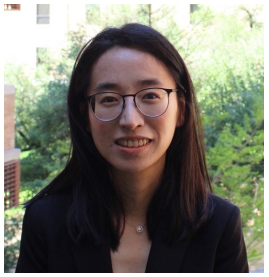
FairQuant: Certifying and Quantifying Fairness of Deep Neural Networks

ICSE 2025: Wednesday, April 30, 2025

Brian Hyeongseok Kim*, Jingbo Wang[+], and Chao Wang*



*University of Southern California, [+]Purdue University

USC Viterbi School of Engineering

University of Southern California

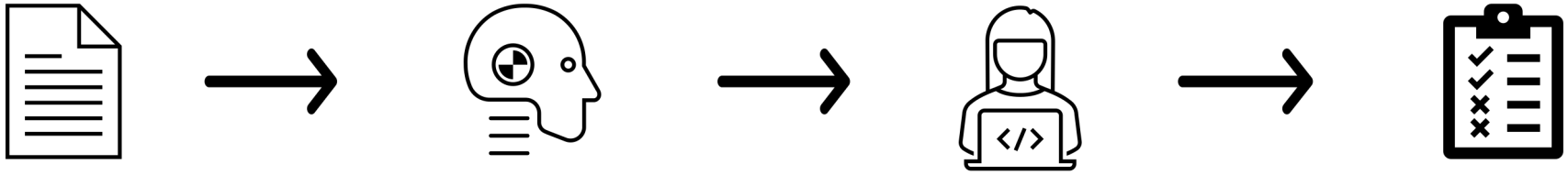
Motivation

Public Policy

Banking

Criminal Justice

Motivation



Training Data | Learning Algorithm | Trained Model | Does φ hold?

Motivation

Existing verifier for individual fairness [1]

- SMT solver → limited scalability
- Qualitative → done after finding one counterexample

Existing verifiers for local robustness [2, 3]

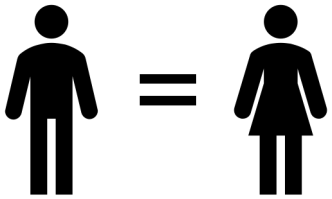
- Abstract interpretation → scalable, sound
- Cannot verify individual fairness

Our Contribution: FairQuant

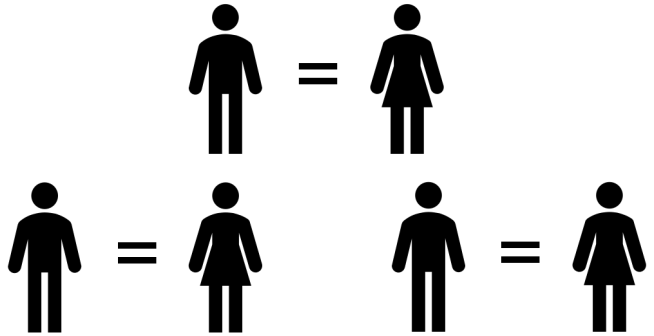
- Verify **individual fairness** of a neural network via **abstract interpretation**
- **Quantitatively** measure the degree of individual fairness across the input domain

[1] Biswas and Rajan. 2023. Fairify: Fairness Verification of Neural Networks. In 45th International Conference on Software Engineering.
[2] Wang et al. 2018. Formal security analysis of neural networks using symbolic intervals. In 27th USENIX Security Symposium.
[3] Wang et al. 2018. Efficient formal safety analysis of neural networks. In 32nd International Conference on Neural Information Processing Systems.

Individual Fairness



(Local) Individual Fairness
for a given input



(Global) Individual Fairness
for the entire input domain

Individual Fairness

Given: classifier f, protected attribute x_j, input domain X

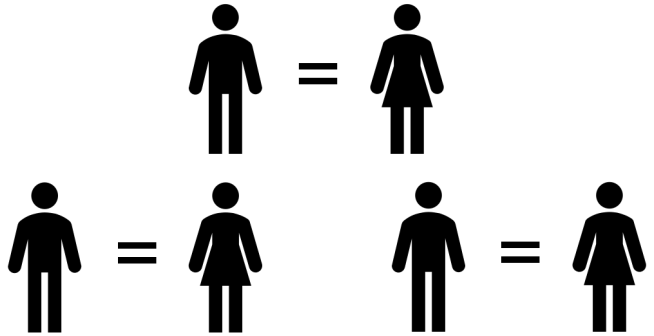
f satisfies individual fairness

with regards to x_j for the domain X

if and only if

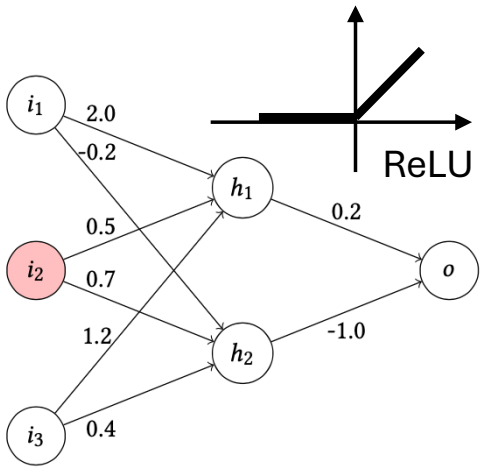
$f(x) = f(x')$ for all $x, x' \in X$

where x and x' differ only in x_j.



(Global) Individual Fairness
for the entire input domain

Example



Input:

$x_1 \in \{1, 2, 3, 4, 5\}$	Interview score
$x_2 \in \{0, 1\}$	Gender, protected attribute x_j
$x_3 \in \{0, 1, 2, 3, 4, 5\}$	Years of experience

Output

$f(x) = \{0,1\}$	Binary classification

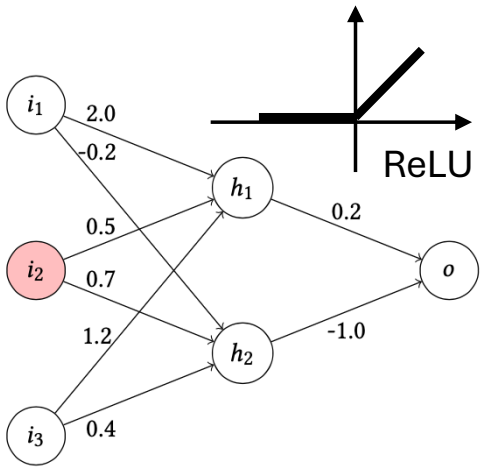
$x = [5, 0, 5]$ $\quad f(x) = 1$
$x' = [5, 1, 5]$ $\quad f(x') = 1$

$f(x) = f(x')$? ✓

$x = [1, 1, 3]$ $\quad f(x) = 1$
$x' = [1, 0, 3]$ $\quad f(x') = 0$

$f(x) = f(x')$? ✗

Example



Input:

$x_1 \in \{1, 2, 3, 4, 5\}$	Interview score
$x_2 \in \{0, 1\}$	Gender, protected attribute x_j
$x_3 \in \{0, 1, 2, 3, 4, 5\}$	Years of experience

Output

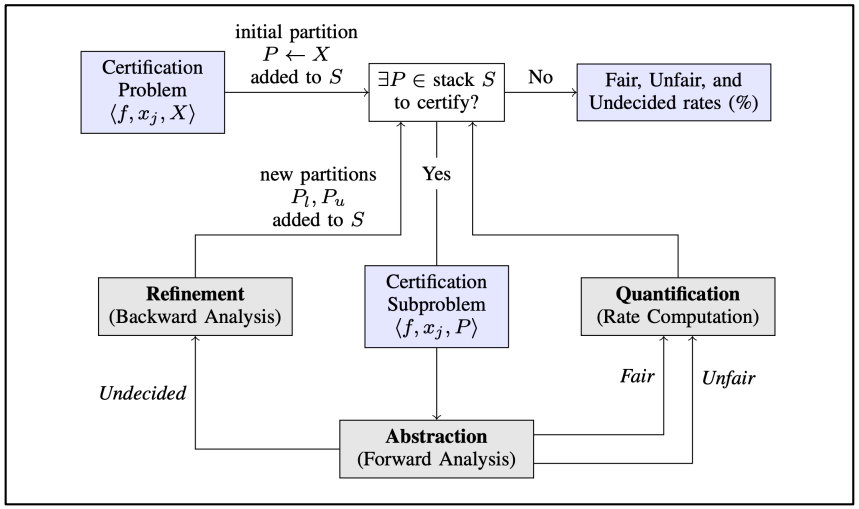
$f(x) = \{0,1\}$	Binary classification

100% fair or unfair?

Scalability?

Overview

1. Our Method



2. Evaluation

Dataset	DNN	Fairify [12]						FairQuant (new)					
		Time	Cex	#Cex	Cer%	Fal%	Und%	Time	Cex	#Cex	Cer%	Fal%	Und%
Bank	BM-1	30m	✓	11	10.00	0	90.00	4.82s	✓	2820	94.23	0	5.76
	BM-2	31m	✓	28	16.07	0	83.93	3.23s	✓	2479	93.41	0	6.58
	BM-3	31m	✓	27	19.60	0	81.40	1.21s	✓	1864	95.69	0	4.30
	BM-4	35m	✓	4	3.72	0	96.18	71.12s	✓	5135	87.03	0	12.96
	BM-5	23m	✓	114	77.25	0	22.75	1.03s	✓	1474	96.27	0	3.72
	BM-6	12m	✓	155	69.41	0	30.59	0.44s	✓	1426	96.44	0	3.55
	BM-7	30m	✓	57	9.41	0	90.59	12.26s	✓	7017	83.65	0	16.34
	BM-8	30m	✓	1	0.98	0	99.02	18.99s	✓	3074	90.75	0	9.24
German	GC-1	32m	✓	22	0	0	100	9.73s	✓	31585	32.67	0	67.33
	GC-2	33m	✓	6	0	0	100	31.72s	✓	32655	42.21	0	57.79
	GC-3	8m	✓	194	2.98	0	97.02	6.77s	✓	25963	58.44	0	41.55
	GC-4	4m	✓	2	99.00	0	1.00	0.29s	✓	77	99.65	0	0.34
	GC-5	30m	✗	0	0	0	100	1.24s	✓	9	99.80	0	0.19
Adult	AC-1	32m	✓	3	0.03	0	99.97	3.23s	✓	6151	90.68	0	9.31
	AC-2	31m	✓	9	0.01	0	99.99	30.04s	✓	13008	79.93	0	20.06
	AC-3	32m	✓	20	0	0	100	37.12s	✓	60494	33.29	0	66.70
	AC-4	36m	✗	0	0	0	100	8m	✓	61324	24.79	0	75.20
	AC-5	33m	✗	0	0	0	100	4m	✓	71012	19.12	0	80.87
	AC-6	33m	✓	4	0.01	0	99.99	10.20s	✓	31593	58.82	0	41.17
	AC-7	30m	✗	0	0.01	0	99.99	4m	✓	25588	31.72	0	68.27
	AC-8	30m	✓	39	0.03	0	99.97	11.18s	✓	26179	66.50	0	33.49
	AC-9	30m	✓	126	0.64	0	99.36	3.50s	✓	5470	91.13	0	8.86
	AC-10	32m	✓	8	0.03	0	99.97	5.01s	✓	9033	87.65	0	12.34
	AC-11	30m	✗	0	0	0	100	36.44s	✓	24516	58.01	0	41.98
	AC-12	30m	✗	0	0.02	0	99.98	0.91s	✓	8824	70.82	0	29.17
Compas	compas-1	17m	✓	2	80.00	0.32	19.68	0.01s	✓	17	97.27	2.72	0
	compas-2	31m	✗	0	0	0	100	0.01s	✓	15	97.59	2.40	0
	compas-3	30m	✗	0	0	0	100	0.30s	✓	12	98.07	1.92	0
	compas-4	30m	✗	0	0	0	100	0.01s	✓	14	97.75	2.24	0
	compas-5	T/O	✗	0	0	0	100	5.24s	✓	11	98.23	1.76	0
	compas-6	M/O	✗	0	0	0	100	9.19s	✓	12	98.07	1.92	0
	compas-7	M/O	✗	0	0	0	100	101.25s	✓	15	97.59	2.40	0

Our Method

FairQuant

Abstraction	Symbolic Forward Analysis
Refinement	Iterative Backward Analysis
Quantification	Fairness Rate Calculation

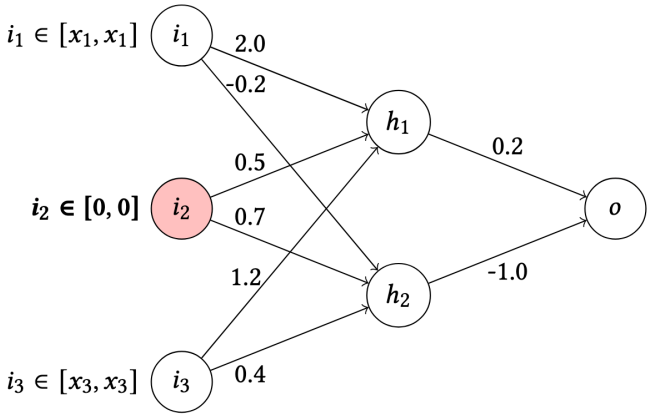
Our Method

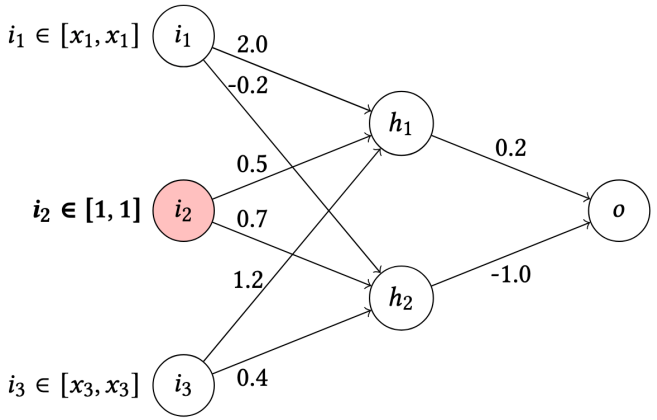
FairQuant

Abstraction	Symbolic Forward Analysis
Refinement	Iterative Backward Analysis
Quantification	Fairness Rate Calculation

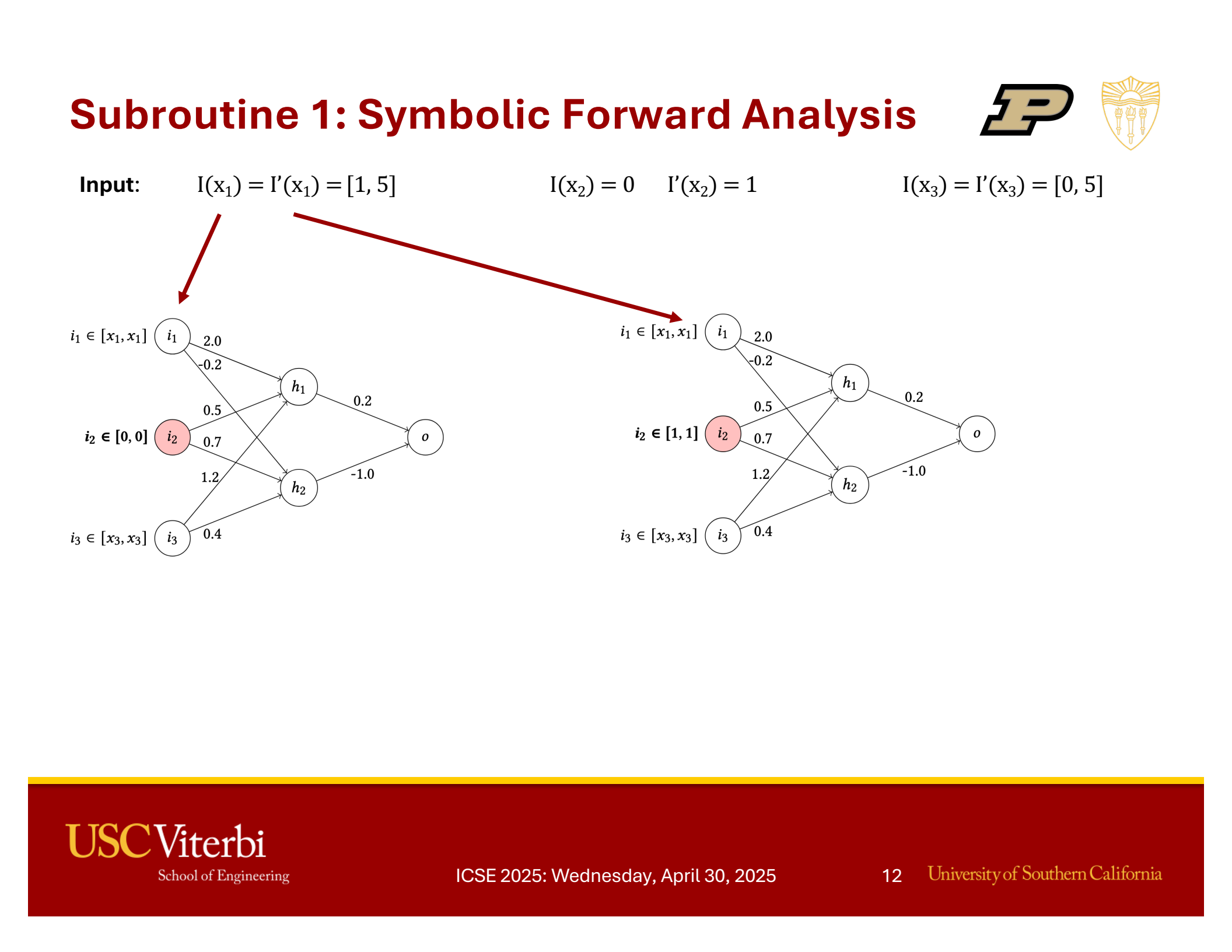
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$





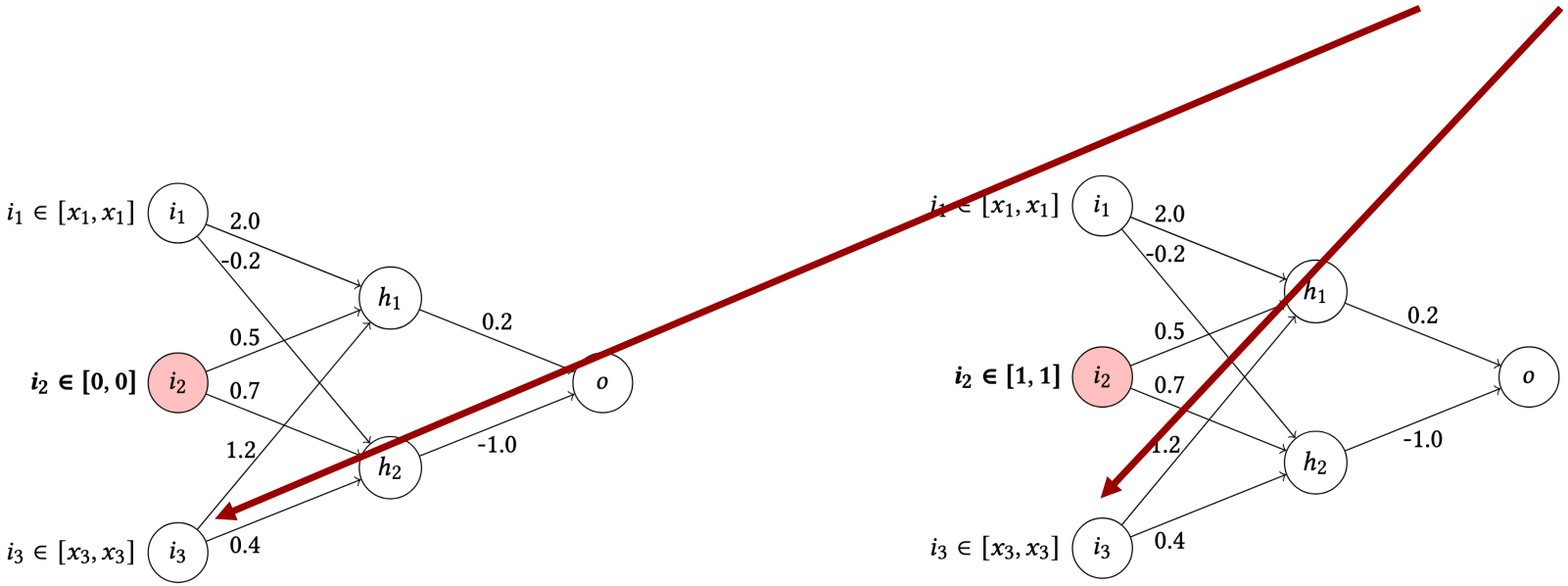
Subroutine 1: Symbolic Forward Analysis



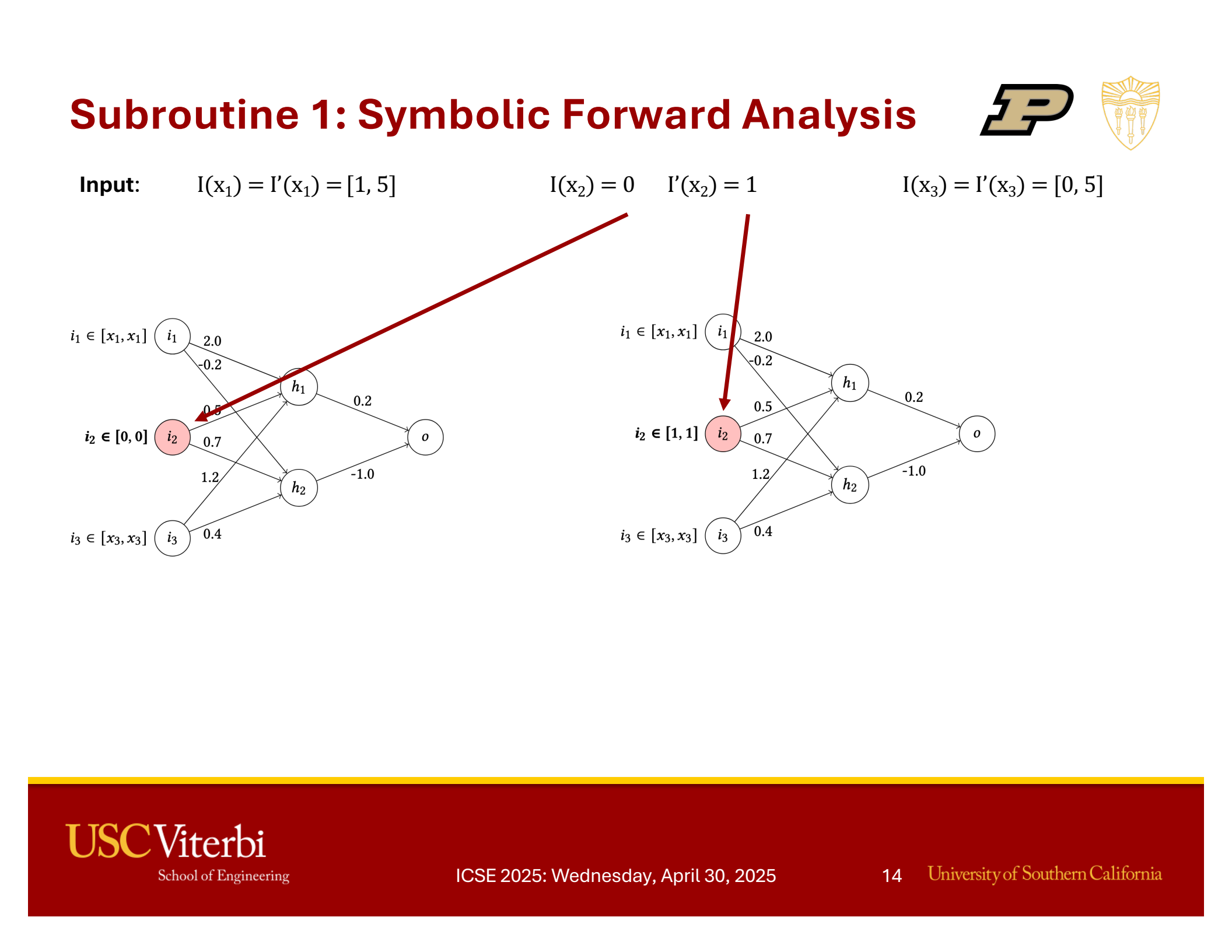
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



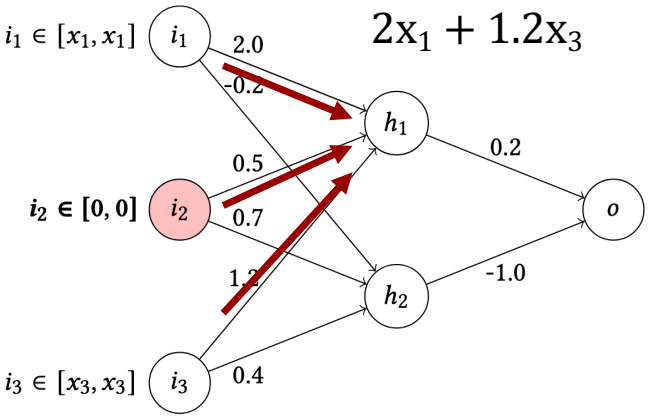
Subroutine 1: Symbolic Forward Analysis



Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

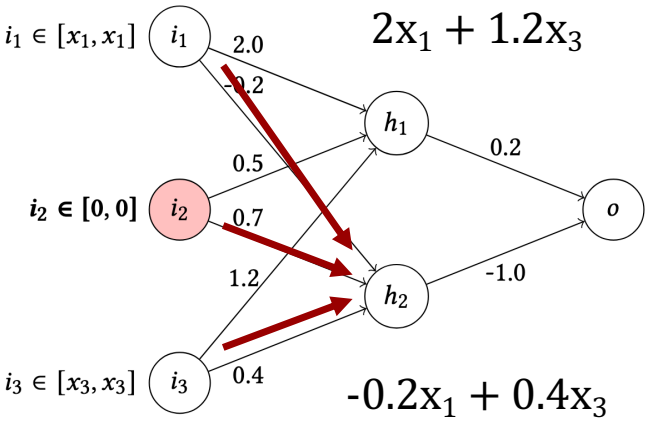
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



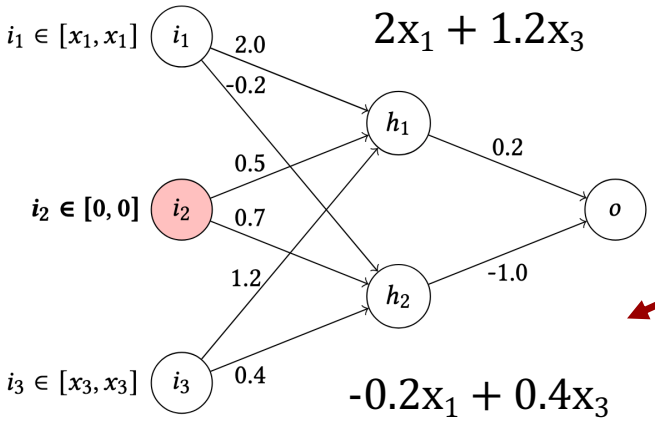
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

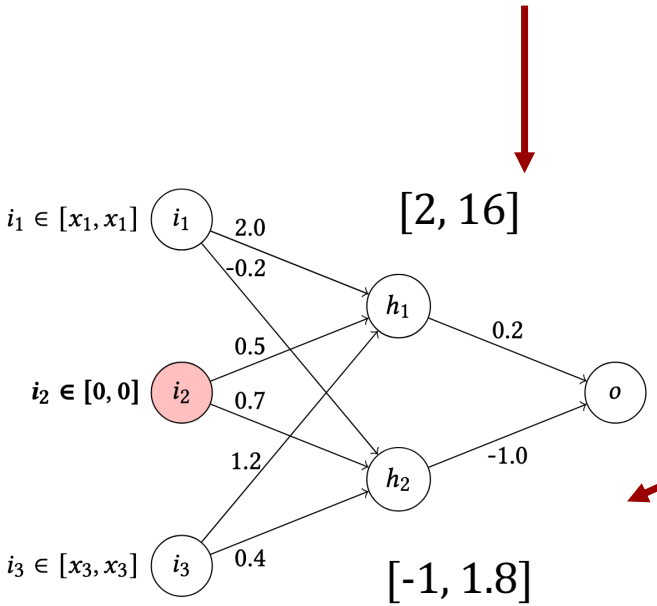


$2x_1 + 1.2x_3$

$-0.2x_1 + 0.4x_3$

Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

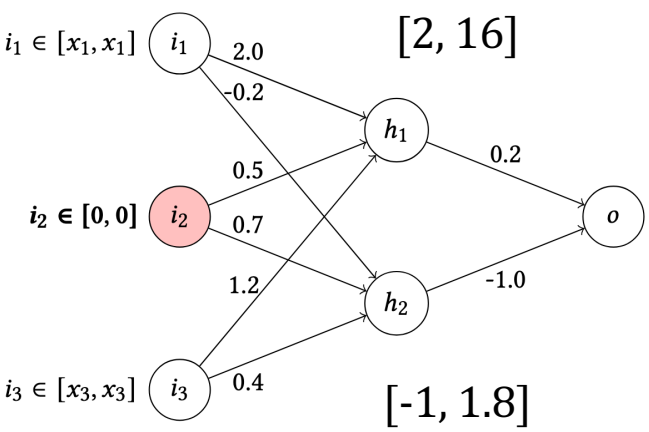


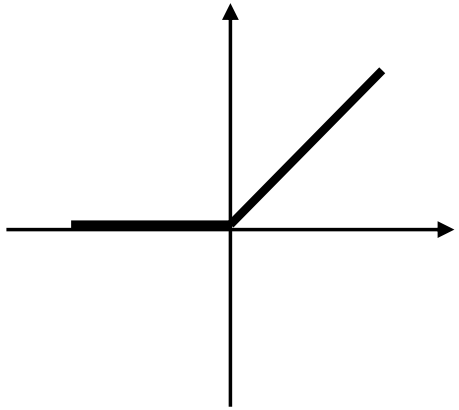
[2, 16]

[-1, 1.8]

Subroutine 1: Symbolic Forward Analysis

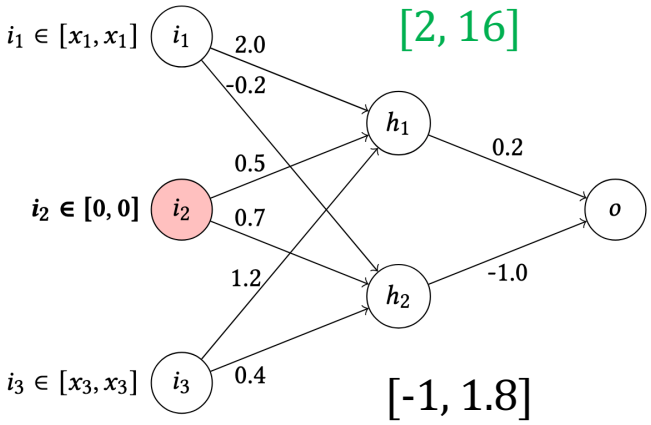
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

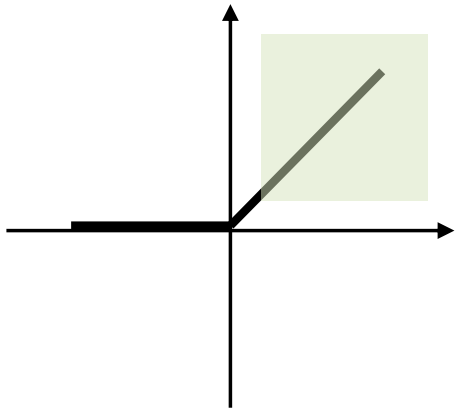




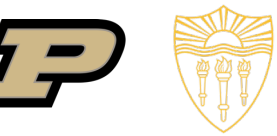
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

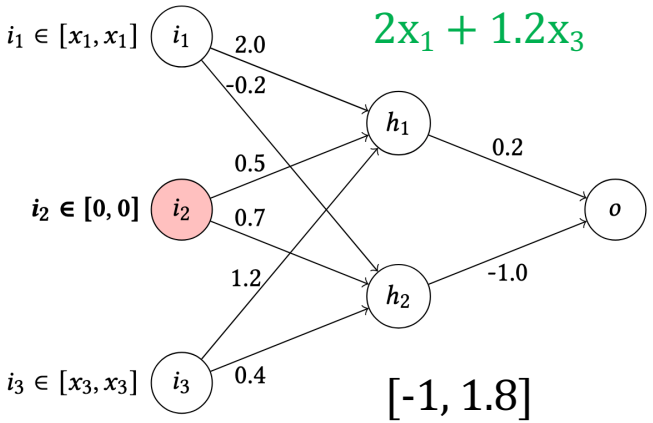


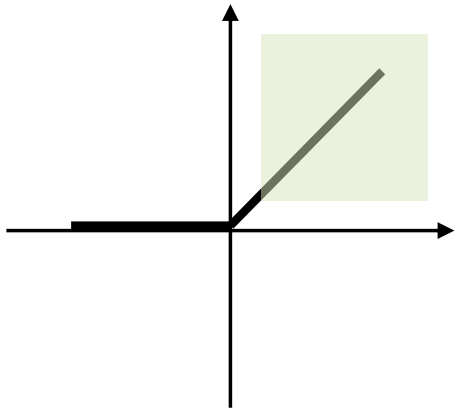


Subroutine 1: Symbolic Forward Analysis



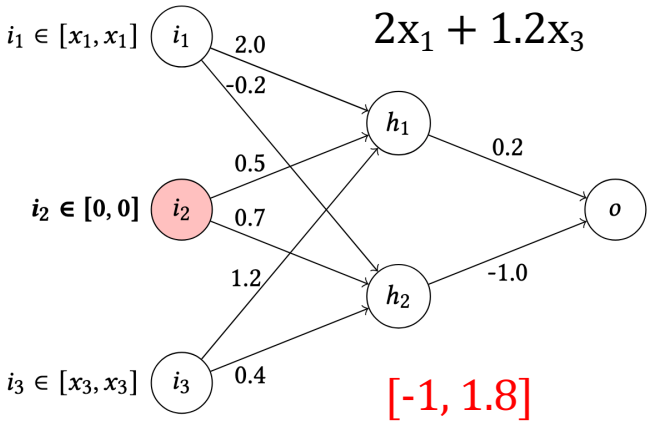
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

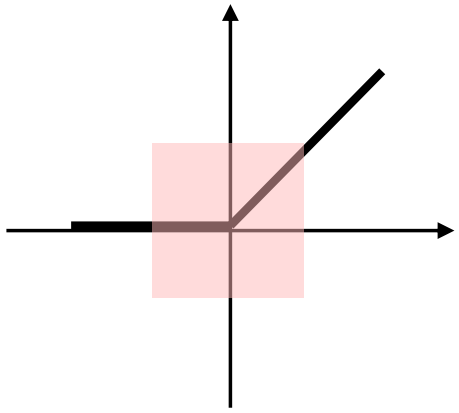




Subroutine 1: Symbolic Forward Analysis

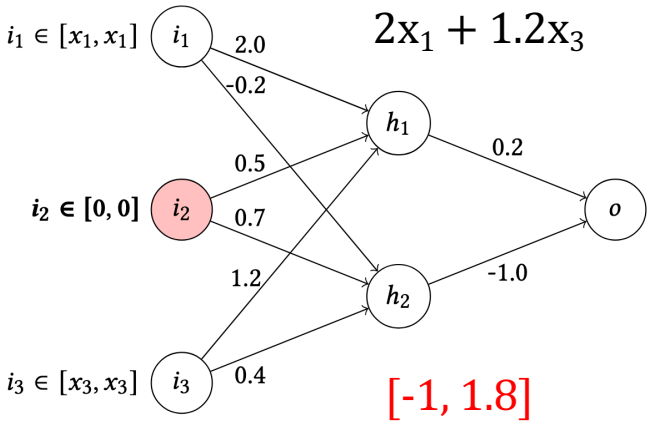
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

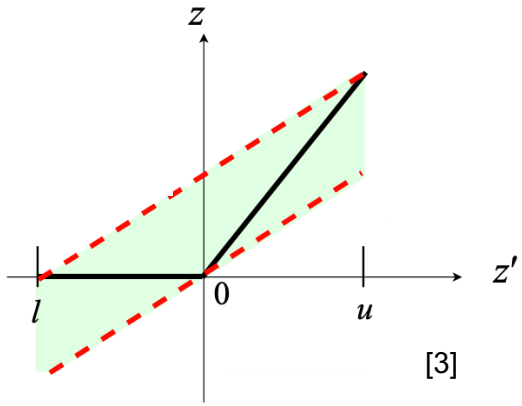




Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



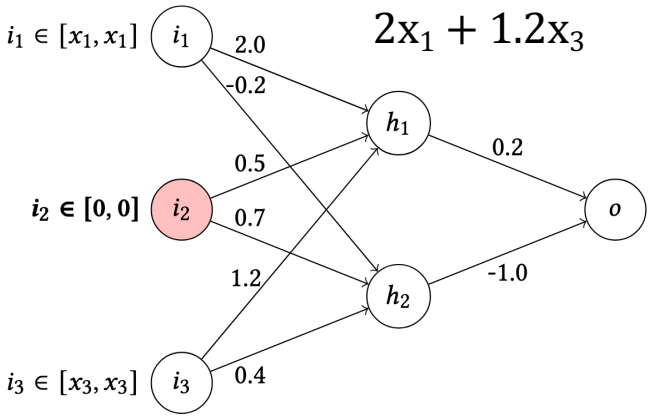


[3]

[3] Wang et al. 2018. Efficient formal safety analysis of neural networks. In 32nd International Conference on Neural Information Processing Systems.

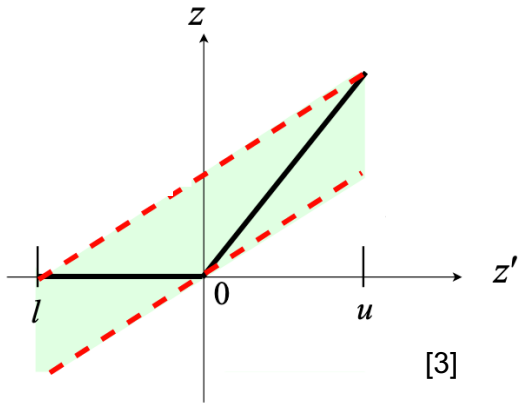
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



$2x_1 + 1.2x_3$

$[-0.128x_1 + 0.257x_3 ,$
$-0.128x_1 + 0.257x_3 + 0.643]$

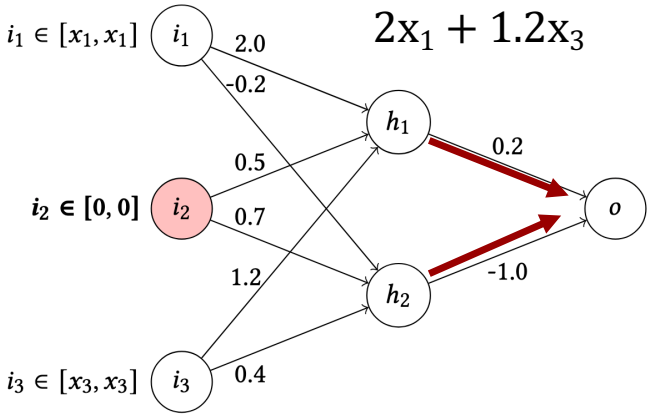


[3]

[3] Wang et al. 2018. Efficient formal safety analysis of neural networks. In 32nd International Conference on Neural Information Processing Systems.

Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



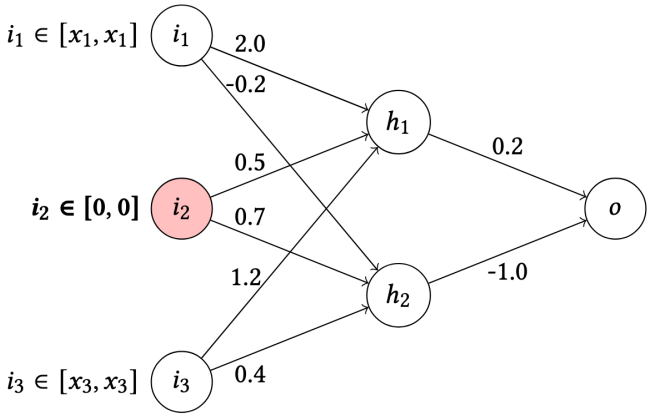
$2x_1 + 1.2x_3$

$[0.528x_1 - 0.017x_3 - 0.643,$
$0.528x_1 - 0.017x_3]$

$[-0.128x_1 + 0.257x_3 ,$
$-0.128x_1 + 0.257x_3 + 0.643]$

Subroutine 1: Symbolic Forward Analysis

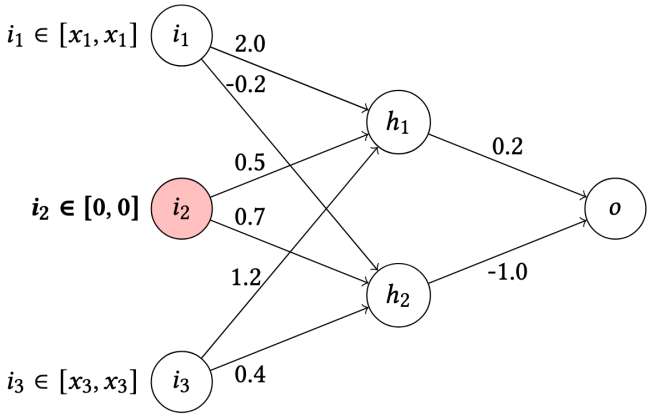
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



$[0.528x_1 - 0.017x_3 - 0.643,$
$0.528x_1 - 0.017x_3]$

Subroutine 1: Symbolic Forward Analysis

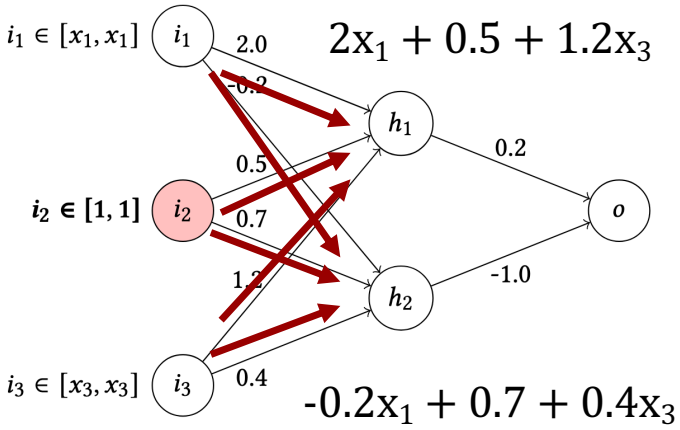
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



$O = [-0.2, 2.64]$

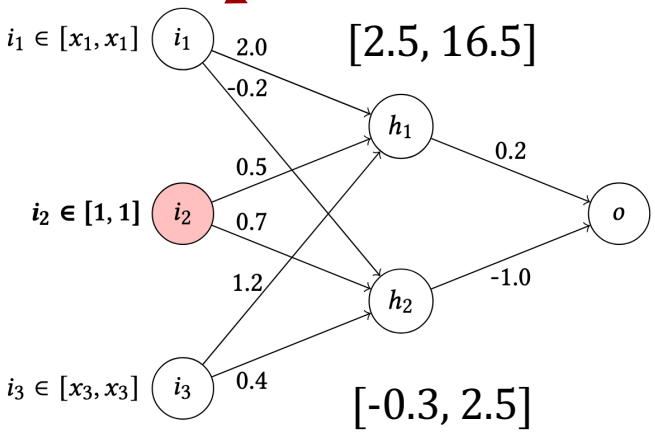
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



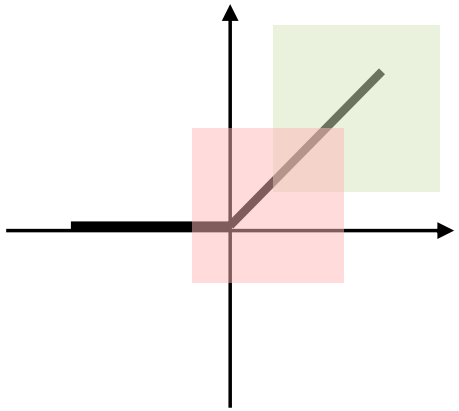
Subroutine 1: Symbolic Forward Analysis

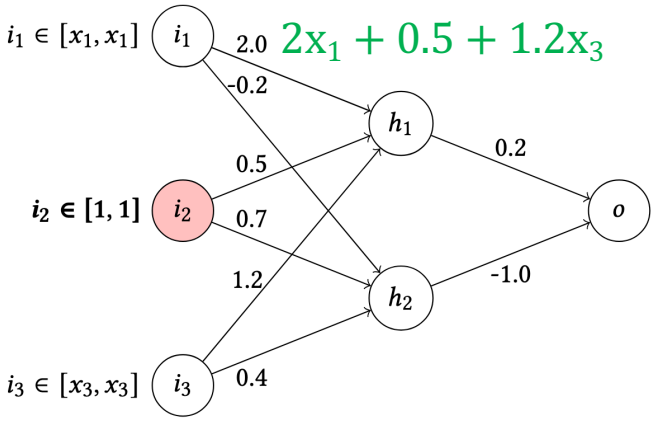
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



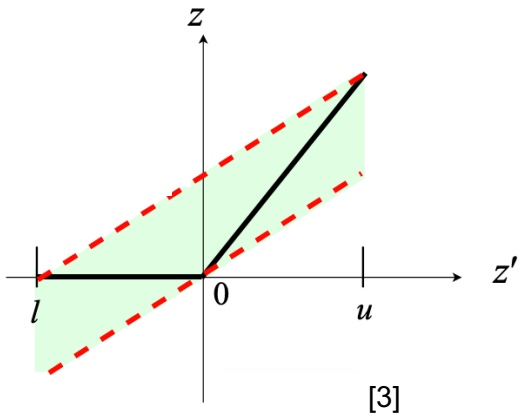
Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$





$[-0.178x_1 + 0.357x_3 + 0.625,$
$-0.178x_1 + 0.357x_3 + 0.893]$

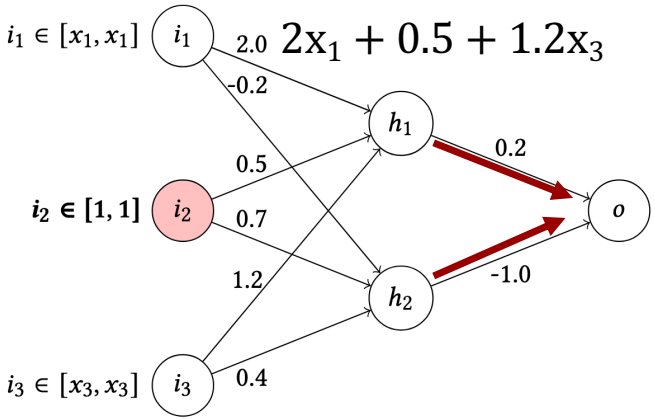


[3]

[3] Wang et al. 2018. Efficient formal safety analysis of neural networks. In 32nd International Conference on Neural Information Processing Systems.

Subroutine 1: Symbolic Forward Analysis

Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



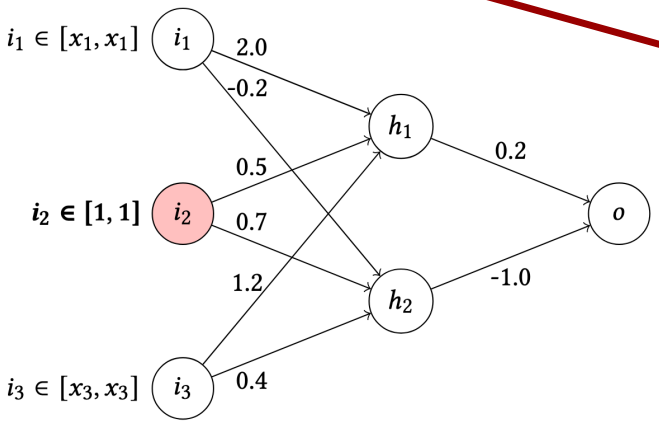
$2x_1 + 0.5 + 1.2x_3$

$[0.578x_1 - 0.117x_3 - 0.793,$
$0.578x_1 - 0.017x_3 - 0.525]$

$[-0.178x_1 + 0.357x_3 + 0.625,$
$-0.178x_1 + 0.357x_3 + 0.893]$

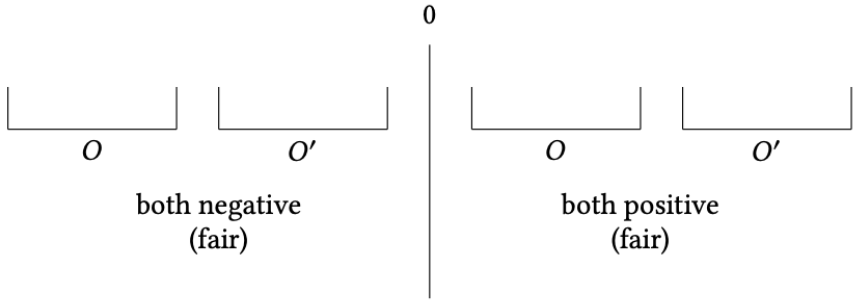
Subroutine 1: Symbolic Forward Analysis

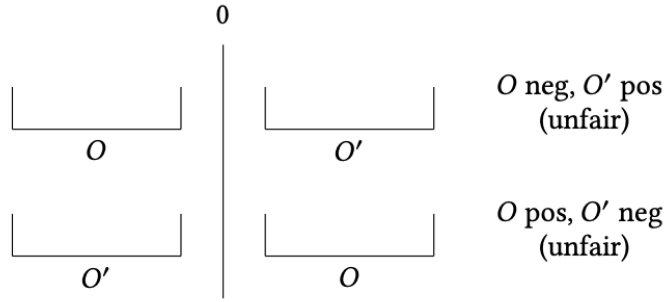
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$



$O' = [-0.8, 2.368]$

Subroutine 1: Symbolic Forward Analysis





#1: $O = [-2, -1]$ $O' = [-3, -2]$

#2: $O = [1, 2]$ $O' = [3, 4]$

Fair

#3: $O = [-2, -1]$ $O' = [3, 4]$

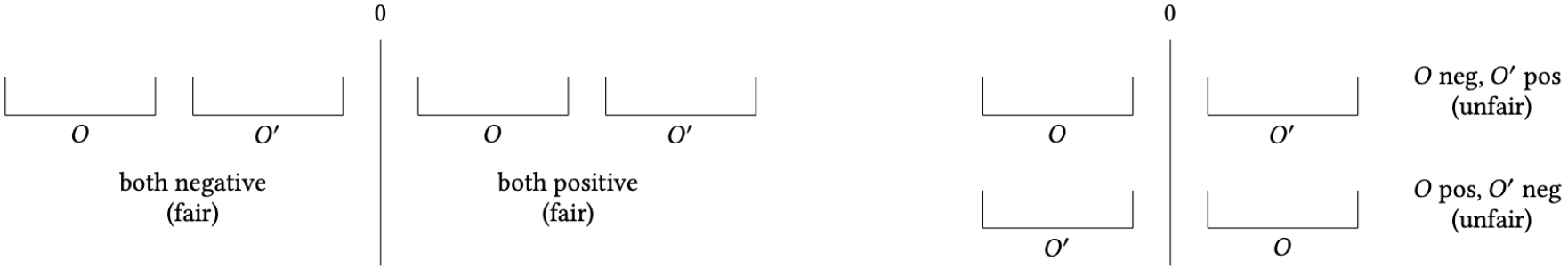
#4: $O' = [-2, -1]$ $O = [3, 4]$

Unfair

$O = [-1, 2]$?

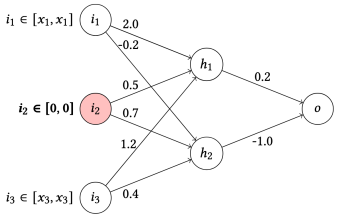
Undecided

Subroutine 1: Symbolic Forward Analysis

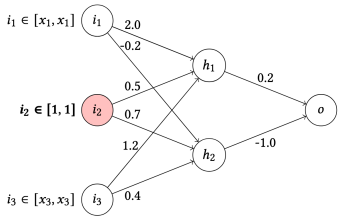


Our Example

O = [-0.2, 2.64]



O’ = [-0.8, 2.368]



Undecided

Our Method

FairQuant

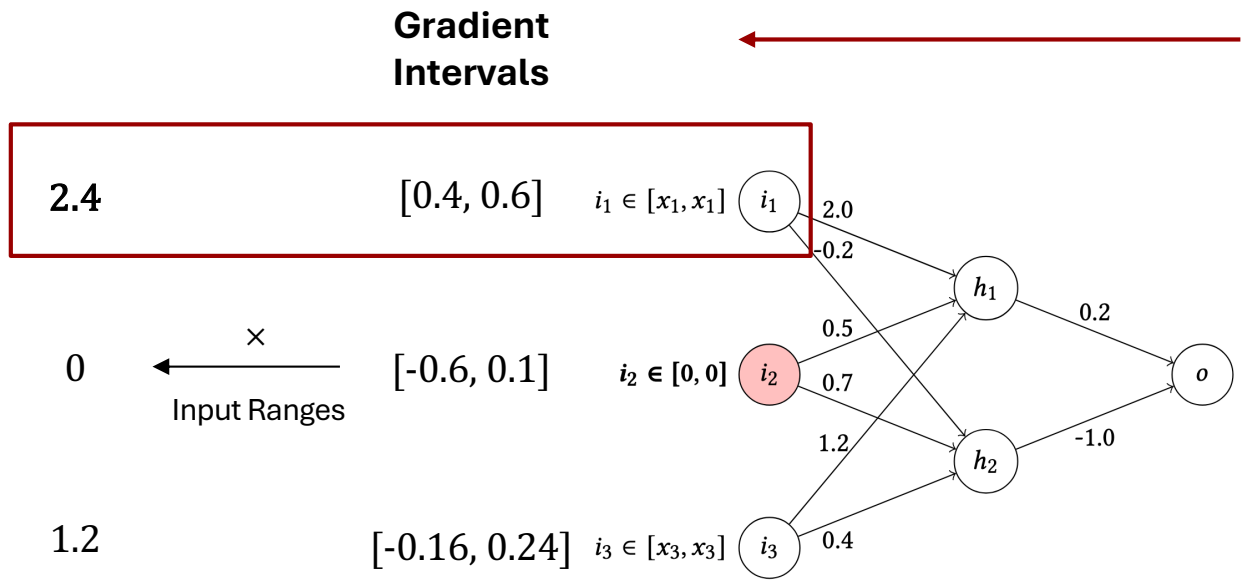
Abstraction	Symbolic Forward Analysis
Refinement	Iterative Backward Analysis
Quantification	Fairness Rate Calculation

Subroutine 2: Iterative Backward Analysis

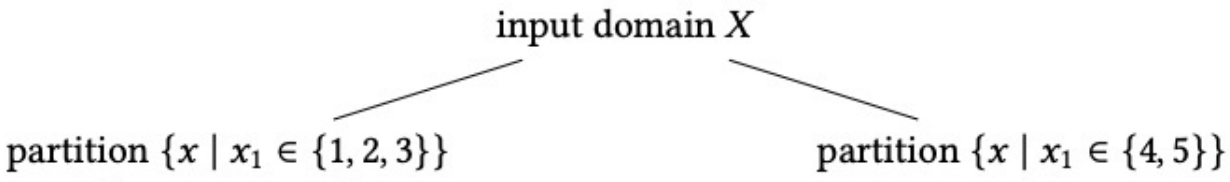
Input: $I(x_1) = I'(x_1) = [1, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

Choose the "most significant" input feature

Small disjoint partitions → More precise forward analysis



Subroutine 2: Iterative Backward Analysis



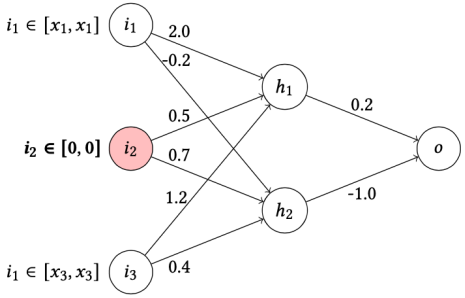
Our Method

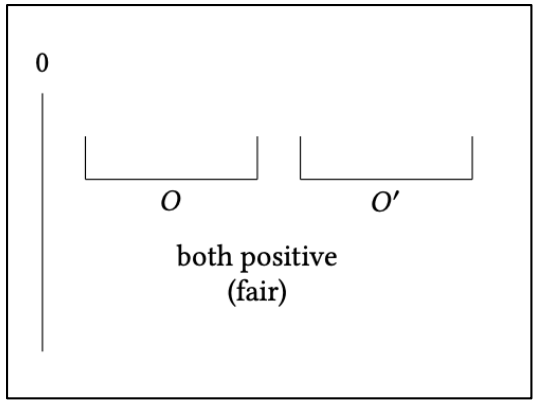
FairQuant

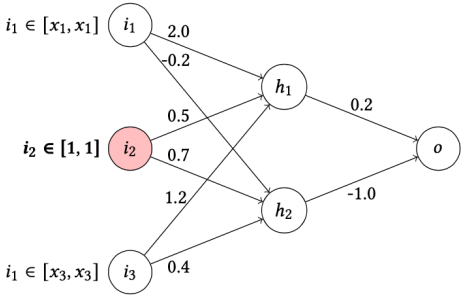
Abstraction	Symbolic Forward Analysis
Refinement	Iterative Backward Analysis
Quantification	Fairness Rate Calculation

Subroutine 3: Fairness Rate Calculation

Input: $I(x_1) = I'(x_1) = [4, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$







$O = [1.49, 2.65]$

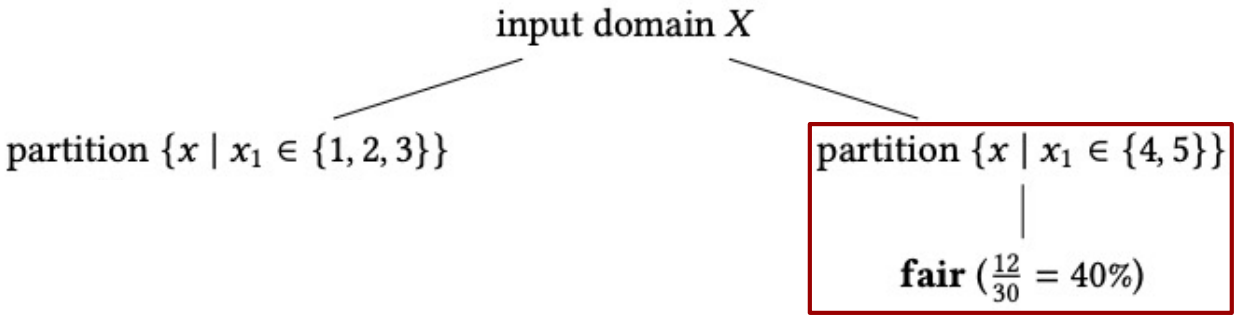
$O' = [1.0, 2.36]$

Subroutine 3: Fairness Rate Calculation

Input: $I(x_1) = I'(x_1) = [4, 5]$ $I(x_2) = 0$ $I'(x_2) = 1$ $I(x_3) = I'(x_3) = [0, 5]$

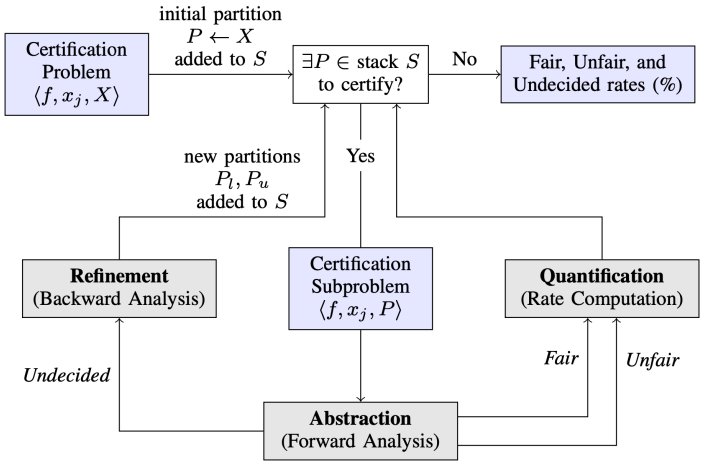
of (x, x') pairs in this partition P $= 12$
of (x, x') pairs in the domain X $= 30$

Certified rate $+= 40\%$
Undecided rate $-= 40\%$



Overview

1. Our Method



2. Evaluation

Dataset	DNN	Fairify [12]					FairQuant (new)				
		Time	Cex	#Cex	Cer%	Fal%	Und%	Time	Cex	#Cex	Cer%	Fal%	Und%
Bank	BM-1	30m	✓	11	10.00	0	90.00	4.82s	✓	2820	94.23	0	5.76
	BM-2	31m	✓	28	16.07	0	83.93	3.23s	✓	2479	93.41	0	6.58
	BM-3	31m	✓	27	19.60	0	81.40	1.21s	✓	1864	95.69	0	4.30
	BM-4	35m	✓	4	3.72	0	96.18	71.12s	✓	5135	87.03	0	12.96
	BM-5	23m	✓	114	77.25	0	22.75	1.03s	✓	1474	96.27	0	3.72
	BM-6	12m	✓	155	69.41	0	30.59	0.44s	✓	1426	96.44	0	3.55
	BM-7	30m	✓	57	9.41	0	90.59	12.26s	✓	7017	83.65	0	16.34
	BM-8	30m	✓	1	0.98	0	99.02	18.99s	✓	3074	90.75	0	9.24
German	GC-1	32m	✓	22	0	0	100	9.73s	✓	31585	32.67	0	67.33
	GC-2	33m	✓	6	0	0	100	31.72s	✓	32655	42.21	0	57.79
	GC-3	8m	✓	194	2.98	0	97.02	6.77s	✓	25963	58.44	0	41.55
	GC-4	4m	✓	2	99.00	0	1.00	0.29s	✓	77	99.65	0	0.34
	GC-5	30m	✗	0	0	0	100	1.24s	✓	9	99.80	0	0.19
Adult	AC-1	32m	✓	3	0.03	0	99.97	3.23s	✓	6151	90.68	0	9.31
	AC-2	31m	✓	9	0.01	0	99.99	30.04s	✓	13008	79.93	0	20.06
	AC-3	32m	✓	20	0	0	100	37.12s	✓	60494	33.29	0	66.70
	AC-4	36m	✗	0	0	0	100	8m	✓	61324	24.79	0	75.20
	AC-5	33m	✗	0	0	0	100	4m	✓	71012	19.12	0	80.87
	AC-6	33m	✓	4	0.01	0	99.99	10.20s	✓	31593	58.82	0	41.17
	AC-7	30m	✗	0	0.01	0	99.99	4m	✓	25588	31.72	0	68.27
	AC-8	30m	✓	39	0.03	0	99.97	11.18s	✓	26179	66.50	0	33.49
	AC-9	30m	✓	126	0.64	0	99.36	3.50s	✓	5470	91.13	0	8.86
	AC-10	32m	✓	8	0.03	0	99.97	5.01s	✓	9033	87.65	0	12.34
	AC-11	30m	✗	0	0	0	100	36.44s	✓	24516	58.01	0	41.98
	AC-12	30m	✗	0	0.02	0	99.98	0.91s	✓	8824	70.82	0	29.17
Compas	compas-1	17m	✓	2	80.00	0.32	19.68	0.01s	✓	17	97.27	2.72	0
	compas-2	31m	✗	0	0	0	100	0.01s	✓	15	97.59	2.40	0
	compas-3	30m	✗	0	0	0	100	0.30s	✓	12	98.07	1.92	0
	compas-4	30m	✗	0	0	0	100	0.01s	✓	14	97.75	2.24	0
	compas-5	T/O	✗	0	0	0	100	5.24s	✓	11	98.23	1.76	0
	compas-6	M/O	✗	0	0	0	100	9.19s	✓	12	98.07	1.92	0
	compas-7	M/O	✗	0	0	0	100	101.25s	✓	15	97.59	2.40	0

Evaluation

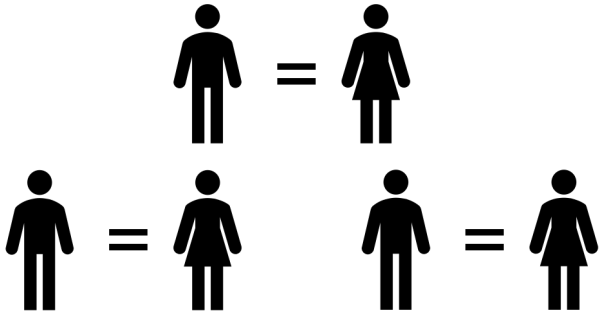
FairQuant vis-a-vis **Fairify** [1]

32 neural networks

- 25 existing networks from Fairify ≤ 300 neurons
- 7 newly trained networks up to 10000 neurons

Fairness datasets

1. Bank Marketing
2. German Credit
3. Adult Census
4. Compas Recidivism

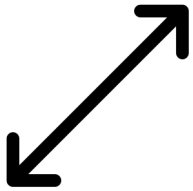


[1] Biswas and Rajan. 2023. Fairify: Fairness Verification of Neural Networks. In 45th International Conference on Software Engineering.

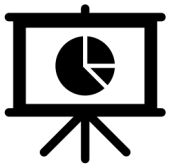
Evaluation

Is FairQuant...

1. More Accurate?



2. More Scalable?



3. More Informative?

Evaluation

1. More Accurate?

Fairify

20 / 32 networks found with a counterexample

FairQuant ✓

32/32 networks found with a counterexample

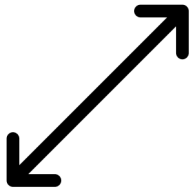
Method	BM-1	GC-1	AC-1	compas-1
Fairify	11	22	3	2
FairQuant	**2820**	**31585**	**6151**	**17**

# of counterexamples (↑) on selected models

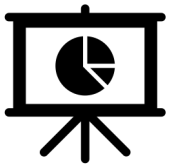
Evaluation

Is FairQuant...

1. More Accurate?



2. More Scalable?



3. More Informative?

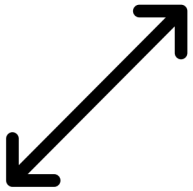
Evaluation

Fairify

Often runs until timeout
Cannot verify a single partition for 3 largest networks

FairQuant ✓

Always finishes before timeout
Often takes <1 – 100 seconds



2. More Scalable?

Method	BM-1	GC-1	AC-1	compas-1
Fairify	30 min	32 min	32 min	17 min
FairQuant	**4 sec**	**9 sec**	**3 sec**	**0.01 sec**

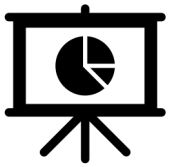
Verification runtime (↓) on selected models, T/O = 30 min

Evaluation

Is FairQuant...

1. More Accurate?

2. More Scalable?



3. More Informative?

Evaluation

Fairify

Only UNSAT partitions provide Certification

FairQuant ✓

Quantitative results for each partition

3. More Informative?

Method	BM-1	GC-1	AC-1	compas-1
Fairify	10% / 90%	0% / 100%	0.03% / 99.97%	80% / 19.68%
FairQuant	**94.23% / 5.76%**	**32.67% / 67.33%**	**90.68% / 9.31%**	**97.27% / 0%**

Percentages of Certified (↑) / Undecided (↓) rates on selected models

Evaluation

Is FairQuant...

1. More Accurate?

2. More Scalable?

3. More Informative?

Conclusion

- We introduce **FairQuant**, a novel method for verifying global individual fairness of neural networks.
- FairQuant relies on **abstract interpretation** and **iterative refinement** to achieve soundness, scalability, and accuracy.
- We provide a **quantitative** framework to measure the degree of individual fairness achieved by neural networks.

Thank You!

Any Questions?

Brian Hyeongseok Kim
brian.hs.kim@usc.edu

arXiv