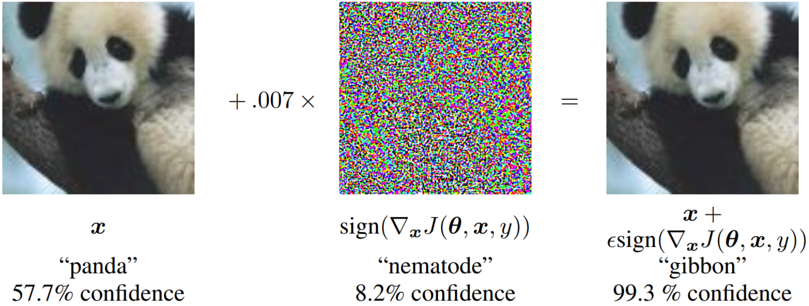
Zonotope verification of Monotone operator equilibrium models

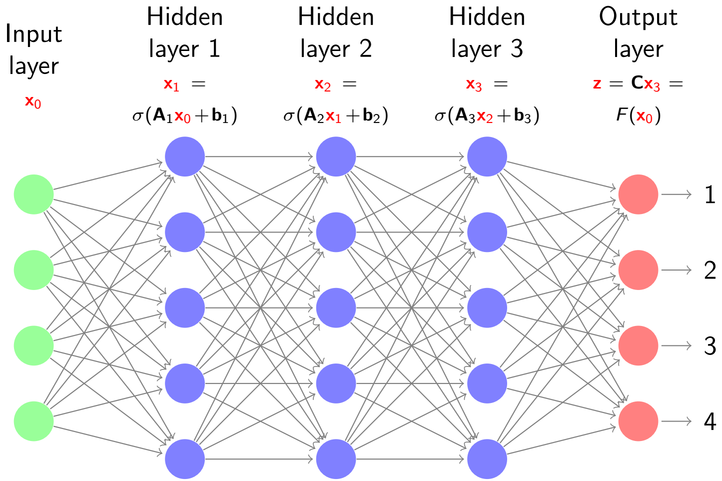
http://arxiv.org/abs/2110.08260

Robustness of neural network



Adversarial example of neural network, Ian Goodfellow et al., 2015.

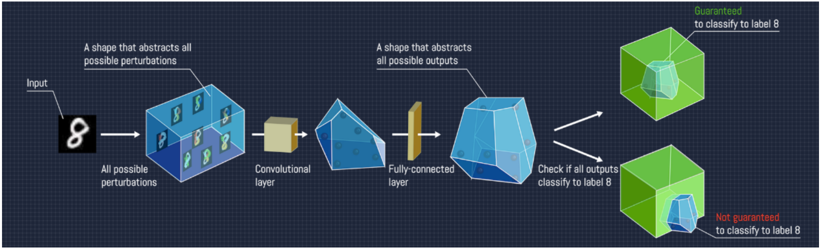
Deep neural networks (DNNs)



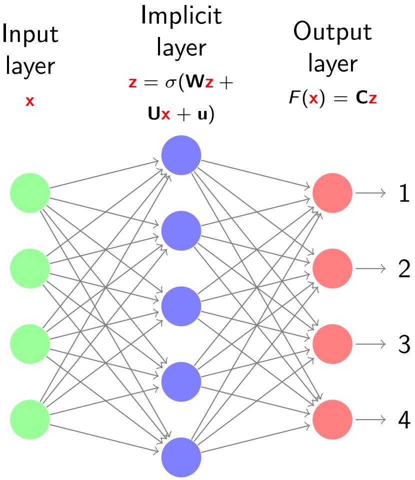
Fully-connected DNN with activation function σ.

Zonotope verificaiton of DNNs

Input $\rightarrow$ Transformation $\rightarrow$ Approximation $\rightarrow$ Output.



Monotone operator equilibrium models (monDEQs)



Fully-connected monDEQ with activation function σ, $\mathbf{I} - \mathbf{W}$ is strongly monotone.

Splitting methods for fixed-point iteration

Fixed-point equation: $\mathbf{z} = \sigma(\mathbf{W}\mathbf{z} + \mathbf{U}\mathbf{x} + \mathbf{u})$

- Forward-Backward Splitting (FB): $\mathbf{z}_0 = 0$,

$$\mathbf{z}_{n+1} = g_\alpha^{FB}(\mathbf{x}, \mathbf{z}_n) = \sigma((1-\alpha)\mathbf{z}_n + \alpha(\mathbf{W}\mathbf{z}_n + \mathbf{U}\mathbf{x} + \mathbf{u}))$$

  converges if $0 < \alpha < 2m/\|\mathbf{I} - \mathbf{W}\|_2^2$.

- Peaceman-Rachford Splitting (PR): $\mathbf{z}_0 = \mathbf{u}_0 = 0$,

$$\begin{aligned}
\mathbf{u}'_{n+1} &= 2\mathbf{z}_n - \mathbf{u}_n \\
\mathbf{z}'_{n+1} &= (\mathbf{I} + \alpha(\mathbf{I} - \mathbf{W}))^{-1}(\mathbf{u}'_{n+1} + \alpha(\mathbf{U}\mathbf{x} + \mathbf{u})) \\
\mathbf{u}_{n+1} &= 2\mathbf{z}'_{n+1} - \mathbf{u}'_{n+1} \\
\mathbf{z}_{n+1} &= \sigma(\mathbf{u}_{n+1}) \\
[\mathbf{z}_{n+1}, \mathbf{u}_{n+1}] &= g_\alpha^{PR}(\mathbf{x}, \mathbf{z}_n, \mathbf{u}_n)
\end{aligned}$$

  converges for any $\alpha > 0$.

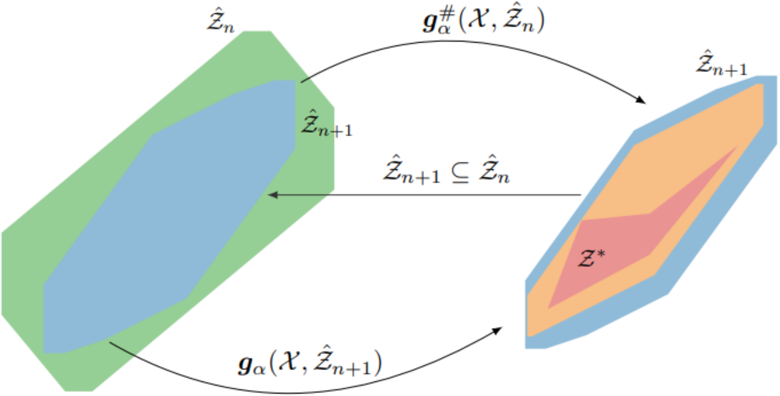
Splitting methods on sets of points

$$\text{Input} \to \overbrace{\textit{Iteration} \to \textit{Approximation}}^{\textit{repeat}} \to \text{Output.}$$

- Denote by g_α the exact iteration, $g_\alpha^\#$ the over-approximation operation.
- Denote by $\mathcal{Z}_n$ (resp. $\mathcal{U}_n$) the exact set, $\hat{\mathcal{Z}}_n$ (resp. $\hat{\mathcal{U}}_n$) the over-approximation set, $\mathcal{Z}^*$ the fixed-point set.
- **(Fixed-Point contraction)**. Let $[\hat{\mathcal{Z}}_{n+1}, \hat{\mathcal{U}}_{n+1}] = g_\alpha^\#(\mathcal{X}, \hat{\mathcal{Z}}_n, \hat{\mathcal{U}}_n)$ be closed sets over-approximating $\mathbf{z}_{n+1}$ and $\mathbf{u}_{n+1}$ obtained by applying the solver iteration $n+1$ times for some $\mathbf{z}_0, \mathbf{u}_0$ and all inputs $\mathbf{x} \in \mathcal{X}$. Then:

$$\boxed{\hat{\mathcal{Z}}_{n+1} \subseteq \hat{\mathcal{Z}}_n, \hat{\mathcal{U}}_{n+1} \subseteq \hat{\mathcal{U}}_n} \Rightarrow \boxed{\mathcal{Z}_j \subseteq \hat{\mathcal{Z}}_{n+1}, \forall j > n} \Rightarrow \boxed{\mathcal{Z}^* \subseteq \hat{\mathcal{Z}}_{n+1}}$$

Splitting methods on sets of points



Input to iteration step Output of iteration step

M-zonotope: inclusion checking and propagation

Mixed(M)-zonotope = zonotope + hyper-box + center:

$$\hat{\mathcal{Z}} = \mathbf{A}\mathbf{x} + \mathrm{diag}(\mathbf{b})\mathbf{y} + \mathbf{c} \subseteq \mathbb{R}^p$$

- zonotope $\mathbf{A}\mathbf{x}$: $\mathbf{A} \in \mathbb{R}^{p\times k}, \mathbf{x} \in [-1,1]^k$.
- hyper-box $\mathrm{diag}(\mathbf{b})\mathbf{y}$: $\mathbf{b} \in \mathbb{R}^p_+, \mathbf{y} \in [-1,1]^k$.
- center $\mathbf{c} \in \mathbb{R}^p$.

	Box	Zonotope	M-zonotope
precision	☹	☺	☺
Inclusion checking	☺	☹	☺

SemiSDP v.s. Zonotope

- First 100 test examples of MNIST dataset.
- ε: range of perturbation; n: number of successfully certified examples; t: average computation time.
- SemiSDP: SDP-based method by (Chen et al., 2021); CRAFT: zonotope-based method by (Müller et al., 2021).

ε	SemiSDP		CRAFT	
	n	$t(s)$	n	$t(s)$
0.10	0	1350	0	**9.75**
0.05	24	1350	30	**15.75**
0.01	99	1350	99	**1.4**