CS781: A Quick Primer on Abstract Interpretation for Neural Networks

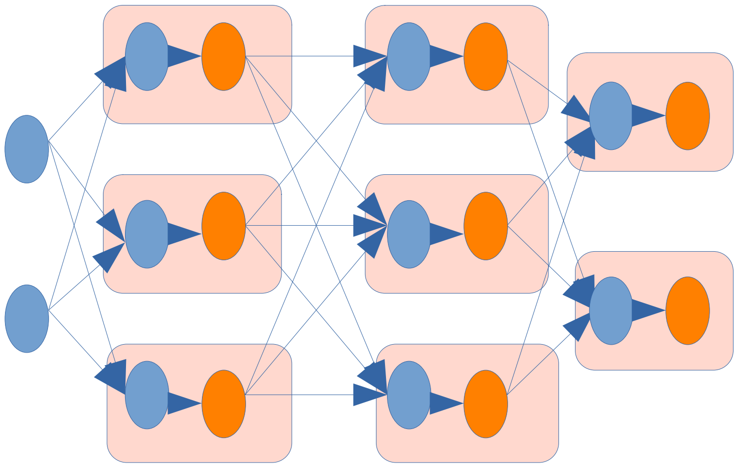
Supratik Chakraborty
IIT Bombay

A Simple Abstract Domain

Interval Abstract Domain

- Simplest domain for analyzing numerical programs
- Represent values of each variable separately using intervals
- Example:

Pre-condition on (x1, x2)



Post-condition on (x17, x18)

Represent values of inputs by intervals,

Compute values of hidden layer nodes and outputs as intervals

Interval Abstract Domain

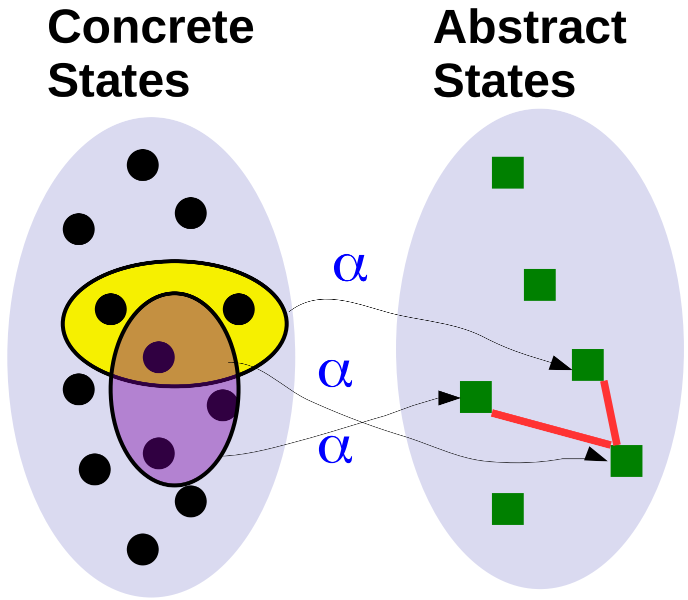
- Abstract states: intervals of values of x, (ignore values of y)
  - [-10, 7]: { (x, y) | -10 <= x <= 7 }
  - (-∞, 20]: { (x, y) | x <= 20 }

- $\sqsubseteq$ relation: Inclusion of intervals
  - [-10, 7] $\sqsubseteq$ [-20, 9]
- $\sqcup$ and $\sqcap$: union and intersection of intervals
  - [-10, 9] $\sqcup$ [-20, 7] = [-20, 9]
  - [-10, 9] $\sqcap$ [-20, 7] = [-10, 7]
- $\perp$ is empty interval of x
- $\top$ is (-∞, +∞)

Interval Abstract Domain

- Abstract states: intervals of values of x, (ignore values of y)
  - [-10, 7]: { (x, y) | -10 <= x <= 7 }
  - (-∞, 20]: { (x, y) | x <= 20 }

- $\sqsubseteq$ relation: Inclusion of intervals
  - [-10, 7] $\sqsubseteq$ [-20, 9]
- $\sqcup$ and $\sqcap$: union and intersection
  - [-10, 9] $\sqcup$ [-20, 7] = [-20, 9]
  - [-10, 9] $\sqcap$ [-20, 7] = [-10, 7]
- $\perp$ is empty interval of x
- $\top$ is (-∞, +∞)



α({(1, 3), (2, 4), (5, 7)}) = [1, 5]
α({(5, 7), (7, 6), (9, 10)}) = [5, 9]
α({(5, 7)}) = [5, 5]

Interval Abstract Domain

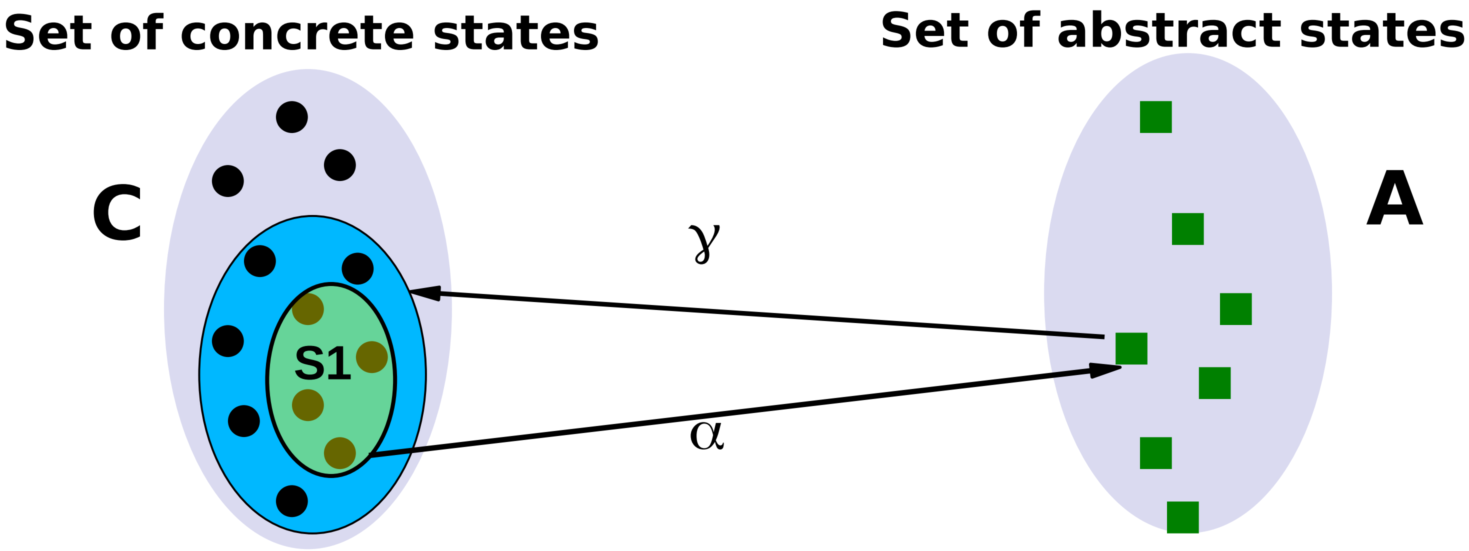
- Abstract states: pairs of intervals (one for x, y)
  - ([-10, 7] , (-1, 20])
  - $\sqsubseteq$ relation: Inclusion of intervals

    ([-10, 7] , (-1, 20]) $\sqsubseteq$ ([-20, 9], (-1, +∞))
  - $\sqcup$ and $\sqcap$: union and intersection of intervals
  - ([-10, 9] , (-1, 20]) $\sqcap$ ([-20, 7], [3,+∞)) = ([-10, 7], [3, 20])
  - ([-10, 9], (-1, 20]) $\sqcup$ ([-20, 7], [3,+∞)) = ([-20, 9],(-1,+∞))

  - $\perp$ is empty interval of x and y
  - $\top$ is ((-∞, +∞), (-∞, +∞))

Desirable Properties of α and γ

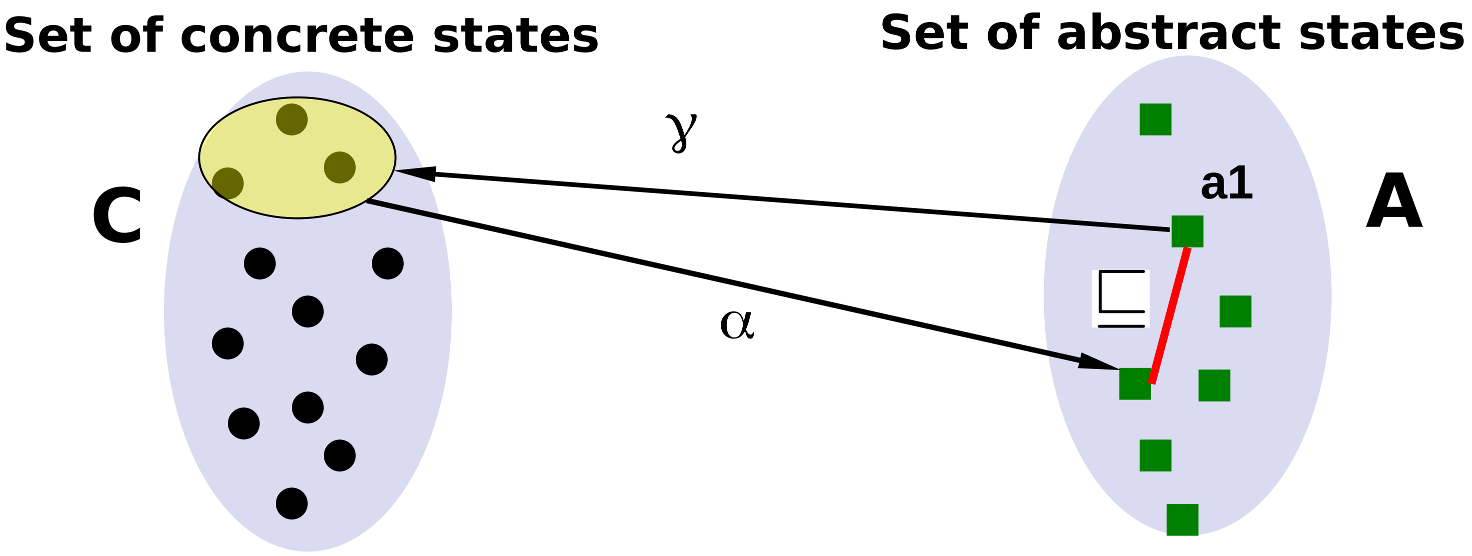
For all $S_1 \subseteq \mathcal{C}$ $\qquad S_1 \subseteq \gamma(\alpha(S_1))$



Desirable Properties of α and γ

$$S_1 \subseteq \gamma(\alpha(S_1)) \quad \text{forall} \quad S_1 \subseteq \mathcal{C}$$

$$\alpha(\gamma(a_1)) \sqsubseteq a_1 \quad \text{forall} \quad a_1 \in \mathcal{A}$$

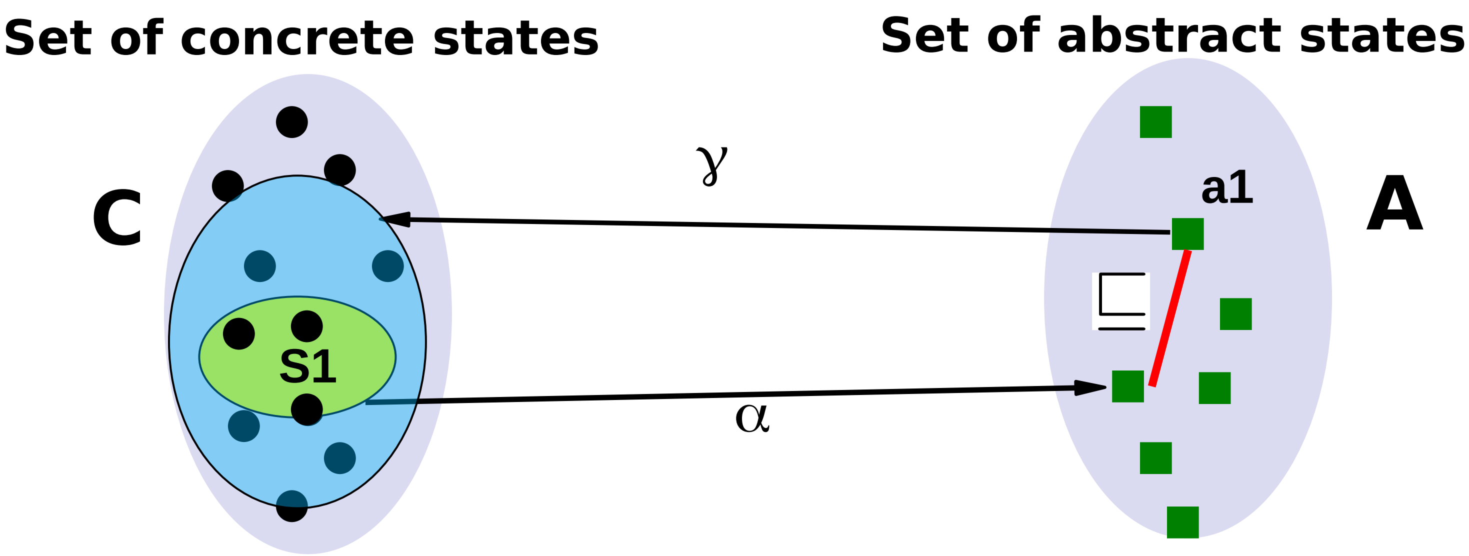


α and γ form a Galois connection

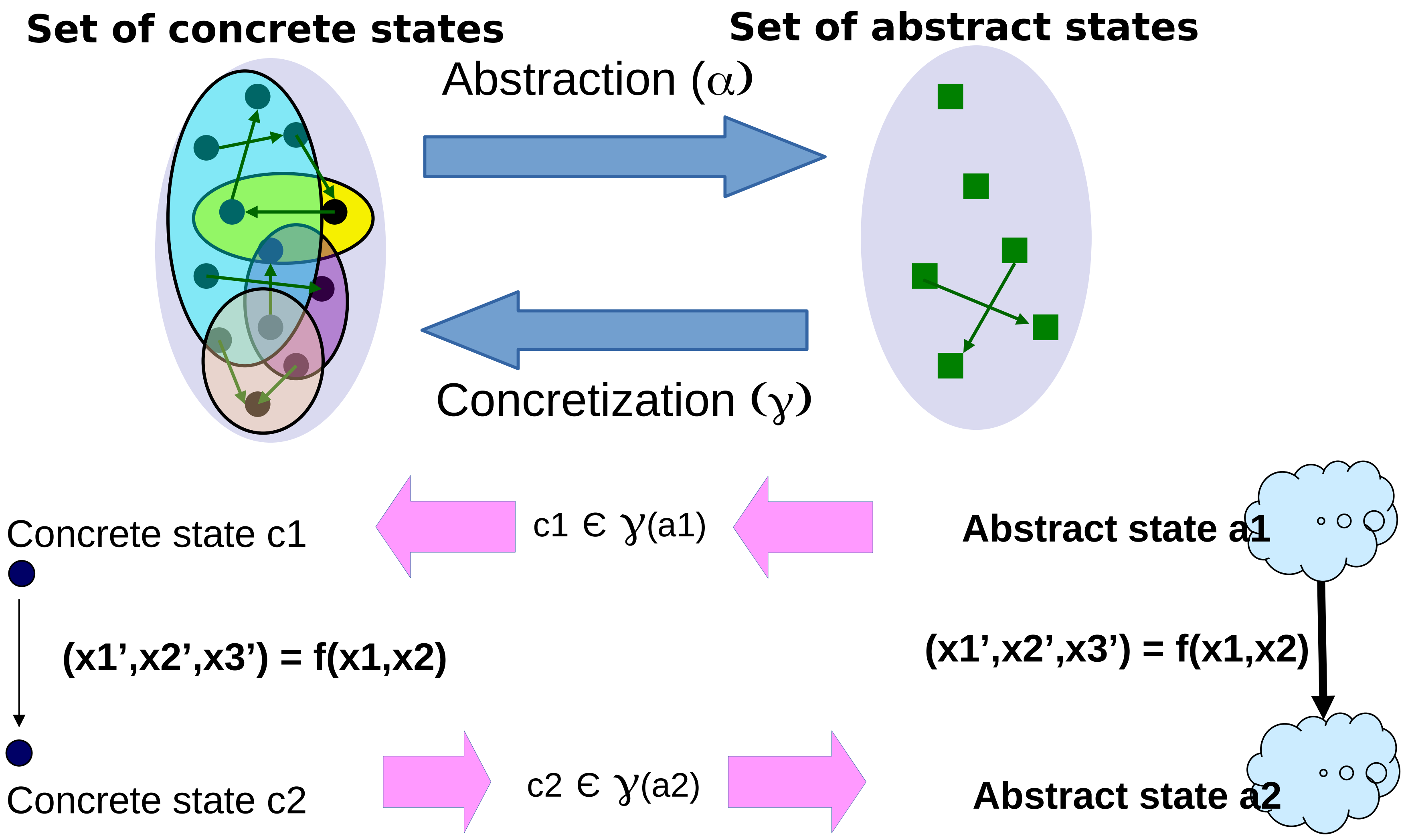
Desirable Properties of α and γ

- α and γ form a Galois connection
  - Second (equivalent) view:

$$\alpha(S_1) \sqsubseteq a_1 \Leftrightarrow S_1 \subseteq \gamma(a_1) \text{ for all } S_1 \subseteq S,\ a_1 \in \mathcal{A}$$



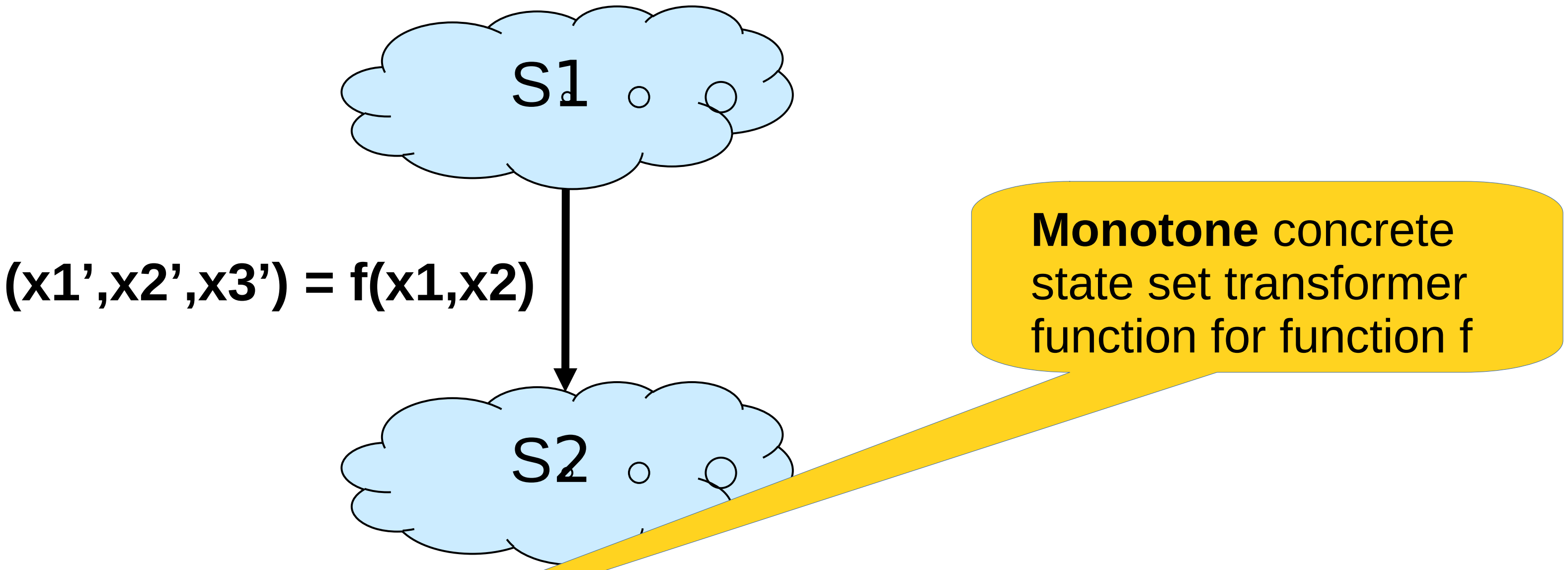
Computing Abstract State Transitions



Computing Abstract State Transitions

- Concrete state set transformer function
  - Example:

$S1 = \{ (x1, x2, x3) \mid \ldots\ldots \}$: set of concr. states



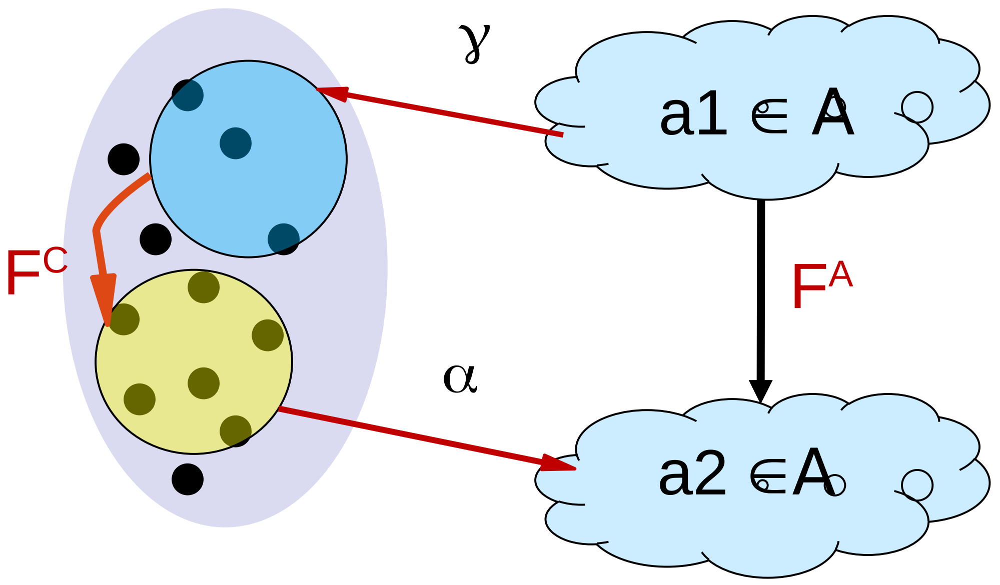
$S2 = \{(x1', x2', x3') \mid \exists\ (x1, x2, x3) \in S1,\ (x1',x2',x3') = f(x1,x2)\}$

$= F^C(S1)$: set of concrete states

Computing Abstract State Transitions

- Abstract state transformer function
  - Example:

Set of concrete states



$a2 = \alpha(F^C (\gamma\ (a1)))$ ideally, but $F^A(a1) \sqsupseteq \alpha(F^C (\gamma\ (a1)))$ often used

Summary

- Abstract interpretation is a general framework for analysis of state transition systems
- Widely used for verification and static analysis of programs
- Recent applications in neural network analysis
- Choice of right abstraction crucial to success
  - Balance between precision and efficiency

This lecture should help you understand the paper "An Abstract Domain for Certifying Neural Networks" by Singh et al. better