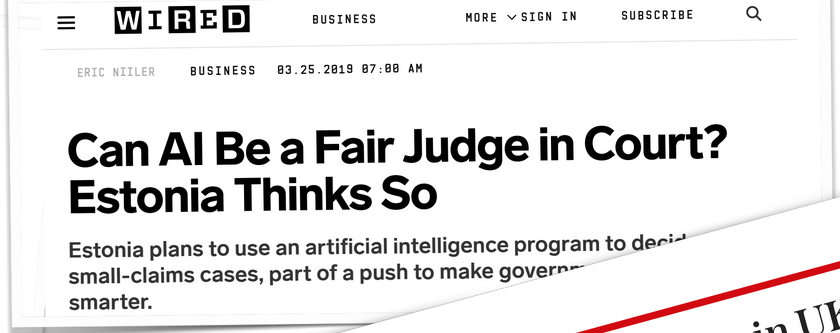
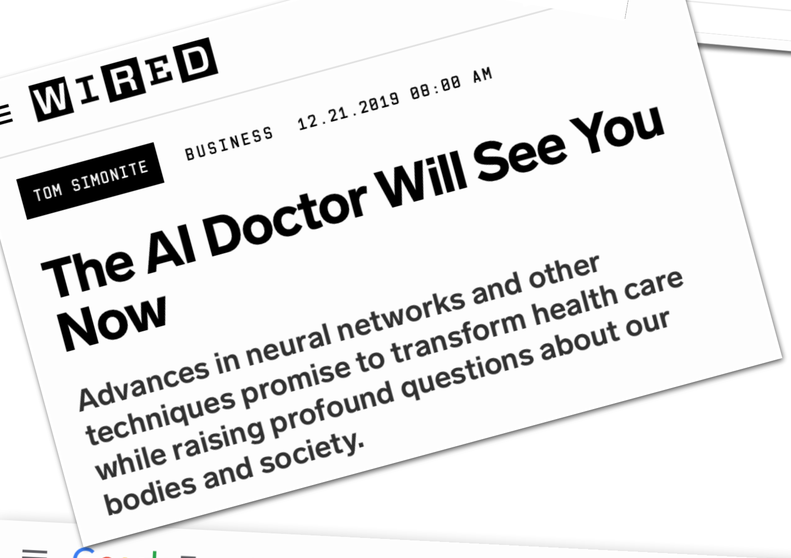
Reduced Products of Abstract Domains for Fairness Certification of Neural Networks

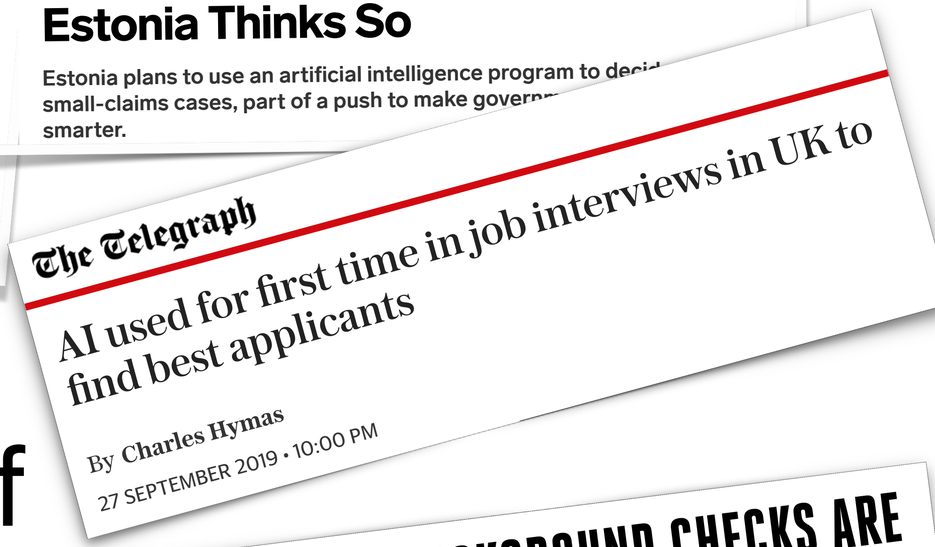
Denis Mazzucato and Caterina Urban

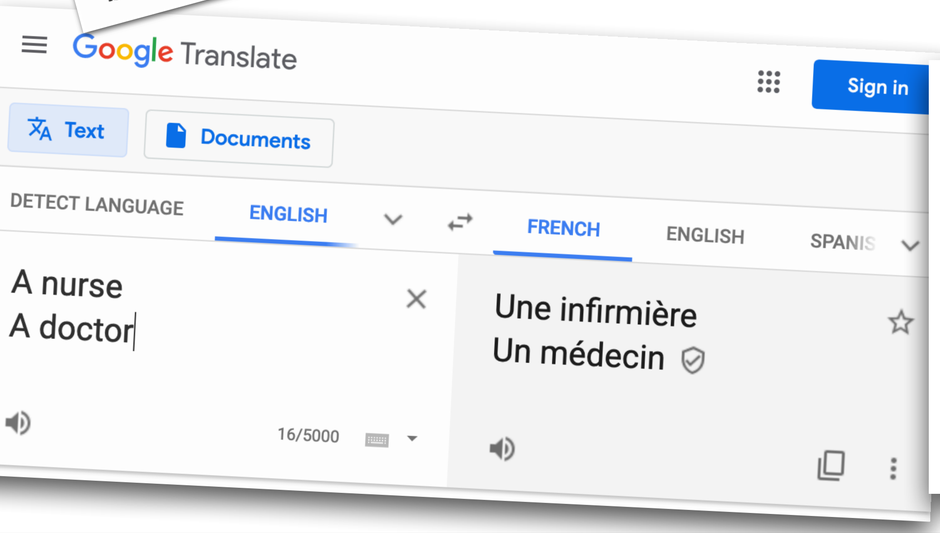
SAS 2021 - June 16th, 2022

Social Impact of Machine Learning

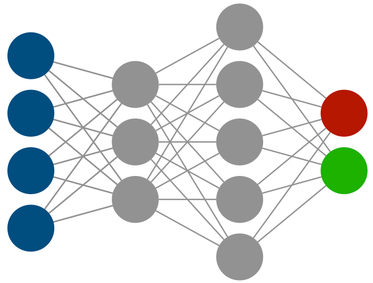




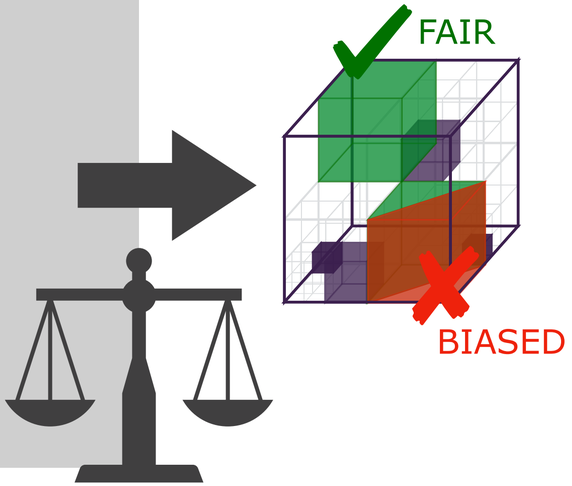


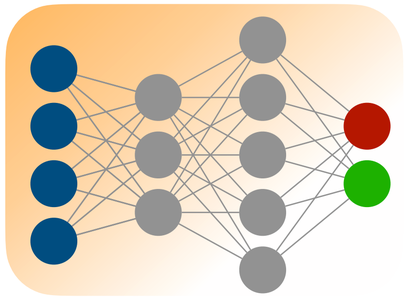



Artificial
Intelligence Act
April 2021



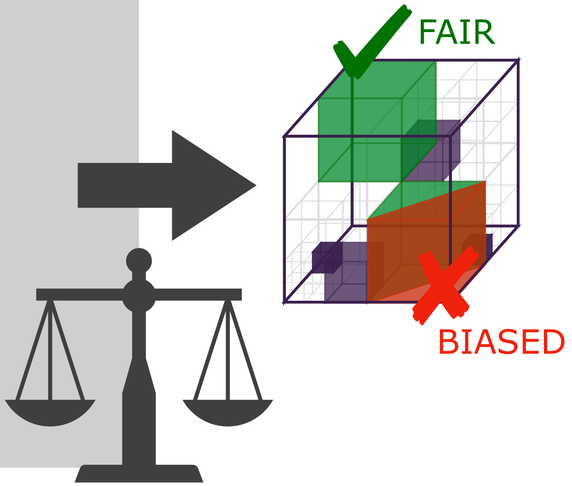
Libra



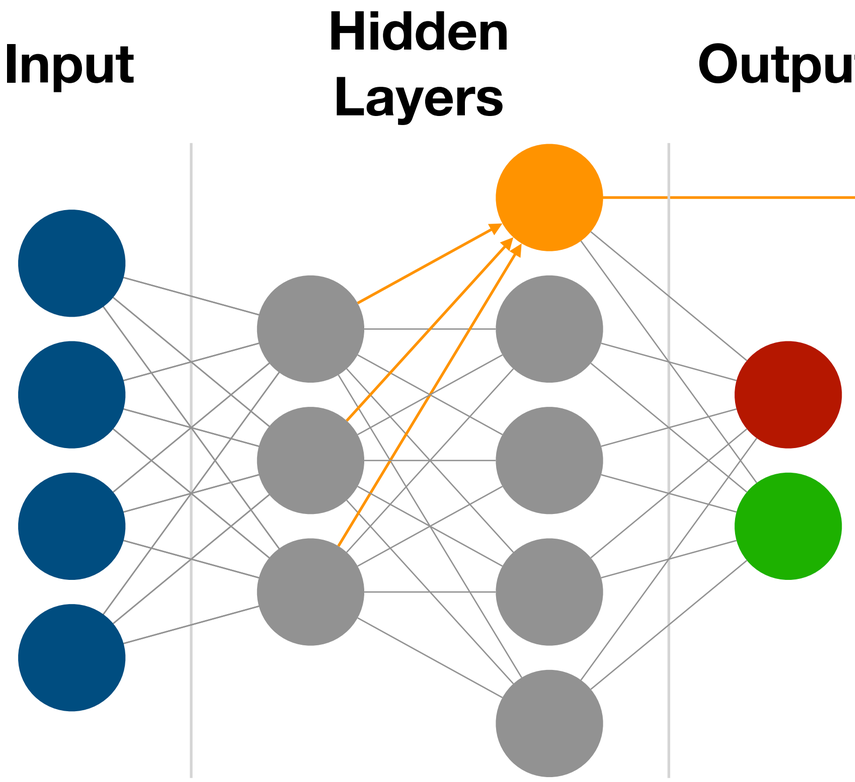


Neural Network

Libra

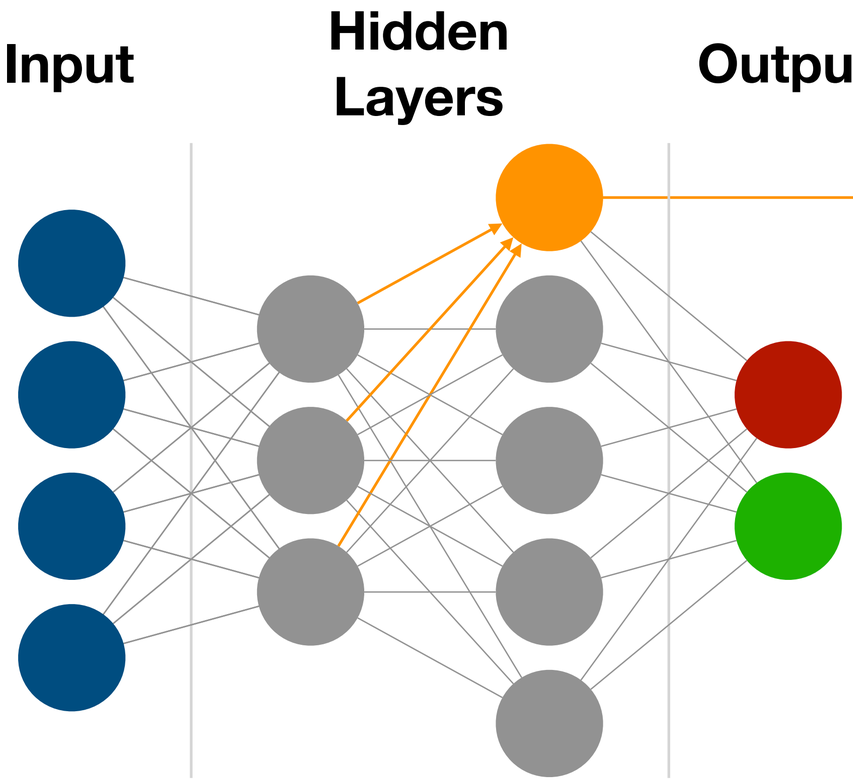


Feed-Forward **Neural Networks** with **ReLU Activations**



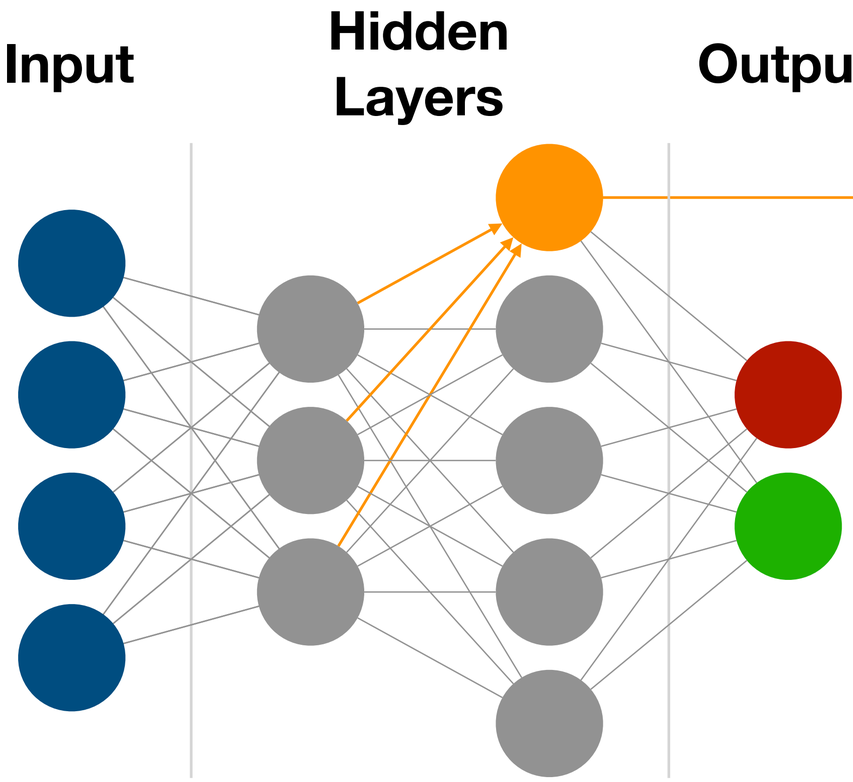
$$x_{i,j} = \text{ReLU}(\sum_k w_k \cdot x_{i-1,k} + b_{i,j})$$

Feed-Forward **Neural Networks** with **ReLU Activations**



$$x_{i,j} = \text{ReLU}(\sum_k w_k \cdot x_{i-1,k} + b_{i,j})$$
$$= \max(\sum_k w_k \cdot x_{i-1,k} + b_{i,j}, 0)$$

Feed-Forward **Neural Networks** with **ReLU Activations**

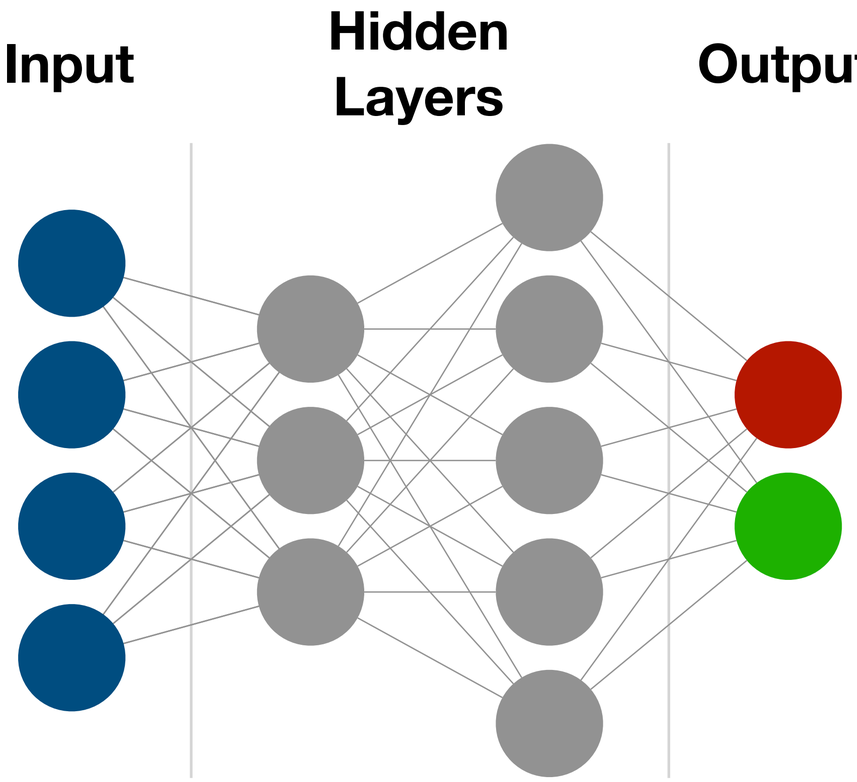


$$x_{i,j} = \text{ReLU}(\sum_k w_k \cdot x_{i-1,k} + b_{i,j})$$

$$= \max(\underbrace{\sum_k w_k \cdot x_{i-1,k} + b_{i,j}}, 0)$$

$$\hat{x}_{i,j} = \sum_k w_k \cdot x_{i-1,k} + b_{i,j}$$

Feed-Forward **Neural Networks** with **ReLU Activations**

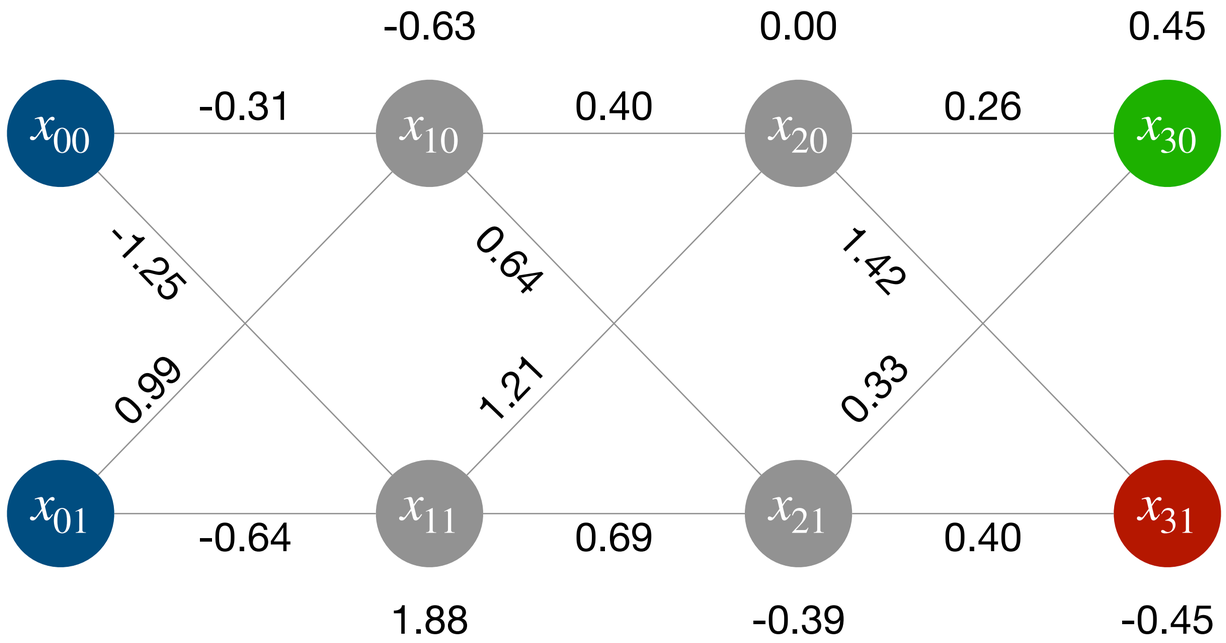


Classification Outcome

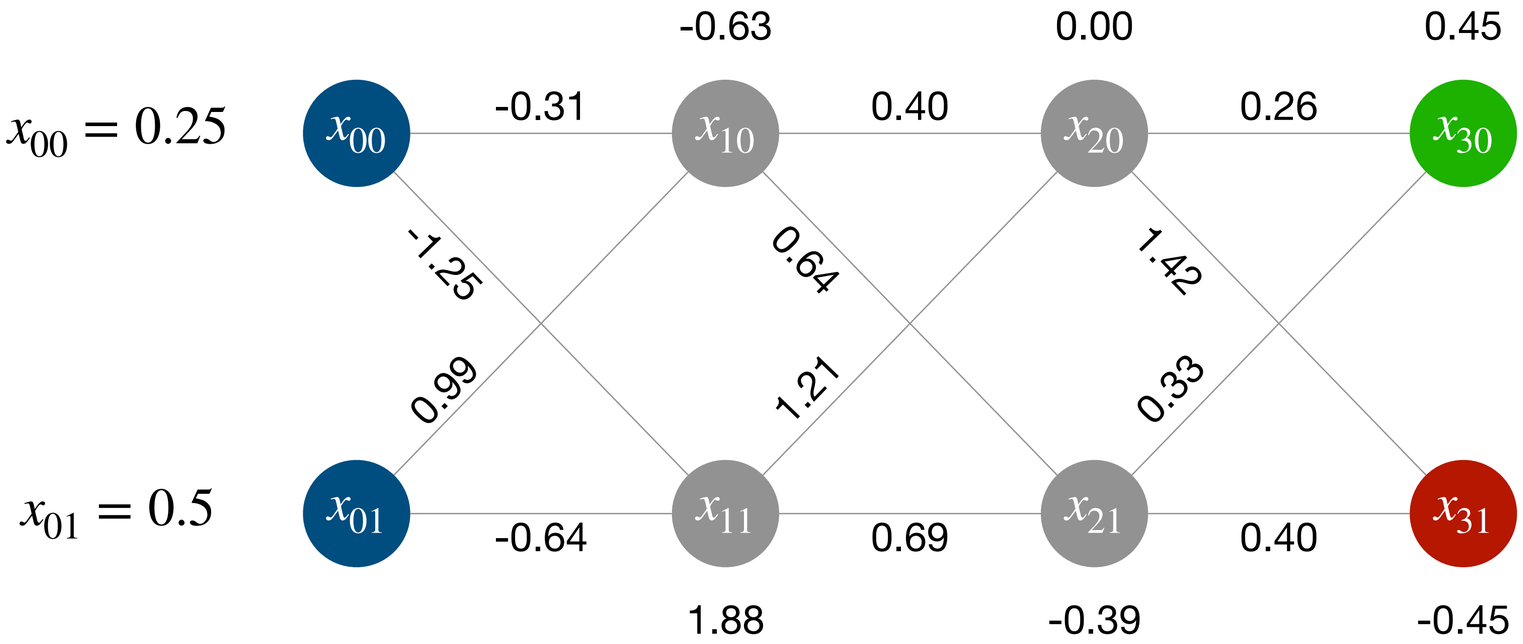
$$\underset{j}{\text{argmax}}\ x_{i,j}$$

for $x_{i,j}$ outputs

Feed-Forward **Neural Networks** with **ReLU Activations**

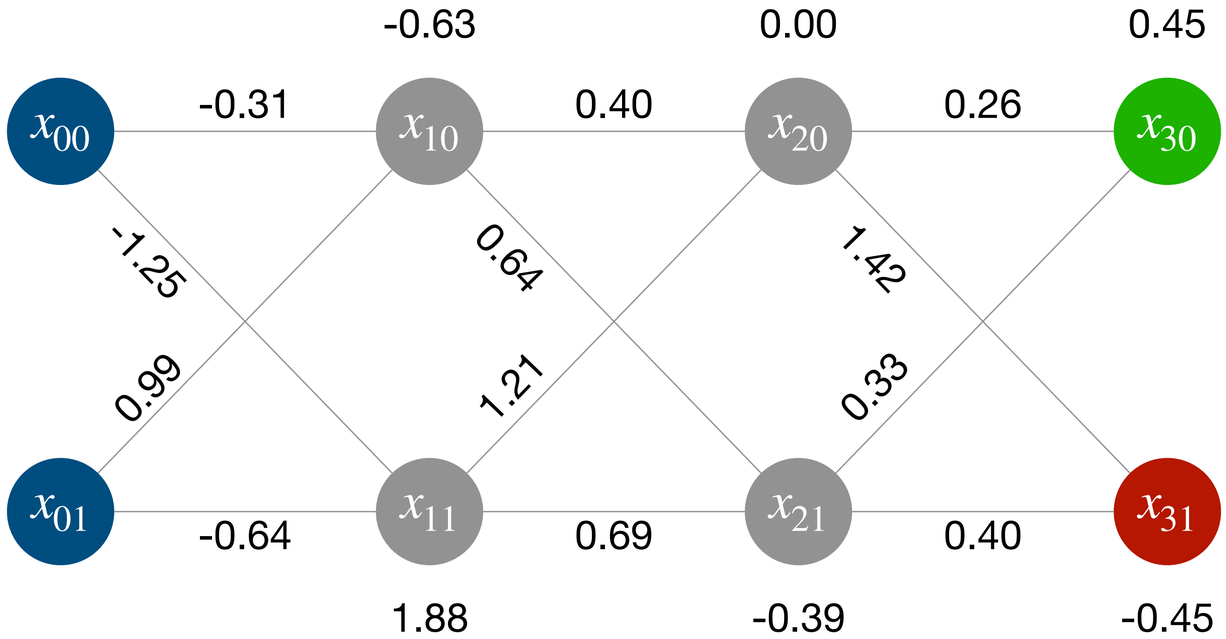


Feed-Forward **Neural Networks** with **ReLU Activations**



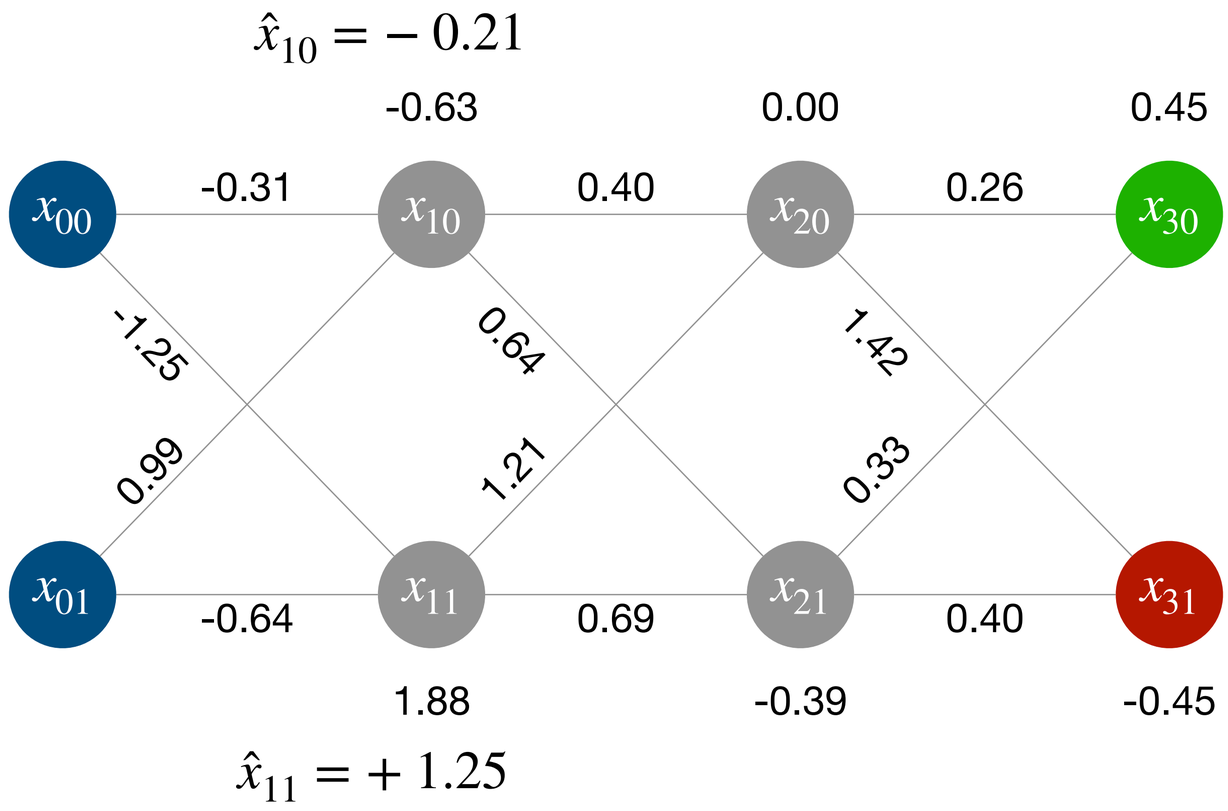
Feed-Forward **Neural Networks** with **ReLU Activations**

$$\hat{x}_{10} = -0.31 \cdot x_{00} + 0.99 \cdot x_{01} - 0.63$$



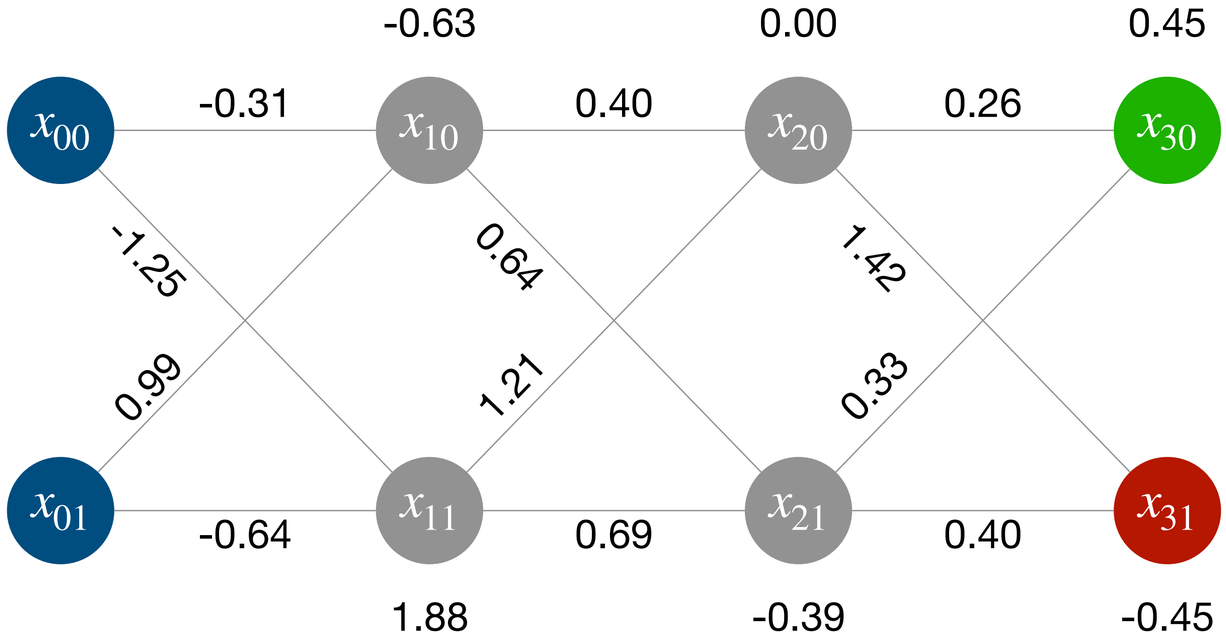
$$\hat{x}_{11} = -1.25 \cdot x_{00} - 0.64 \cdot x_{01} + 1.88$$

Feed-Forward **Neural Networks** with **ReLU Activations**



Feed-Forward **Neural Networks** with **ReLU Activations**

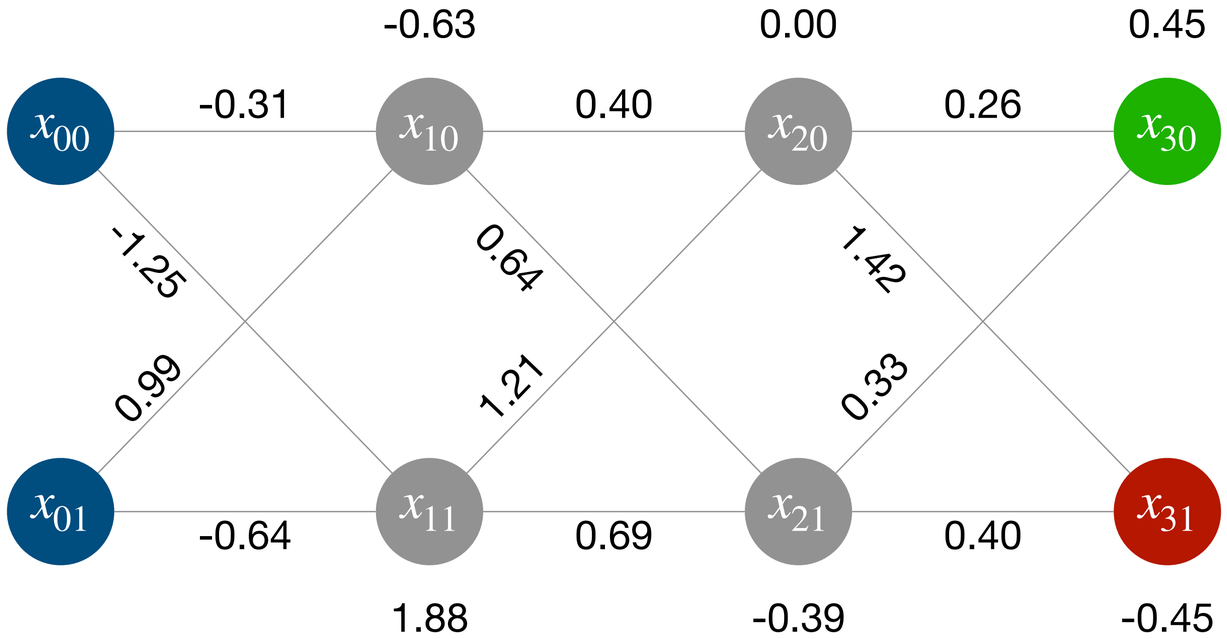
$$x_{10} = \max(-0.21, 0) = 0$$



$$x_{11} = \max(+1.25, 0) = +1.25$$

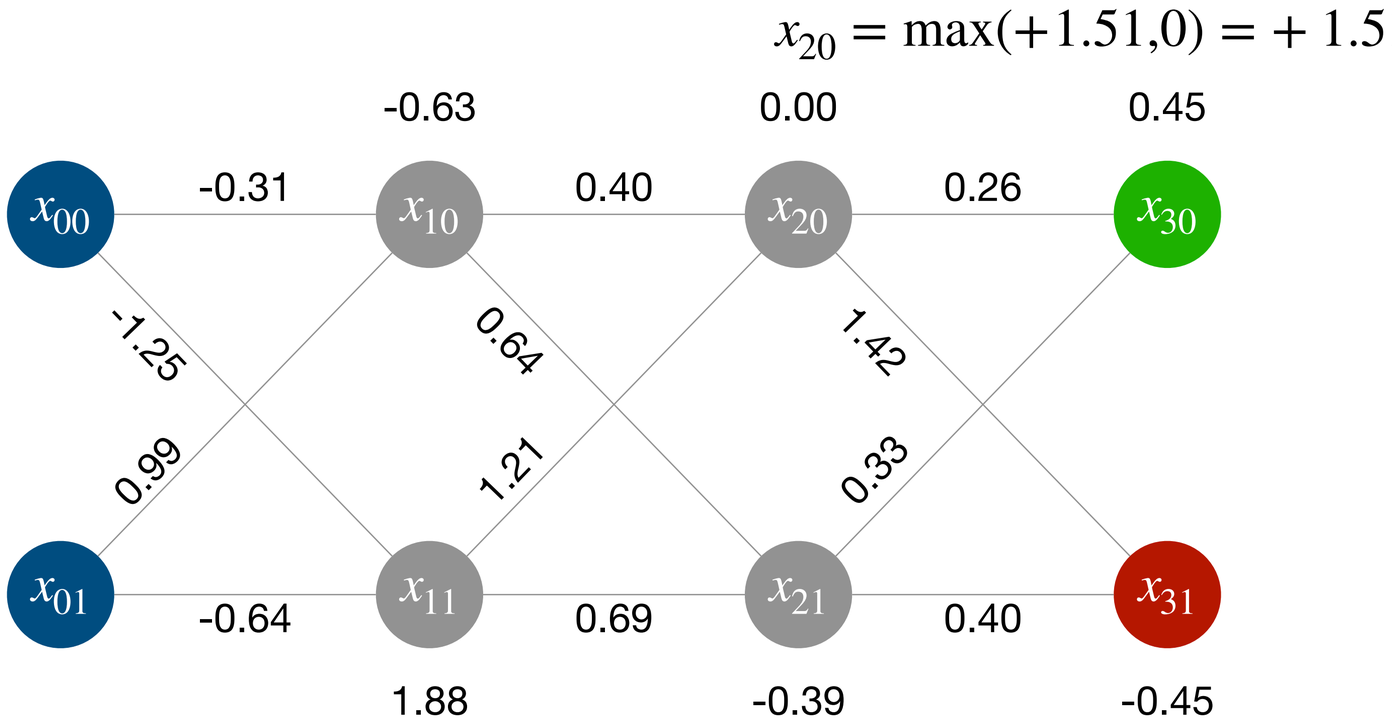
Feed-Forward **Neural Networks** with **ReLU Activations**

$$\hat{x}_{20} = 0.4 \cdot x_{10} + 1.21 \cdot x_{11} = +1.51$$



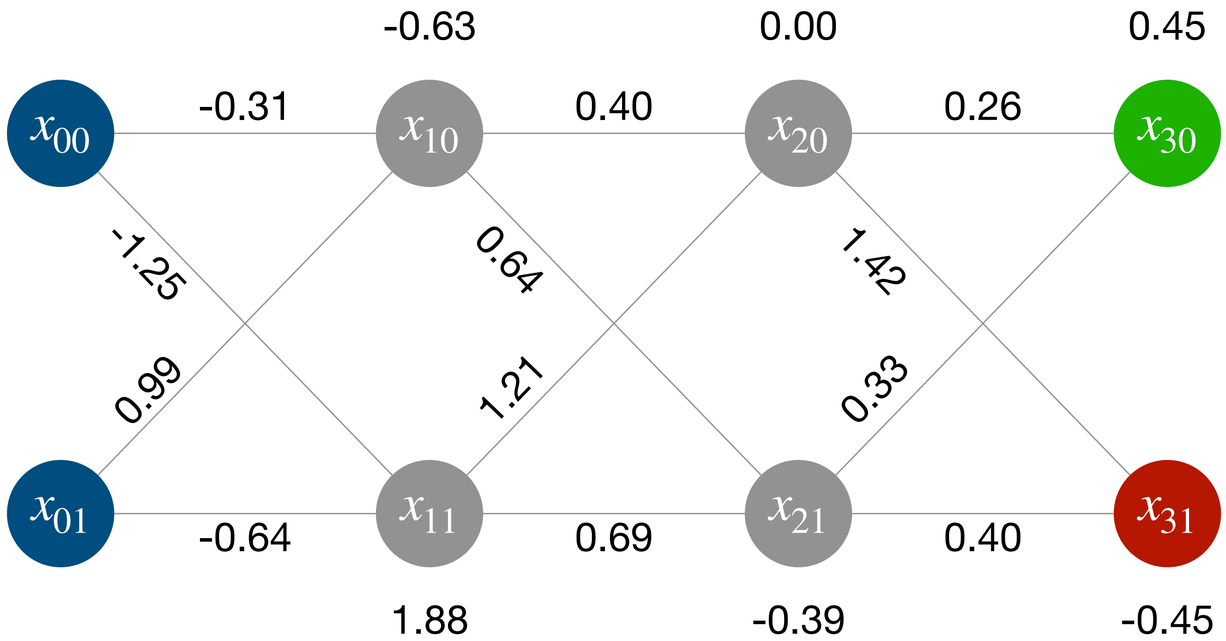
$$\hat{x}_{21} = 0.64 * x10 + 0.69 * x11 - 0.39 = +0.47$$

Feed-Forward **Neural Networks** with **ReLU Activations**



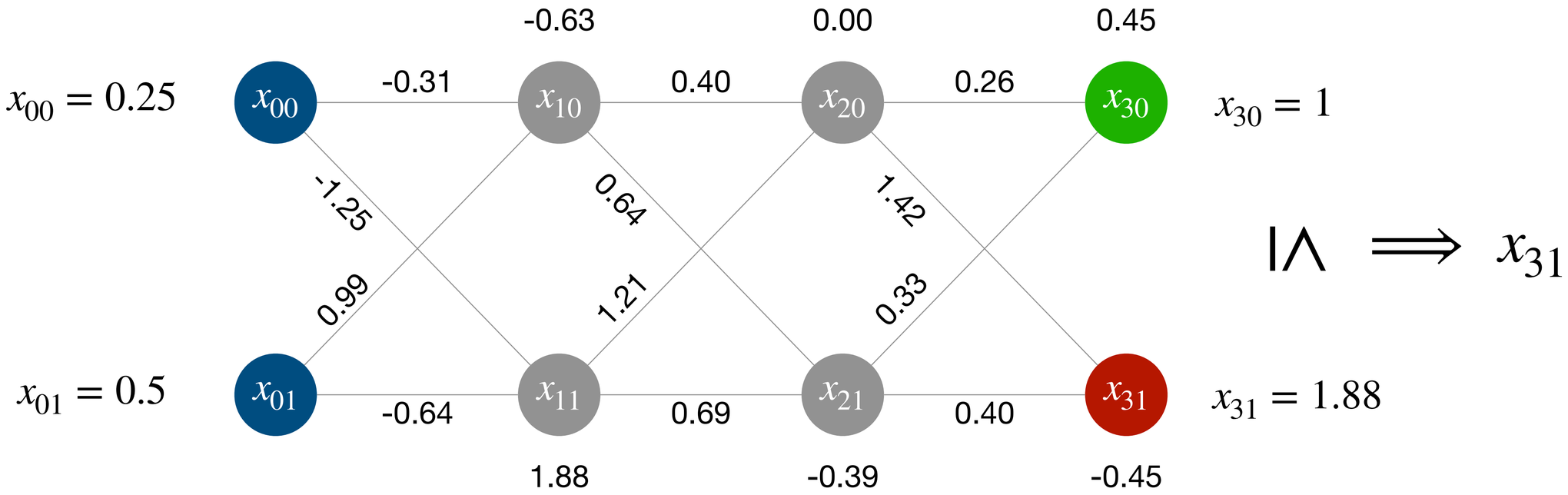
Feed-Forward **Neural Networks** with **ReLU Activations**

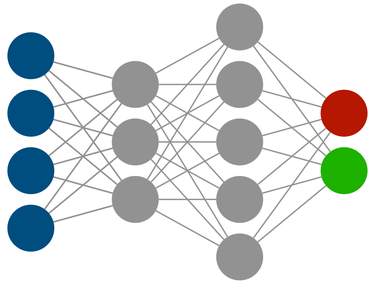
$$x_{30} = 0.26 \cdot x_{20} + 0.33 \cdot x_{21} + 0.45$$
$$= 1$$



$$x_{31} = 1.42 \cdot x_{20} + 0.4 \cdot x_{21} - 0.45$$
$$= 1.88$$

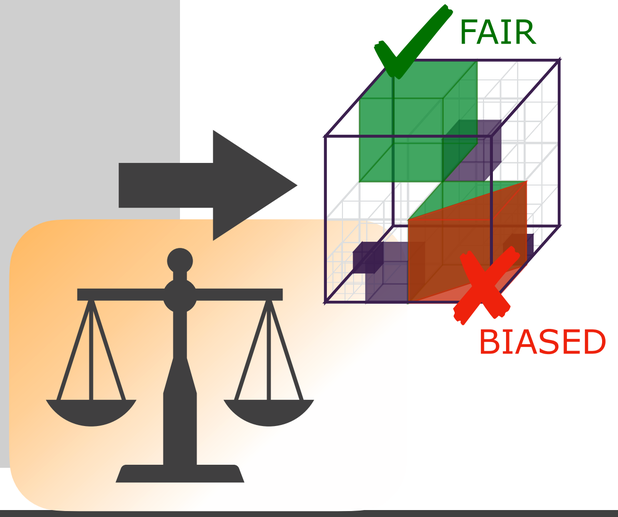
Feed-Forward **Neural Networks** with **ReLU Activations**





Neural Network

Libra

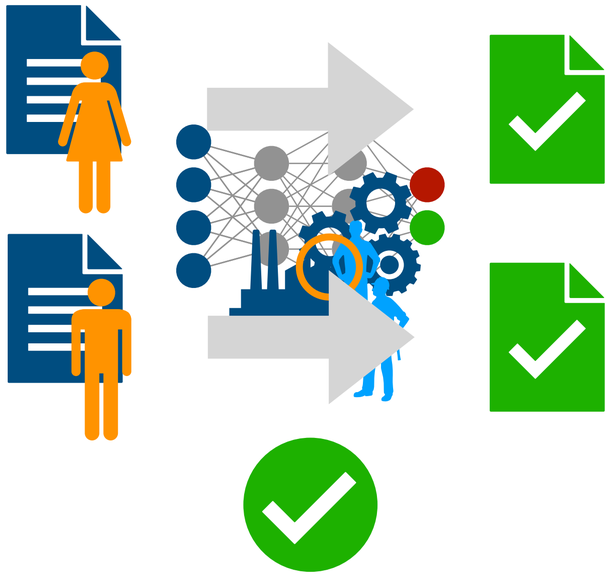


Dependency **Fairness**

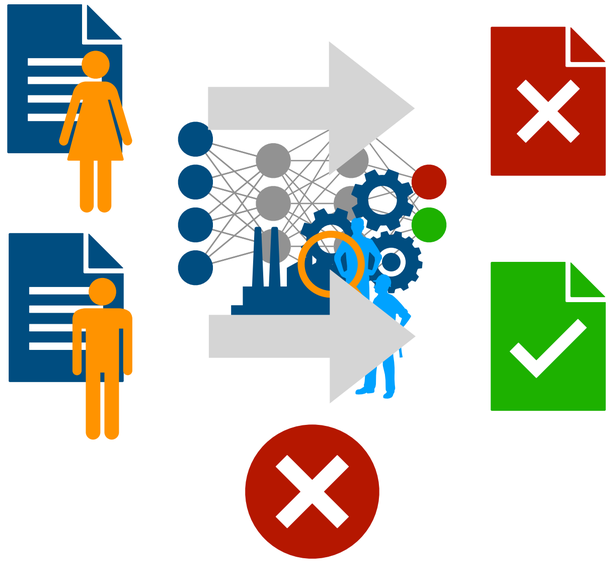
The classification outcome is
Independent on the
Sensitive Features

Recruiting Process

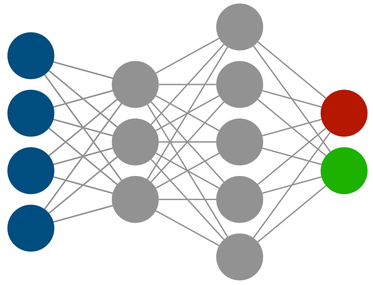
Recruiting Process



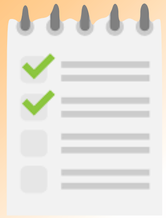
Fair



Unfair

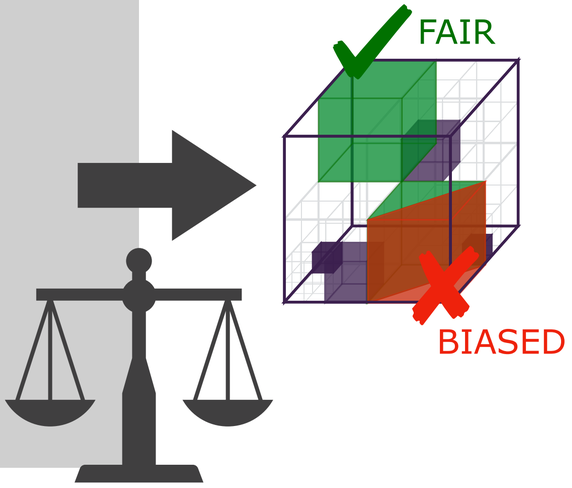


Neural Network

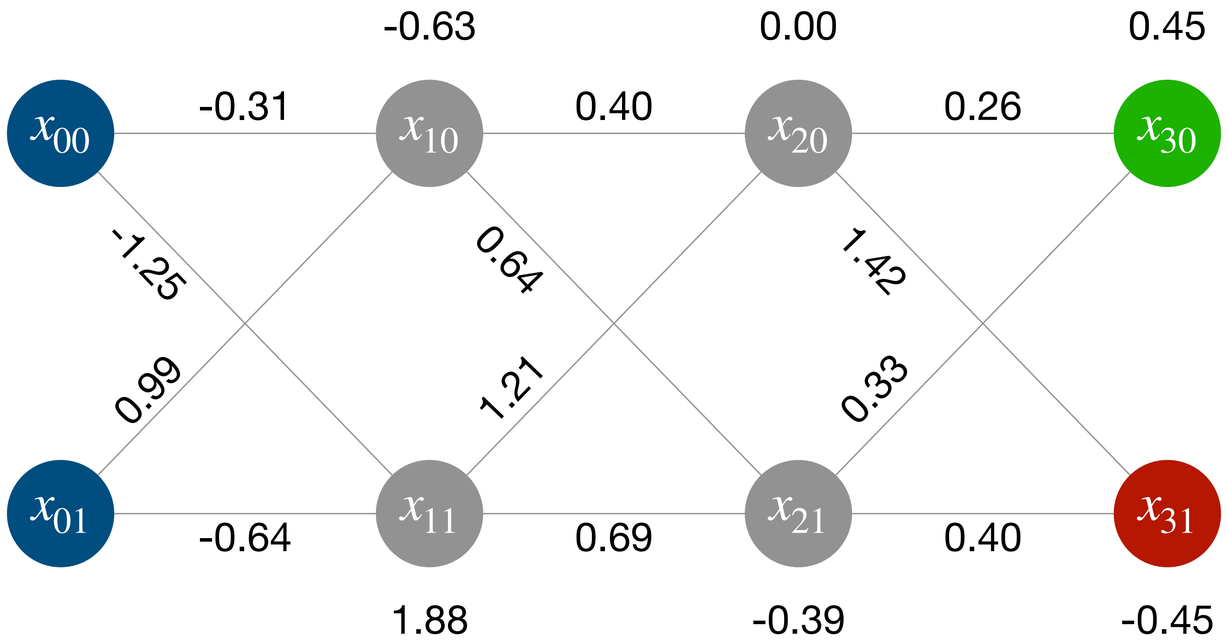


Specification

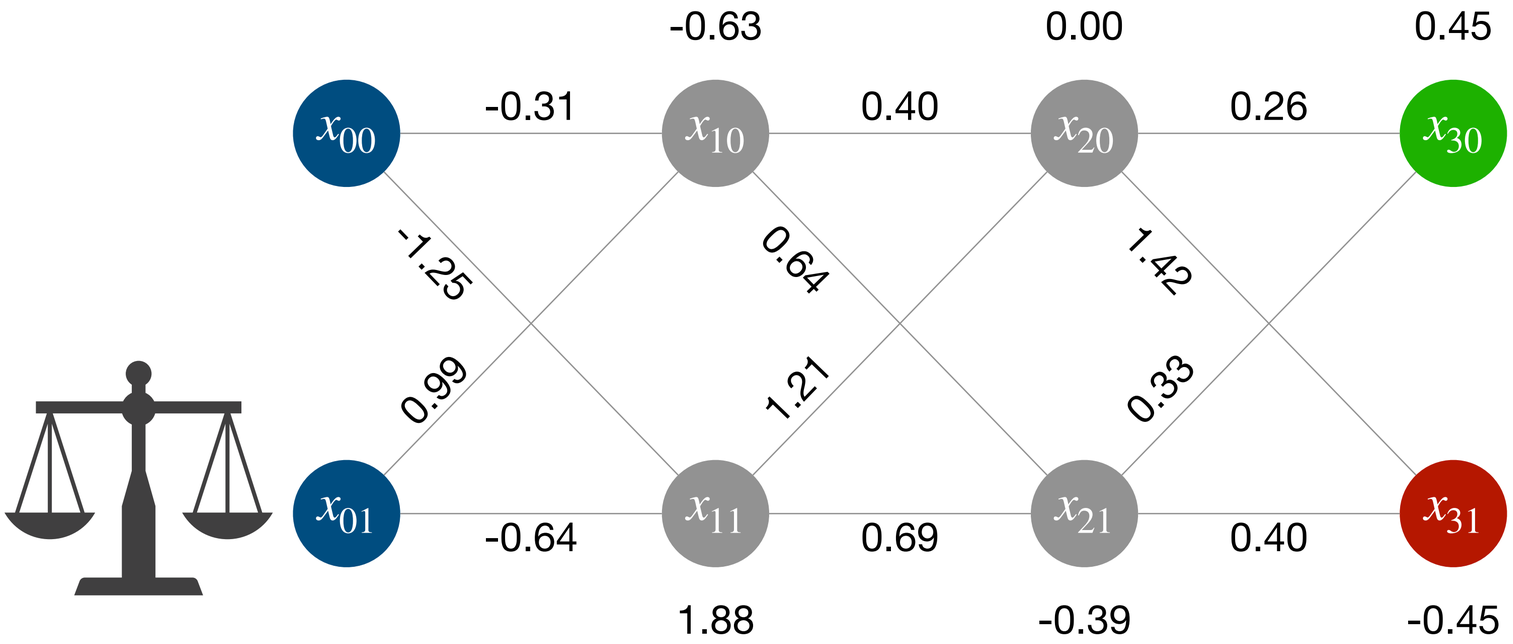
Libra

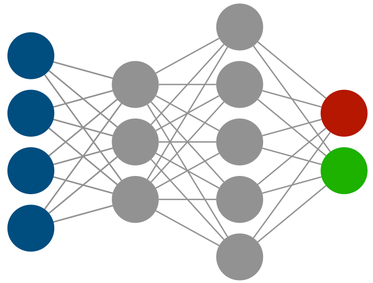


Specification

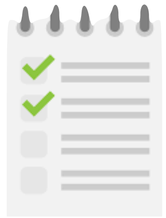


Specification

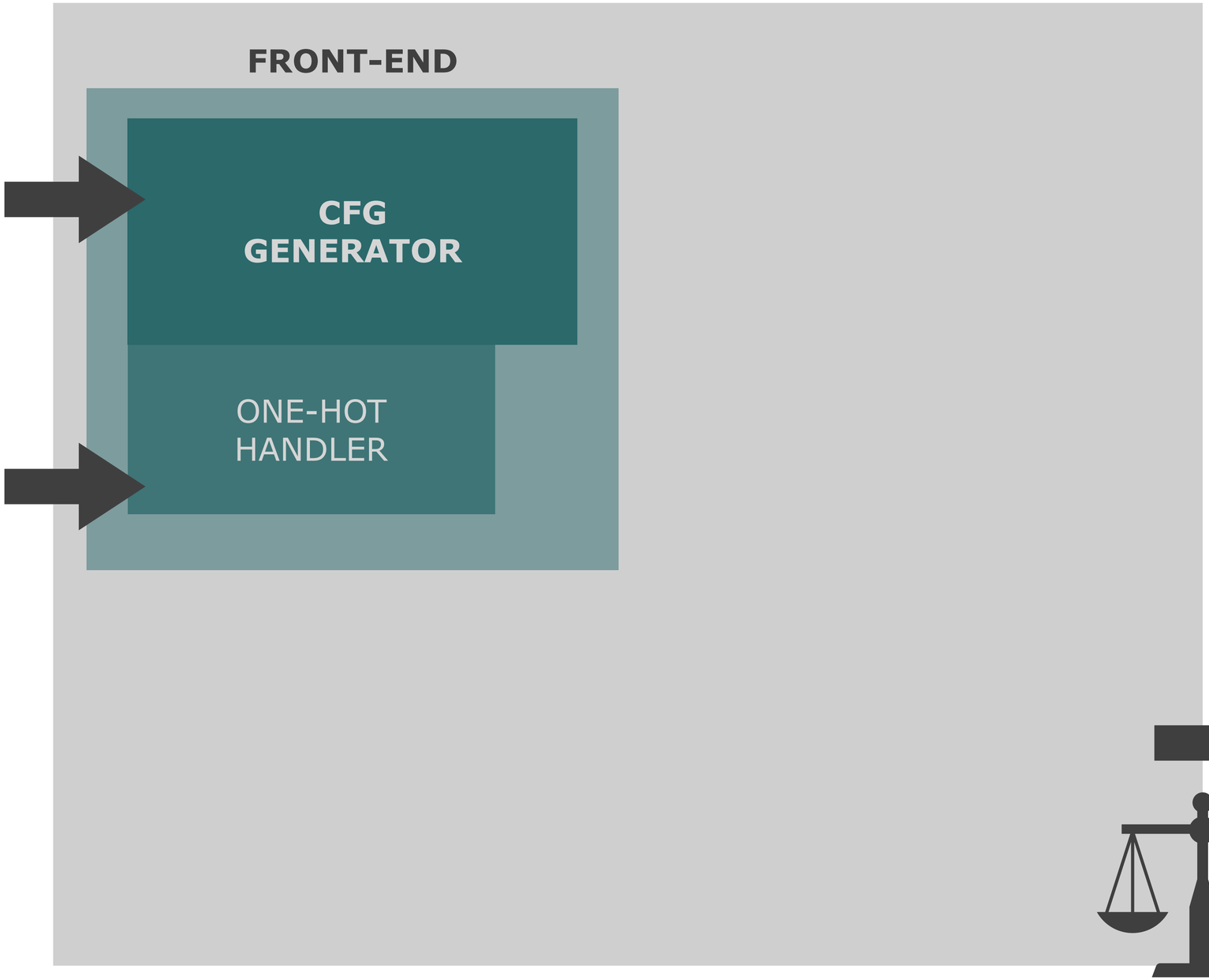


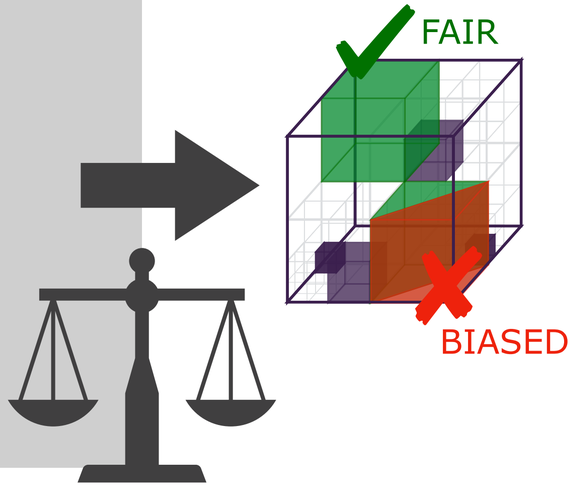


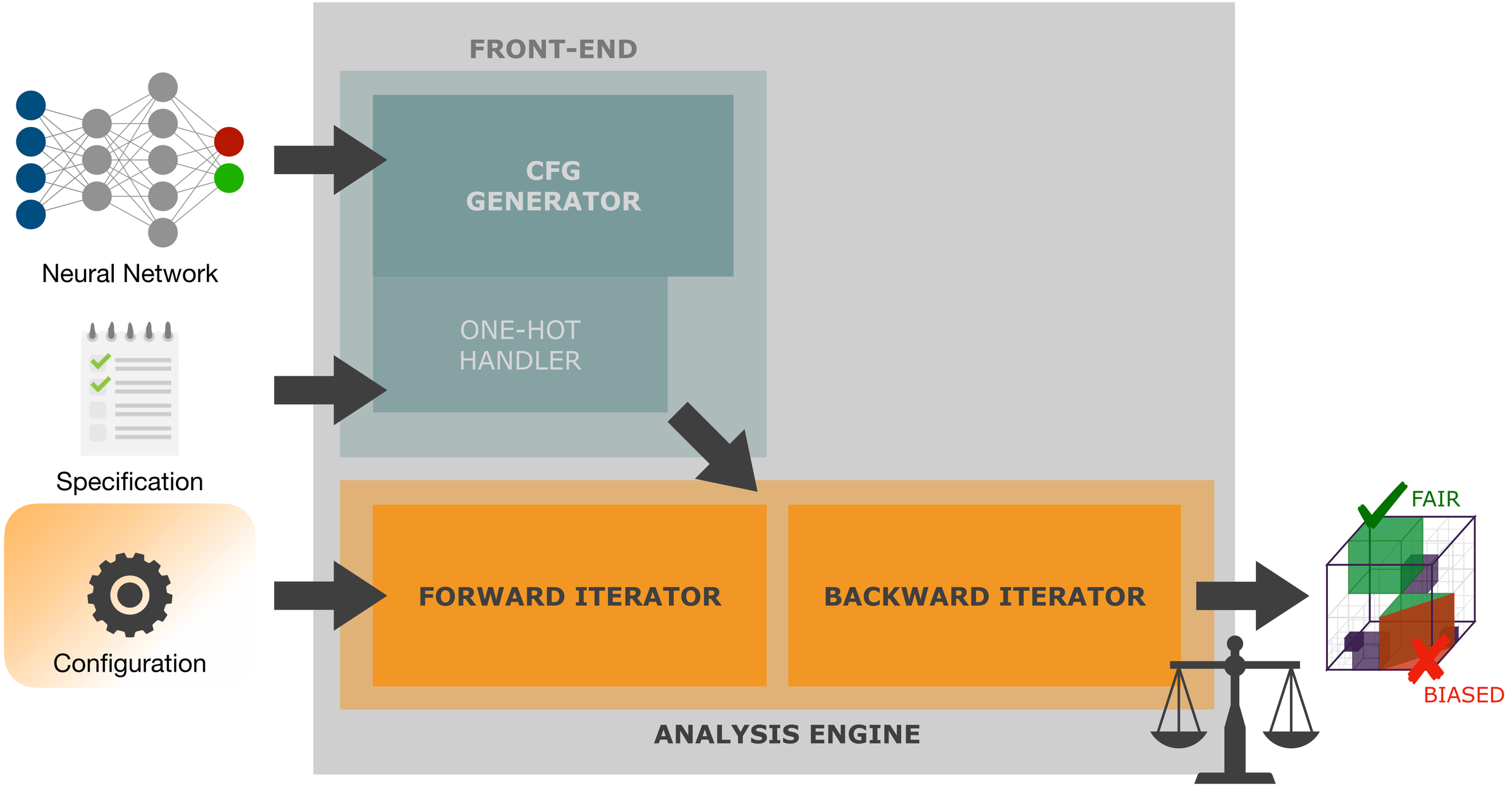
Neural Network



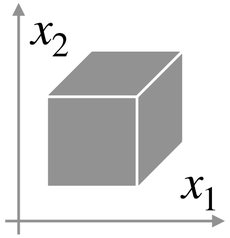
Specification



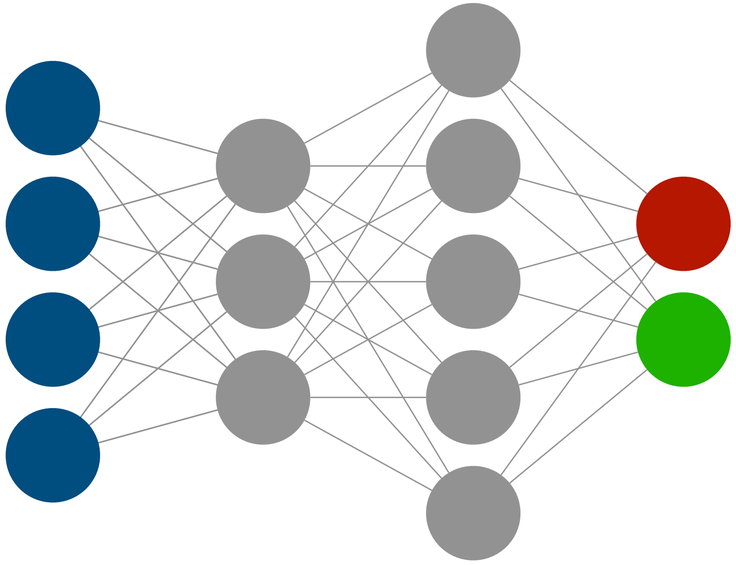


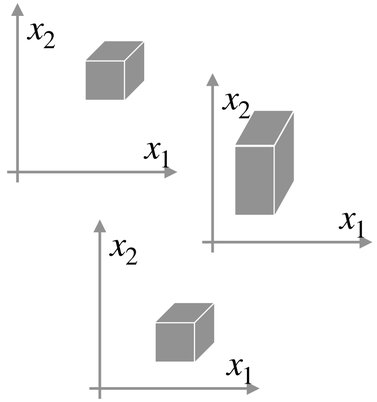

FRONT-END
CFG GENERATOR
ONE-HOT HANDLER
Neural Network
Specification
Configuration
FORWARD ITERATOR
BACKWARD ITERATOR
ANALYSIS ENGINE
FAIR
BIASED

Cheap **Forward Pre-Analysis**

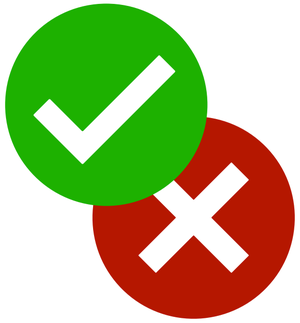


Input Space



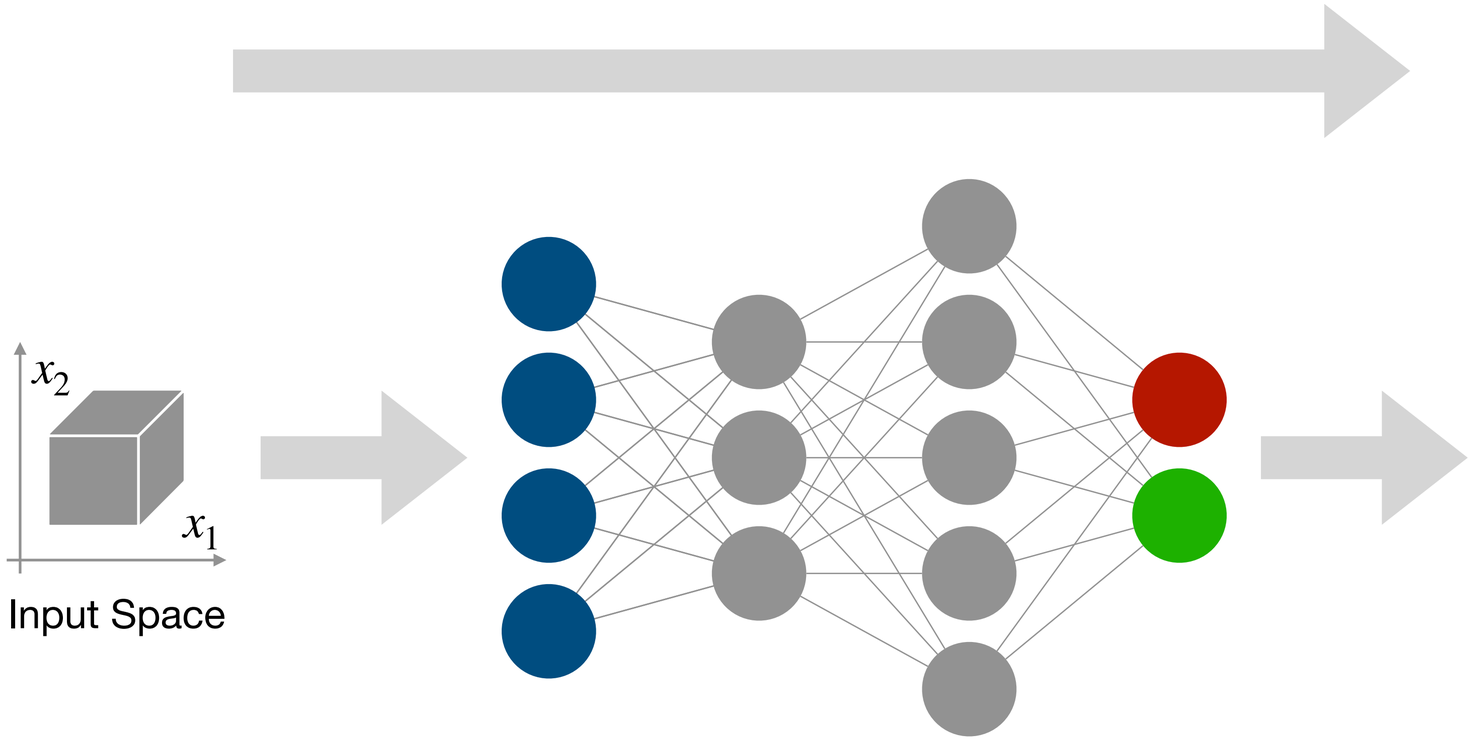


Partitions



Exact **Backward Analysis**
using **Polyhedra**

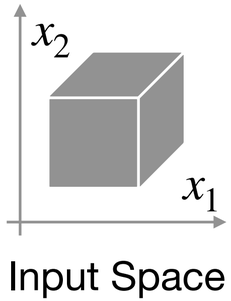
Cheap **Forward Pre-Analysis**



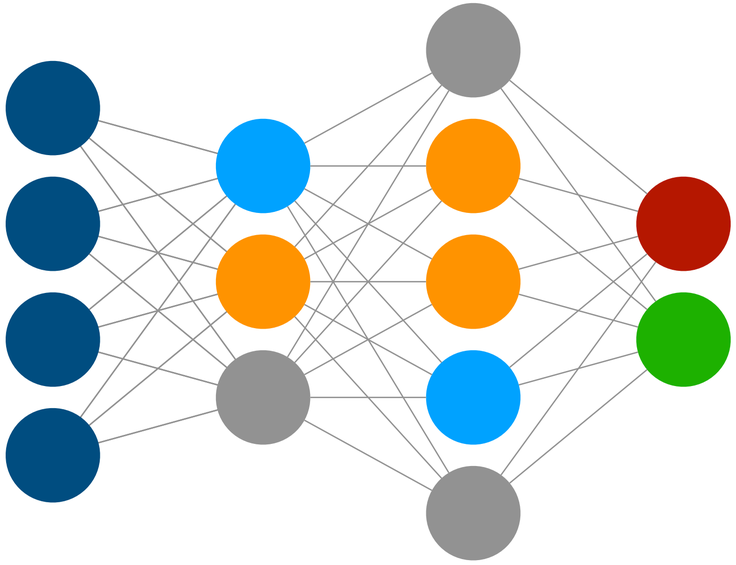
- **Fair**

Propagate the partition through the network via **abstract domains**

Cheap **Forward Pre-Analysis**



Input Space



- **Fair**
- Or, …

Activation Patterns

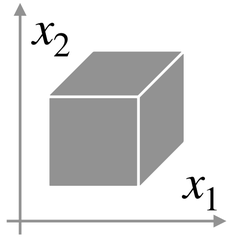
Always Negative $\max(\hat{x},0) = 0$

Always Positive $\max(\hat{x},0) = \hat{x}$

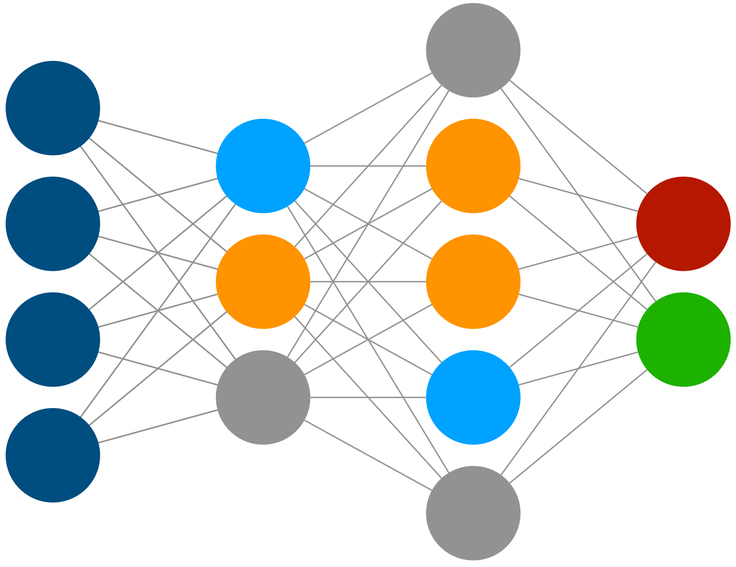
Uncertain $\hat{x} \in [a, b]$ for $a < 0 < b$

Fixed Activation Status

Cheap **Forward Pre-Analysis**

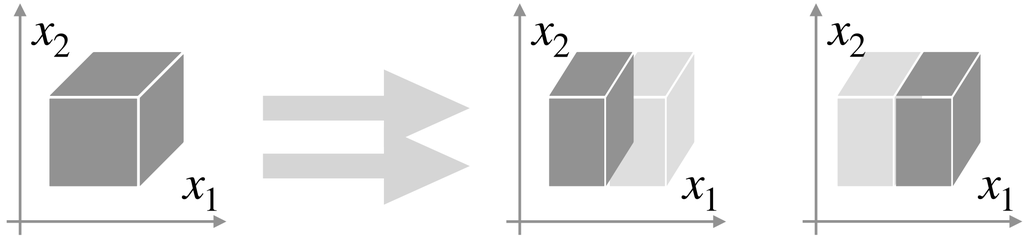


Input Space



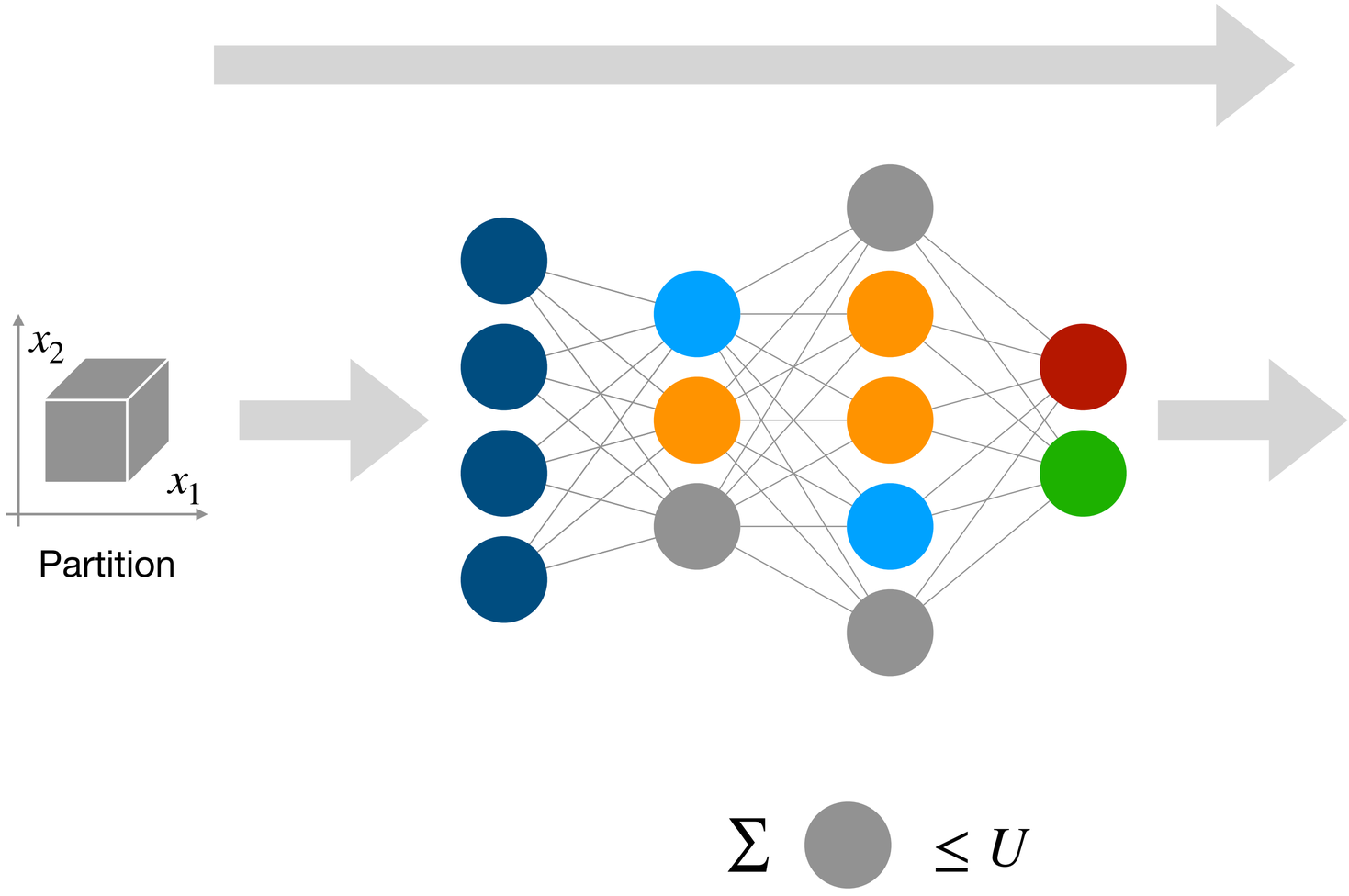
- **Fair**
  - **Partitioned**

$\sum$ ● $\geq U$



Along non-sensitive features only

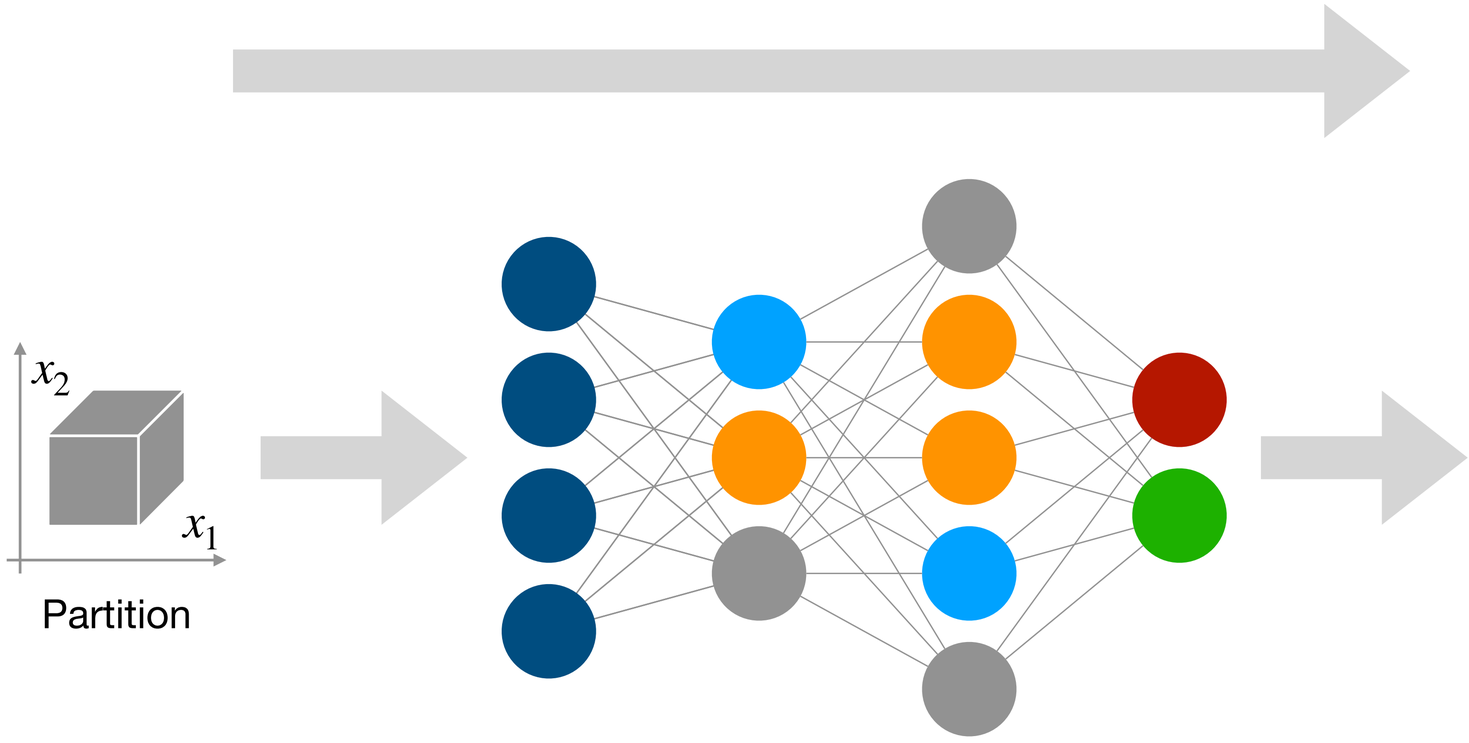
Cheap **Forward Pre-Analysis**



- **Fair**
- **Partitioned**
  - **Feasible**

$$\sum \bigcirc \leq U$$

Cheap **Forward Pre-Analysis**

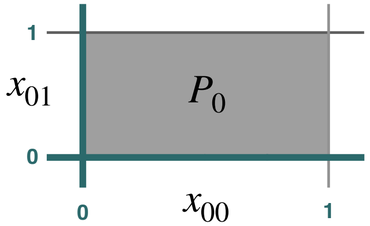


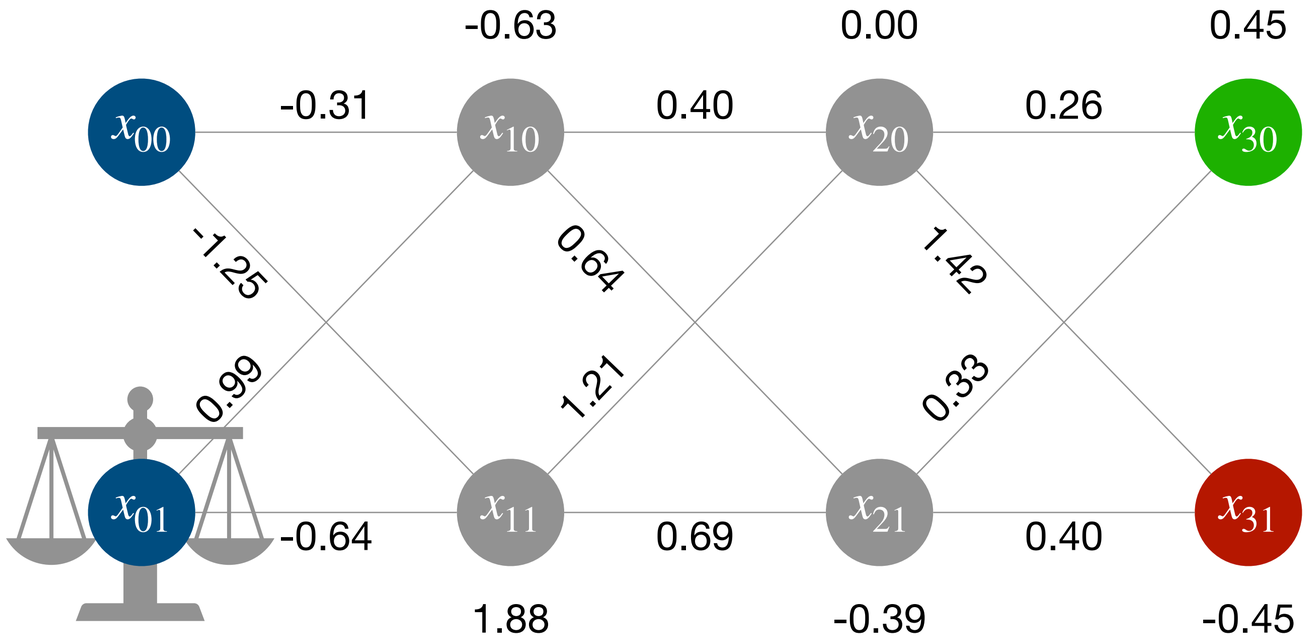
- **Fair**
- **Partitioned**
- **Feasible**
- ▸ **Excluded**

$\sum$ ● $\geq U$, and the partition becomes smaller than L

Forward Analysis

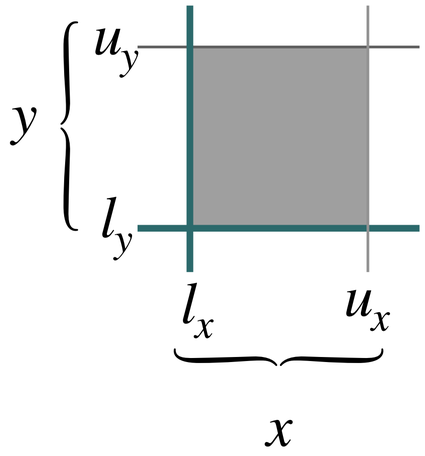
$L = 0.25, U = 2$



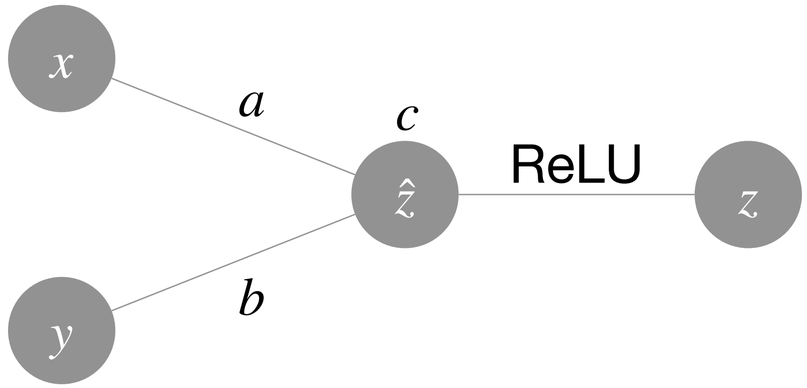


Forward Analysis

Boxes domain



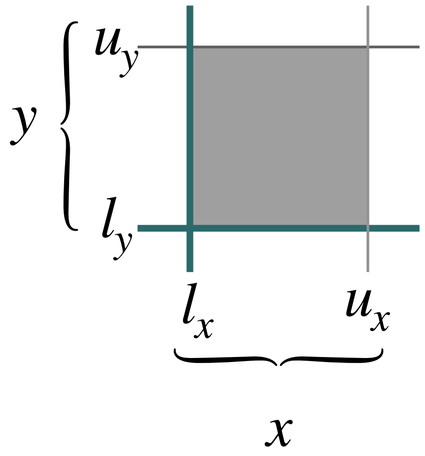
$x \in [l_x, u_x]$



$y \in [l_y, u_y]$

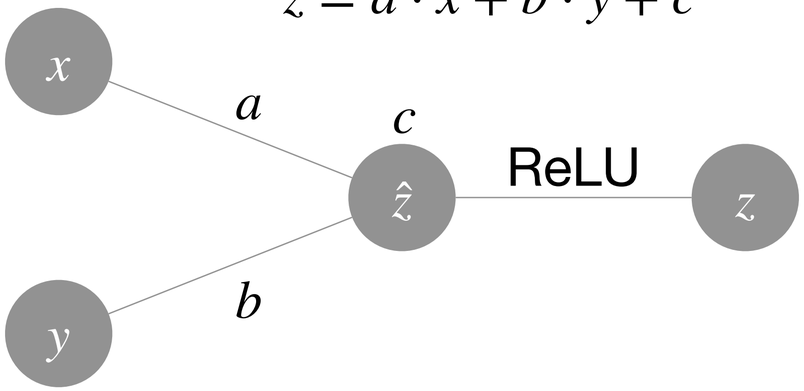
Forward Analysis

Boxes domain



$x \in [l_x, u_x]$

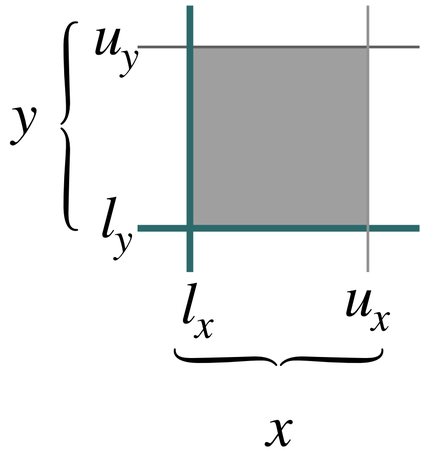
$\hat{z} = a \cdot x + b \cdot y + c$

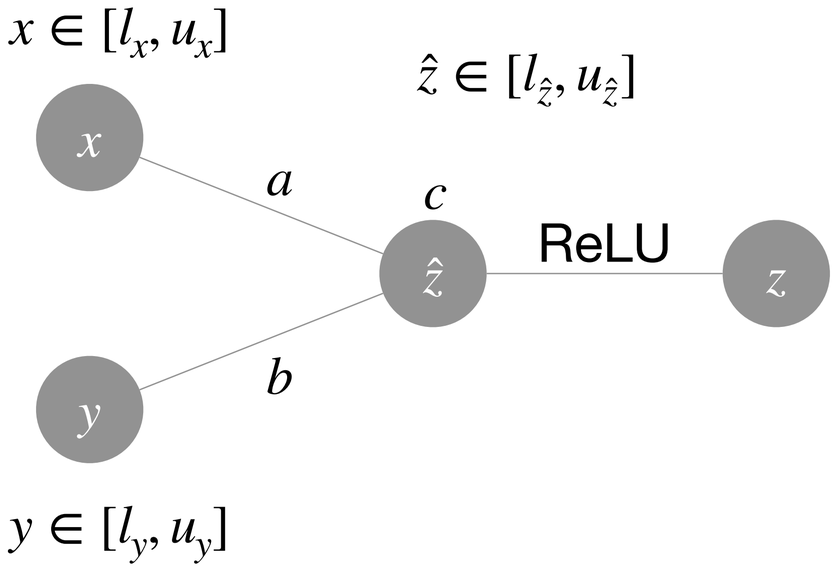


$y \in [l_y, u_y]$

Forward Analysis

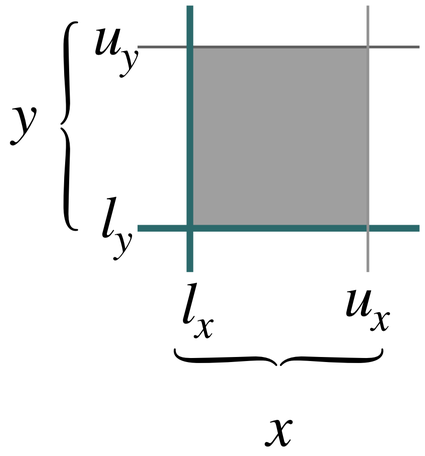
Boxes domain





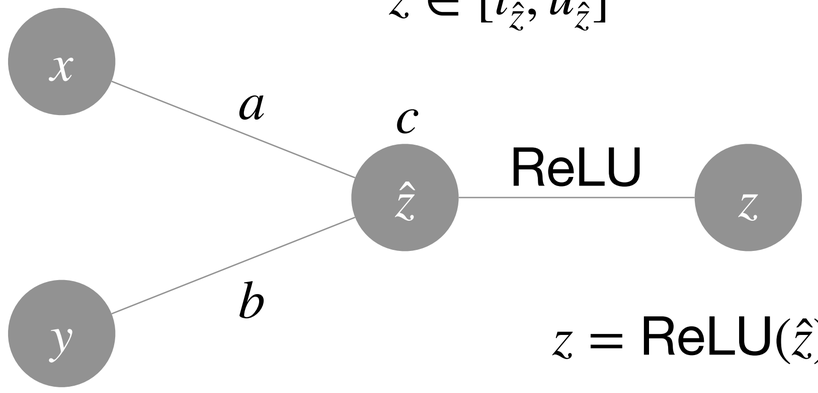
Forward Analysis

Boxes domain



$x \in [l_x, u_x]$

$\hat{z} \in [l_{\hat{z}}, u_{\hat{z}}]$

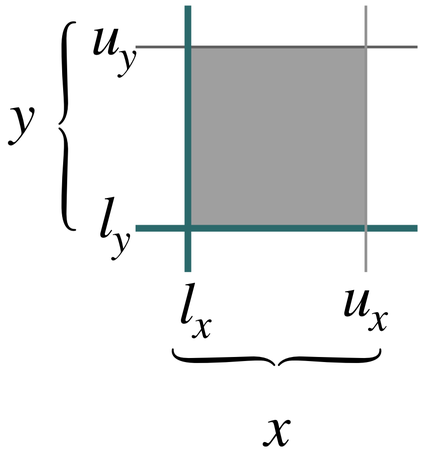


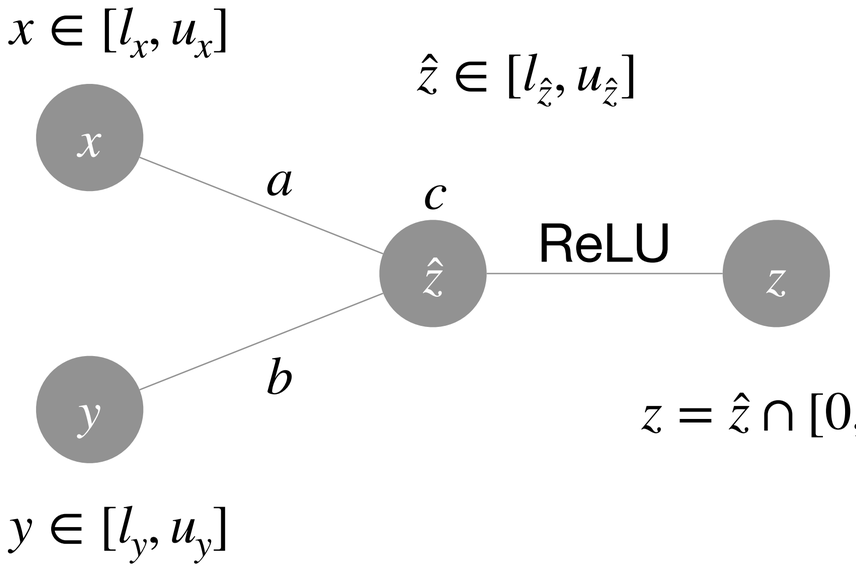
$y \in [l_y, u_y]$

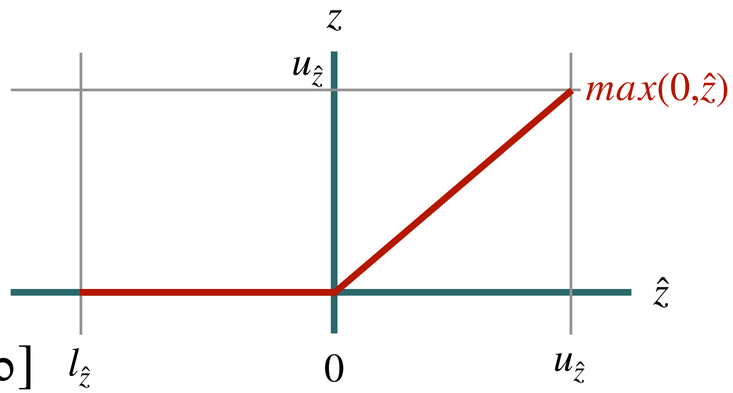
$z = \text{ReLU}(\hat{z}) = \max(\hat{z}, 0)$

Forward Analysis

Boxes domain



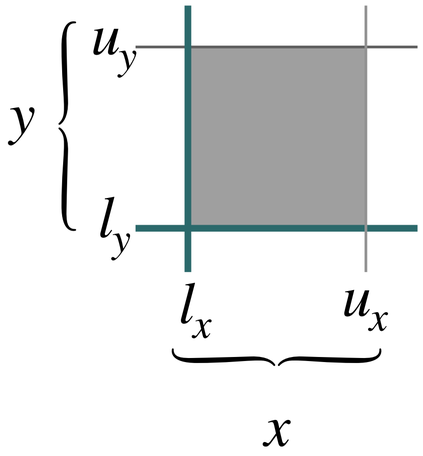


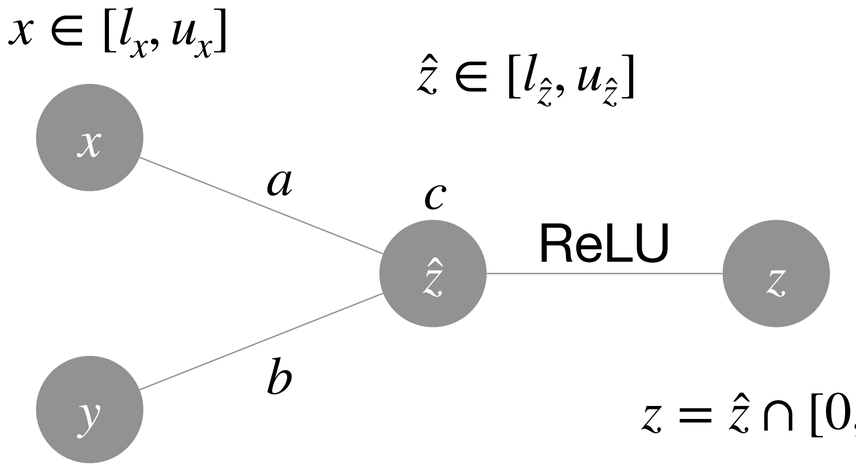


$z = \hat{z} \cap [0, +\infty]$

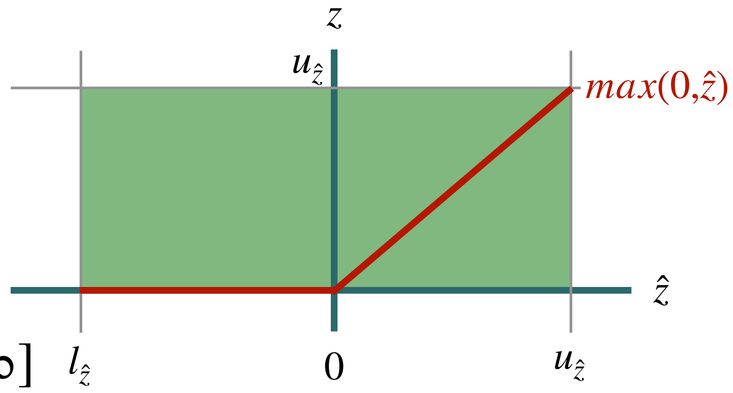
Forward Analysis

Boxes domain





$z = \hat{z} \cap [0, +\infty]$

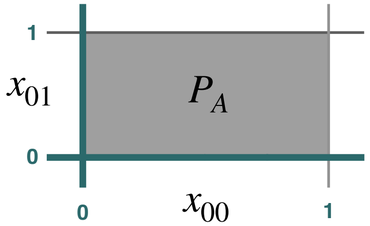


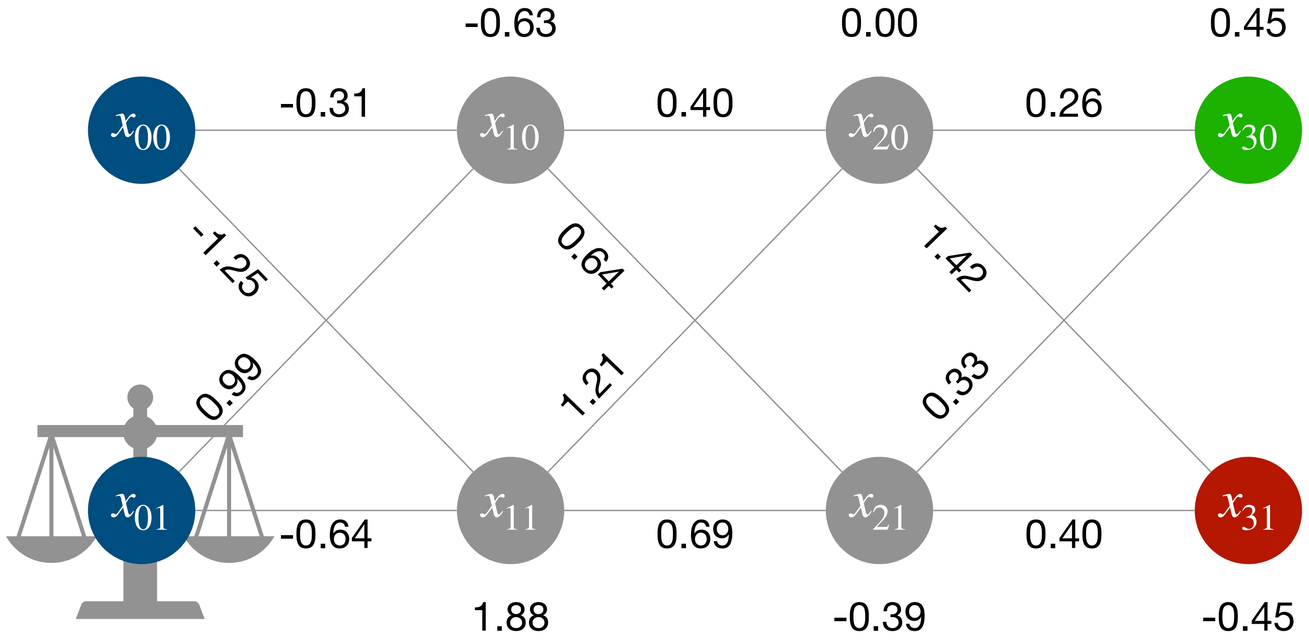
$z \in [0, u_{\hat{z}}]$

$0 \leq z \leq u_{\hat{z}}$

Forward Analysis

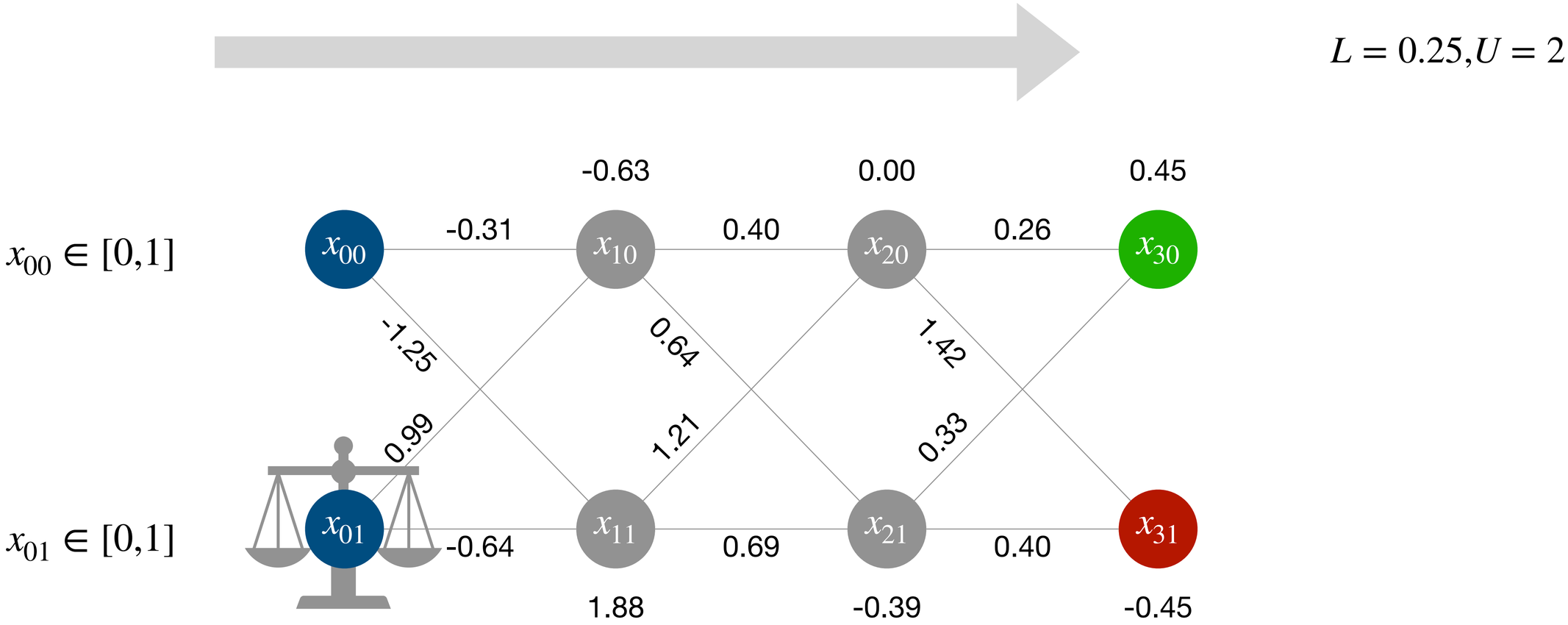
$L = 0.25, U = 2$



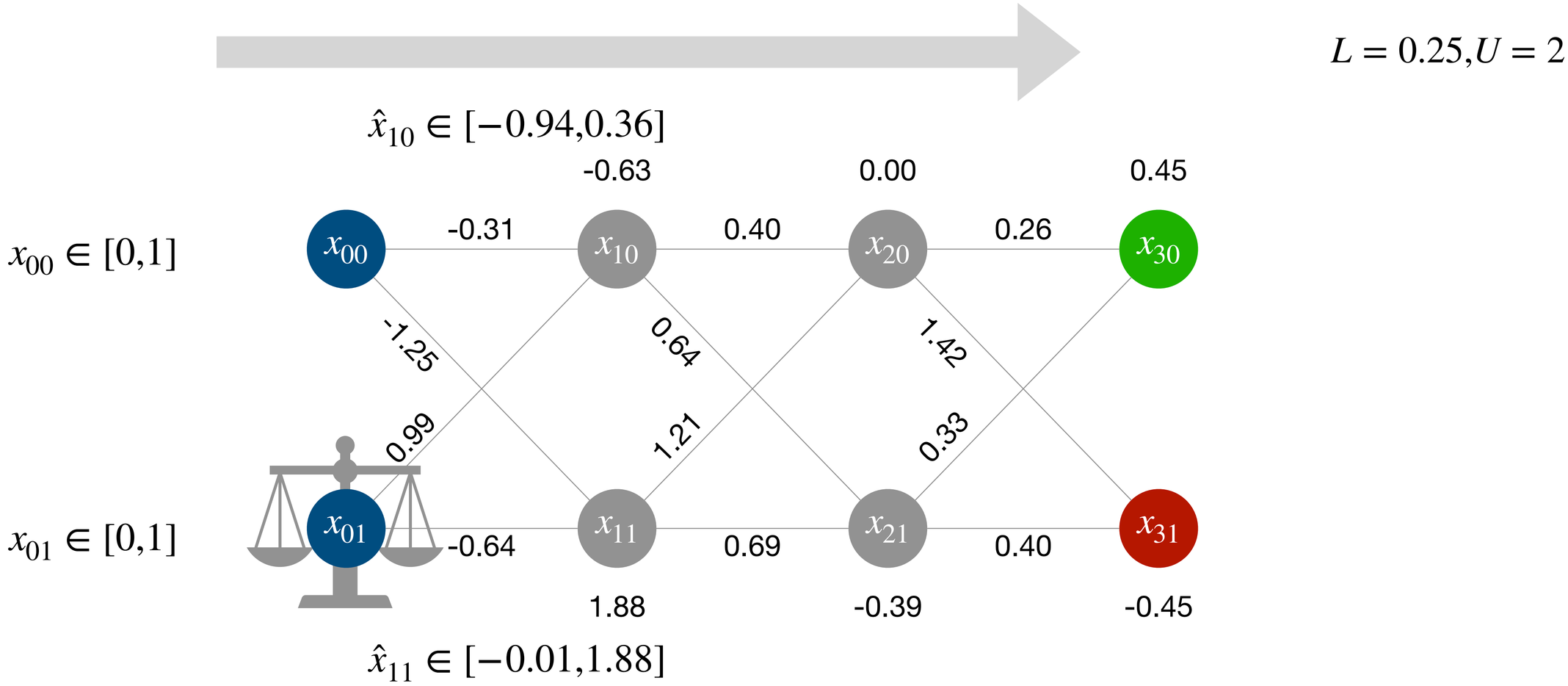


Forward Analysis

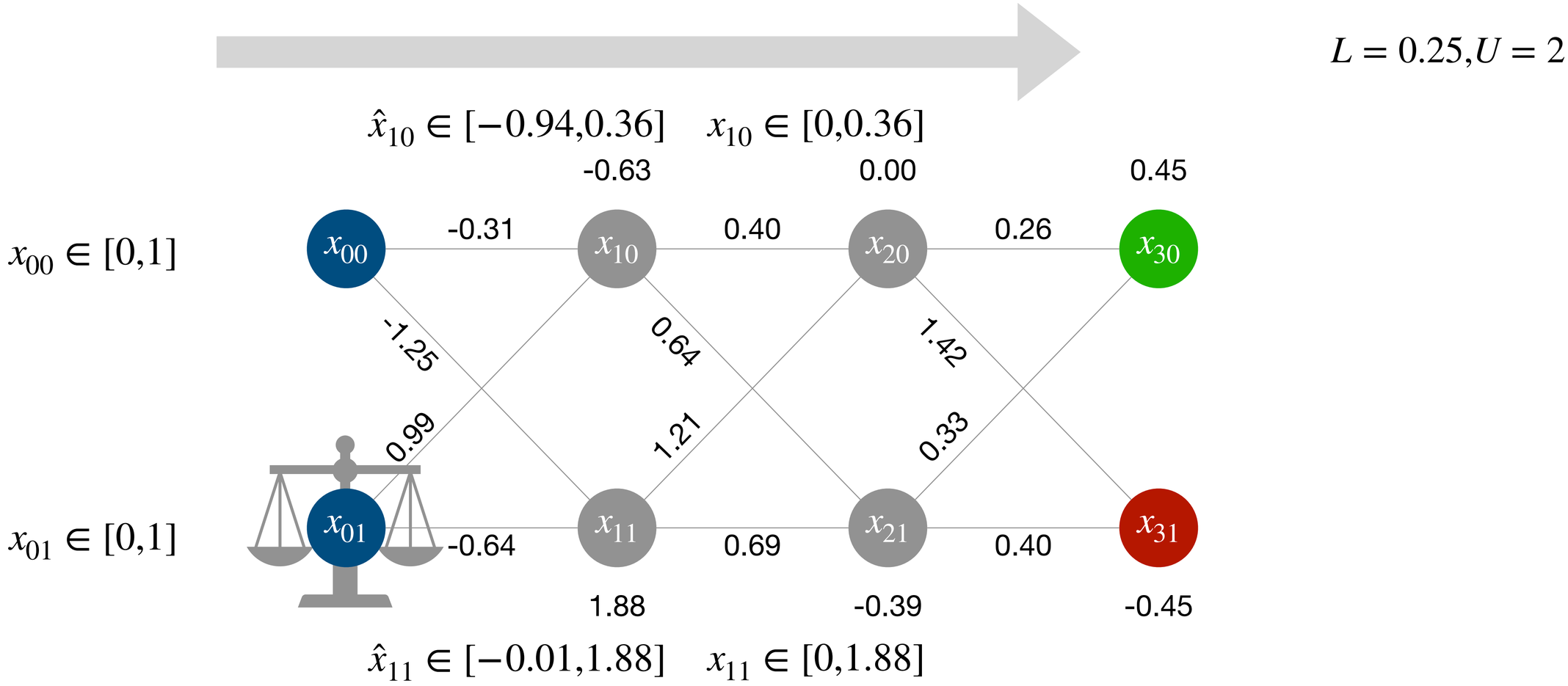
$L = 0.25, U = 2$



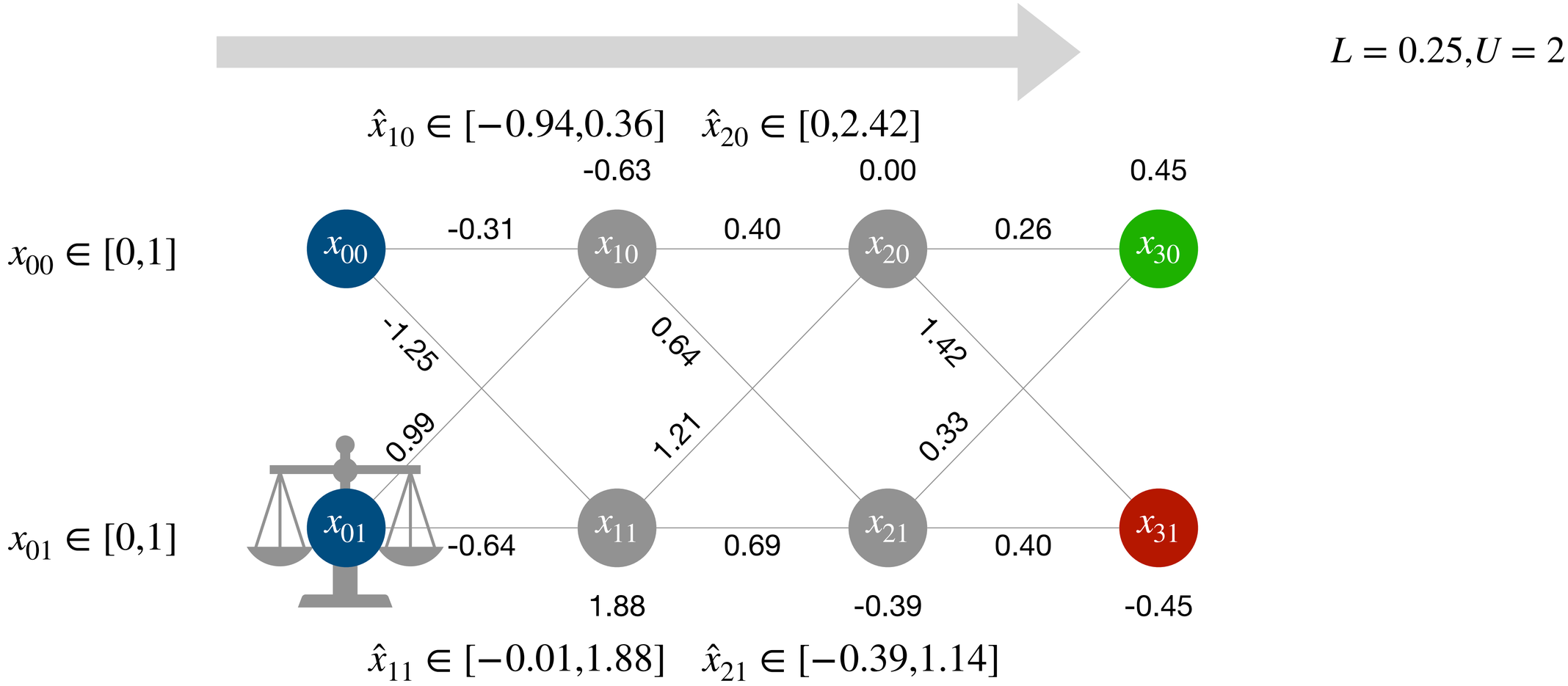
Forward Analysis



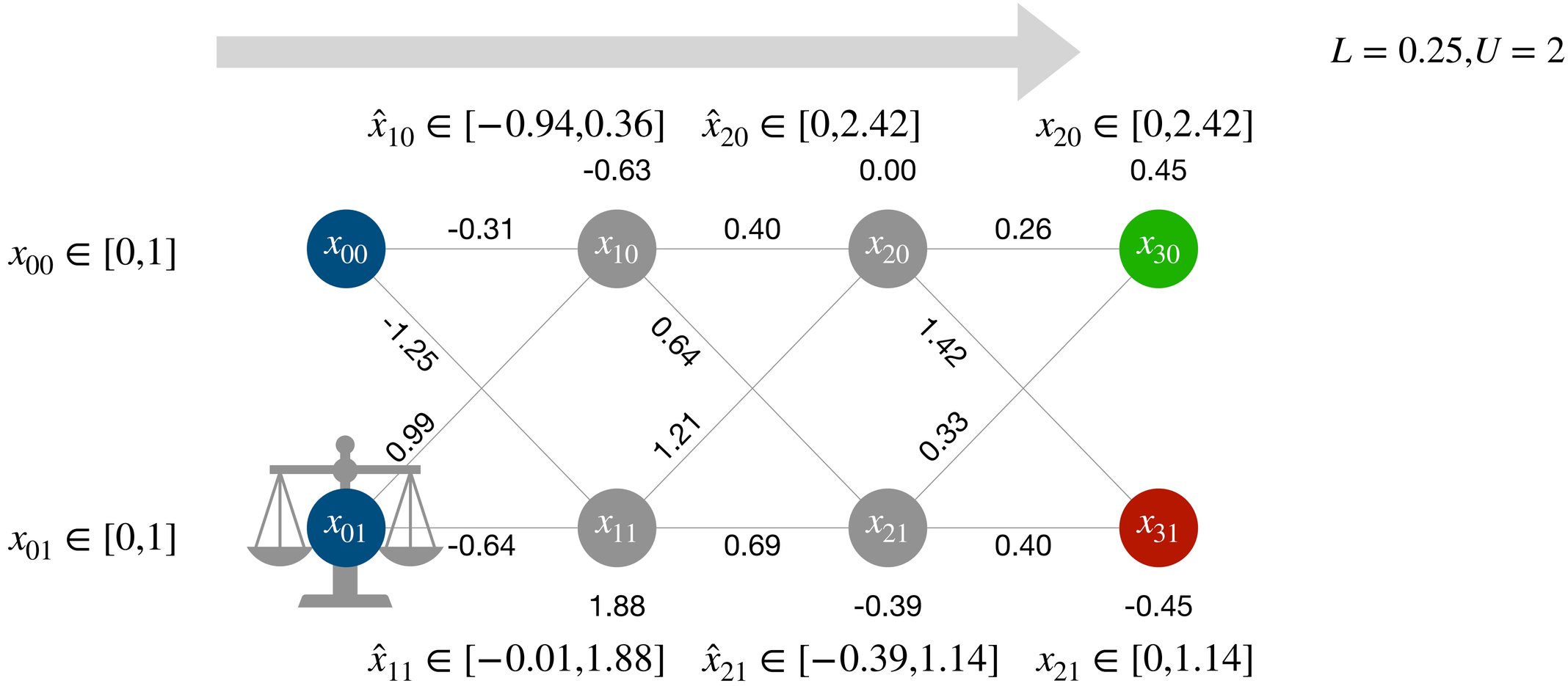
Forward Analysis



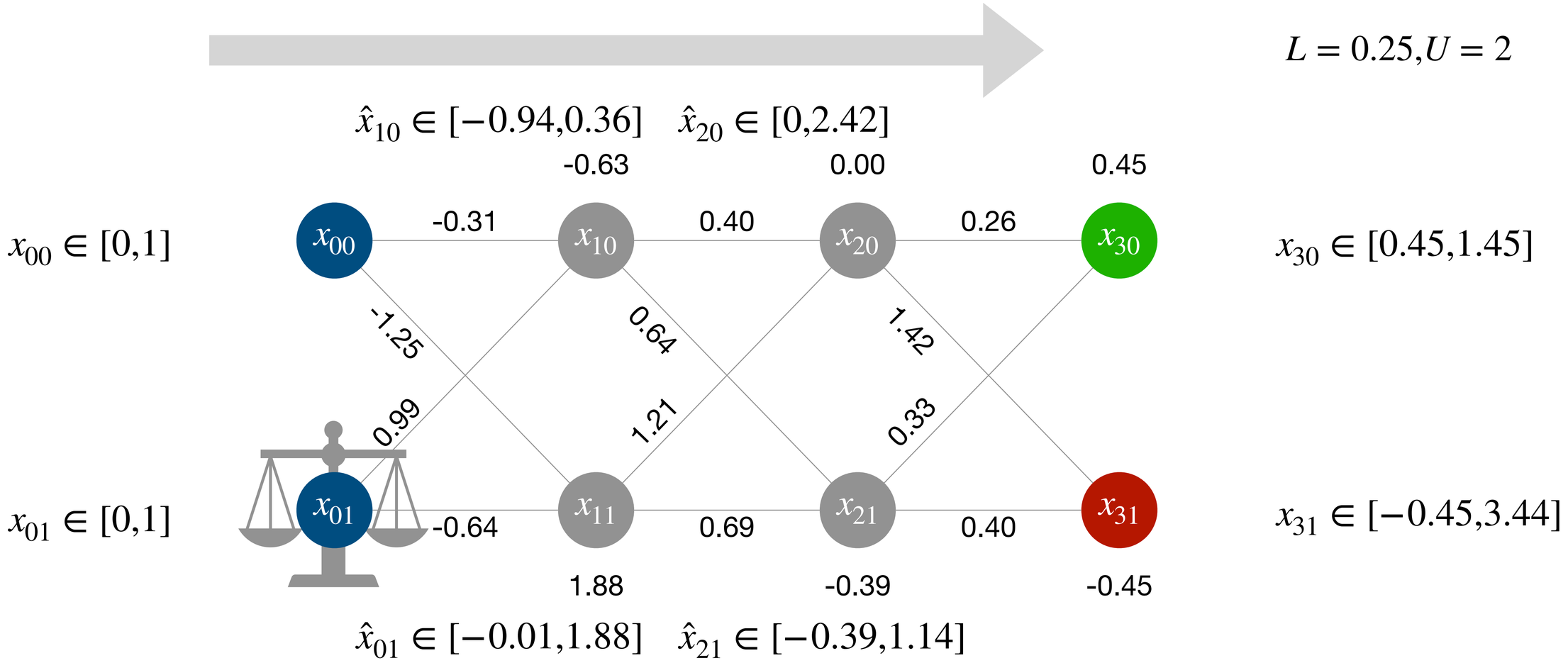
Forward Analysis



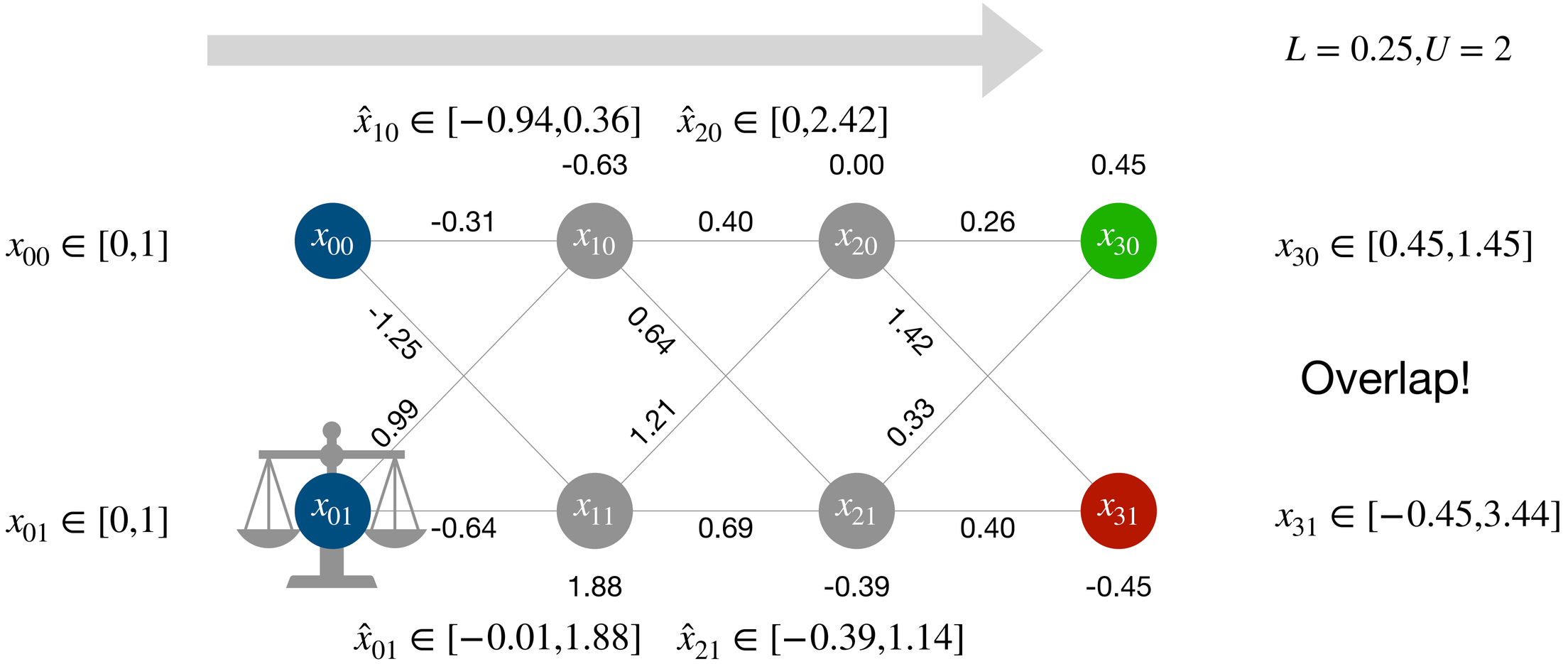
Forward Analysis



Forward Analysis

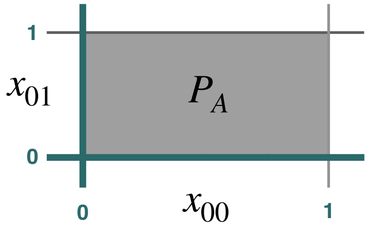


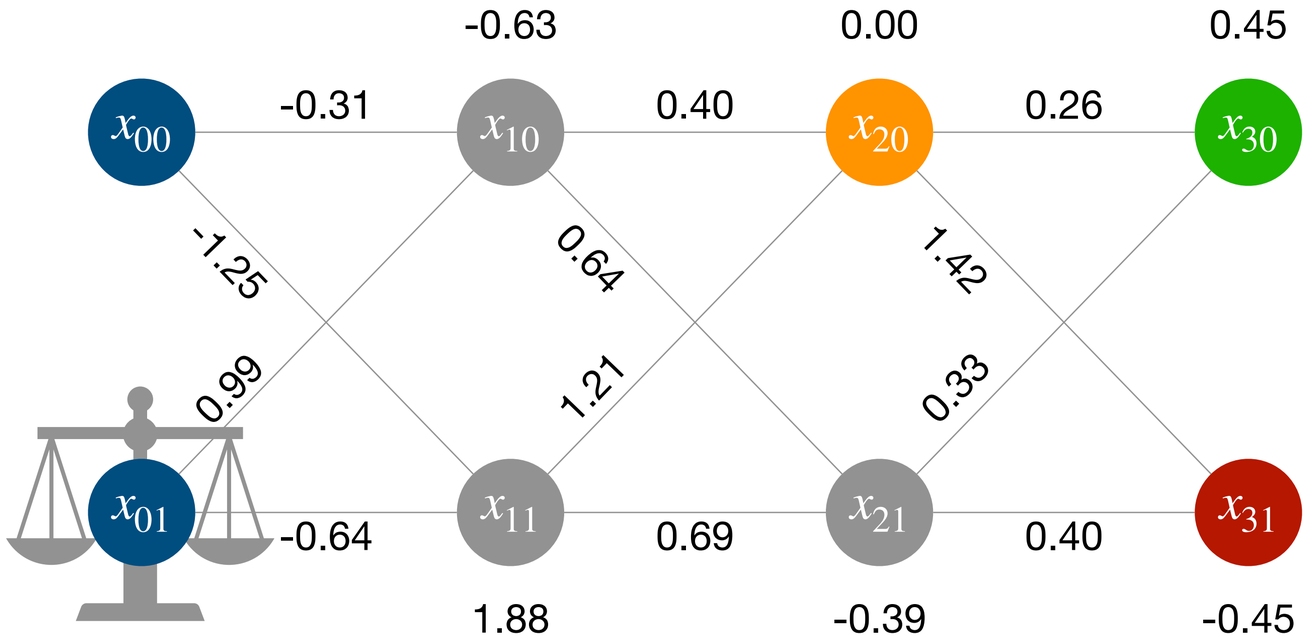
Forward Analysis



Forward Analysis

$L = 0.25, U = 2$

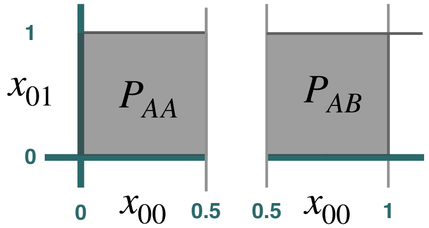


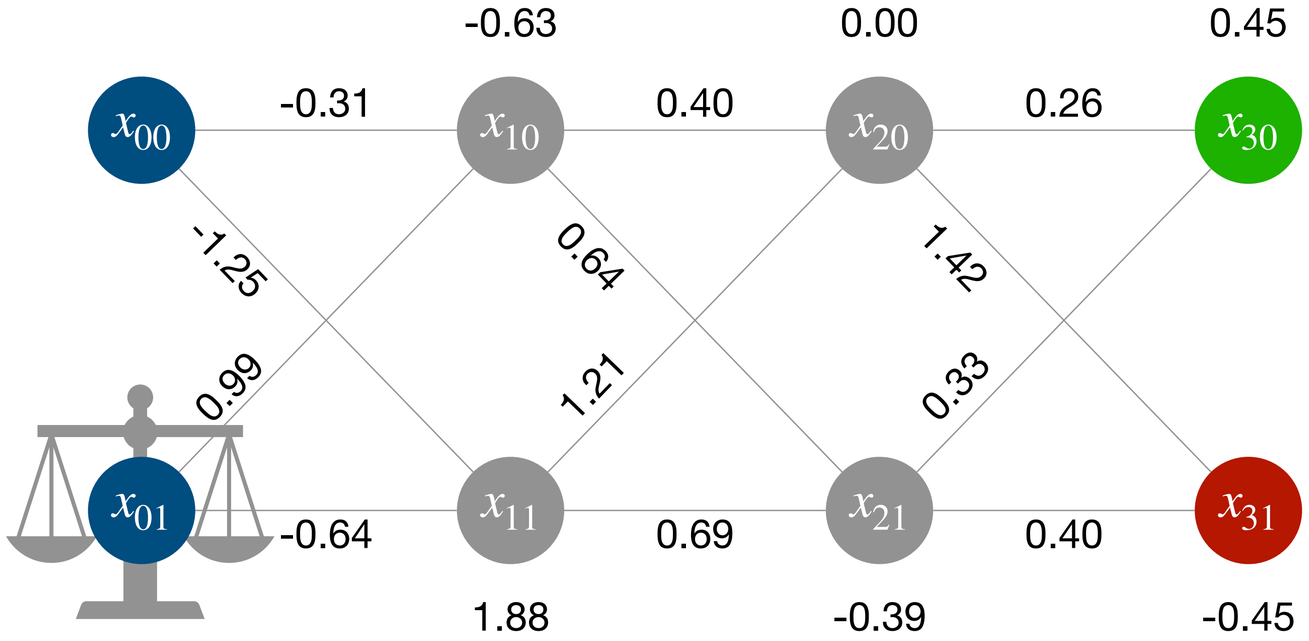


Activation Pattern

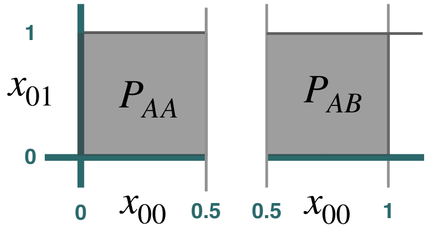
Forward Analysis

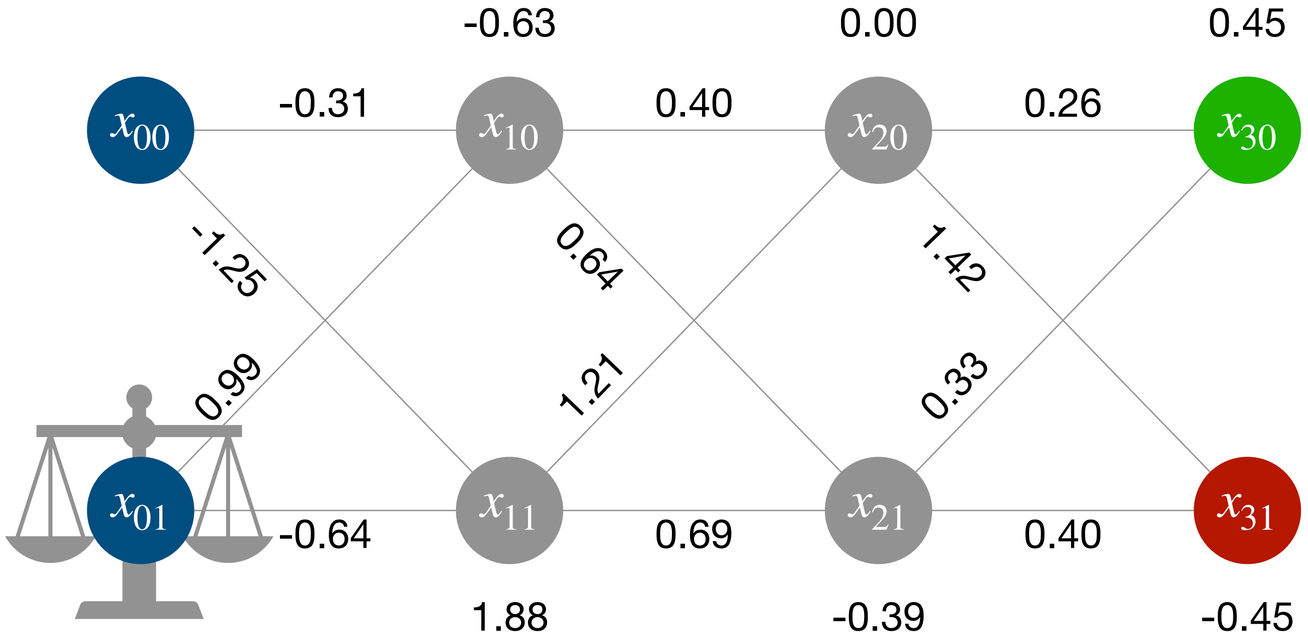
$L = 0.25, U = 2$



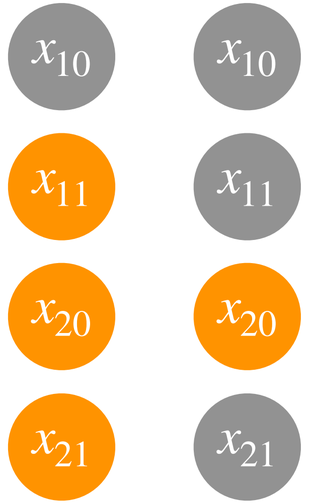


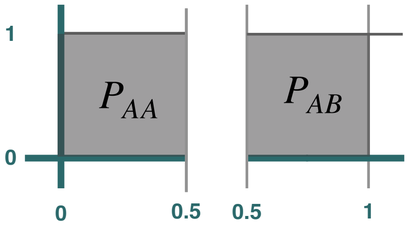
Forward Analysis





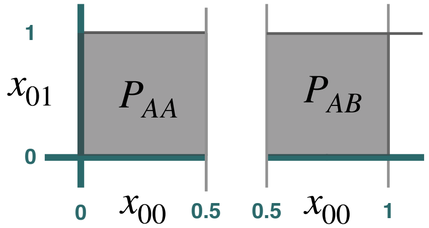
$L = 0.25, U = 2$

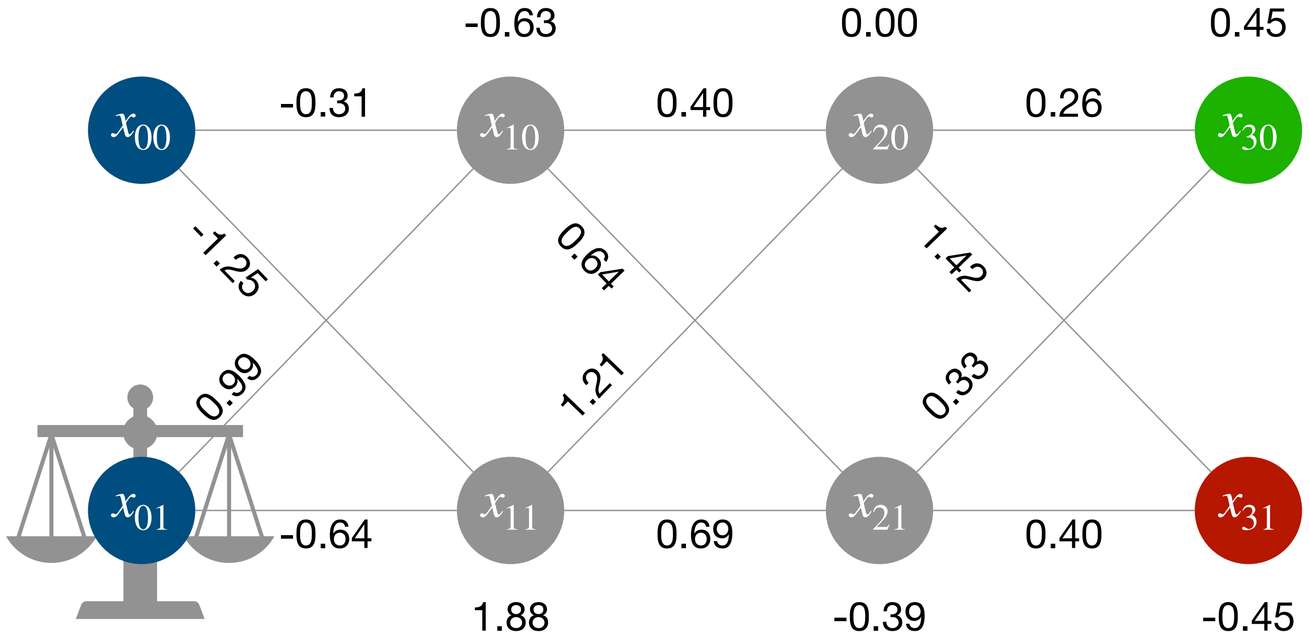


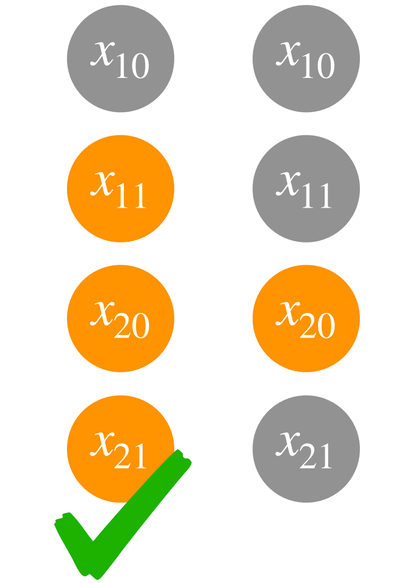


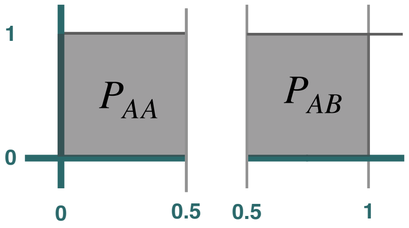
Forward Analysis

$L = 0.25, U = 2$



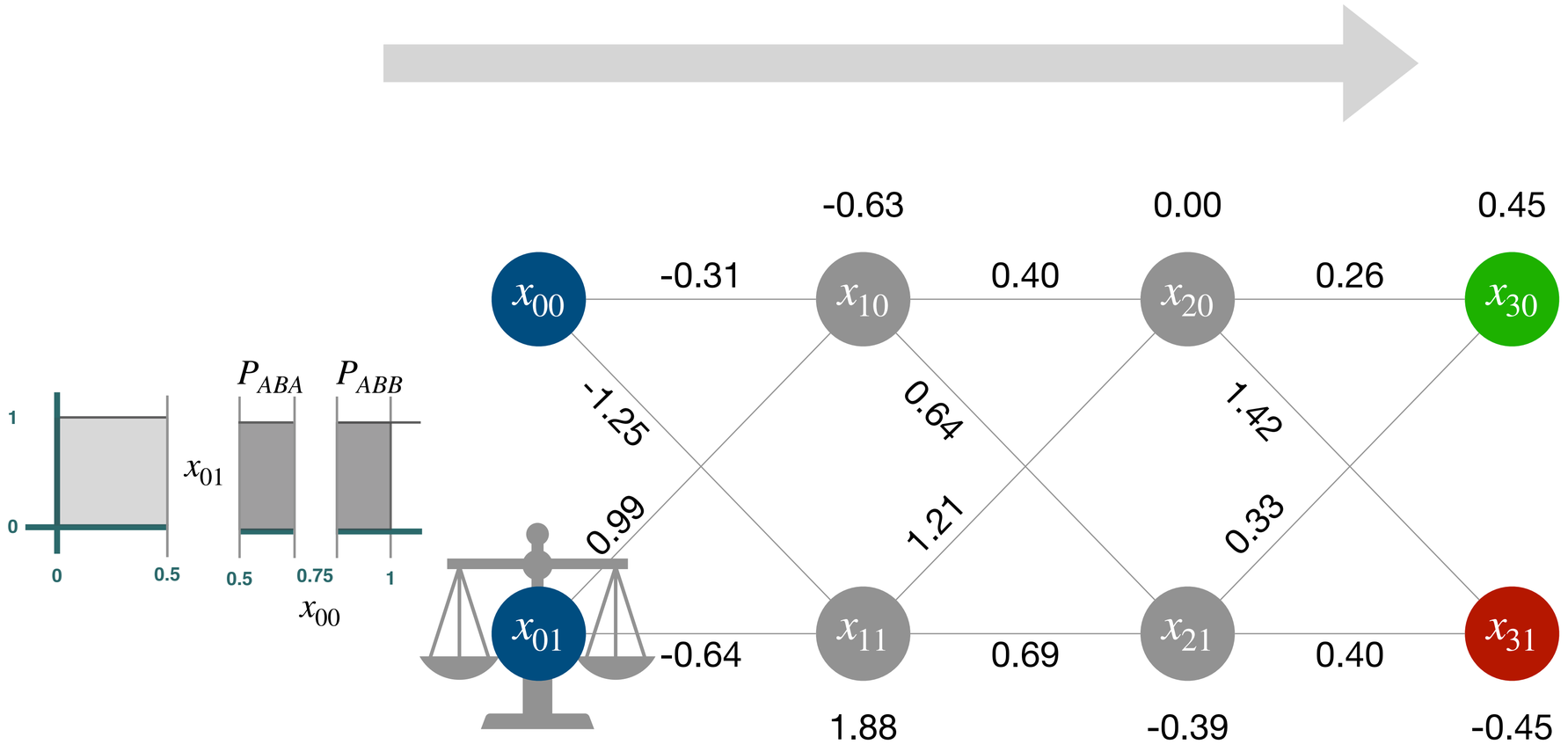




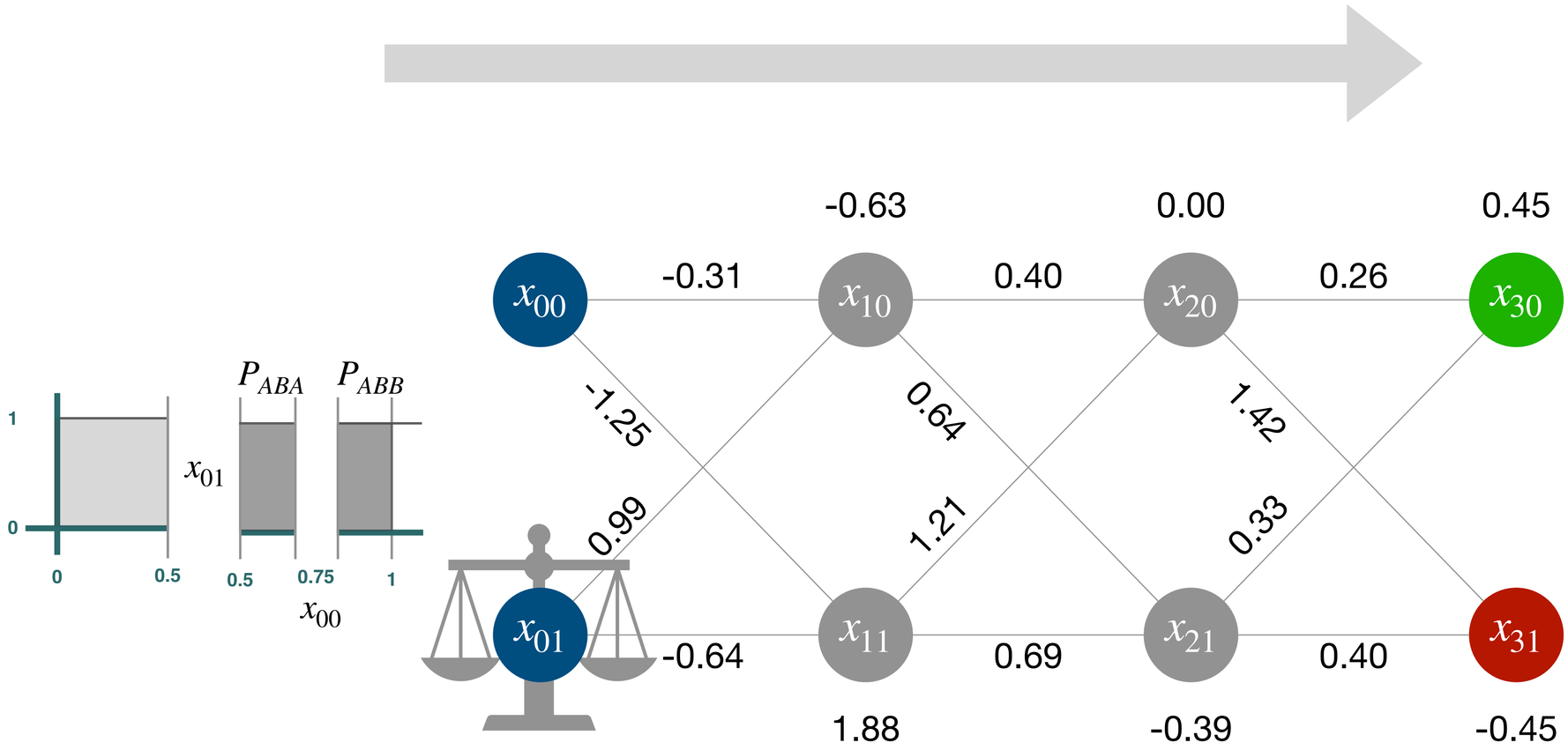


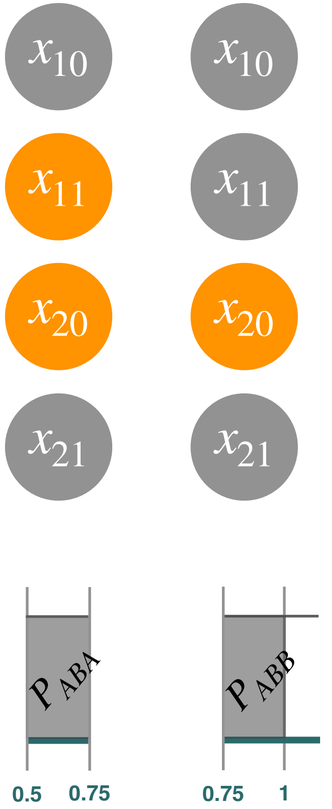
Forward Analysis

$L = 0.25, U = 2$



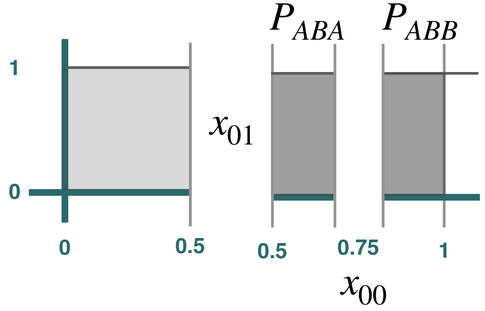
Forward Analysis

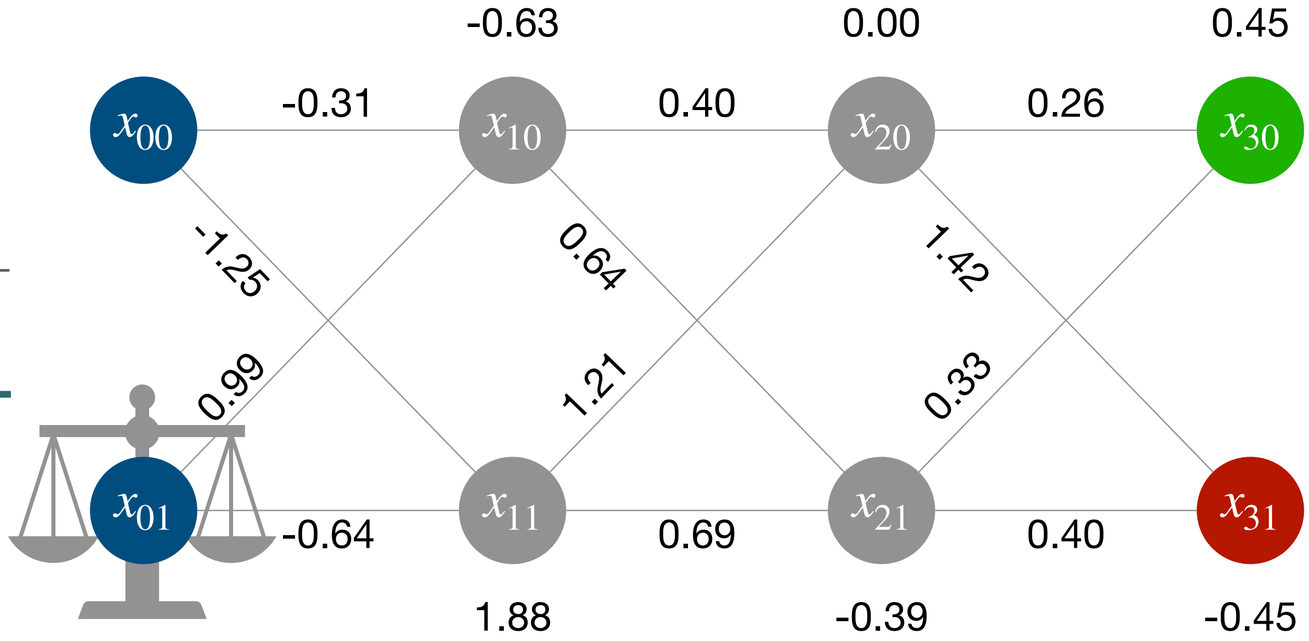




Forward Analysis

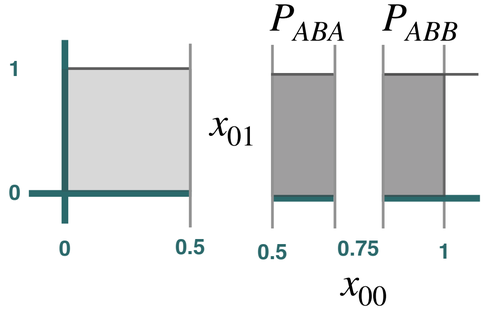
$L = 0.25, U = 2$

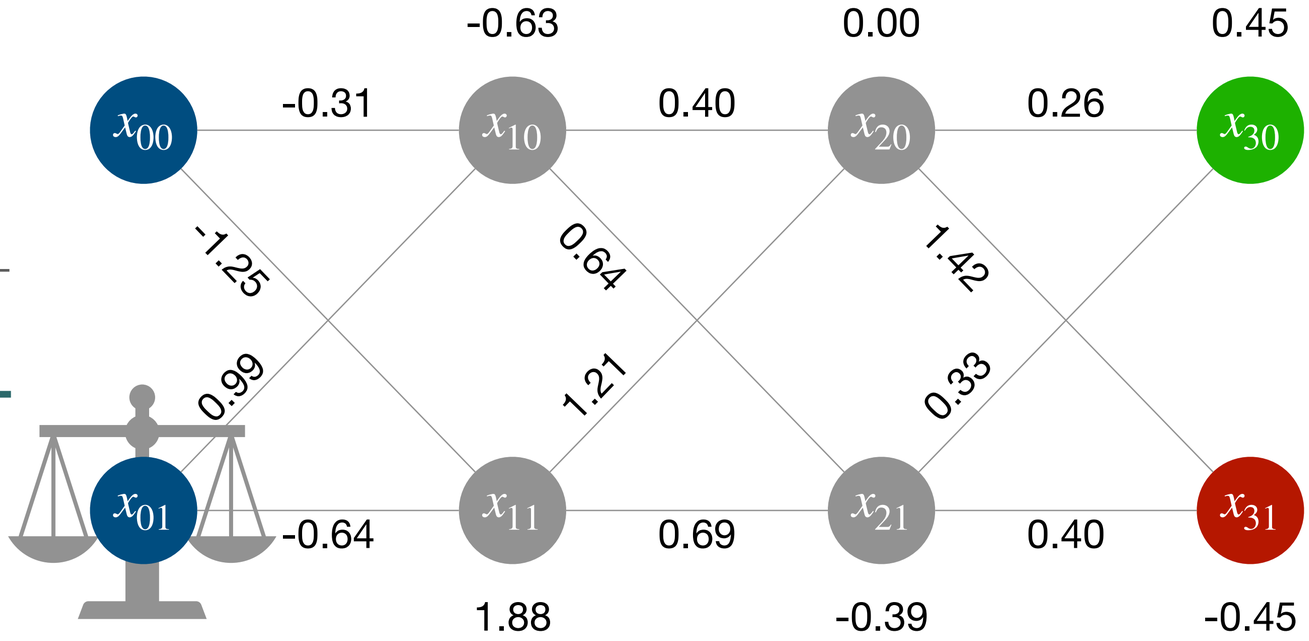


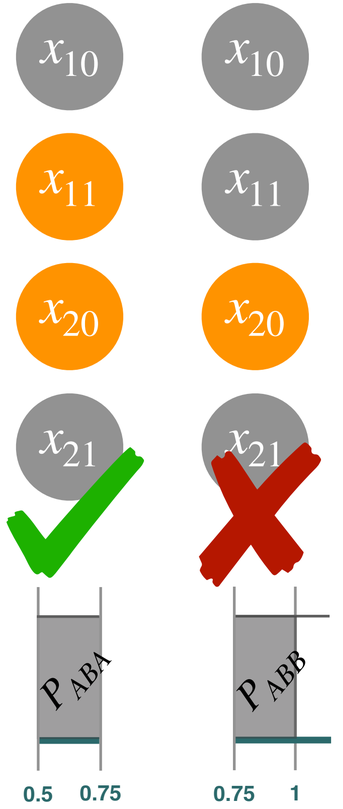


Forward Analysis

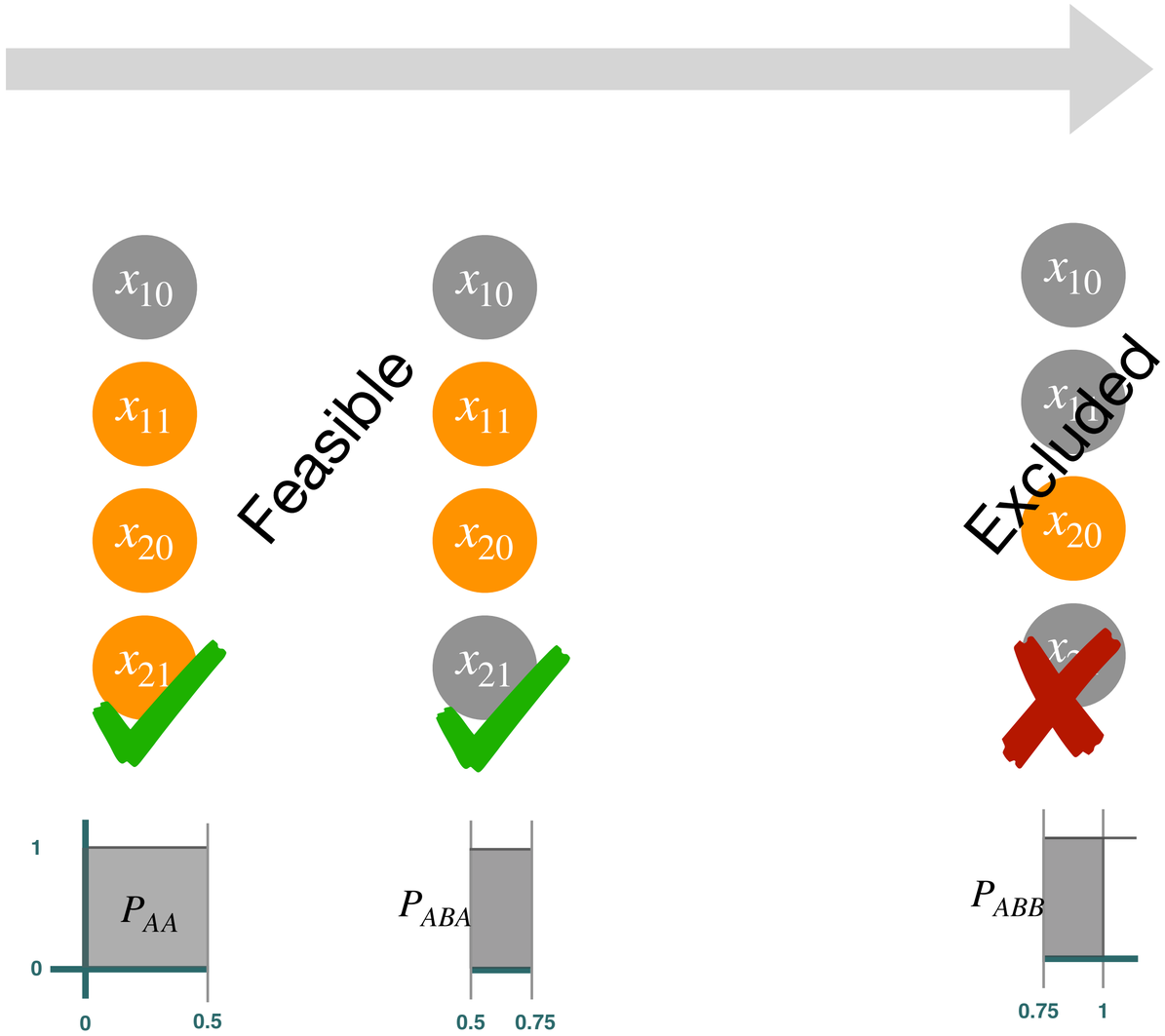
$L = 0.25, U = 2$



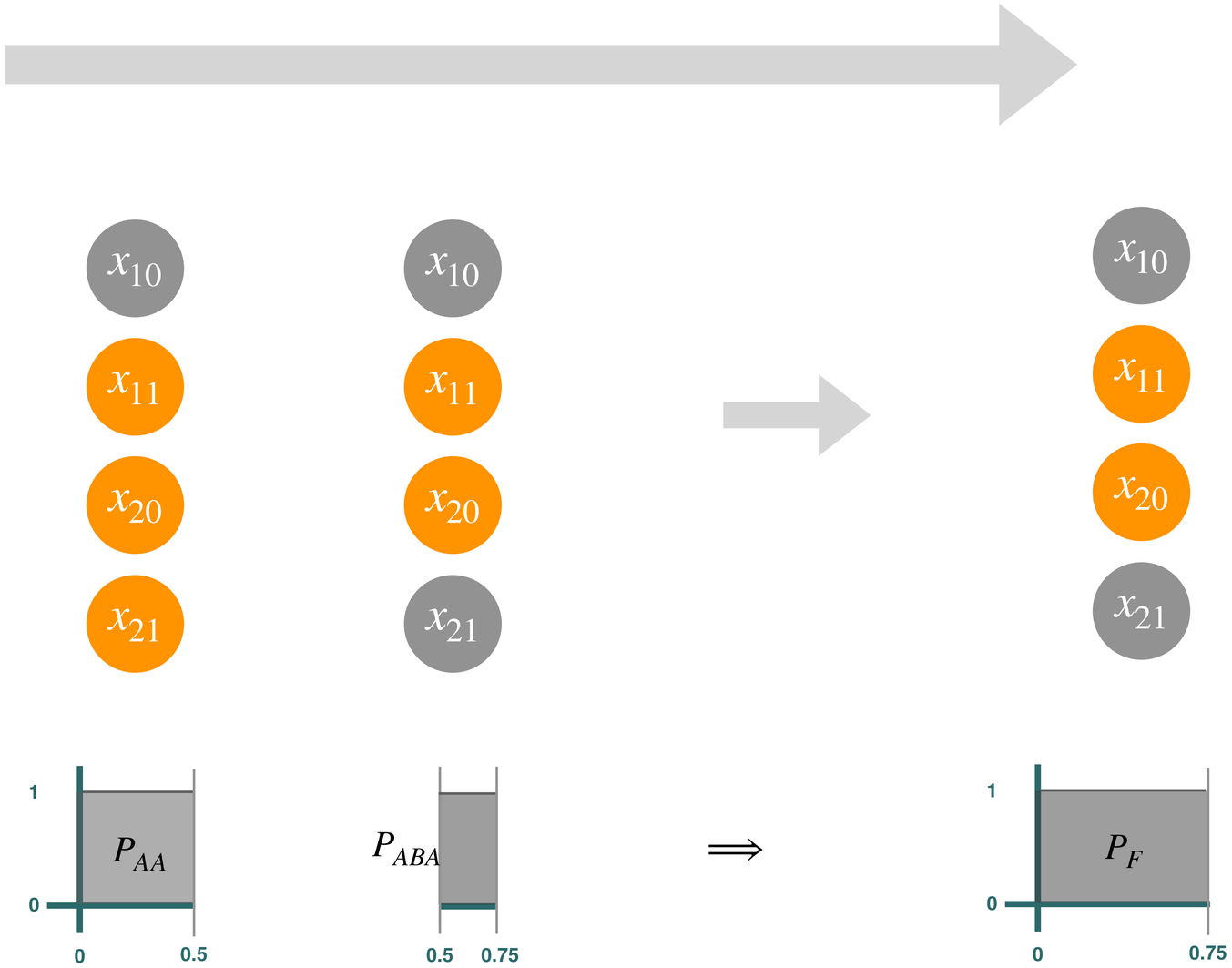




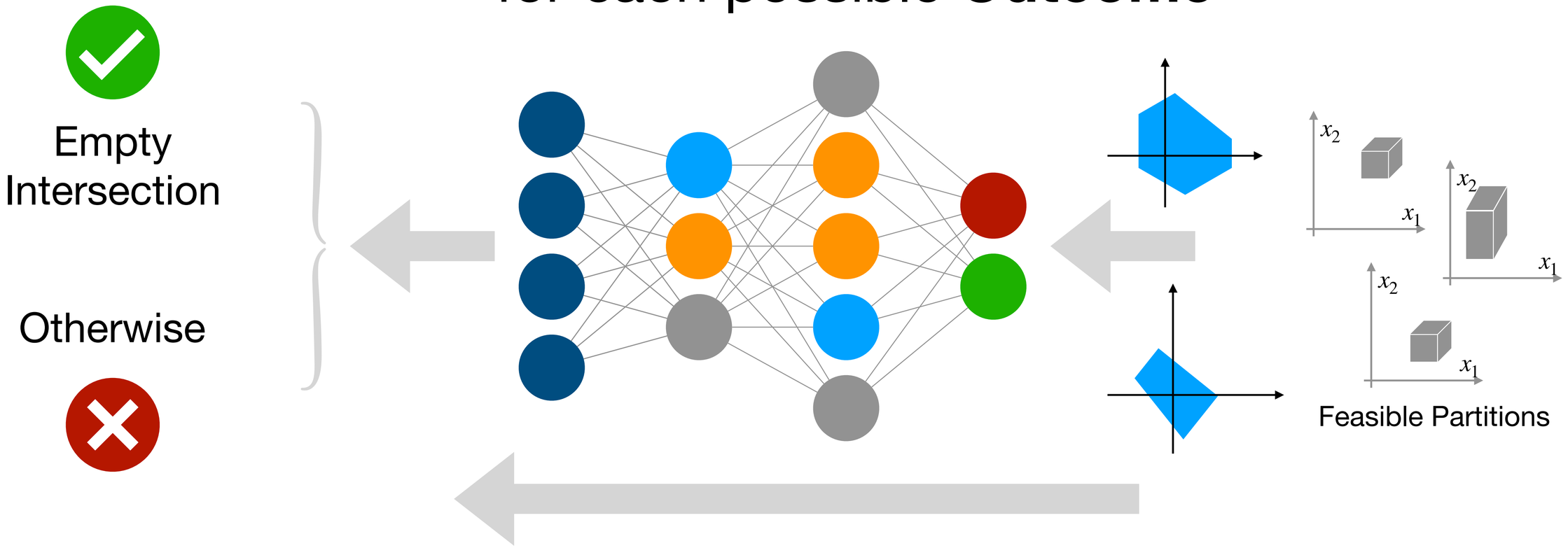
Forward Analysis



Forward Analysis



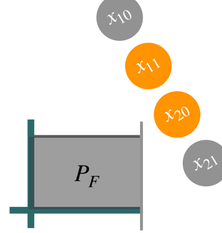
Proceed **Backwards**
for each **Feasible** partitions
for each possible **Outcome**

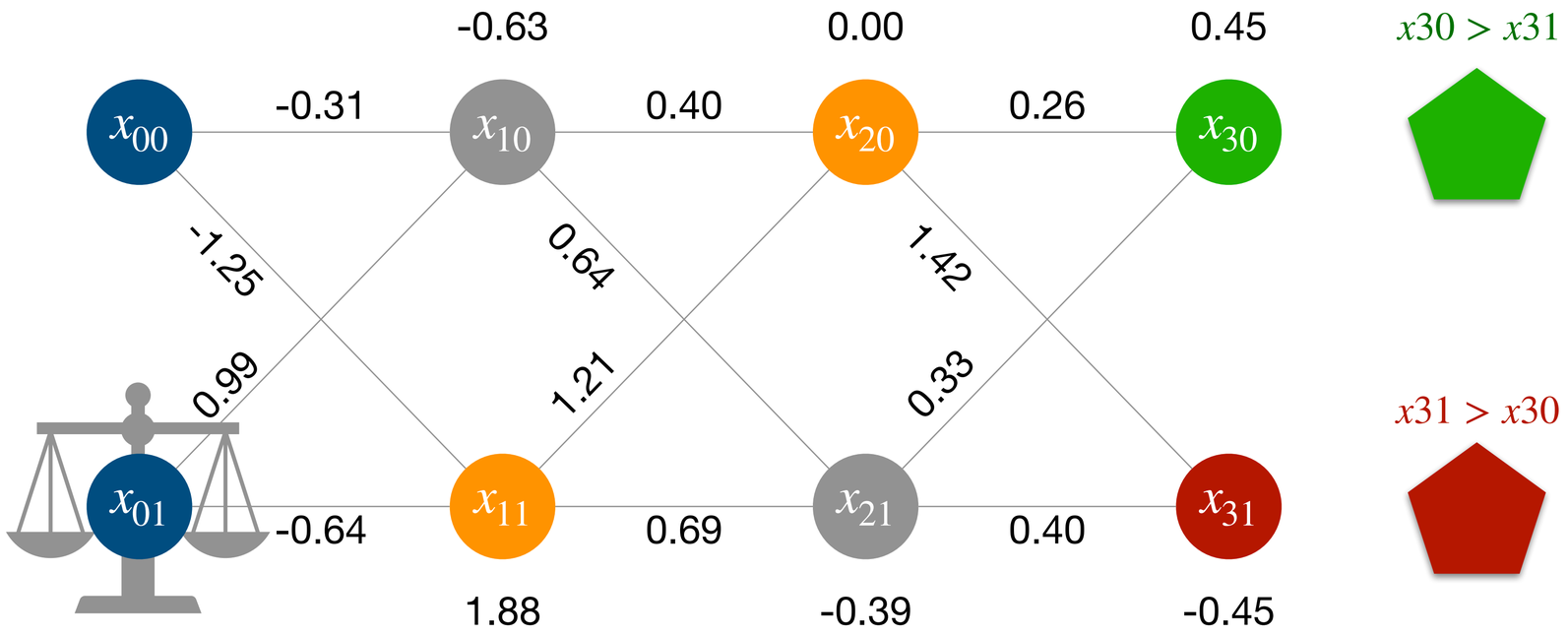


Exact **Backward Analysis**
using **Polyhedra**

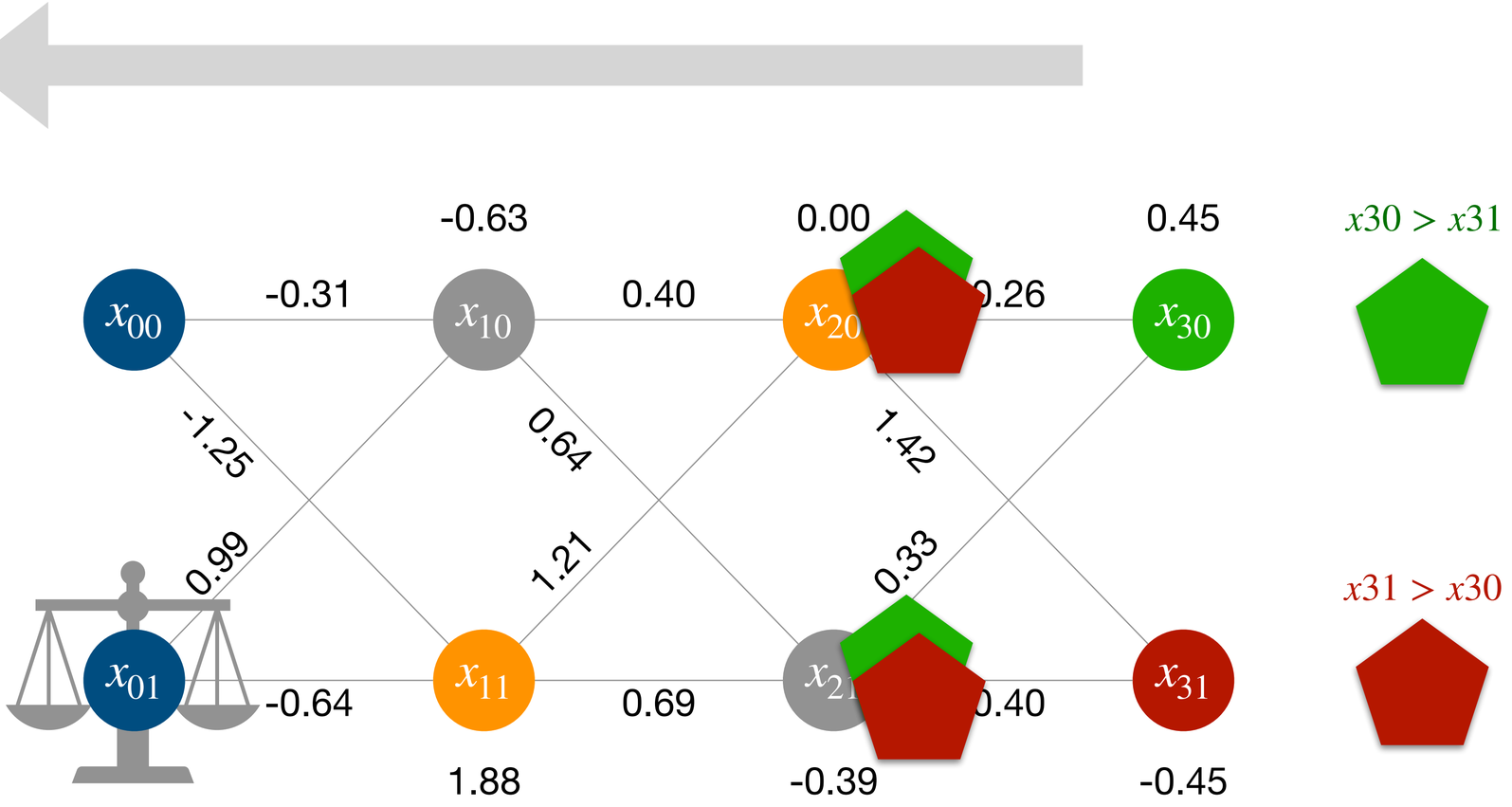
Backward Analysis

From the forward analysis

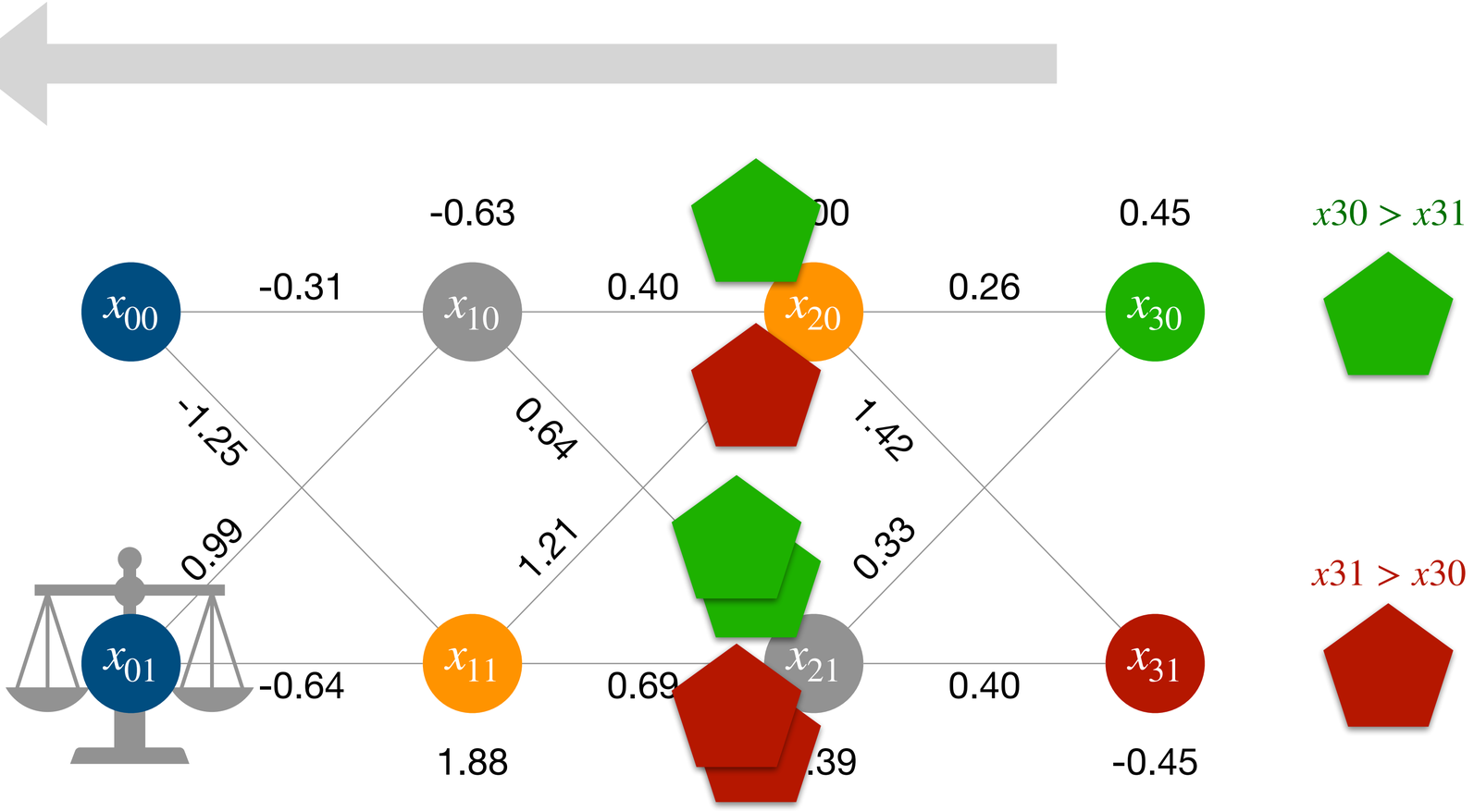




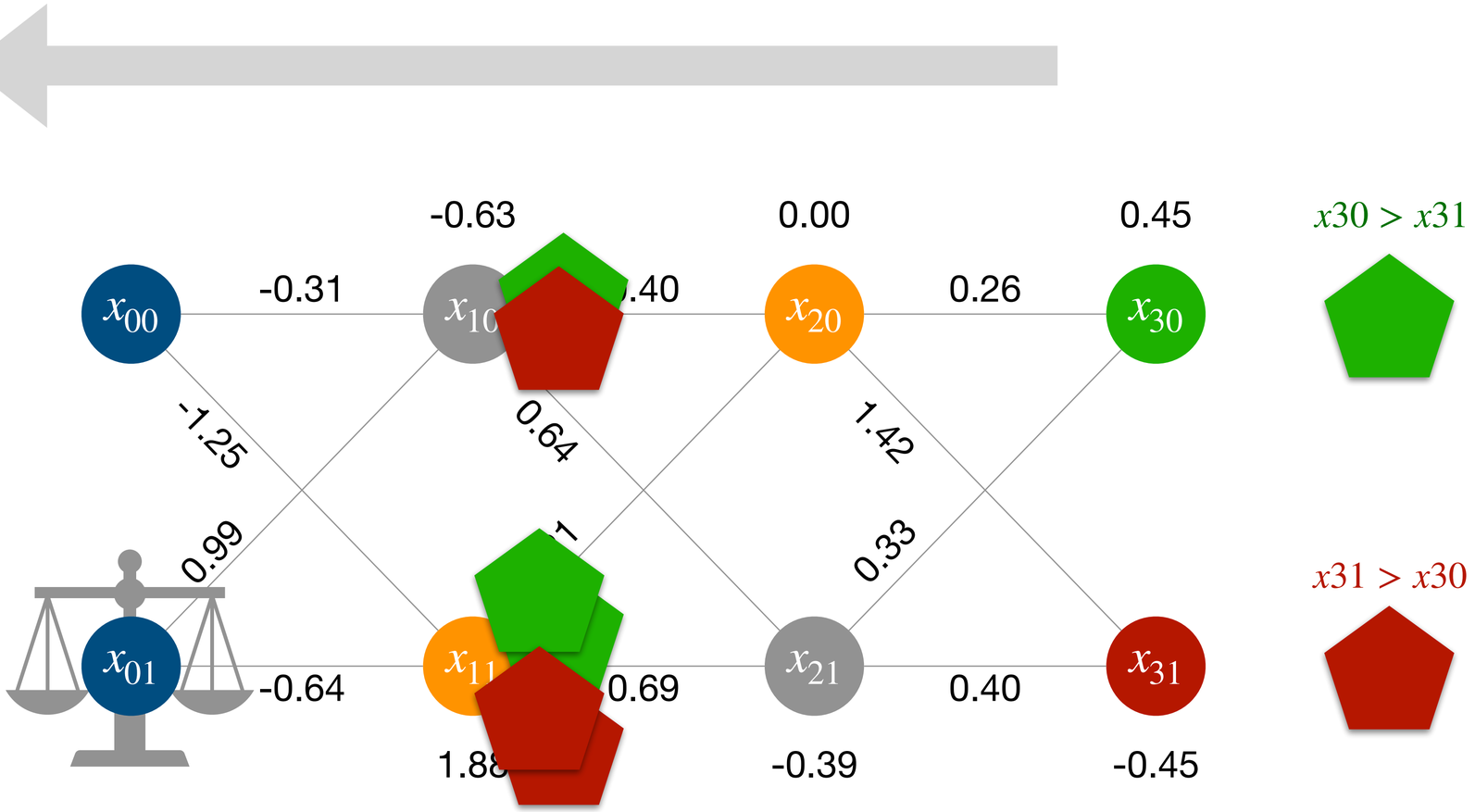
Backward Analysis



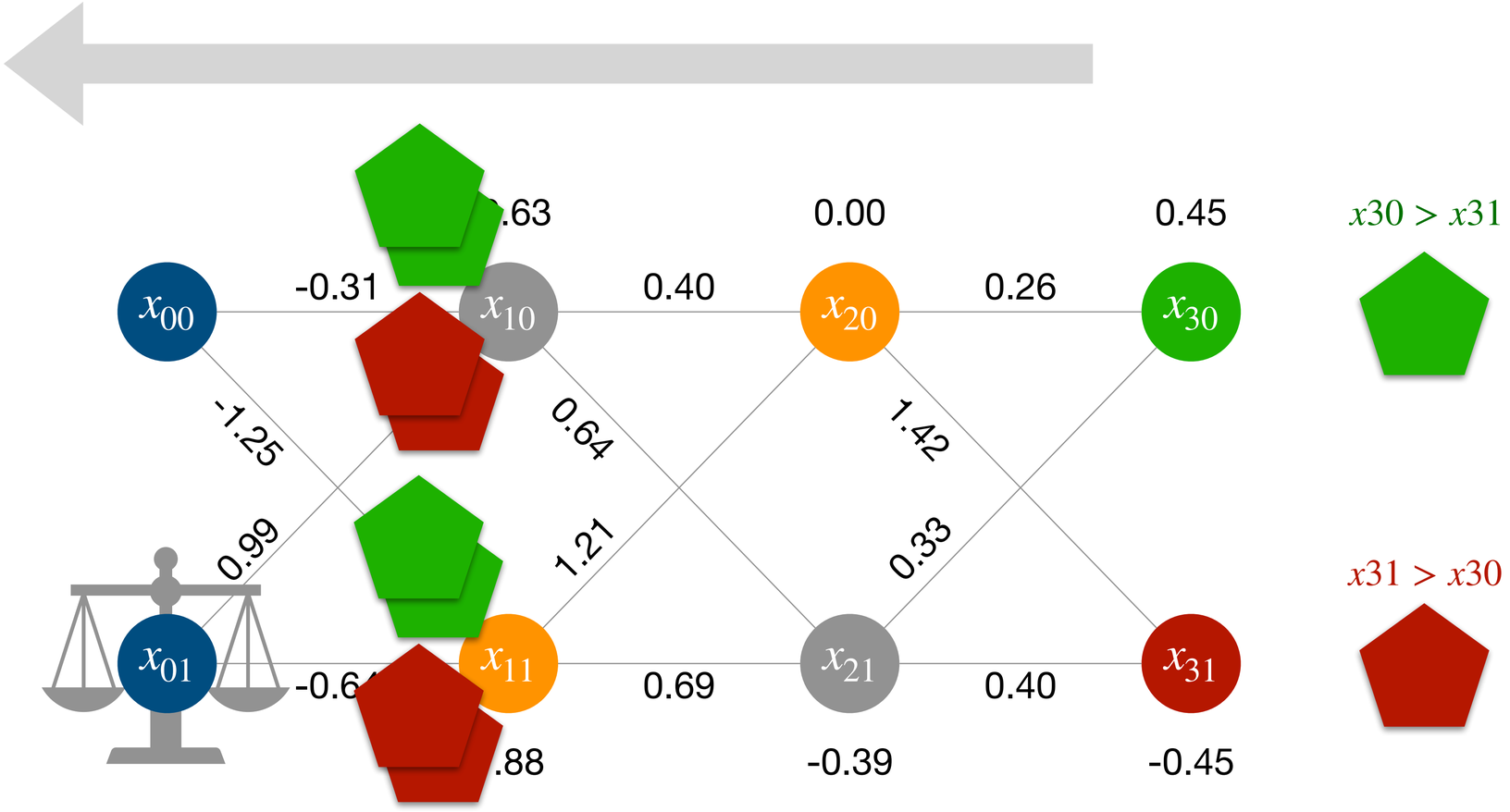
Backward Analysis



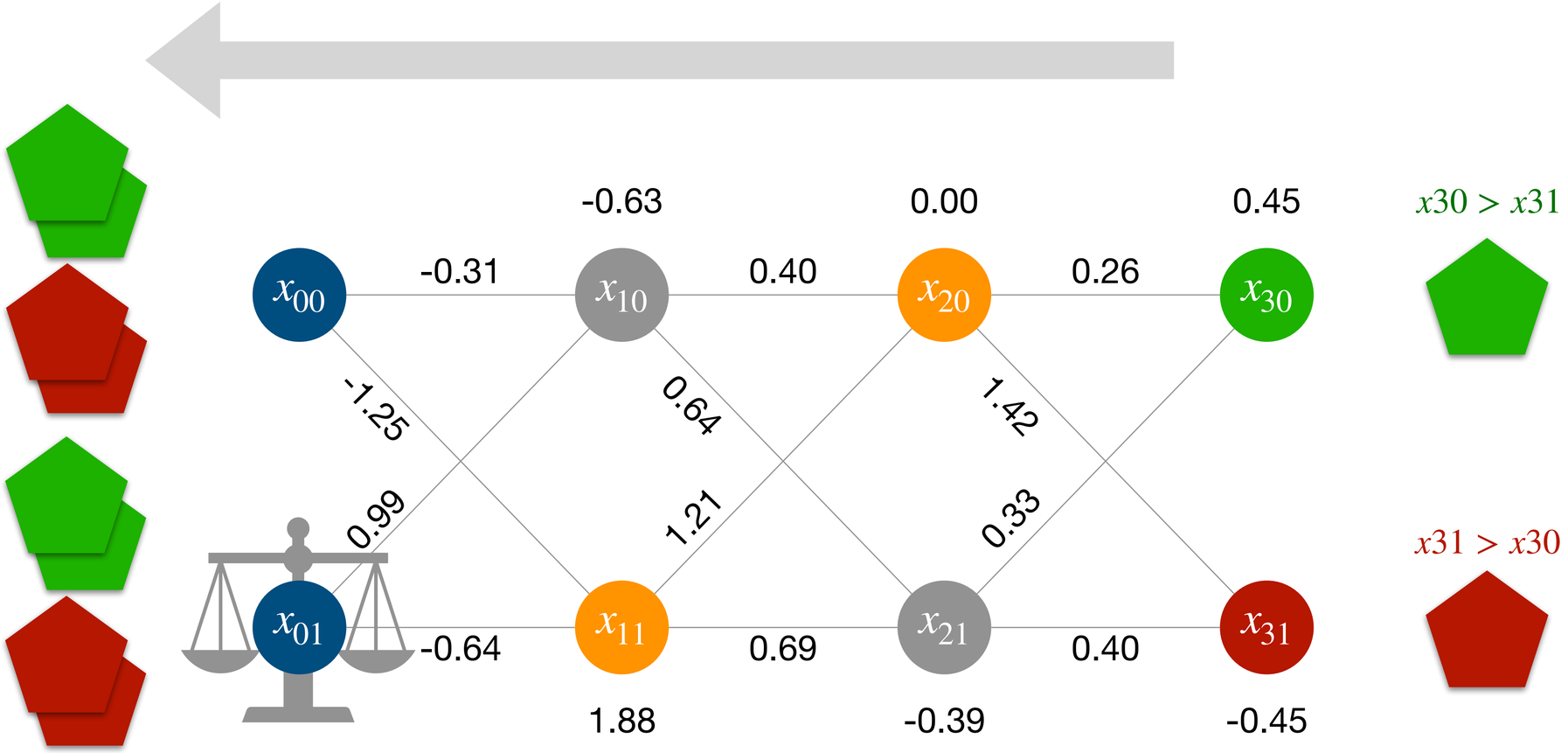
Backward Analysis



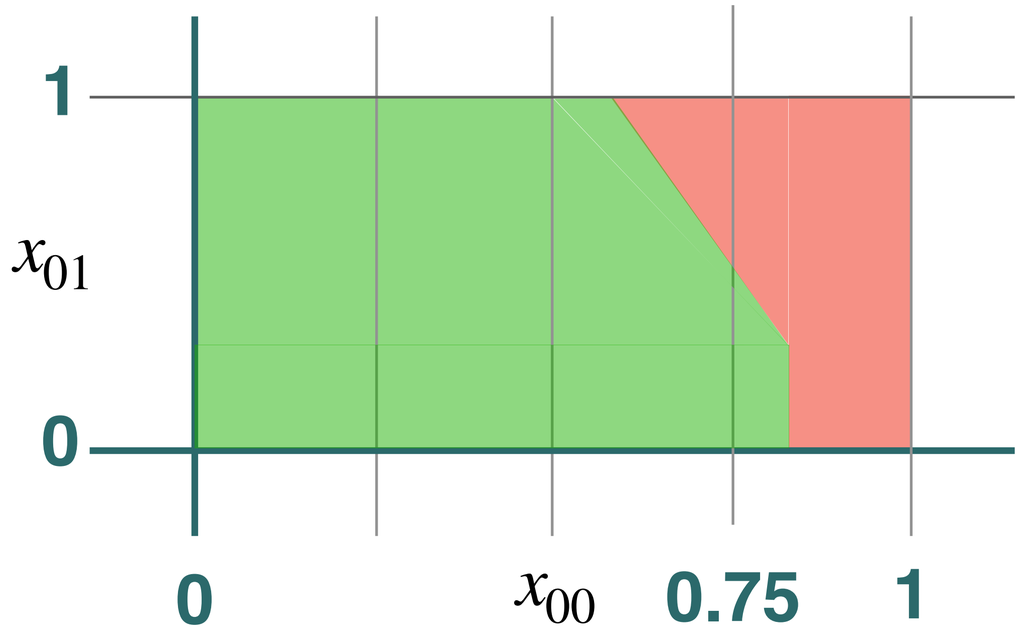
Backward Analysis



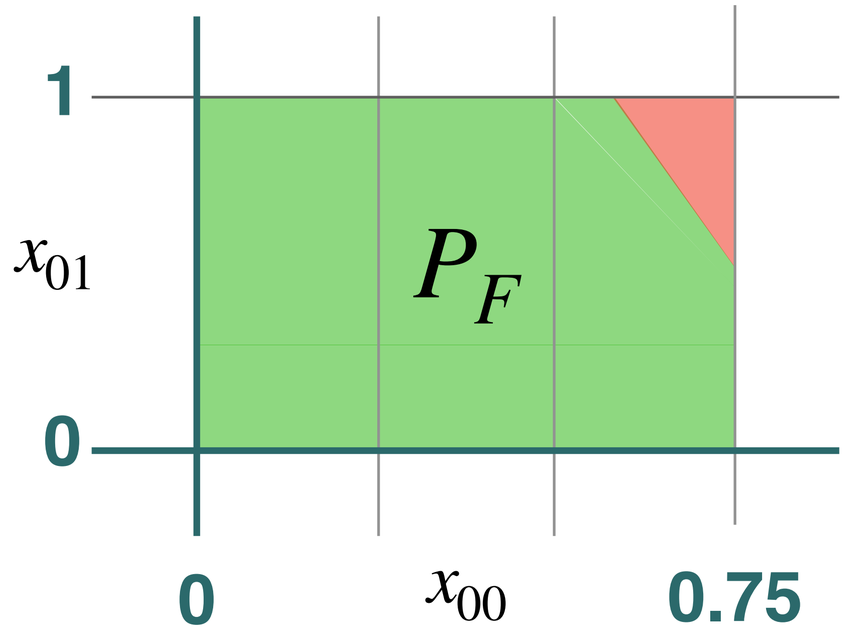
Backward Analysis



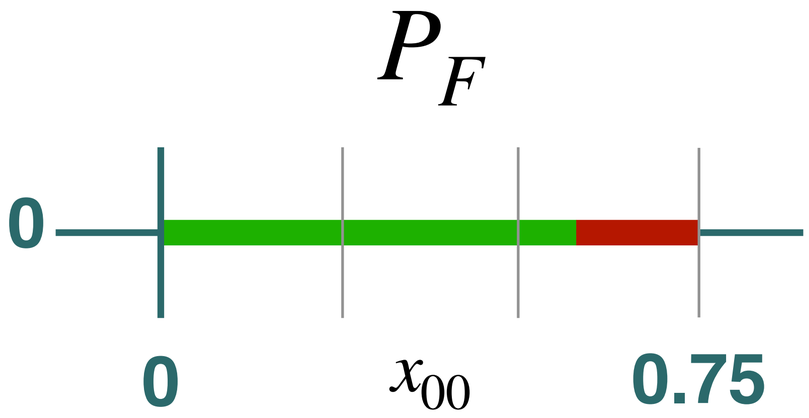
Analysis Result



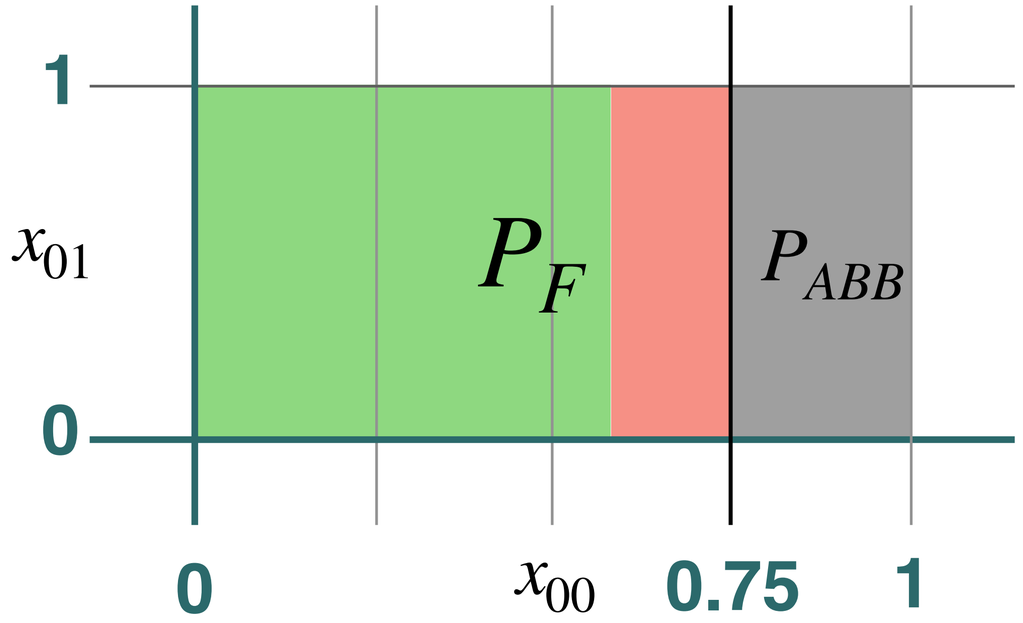
Analysis Result

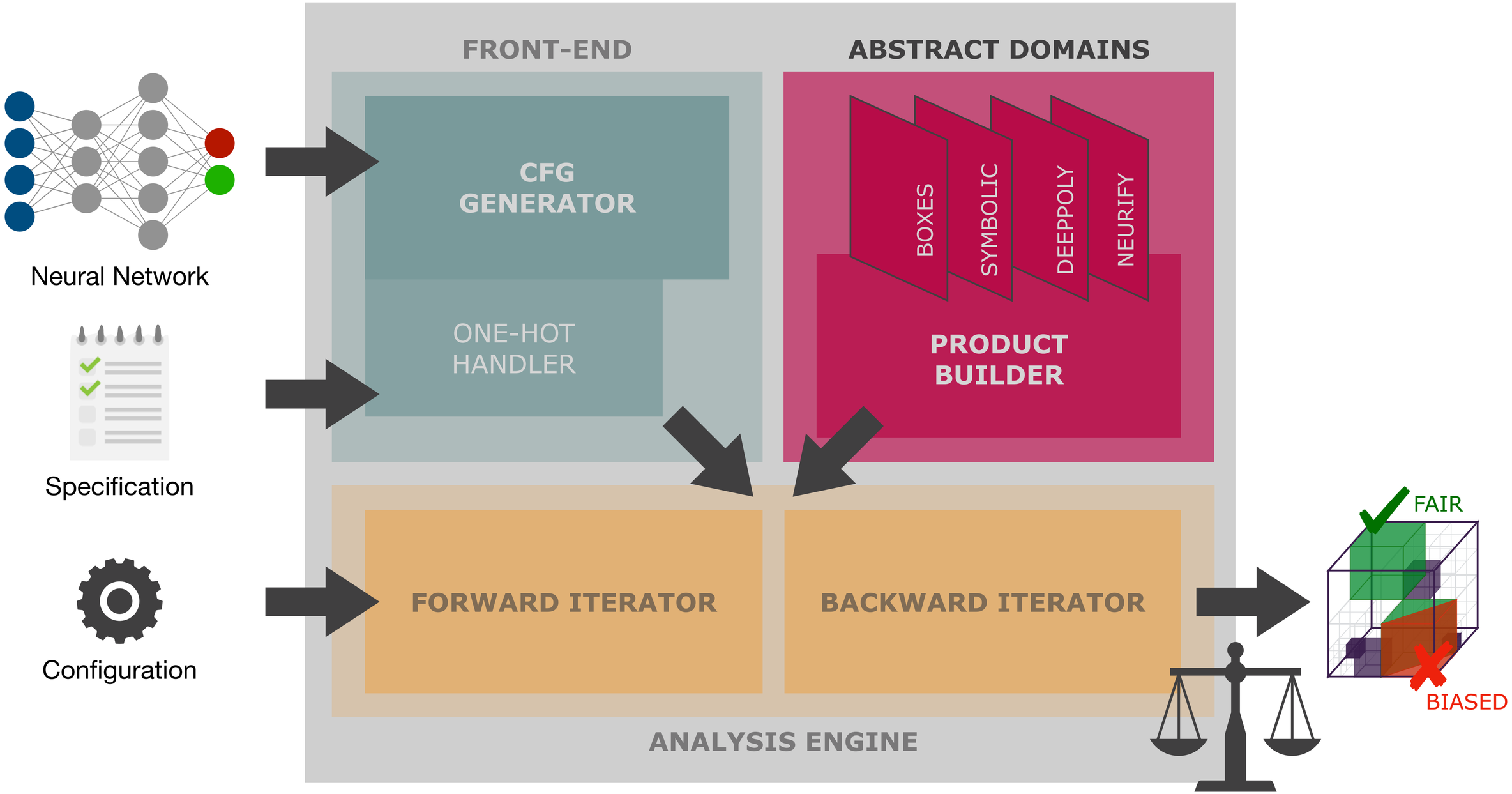


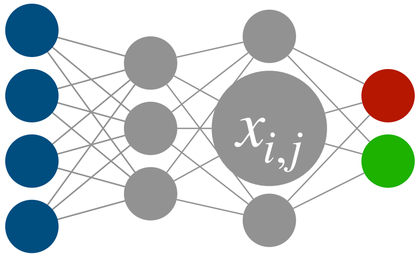
Analysis Result



Analysis Result




FRONT-END
ABSTRACT DOMAINS
CFG GENERATOR
ONE-HOT HANDLER
BOXES
SYMBOLIC
DEEPPOLY
NEURIFY
PRODUCT BUILDER
FORWARD ITERATOR
BACKWARD ITERATOR
ANALYSIS ENGINE
Neural Network
Specification
Configuration
FAIR
BIASED



Symbolic

Li et al. @ SAS 2019

$$[l, u]$$

$$\sum_k m_k \cdot x_k + q$$

DeepPoly

Singh et al. @ POPL 2019

$$[l, u]$$

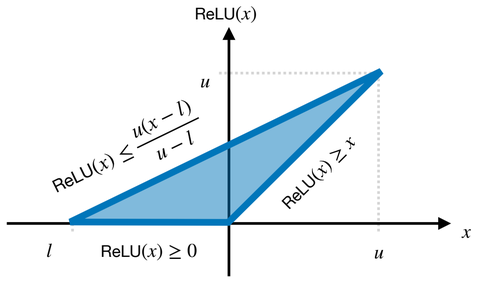
$$[\text{eq}_{\text{low}}, \text{eq}_{\text{up}}]$$

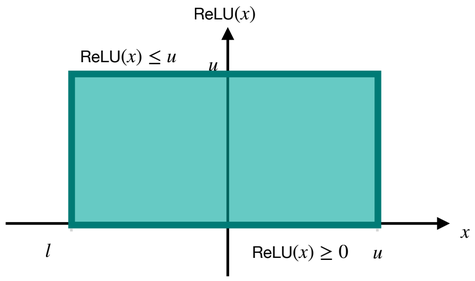
Neurify

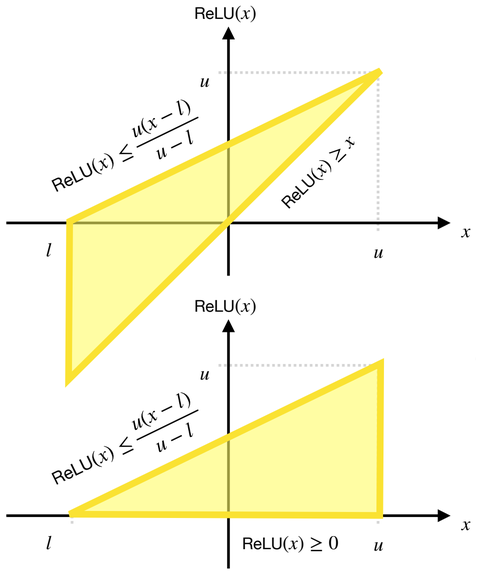
Wang et al. @ NeurIPS 2018

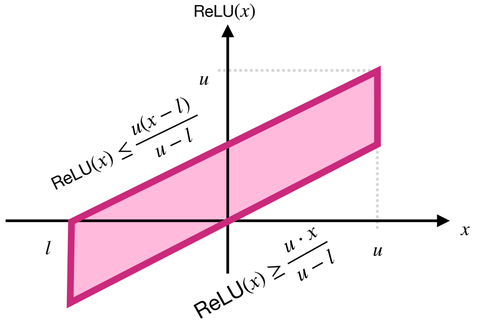
$$[l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$

$$[\text{eq}_{\text{low}}, \text{eq}_{\text{up}}]$$









Reduced Product

Mazzucato, Urban @ SAS 2021

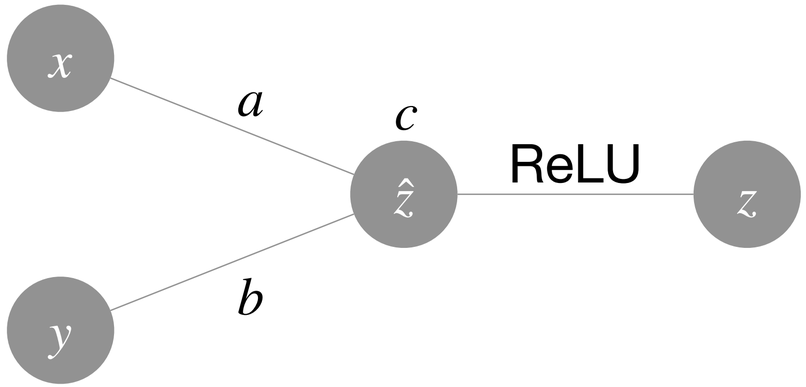
Symbolic

Li et al. @ SAS 2019

$$\hat{z} = a \cdot x + b \cdot y + c$$

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$



$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

Symbolic

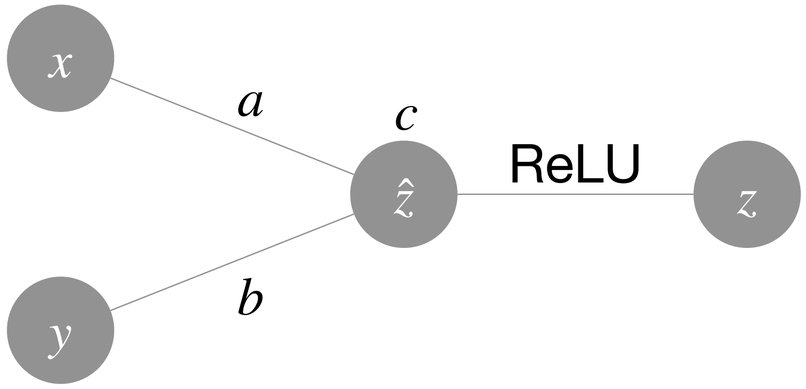
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

Symbolic

Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

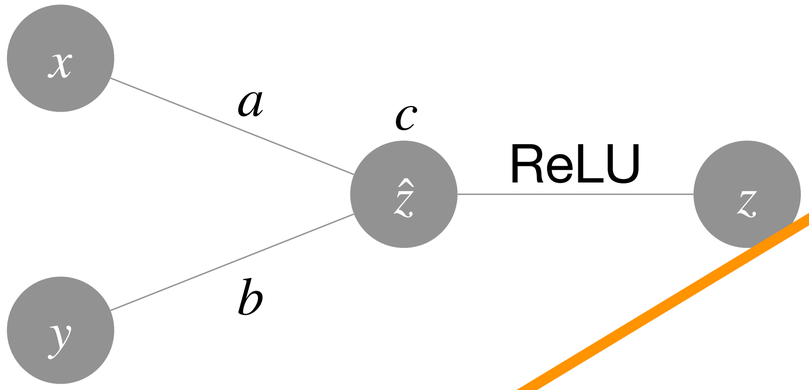
$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$



Symbolic

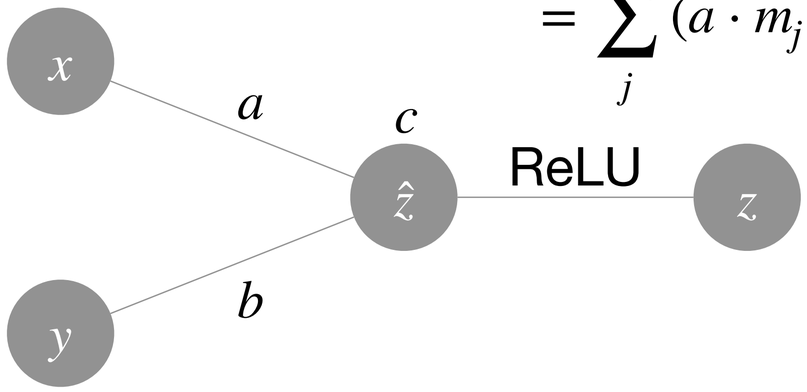
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

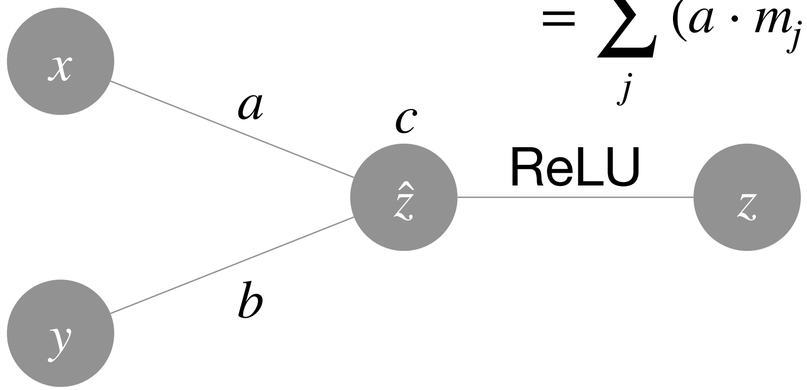
$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c$$

Symbolic

Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$



$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c \; = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

Symbolic

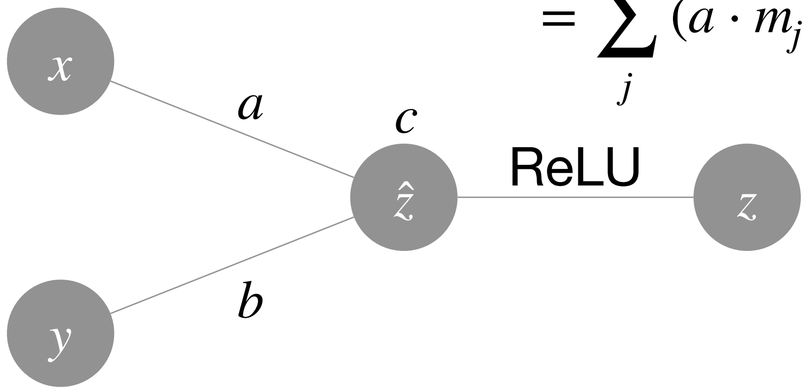
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c \; = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

Dependency Problem

$$x \in [0,2], x \leftarrow m + n$$

$$y \in [0,2], y \leftarrow m + n$$

$$\rightarrow \quad z = x - y$$

Symbolic

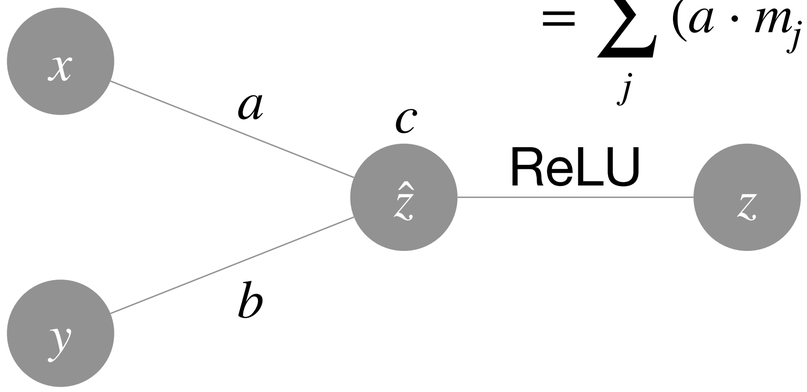
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c \; = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

Dependency Problem

$$x \in [0,2], x \leftarrow m + n$$

$$y \in [0,2], y \leftarrow m + n$$

$$\rightarrow \; z = x - y$$

Boxes

$$\rightarrow \; z \in [-2,2]$$

Symbolic

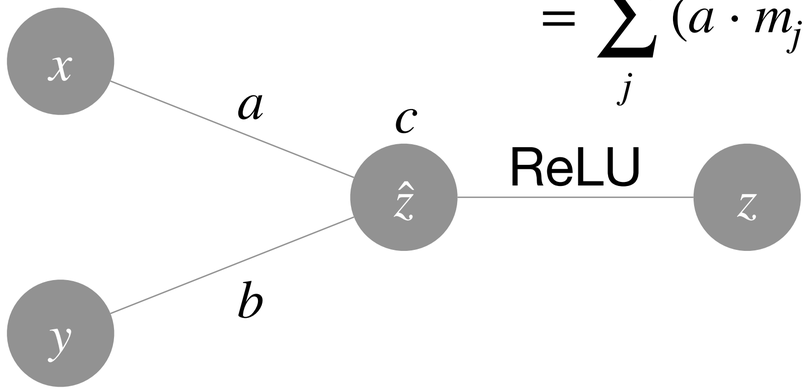
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c \; = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

Dependency Problem

$$x \in [0,2], x \leftarrow m + n$$

$$y \in [0,2], y \leftarrow m + n$$

$$z = x - y$$

Symbolic

$$z = (m + n) - (m + n)$$
$$= (1 - 1) \cdot m + (1 - 1) \cdot n$$
$$= 0 \cdot m + 0 \cdot n = 0 \in [0,0]$$

Boxes

$$z \in [-2,2]$$

Symbolic

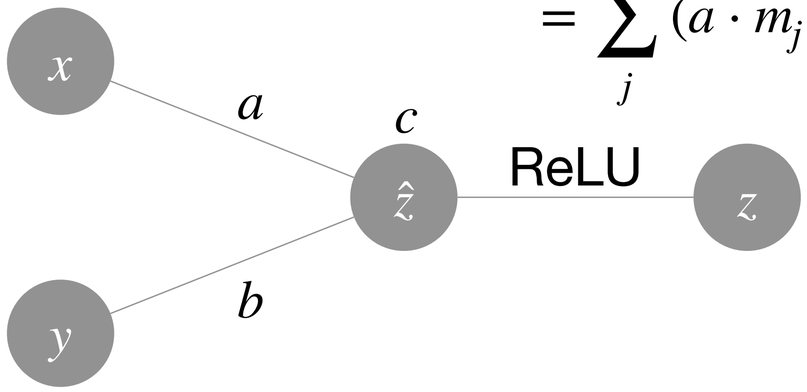
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c$$

$$= a \cdot (\sum_j m_j \cdot x_j + p_j) + b \cdot (\sum_j n_j \cdot x_j + q_j) + c$$

$$= \sum_j (a \cdot m_j + b \cdot n_j) \cdot x_j + a \cdot p_j + b \cdot q_j + c = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

Dependency Problem

$$x \in [0,2], x \leftarrow m + n$$

$$y \in [0,2], y \leftarrow m + n$$

$$z = x - y$$

Symbolic

$$z = (m + n) - (m + n)$$
$$= (1 - 1) \cdot m + (1 - 1) \cdot n$$
$$= 0 \cdot m + 0 \cdot n = 0 \in [0,0]$$

$$\cap$$

Boxes

$$z \in [-2,2]$$

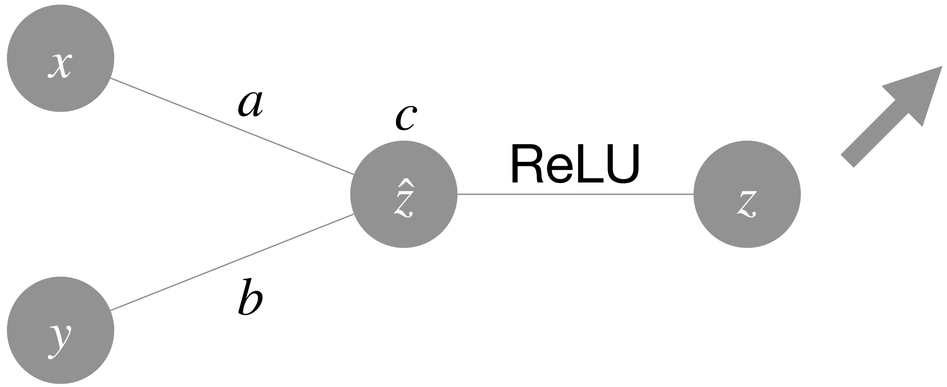
Symbolic

Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$\hat{z} = a \cdot x + b \cdot y + c = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$



$$u_{\hat{z}} \leq 0 \implies z \leftarrow 0$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

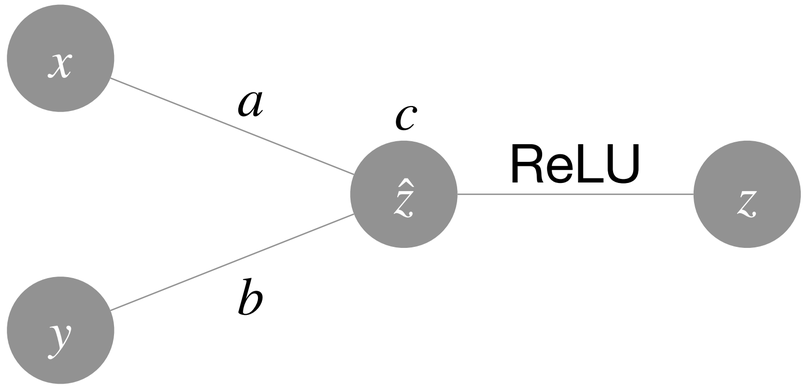
Symbolic

Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$\hat{z} = a \cdot x + b \cdot y + c = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$



$$u_{\hat{z}} \leq 0 \implies z \leftarrow 0$$

$$l_{\hat{z}} \geq 0 \implies z \leftarrow a \cdot x + b \cdot y + c$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

Symbolic

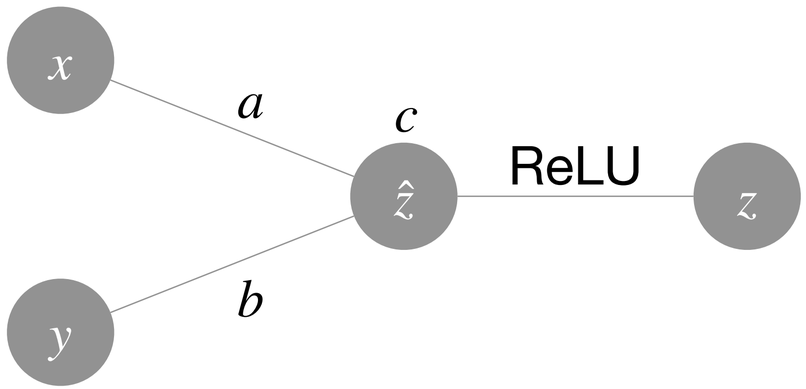
Li et al. @ SAS 2019

$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$



$$\hat{z} = a \cdot x + b \cdot y + c = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

$$u_{\hat{z}} \leq 0 \implies z \leftarrow 0$$

$$l_{\hat{z}} \geq 0 \implies z \leftarrow a \cdot x + b \cdot y + c$$

Symbolic

Li et al. @ SAS 2019

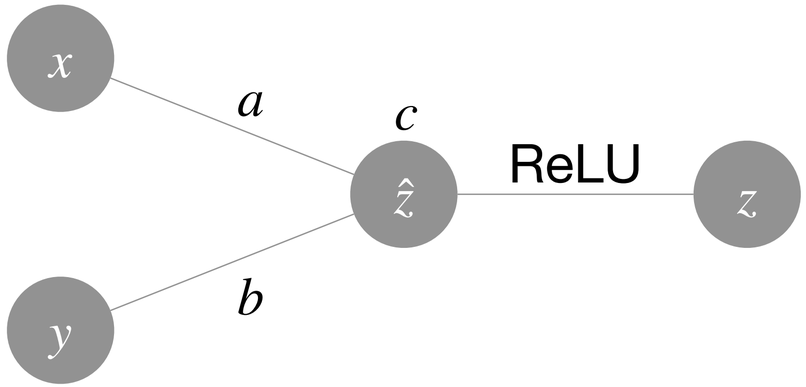
$$x \in [l_x, u_x]$$

$$x \leftarrow \sum_j m_j \cdot x_j + p_j$$

$$y \in [l_y, u_y]$$

$$y \leftarrow \sum_j n_j \cdot x_j + q_j$$

$$\hat{z} = a \cdot x + b \cdot y + c = \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$



$$u_{\hat{z}} \leq 0 \implies z \leftarrow 0$$

$$l_{\hat{z}} \geq 0 \implies z \leftarrow a \cdot x + b \cdot y + c$$

$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \implies z \leftarrow s \wedge s \in [0, u_{\hat{z}}]$$

s is a fresh variable

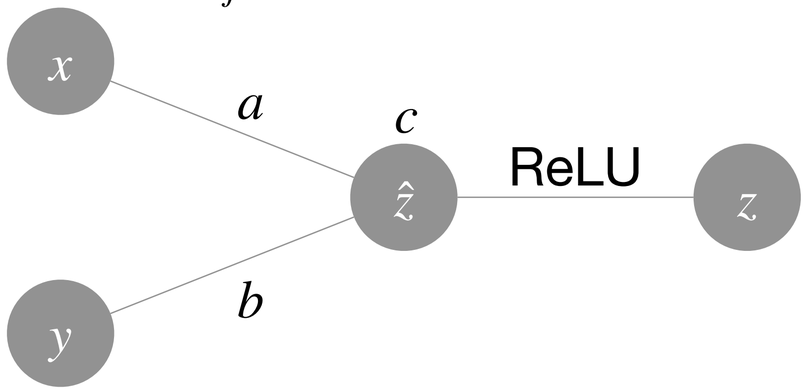
DeepPoly

Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$

$$\hat{z} = a \cdot x + b \cdot y + c$$



$$y \in [l_y, u_y]$$

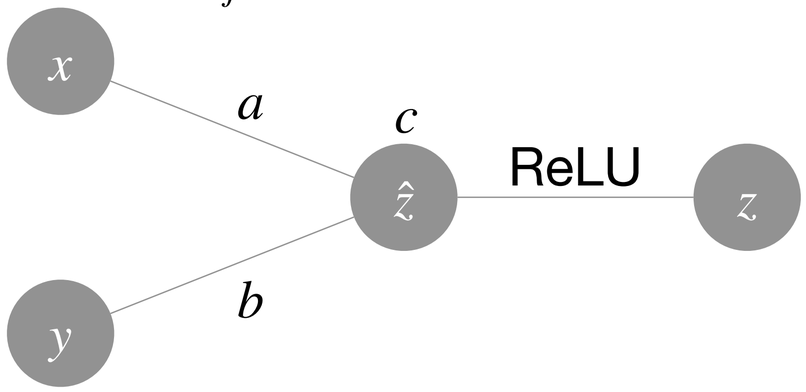
$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

DeepPoly

Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$a \geq 0 \wedge b \geq 0 \implies a \cdot (\sum_j \underline{m}_j \cdot x_j + \underline{p}_j) + b \cdot (\sum_j \underline{n}_j \cdot x_j + \underline{q}_j) + c \leq$$

$$\hat{z} \leq a \cdot (\sum_j \overline{m}_j \cdot x_j + \overline{p}_j) + b \cdot (\sum_j \overline{n}_j \cdot x_j + \overline{q}_j) + c$$

DeepPoly

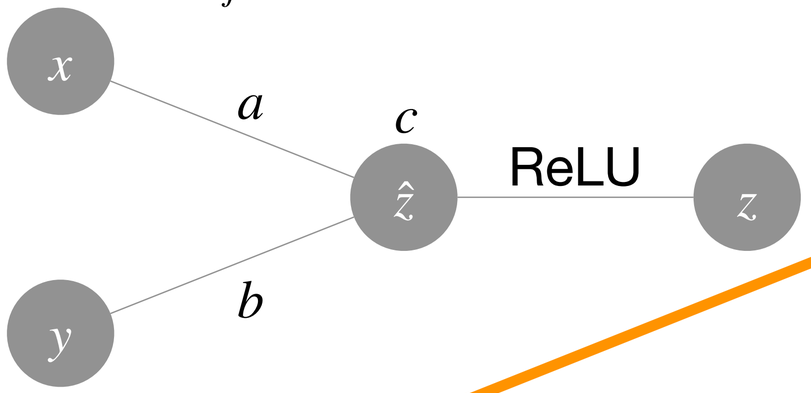
Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$

$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$



$$a \geq 0 \wedge b \geq 0 \implies a \cdot (\sum_j \underline{m}_j \cdot x_j + \underline{p}_j) + b \cdot (\sum_j \underline{n}_j \cdot x_j + \underline{q}_j) + c \leq$$

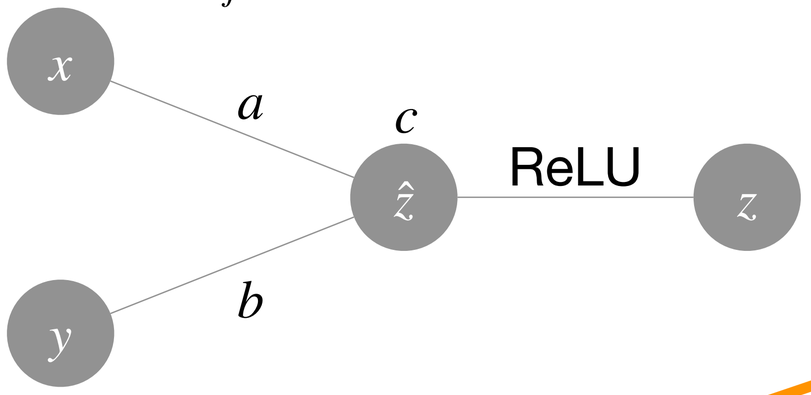
$$\hat{z} \leq a \cdot (\sum_j \overline{m}_j \cdot x_j + \overline{p}_j) + b \cdot (\sum_j \overline{n}_j \cdot x_j + \overline{q}_j) + c$$

DeepPoly

Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$a \geq 0 \wedge b \geq 0 \implies a \cdot (\sum_j \underline{m}_j \cdot x_j + \underline{p}_j) + b \cdot (\sum_j \underline{n}_j \cdot x_j + \underline{q}_j) + c \leq$$

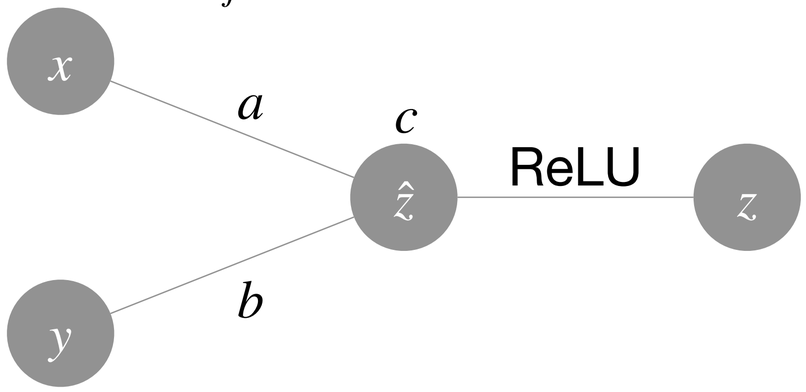
$$\hat{z} \leq a \cdot (\sum_j \overline{m}_j \cdot x_j + \overline{p}_j) + b \cdot (\sum_j \overline{n}_j \cdot x_j + \overline{q}_j) + c$$

DeepPoly

Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$a \geq 0 \wedge b \geq 0 \implies a \cdot (\sum_j \underline{m}_j \cdot x_j + \underline{p}_j) + b \cdot (\sum_j \underline{n}_j \cdot x_j + \underline{q}_j) + c \leq$$

$$\hat{z} \leq a \cdot (\sum_j \overline{m}_j \cdot x_j + \overline{p}_j) + b \cdot (\sum_j \overline{n}_j \cdot x_j + \overline{q}_j) + c$$

$$\iff \sum_j (a \cdot \underline{m}_j + b \cdot \underline{n}_j) \cdot x_j + a \cdot \underline{p}_j + b \cdot \underline{q}_j + c \leq$$

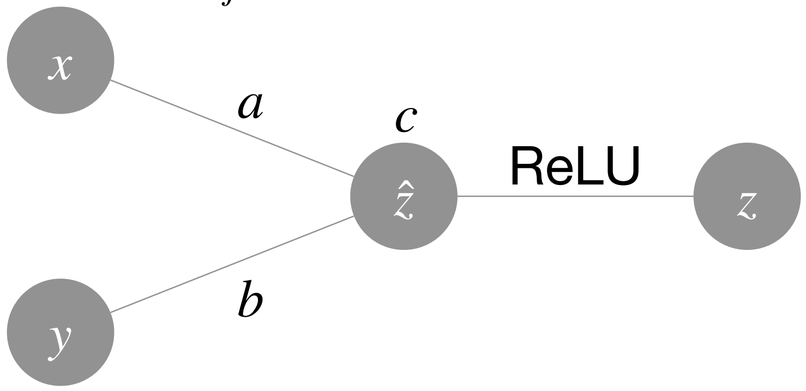
$$\hat{z} \leq \sum_j (a \cdot \overline{m}_j + b \cdot \overline{n}_j) \cdot x_j + a \cdot \overline{p}_j + b \cdot \overline{q}_j + c$$

DeepPoly

Singh et al. @ POPL 2019

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$a \geq 0 \wedge b \geq 0 \implies a \cdot (\sum_j \underline{m}_j \cdot x_j + \underline{p}_j) + b \cdot (\sum_j \underline{n}_j \cdot x_j + \underline{q}_j) + c \leq$$

$$\hat{z} \leq a \cdot (\sum_j \overline{m}_j \cdot x_j + \overline{p}_j) + b \cdot (\sum_j \overline{n}_j \cdot x_j + \overline{q}_j) + c$$

$$\iff \sum_j (a \cdot \underline{m}_j + b \cdot \underline{n}_j) \cdot x_j + a \cdot \underline{p}_j + b \cdot \underline{q}_j + c \leq$$

$$\hat{z} \leq \sum_j (a \cdot \overline{m}_j + b \cdot \overline{n}_j) \cdot x_j + a \cdot \overline{p}_j + b \cdot \overline{q}_j + c$$

$$\iff \ldots \in [l_{\hat{z}}, u_{\hat{z}}]$$

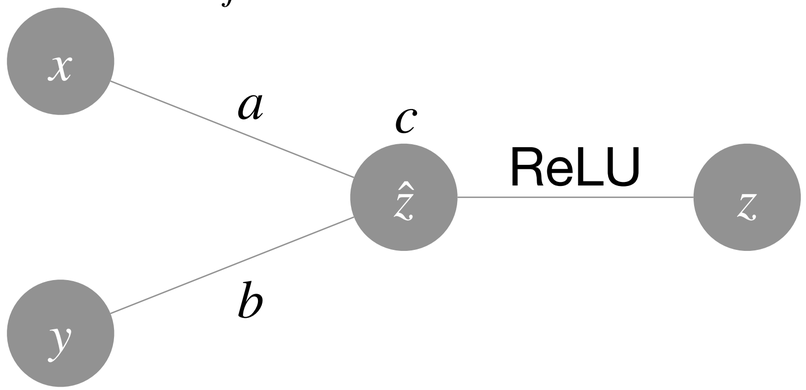
DeepPoly

Singh et al. @ POPL 2019

$$a \cdot x + b \cdot y + c \leq \hat{z} \leq a \cdot x + b \cdot y + c$$

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$u_{\hat{z}} \leq 0 \implies 0 \leq z \leq 0$$

$$l_{\hat{z}} \geq 0 \implies a \cdot x + b \cdot y + c \leq z \leq a \cdot x + b \cdot y + c$$

$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

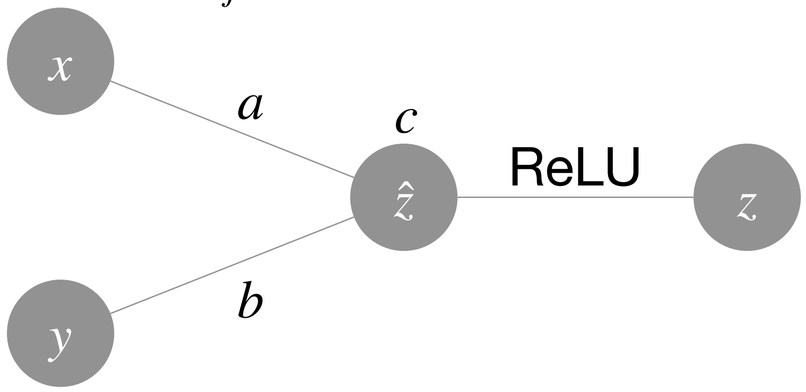
DeepPoly

Singh et al. @ POPL 2019

$$a \cdot x + b \cdot y + c \leq \hat{z} \leq a \cdot x + b \cdot y + c$$

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$u_{\hat{z}} \leq 0 \implies 0 \leq z \leq 0$$

$$l_{\hat{z}} \geq 0 \implies a \cdot x + b \cdot y + c \leq z \leq a \cdot x + b \cdot y + c$$

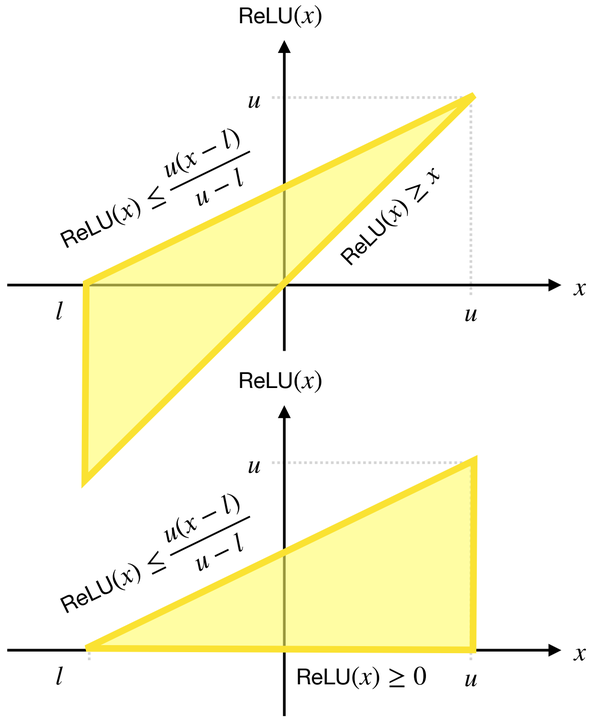
$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \implies ?$$

$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

DeepPoly

Singh et al. @ POPL 2019



$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \wedge -l_{\hat{z}} \leq u_{\hat{z}} \implies \begin{cases} z \leq \dfrac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} \\ z \geq \hat{z} \end{cases}$$

$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \wedge -l_{\hat{z}} > u_{\hat{z}} \implies \begin{cases} z \leq \dfrac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} \\ z \geq 0 \end{cases}$$

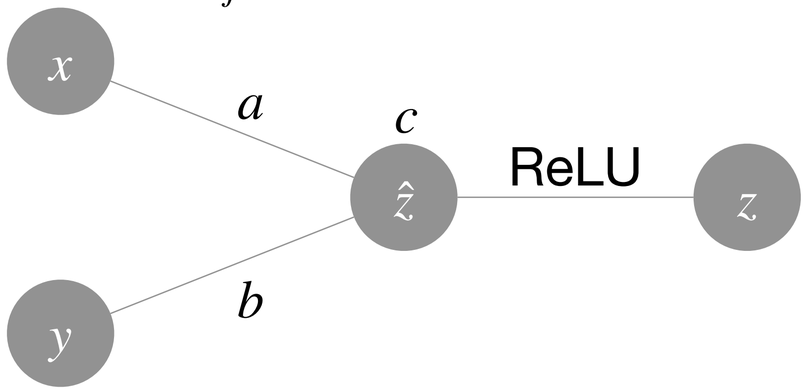
DeepPoly

Singh et al. @ POPL 2019

$$a \cdot x + b \cdot y + c \leq \hat{z} \leq a \cdot x + b \cdot y + c$$

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$u_{\hat{z}} \leq 0 \implies 0 \leq z \leq 0$$

$$l_{\hat{z}} \geq 0 \implies a \cdot x + b \cdot y + c \leq z \leq a \cdot x + b \cdot y + c$$

$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \implies \begin{cases} \hat{z} \leq z \leq \frac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} & -l_{\hat{z}} \leq u_{\hat{z}} \\ 0 \leq z \leq \frac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} & -l_{\hat{z}} > u_{\hat{z}} \end{cases}$$

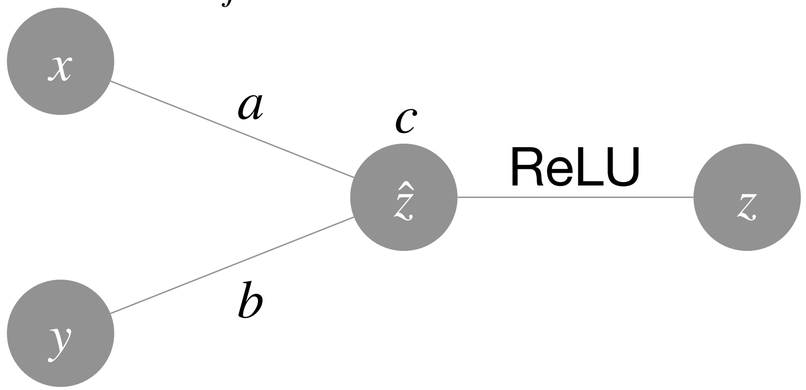
DeepPoly

Singh et al. @ POPL 2019

$$a \cdot x + b \cdot y + c \leq \hat{z} \leq a \cdot x + b \cdot y + c$$

$$x \in [l_x, u_x]$$

$$\sum_j \underline{m}_j \cdot x_j + \underline{p}_j \leq x \leq \sum_j \overline{m}_j \cdot x_j + \overline{p}_j$$



$$y \in [l_y, u_y]$$

$$\sum_j \underline{n}_j \cdot y_j + \underline{q}_j \leq y \leq \sum_j \overline{n}_j \cdot y_j + \overline{q}_j$$

$$u_{\hat{z}} \leq 0 \implies 0 \leq z \leq 0$$

$$l_{\hat{z}} \geq 0 \implies a \cdot x + b \cdot y + c \leq z \leq a \cdot x + b \cdot y + c$$

$$l_{\hat{z}} \leq 0 \wedge u_{\hat{z}} \geq 0 \implies \begin{cases} \hat{z} \leq z \leq \frac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} & -l_{\hat{z}} \leq u_{\hat{z}} \\ 0 \leq z \leq \frac{u_{\hat{z}}(\hat{z} - l_{\hat{z}})}{u_{\hat{z}} - l_{\hat{z}}} & -l_{\hat{z}} > u_{\hat{z}} \end{cases}$$

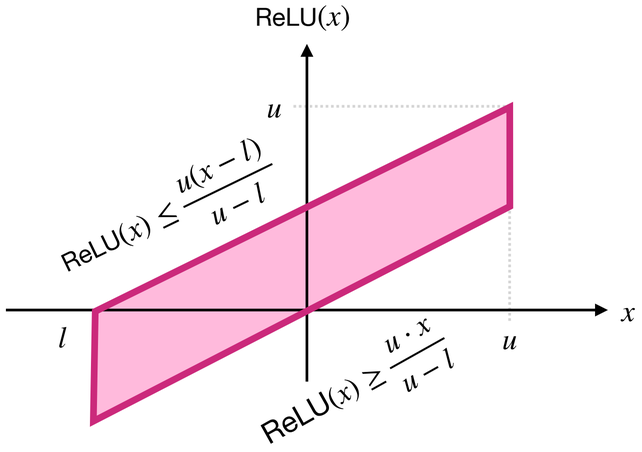
$$\bigcap$$

Symbolic $\implies z \leftarrow s$

Neurify

Wang et al. @ NeurIPS 2018

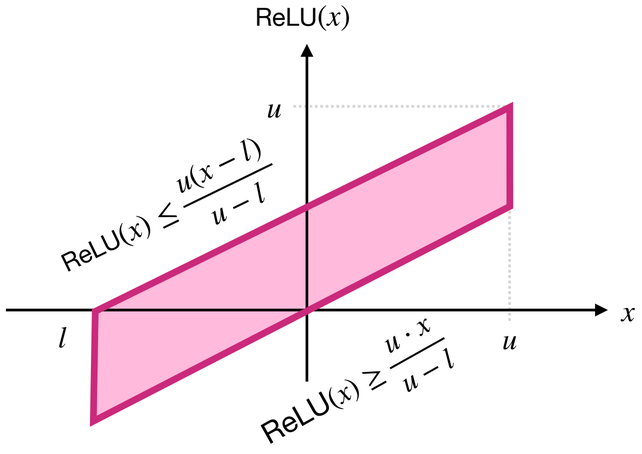
$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$

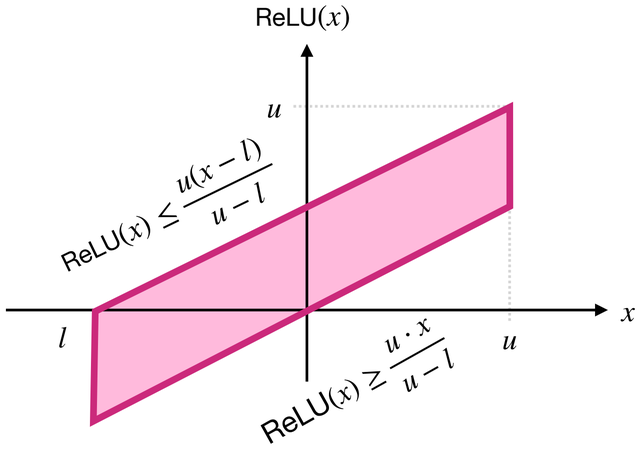


$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

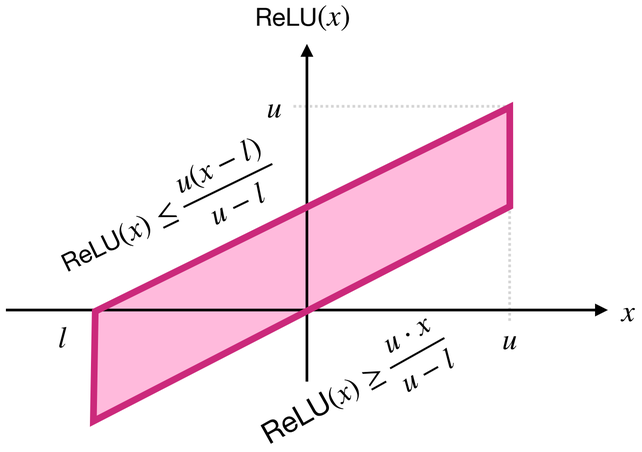
$$l_{\text{low}} \leq 0 \wedge l_{\text{up}} \geq 0 \implies \underline{\text{ReLU}}(\hat{z}) \leq z$$

$$l_{\text{low}} \leq 0 \wedge u_{\text{up}} \geq 0 \wedge$$

Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

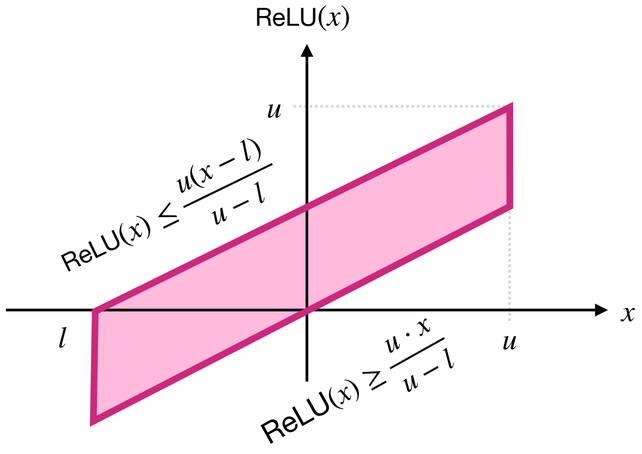
$$l_{\text{low}} \leq 0 \land l_{\text{up}} \geq 0 \implies \underline{\text{ReLU}}(\hat{z}) \leq z$$
$$l_{\text{low}} \leq 0 \land l_{\text{up}} \leq 0 \implies 0 \leq z$$

$$l_{\text{low}} \leq 0 \land u_{\text{up}} \geq 0 \land$$

Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

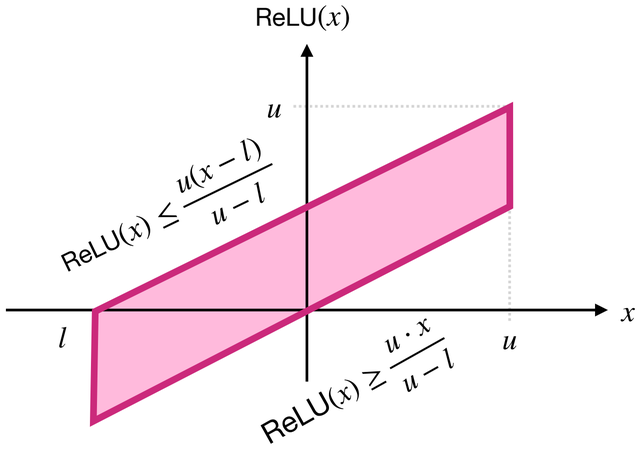
$$l_{\text{low}} \leq 0 \wedge u_{\text{up}} \geq 0 \wedge$$

$$l_{\text{low}} \leq 0 \wedge l_{\text{up}} \geq 0 \implies \underline{\text{ReLU}}(\hat{z}) \leq z$$
$$l_{\text{low}} \leq 0 \wedge l_{\text{up}} \leq 0 \implies 0 \leq z$$
$$u_{\text{low}} \leq 0 \wedge u_{\text{up}} \geq 0 \implies z \leq \overline{\text{ReLU}}(\hat{z})$$

Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

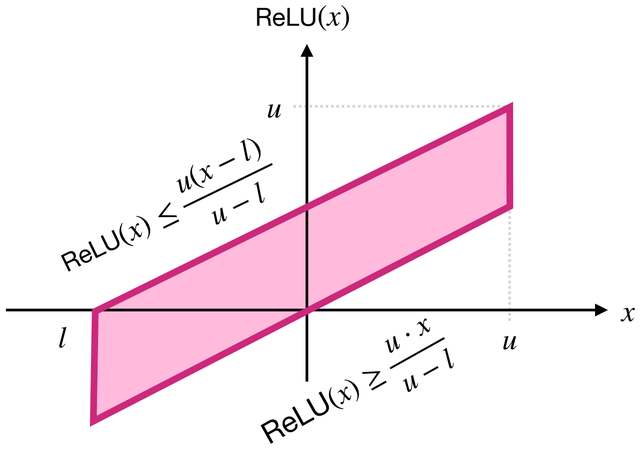
$$l_{\text{low}} \leq 0 \land u_{\text{up}} \geq 0 \land$$

$$\begin{aligned}
l_{\text{low}} \leq 0 \land l_{\text{up}} \geq 0 &\implies \underline{\text{ReLU}}(\hat{z}) \leq z \\
l_{\text{low}} \leq 0 \land l_{\text{up}} \leq 0 &\implies 0 \leq z \\
u_{\text{low}} \leq 0 \land u_{\text{up}} \geq 0 &\implies z \leq \overline{\text{ReLU}}(\hat{z}) \\
u_{\text{low}} \geq 0 \land u_{\text{up}} \geq 0 &\implies z \leq \hat{z}
\end{aligned}$$

Neurify

Wang et al. @ NeurIPS 2018

$$z \leftarrow \text{ReLU}(\hat{z}) \qquad \hat{z} \in [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$



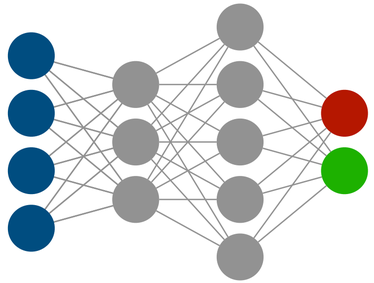
$$\underline{\text{ReLU}}(\hat{z}) \leq z \leq \overline{\text{ReLU}}(\hat{z})$$

$l_{\text{low}} \leq 0 \wedge u_{\text{up}} \geq 0 \wedge$

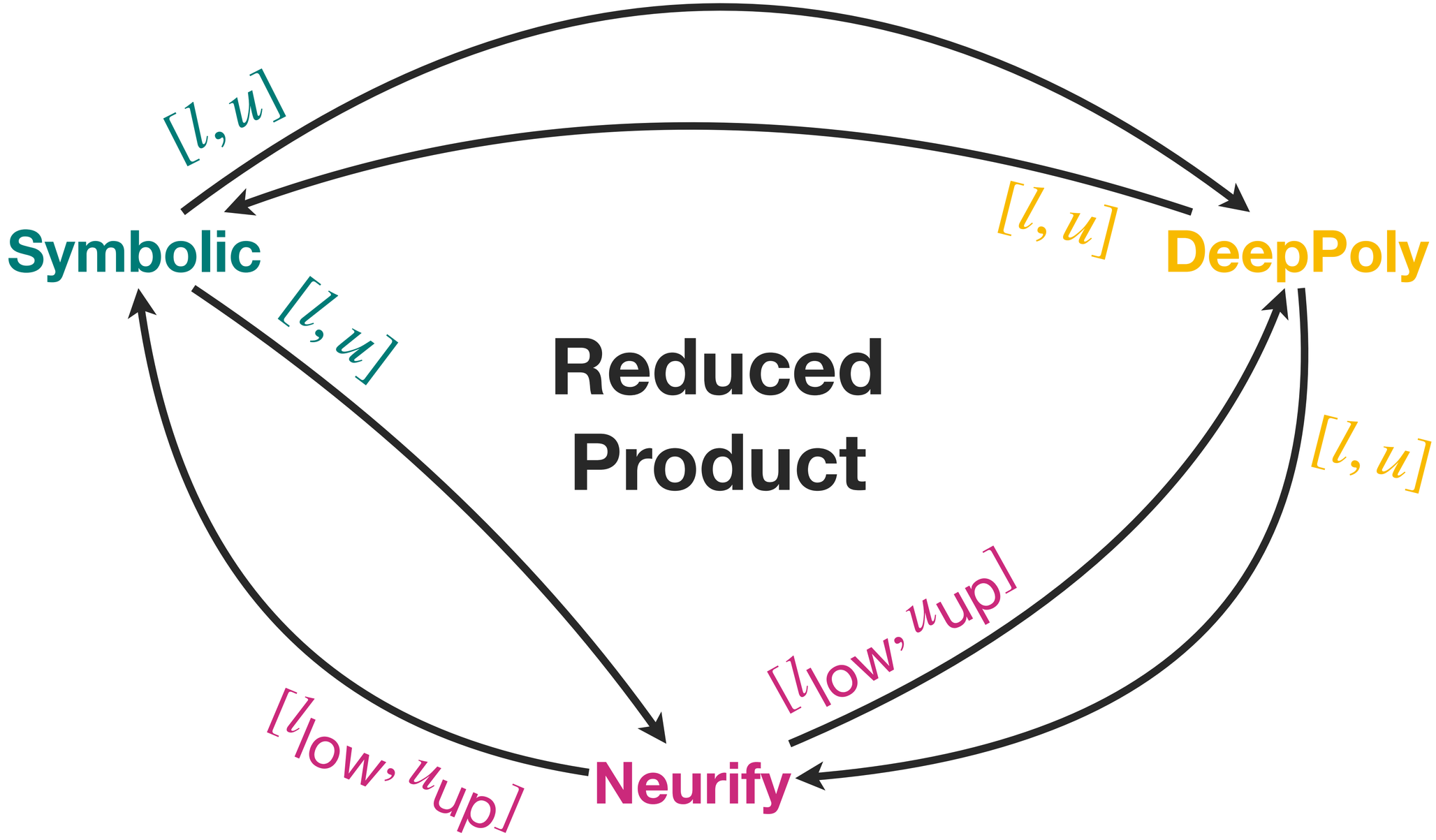
$$l_{\text{low}} \leq 0 \wedge l_{\text{up}} \geq 0 \implies \underline{\text{ReLU}}(\hat{z}) \leq z$$
$$l_{\text{low}} \leq 0 \wedge l_{\text{up}} \leq 0 \implies 0 \leq z$$ ← No relaxation
$$u_{\text{low}} \leq 0 \wedge u_{\text{up}} \geq 0 \implies z \leq \overline{\text{ReLU}}(\hat{z})$$
$$u_{\text{low}} \geq 0 \wedge u_{\text{up}} \geq 0 \implies z \leq \hat{z}$$ ← No relaxation

Precision-vs-Scalability

L	U	Symbolic	DeepPoly	Neurify
0.5	3	48.78%	49.01%	46.49%
	5	56.11%	56.15%	53.06%
0.25	3	83.63%	81.82%	81.40%
	5	91.67%%	91.58%	92.33%



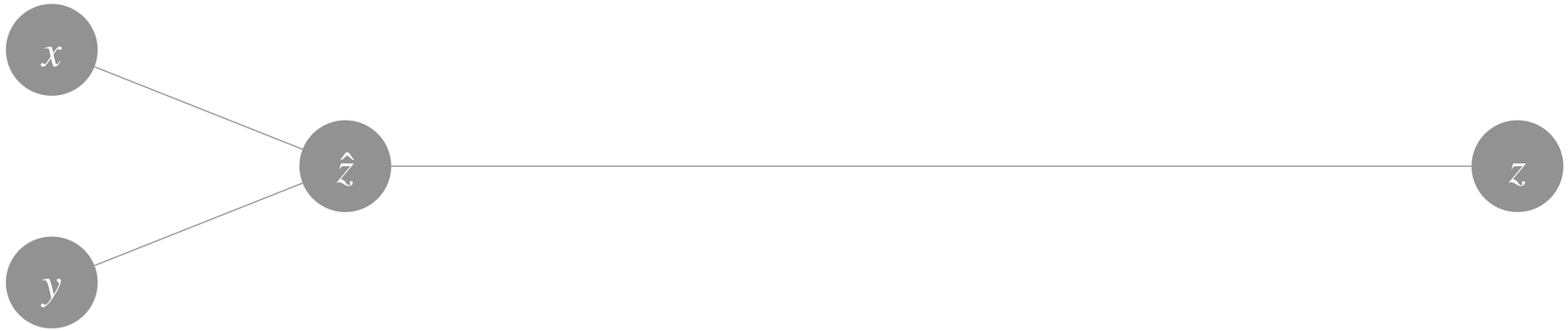
- 4 Hidden Layers
- 5 Neuron per Layer
- 23 inputs $\in [0,1]$
- 2 Output classes


Symbolic
DeepPoly
Neurify
Reduced
Product
[l, u]
[l, u]
[l, u]
[l, u]
[l_low, u_up]
[l_low, u_up]

Reduced Product

Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] & \textbf{Symbolic} \\ [l_x, u_x] & \textbf{DeepPoly} \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] & \textbf{Neurify} \end{cases}$$



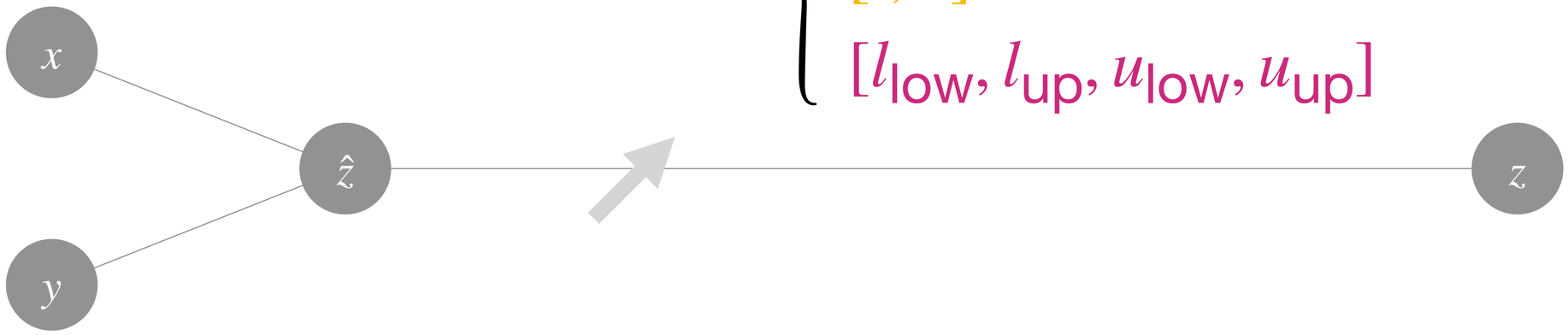
$$x \in \begin{cases} [l_y, u_y] & \textbf{Symbolic} \\ [l_y, u_y] & \textbf{DeepPoly} \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] & \textbf{Neurify} \end{cases}$$

Reduced Product

Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

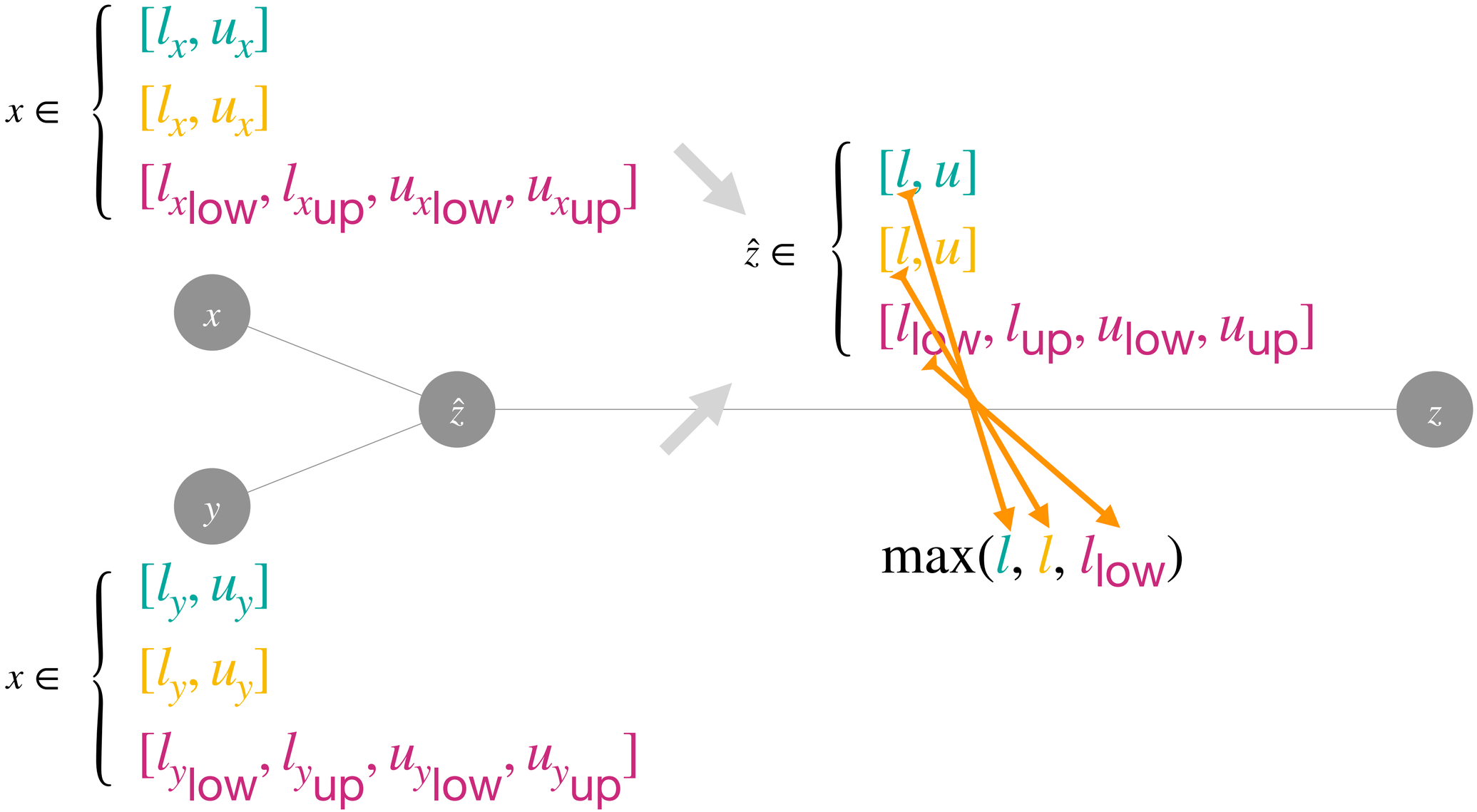
$$\hat{z} \in \begin{cases} [l, u] \\ [l, u] \\ [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}] \end{cases}$$



$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

Reduced Product

Mazzucato, Urban @ SAS 2021



Reduced Product

Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

$$\hat{z} \in \begin{cases} [l, u] \\ [l, u] \\ [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}] \end{cases}$$

$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

$$[\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})]$$

Reduced Product

Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

$$\hat{z} \in \begin{cases} [l, u] \\ [l, u] \\ [l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}] \end{cases}$$

$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

$$\hat{z} \in [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})]$$

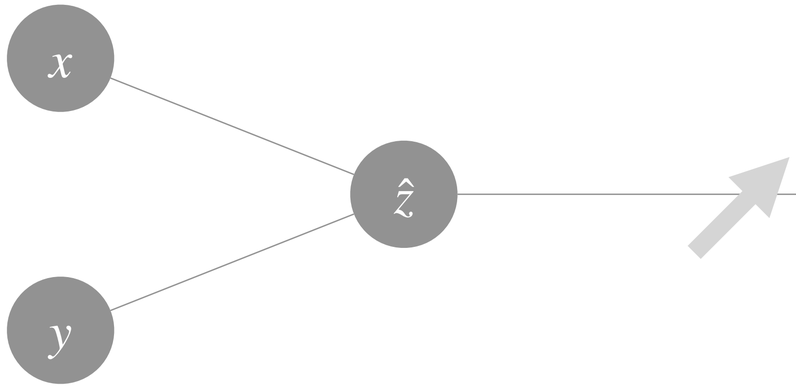
Reduced Product

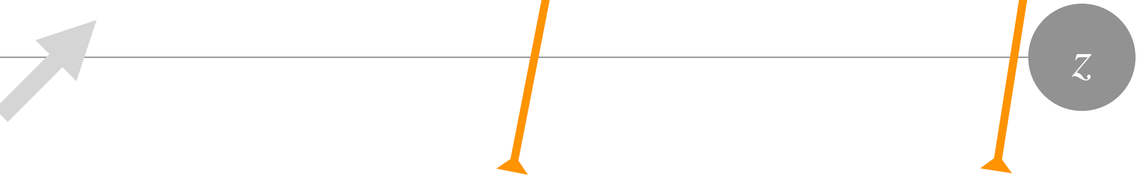
Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

$$\hat{z} \in \begin{cases} [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \end{cases}$$

Symbolic





$$\hat{z} \in [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})]$$

$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

Reduced Product

Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

$$\hat{z} \in \begin{cases} [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \\ [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \end{cases}$$

DeepPoly

$$\hat{z} \in [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})]$$

$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

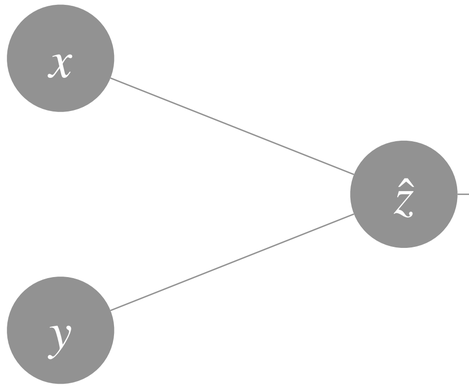
Reduced Product

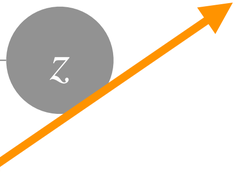
Mazzucato, Urban @ SAS 2021

$$x \in \begin{cases} [l_x, u_x] \\ [l_x, u_x] \\ [l_{x\text{low}}, l_{x\text{up}}, u_{x\text{low}}, u_{x\text{up}}] \end{cases}$$

$$\hat{z} \in \begin{cases} [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \\ [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \\ [\max(l, l, l_{\text{low}}), \ldots, \ldots, \min(u, u, u_{\text{up}})] \end{cases}$$

Neurify



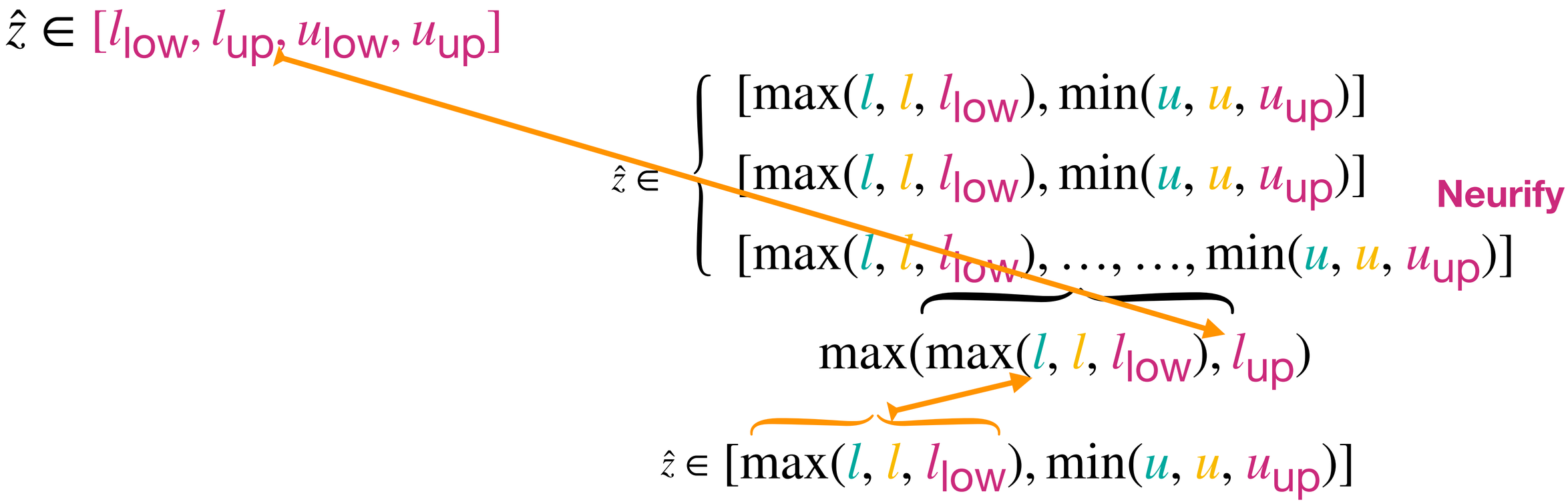


$$\hat{z} \in [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})]$$

$$x \in \begin{cases} [l_y, u_y] \\ [l_y, u_y] \\ [l_{y\text{low}}, l_{y\text{up}}, u_{y\text{low}}, u_{y\text{up}}] \end{cases}$$

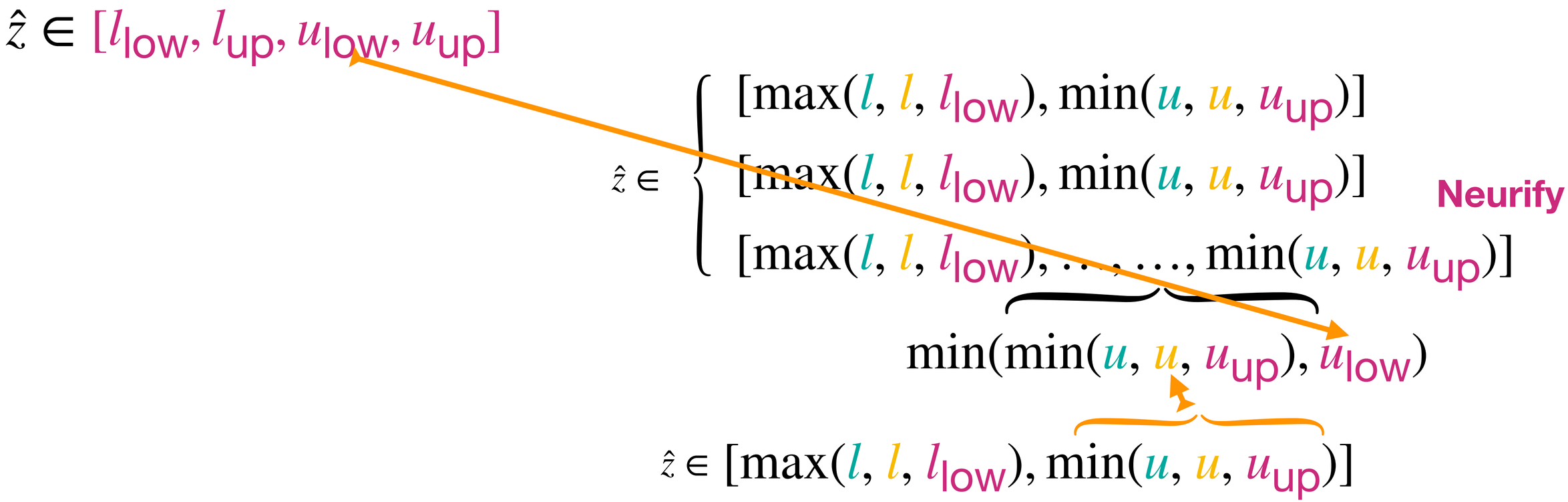
Reduced Product

Mazzucato, Urban @ SAS 2021



Reduced Product

Mazzucato, Urban @ SAS 2021



Reduced Product

Mazzucato, Urban @ SAS 2021

$$\hat{z} \in \begin{cases} [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \\ [\max(l, l, l_{\text{low}}), \min(u, u, u_{\text{up}})] \\ [\max(l, l, l_{\text{low}}), \max(\max(l, l, l_{\text{low}}), l_{\text{up}}), \\ \quad \min(\min(u, u, u_{\text{up}}), u_{\text{low}}), \min(u, u, u_{\text{up}})] \end{cases}$$

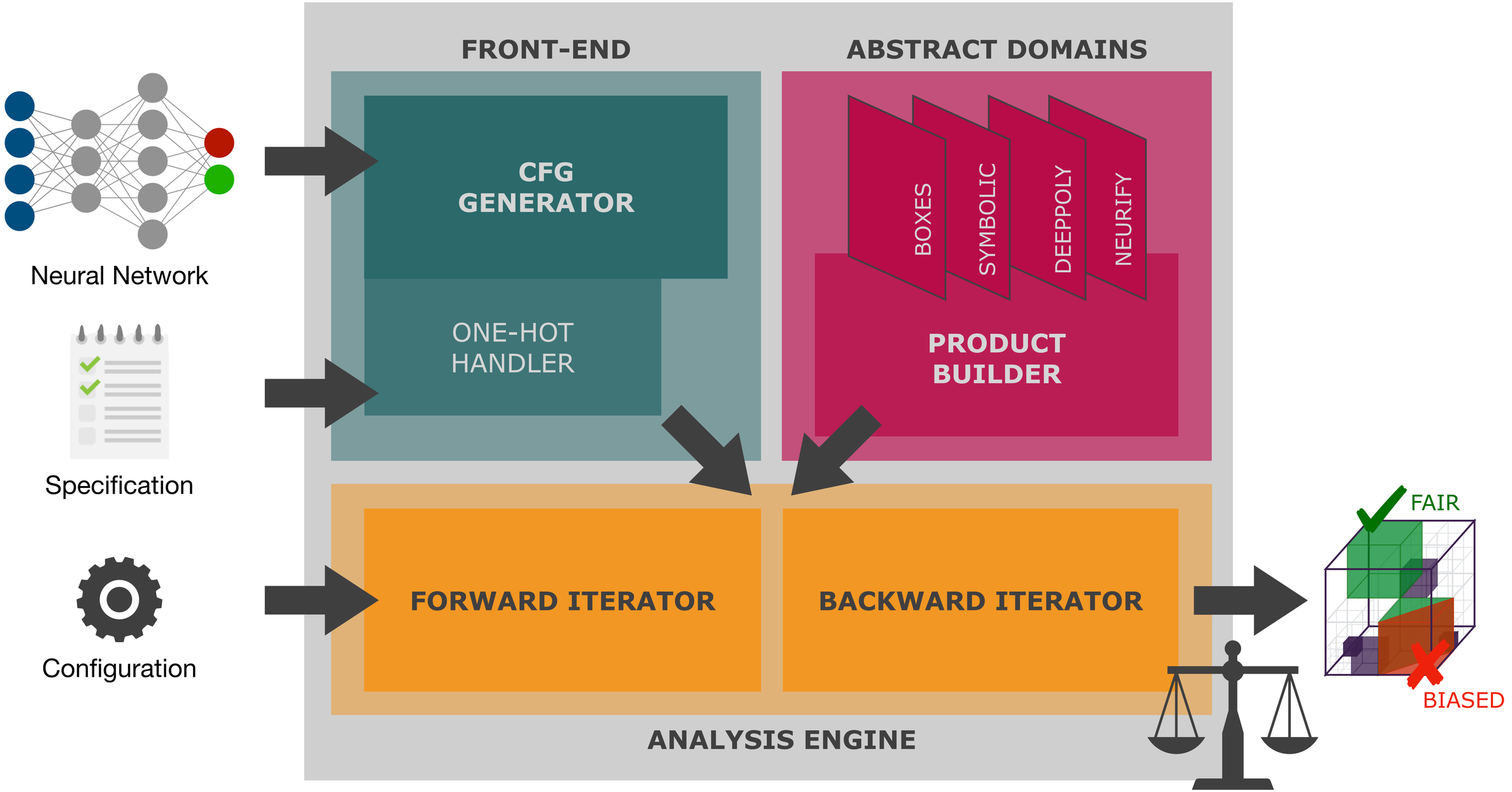
Symbolic

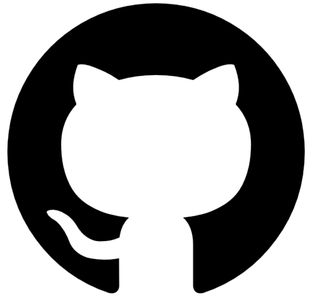
DeepPoly

Neurify

Precision-vs-Scalability

L	U	Symbolic	DeepPoly	Neurify	Product	
0.5	3	48.78%	49.01%	46.49%	59.20%	+10.3%
	5	56.11%	56.15%	53.06%	68.23%	+11.9%
0.25	3	83.63%	81.82%	81.40%	87.04%	+3.4%
	5	91.67%%	91.58%	92.33%	95.48%	+3.2%


Neural Network
Specification
Configuration
FRONT-END
ABSTRACT DOMAINS
CFG GENERATOR
ONE-HOT HANDLER
BOXES
SYMBOLIC
DEEPPOLY
NEURIFY
PRODUCT BUILDER
FORWARD ITERATOR
BACKWARD ITERATOR
ANALYSIS ENGINE
FAIR
BIASED



Check it out on **GitHub**!
https://github.com/caterinaurban/libra

zenodo

Ready-to-go **Docker image*** at
https://doi.org/10.5281/zenodo.4737450

* no installation needed!