Neural Arithmetic Units

By Andreas Madsen and Alexander Rosenberg Johansen
@andreas_madsen @AlexRoseJo

ICLR 2020, spotlight awarded paper

paper: https://openreview.net/forum?id=H1gNOeHKPS
code: https://github.com/AndreasMadsen/stable-nalu

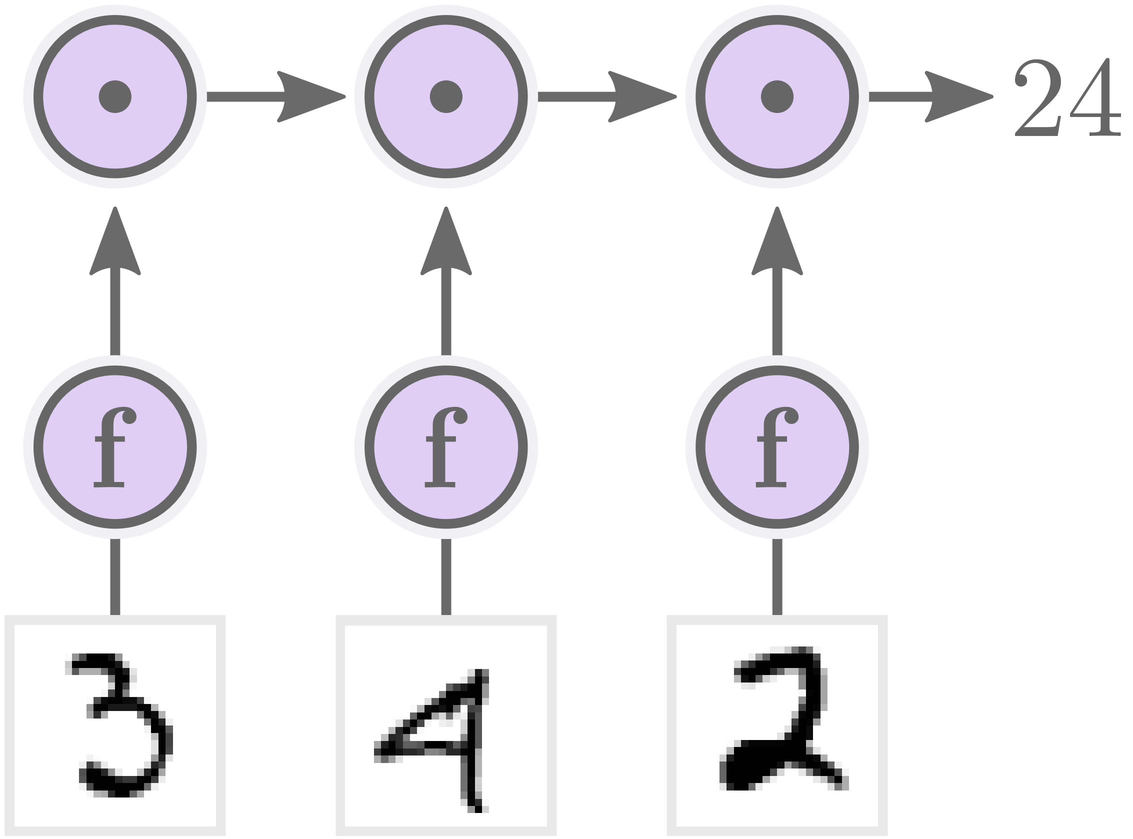
Arithmetic Extrapolation

- Neural networks are great at interpolation but can rarely extrapolate.
- Arithmetic extrapolation assumes there is an underlying function partially composed of simple arithmetics.
- Simple arithmetics might occur in:
    - Physical modelling
    - Financial modelling
    - NLP Q&A tasks

Direct arithmetic example:

$$t = (x_1 + x_2) \cdot (x_1 + x_2 + x_3 + x_4) \text{ for } x \in \mathbb{R}^4$$

MNIST example:



NALU by Andrew Trask, et. al

- Central idea: Learn the underlying function, by being able to represent it exactly and having constrained (and biased) weights.
- NALU has 3 components[1]:
  - Addition/Subtraction component
  - Multiplication/Division component
  - Gating component

σ controls scale (0, 1)

tanh controls sign (-1, 1)

$$W_{h_\ell, h_{\ell-1}} = \tanh(\hat{W}_{h_\ell, h_{\ell-1}})\sigma(\hat{M}_{h_\ell, h_{\ell-1}})$$

$$\text{NAC}_+ : \ z_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell, h_{\ell-1}} z_{h_{\ell-1}}$$

to avoid issues with log(·)

$$\text{NAC}_\bullet : \ z_{h_\ell} = \exp\left(\sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell, h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon)\right)$$

$$g_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} G_{h_\ell, h_{\ell-1}} z_{h_{\ell-1}}$$

$$\text{NALU} : \ z_{h_\ell} = g_{h_\ell} \odot \text{NAC}_+(\mathbf{z}_{\ell-1})_{h_\ell} + (1 - g_{h_\ell}) \odot \text{NAC}_\bullet(\mathbf{z}_{\ell-1})_{h_\ell}$$

gating choose which operation to use

1) Neural Arithmetic Logic Units (NALU), by Andrew Trask et. al, NeurIPS 2018.

Measuring Arithmetic Extrapolation

To extrapolate in an arithmetic task, the exact solution must be found.

In our NeurIPS workshop paper, we therefore argue consistency is a primary concern [2].

2) Measuring Arithmetic Extrapolation Performance by Andreas Madsen and Alexander R. Johansen. SEDL at NeurIPS 2019.

Direct arithmetic example:

$$t = (x_1 + x_2) \cdot (x_1 + x_2 + x_3 + x_4) \text{ for } x \in \mathbb{R}^4$$

Op	Model	Success	Solved at iteration step		Sparsity error
		Rate	Median	Mean	Mean
$\times$	NAC$_\bullet$	$13\%\,^{+8\%}_{-5\%}$	$5.5 \cdot 10^4$	$5.9 \cdot 10^4\,^{+7.8\cdot10^3}_{-6.6\cdot10^3}$	$7.5 \cdot 10^{-6}\,^{+2.0\cdot10^{-6}}_{-2.0\cdot10^{-6}}$
	NALU	$26\%\,^{+9\%}_{-8\%}$	$7.0 \cdot 10^4$	$7.8 \cdot 10^4\,^{+6.2\cdot10^3}_{-8.6\cdot10^3}$	$9.2 \cdot 10^{-6}\,^{+1.7\cdot10^{-6}}_{-1.7\cdot10^{-6}}$
	NMU	$\mathbf{94\%}\,^{+3\%}_{-6\%}$	$\mathbf{1.4 \cdot 10^4}$	$\mathbf{1.4 \cdot 10^4}\,^{+2.2\cdot10^2}_{-2.1\cdot10^2}$	$\mathbf{2.6 \cdot 10^{-8}}\,^{+6.4\cdot10^{-9}}_{-6.4\cdot10^{-9}}$

Analysis of issues

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_{\bullet}$
  - $\mathrm{NAC}_{\bullet}$ can’t be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Weight issues

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_{\bullet}$
  - $\mathrm{NAC}_{\bullet}$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Expectation of the gradient w.r.t. **M**:

$$E\left[\frac{\partial \mathcal{L}}{\partial \hat{M}_{h_{\ell-1},h_\ell}}\right] = E\left[\frac{\partial \mathcal{L}}{\partial W_{h_{\ell-1},h_\ell}}\right] E\left[\tanh(\hat{W}_{h_{\ell-1},h_\ell})\right] E\left[\sigma'(\hat{M}_{h_{\ell-1},h_\ell})\right] = 0$$

Weight issues

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_\bullet$
  - $\mathrm{NAC}_\bullet$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Expectation of the gradient w.r.t. **M**:

$$E\left[\frac{\partial \mathcal{L}}{\partial \hat{M}_{h_{\ell-1},h_\ell}}\right] = E\left[\frac{\partial \mathcal{L}}{\partial W_{h_{\ell-1},h_\ell}}\right] E\left[\tanh(\hat{W}_{h_{\ell-1},h_\ell})\right] E\left[\sigma'(\hat{M}_{h_{\ell-1},h_\ell})\right] = 0$$

Evaluate sparsity error on valid solutions:

$$E_{\text{sparsity}} = \max_{h_{\ell-1},h_\ell} \min(|W_{h_{\ell-1},h_\ell}|, |1 - |W_{h_{\ell-1},h_\ell}||)$$

Op	Model	Success Rate	Solved at iteration step: Median	Solved at iteration step: Mean	Sparsity error: Mean
	NAC_+	$\mathbf{100\%}~^{+0\%}_{-4\%}$	$2.5 \cdot 10^5$	$4.9 \cdot 10^5~^{+5.2 \cdot 10^4}_{-4.5 \cdot 10^4}$	$2.3 \cdot 10^{-1}~^{+6.5 \cdot 10^{-3}}_{-6.5 \cdot 10^{-3}}$
	Linear	$\mathbf{100\%}~^{+0\%}_{-4\%}$	$6.1 \cdot 10^4$	$\mathbf{6.3 \cdot 10^4}~^{+2.5 \cdot 10^3}_{-3.3 \cdot 10^3}$	$2.5 \cdot 10^{-1}~^{+3.6 \cdot 10^{-4}}_{-3.6 \cdot 10^{-4}}$
$+$	NALU	$14\%~^{+8\%}_{-5\%}$	$1.5 \cdot 10^6$	$1.6 \cdot 10^6~^{+3.8 \cdot 10^5}_{-3.3 \cdot 10^5}$	$1.7 \cdot 10^{-1}~^{+2.7 \cdot 10^{-2}}_{-2.5 \cdot 10^{-2}}$
	NAU	$\mathbf{100\%}~^{+0\%}_{-4\%}$	$\mathbf{1.8 \cdot 10^4}$	$3.9 \cdot 10^5~^{+4.5 \cdot 10^4}_{-3.7 \cdot 10^4}$	$\mathbf{3.2 \cdot 10^{-5}}~^{+1.3 \cdot 10^{-5}}_{-1.3 \cdot 10^{-5}}$
	NAC_+	$\mathbf{100\%}~^{+0\%}_{-4\%}$	$9.0 \cdot 10^3$	$3.7 \cdot 10^5~^{+3.8 \cdot 10^4}_{-3.8 \cdot 10^4}$	$2.3 \cdot 10^{-1}~^{+5.4 \cdot 10^{-3}}_{-5.4 \cdot 10^{-3}}$
	Linear	$7\%~^{+7\%}_{-4\%}$	$3.3 \cdot 10^6$	$1.4 \cdot 10^6~^{+7.0 \cdot 10^5}_{-6.1 \cdot 10^5}$	$1.8 \cdot 10^{-1}~^{+7.2 \cdot 10^{-2}}_{-5.8 \cdot 10^{-2}}$
$-$	NALU	$14\%~^{+8\%}_{-5\%}$	$1.9 \cdot 10^6$	$1.9 \cdot 10^6~^{+4.4 \cdot 10^5}_{-4.5 \cdot 10^5}$	$2.1 \cdot 10^{-1}~^{+2.2 \cdot 10^{-2}}_{-2.2 \cdot 10^{-2}}$
	NAU	$\mathbf{100\%}~^{+0\%}_{-4\%}$	$\mathbf{5.0 \cdot 10^3}$	$\mathbf{1.6 \cdot 10^5}~^{+1.7 \cdot 10^4}_{-1.6 \cdot 10^4}$	$\mathbf{6.6 \cdot 10^{-2}}~^{+2.5 \cdot 10^{-2}}_{-1.9 \cdot 10^{-2}}$

Neural Addition Unit

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_\bullet$
  - $\mathrm{NAC}_\bullet$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Our solution, **N**eural **A**ddition **U**nit:

clipped linear weights

$$W_{h_{\ell-1},h_\ell} = \min(\max(W_{h_{\ell-1},h_\ell}, -1), 1),$$

$$\mathcal{R}_{\ell,\text{sparse}} = \frac{1}{H_\ell \cdot H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \min\left(|W_{h_{\ell-1},h_\ell}|, 1 - |W_{h_{\ell-1},h_\ell}|\right)$$

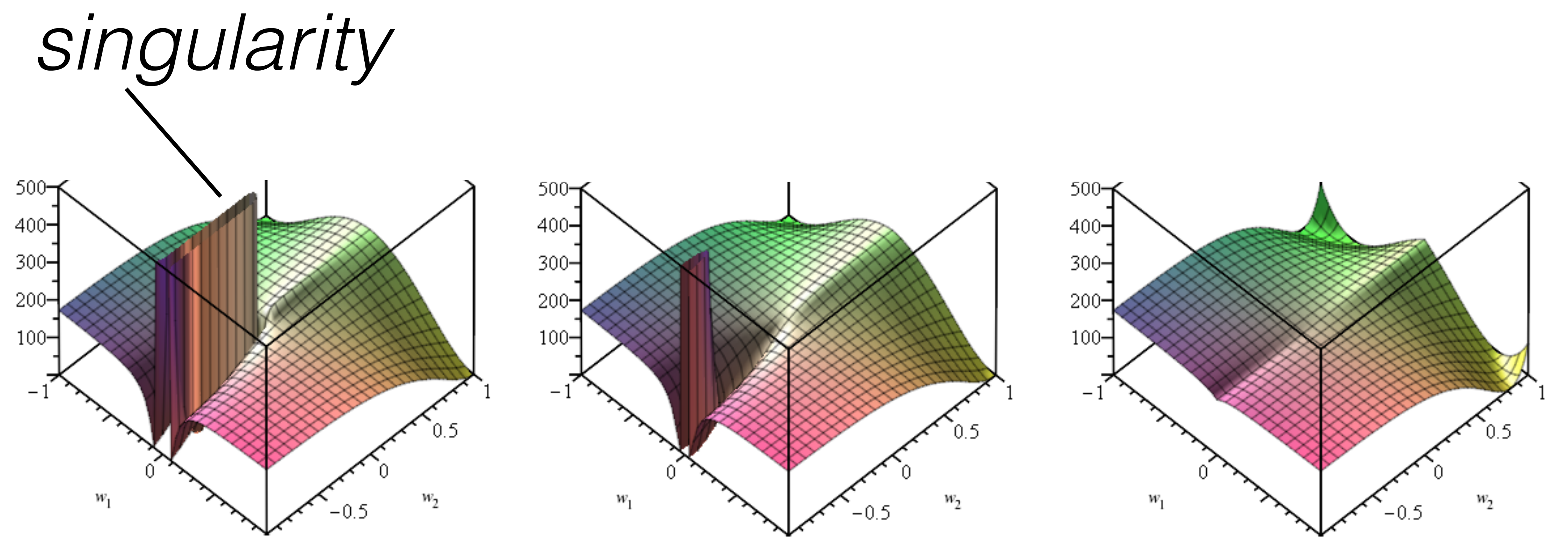
sparsity regularizer

$$\text{NAU}: \; z_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} z_{h_{\ell-1}}$$

Multiplication issues

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_{\bullet}$
  - $\mathrm{NAC}_{\bullet}$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Division causes a singularity in the loss curvature



(a) $\mathrm{NAC}_{\bullet}$ with $\epsilon = 10^{-7}$ (b) $\mathrm{NAC}_{\bullet}$ with $\epsilon = 0.1$ (c) $\mathrm{NAC}_{\bullet}$ with $\epsilon = 1$

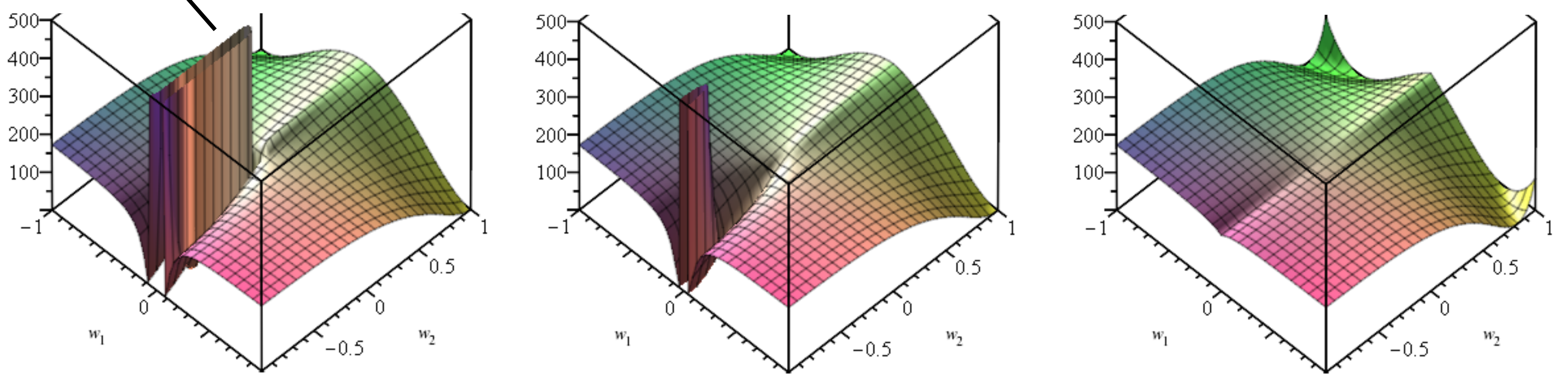
Figure 2: RMS loss curvature for a NAC_{+} unit followed by a $\mathrm{NAC}_{\bullet}$. The weight matrices are constrained to $\mathbf{W}_1 = \begin{bmatrix} w_1 & w_1 & 0 & 0 \\ w_1 & w_1 & w_1 & w_1 \end{bmatrix}$, $\mathbf{W}_2 = \begin{bmatrix} w_2 & w_2 \end{bmatrix}$. The problem is $(x_1 + x_2) \cdot (x_1 + x_2 + x_3 + x_4)$ for $x = (1, 1.2, 1.8, 2)$. The solution is $w_1 = w_2 = 1$ in (a), with many unstable alternatives.

Multiplication issues

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\mathrm{NAC}_\bullet$
  - $\mathrm{NAC}_\bullet$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

Division causes a singularity in the loss curvature

singularity



(a) $\mathrm{NAC}_\bullet$ with $\epsilon = 10^{-7}$ (b) $\mathrm{NAC}_\bullet$ with $\epsilon = 0.1$ (c) $\mathrm{NAC}_\bullet$ with $\epsilon = 1$

Figure 2: RMS loss curvature for a NAC_+ unit followed by a $\mathrm{NAC}_\bullet$. The weight matrices are constrained to $\mathbf{W}_1 = \begin{bmatrix} w_1 & w_1 & 0 & 0 \\ w_1 & w_1 & w_1 & w_1 \end{bmatrix}$, $\mathbf{W}_2 = \begin{bmatrix} w_2 & w_2 \end{bmatrix}$. The problem is $(x_1+x_2)\cdot(x_1+x_2+x_3+x_4)$ for $x = (1, 1.2, 1.8, 2)$. The solution is $w_1 = w_2 = 1$ in (a), with many unstable alternatives.

Second order Taylor approximation

$$E[z_{h_\ell}] \approx \left(1 + \frac{1}{2} Var[W_{h_\ell, h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{H_{\ell-1}} \Rightarrow E[z_{h_\ell}] > 1.$$

The desirable is E[z] = 0

Neural Multiplication Unit

- weight issues:
  - gradient w.r.t. **M** is expected to be zero
  - assumed bias towards (-1, 0, 1) does not exists
- Multiplication issues:
  - singularities for w < 0 in $\text{NAC}_{\bullet}$
  - $\text{NAC}_{\bullet}$ can't be initialized optimally
  - (no multiplication of negative numbers)
- Gatting issues:
  - gating does not converge consistently

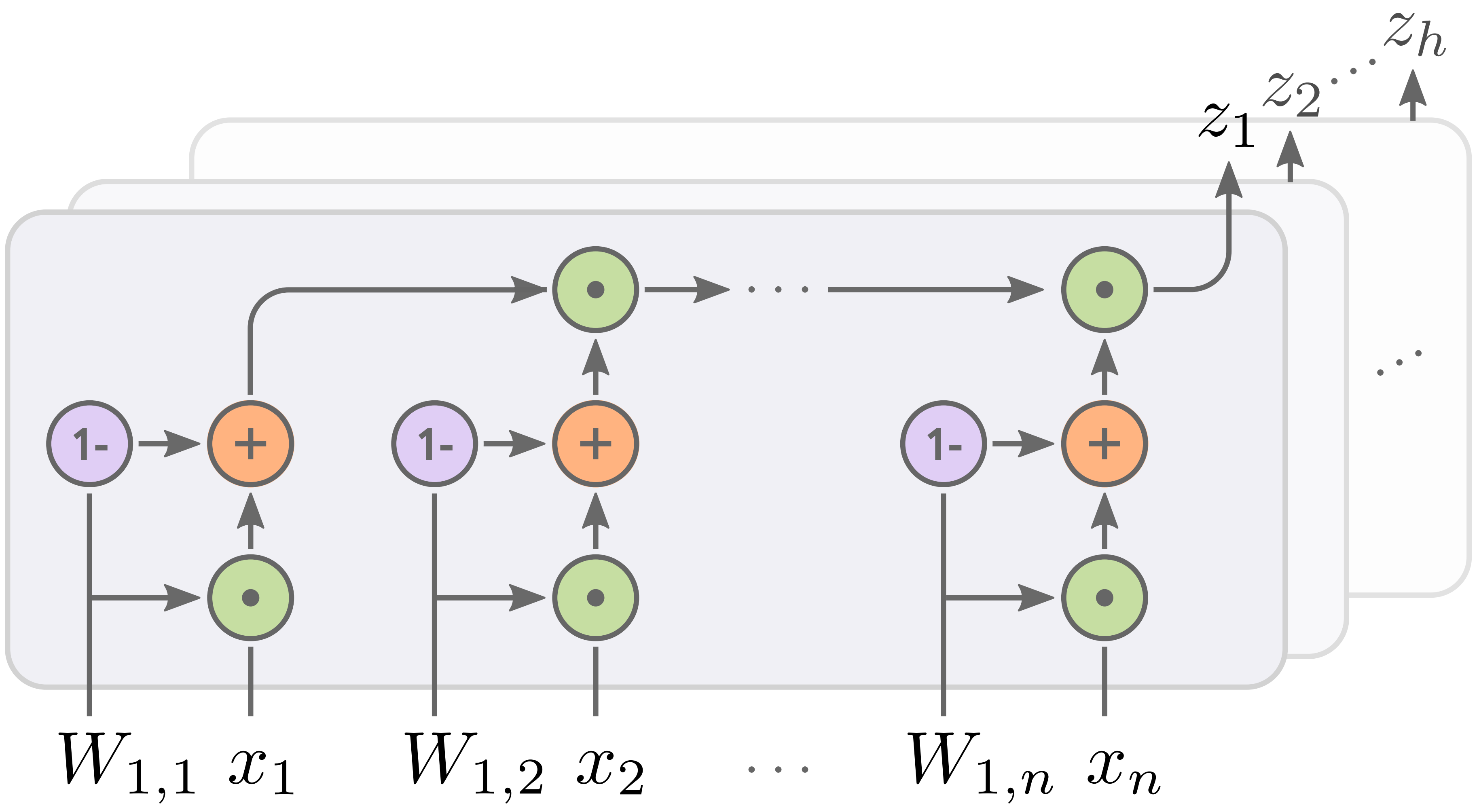
Our solution, **N**eural **M**ultiplication **U**nit:

$$W_{h_{\ell-1},h_\ell} = \min(\max(W_{h_{\ell-1},h_\ell}, 0), 1),$$

$$\mathcal{R}_{\ell,\text{sparse}} = \frac{1}{H_\ell \cdot H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \min\left(W_{h_{\ell-1},h_\ell}, 1 - W_{h_{\ell-1},h_\ell}\right)$$

$$\text{NMU}: \; z_{h_\ell} = \prod_{h_{\ell-1}=1}^{H_{\ell-1}} \left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right)$$

essentially a linear gate between 1 and z



Results

higher success-rate

Table 2: Comparison of: success-rate, first iteration reaching success, and sparsity error, all with 95% confidence interval on the "arithmetic datasets" task. Each value is a summary of 100 different seeds.

Op	Model	Success	Solved at iteration step		Sparsity error
		Rate	Median	Mean	Mean
$\times$	$\mathrm{NAC}_{\bullet}$	$31\%\,^{+10\%}_{-8\%}$	$2.8 \cdot 10^{6}$	$3.0 \cdot 10^{6}\,^{+2.9 \cdot 10^{5}}_{-2.4 \cdot 10^{5}}$	$5.8 \cdot 10^{-4}\,^{+4.8 \cdot 10^{-4}}_{-2.6 \cdot 10^{-4}}$
	NALU	$0\%\,^{+4\%}_{-0\%}$	—	—	—
	NMU	$\mathbf{98\%}\,^{+1\%}_{-5\%}$	$\mathbf{1.4 \cdot 10^{6}}$	$\mathbf{1.5 \cdot 10^{6}}\,^{+5.0 \cdot 10^{4}}_{-6.6 \cdot 10^{4}}$	$\mathbf{4.2 \cdot 10^{-7}}\,^{+2.9 \cdot 10^{-8}}_{-2.9 \cdot 10^{-8}}$
$+$	NAC_{+}	$\mathbf{100\%}\,^{+0\%}_{-4\%}$	$2.5 \cdot 10^{5}$	$4.9 \cdot 10^{5}\,^{+5.2 \cdot 10^{4}}_{-4.5 \cdot 10^{4}}$	$2.3 \cdot 10^{-1}\,^{+6.5 \cdot 10^{-3}}_{-6.5 \cdot 10^{-3}}$
	Linear	$\mathbf{100\%}\,^{+0\%}_{-4\%}$	$6.1 \cdot 10^{4}$	$\mathbf{6.3 \cdot 10^{4}}\,^{+2.5 \cdot 10^{3}}_{-3.3 \cdot 10^{3}}$	$2.5 \cdot 10^{-1}\,^{+3.6 \cdot 10^{-4}}_{-3.6 \cdot 10^{-4}}$
	NALU	$14\%\,^{+8\%}_{-5\%}$	$1.5 \cdot 10^{6}$	$1.6 \cdot 10^{6}\,^{+3.8 \cdot 10^{5}}_{-3.3 \cdot 10^{5}}$	$1.7 \cdot 10^{-1}\,^{+2.7 \cdot 10^{-2}}_{-2.5 \cdot 10^{-2}}$
	NAU	$\mathbf{100\%}\,^{+0\%}_{-4\%}$	$\mathbf{1.8 \cdot 10^{4}}$	$3.9 \cdot 10^{5}\,^{+4.5 \cdot 10^{4}}_{-3.7 \cdot 10^{4}}$	$\mathbf{3.2 \cdot 10^{-5}}\,^{+1.3 \cdot 10^{-5}}_{-1.3 \cdot 10^{-5}}$
$-$	NAC_{+}	$\mathbf{100\%}\,^{+0\%}_{-4\%}$	$9.0 \cdot 10^{3}$	$3.7 \cdot 10^{5}\,^{+3.8 \cdot 10^{4}}_{-3.8 \cdot 10^{4}}$	$2.3 \cdot 10^{-1}\,^{+5.4 \cdot 10^{-3}}_{-5.4 \cdot 10^{-3}}$
	Linear	$7\%\,^{+7\%}_{-4\%}$	$3.3 \cdot 10^{6}$	$1.4 \cdot 10^{6}\,^{+7.0 \cdot 10^{5}}_{-6.1 \cdot 10^{5}}$	$1.8 \cdot 10^{-1}\,^{+7.2 \cdot 10^{-2}}_{-5.8 \cdot 10^{-2}}$
	NALU	$14\%\,^{+8\%}_{-5\%}$	$1.9 \cdot 10^{6}$	$1.9 \cdot 10^{6}\,^{+4.4 \cdot 10^{5}}_{-4.5 \cdot 10^{5}}$	$2.1 \cdot 10^{-1}\,^{+2.2 \cdot 10^{-2}}_{-2.2 \cdot 10^{-2}}$
	NAU	$\mathbf{100\%}\,^{+0\%}_{-4\%}$	$\mathbf{5.0 \cdot 10^{3}}$	$\mathbf{1.6 \cdot 10^{5}}\,^{+1.7 \cdot 10^{4}}_{-1.6 \cdot 10^{4}}$	$\mathbf{6.6 \cdot 10^{-2}}\,^{+2.5 \cdot 10^{-2}}_{-1.9 \cdot 10^{-2}}$

sparser solution

faster convergence

Results

higher success-rate

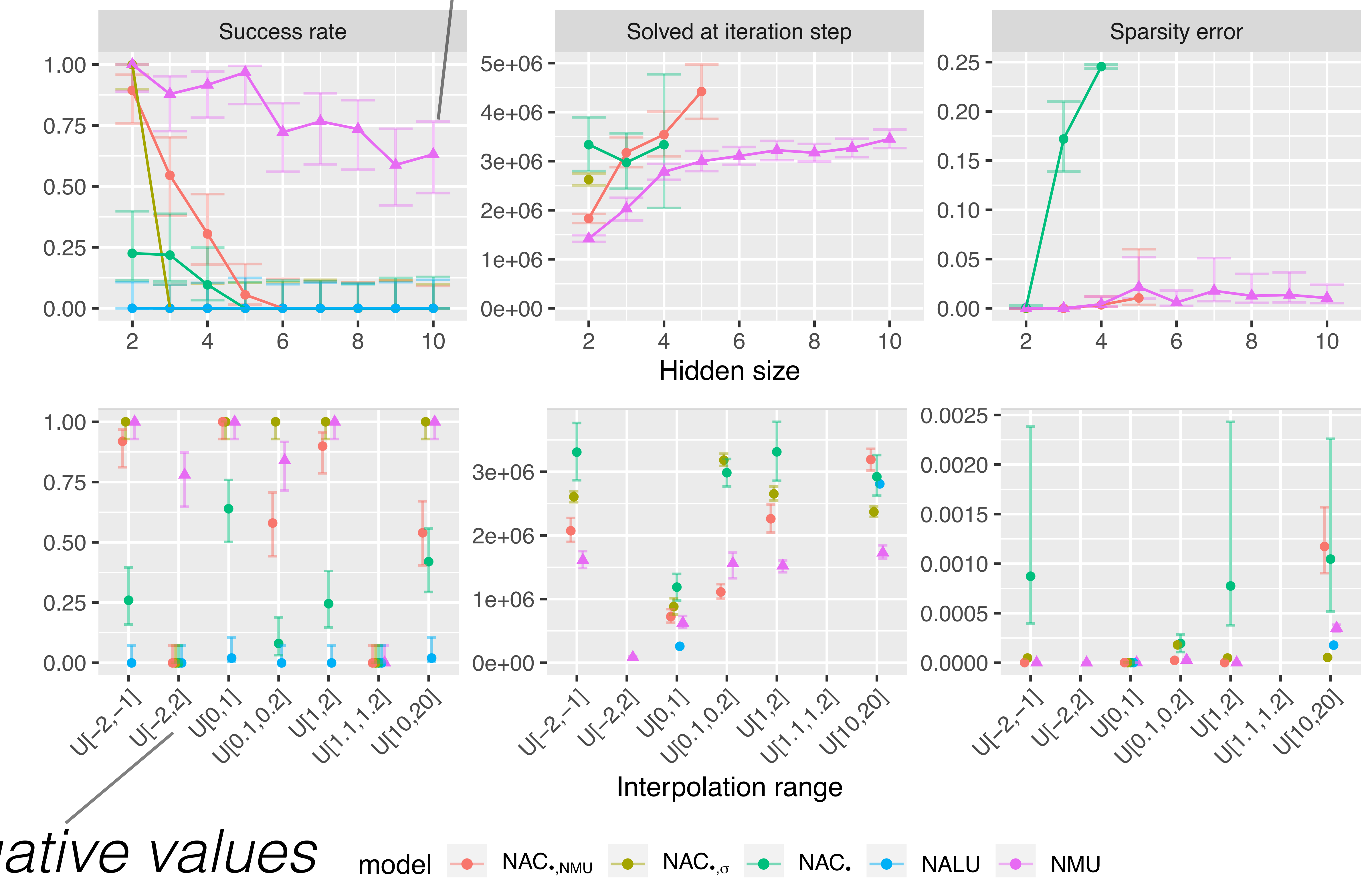
Table 2: Comparison of: success-rate, first iteration reaching success, and sparsity error, all with 95% confidence interval on the "arithmetic datasets" task. Each value is a summary of 100 different seeds.

Op	Model	Success	Solved at iteration step		Sparsity error
		Rate	Median	Mean	Mean
$\times$	NAC$_\bullet$	$31\%^{+10\%}_{-8\%}$	$2.8\cdot10^{6}$	$3.0\cdot10^{6}\,^{+2.9\cdot10^{5}}_{-2.4\cdot10^{5}}$	$5.8\cdot10^{-4}\,^{+4.8\cdot10^{-4}}_{-2.6\cdot10^{-4}}$
	NALU	$0\%^{+4\%}_{-0\%}$	—	—	—
	NMU	$\mathbf{98\%}^{+1\%}_{-5\%}$	$\mathbf{1.4\cdot10^{6}}$	$\mathbf{1.5\cdot10^{6}}\,^{+5.0\cdot10^{4}}_{-6.6\cdot10^{4}}$	$\mathbf{4.2\cdot10^{-7}}\,^{+2.9\cdot10^{-8}}_{-2.9\cdot10^{-8}}$
$+$	NAC$_+$	$\mathbf{100\%}^{+0\%}_{-4\%}$	$2.5\cdot10^{5}$	$4.9\cdot10^{5}\,^{+5.2\cdot10^{4}}_{-4.5\cdot10^{4}}$	$2.3\cdot10^{-1}\,^{+6.5\cdot10^{-3}}_{-6.5\cdot10^{-3}}$
	Linear	$\mathbf{100\%}^{+0\%}_{-4\%}$	$6.1\cdot10^{4}$	$\mathbf{6.3\cdot10^{4}}\,^{+2.5\cdot10^{3}}_{-3.3\cdot10^{3}}$	$2.5\cdot10^{-1}\,^{+3.6\cdot10^{-4}}_{-3.6\cdot10^{-4}}$
	NALU	$14\%^{+8\%}_{-5\%}$	$1.5\cdot10^{6}$	$1.6\cdot10^{6}\,^{+3.8\cdot10^{5}}_{-3.3\cdot10^{5}}$	$1.7\cdot10^{-1}\,^{+2.7\cdot10^{-2}}_{-2.5\cdot10^{-2}}$
	NAU	$\mathbf{100\%}^{+0\%}_{-4\%}$	$\mathbf{1.8\cdot10^{4}}$	$3.9\cdot10^{5}\,^{+4.5\cdot10^{4}}_{-3.7\cdot10^{4}}$	$\mathbf{3.2\cdot10^{-5}}\,^{+1.3\cdot10^{-5}}_{-1.3\cdot10^{-5}}$
$-$	NAC$_+$	$\mathbf{100\%}^{+0\%}_{-4\%}$	$9.0\cdot10^{3}$	$3.7\cdot10^{5}\,^{+3.8\cdot10^{4}}_{-3.8\cdot10^{4}}$	$2.3\cdot10^{-1}\,^{+5.4\cdot10^{-3}}_{-5.4\cdot10^{-3}}$
	Linear	$7\%^{+7\%}_{-4\%}$	$3.3\cdot10^{6}$	$1.4\cdot10^{6}\,^{+7.0\cdot10^{5}}_{-6.1\cdot10^{5}}$	$1.8\cdot10^{-1}\,^{+7.2\cdot10^{-2}}_{-5.8\cdot10^{-2}}$
	NALU	$14\%^{+8\%}_{-5\%}$	$1.9\cdot10^{6}$	$1.9\cdot10^{6}\,^{+4.4\cdot10^{5}}_{-4.5\cdot10^{5}}$	$2.1\cdot10^{-1}\,^{+2.2\cdot10^{-2}}_{-2.2\cdot10^{-2}}$
	NAU	$\mathbf{100\%}^{+0\%}_{-4\%}$	$\mathbf{5.0\cdot10^{3}}$	$\mathbf{1.6\cdot10^{5}}\,^{+1.7\cdot10^{4}}_{-1.6\cdot10^{4}}$	$\mathbf{6.6\cdot10^{-2}}\,^{+2.5\cdot10^{-2}}_{-1.9\cdot10^{-2}}$

sparser solution

faster convergence

better at larger hidden-sizes



supports negative values

Figure 3: Multiplication task results when varying the hidden input-size and when varying the input-range. Extrapolation ranges are defined in Appendix C.4.

Not covered in this presentation

- Theory:
  - Initialization of NMU
  - Gradients of NMU
  - Regularization scaling
- Discussions:
  - Gating issues (Appendix C.5)
  - Effect of shared weights in NALU
  - Measuring performance
- Experiments:
  - Detailed ablation study
  - Effect of dataset parameters
  - Results for MNIST experiment
  - Hyperparameter optimization
  - Complete comparison of all models on all arithmetic problems

Neural Arithmetic Units

By Andreas Madsen and Alexander Rosenberg Johansen
@andreas_madsen @AlexRoseJo

ICLR 2020, spotlight awarded paper

paper: https://openreview.net/forum?id=H1gNOeHKPS
code: https://github.com/AndreasMadsen/stable-nalu