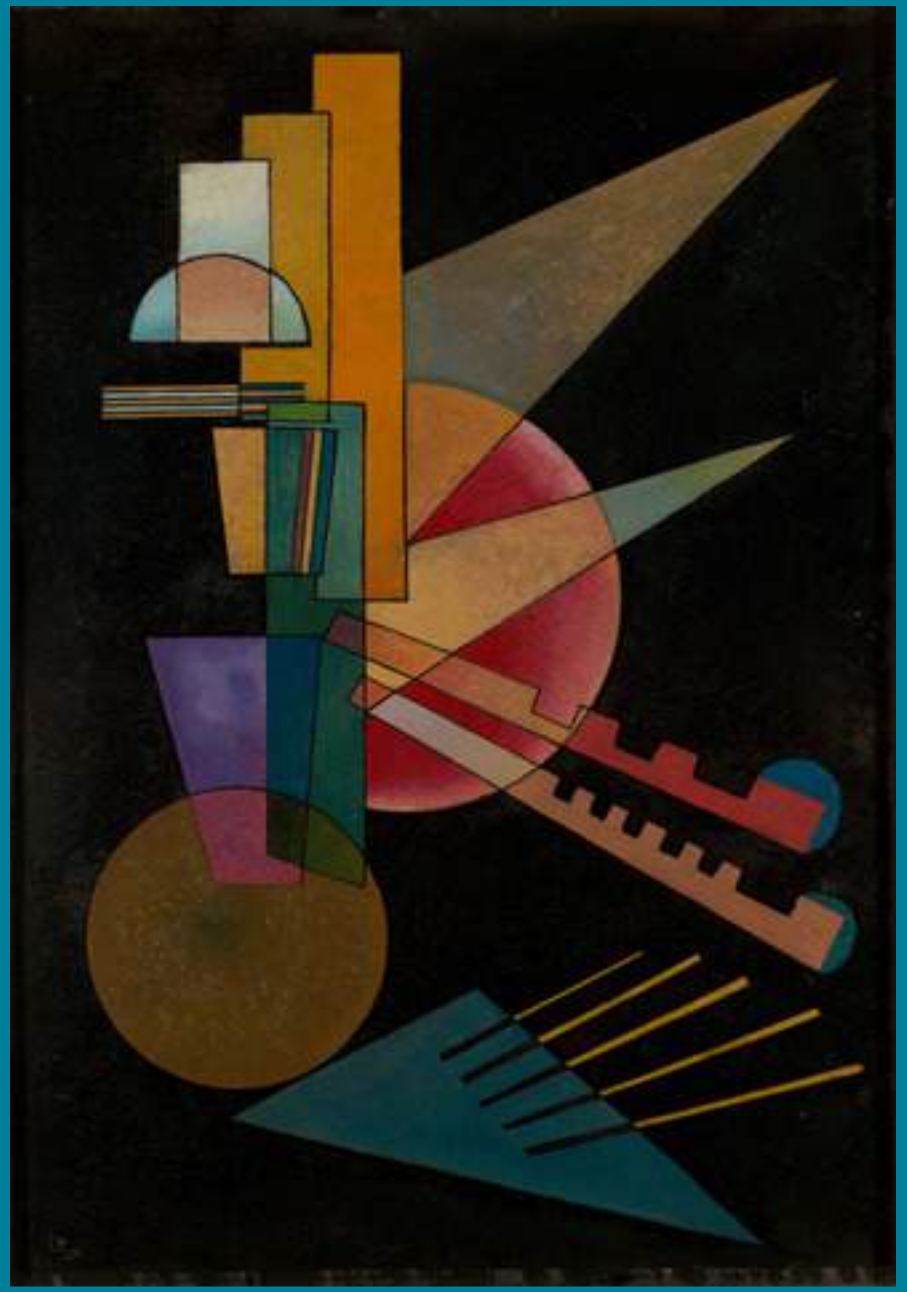


Kandinsky - Abstract Interpretation, 1925

Abstract Interpretation

and Applications in Security, Data Science, and Machine Learning

OPLSS 2025

Caterina Urban
Inria & École Normale Supérieure | Université PSL

Static Analysis of Safety Properties

The Art of Losing Precision

No Surprises, Please

What Could
Possibly Go Right?

It's Complicated

Trace Properties

$T \in \mathscr{P}(\Sigma^\infty)$

Safety Properties = “Nothing Bad Ever Happens”

Example

- Any State Property $S \in \mathscr{P}(\Sigma)$: $T \stackrel{\text{def}}{=} S^\infty$

Safety Property Verification

- T can be **verified** by **exhaustive testing**

$\mathscr{T}_p(I) \subseteq T$

- T can be **falsified** by finding **a single finite execution not in** T

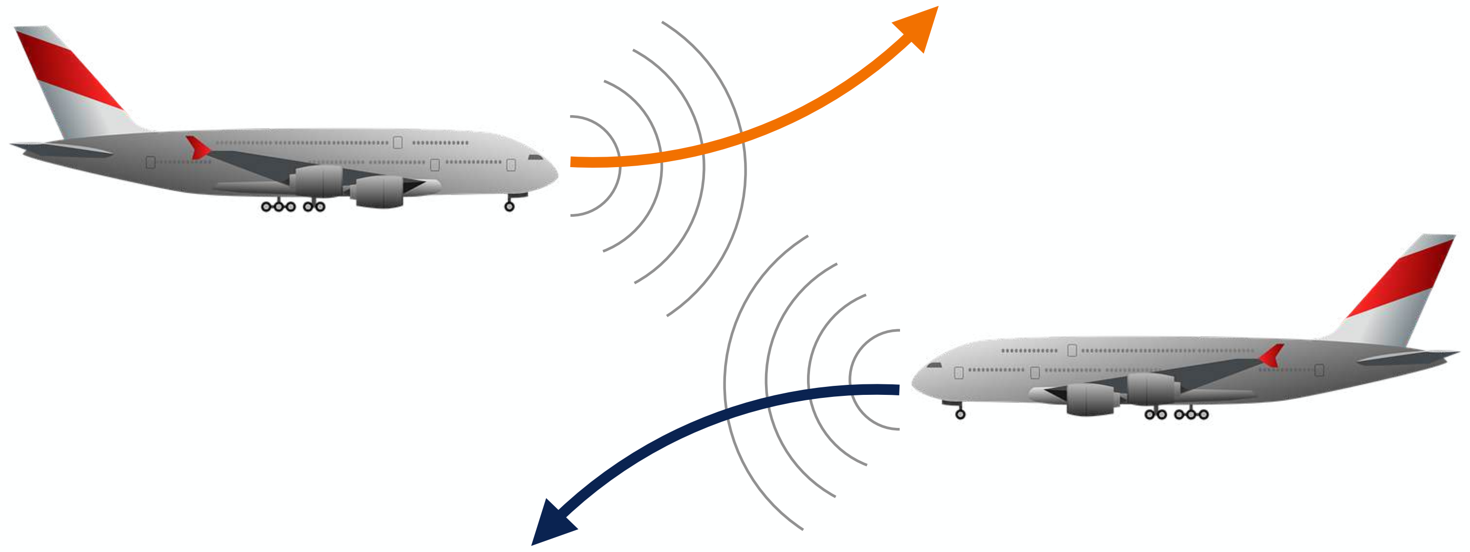
Machine Learning Safety

Machine Learning in High-Stakes Systems

data

ML software

perform tasks that are impossible using explicit programming



act as surrogate model

automate decision-making

Machine Learning in High-Stakes Systems

[1] STAT+ [2]

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

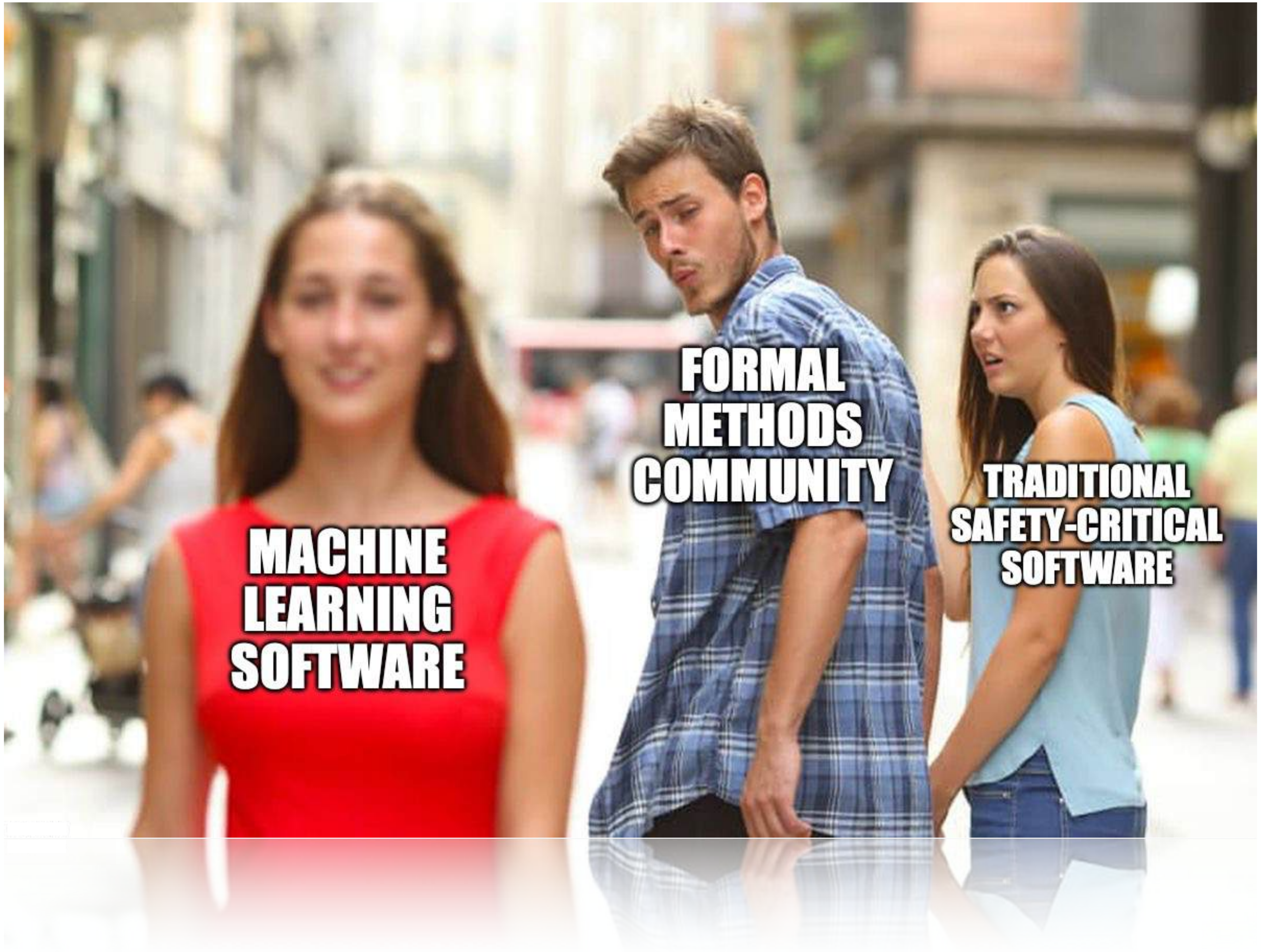
By Casey Ross[3] @caseymross[4] *and* Ike Swetlitz

July 25, 2018

A self-driving Uber ran a red light last December, contrary to company claims

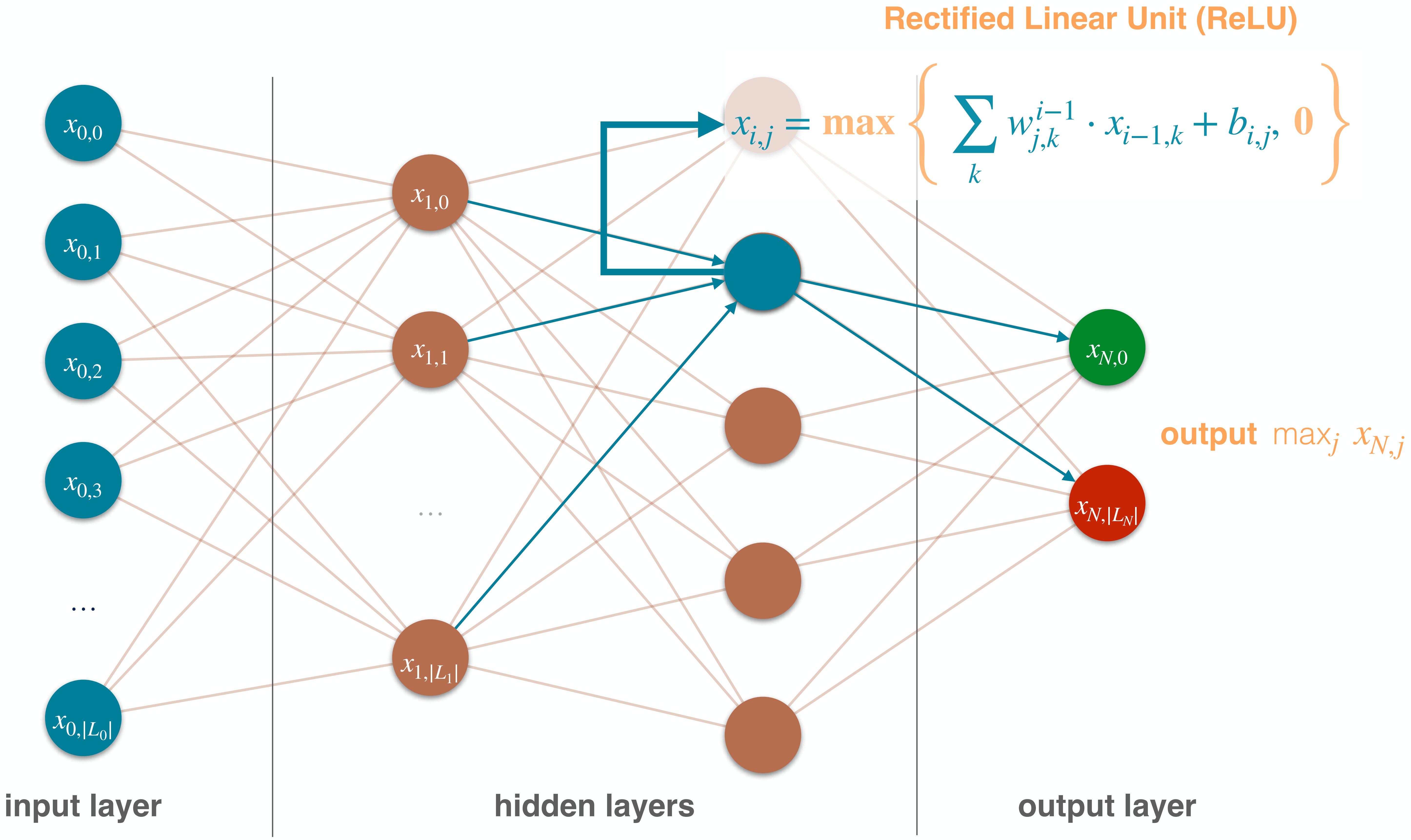
Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash

Richard Gonzales November 7, 201910:57 PM ET

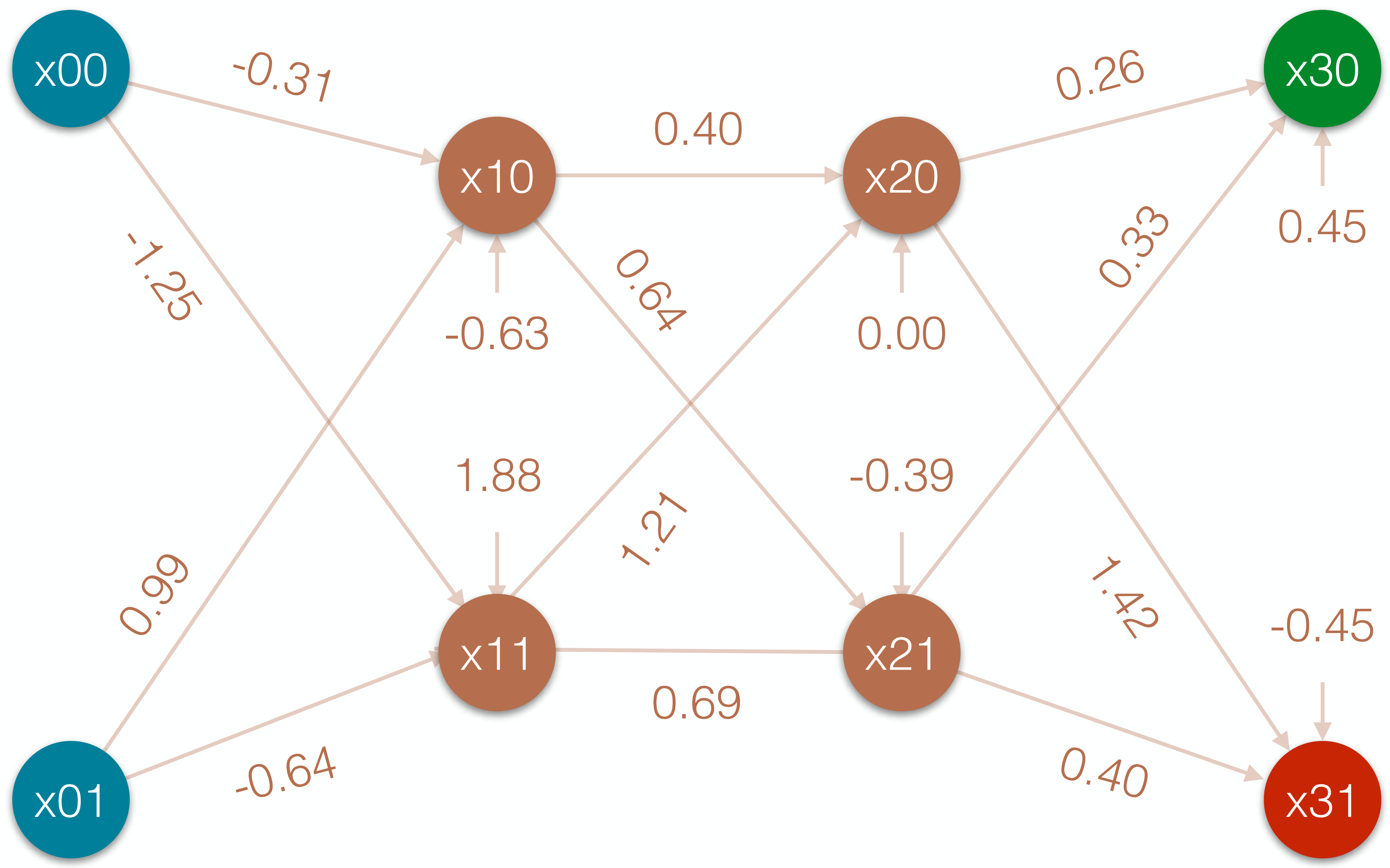

FORMAL METHODS COMMUNITY
MACHINE LEARNING SOFTWARE
TRADITIONAL SAFETY-CRITICAL SOFTWARE

Neural Networks

Feed-Forward ReLU-Activated Neural Networks



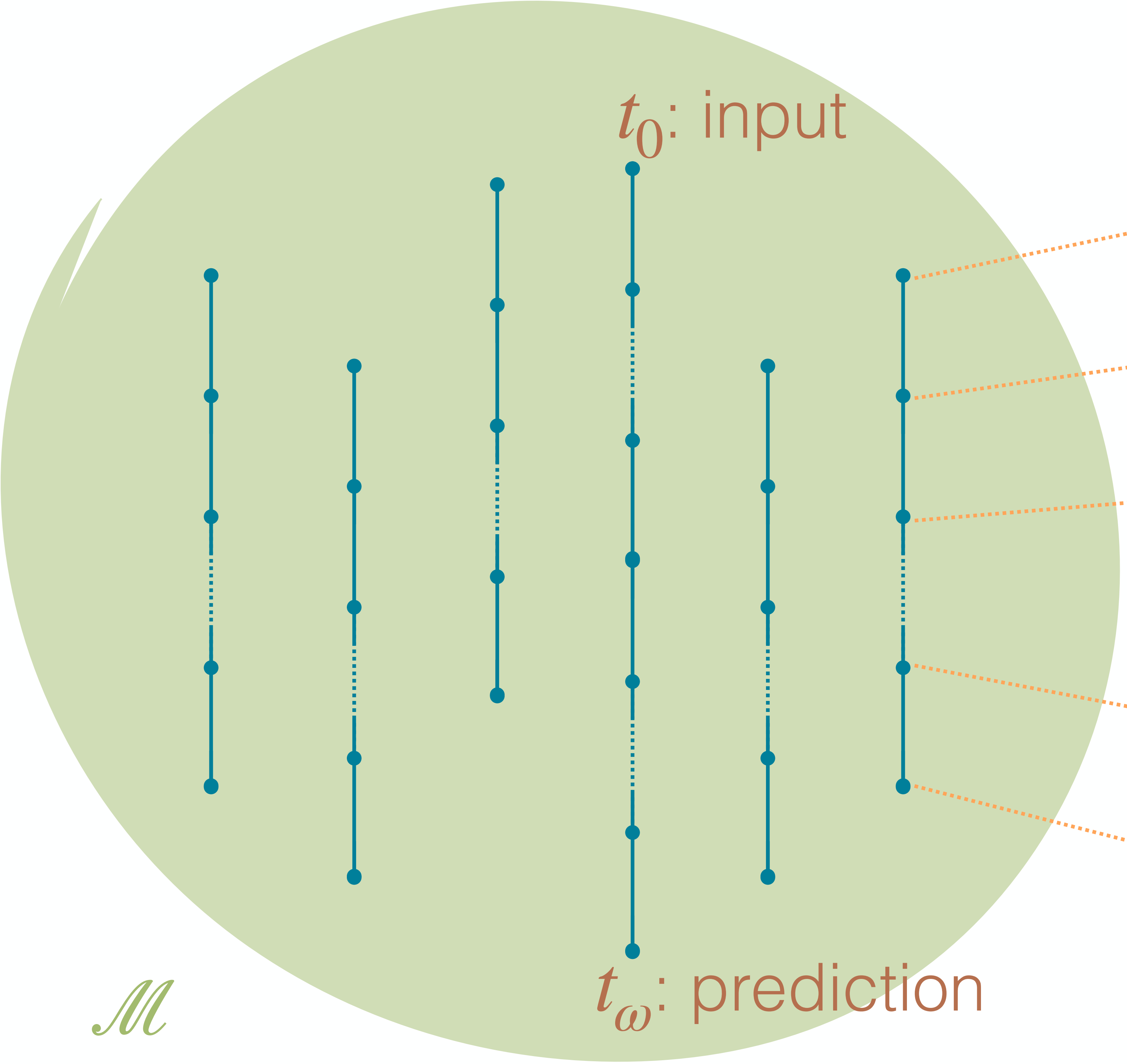
Neural Networks as Programs




```
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 ‘ ’
```

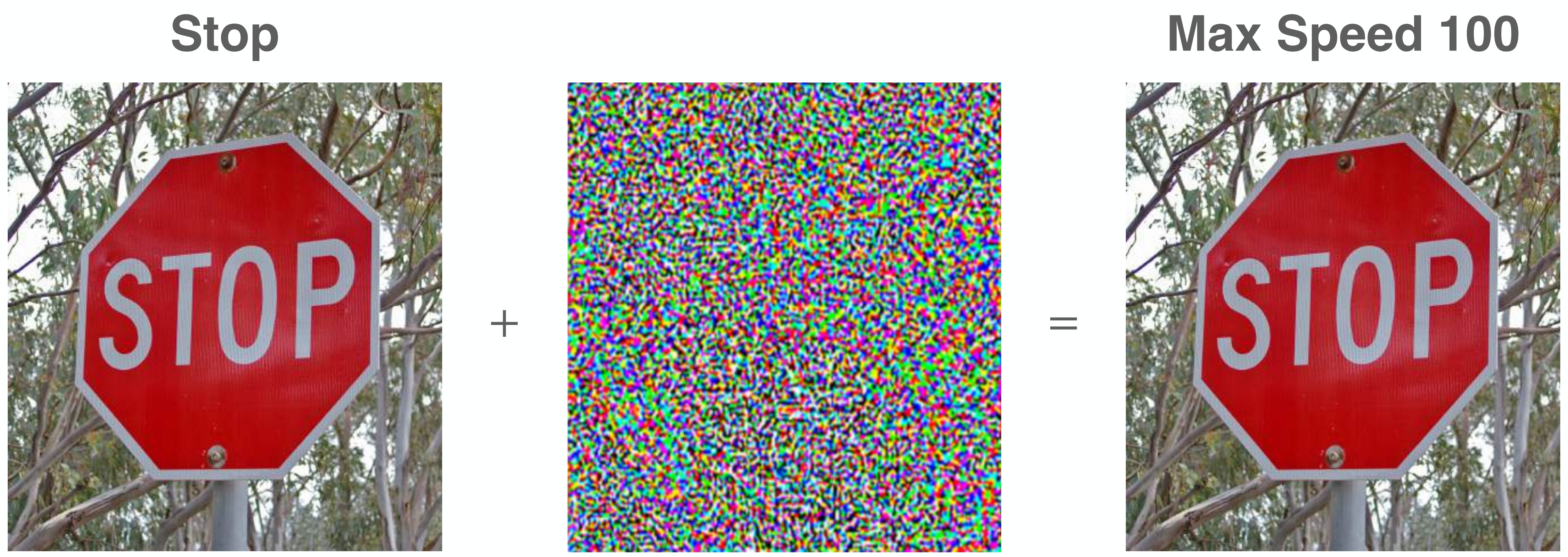
Neural Networks as Programs

Maximal Trace Semantics

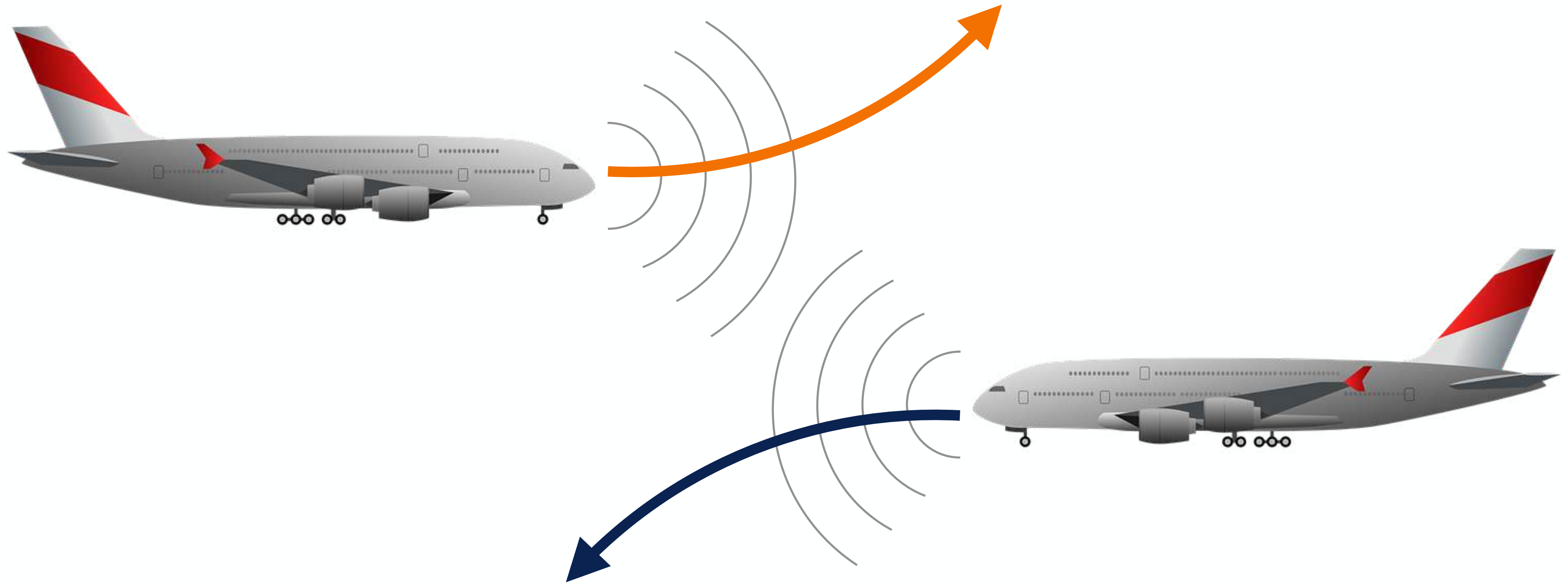



```
x00 = input()
x01 = input()

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 ‘ ’
```

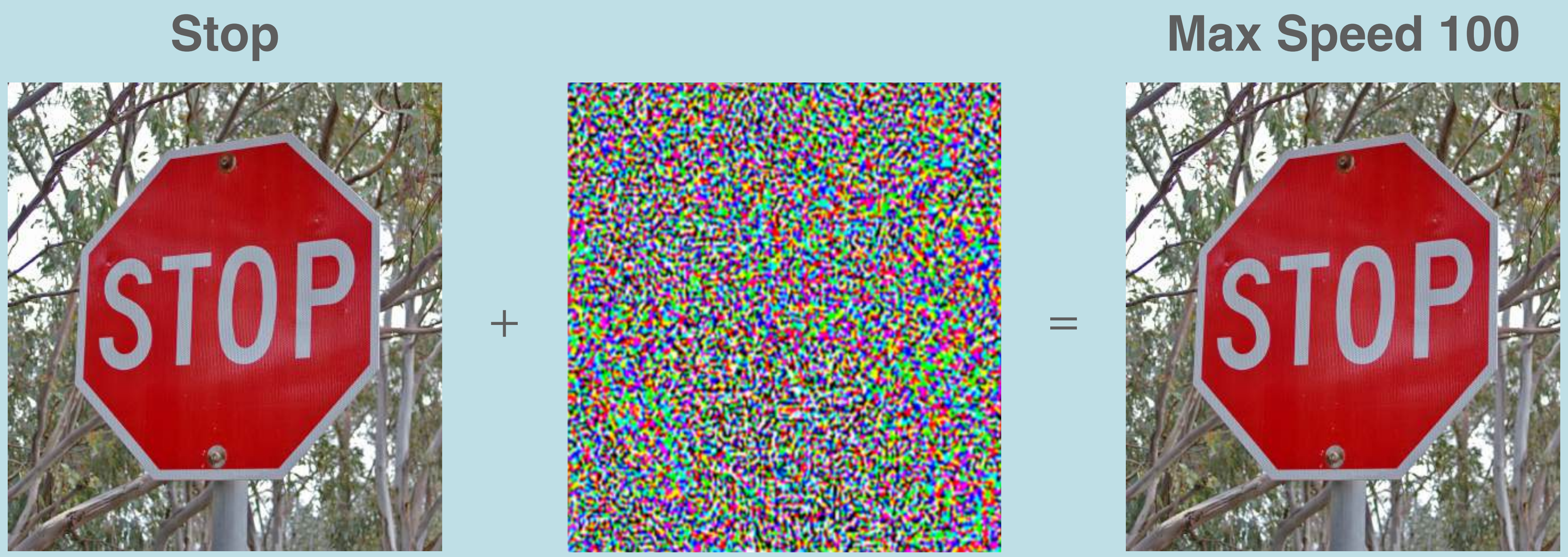
Stability



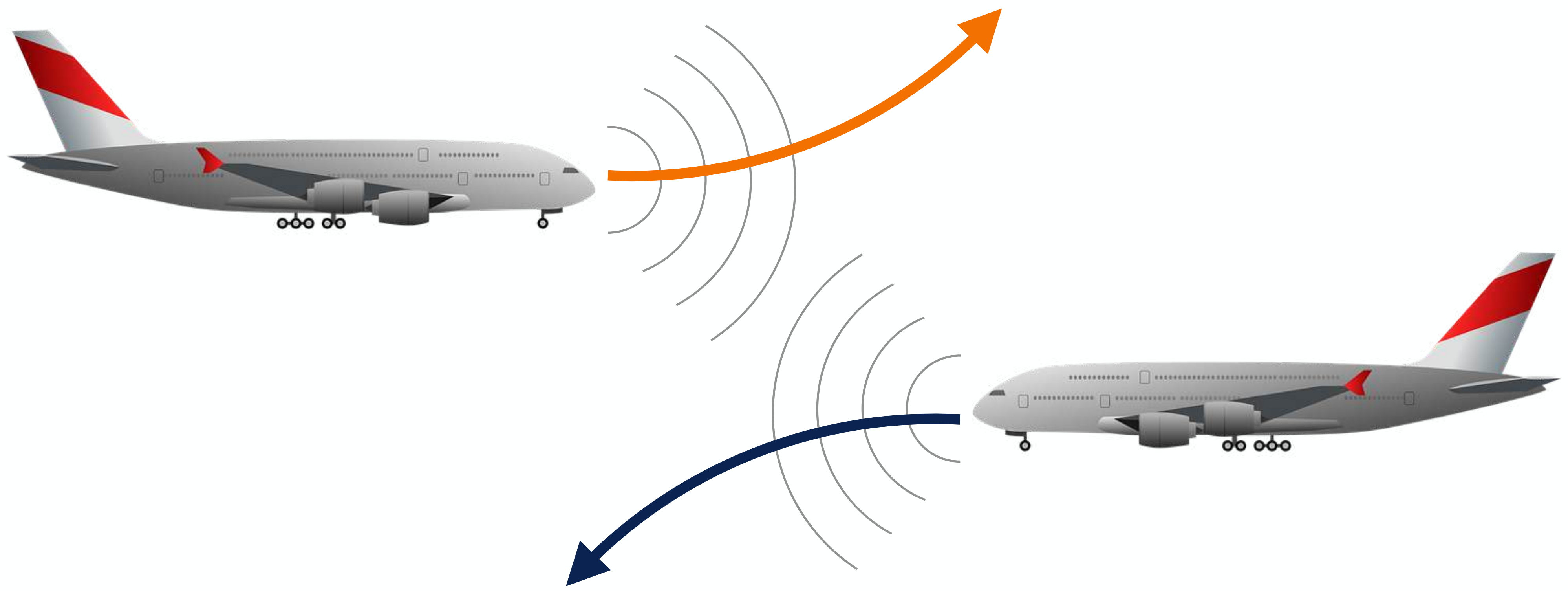
Safety



Stability



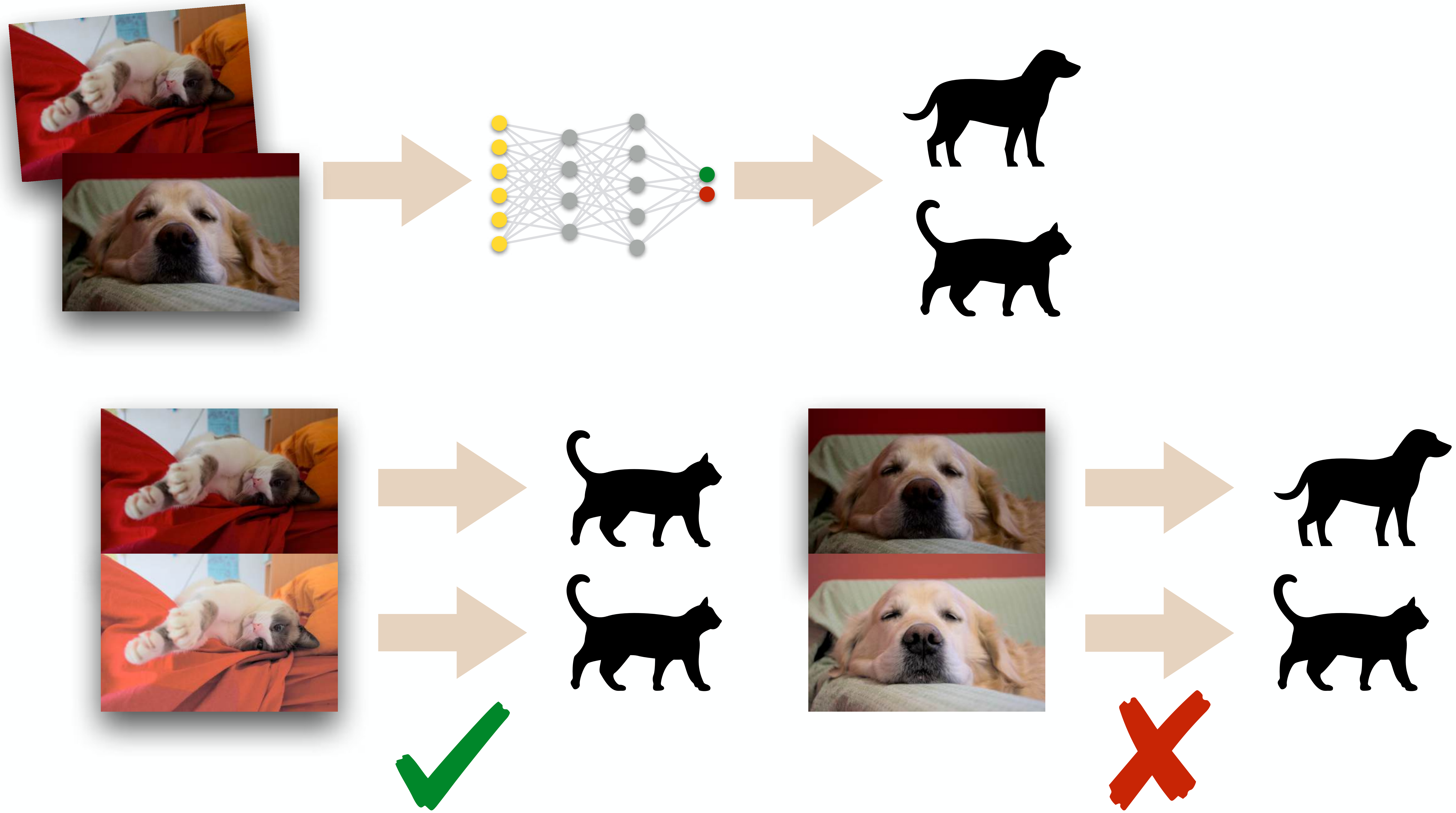
Safety



Prediction Stability

Local Prediction Stability

Prediction is Unaffected by Input Perturbations

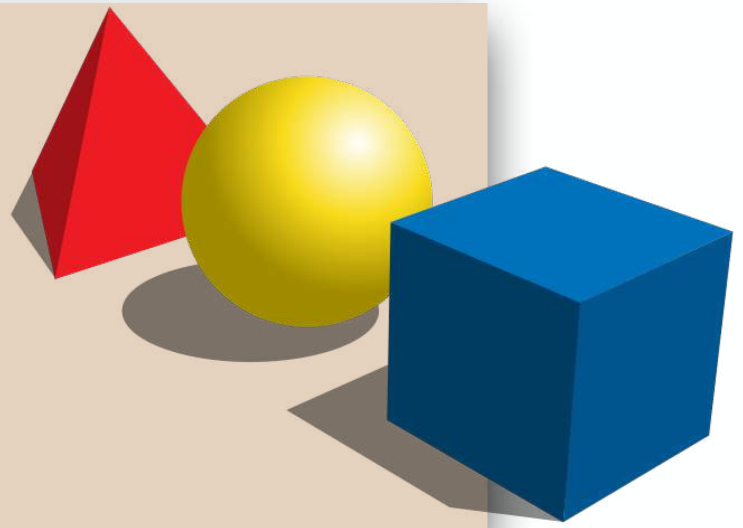


Static Local Prediction Stability Analysis

3-Step Recipe

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

Local Prediction Stability

Distance-Based Perturbations

$P_{\delta,\epsilon}(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{x}' \in \mathbb{R}^{|L_0|} \mid \delta(\mathbf{x}, \mathbf{x}') \leq \epsilon\}$

Example (L_∞ distance): $P_{\infty,\epsilon}(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{x}' \in \mathbb{R}^{|L_0|} \mid \max_i |\mathbf{x}_i - \mathbf{x}'_i| \leq \epsilon\}$

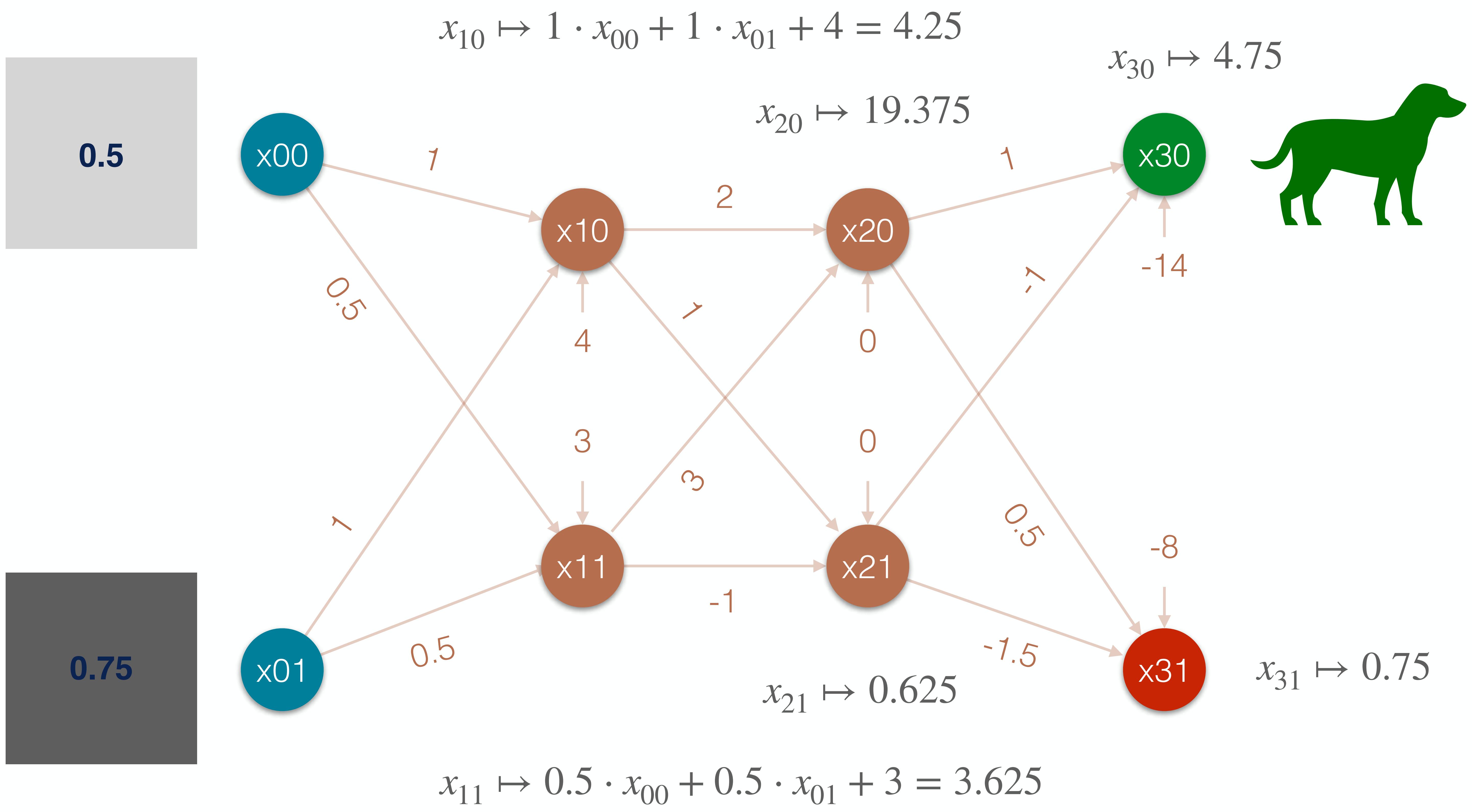
classification of $\mathbf{x}$

$$\mathscr{R}_{\mathbf{x}}^{\delta,\epsilon} \stackrel{\text{def}}{=} \{t \in \Sigma^* \mid t_0 \in P_{\delta,\epsilon}(\mathbf{x}) \Rightarrow t_\omega = C(\mathbf{x})\}$$

$\mathscr{R}_{\mathbf{x}}^{\delta,\epsilon}$ is the set of all traces that are **prediction stable** in the neighborhood $P_{\delta,\epsilon}(\mathbf{x})$ of a given input $\mathbf{x}$

Static Local Prediction Stability

Example



$$P(\langle 0.5, 0.75 \rangle) \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathscr{R} \times \mathscr{R} \mid 0 \leq \mathbf{x}_0 \leq 1 \wedge 0 \leq \mathbf{x}_1 \leq 1\}$$

Static Local Prediction Stability Analysis

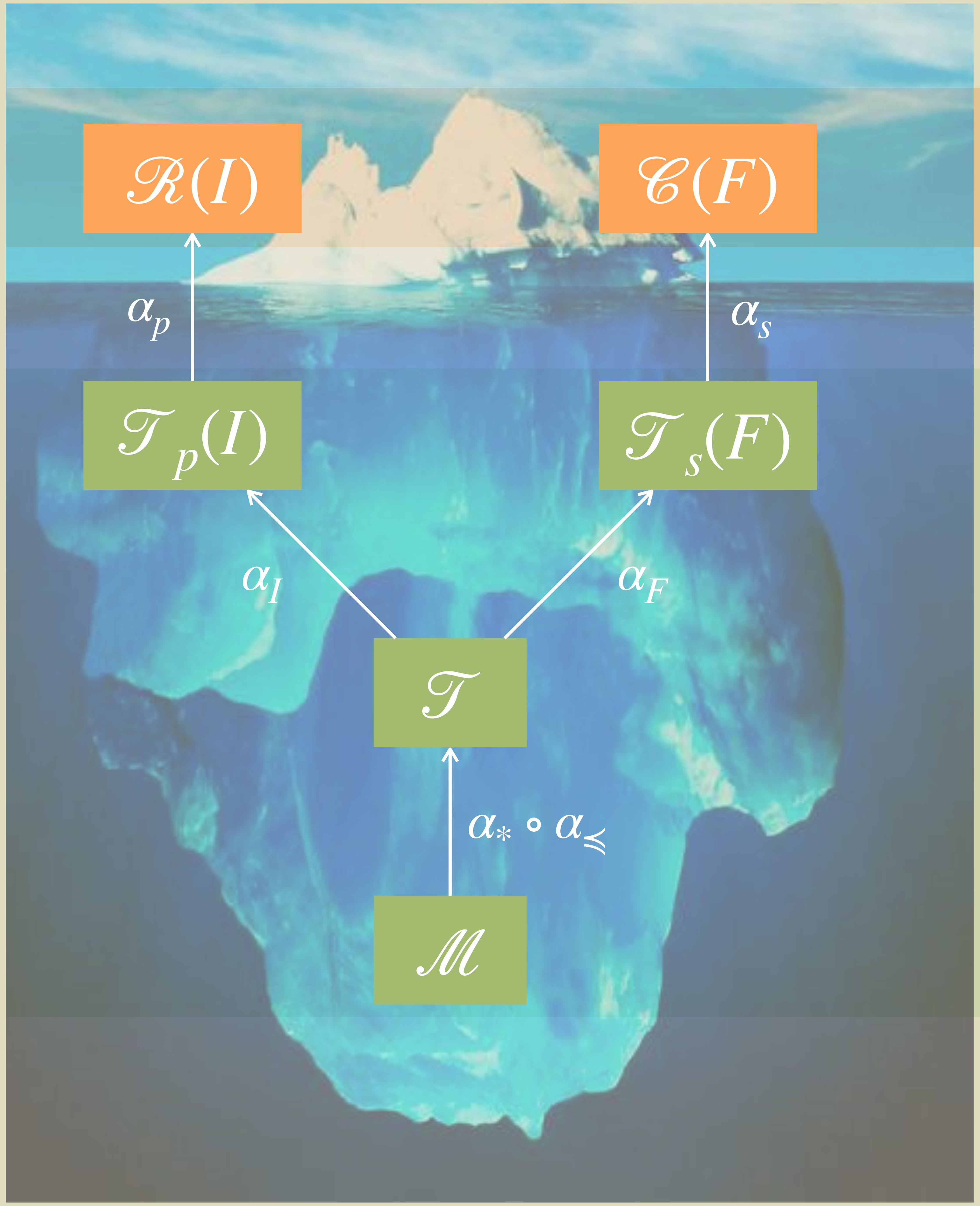
Concrete Semantics

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

concrete semantics
mathematical models of the program behavior

Hierarchy of Semantics



Forward/Backward Reachability Semantics

Forward/Backward Reachable State Abstraction

Prefix/Suffix Trace Semantics

Prefix/Suffix Trace Abstraction

Partial Finite Trace Semantics

Partial Finite Trace Abstraction

Maximal Trace Semantics

Local Prediction Stability

Distance-Based Perturbations

$P_{\delta,\epsilon}(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{x}' \in \mathbb{R}^{|L_0|} \mid \delta(\mathbf{x}, \mathbf{x}') \leq \epsilon\}$

Example (L_∞ distance): $P_{\infty,\epsilon}(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{x}' \in \mathbb{R}^{|L_0|} \mid \max_i |\mathbf{x}_i - \mathbf{x}'_i| \leq \epsilon\}$

classification of $\mathbf{x}$

$$\mathcal{R}_{\mathbf{x}}^{\delta,\epsilon} \stackrel{\text{def}}{=} \{t \in \Sigma^* \mid t_0 \in P_{\delta,\epsilon}(\mathbf{x}) \Rightarrow t_\omega = C(\mathbf{x})\}$$

$\mathcal{R}_{\mathbf{x}}^{\delta,\epsilon}$ is the set of all traces that are **prediction stable** in the neighborhood $P_{\delta,\epsilon}(\mathbf{x})$ of a given input $\mathbf{x}$

Theorem

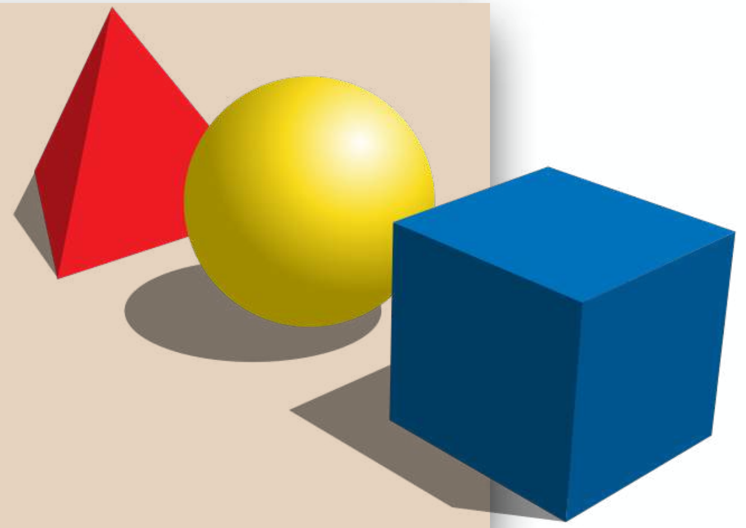
$$M \models \mathcal{R}_{\mathbf{x}}^{\delta,\epsilon} \Leftrightarrow \mathcal{M}[\![M]\!] \subseteq \mathcal{R}_{\mathbf{x}}^{\delta,\epsilon} \Leftrightarrow \mathcal{T}_p(P_{\delta,\epsilon}(\mathbf{x}))[\![M]\!] \subseteq \mathcal{R}_{\mathbf{x}}^{\delta,\epsilon}$$

Static Local Prediction Stability Analysis

Abstract Semantics

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

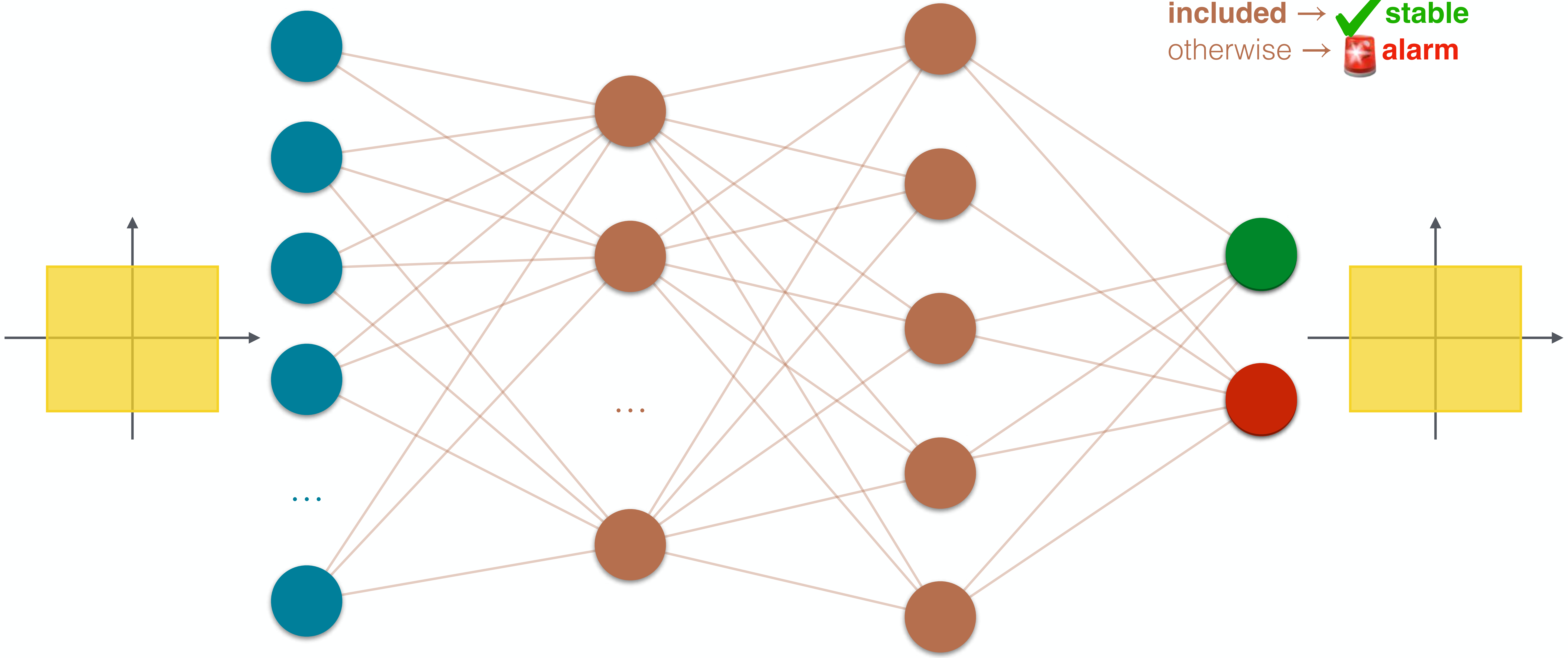


concrete semantics
mathematical models of the program behavior

Static Local Prediction Stability Analysis

Abstract Prefix Trace Semantics

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



① proceed **forwards from an abstraction** $P^{\#}_{\delta,\epsilon}$ of $P_{\delta,\epsilon}$

Theorem

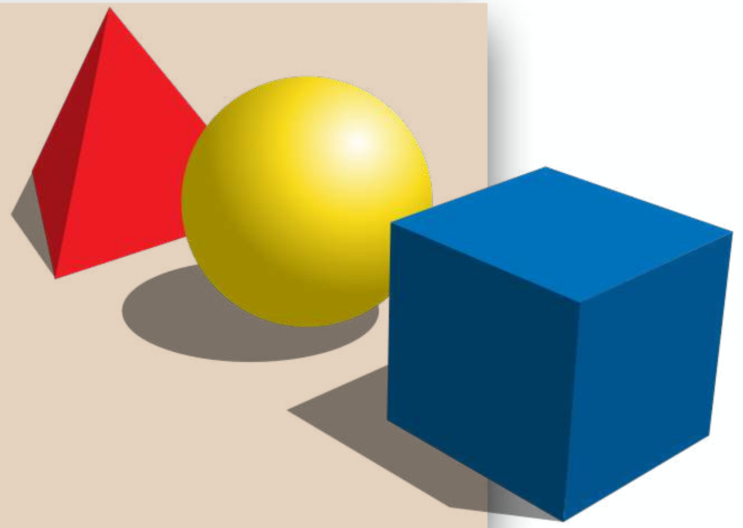
$$\mathcal{T}_p(P_{\delta,\epsilon}(\mathbf{x}))[\![M]\!] \subseteq \mathcal{T}^{\#}_p(P^{\#}_{\delta,\epsilon}(\mathbf{x}))[\![M]\!] \subseteq \mathcal{R}^{\delta,\epsilon}_{\mathbf{x}} \Rightarrow M \models \mathcal{R}^{\delta,\epsilon}_{\mathbf{x}}$$

Static Local Prediction Stability Analysis

Abstract Domain #1: Interval Domain

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

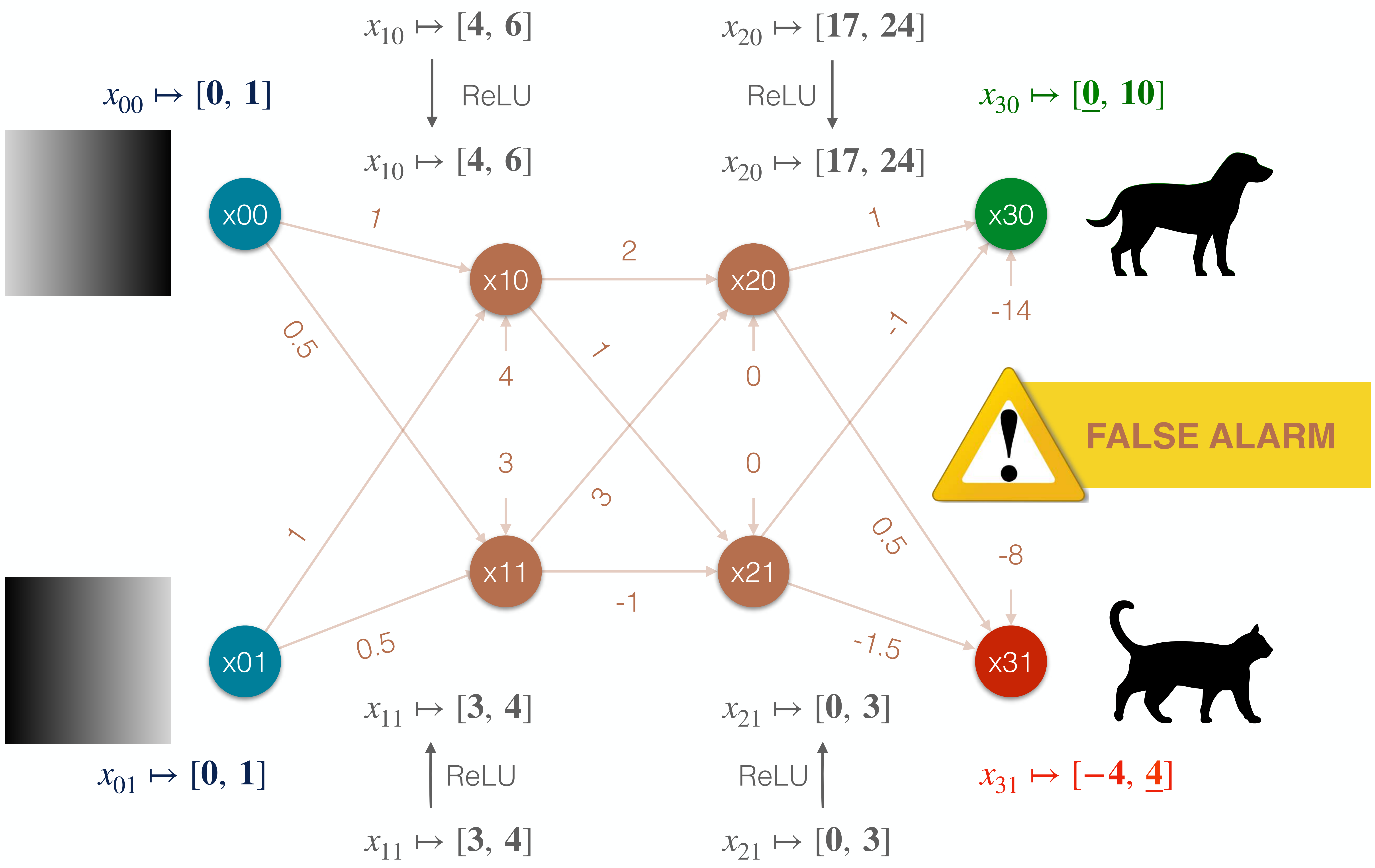


concrete semantics
mathematical models of the program behavior

Interval Domain

Example

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



Static Analysis by Abstract Interpretation

~~9.95€~~ 10€

~~35.85€~~ 40€

~~24.95€~~ 25€

~~4.85€~~ 5€

€ 10 +
€ 40 +
€ 25 +
€ 5

€ 80

€ 9.95 +
€ 35.85 +
€ 24.95 +
€ 4.85

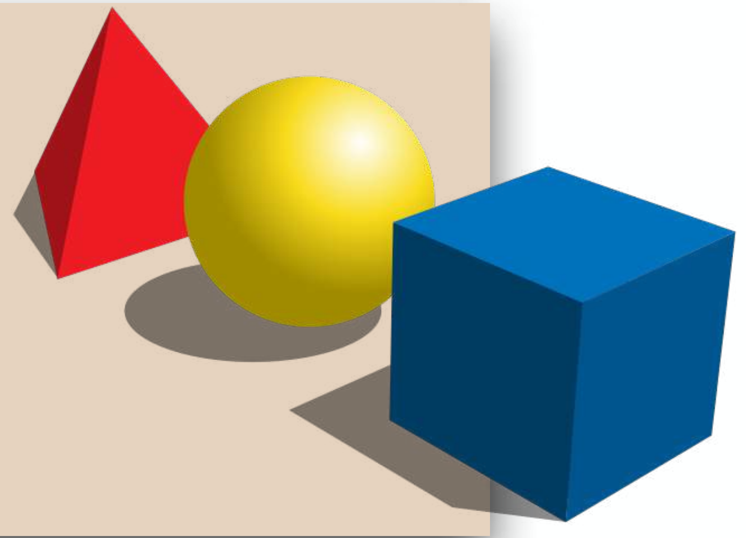
€ 75.60

Static Local Prediction Stability Analysis

Abstract Domain #2: Symbolic Domain

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

Symbolic Domain [Li et al. @ SAS 2019]

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

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

$x_{i-1,0} \mapsto \mathbf{E_{i-1,0}}$

$\dots$

$x_{i-1,j} \mapsto \mathbf{E_{i-1,j}}$

$\dots$

$$\xrightarrow{x_{i,j} = \sum_k w_{j,k}^{i-1} \cdot x_{i-1,k} + b_{i,j}}$$

$$x_{i,j} \mapsto \sum_k w_{j,k}^{i-1} \cdot \mathbf{E_{i-1,k}} + b_{i,j}$$

$$x_{i,j} \mapsto \begin{cases} \mathbf{E_{i,j}} \\ [\mathbf{a}, \mathbf{b}] \end{cases} \xrightarrow{\text{ReLU}}$$

$$x_{i,j} \mapsto \begin{cases} \mathbf{E_{i,j}} \\ [\mathbf{a}, \mathbf{b}] \end{cases} \quad 0 \leq a$$

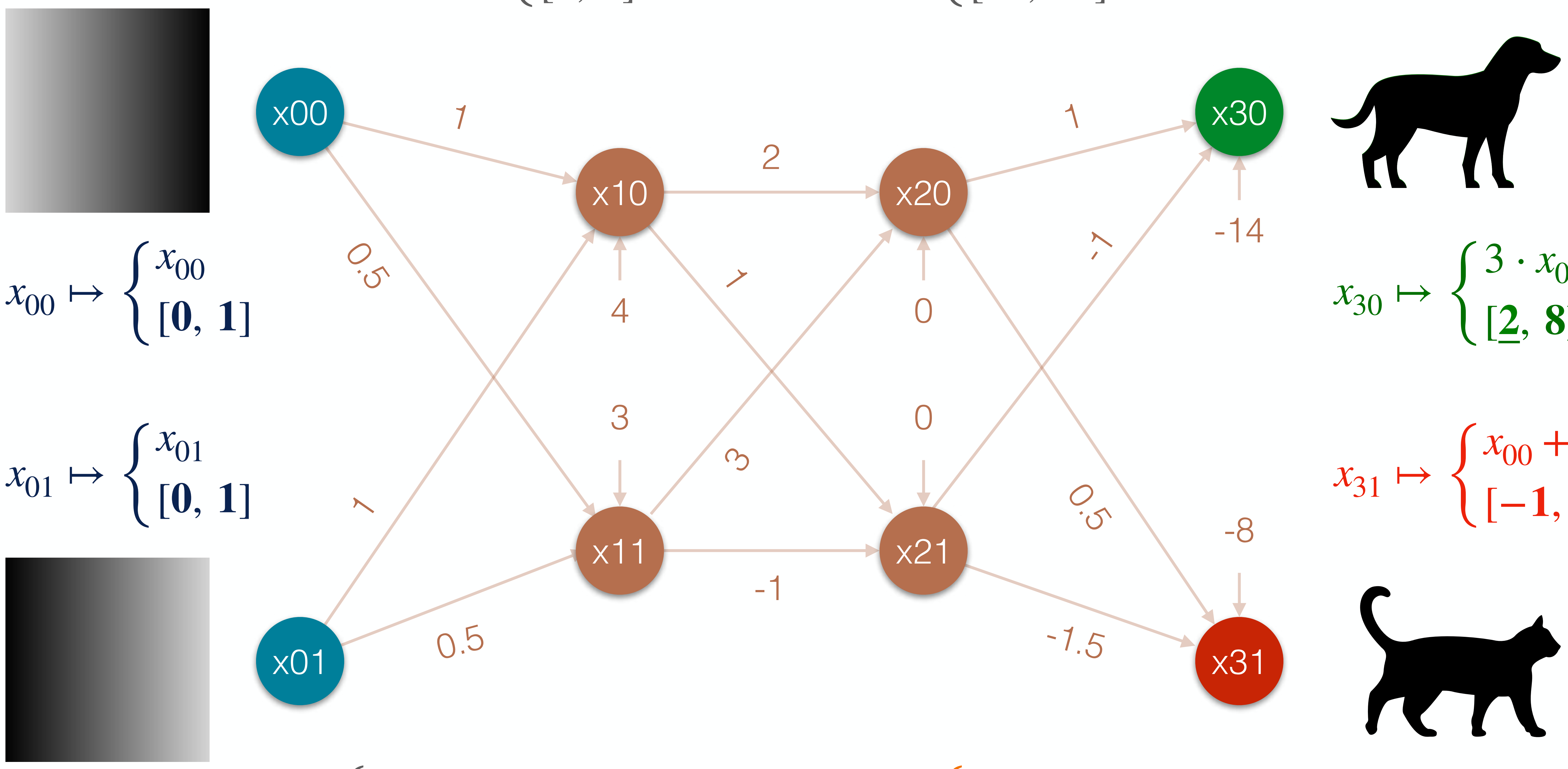
$$x_{i,j} \mapsto \begin{cases} \mathbf{x_{i,j}} \\ [\mathbf{0}, \mathbf{b}] \end{cases} \quad a < 0 \land 0 < b$$

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

Symbolic Domain

Example

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



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

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

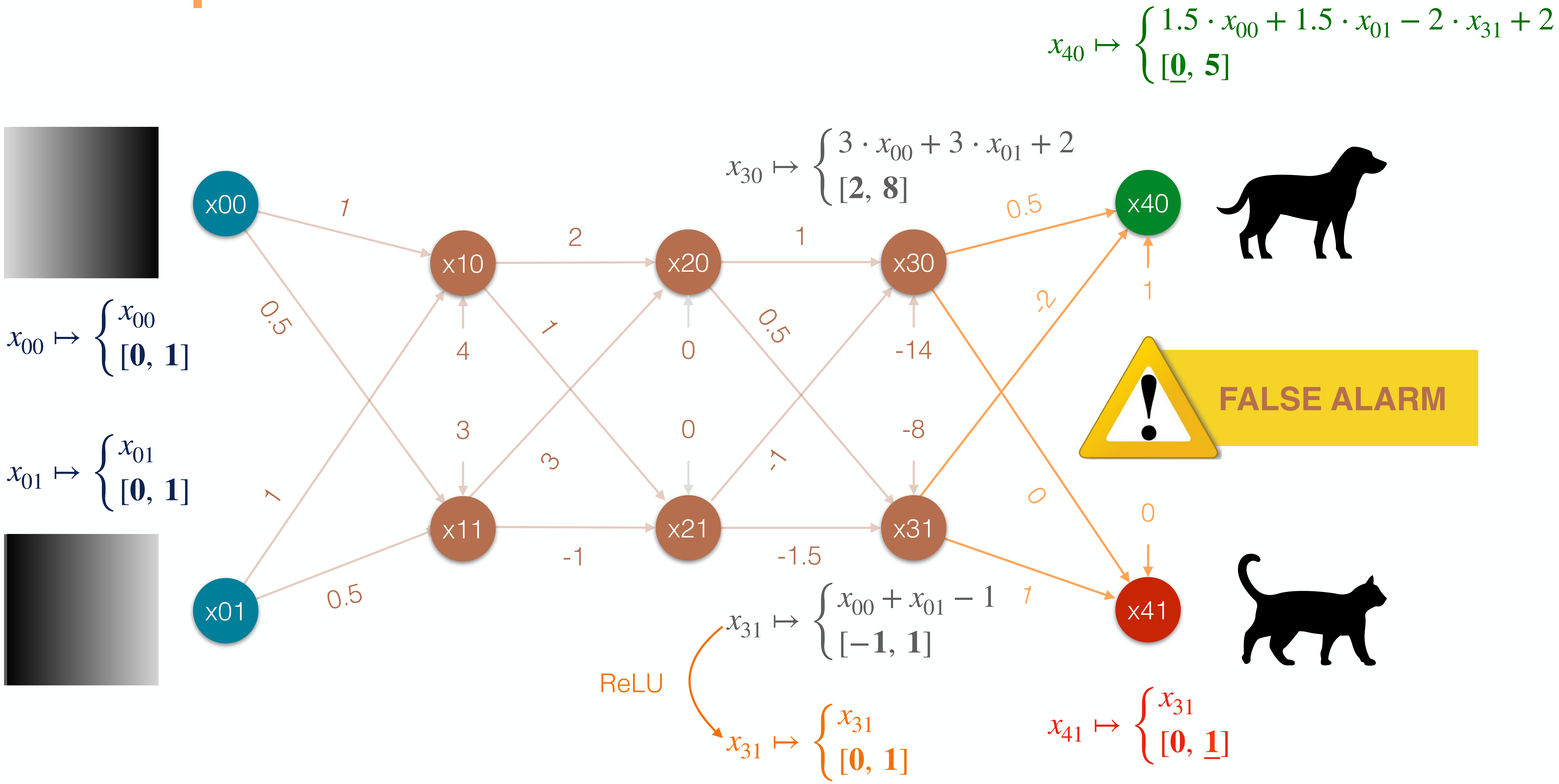
$$x_{30} \mapsto \begin{cases} 3 \cdot x_{00} + 3 \cdot x_{01} + 2 \\ [\underline{\mathbf{2}}, \mathbf{8}] \end{cases}$$

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

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

Symbolic Domain

Modified Example

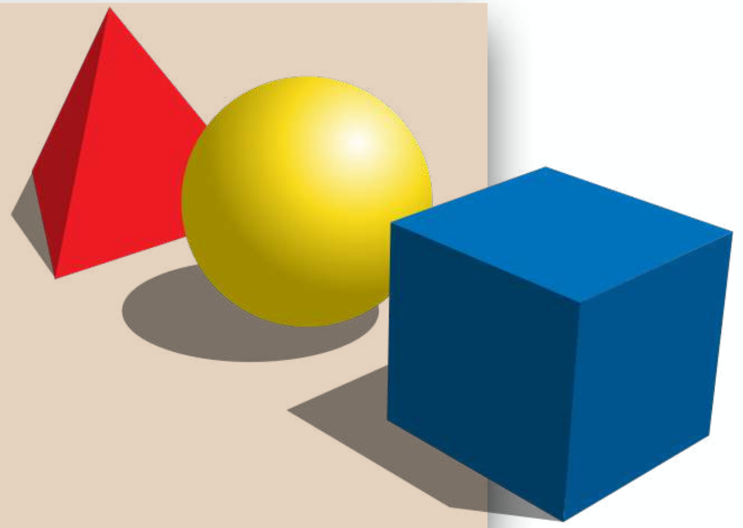


Static Local Prediction Stability Analysis

Abstract Domain #3: DeepPoly Domain

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

DeepPoly Domain [Singh et al. @ POPL 2019]

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 $a < 0 \land 0 < b$, $b \leq -a$:

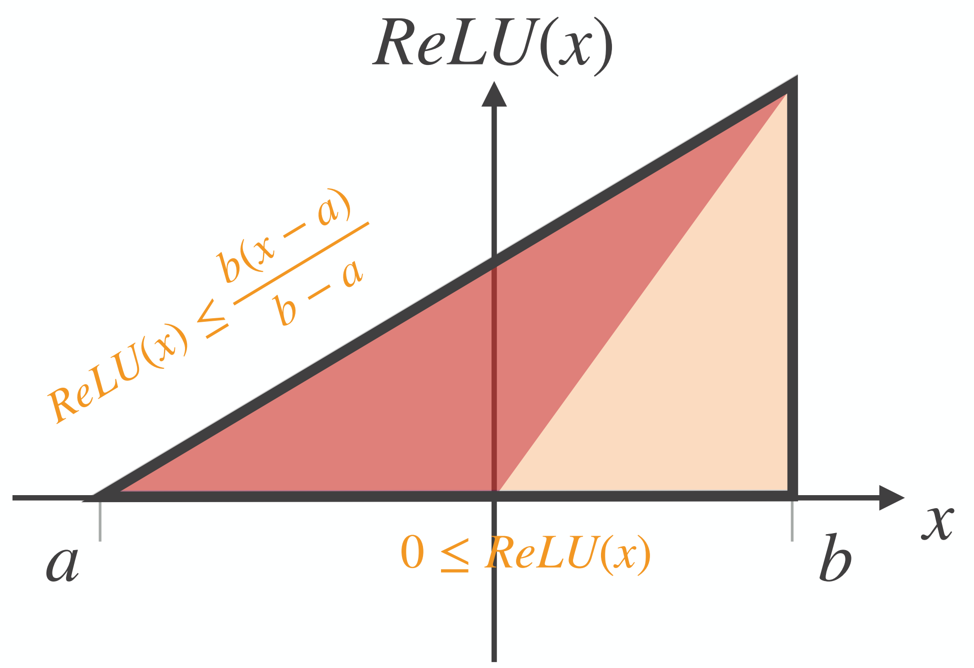
$$x_{i,j} \mapsto \begin{cases} \left[\mathbf{0}, \dfrac{\mathbf{b(x_{i,j} - a)}}{\mathbf{b - a}}\right] \\ [\mathbf{0}, \mathbf{b}] \end{cases}$$

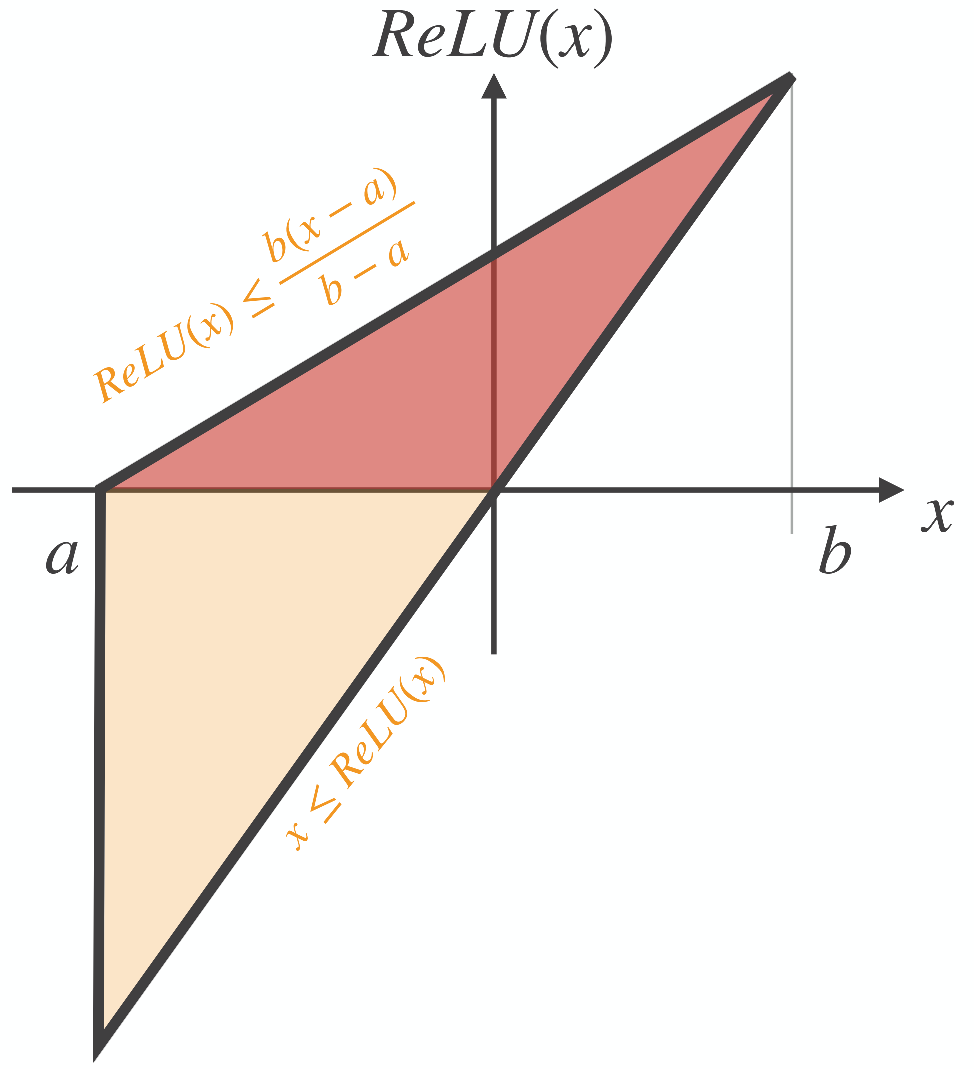
ReLU $a < 0 \land 0 < b$, $-a < b$:

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

ReLU $b \leq 0$:

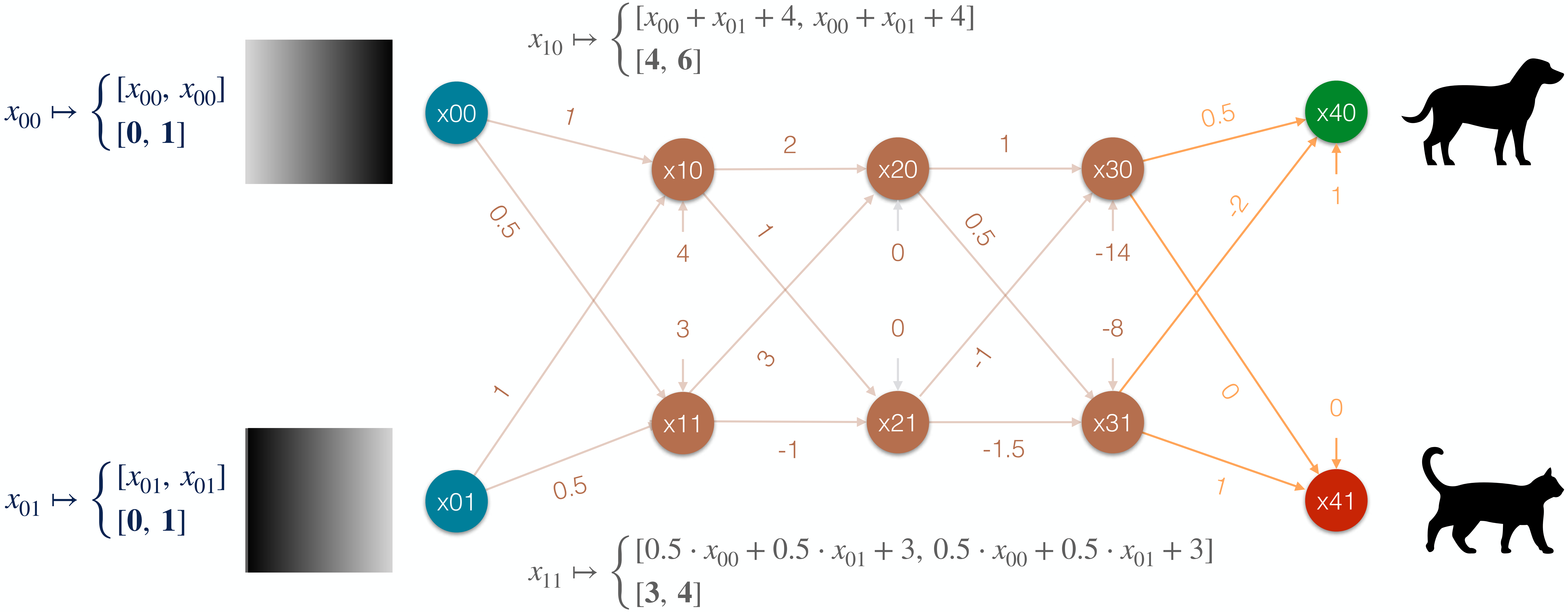
$$x_{i,j} \mapsto \begin{cases} [\mathbf{0}, \mathbf{0}] \\ [\mathbf{0}, \mathbf{0}] \end{cases}$$





DeepPoly Domain

Example

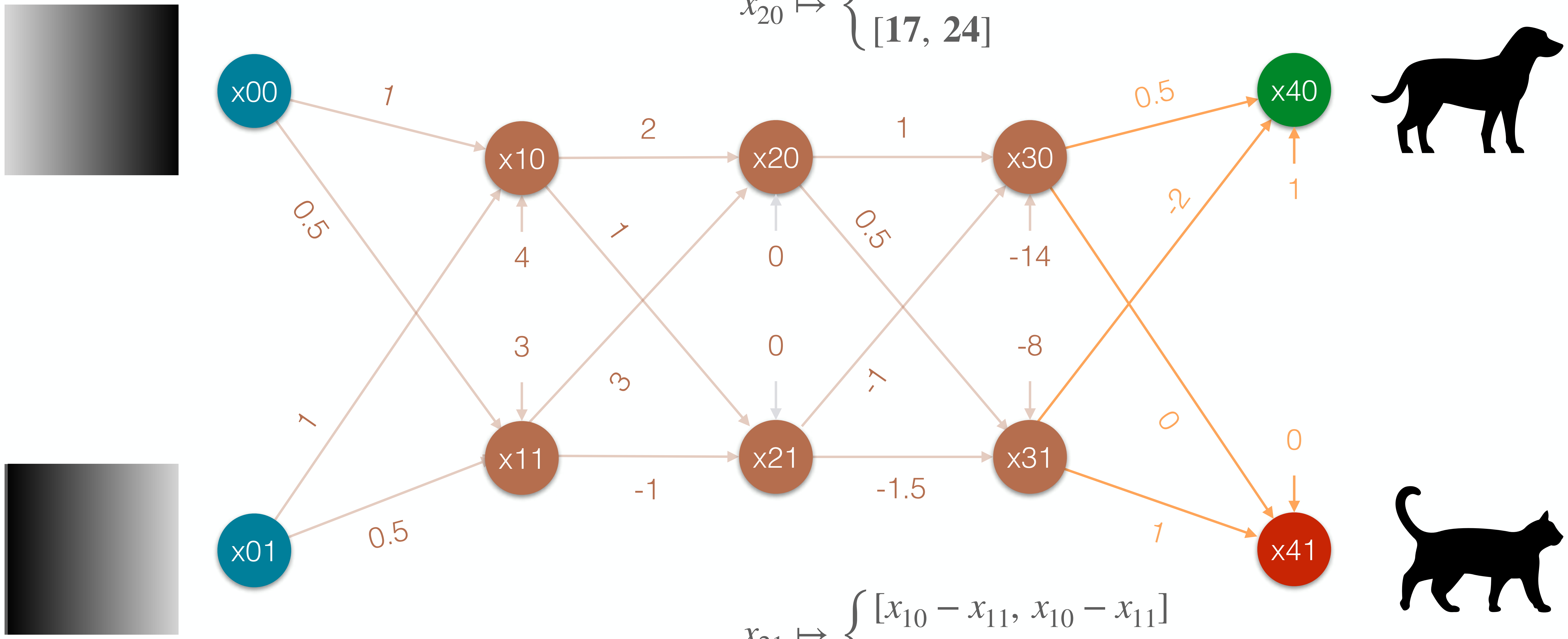


DeepPoly Domain

Example

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

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



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

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

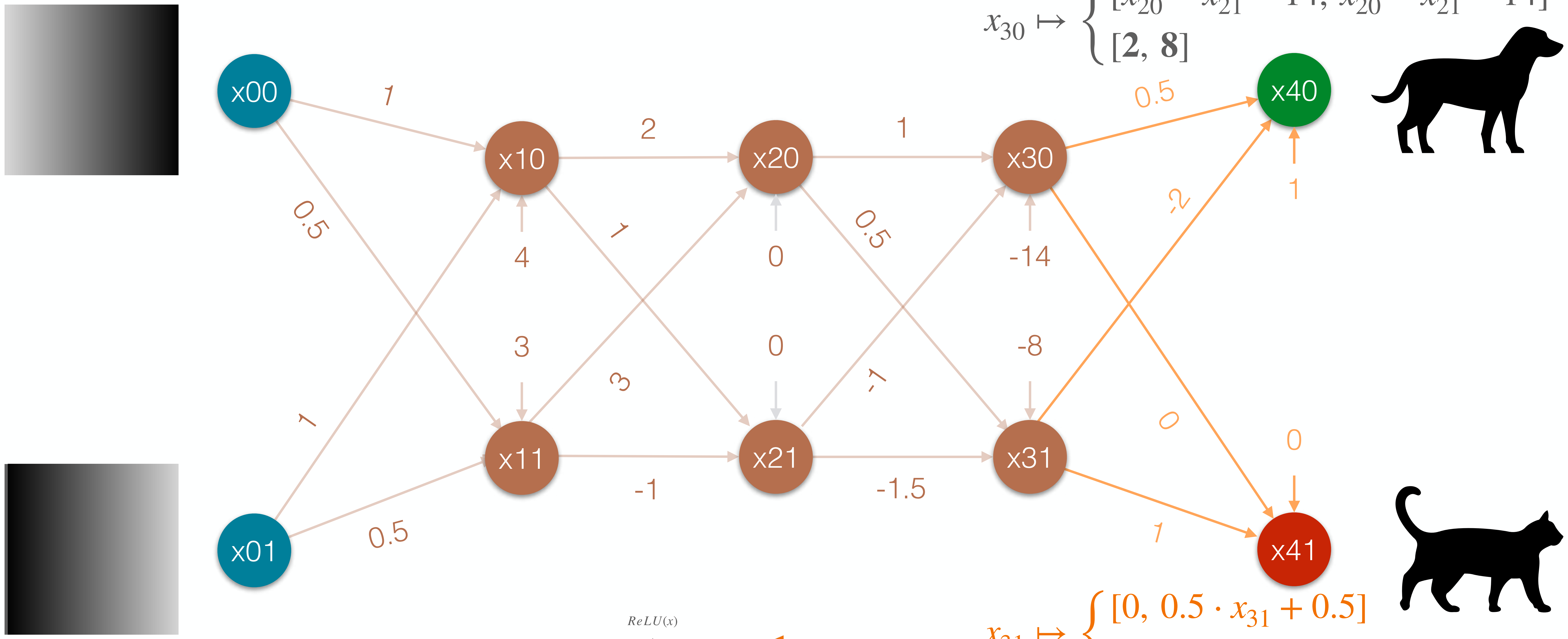
DeepPoly Domain

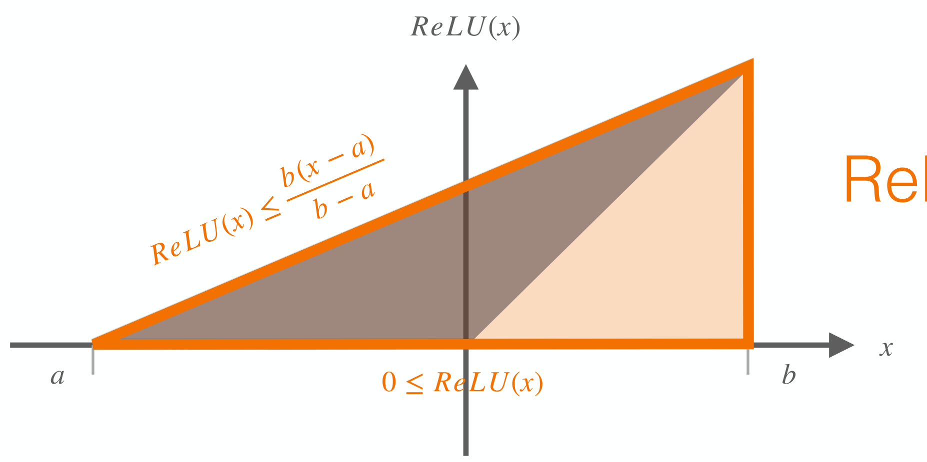
Example

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

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

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





ReLU

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

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

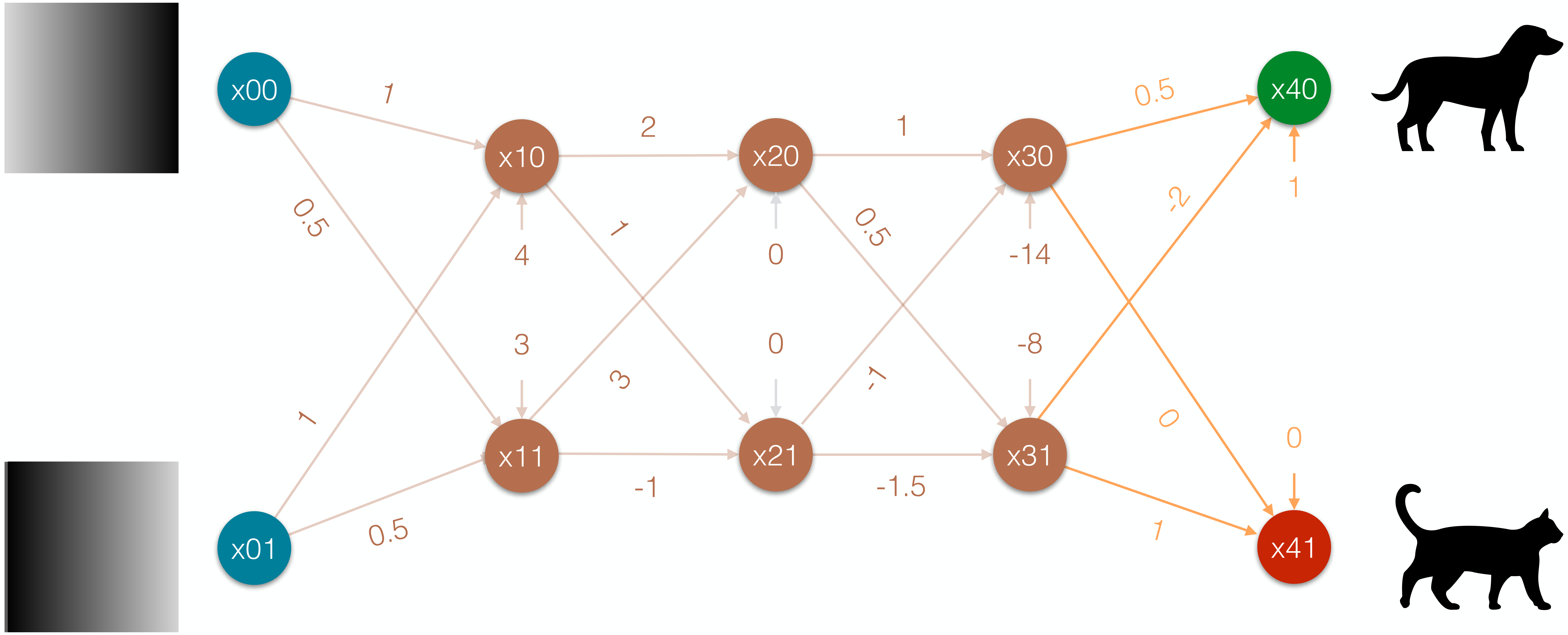
DeepPoly Domain

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 Domain

Back-Substitution

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

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

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

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

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

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

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

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

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

DeepPoly Domain

Back-Substitution

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

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

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

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

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

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

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

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

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

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

DeepPoly Domain

Back-Substitution

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

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

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

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

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

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

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

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

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

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

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

DeepPoly Domain

Back-Substitution

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

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

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

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

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

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

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

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

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

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

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

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

DeepPoly Domain

Partial Back-Substitution

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

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

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

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

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

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

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

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

$$
\begin{aligned}
x_{40} &\mapsto \begin{cases} [0.5 \cdot x_{30} - 2 \cdot x_{31} + 1,\ 0.5 \cdot x_{30} - 2 \cdot x_{31} + 1] \\ [\underline{\mathbf{0}},\ \mathbf{5}] \end{cases} \\
&\mapsto \begin{cases} [x_{21} + 1,\ 0.5 \cdot x_{20} - 0.5 \cdot x_{21} - 6] \\ [\mathbf{2},\ \underline{\mathbf{5.5}}] \end{cases} \\
&\mapsto \begin{cases} [x_{10} - x_{11} + 1,\ 0.5 \cdot x_{10} + 2 \cdot x_{11} - 6] \\ [\underline{\mathbf{1}},\ \mathbf{5}] \end{cases} \\
&\mapsto \begin{cases} [0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 2,\ 1.5 \cdot x_{00} + 1.5 \cdot x_{11} + 2] \\ [\mathbf{2},\ \mathbf{5}] \end{cases}
\end{aligned}
$$

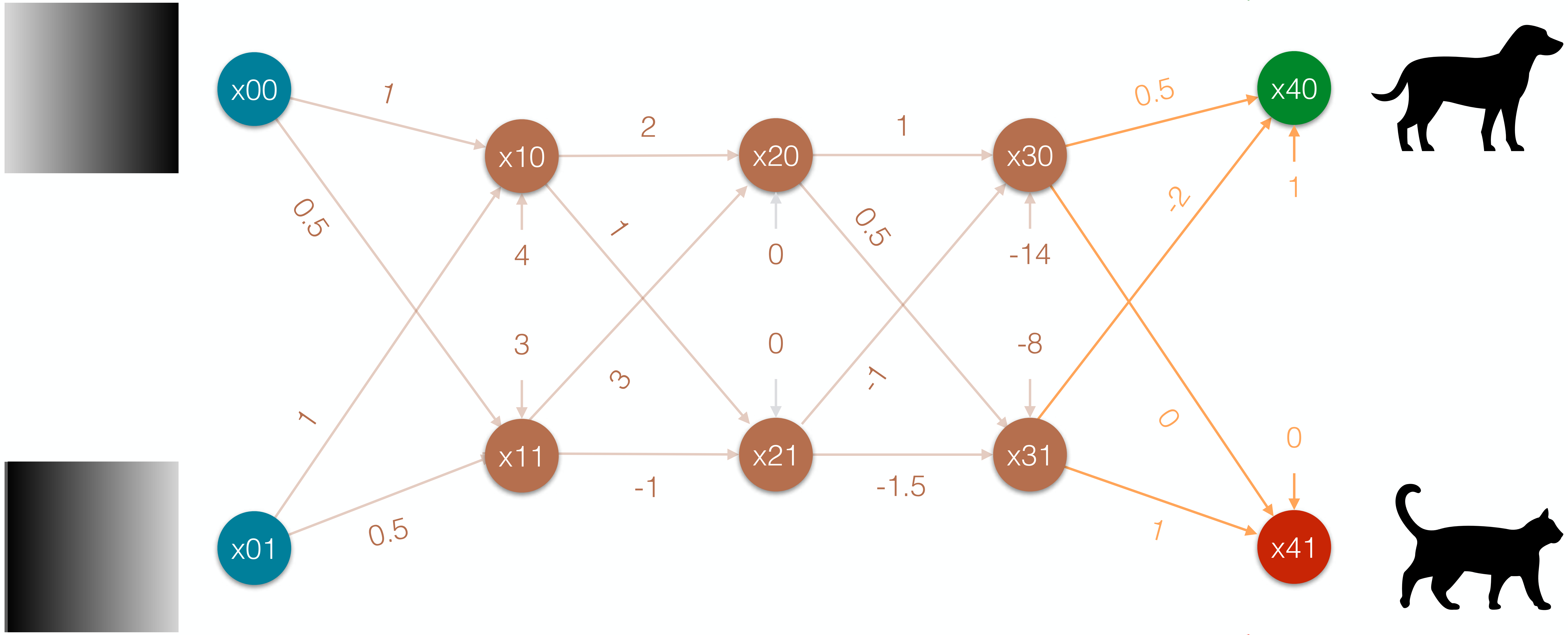
DeepPoly Domain

Example

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

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

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



$$x_{41} \mapsto \begin{cases} [x_{31},\ x_{31}] \end{cases}$$

DeepPoly Domain

Back-Substitution

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

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

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

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

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

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

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

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

$$x_{41} \mapsto \begin{cases} [x_{31}, x_{31}] \\ \end{cases}$$

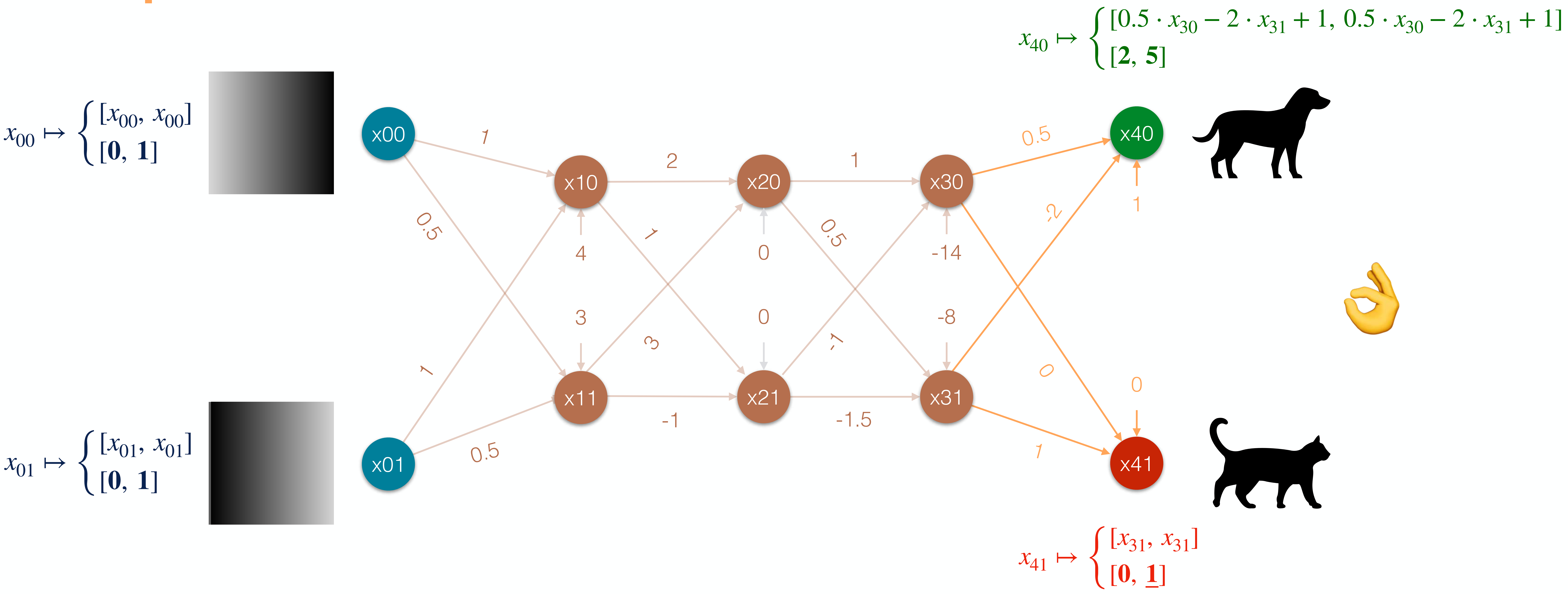
$$\mapsto \begin{cases} [0, 0.25 \cdot x_{20} - 0.75 \cdot x_{21} - 3.5] \\ \end{cases}$$

$$\mapsto \begin{cases} [0, \; - 0.25 \cdot x_{10} + 1.5 \cdot x_{11} - 3.5] \\ \end{cases}$$

$$\mapsto \begin{cases} [0, 0.5 \cdot x_{00} + 0.5 \cdot x_{01}] \\ [\mathbf{0}, \underline{\mathbf{1}}] \end{cases}$$

DeepPoly Domain

Example



DeepPoly Domain

Maintaining Symbolic Bounds wrt the Inputs (“à la Symbolic”)

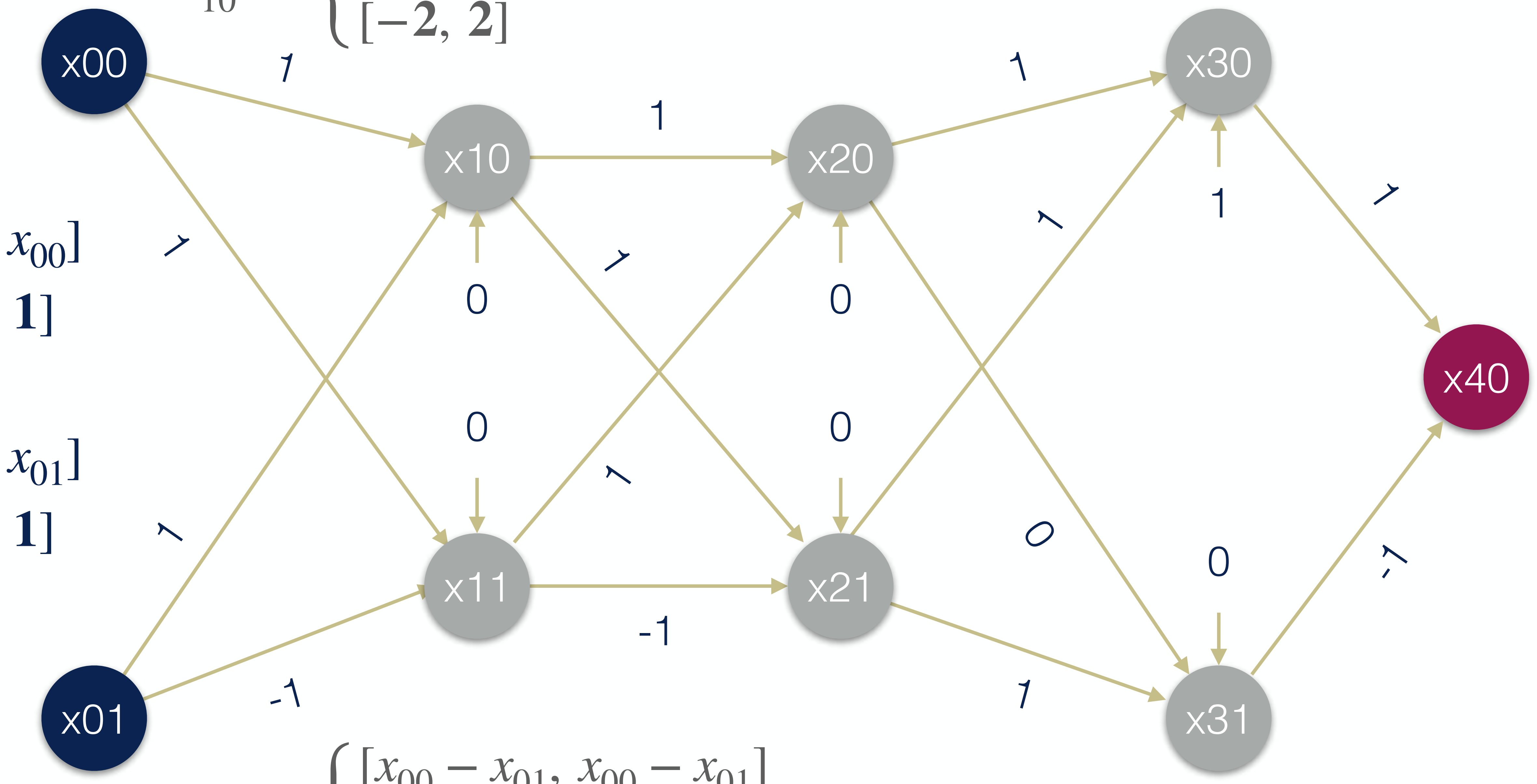
$x_{10} \mapsto [0,\ 0.5 \cdot x_{10} + 1] \to [0,\ 0.5 \cdot (x_{00} + x_{01}) + 1] = [0,\ 0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 1]$

ReLU

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

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

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



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

ReLU

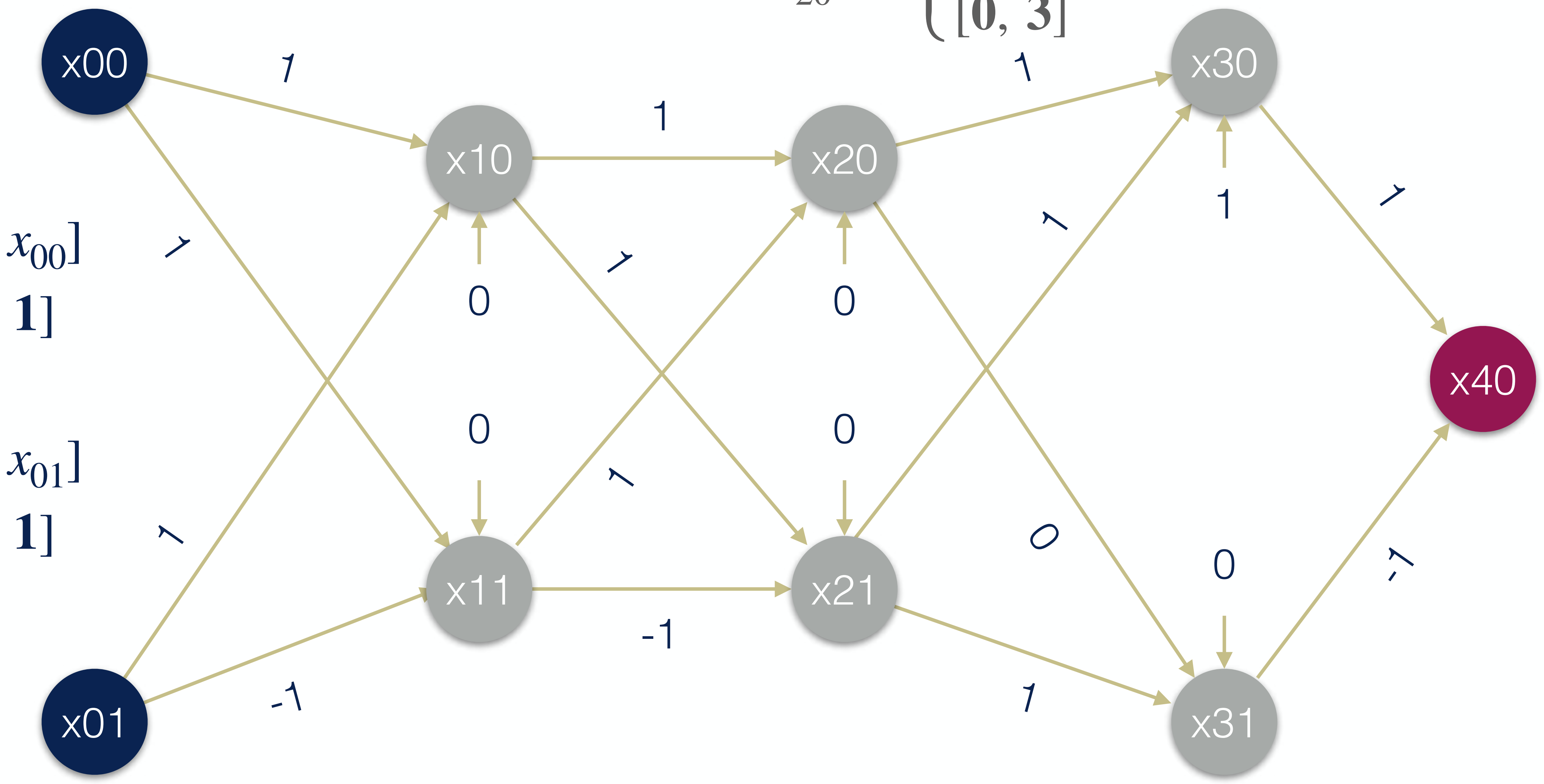
$x_{11} \mapsto [0,\ 0.5 \cdot x_{11} + 1] \to [0,\ 0.5 \cdot x_{00} - 0.5 \cdot x_{01} + 1]$

DeepPoly Domain

Maintaining Symbolic Bounds wrt the Inputs ("à la Symbolic")

$x_{10} \mapsto [0,\ 0.5 \cdot x_{10} + 1] \to [0,\ 0.5 \cdot (x_{00} + x_{01}) + 1] = [0,\ 0.5 \cdot x_{00} + 0.5 \cdot x_{01} + 1]$

$x_{20} \mapsto \begin{cases} [x_{10} + x_{11},\ x_{10} + x_{11}] \to [0,\ x_{00} + 2] \\ [\mathbf{0},\ \mathbf{3}] \end{cases}$



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

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

$x_{21} \mapsto \begin{cases} [x_{10} - x_{11},\ x_{10} - x_{11}] \to [-0.5 \cdot x_{00} + 0.5 \cdot x_{01} - 1,\ 0.5 \cdot x_{00} - 0.5 \cdot x_{01} + 1] \\ [\mathbf{-2},\ \mathbf{2}] \end{cases}$

$x_{11} \mapsto [0,\ 0.5 \cdot x_{11} + 1] \to [0,\ 0.5 \cdot x_{00} - 0.5 \cdot x_{01} + 1]$

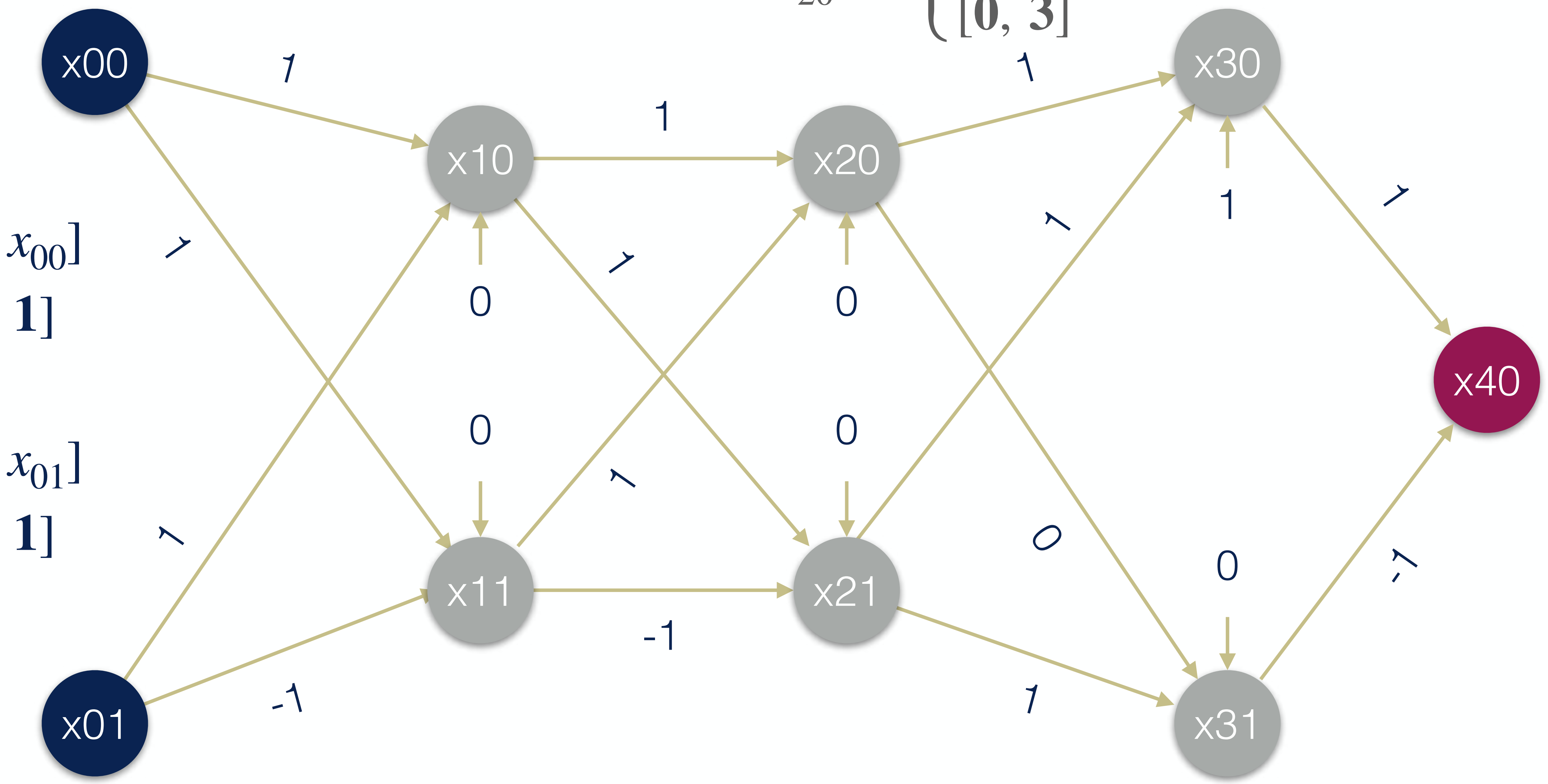
DeepPoly Domain

Maintaining Symbolic Bounds wrt the Inputs (“à la Symbolic”)

$$x_{20} \mapsto \begin{cases} [x_{10} + x_{11},\ x_{10} + x_{11}] \rightarrow [0,\ x_{00} + 2] \\ [\mathbf{0},\ \mathbf{3}] \end{cases}$$

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

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



$$x_{21} \mapsto \begin{cases} [x_{10} - x_{11},\ x_{10} - x_{11}] \rightarrow [-0.5 \cdot x_{00} + 0.5 \cdot x_{01} - 1,\ 0.5 \cdot x_{00} - 0.5 \cdot x_{01} + 1] \\ [\mathbf{-2},\ \mathbf{2}] \end{cases}$$

ReLU

$$x_{21} \mapsto [0,\ 0.5 \cdot x_{21} + 1] \rightarrow [0,\ 0.25 \cdot x_{00} - 0.25 \cdot x_{01} + 1.5]$$

DeepPoly Domain

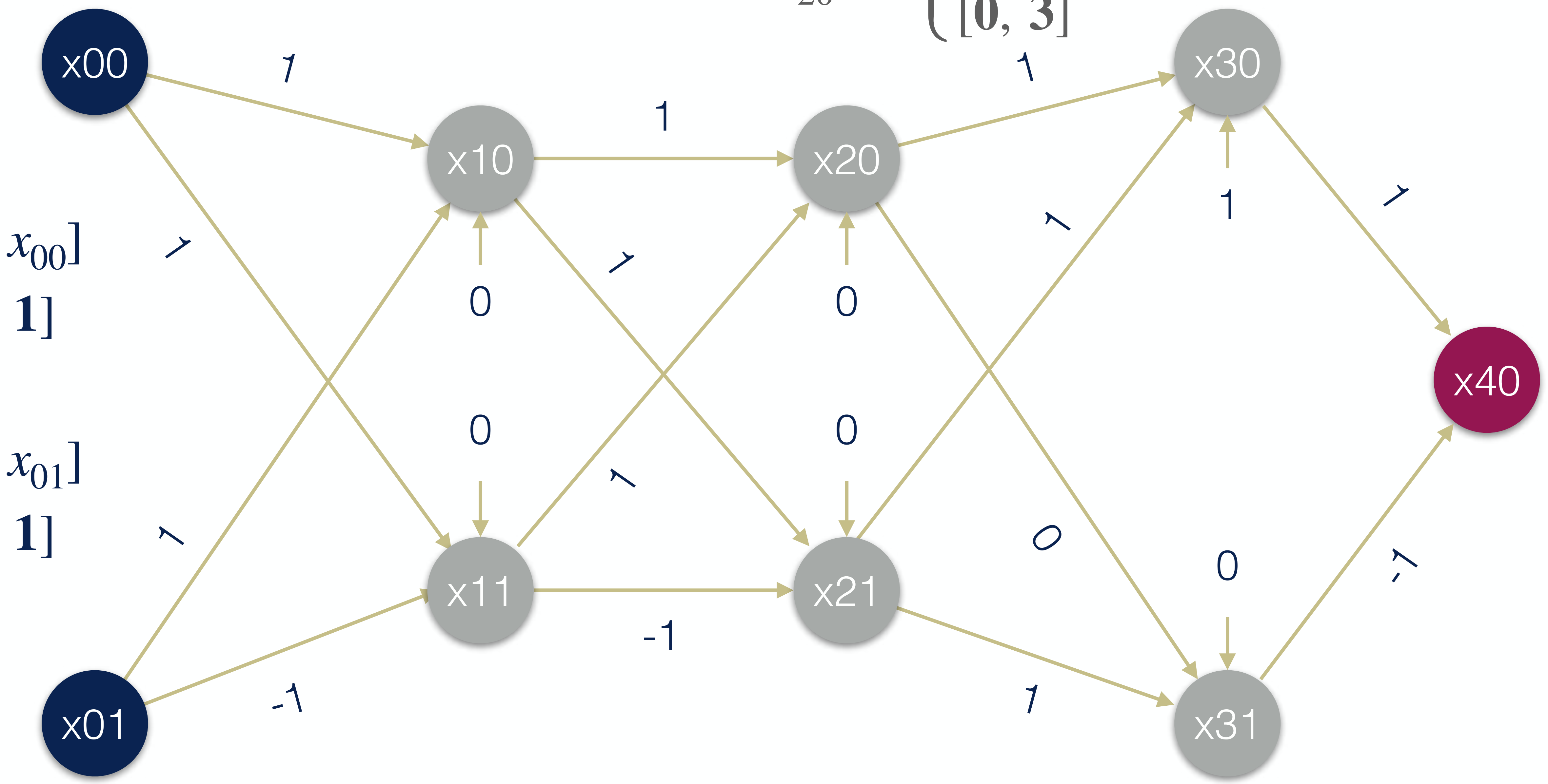
Maintaining Symbolic Bounds wrt the Inputs (“à la Symbolic”)

$$x_{30} \mapsto \begin{cases} [x_{20} + x_{21} + 1,\ x_{20} + x_{21} + 1] \rightarrow [1,\ 1.25 \cdot x_{00} - 0.25 \cdot x_{01} + 4.5] \\ [\mathbf{1},\ \underline{\mathbf{6}}] \quad \leftarrow [\mathbf{1}, \mathbf{5.5}] \text{ with back-substitution} \end{cases}$$

$$x_{20} \mapsto \begin{cases} [x_{10} + x_{11},\ x_{10} + x_{11}] \rightarrow [0,\ x_{00} + 2] \\ [\mathbf{0},\ \mathbf{3}] \end{cases}$$

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

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



$$x_{31} \mapsto \begin{cases} [x_{21},\ x_{21}] \rightarrow [0,\ 0.25 \cdot x_{00} - 0.25 \cdot x_{01} + 1.5] \\ [\mathbf{0},\ \mathbf{2}] \end{cases}$$

$x_{21} \mapsto [0,\ 0.5 \cdot x_{21} + 1] \rightarrow [0,\ 0.25 \cdot x_{00} - 0.25 \cdot x_{01} + 1.5]$

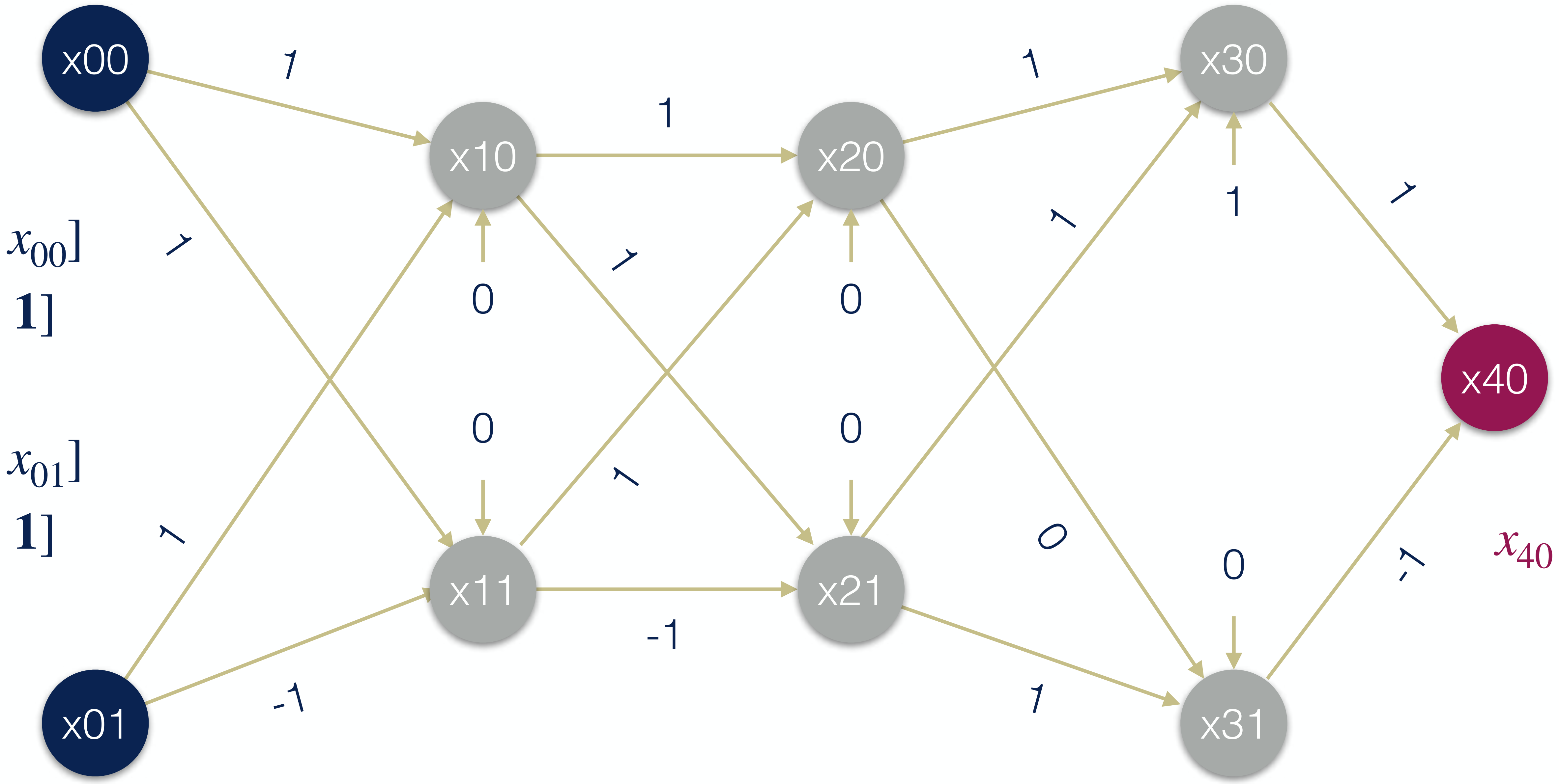
DeepPoly Domain

Maintaining Symbolic Bounds wrt the Inputs (“à la Symbolic”)

$x_{30} \mapsto \begin{cases} [x_{20} + x_{21} + 1,\ x_{20} + x_{21} + 1] \to [1,\ 1.25 \cdot x_{00} - 0.25 \cdot x_{01} + 4.5] \\ [\mathbf{1},\ \underline{\mathbf{6}}] \end{cases}$ $\leftarrow [\mathbf{1}, \mathbf{5.5}]$ with back-substitution

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

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



$x_{40} \mapsto \begin{cases} \cdots \\ [\mathbf{-1},\ \underline{\mathbf{6}}] \end{cases}$ $\leftarrow [\mathbf{1}, \mathbf{4}]$ with back-substitution

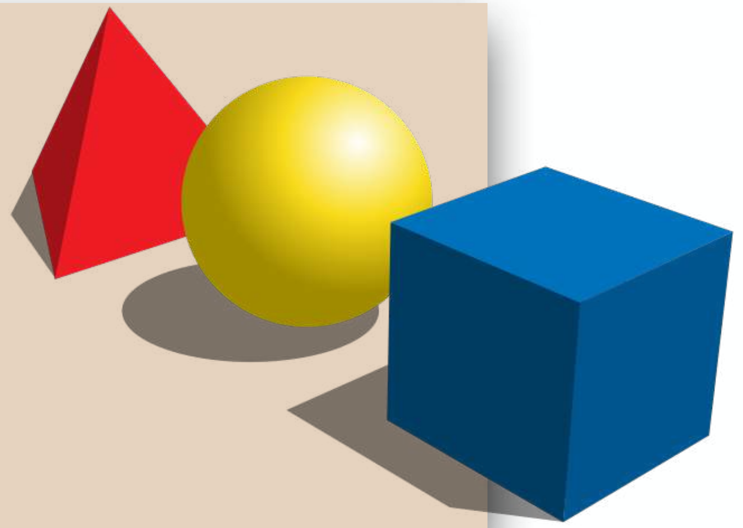
$x_{31} \mapsto \begin{cases} [x_{21},\ x_{21}] \to [0,\ 0.25 \cdot x_{00} - 0.25 \cdot x_{01} + 1.5] \\ [\mathbf{0},\ \mathbf{2}] \end{cases}$

Static Local Prediction Stability Analysis

Going Farther: Multi-Neuron Abstractions

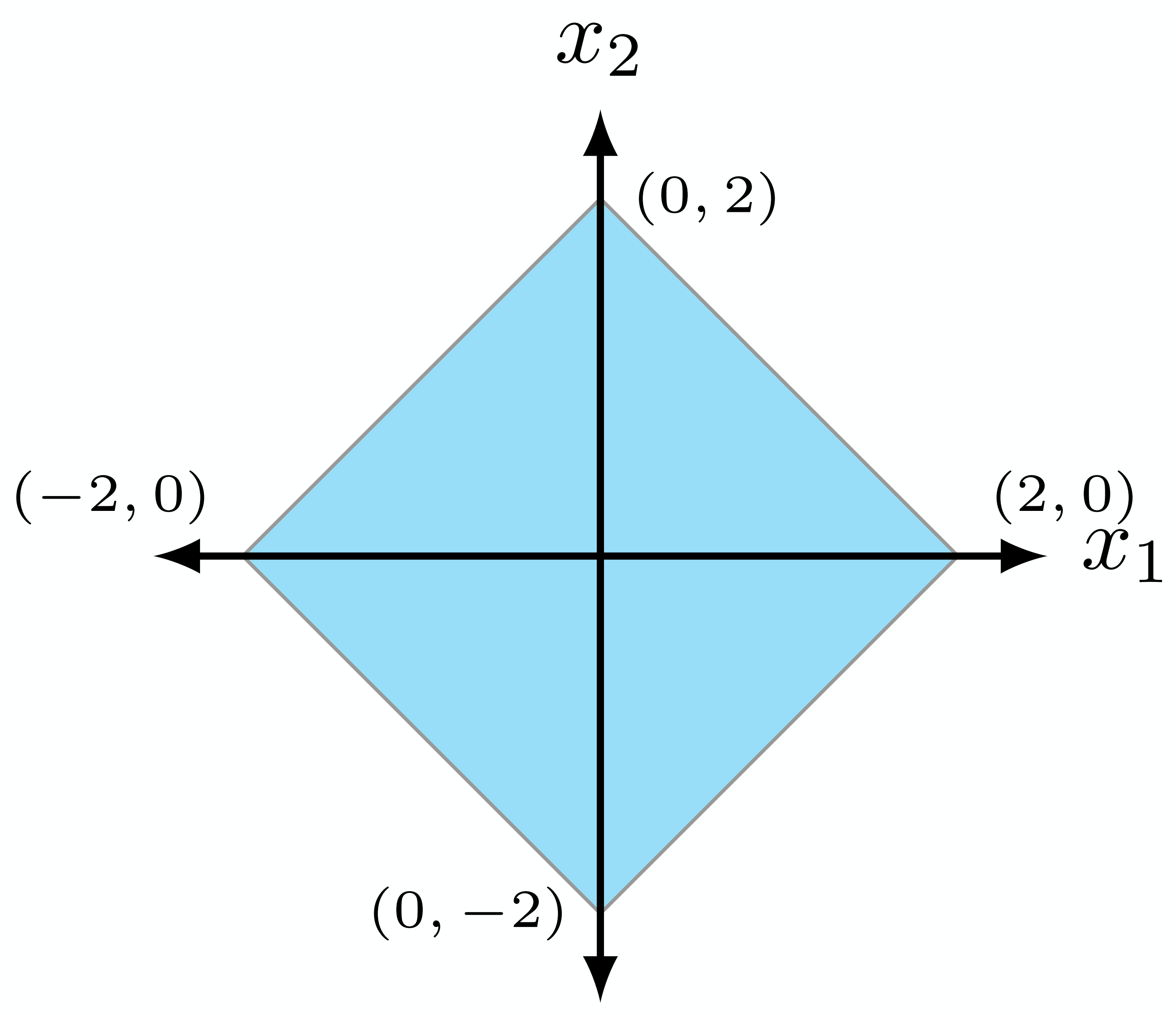
practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

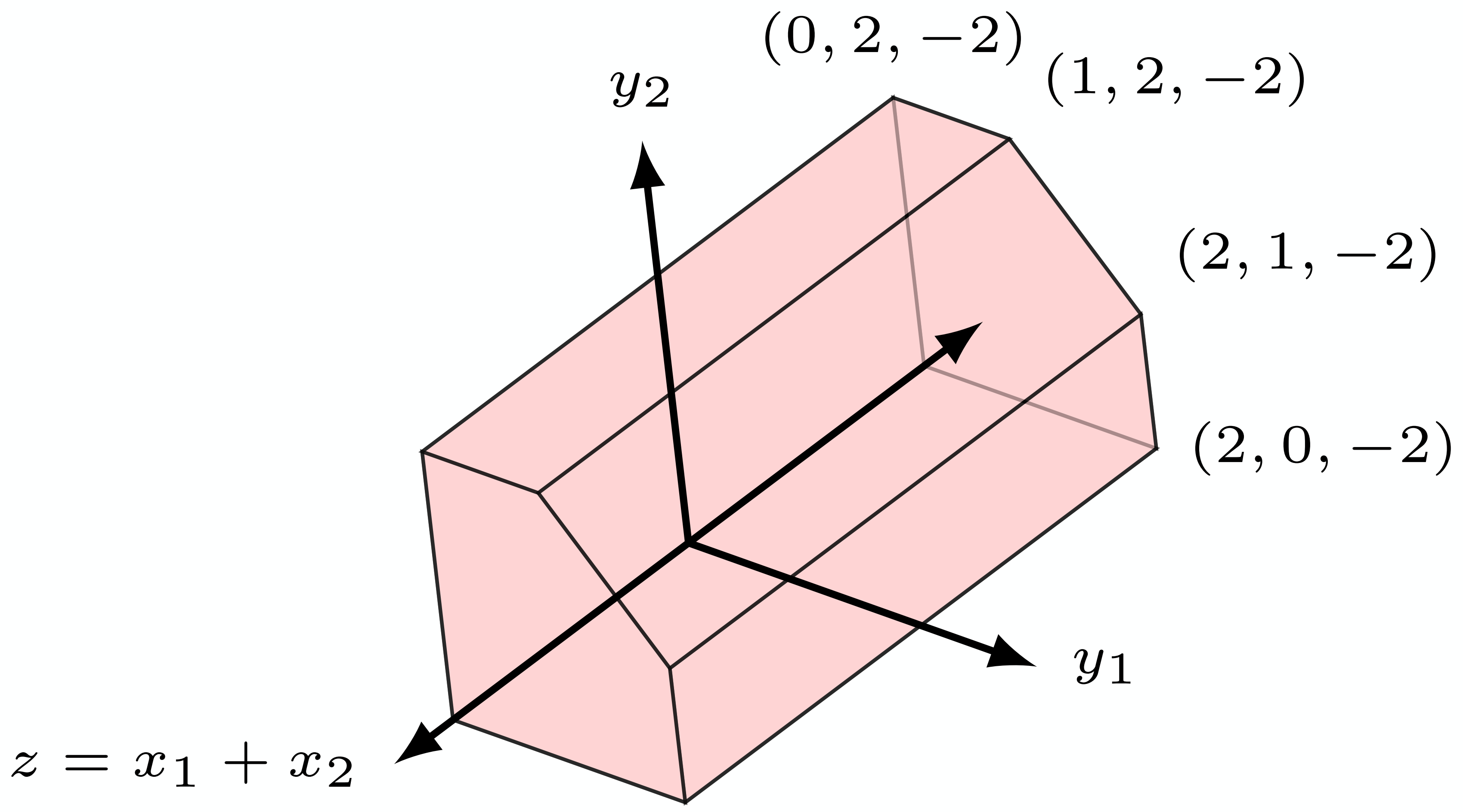


concrete semantics
mathematical models of the program behavior

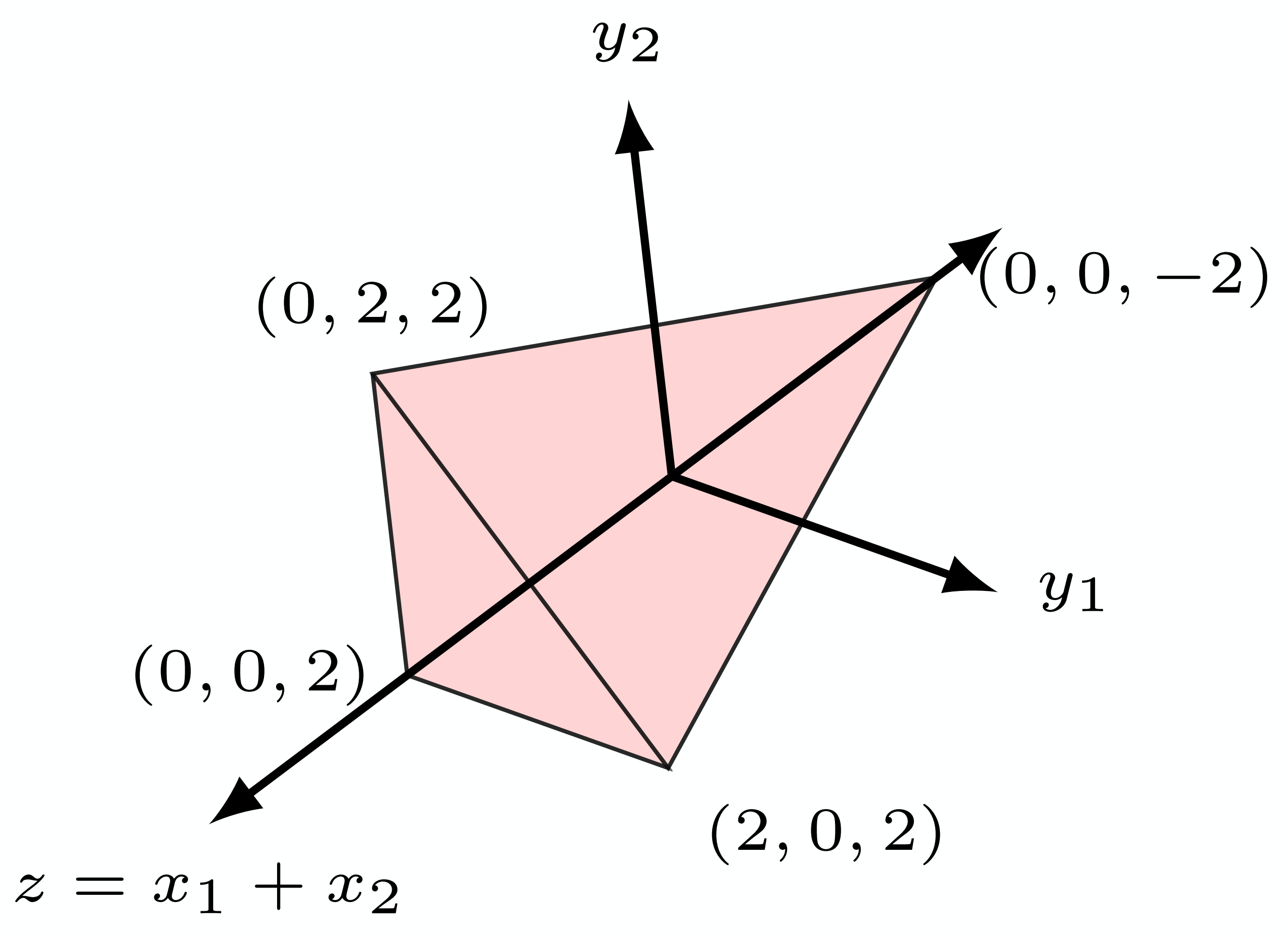
Multi-Neuron Abstractions [Singh et al. @ NeurIPS 2019]



(a) Input shape



(b) 1-ReLU



(c) 2-ReLU

Static Local Prediction Stability Analysis

An Exercise Worth Trying 😉

practical tools
targeting specific programs

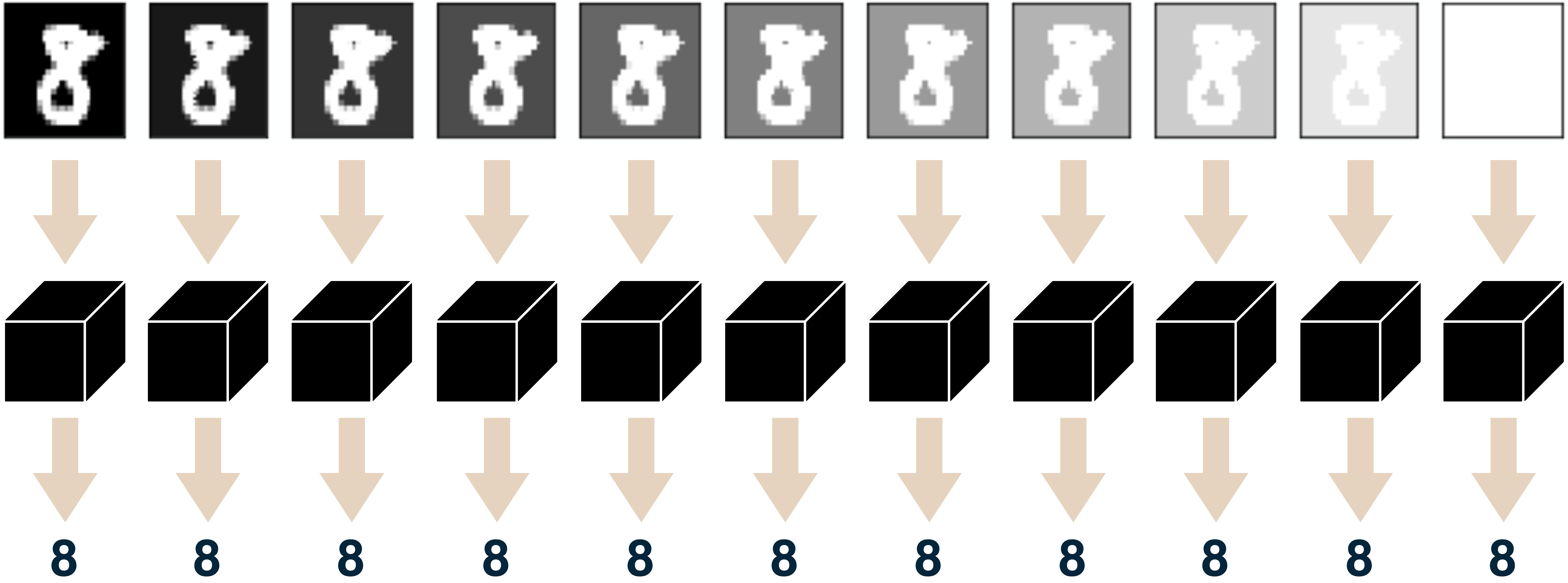
abstract semantics, abstract domains
algorithmic approaches to decide program properties

concrete semantics
mathematical models of the program behavior

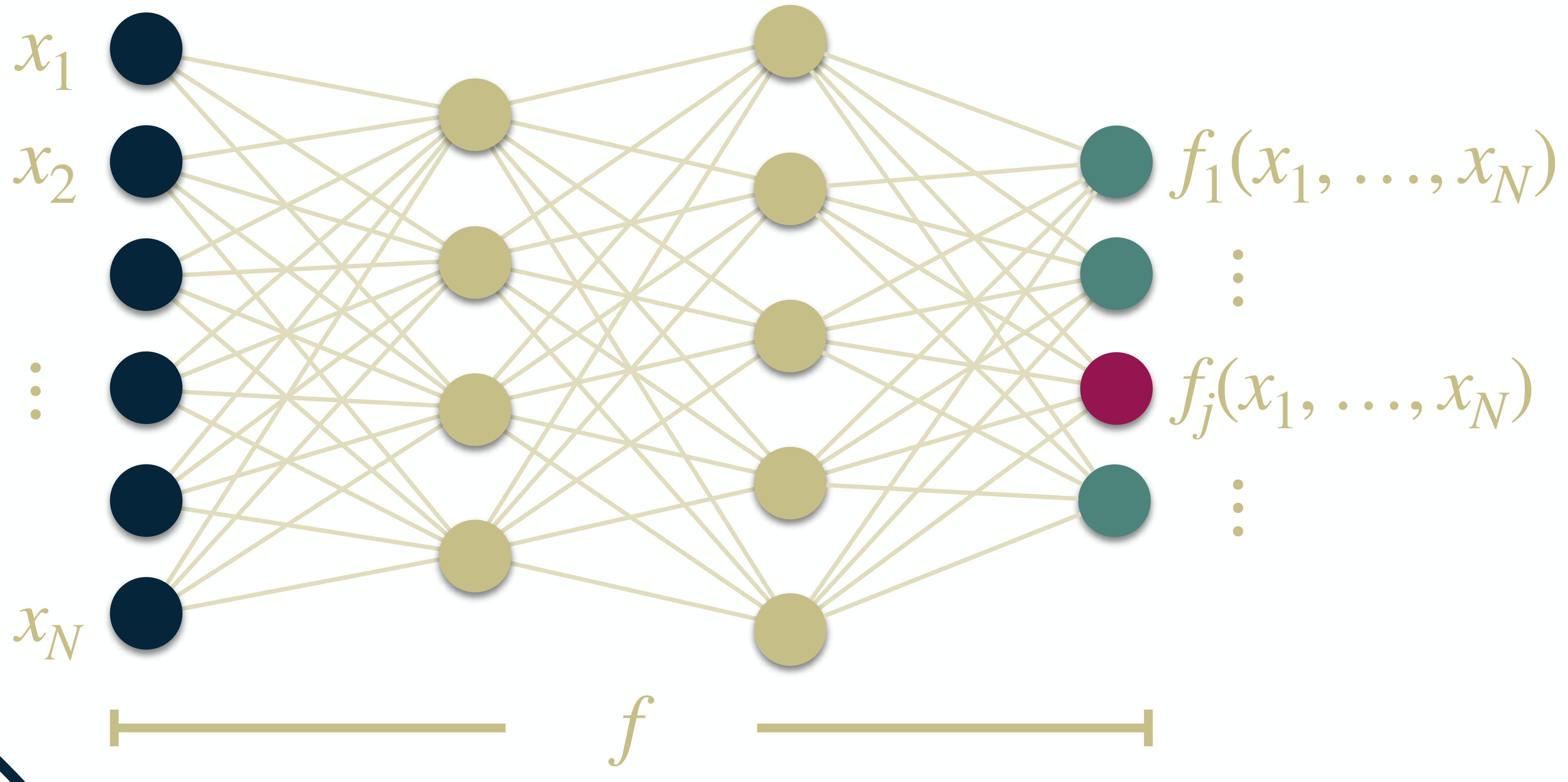
Saliency Map Stability

Local Prediction Stability

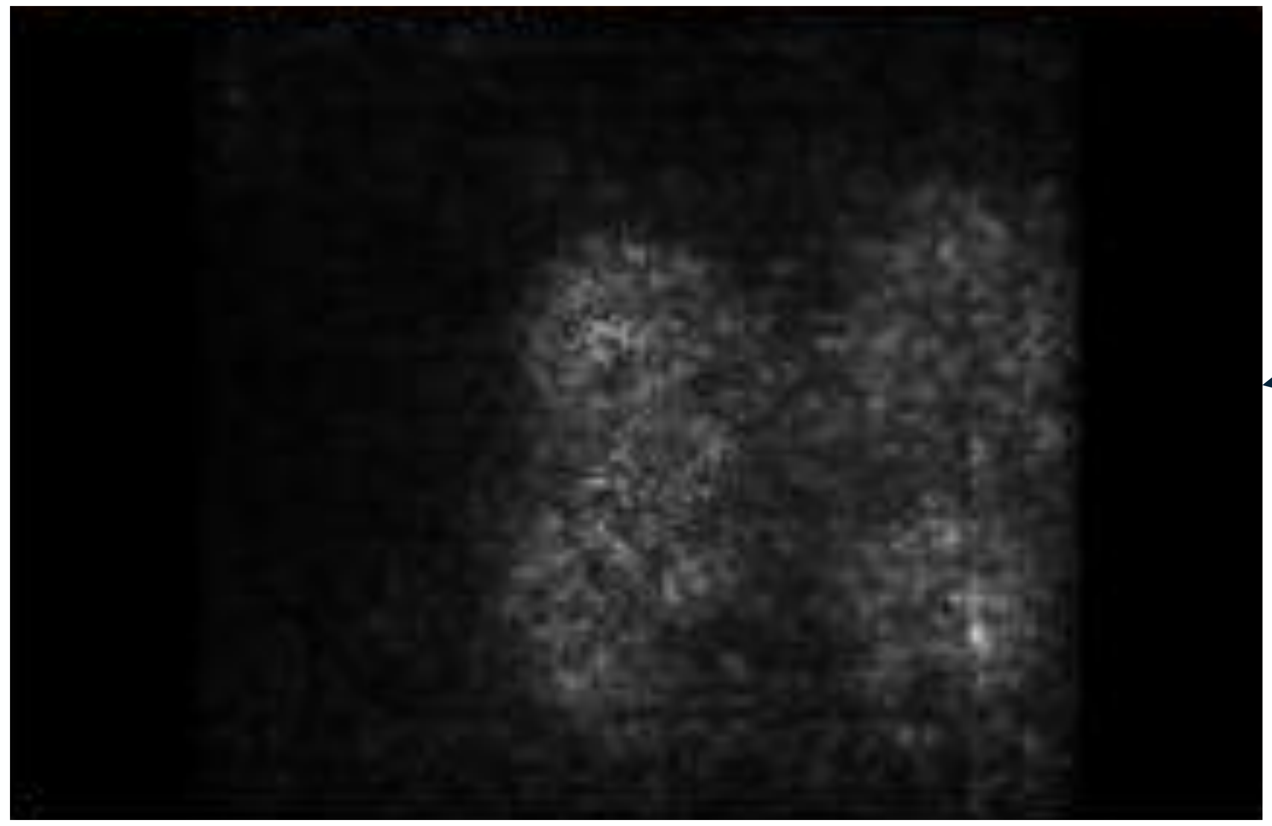
Not Enough!



Saliency Maps [Simonyan & al. @ ICLR 2014]



Dog



$$\text{map}_j(x) = \left| \frac{\partial f_j(x_1, \ldots, x_N)}{\partial x_1} \right| \; \ldots \; \left| \frac{\partial f_j(x_1, \ldots, x_N)}{\partial x_N} \right|$$

Local Saliency Map Stability

Reading Suggestion

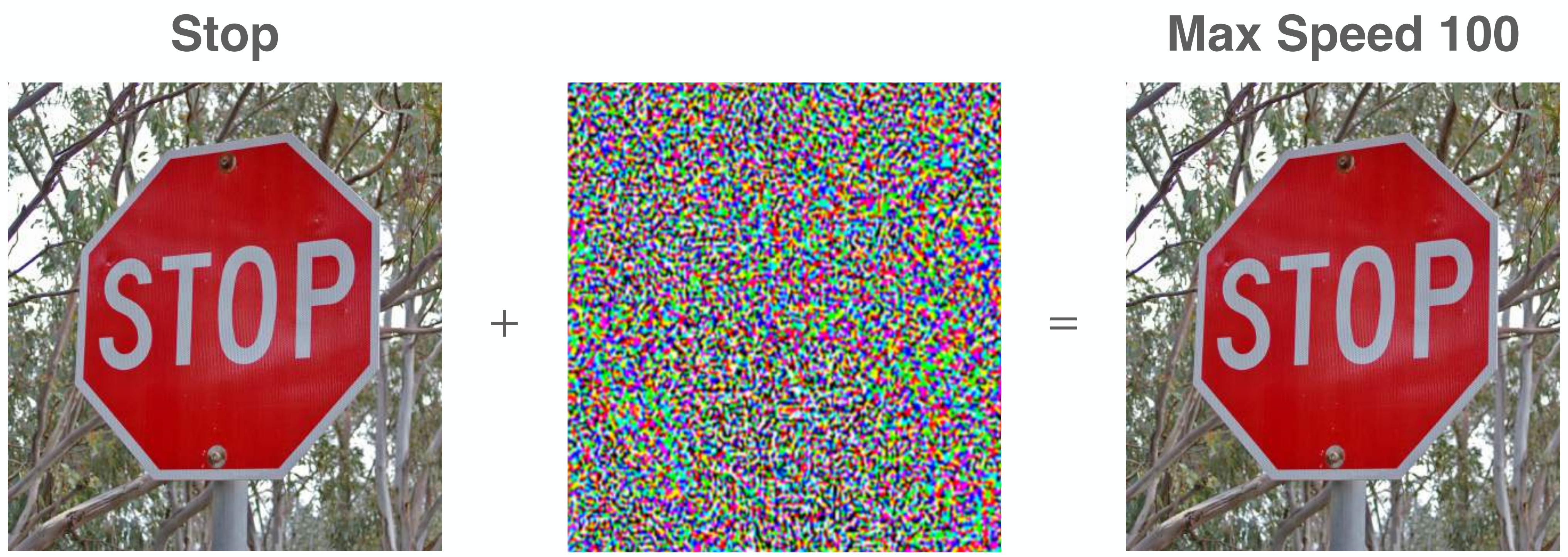
Verifying Attention Robustness of Deep Neural Networks against Semantic Perturbations

Satoshi Munakata[1], Caterina Urban[2], Haruki Yokoyama[1], Koji Yamamoto[1], and Kazuki Munakata[1]

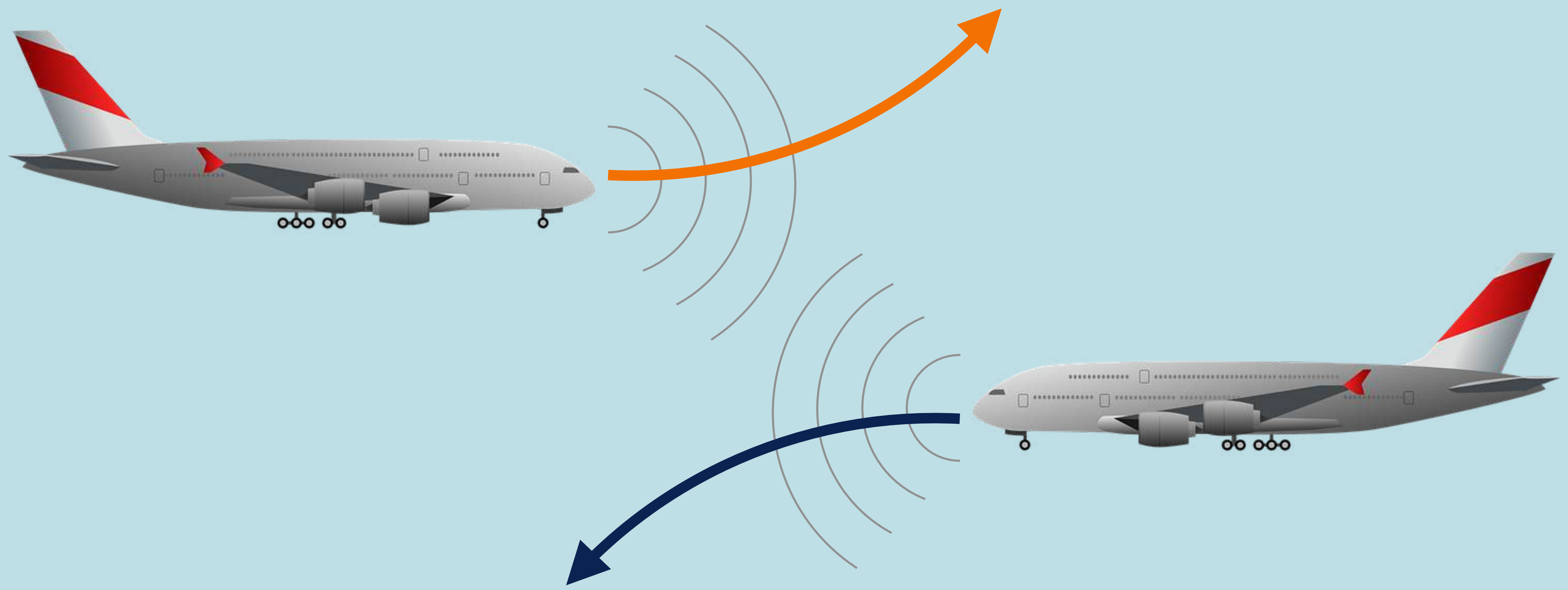
[1] Fujitsu, Kanagawa, Japan
[2] Inria & ENS | PSL & CNRS, Paris, France

Abstract. It is known that deep neural networks (DNNs) classify an input image by paying particular attention to certain specific pixels; a graphical representation of the magnitude of attention to each pixel is called a *saliency-map*. Saliency-maps are used to check the validity of the classification decision basis, e.g., it is not a valid basis for classification if a DNN pays more attention to the background rather than the subject of an image. Semantic perturbations can significantly change the saliency-map. In this work, we propose the first verification method for *attention robustness*, i.e., the local robustness of the changes in the saliency-map against combinations of semantic perturbations. Specifically, our method determines the range of the perturbation parameters (e.g., the brightness change) that maintains the difference between the actual saliency-map change and the expected saliency-map change below a given threshold value. Our method is based on activation region traversals, focusing on the outermost robust boundary for scalability on larger DNNs. We empirically evaluate the effectiveness and performance of our method on DNNs trained on popular image classification datasets.

Stability



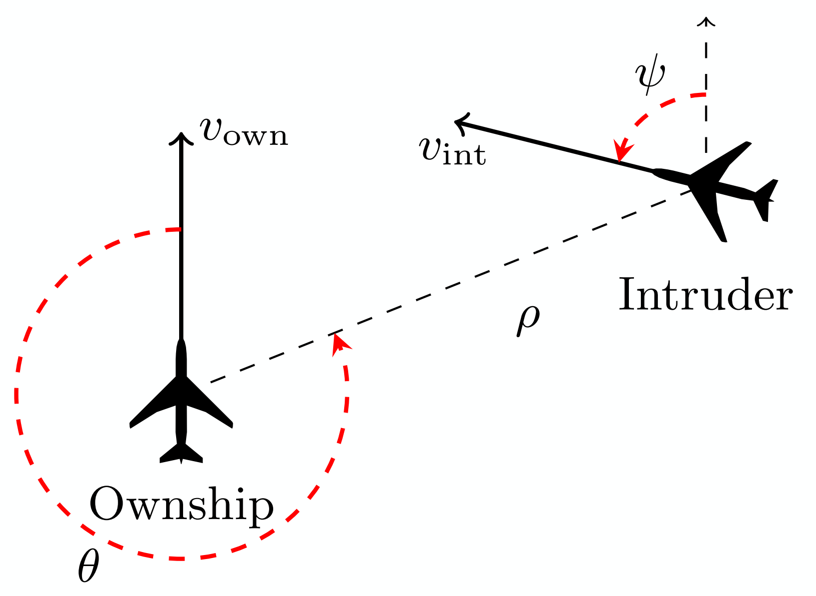
Safety



ACAS Xu

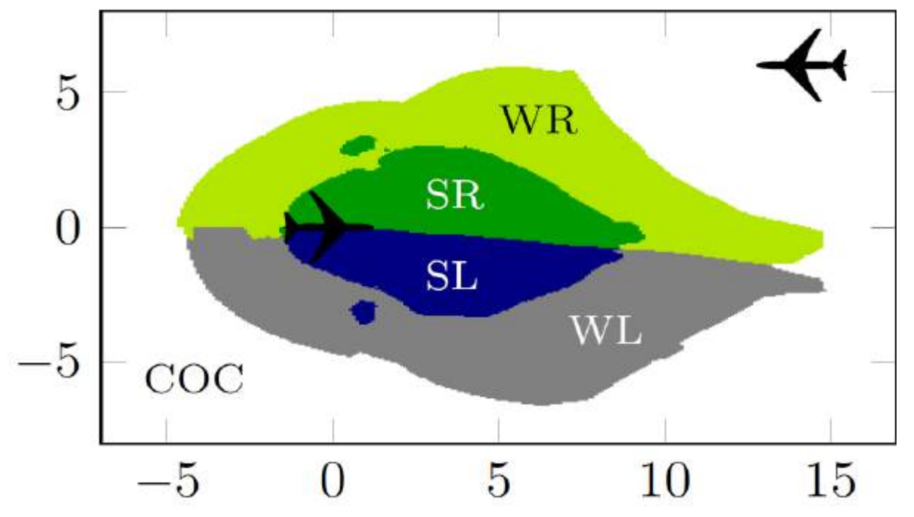
Airborne Collision Avoidance System for Unmanned Aircraft

implemented using **45 feed-forward fully-connected ReLU networks**



5 input **sensor measurements**

- ρ: distance from ownship to intruder
- θ: angle to intruder relative to ownship heading direction
- ψ: heading angle to intruder relative to ownship heading direction
- v_{own}: speed of ownship
- v_{int}: speed of intruder

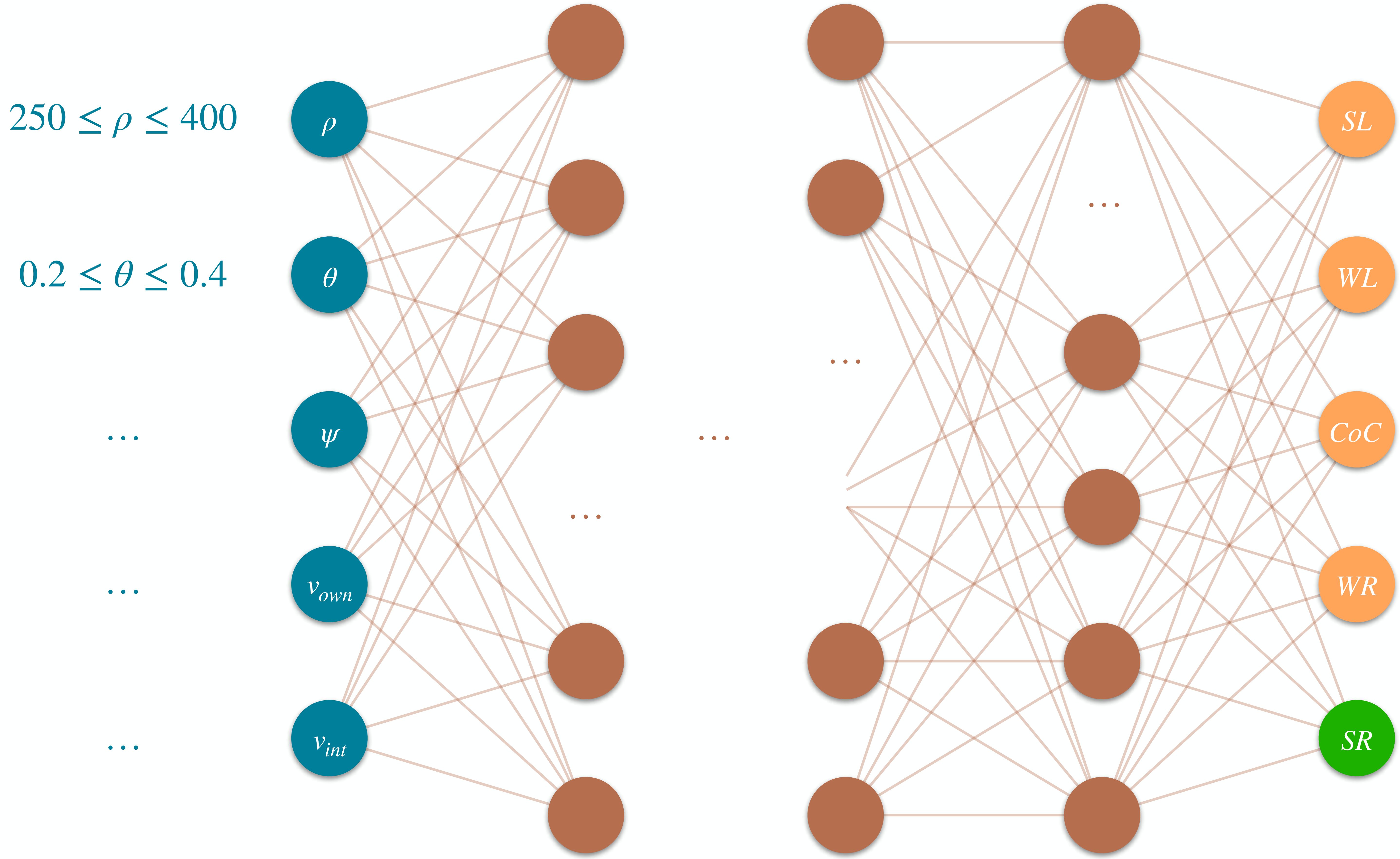


5 output **horizontal advisories**

- Strong Left
- Weak Left
- Clear of Conflict
- Weak Right
- Strong Right

ACAS Xu Safety Properties

Example: "if intruder is near and approaching from the left, go Strong Right"

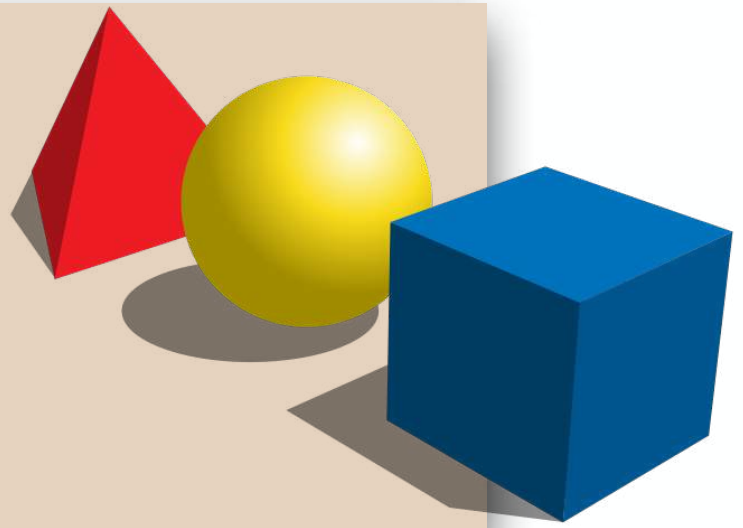


Static Safety Analysis

3-Step Recipe

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

Safety

Input-Output Properties

$\mathbf{I}$: input specification

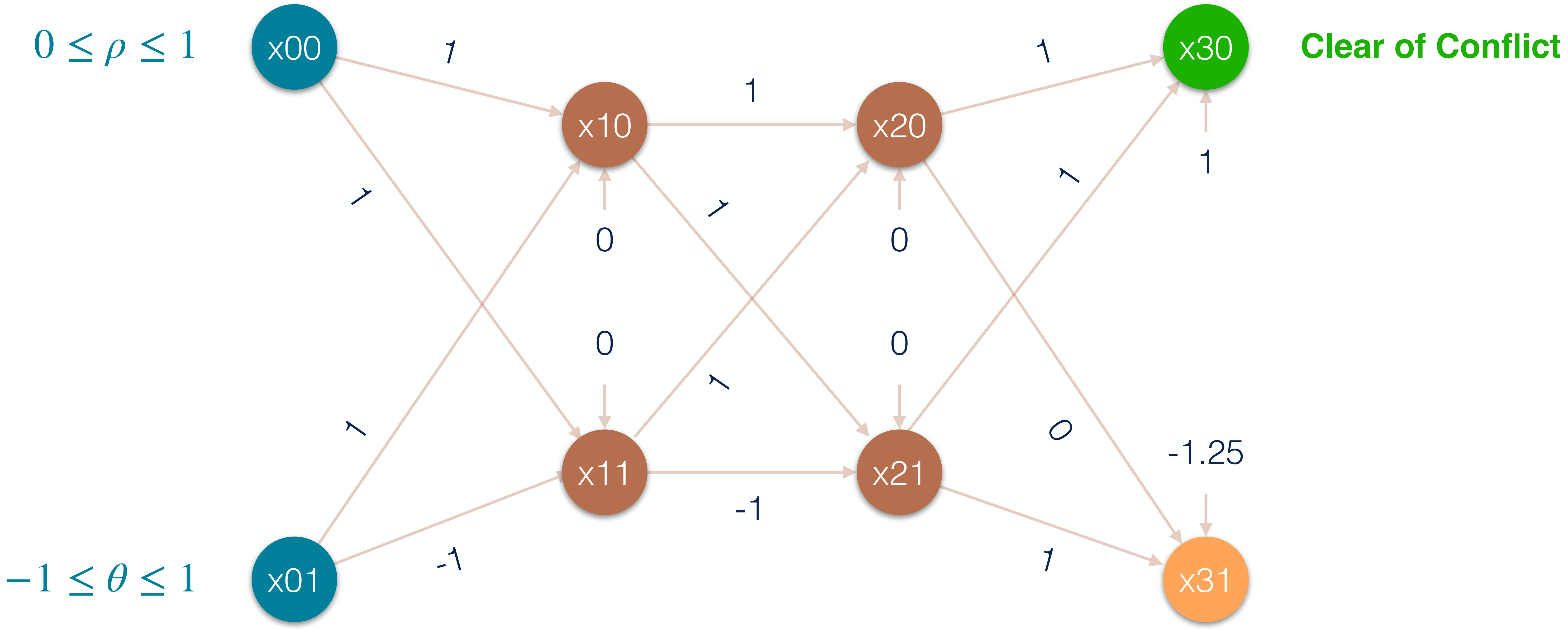
$\mathbf{O}$: output specification

$$\mathscr{R}_{\mathbf{x}}^{\delta,\epsilon} \stackrel{\text{def}}{=} \{t \in \Sigma^* \mid t_0 \vDash \mathbf{I} \Rightarrow t_\omega \vDash \mathbf{O}\}$$

$\mathscr{S}_{\mathbf{O}}^{\mathbf{I}}$ is the set of all traces that are **satisfy** the input and output specifications $\mathbf{I}$ and $\mathbf{O}$

Safety

Example



Static Safety Analysis

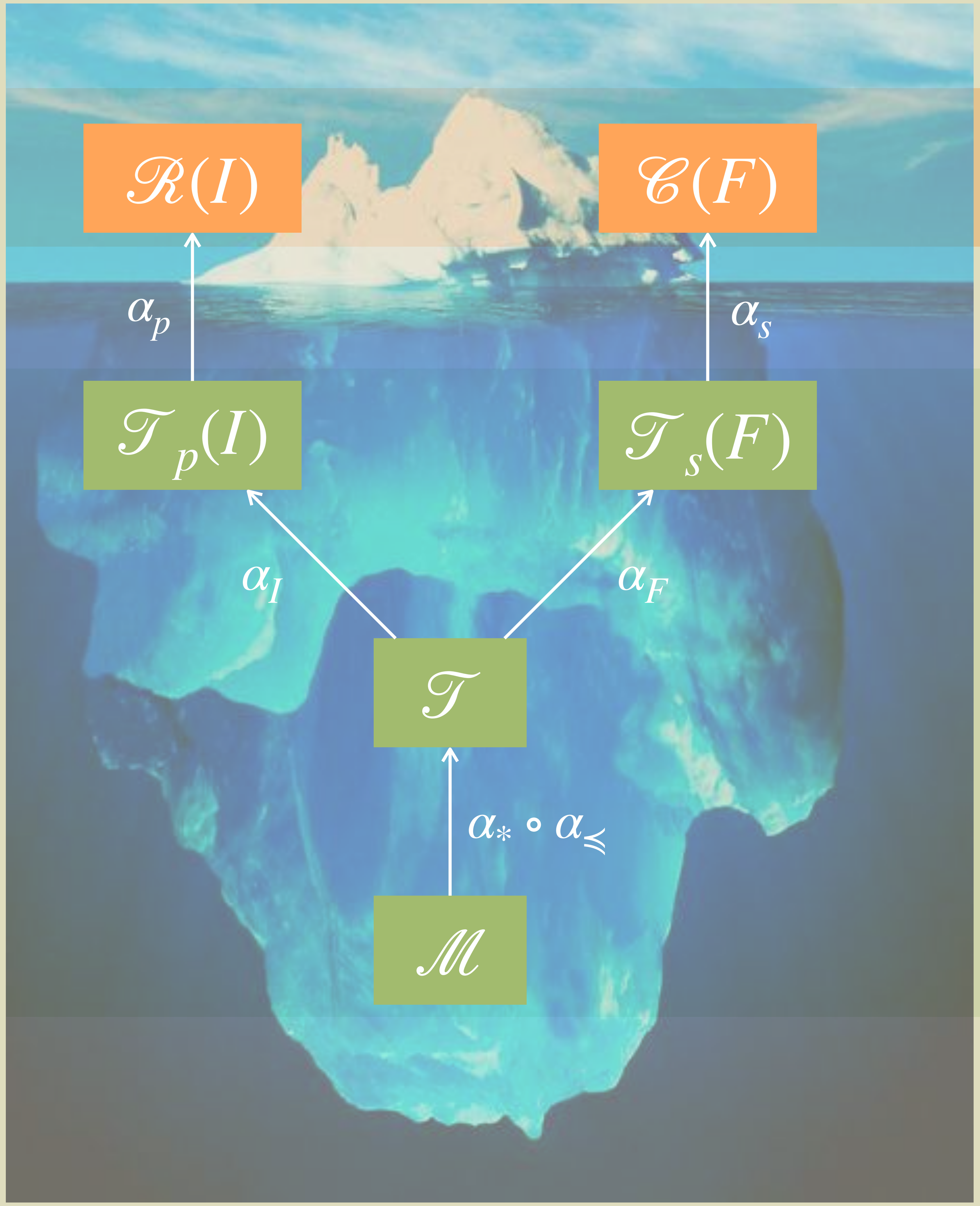
Concrete Semantics

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

concrete semantics
mathematical models of the program behavior

Hierarchy of Semantics



Forward/Backward Reachability Semantics

Forward/Backward Reachable State Abstraction

Prefix/Suffix Trace Semantics

Prefix/Suffix Trace Abstraction

Partial Finite Trace Semantics

Partial Finite Trace Abstraction

Maximal Trace Semantics

Safety

Input-Output Properties

$\mathbf{I}$: input specification

$\mathbf{O}$: output specification

$$\mathscr{R}_{\mathbf{x}}^{\delta,\epsilon} \stackrel{\text{def}}{=} \{t \in \Sigma^* \mid t_0 \models \mathbf{I} \Rightarrow t_\omega \models \mathbf{O}\}$$

$\mathscr{S}_{\mathbf{O}}^{\mathbf{I}}$ is the set of all traces that **satisfy** the input and output specifications $\mathbf{I}$ and $\mathbf{O}$

Theorem

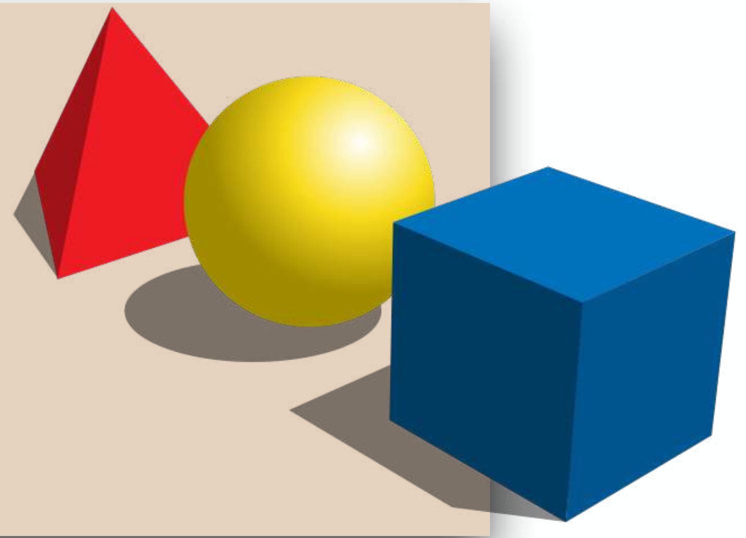
$$M \models \mathscr{S}_{\mathbf{O}}^{\mathbf{I}} \Leftrightarrow \mathscr{M}[\![M]\!] \subseteq \mathscr{S}_{\mathbf{O}}^{\mathbf{I}} \Leftrightarrow \mathscr{T}_p(\mathbf{I})[\![M]\!] \subseteq \mathscr{S}_{\mathbf{O}}^{\mathbf{I}}$$

Static Safety Analysis

Abstract Semantics

practical tools
targeting specific programs

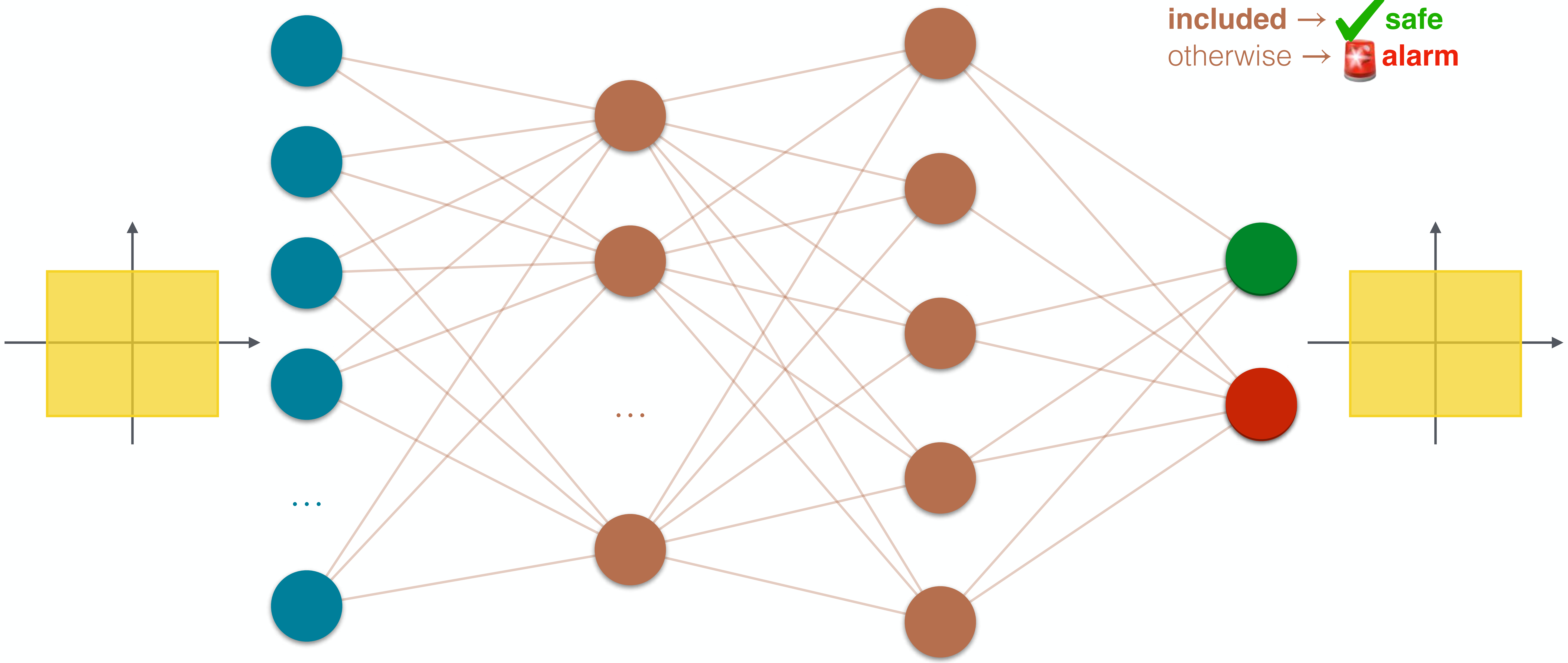
abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

Static Safety Analysis

Abstract Prefix Trace Semantics



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

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

Theorem

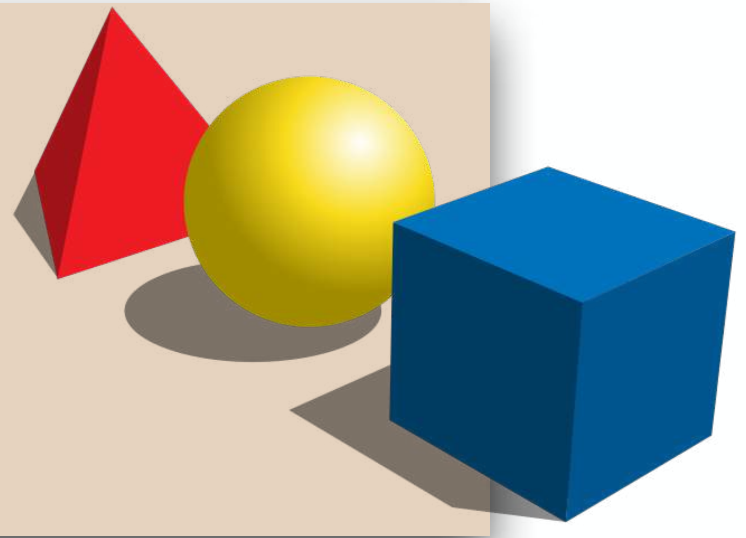
$$\mathscr{T}_p(\mathbf{I}))[\![M]\!] \subseteq \mathscr{T}_p^\#(\mathbf{I}^\#))[\![M]\!] \subseteq \mathscr{S}_{\mathbf{O}}^{\mathbf{I}} \Rightarrow M \models \mathscr{S}_{\mathbf{O}}^{\mathbf{I}}$$

Static Safety Analysis

Abstract Domain #3: DeepPoly Domain

practical tools
targeting specific programs

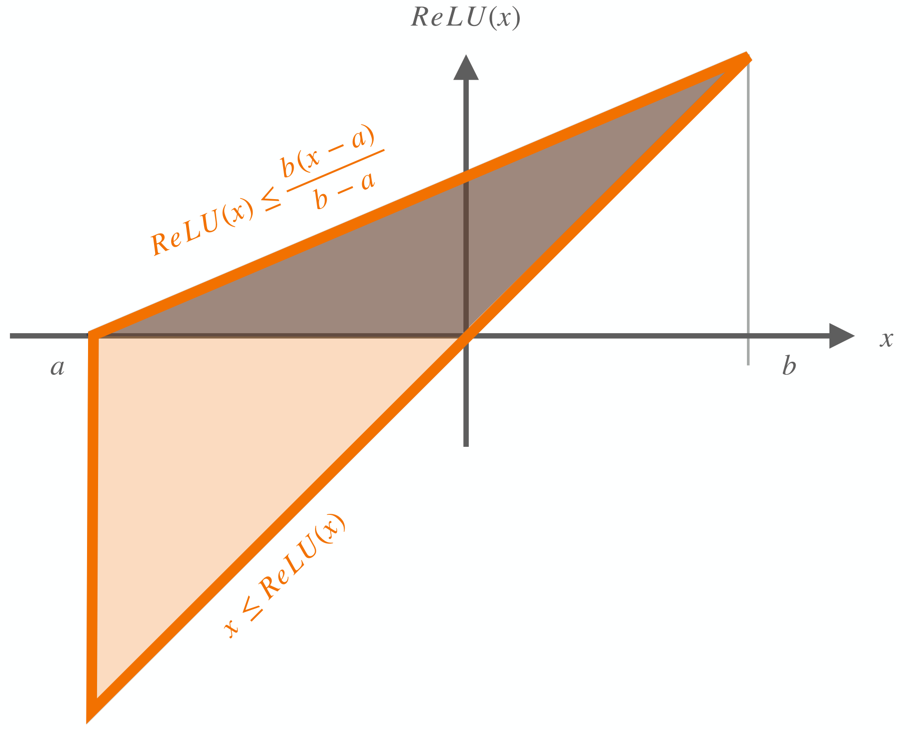
abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

DeepPoly Domain

Example



ReLU

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

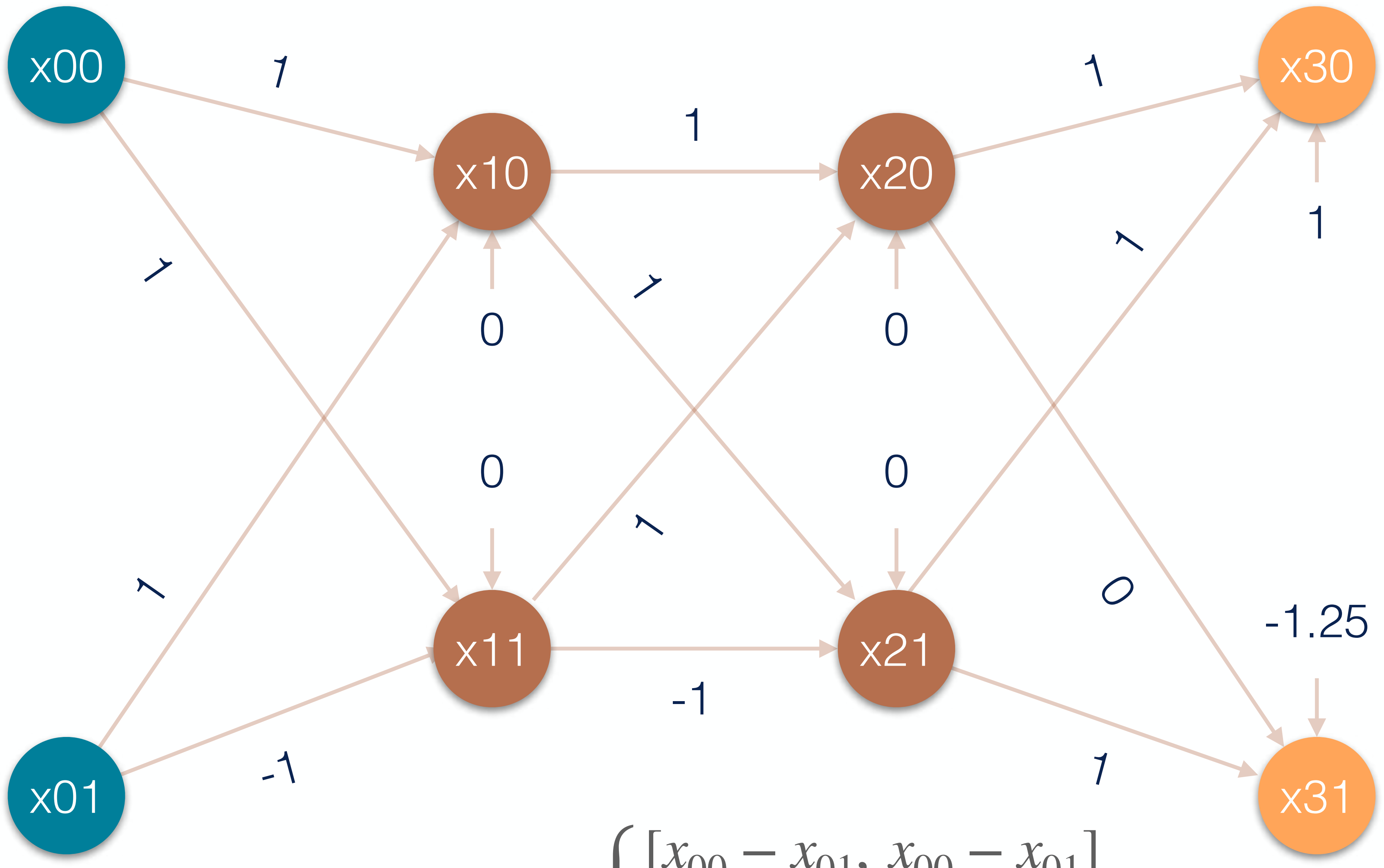
$x_{10} \mapsto \begin{cases} [x_{00} + x_{01},\ x_{00} + x_{01}] \\ [\mathbf{-1},\ \mathbf{2}] \end{cases}$

$0 \leq \rho \leq 1$

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

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

$-1 \leq \theta \leq 1$



Clear of Conflict

Strong Turn

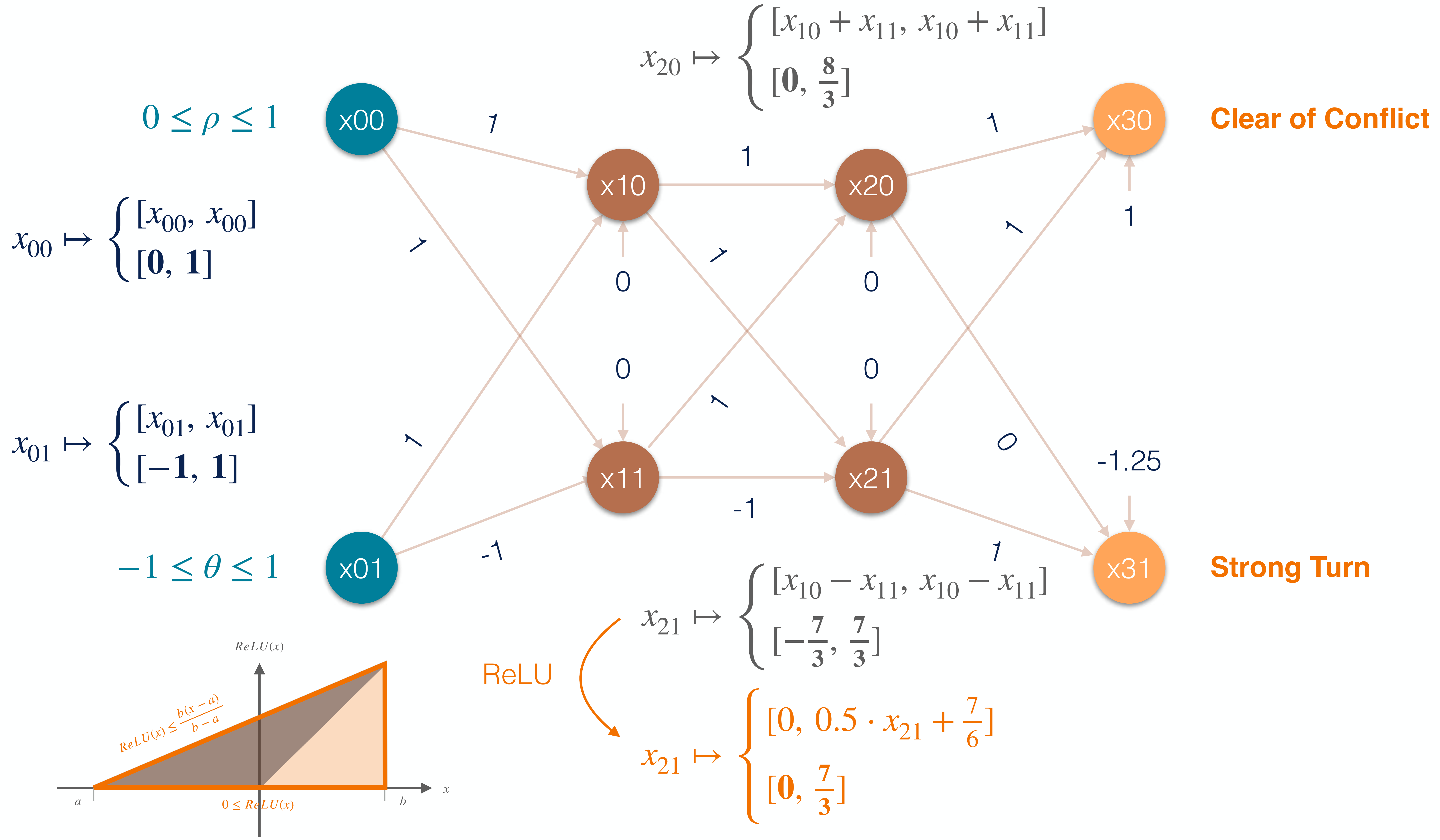
$x_{11} \mapsto \begin{cases} [x_{00} - x_{01},\ x_{00} - x_{01}] \\ [\mathbf{-1},\ \mathbf{2}] \end{cases}$

ReLU

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

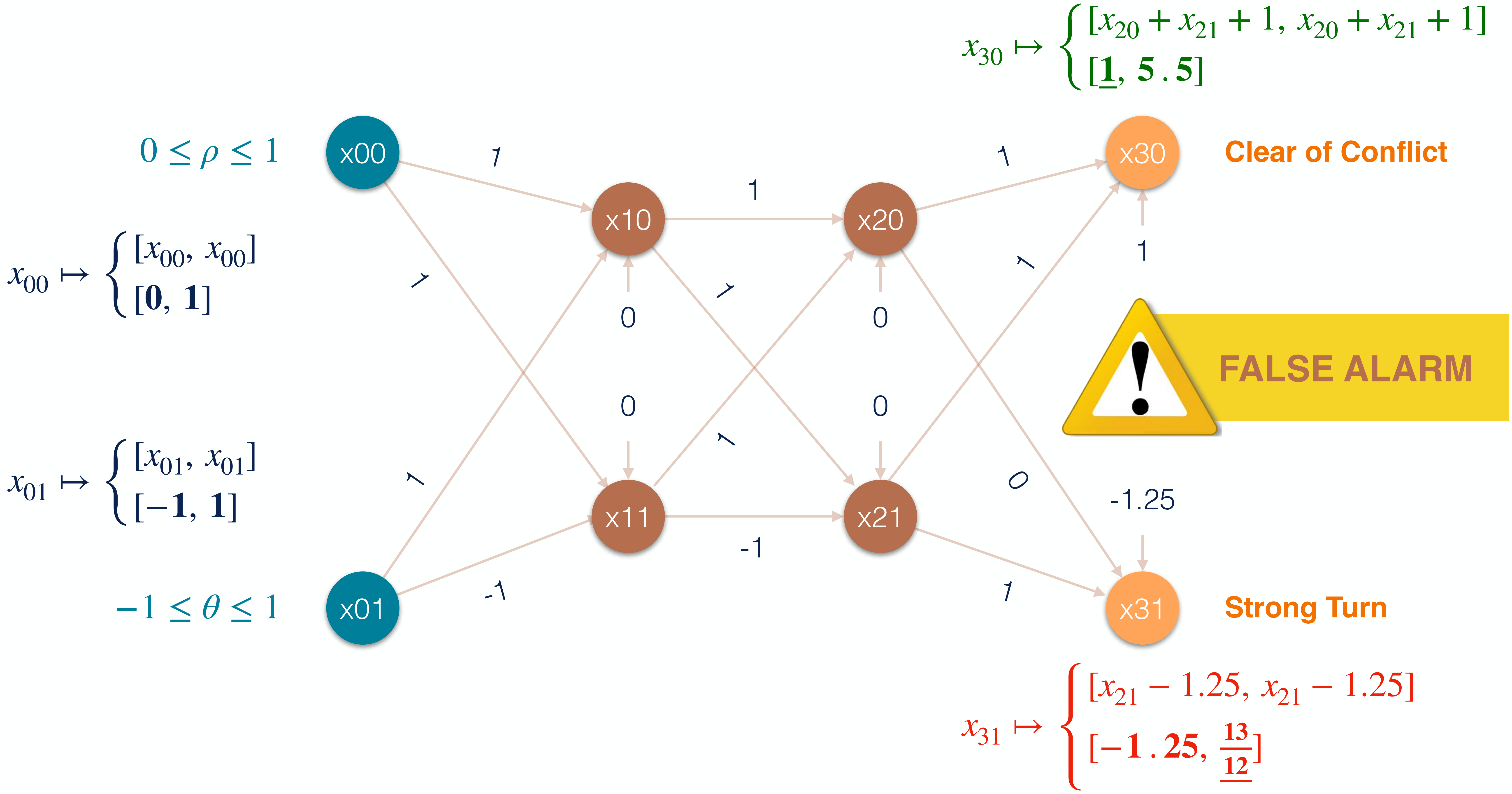
DeepPoly Domain

Example



DeepPoly Domain

Example

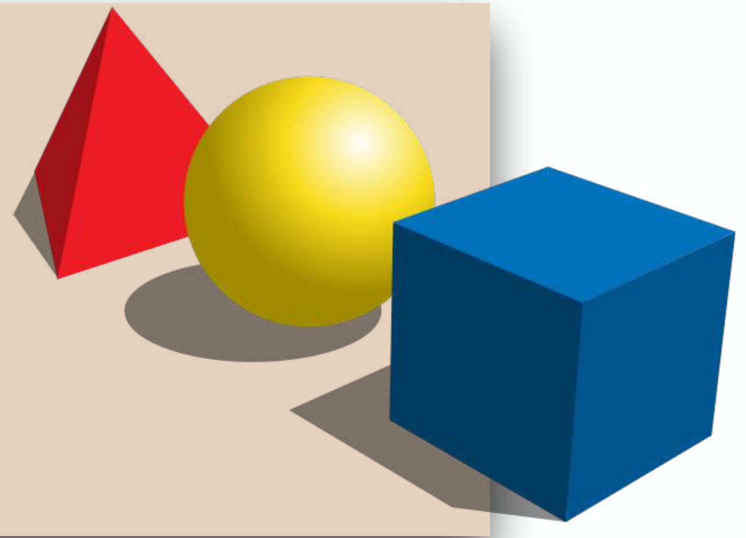


Static Safety Analysis

Abstract Domain #2: Symbolic Domain

practical tools
targeting specific programs

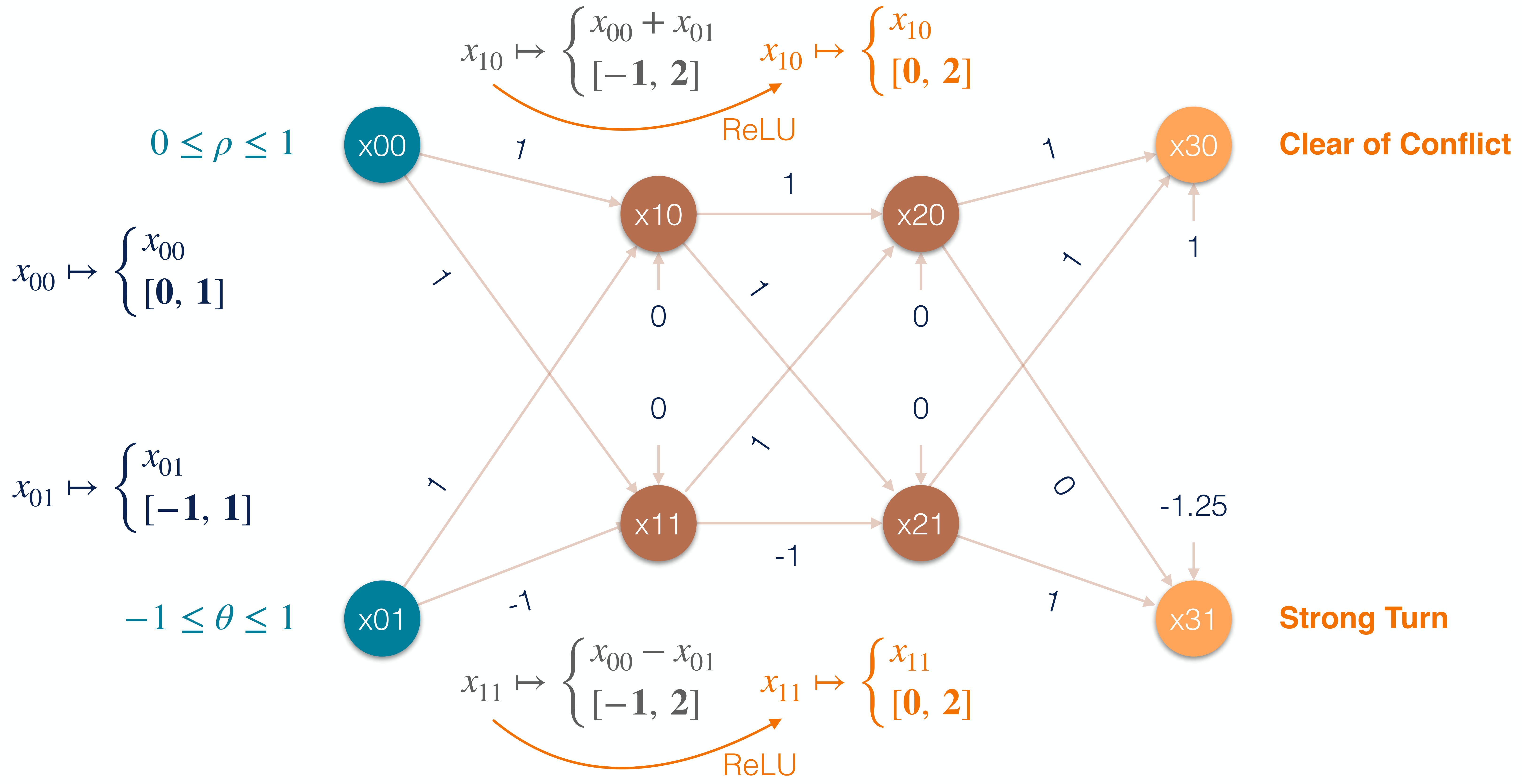
abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

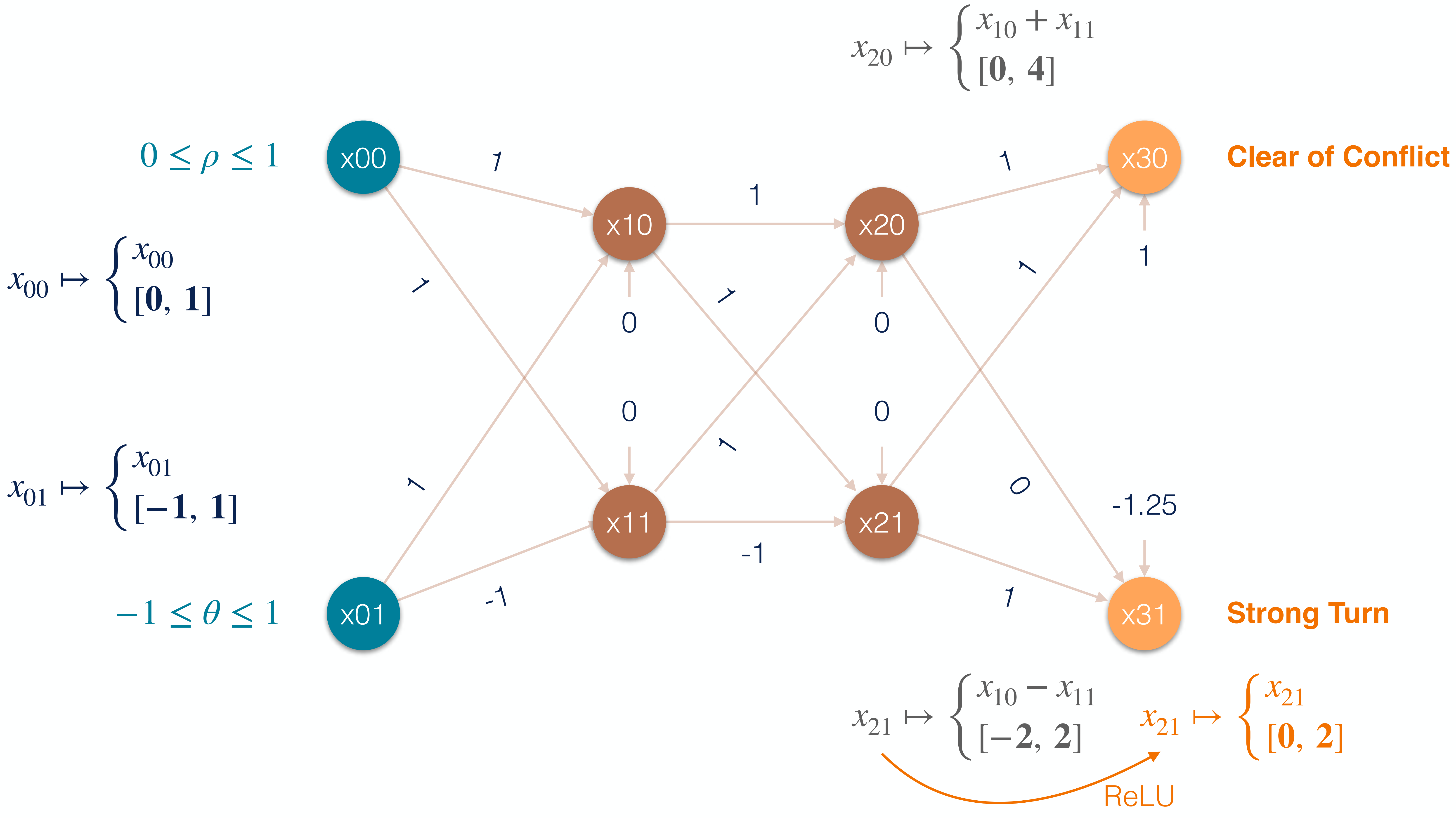
Symbolic Domain

Example



Symbolic Domain

Example



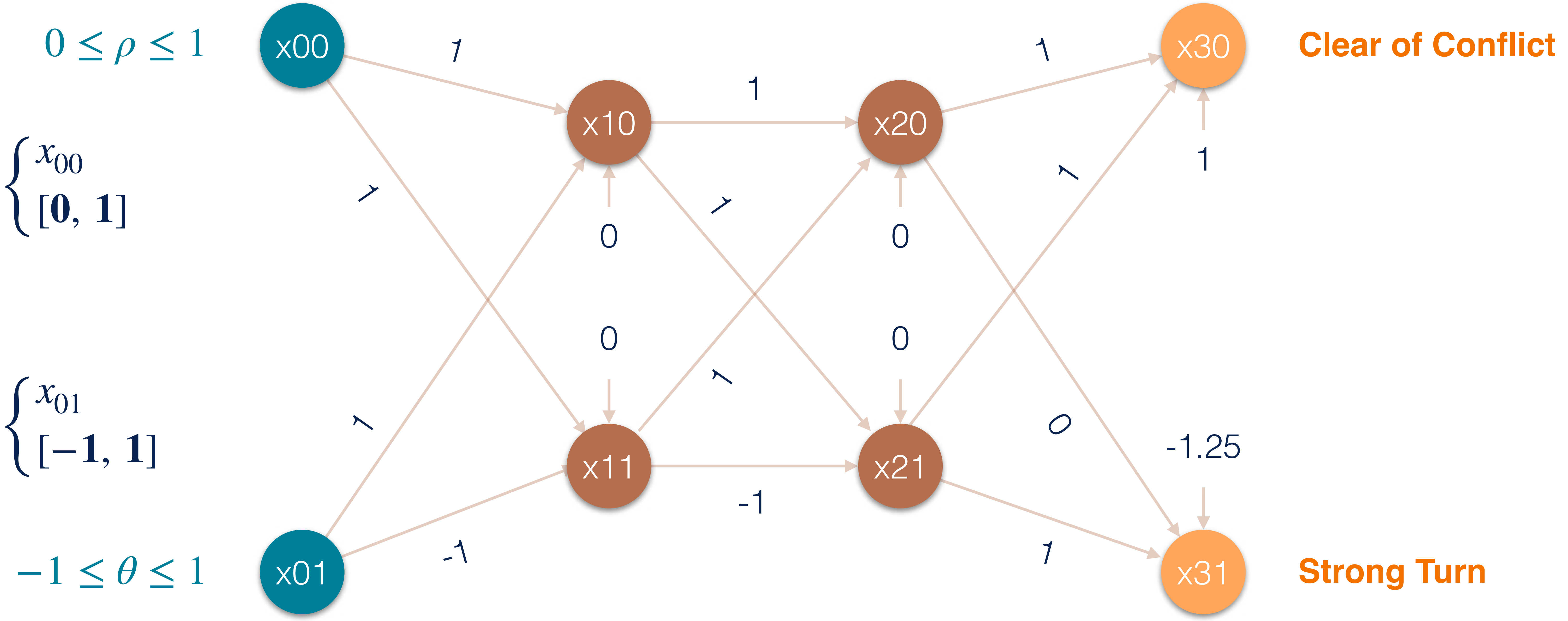
Symbolic Domain

Example

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

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

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



👌

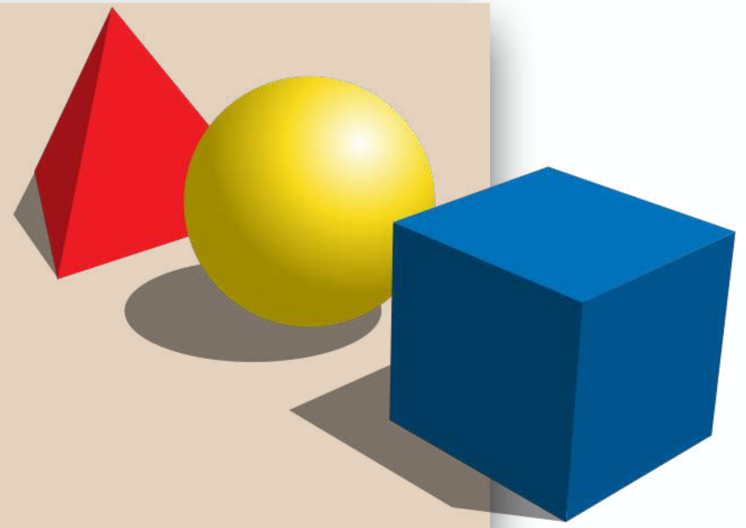
$$x_{31} \mapsto \begin{cases} x_{21} - 1.25 \\ [\mathbf{-1.25},\ \underline{\mathbf{0.75}}] \end{cases} \leftarrow [\mathbf{-1.25}, \underline{\frac{\mathbf{13}}{\mathbf{12}}}] \text{ with DeepPoly}$$

Static Safety Analysis

Abstract Domain #4: Reduced Product Domain

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties



concrete semantics
mathematical models of the program behavior

Reduced Product Domain

Symbolic Domain & DeepPoly Domain

$[a_d, b_d]$

Symbolic

$[\max(a_s, a_d), \min(b_s, b_u)]$

DeepPoly

$[\max(a_s, a_d), \min(b_s, b_u)]$

$[a_s, b_s]$

Reduced Product Domain

Example

$$x_{10} \mapsto \begin{cases} x_{10} & \to [\mathbf{0},\ \mathbf{2}] \\ [x_{10},\ \frac{2}{3} \cdot x_{10} + \frac{2}{3}] & \to [\mathbf{-1},\ \mathbf{2}] \\ [\mathbf{0},\ \mathbf{2}] \end{cases}$$

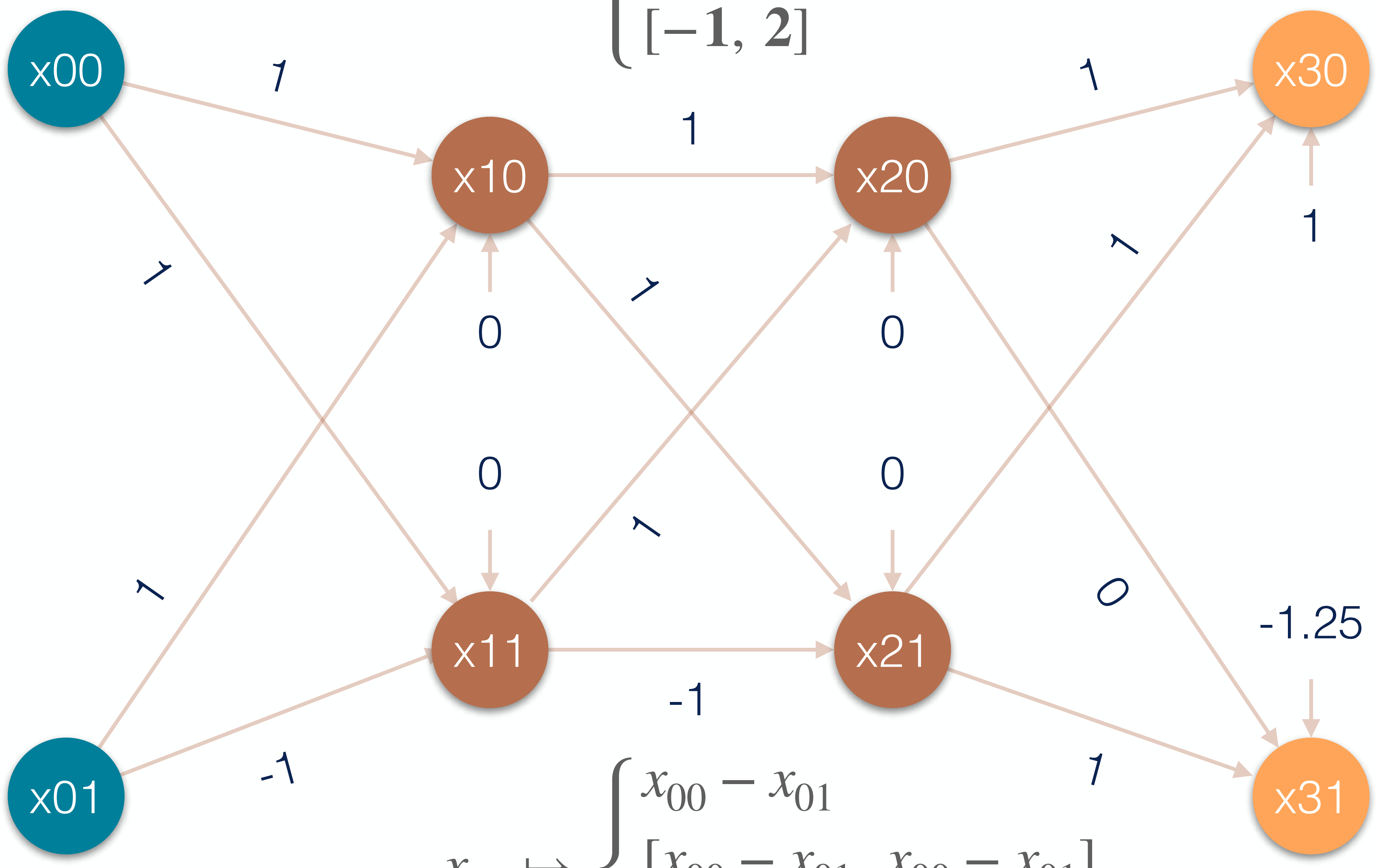
$$x_{10} \mapsto \begin{cases} x_{00} + x_{01} \\ [x_{00} + x_{01},\ x_{00} + x_{01}] \\ [\mathbf{-1},\ \mathbf{2}] \end{cases}$$

$0 \leq \rho \leq 1$

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

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

$-1 \leq \theta \leq 1$



Clear of Conflict

Strong Turn

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

ReLU

$$x_{11} \mapsto \begin{cases} x_{11} & \to [\mathbf{0},\ \mathbf{2}] \\ [x_{11},\ \frac{2}{3} \cdot x_{11} + \frac{2}{3}] & \to [\mathbf{-1},\ \mathbf{2}] \\ [\mathbf{0},\ \mathbf{2}] \end{cases}$$

Reduced Product Domain

Example

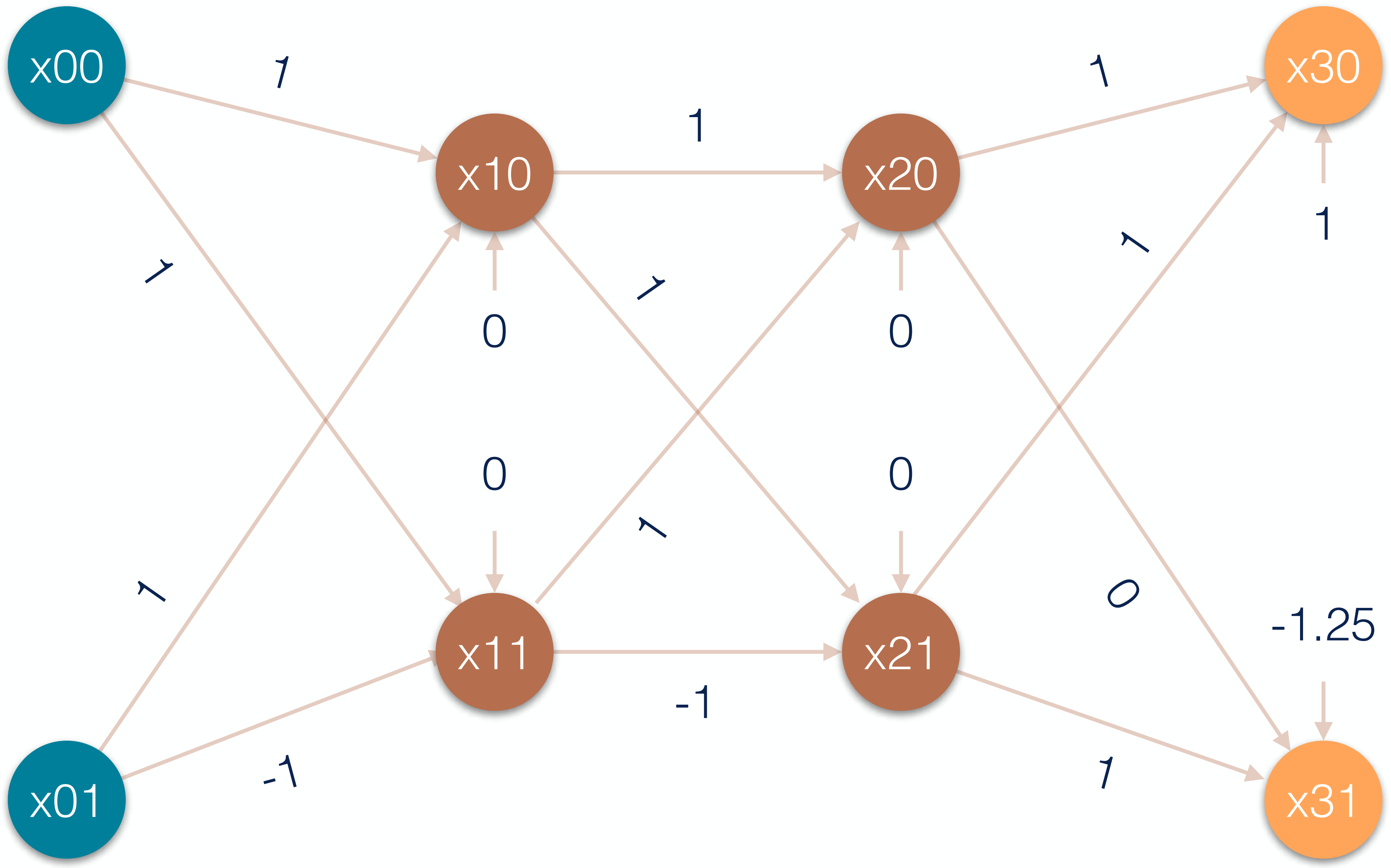
$$x_{20} \mapsto \begin{cases} x_{10} + x_{11} & \rightarrow [\mathbf{0},\ \mathbf{4}] \\ [x_{10} + x_{11},\ x_{10} + x_{11}] & \rightarrow [\mathbf{0},\ \mathbf{\tfrac{8}{3}}] \\ [\mathbf{0},\ \mathbf{\tfrac{8}{3}}] & \end{cases}$$

$0 \leq \rho \leq 1$

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

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

$-1 \leq \theta \leq 1$



Clear of Conflict

Strong Turn

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

ReLU

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

Reduced Product Domain

Example

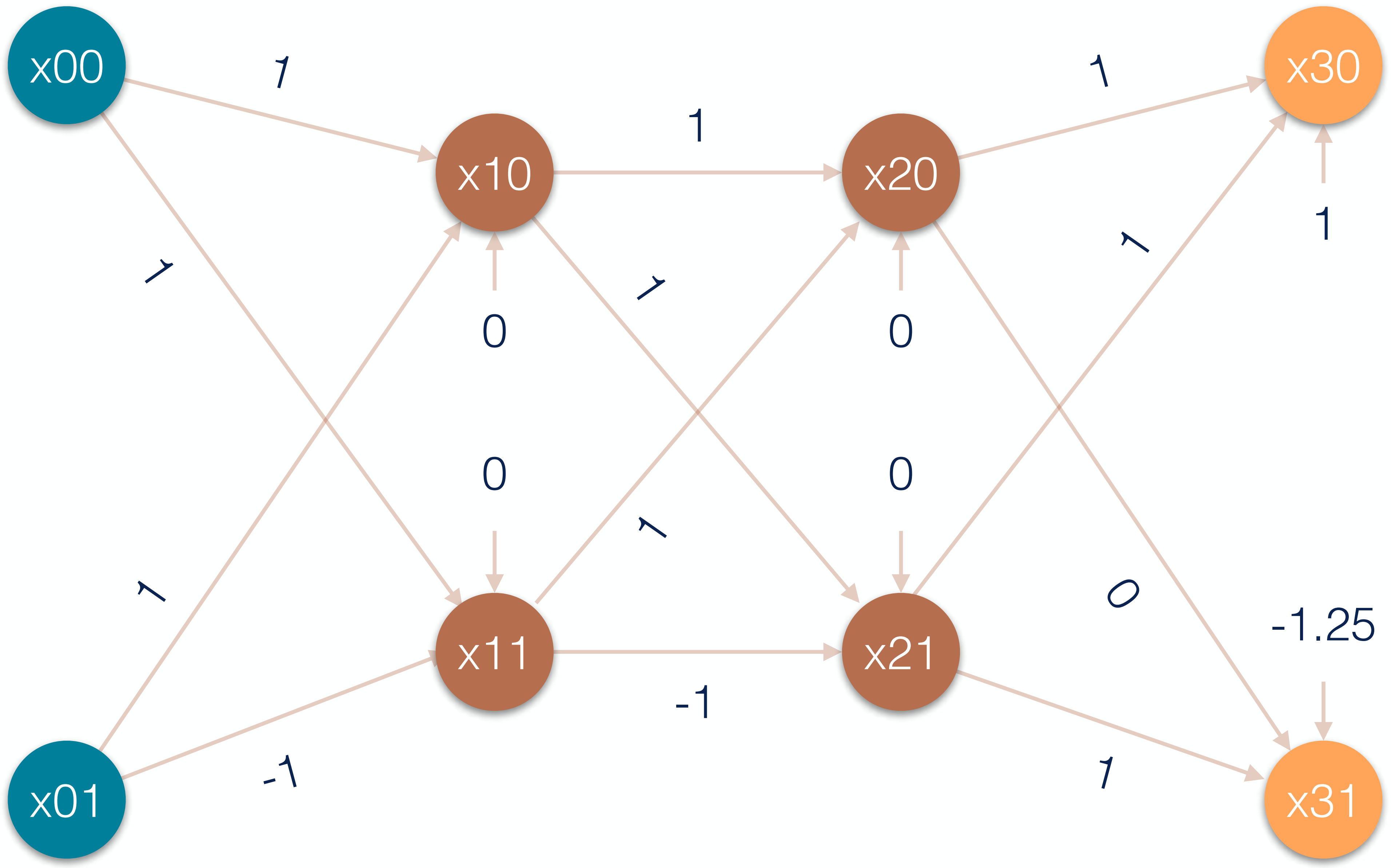
$$x_{30} \mapsto \begin{cases} x_{10} + x_{11} + x_{21} + 1 & \to [\mathbf{1}, \frac{\mathbf{20}}{\mathbf{3}}] \\ [x_{20} + x_{21} + 1,\ x_{20} + x_{21} + 1] & \to [\mathbf{1}, \frac{\mathbf{16}}{\mathbf{3}}] \\ [\underline{\mathbf{1}}, \frac{\mathbf{16}}{\mathbf{3}}] \end{cases}$$

$0 \leq \rho \leq 1$

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

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

$-1 \leq \theta \leq 1$



Clear of Conflict

Strong Turn

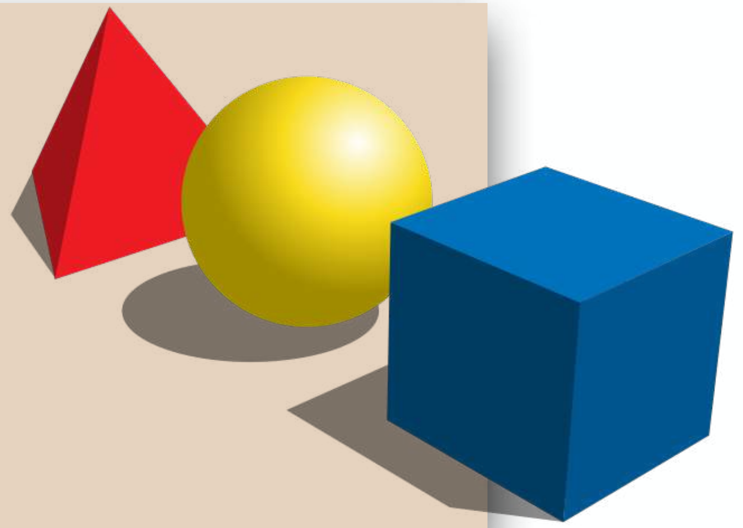
$$x_{31} \mapsto \begin{cases} x_{21} - 1.25 & \to [-\mathbf{1.25}, \mathbf{0.75}] \\ [x_{21} - 1.25,\ x_{21} - 1.25] & \to [-\mathbf{1.25}, \mathbf{0.75}] \\ [-\mathbf{1.25}, \underline{\mathbf{0.75}}] \end{cases}$$

Static Safety Analysis

Going Farther: Complete Methods

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

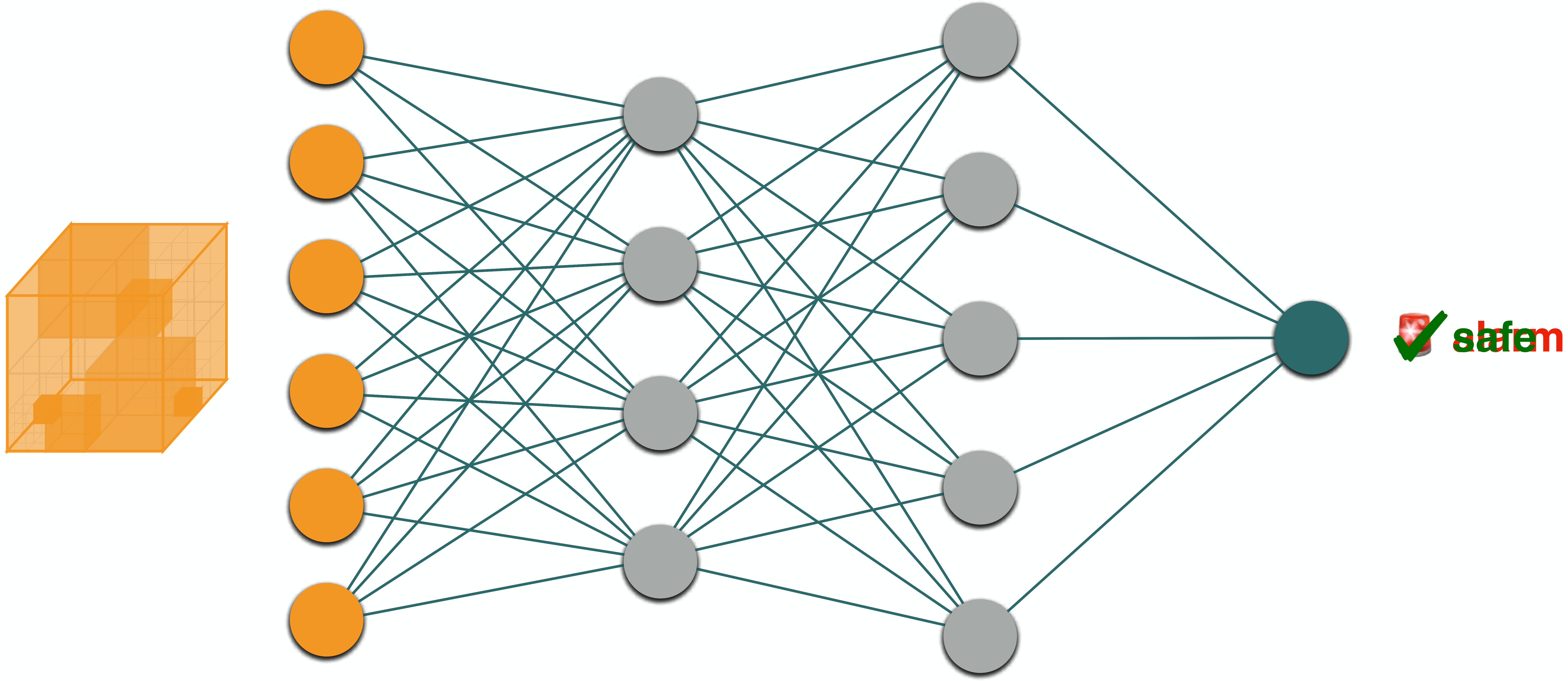


concrete semantics
mathematical models of the program behavior

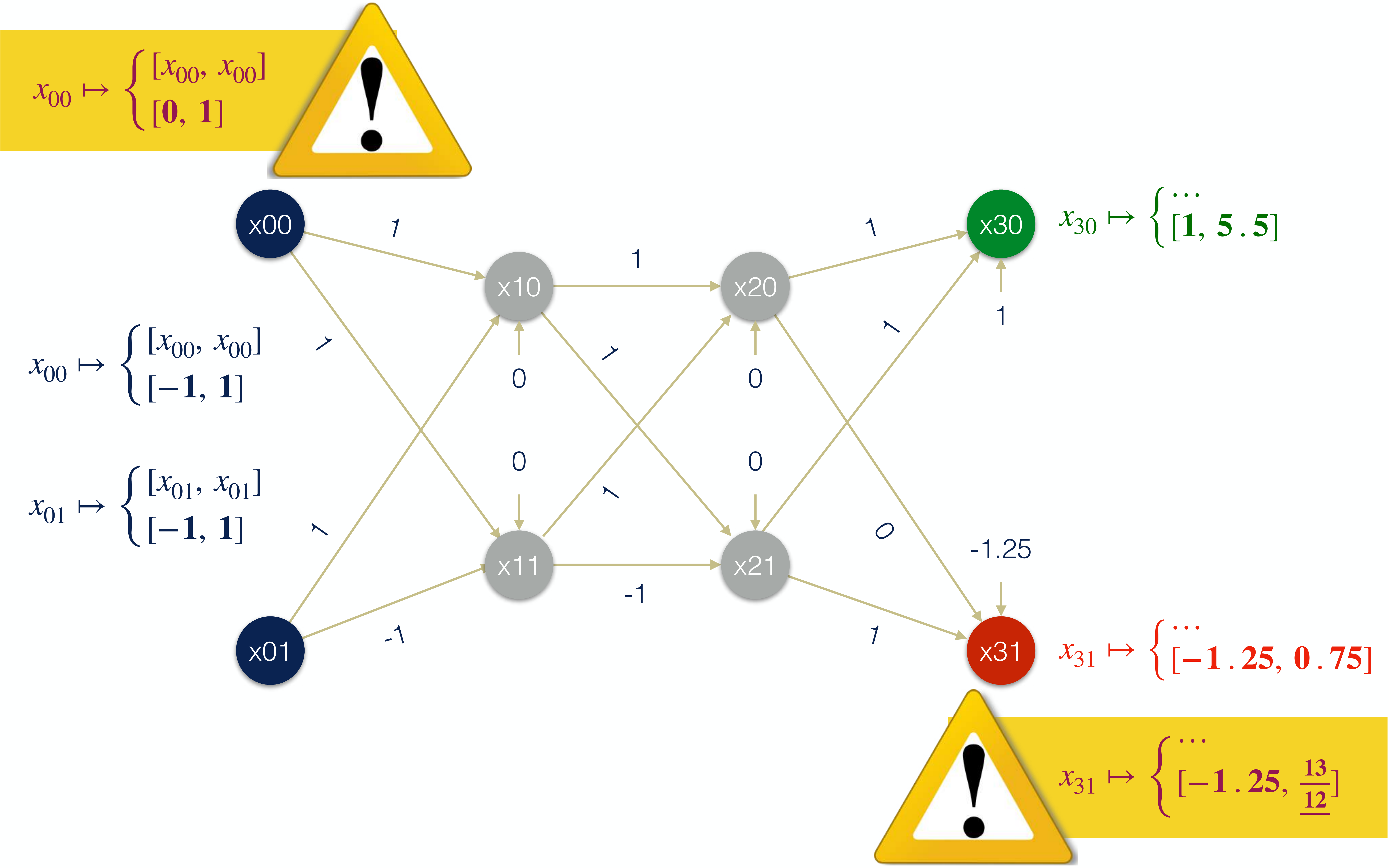
ReluVal[Wang et al. @ USENIX Security 2018]

Asymptotically Complete Method

symbolic propagation
+ iterative input refinement



DeepPoly Domain+ Input Refinement



Neurify

Asymptotically Complete Method

symbolic propagation + convex ReLU approximation + iterative input/ReLU refinement

$$x_{i,j} \mapsto \begin{cases} [\sum_k c_{0,k} \cdot x_{0,k} + c,\ \sum_k d_{0,k} \cdot x_{0,k} + d] & c_{0,k}, c, d_{0,k}, d \in \mathbb{R} \\ [a,\ b] & a, b \in \mathbb{R} \end{cases}$$

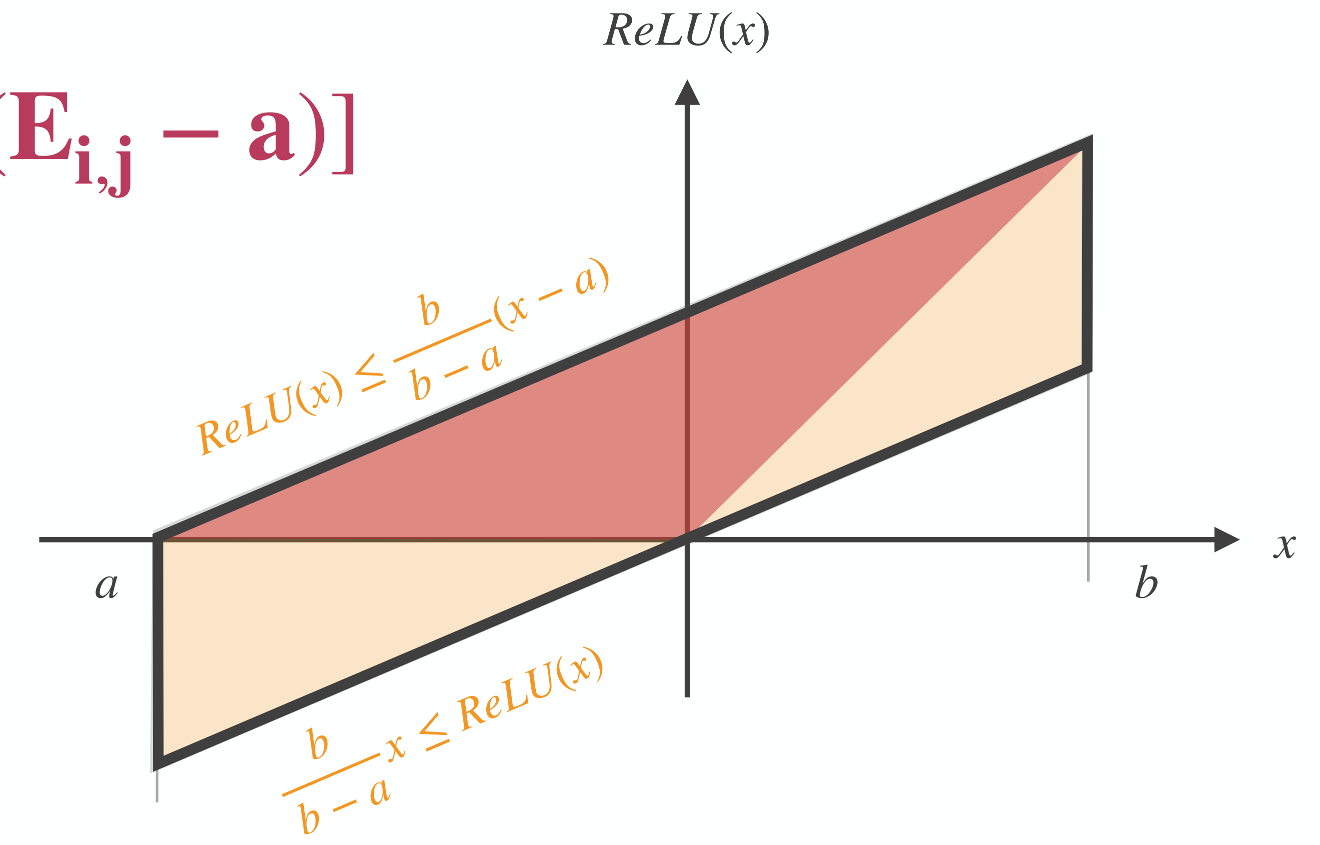
$$x_{i,j} \mapsto \begin{cases} [\mathbf{E_{i,j}}, \mathbf{E_{i,j}}] \\ [\mathbf{a}, \mathbf{b}] \end{cases}$$

ReLU ($0 \leq a$):

$$x_{i,j} \mapsto \begin{cases} [\mathbf{E_{i,j}}, \mathbf{E_{i,j}}] \\ [\mathbf{a}, \mathbf{b}] \end{cases} \qquad 0 \leq a$$

ReLU ($a < 0 \wedge 0 < b$):

$$x_{i,j} \mapsto \begin{cases} [\frac{\mathbf{b}}{\mathbf{b}-\mathbf{a}}\mathbf{E_{i,j}}, \frac{\mathbf{b}}{\mathbf{b}-\mathbf{a}}(\mathbf{E_{i,j}} - \mathbf{a})] \\ [\mathbf{a}, \mathbf{b}] \end{cases} \qquad a < 0 \wedge 0 < b$$

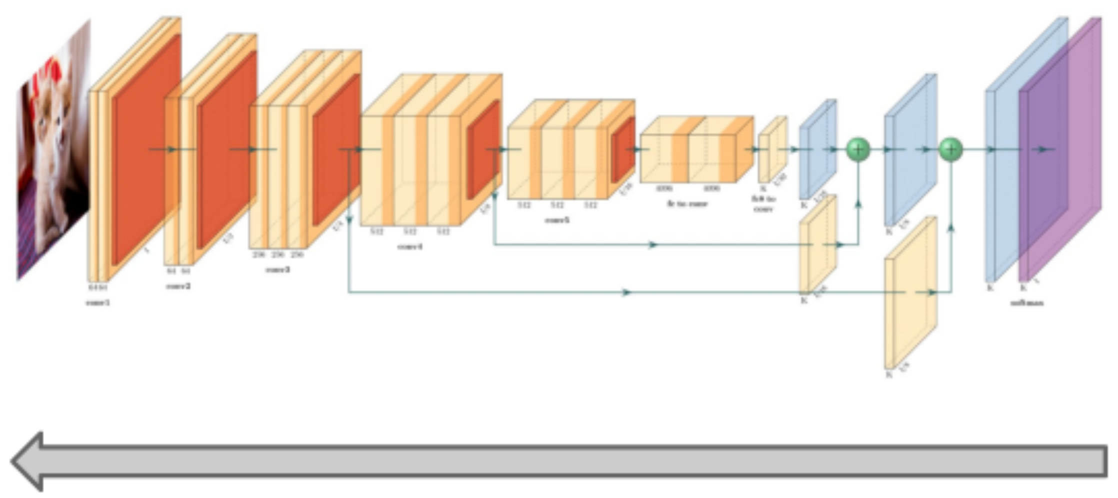


ReLU ($b \leq 0$):

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

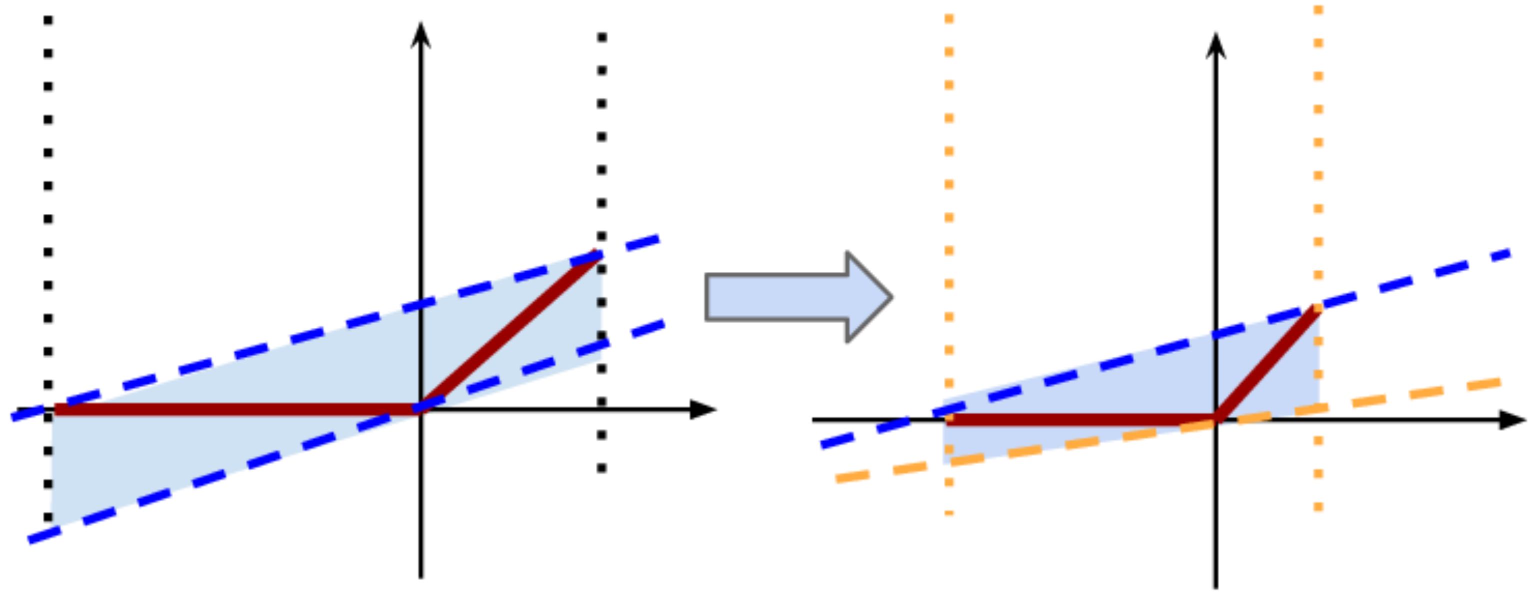
$\alpha\beta$-CROWN

The State of the Art

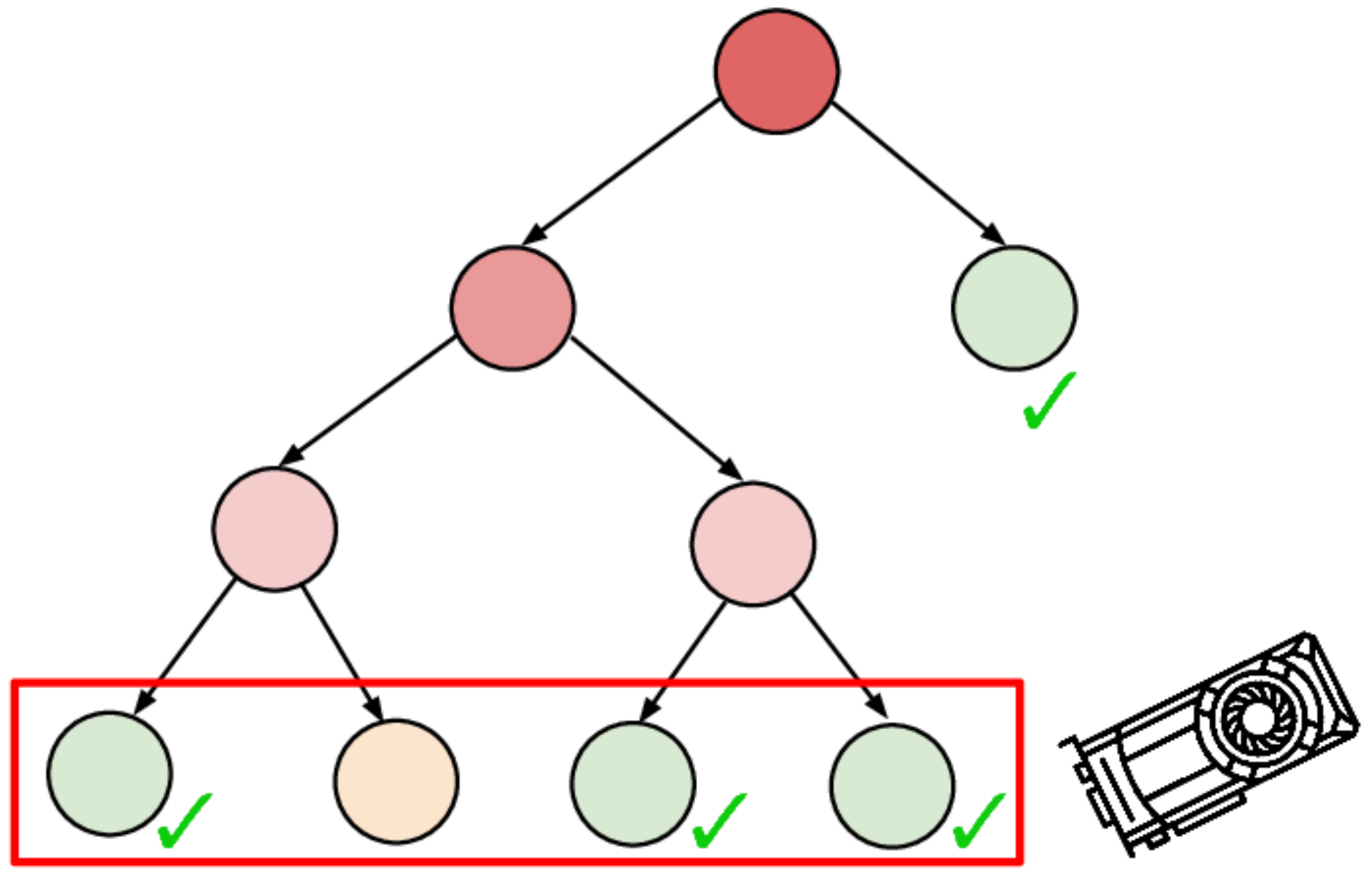


$$\min_{x \in \mathcal{C}} f(x) \geq \min_{x \in \mathcal{C}} \boldsymbol{a}^\top x + c$$

Efficient bound propagation (**CROWN**)



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



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

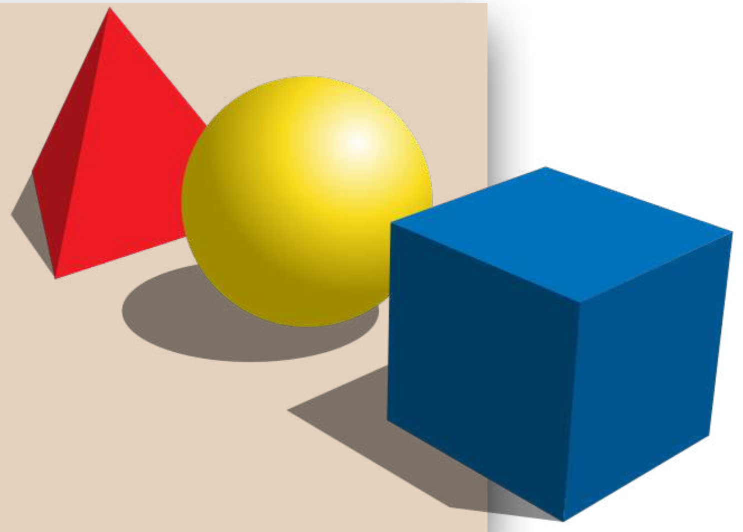
Winner of the International Verification of Neural Networks Competition since 2021

Static Safety Analysis

Going Farther: Other Abstractions

practical tools
targeting specific programs

abstract semantics, abstract domains
algorithmic approaches to decide program properties

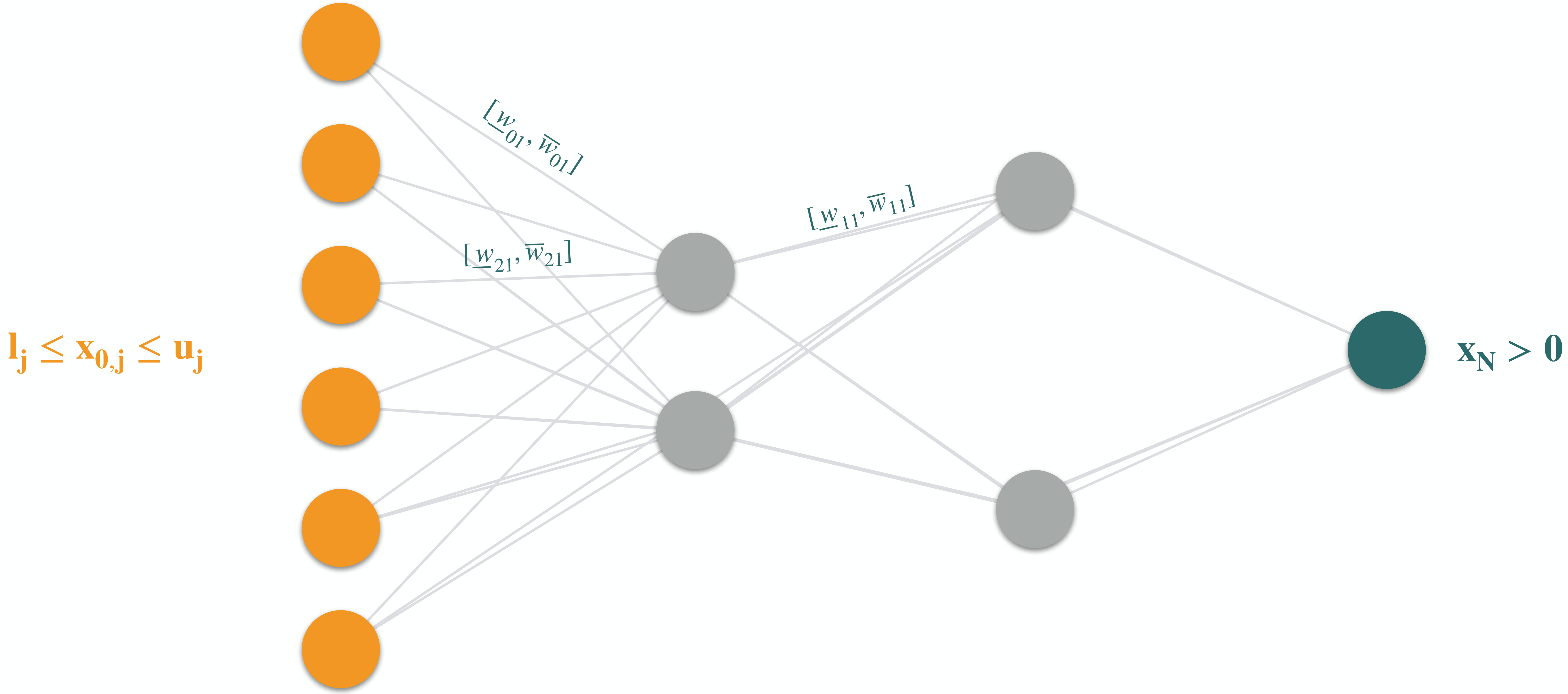


concrete semantics
mathematical models of the program behavior

Interval Neural Networks [Prabhakar and Afza @ NeurIPS 2019]

Related Work

Elboher et al. @ CAV 2020



merge neurons layer-wise based on partitioning strategy + replace weights with intervals