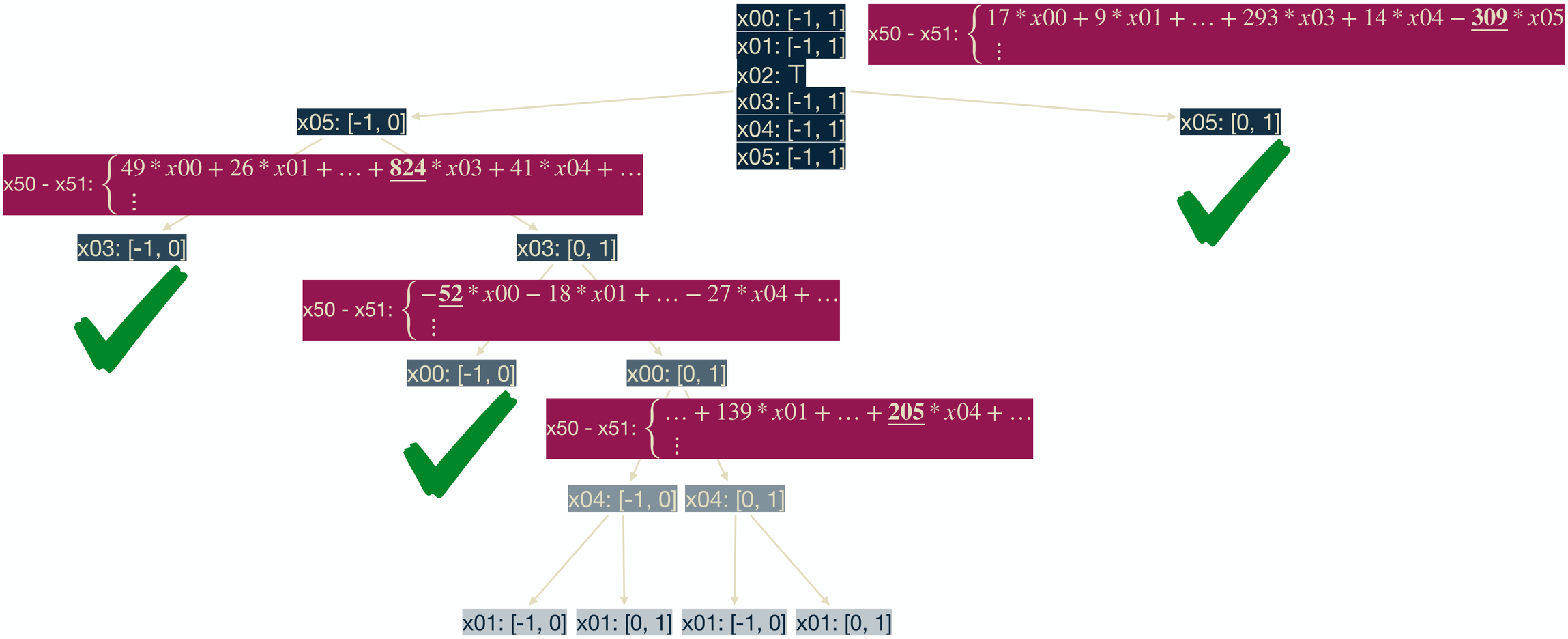
BaB with Input Splitting

PARTITIONING STRATEGY #1: LARGEST RANGE



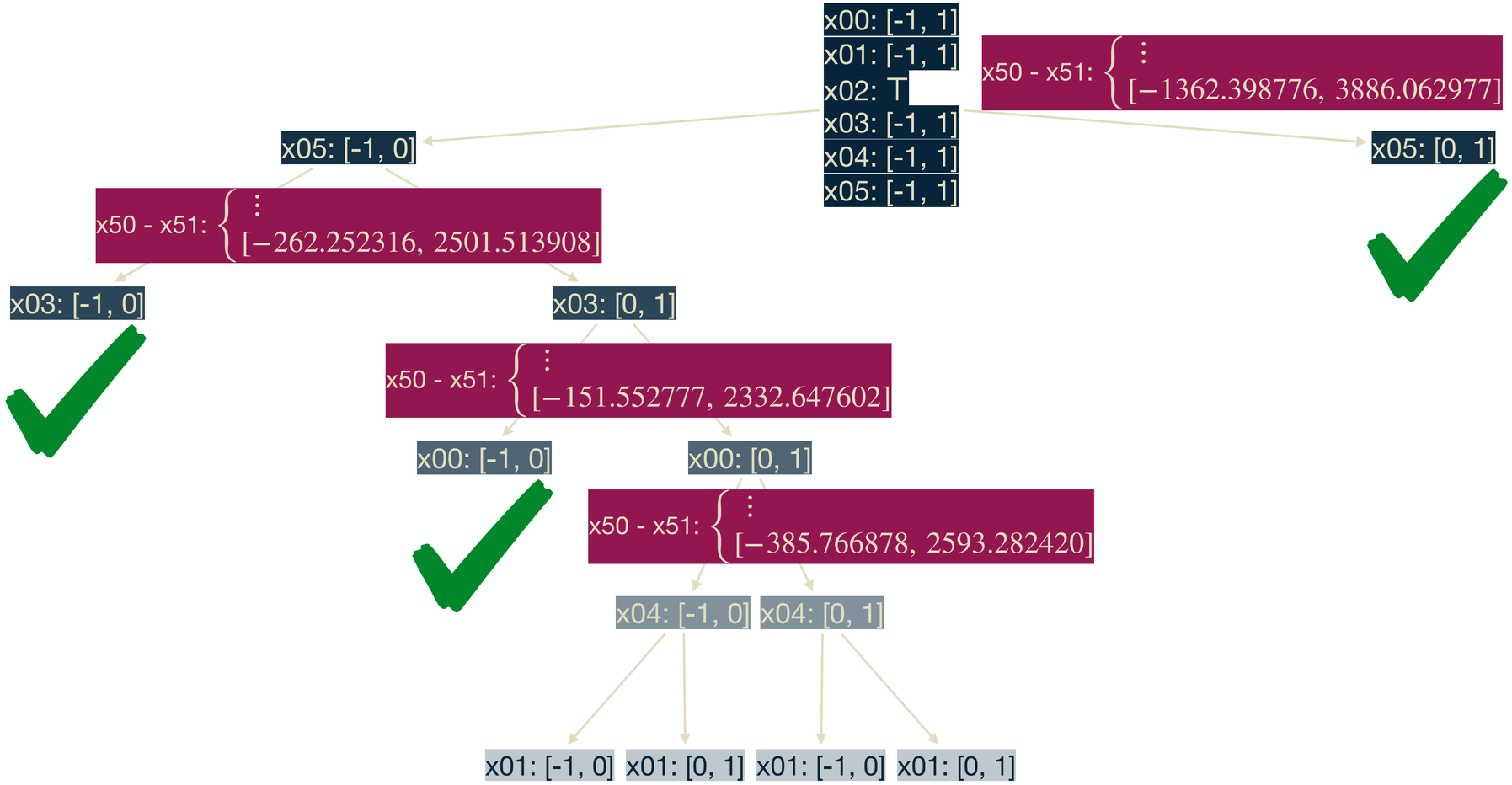
BaB with Input Splitting

PARTITIONING STRATEGY #2: LARGEST COEFFICIENT



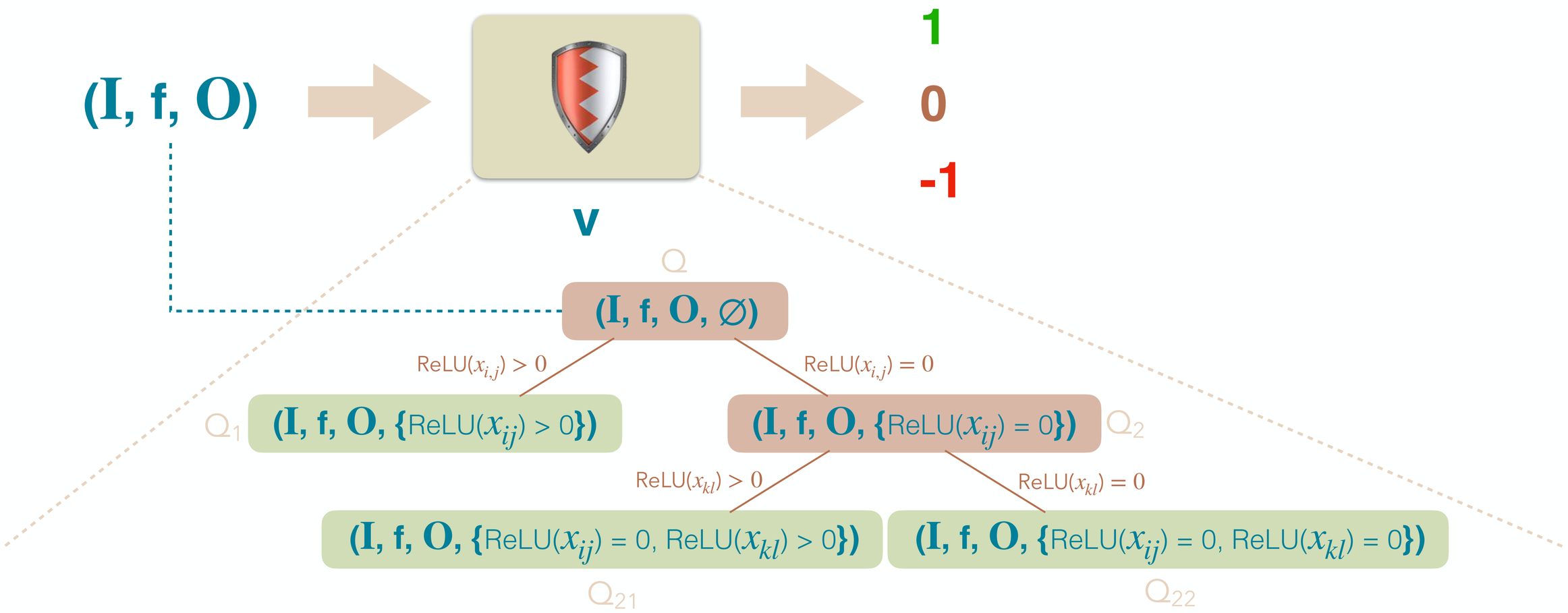
DeePoly Abstract Domain

BAB WITH INPUT SPLITTING ⇏ OUTPUT REFINEMENT



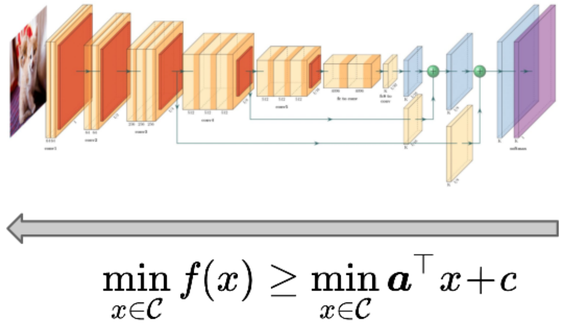
Neural Network Verification

GOING FARTHER: BRANCH-AND-BOUND (BAB) WITH RELU SPLITTING

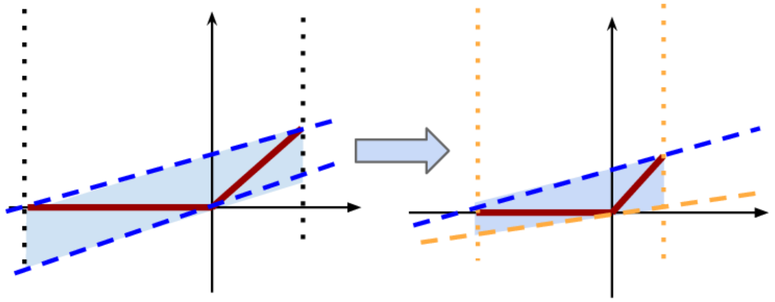


$\alpha\beta$-CROWN

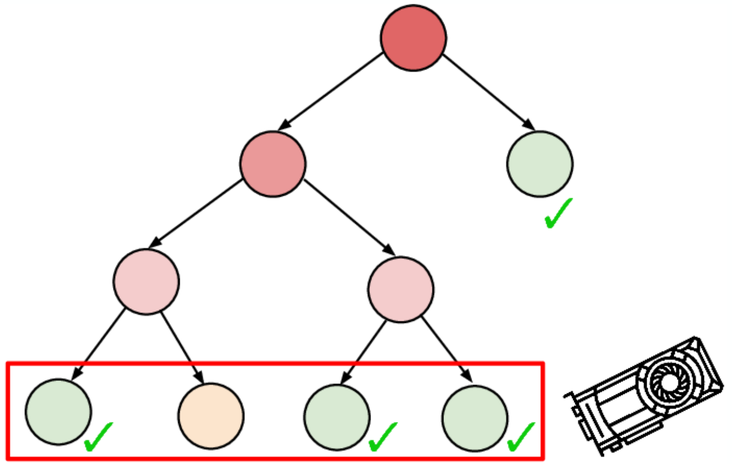
THE STATE OF THE ART



Efficient bound propagation (**CROWN**)



GPU optimized relaxation (***α*-CROWN**)



Parallel branch and bound (***β*-CROWN**)

Winner of the International Verification of Neural Networks Competition since 2021

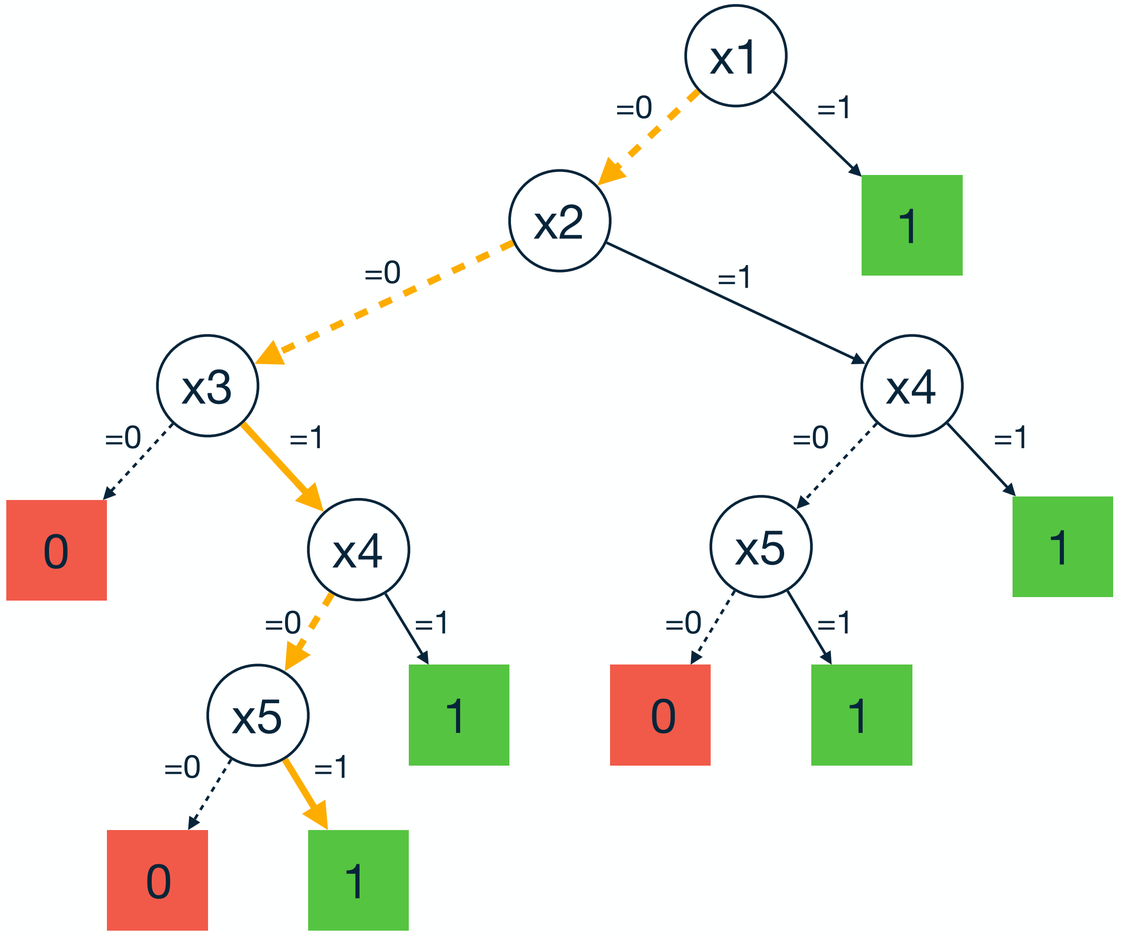
Static Analysis for Trained Models

Verification

Explainability

Abductive Explanation (AXp)

SUBSET-MINIMAL SET OF FEATURES SUFFICIENT TO ENSURE A PREDICTION



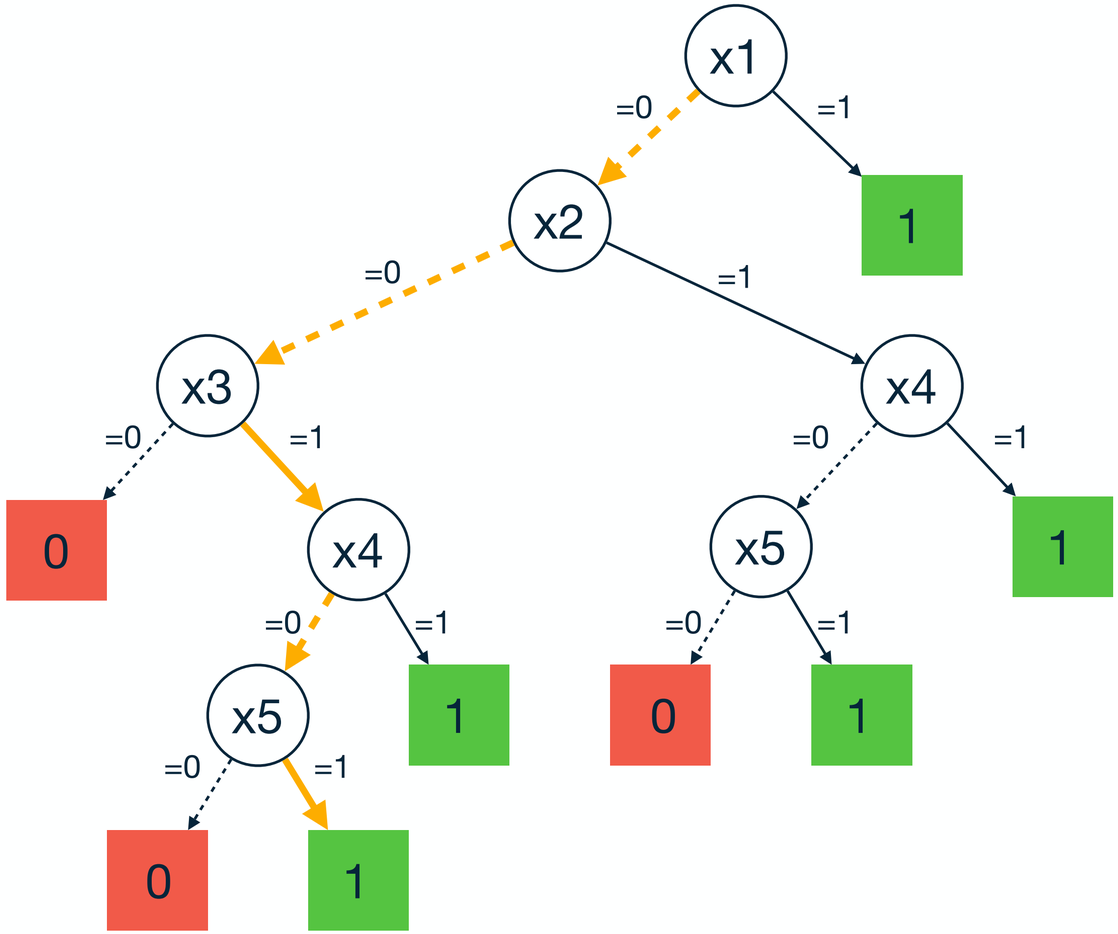
AXp = { x3, x5 }

x3	x5	x1	x2	x4		
1	1	0	0	0	→	1
1	1	0	0	1	→	1
1	1	0	1	0	→	1
1	1	0	1	1	→	1
1	1	1	0	0	→	1
1	1	1	0	1	→	1
1	1	1	1	0	→	1
1	1	1	1	1	→	1

Computing One AXp

DROP (I.E., FREE) INPUT FEATURES WHILE AXp CONDITION HOLDS

SAME PREDICTION



{ 1, 2, 3, 4, 5 } → 1

Free 1: { 2, 3, 4, 5 } → 1

Free 2: { 3, 4, 5 } → 1

Free 3: { 4, 5 } →

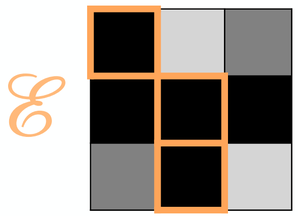
Free 4: { 3, 5 } → 1

Free 5: { 3 } →

AXp = { x3, x5 }

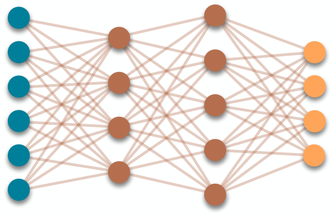
Optimal Robust Explanations

ABDUCTIVE EXPLANATIONS (AXps)



$\mathscr{E}$

$\mathbf{x} \in \mathbb{R}^d$



f

4

$(\mathbf{x}, \overline{\mathscr{E}}$

$\forall i \in \mathscr{E} \cdot$ $\cup \{i\}, \epsilon, \mathbf{f})$

v

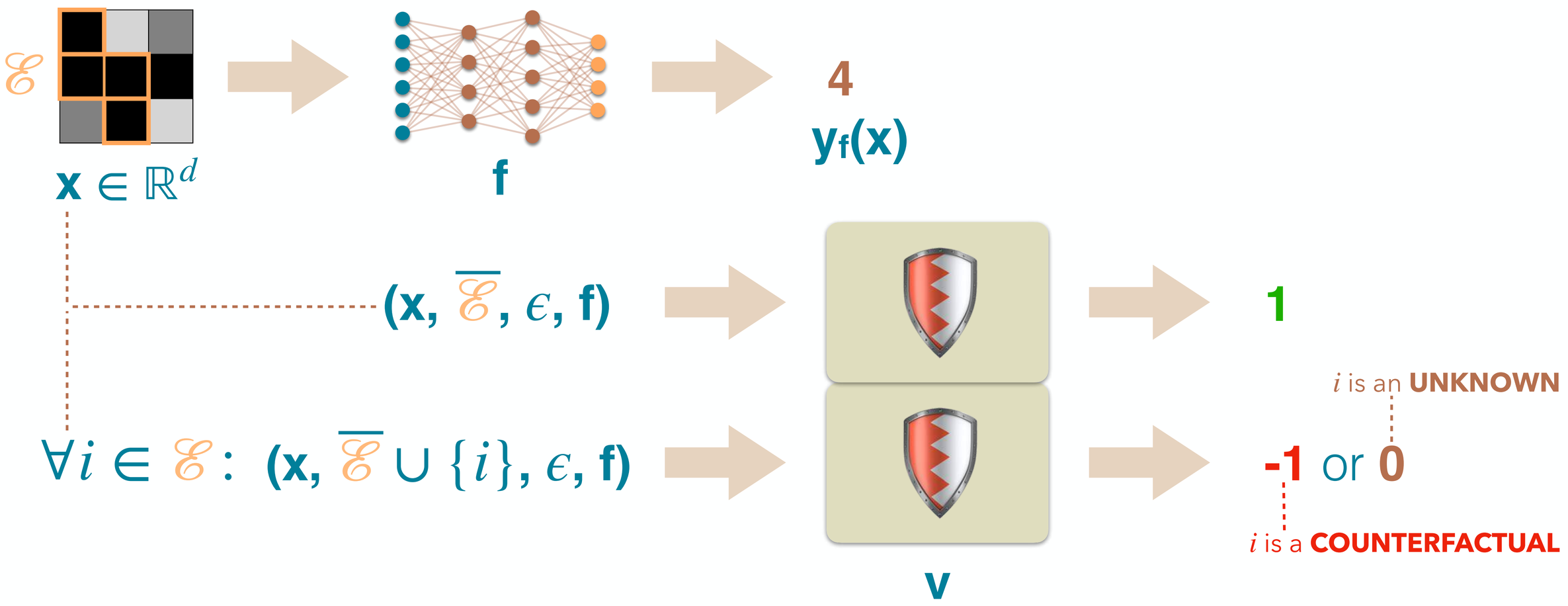
1

-1

Useful Semantic Ideal, but Unrealistic in Practice

~~Optimal~~ Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



Computing ~~Optimal~~ Robust Explanations

DROP (I.E., FREE) INPUT FEATURES WHILE AXp CONDITION HOLDS

ADD TO $\overline{\mathscr{E}}$

```
x00 = float(input())
x01 = float(input())
x02 = float(input())
x03 = float(input())
x04 = float(input())
x05 = float(input())
```

$\mathbf{x}$: x00: [-1, 1], x01: [-1, 1], x02: -1, x03: [-1, 1], x04: [-1, 1], x05: [-1, 1]

```
x10 = ReLU((0.120875)*x00 + (0.065404)*x01 + (0.007862)*x02 + (3.030051)*x03 + (0.101056)*x04 + (-2.103565)*x05 + (1.623834))
x11 = ReLU((0.113805)*x00 + (0.064486)*x01 + (0.007862)*x02 + (3.030051)*x03 + (0.101056)*x04 + (-2.651132)*x05 + (-0.828711))
x12 = ReLU((0.755487)*x00 + (0.224640)*x01 + (0.007862)*x02 + (3.030051)*x03 + (0.101056)*x04 + (1.418635)*x05 + (-0.686885))

x20 = ReLU((1.803209)*x10 + (1.222249)*x11 + (0.007862)*x12 + (3.030051)*x13 + (0.101056)*x14 + (-2.103565)*x15 + (1.623834))
x21 = ReLU((1.958950)*x10 + (2.388245)*x11 + (0.007862)*x12 + (3.030051)*x13 + (0.101056)*x14 + (-2.651132)*x15 + (-0.828711))
x22 = ReLU((1.958103)*x10 + (2.273354)*x11 + (0.007862)*x12 + (3.030051)*x13 + (0.101056)*x14 + (1.418635)*x15 + (-0.686885))

x30 = ReLU((1.735994)*x20 + (0.666507)*x21 + (0.007862)*x22 + (3.030051)*x23 + (0.101056)*x24 + (-2.103565)*x25 + (1.623834))
x31 = ReLU((2.327110)*x20 + (2.685314)*x21 + (0.007862)*x22 + (3.030051)*x23 + (0.101056)*x24 + (-2.651132)*x25 + (-0.828711))
x32 = ReLU((2.147212)*x20 + (2.285599)*x21 + (0.007862)*x22 + (3.030051)*x23 + (0.101056)*x24 + (1.418635)*x25 + (-0.686885))

x40 = ReLU((2.296390)*x30 + (1.980387)*x31 + (0.007862)*x32 + (3.030051)*x33 + (0.101056)*x34 + (-2.103565)*x35 + (1.623834))
x41 = ReLU((-0.552155)*x30 + (-0.828226)*x31 + (0.007862)*x32 + (3.030051)*x33 + (0.101056)*x34 + (-2.651132)*x35 + (-0.828711))
x42 = ReLU((-2.509773)*x30 + (1.199384)*x31 + (0.213429)*x32 + (3.024775)*x33 + (0.101056)*x34 + (1.418635)*x35 + (-0.686885))

x50 = (-2.278012)*x40 + (0.180652)*x41 + (-16.663048)*x42 + (1864)
x51 = (2.278012)*x40 + (-0.180652)*x41 + (16.663048)*x42 + (-1864)
```

LOCAL STABILITY IN $B^{\epsilon}_{\overline{\mathscr{E}}}(\mathbf{x})$

{ x00, x01, x02, x03, x04, x05 } → x51

Free x00: { x01, x02, x03, x04, x05 } → x51

Free x01: { x02, x03, x04, x05 } → x51

Free x02: { x03, x04, x05 } → ✗

Free x03: { x02, x04, x05 } → x51

Free x04: { x02, x05 } → x51

Free x05: { x02 } → x51

AXp = { x02 }

Computing ~~Optimal~~ Robust Explanations

DROP (I.E., FREE) INPUT FEATURES WHILE AXp CONDITION HOLDS

ADD TO $\overline{\mathscr{E}}$

```
x00 = float(input())
x01 = float(input())
x02 = float(input())
x03 = float(input())
x04 = float(input())
x05 = float(input())
```

x: x00: 1, x01: 1, x02: -1, x03: 1, x04: 1, x05: -1

LOCAL STABILITY IN $B^{\epsilon}_{\mathscr{E}}(\mathbf{x})$

INTERVALS

wAXp = { x02, x03, x05 }

SYMBOLIC

wAXp = { x00, x02, x03 }
wAXp = { x02, x03, x05 }

DEEPPOLY

wAXp = { x02, x03 }
wAXp = { x02, x05 }

= PRODUCT

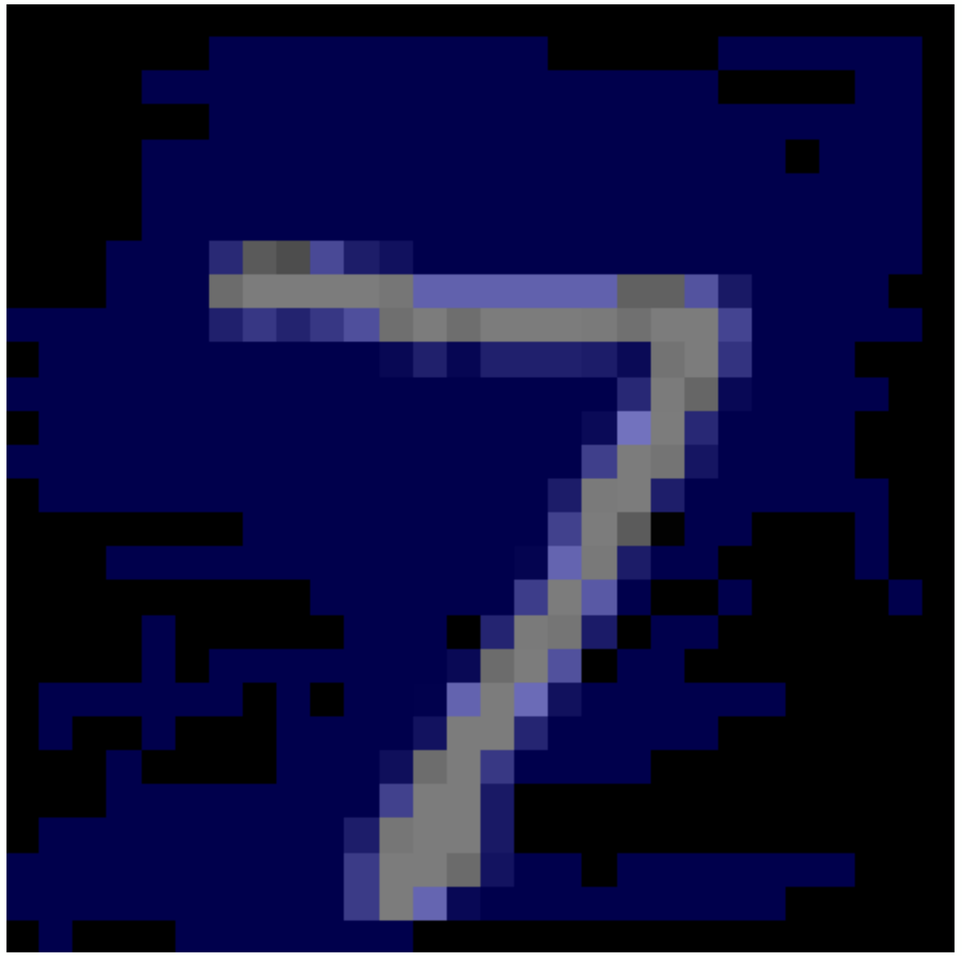
```
x10 = ReLU((0.120875)*x00 + (0.065404)*x01 + (0.097862)*x02 + (2.030051)*x03 + (0.101956)*x04 + (-2.103565)*x05 + (1.623834))
x11 = ReLU((0.113805)*x00 + (0.064486)*x01 + (0.090701)*x02 + (2.123338)*x03 + (0.076374)*x04 + (-1.651132)*x05 + (-0.828711))
x12 = ReLU((0.755487)*x00 + (0.224640)*x01 + (0.344943)*x02 + (2.619876)*x03 + (0.346636)*x04 + (1.418635)*x05 + (-0.686885))

x20 = ReLU((1.803209)*x10 + (1.222249)*x11 + (2.725716)*x12 + (-3.489653))
x21 = ReLU((1.958950)*x10 + (2.388245)*x11 + (2.245851)*x12 + (-3.834811))
x22 = ReLU((1.958103)*x10 + (2.273354)*x11 + (0.662405)*x12 + (-4.211086))

x30 = ReLU((1.735994)*x20 + (0.666507)*x21 + (3.192344)*x22 + (-2.627086))
x31 = ReLU((2.327110)*x20 + (2.685314)*x21 + (1.424807)*x22 + (-3.695113))
x32 = ReLU((2.147212)*x20 + (2.285599)*x21 + (2.665507)*x22 + (-4.299974))

x40 = ReLU((2.296390)*x30 + (1.980387)*x31 + (2.945360)*x32 + (-4.096463))
x41 = ReLU((-0.552155)*x30 + (-0.828226)*x31 + (-0.495998)*x32)
x42 = ReLU((-2.509773)*x30 + (1.199384)*x31 + (-0.245429)*x32 + (5.024773))

x50 = (-2.278012)*x40 + (0.180652)*x41 + (-16.663048)*x42 + (1864)
x51 = (2.278012)*x40 + (-0.180652)*x41 + (16.663048)*x42 + (-1864)
```

Computing ~~Optimal~~ Robust Explanations

FINDING COUNTERFACTUALS RAPIDLY BECOMES INFEASIBLE

Model	MNIST, ϵ=0.25		
	Counterfactuals	**Unknowns**	Time
CNN-3	0.00	247.80	45m

Model	MNIST, ϵ=0.25		
	Counterfactuals	**Unknowns**	Time
CNN-7	0.00	452.00	3h 59m



Computing ~~Optimal~~ Robust Explanations

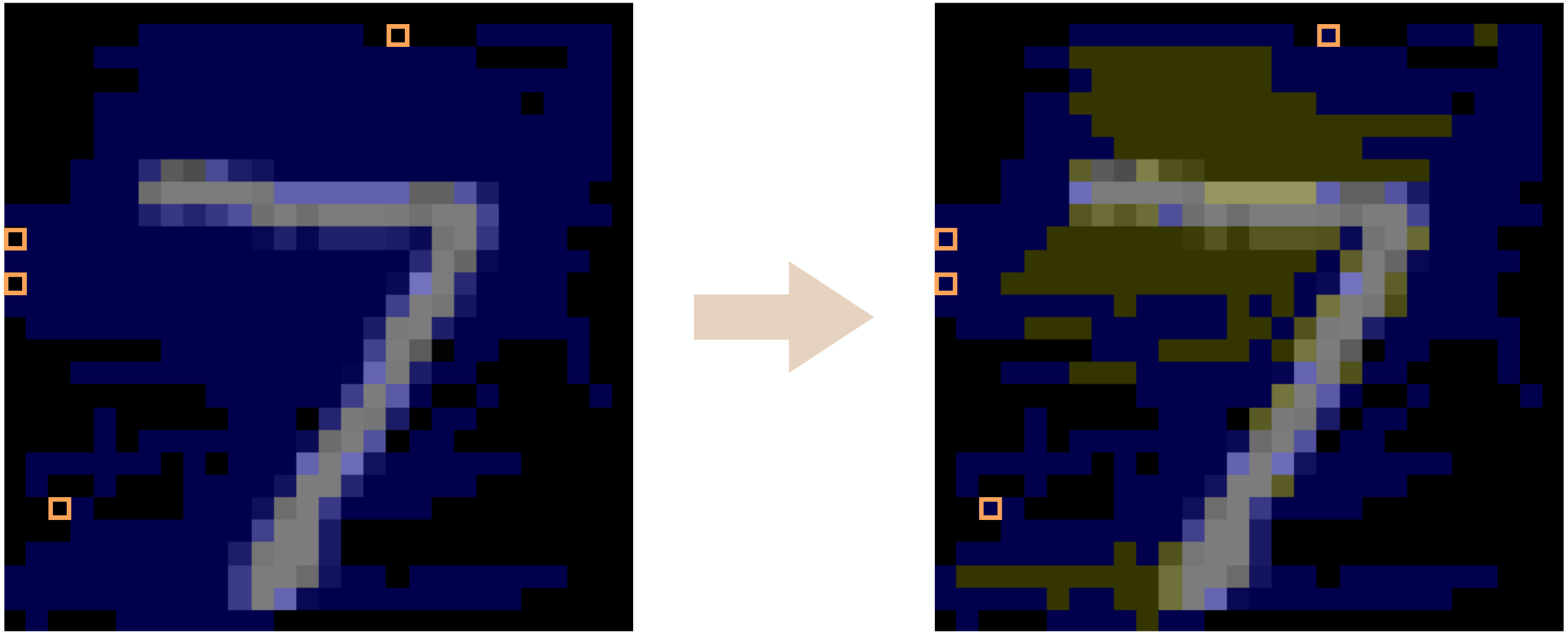
FINDING COUNTERFACTUALS RAPIDLY BECOMES INFEASIBLE

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals	**Unknowns**	Time	**Counterfactuals**	**Unknowns**	Time
CNN-3	0.00	247.80	45m	0.00	461.00	2h 30m

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	Counterfactuals	**Unknowns**	Time	**Counterfactuals**	**Unknowns**	Time
CNN-7	0.00	452.00	3h 59m	0.00	730.67	7h 5m

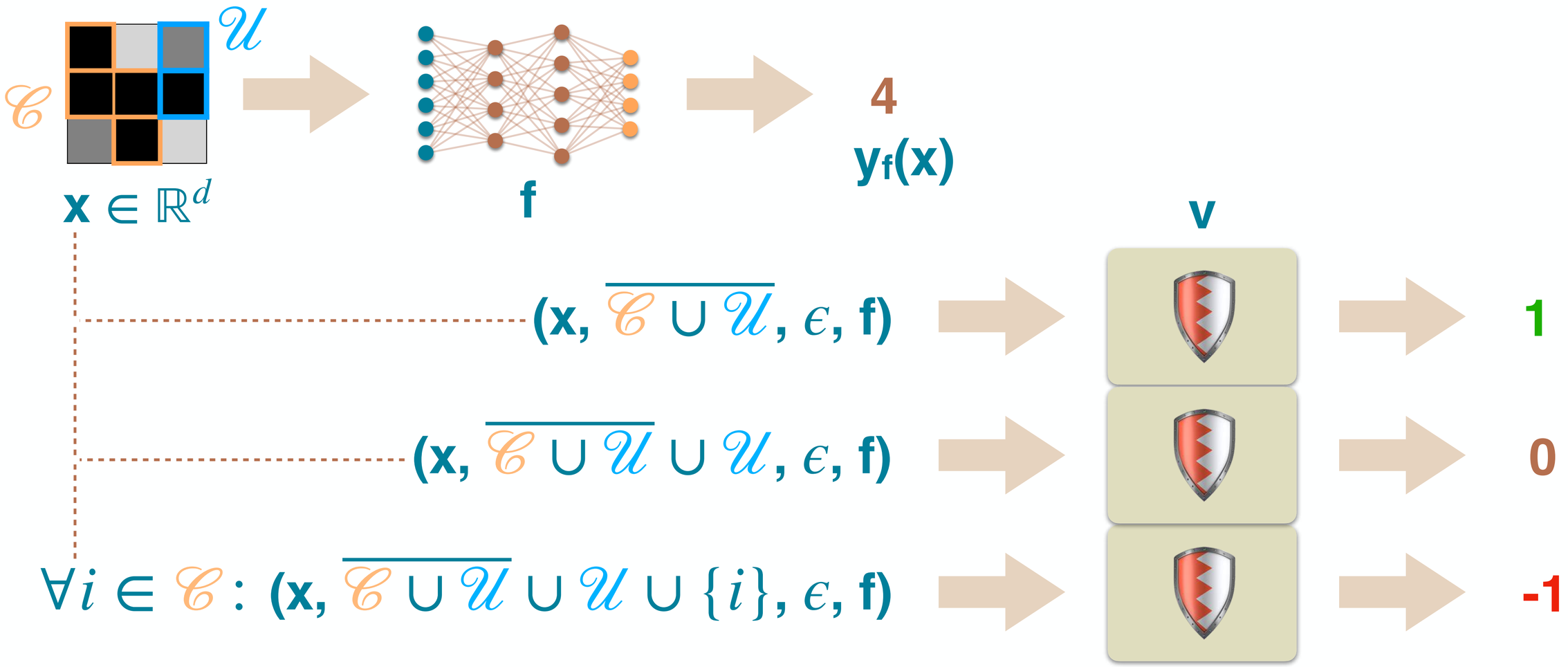
Verifier-Optimal Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



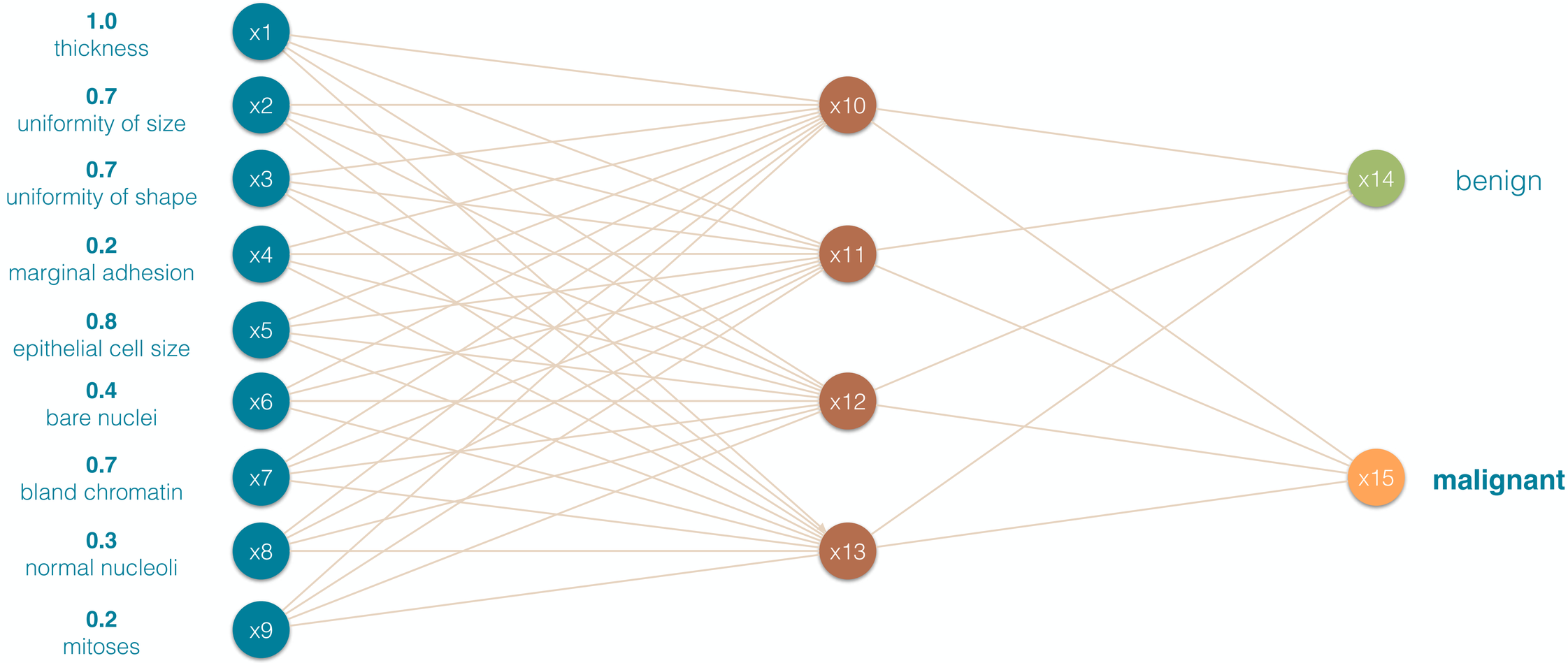
Verifier-Optimal Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



Computing Verifier-Optimal Robust Explanations

EXAMPLE

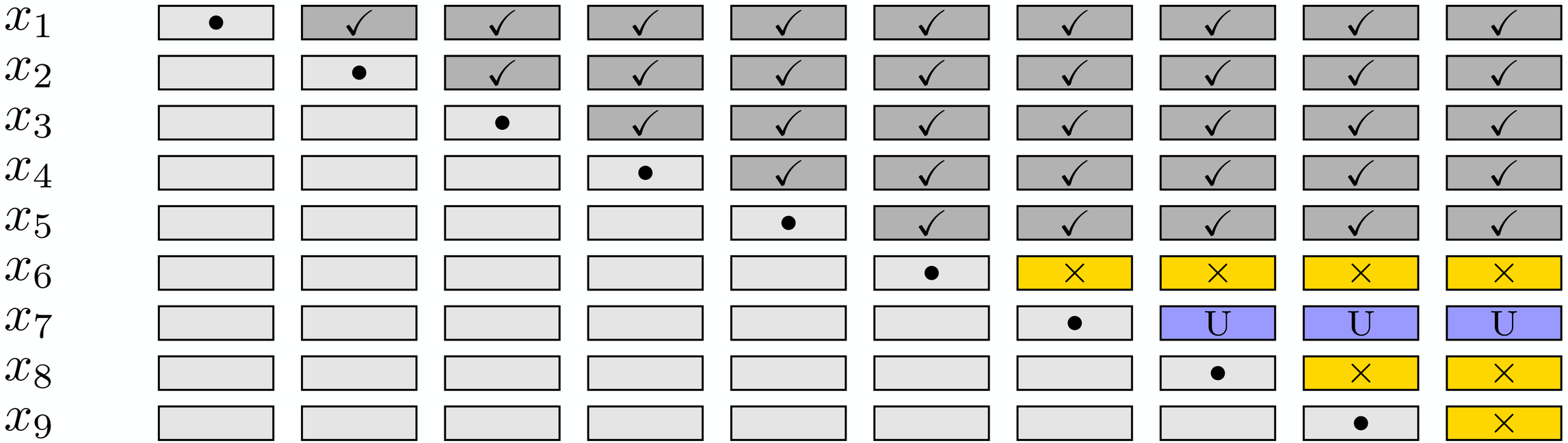


Computing Verifier-Optimal Robust Explanations

DROP (I.E., FREE) INPUT FEATURES WHILE AXp CONDITION HOLDS

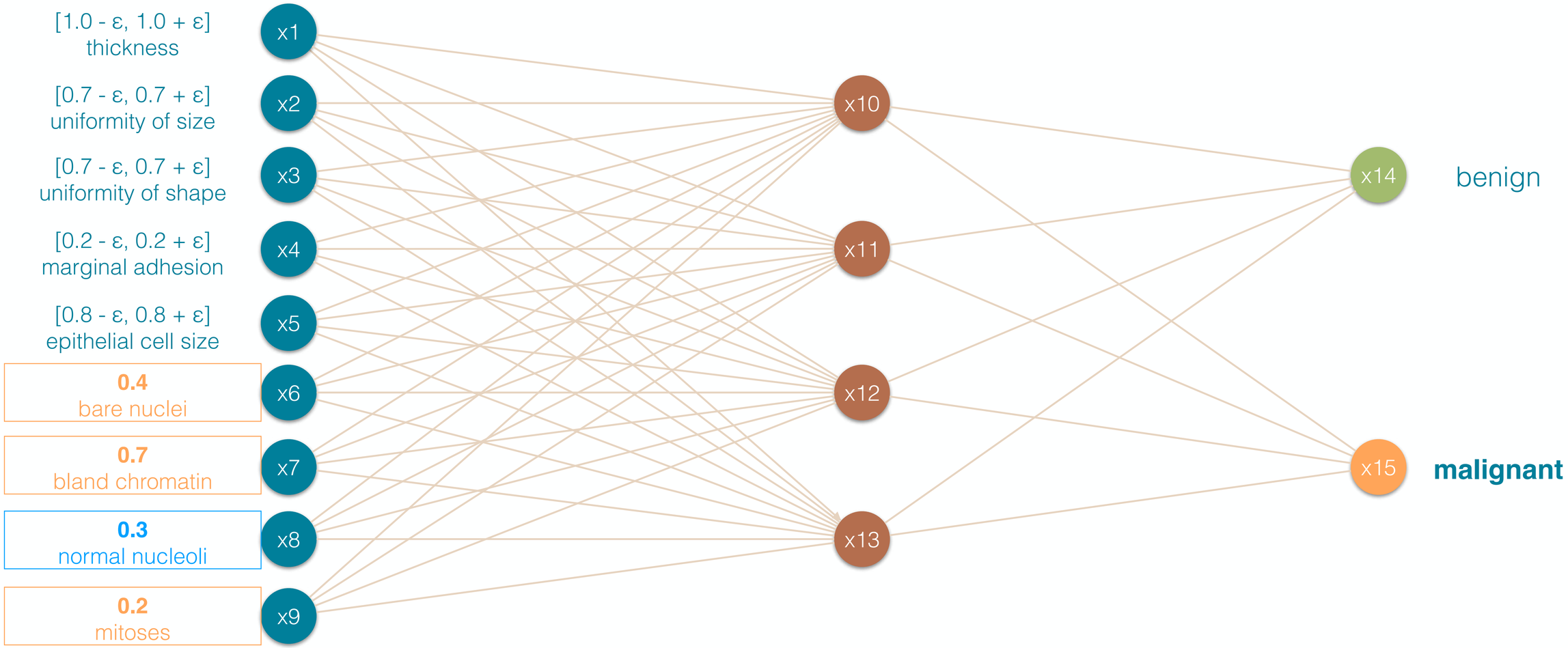
ADD TO $\overline{\mathscr{C} \cup \mathscr{U}}$

LOCAL STABILITY IN $B^{\epsilon=0.6}_{\overline{\mathscr{C} \cup \mathscr{U}}}(\mathbf{x})$



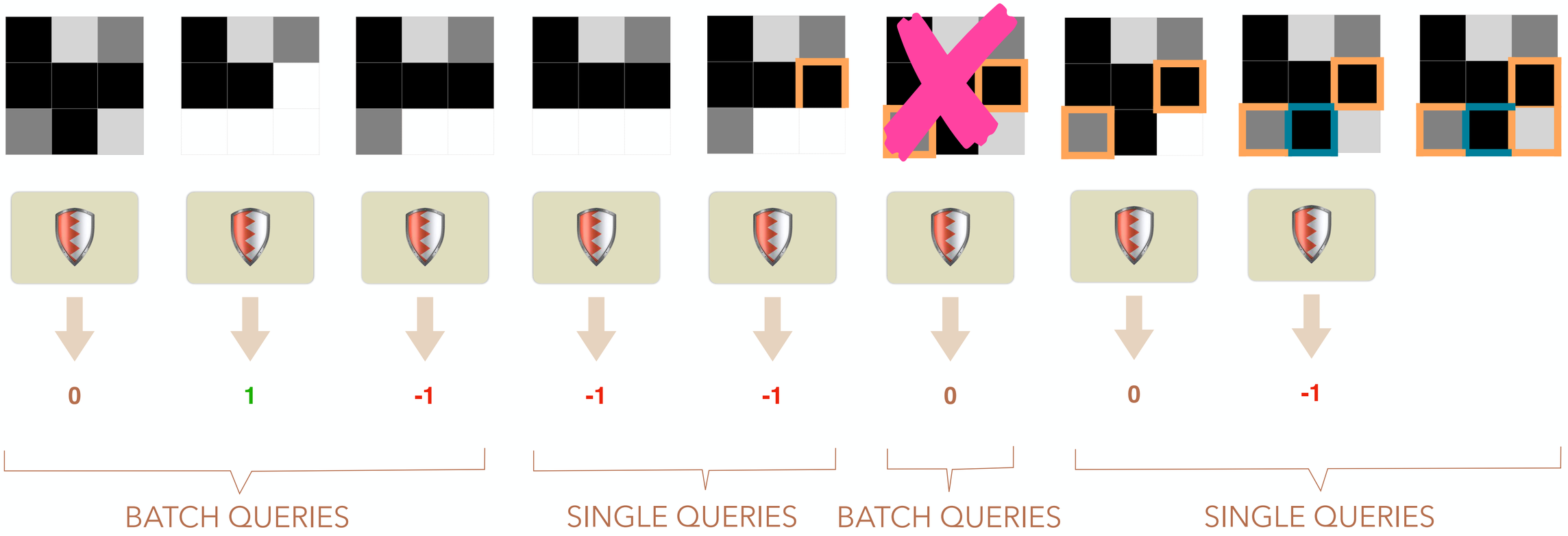
Computing Verifier-Optimal Robust Explanations

EXAMPLE



Computing Verifier-Optimal Robust Explanations

FAVEX (SIMPLIFIED)



Computing Verifier-Optimal Robust Explanations

CNN-3

	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time	**Counterfactuals** $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time
Standard	0.00	247.80	45m	0.00	461.00	2h 30m
Verifier-Optimal	**160.30**	94.40	**10m**	**210.40**	251.70	**19m**

160.30 +
94.40 =
254.70

210.40 +
251.70 =
462.10

Computing Verifier-Optimal Robust Explanations

CNN-7

	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	Counterfactuals $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time	**Counterfactuals** $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time
Standard	0.00	452.00	239m	0.00	730.67	7h 5m
Verifier-Optimal	**207.33**	249.33	**1h 14m**	**467.00**	266.33	**1h 49m**

207.33 + 249.33 = 456.66

467.00 + 266.33 = 733.33

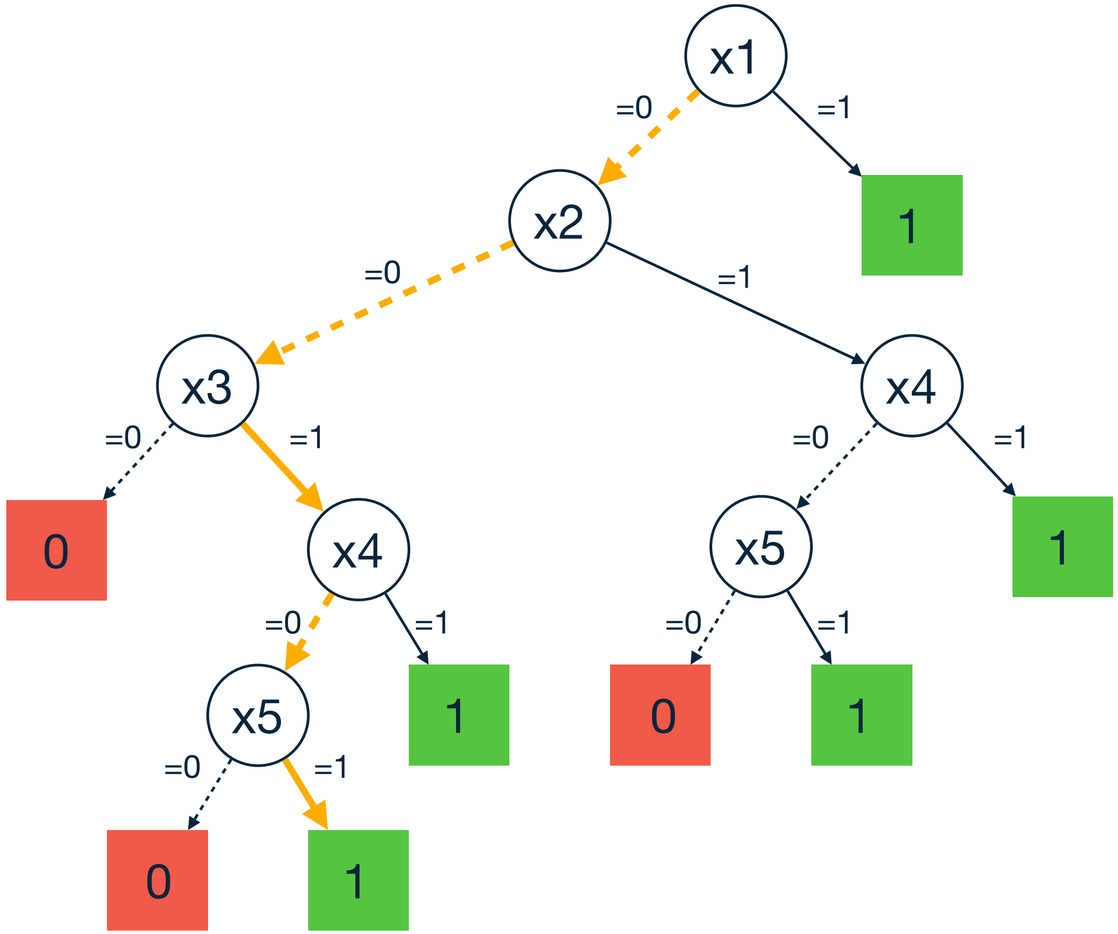
Computing Verifier-Optimal Robust Explanations

TRAVERSAL STRATEGIES

Model	Traversal	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
		$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
CNN-3	VeriX	160.50	122.30	16m	437.20	328.30	32m
	VeriX+	**155.20**	**92.20**	10m	262.00	206.90	15m
	α-FAVEX	181.20	108.70	12m	**210.40**	**251.70**	19m
	FaVeX-IBP	160.30	94.40	10m	250.90	215.50	16m
Model	Traversal	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
		$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
CNN-7	VeriX	123.33	423.67	2h 9m	728.67	216.33	1h 23m
	VeriX+	**196.67**	**232.67**	1h 13m	522.00	213.67	1h 25m
	α-FAVEX	234.67	317.67	1h 29m	467.00	266.33	1h 49m
	FaVeX-IBP	207.33	249.33	1h 14m	**512.67**	**216.67**	1h 25m

Contrastive Explanation (CXp)

SUBSET-MINIMAL SET OF FEATURES SUFFICIENT TO ALTER A PREDICTION



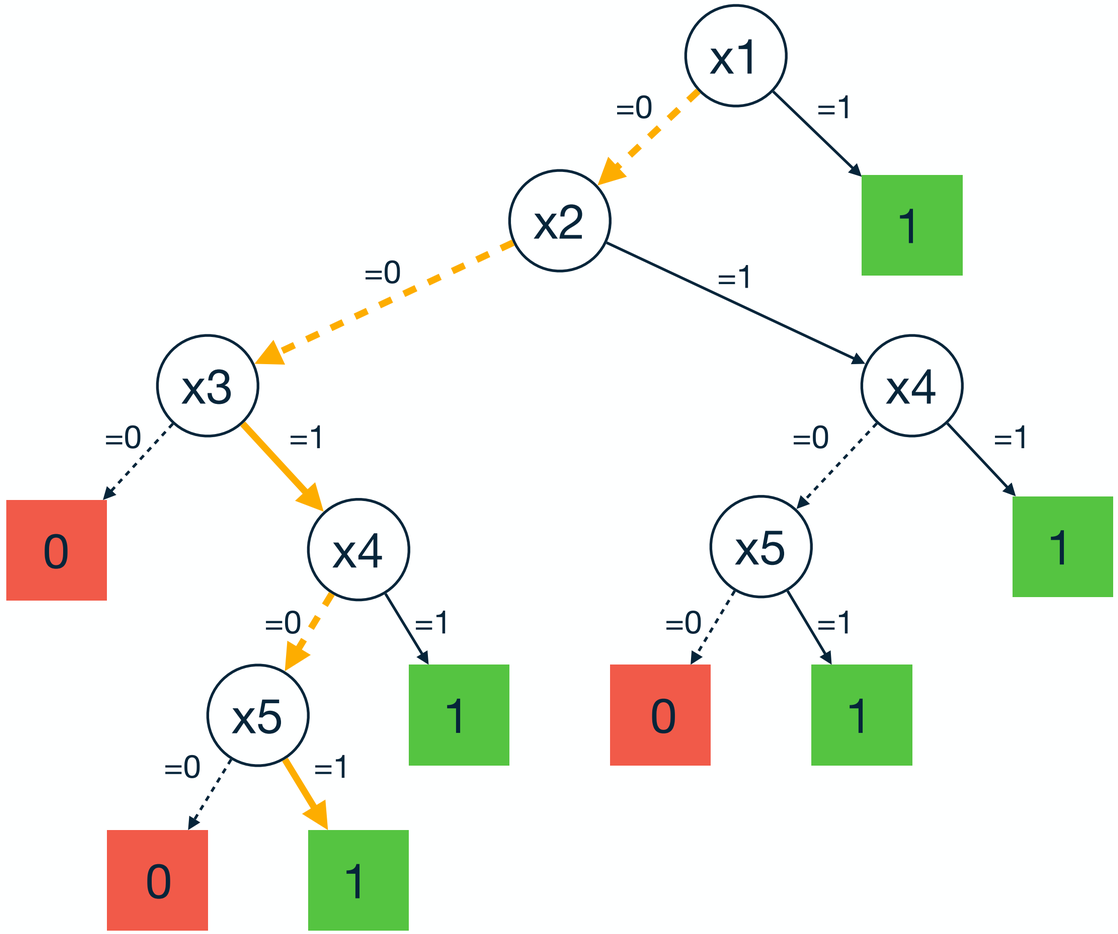
CXp = { x3 }

CXp = { x5 }

Computing One CXp

DROP (I.E., FIX) INPUT FEATURES WHILE CXp CONDITION HOLDS

¬ SAME PREDICTION



{ 1, 2, 3, 4, 5 } → 1 0

Fix 1: { 2, 3, 4, 5 } → 1 0

Fix 2: { 3, 4, 5 } → 1 0

Fix 3: { 4, 5 } → 1 0

Fix 4: { 5 } → 1 0

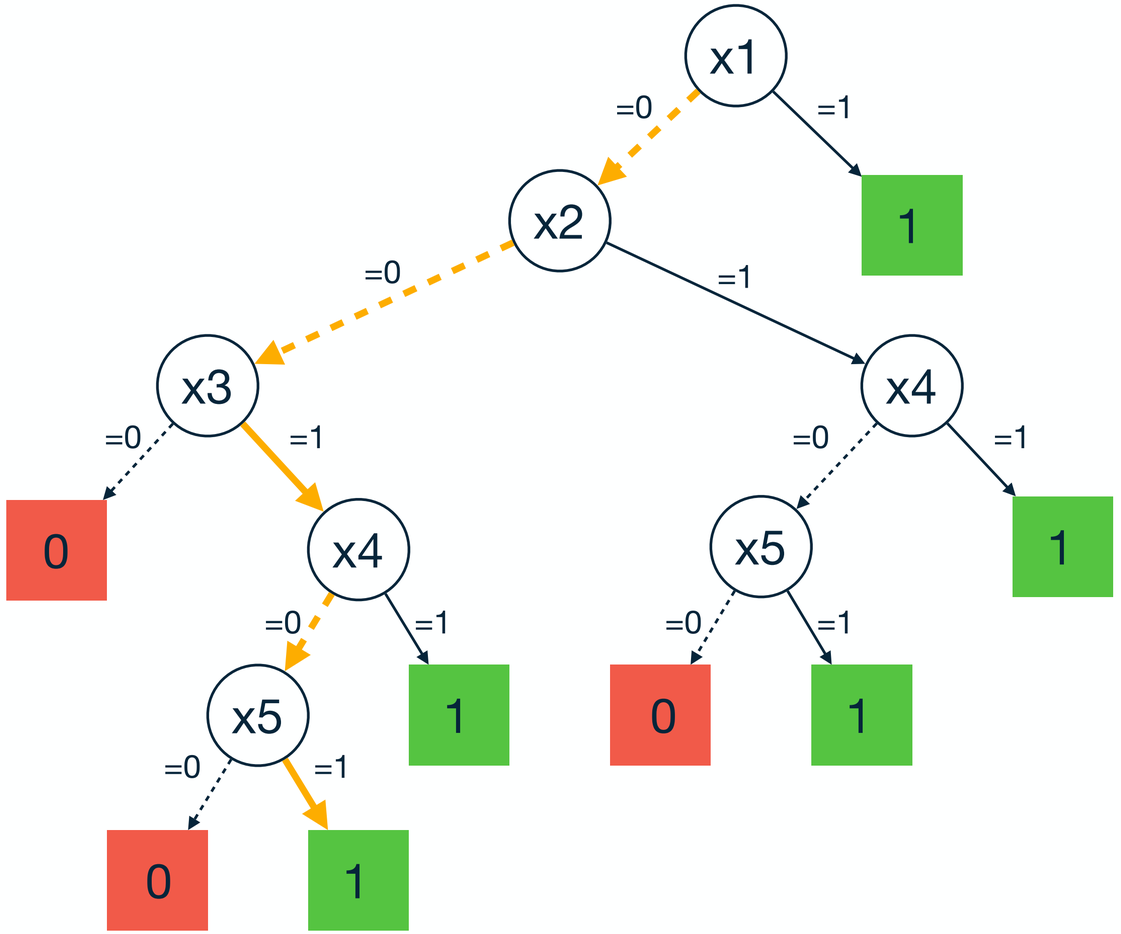
Fix 5: ∅ → ✗

CXp = { x5 }

Computing One CXp

DROP (I.E., FIX) INPUT FEATURES WHILE CXp CONDITION HOLDS

¬ SAME PREDICTION



{ 1, 2, 3, 4, 5 } → 1 0

Fix 5: { 1, 2, 3, 4 } → 1 0

Fix 4: { 1, 2, 3 } → 1 0

Fix 3: { 1, 2 } → ✗

Fix 2: { 1, 3 } → 1 0

Fix 1: { 3 } → 1 0

CXp = { x3 }

Bibliography

[Li19] **Jianlin Li, Jiangchao Liu, Pengfei Yang, Liqian Chen, Xiaowei Huang, and Lijun Zhang**. Analyzing Deep Neural Networks with Symbolic Propagation: Towards Higher Precision and Faster Verification. In SAS, page 296–319, 2019.
symbolic abstract domain

[Singh19] **Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev**. An Abstract Domain for Certifying Neural Networks. In POPL, pages 41:1 - 41:30, 2019.
deeppoly abstract domain

[Urban21] **Caterina Urban and Antoine Miné**. A Review of Formal Methods applied to Machine Learning. https://arxiv.org/abs/2104.02466, 2021.
survey on formal methods for machine learning

Bibliography

[Durand22] **Serge Durand, Augustin Lemesle, Zakaria Chihani, Caterina Urban, François Terrier.** ReCIPH: Relational Coefficients for Input Partitioning Heuristic. In WFVML, 2022.
branch-and-bound with input splitting by largest coefficient

[Marques-Silva21] **João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska**. Explanations for Monotonic Classifiers. In ICML, pages 7469-7479, 2021.
abductive and contrastive explanations

[DePalma26] **Alessandro De Palma, Greta Dolcetti, Caterina Urban.** Faster Verified Explanations for Neural Networks. In ECOOP, 2026.
verifier-optimal robust explanations