Neural Arithmetic Units

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Exact addition, subtraction, multiplication and division present a unique learning challenge for machine learning models. Neural networks can approximate complex functions by learning from labeled data. However, when extrapolating to out-of-distribution samples on arithmetic operations neural networks often fail. Learning the underlying logic, as opposed to an approximation, is crucial for applications such as comparing, counting, and inferring physical models. Our proposed Neural Addition Unit (NAU) and Neural Multiplication Unit (NMU) rely on constrained weights to learn rules and extrapolate well beyond the training distribution. The proposed NAU and NMU are inspired by the underlying arithmetic components of the Neural Arithmetic Logic Unit (NALU). The NAU can perform addition and subtraction using a linear layer of constrained weights. The NMU can perform multiplication using an accumulative product of the input using gating with an identity function to mask out unwanted elements. The weights are optimized with stochastic gradient decent with regularization for sparsity. Through analytic and empirical analysis we justify how the NAU and NMU improve over the Neural Arithmetic Logic Unit (NALU), a linear regression model and a ReLU based multi-layer perceptron (MLP). Our NAU and NMU have fewer parameters, converges more consistently, learns faster and have more meaningful discrete values than the NALU and its arithmetic components.

1 Introduction

When studying intelligence, insects, reptiles and humans have been found to possess neurons with the capacity to hold numbers and do arithmetic operations[**???**]. In our quest to solve intelligence we have put much faith in neural networks, which in turn has provided unparalleled and often superhuman performance in many tasks requiring high cognitive ability [**???**]. However, when using neural networks try to learn simple arithmetic problems, such as counting, multiplication or comparison they systematically fail to extrapolate onto unseen ranges [**???**].

In this paper, we analyze and improve parts of the recently proposed Neural Arithmetic Logic Unit (NALU) [**?**], which we will introduce in section 2. Our contribution is an alternative formulation of the weight constraint with a clipped linear activation, a regularizer that bias towards sparse solutions, and a reformulation of the multiplication unit to be partially linear. All of which significantly improves upon the existing NAC_{+} and $\text{NAC}_{\bullet}$ units as shown through extensive testing on arithmetic constructions.

The NALU is a neural network layer with two sub-units; the NAC_{+} for addition/subtraction and the $\text{NAC}_{\bullet}$ for multiplication/division. The subunits are softly gated between using a sigmoid function. The layer parameters, which are created by a tanh-sigmoid transformation, are learned by observing arithmetic input-output pairs and using backpropagation[**?**].

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

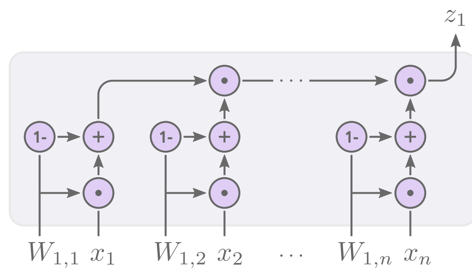


Figure 1: Visualization of NMU for a single output scalar z_1, this construction repeats for every element in the output vector $\mathbf{z}$.

In our work, we present the following findings on the NALU by investigating the parameter transformation, the sub-units and gating mechanism. We find that the mentioned components of the NALU contain analytical or emperical concerns

- The gradients from the weight matrix construction in NAC_+ and $\text{NAC}_\bullet$ have zero expectation.
- The $\text{NAC}_\bullet$ has a treacherous optimization space with unwanted global minimas (as shown in figure 2) and has exploding/vanishing gradients.
- When applying the NAC_+ in isolation, we observe that the wanted weight matrix values of $\{-1, 0, 1\}$ is rarely found.
- Our emperical results show that the NALU is significantly worse than hard-choosing either the NAC_+ or $\text{NAC}_\bullet$, indicating that the gating might not work as intended.

Motivated by these convergence and sparsity issue, we propose alternative formulations of the NAC_+ and $\text{NAC}_\bullet$, which we call the Neural Addition Unit (NAU) and Neural Multiplication Unit (NMU). We choose to avoid the gating mechanism as we see no obvious solution to simultaneously train two vastly different operations with a soft-selection mechanism. We will thus assume that the appropriate operation is already known, or can empirically be found by varying the network architecture (oracle gating).

2 Introducing differentiable binary arithmetic operations

Our goal is to achieve arithmetic operations between the elements of a vector. Such that the output is an addition, subtraction, multiplication, or division of arbitrary elements of a vector $\mathbf{x}$ (e.g. $x_5 + x_1 \cdot x_7$). Formally defined as

$$x_1 \circ_1 x_2 \circ_2 \cdots x_{k-1} \circ_{k-1} x_k \mid (x_1, \ldots, x_k) \in \mathbf{x}, \mathbf{x} \in \mathbb{R}^n, \circ_i \in \{+, -, \times, \div\} \tag{1}$$

The Neural Arithmetic Logic Unit (NALU) [**?**] attempts to solve equation 1 by presenting two sub-units; the NAC_+ and $\text{NAC}_\bullet$ to exclusively represent either the $\{+, -\}$ or the $\{\times, \div\}$ operations. The NALU attempts to have either NAC_+ or $\text{NAC}_\bullet$ selected exclusively, which could require the NALU to be applied multiple times (alternating between NAC_+ and $\text{NAC}_\bullet$) in order to represent the entire space of solutions for equation 1.

The NAC_+ and $\text{NAC}_\bullet$ are defined accordingly,

$$W_{h_\ell, h_{\ell-1}} = \tanh(\hat{W}_{h_\ell, h_{\ell-1}}) \sigma(\hat{M}_{h_\ell, h_{\ell-1}}) \tag{2}$$

$$\text{NAC}_+ : \; z_{h_\ell} = \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell, h_{\ell-1}} z_{h_{\ell-1}} \tag{3}$$

$$\text{NAC}_\bullet : \; z_{h_\ell} = \exp\left(\sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell, h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon) \right) \tag{4}$$

where $\hat{\mathbf{W}}, \hat{\mathbf{M}} \in \mathbb{R}^{H_\ell \times H_{\ell-1}}$ are trainable weight matrices. The matrices are combined using tanh and sigmoid transformation to bias the parameters towards a $\{-1, 0, 1\}$ solution. Having $\{-1, 0, 1\}$

allows a linear layer to exactly emulate the binary $\{+, -\}$ operation between elements of a vector as used when computing the NAC_+. The $\text{NAC}_\bullet$ extends the NAC_+ by using an exponential log transformation, which, with $\{-1, 0, 1\}$ weight values, becomes the $\{\times, \div\}$ operations (within ϵ precision).

The NALU combines these units with a gating mechanism $\mathbf{z} = \mathbf{g} \odot \text{NAC}_+ + (1 - \mathbf{g}) \odot \text{NAC}_\bullet$ given $\mathbf{g} = \sigma(\mathbf{G}\mathbf{x})$. The idea is that the NALU should be a plug-and-play component in a neural network and has the ability to, with stochastic gradient descent and backpropagation, to learn the functionality in equation 1.

2.1 Challenges of the NALU, NAC_+ and $\text{NAC}_\bullet$

To simplify the problem we have chosen to leave out the gating mechanism and focus on the sub-units, assuming "oracle gating". We have not had any consistent success of convergence using the gating mechanism using the NALU or by combining our own proposed sub-units (NAU, NMU), as shown in table 3. We find that gating between NAC_+ and $\text{NAC}_\bullet$ is challenging. This is likely due to the vastly different gradients, causing addition to be learned much faster than multiplication.

2.1.1 Weight matrix construction

The weight matrix construction $\tanh(\hat{W}_{h_{\ell-1},h_\ell})\sigma(\hat{M}_{h_{\ell-1},h_\ell})$ has the following properties that could make convergence challenging using gradient decent.

The loss gradient with respect to the weight matrices can be derived from equation 2.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \hat{W}_{h_{\ell-1},h_\ell}} &= \frac{\partial \mathcal{L}}{\partial W_{h_{\ell-1},h_\ell}} (1 - \tanh^2(\hat{W}_{h_{\ell-1},h_\ell}))\sigma(\hat{M}_{h_{\ell-1},h_\ell}) \\
\frac{\partial \mathcal{L}}{\partial \hat{M}_{h_{\ell-1},h_\ell}} &= \frac{\partial \mathcal{L}}{\partial W_{h_{\ell-1},h_\ell}} \tanh(\hat{W}_{h_{\ell-1},h_\ell})\sigma(\hat{M}_{h_{\ell-1},h_\ell})(1 - \sigma(\hat{M}_{h_{\ell-1},h_\ell}))
\end{aligned} \tag{5}
$$

The gradient $E\left[{}^{\partial\mathcal{L}}/_{\partial\hat{M}_{h_{\ell-1},h_\ell}}\right] = 0$ can be problematic as we prefer zero having a zero mean expectation of our output. Something that can only be ensured with $E[\hat{W}_{h_{\ell-1},h_\ell}] = 0$ [**?**].

In our empirical analysis we find that equation 2 does not create the desired bias for $\{-1, 0, 1\}$, as it doesn't converge towards those values.

To create a bias and prevent the gradient challenges of equation 5 we propose a simple clamped linear construction with an out-of-bound regularizer $\mathcal{R}_{\ell,\text{oob}}$ to force $\hat{W}$ to be within $[-1, 1]$ and ensure that the gradient is always present.

$$
\begin{aligned}
W_{h_{\ell-1},h_\ell} &= \min(\max(\hat{W}_{h_{\ell-1},h_\ell}, -1), 1), \\
\mathcal{R}_{\ell,\text{bias}} &= \frac{1}{H_\ell + H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \hat{W}_{h_{\ell-1},h_\ell}^2 (1 - |\hat{W}_{h_{\ell-1},h_\ell}|)^2 \\
\mathcal{R}_{\ell,\text{oob}} &= \frac{1}{H_\ell + H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \max(|\hat{W}_{h_{\ell-1},h_\ell}| - 1, 0)^2 \\
\text{NAU}: \ z_{h_\ell} &= \sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} z_{h_{\ell-1}} \\
\mathcal{L} &= \hat{\mathcal{L}} + \lambda_{\text{bias}} \mathcal{R}_{\ell,\text{bias}} + \lambda_{\text{oob}} \mathcal{R}_{\ell,\text{oob}}
\end{aligned} \tag{6}
$$

2.1.2 Challenges of division

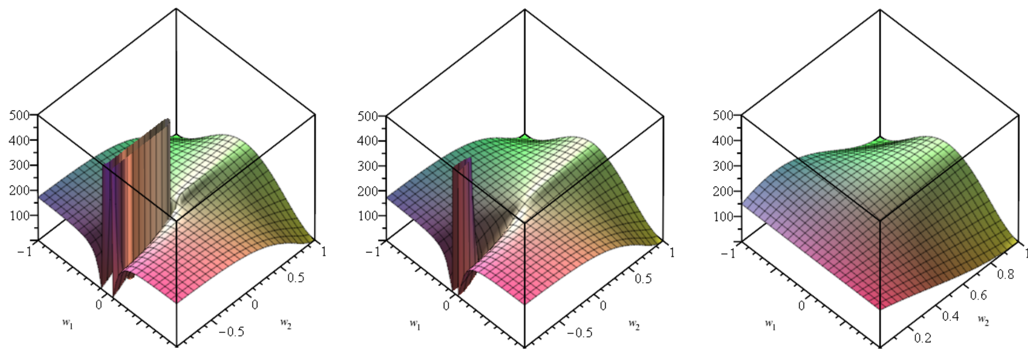
The $\text{NAC}_\bullet$, as formulated in equation 4, has the ability to learn exact multiplication and division of elements from a vector if the weights of $W_{h_{\ell-1},h_\ell}$ are one of $\{-1, 0, 1\}$.

However, backpropagation through the $\text{NAC}_\bullet$ unit reveals that if $|z_{h_{\ell-1}}|$ is near zero, $W_{h_{\ell-1},h_\ell}$ is negative and ϵ is small, the gradient term will explode and oscillate between large positive and large negative values, which can be problematic in optimization [**?**], as visualized in figure 2.

$$\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}} = \sum_{h_\ell=1}^{H_\ell} \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}} = \sum_{h_\ell=1}^{H_\ell} \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} z_{h_\ell} W_{h_\ell, h_{\ell-1}} \frac{\operatorname{sign}(z_{h_{\ell-1}})}{|z_{h_{\ell-1}}| + \epsilon} \tag{7}$$

(see full derivation in Appendix A.2)

This is not an issue for for positive values of $W_{h_{\ell-1},h_\ell}$ (multiplication), as z_{h_ℓ} and $z_{h_{\ell-1}}$ will be correlated causing the terms z_{h_ℓ} and $\frac{\operatorname{sign}(z_{h_{\ell-1}})}{|z_{h_{\ell-1}}|+\epsilon}$ to partially cancel out.



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

Figure 2: RMS loss curvature for a NAC_+ layer followed by either a $\mathrm{NAC}_\bullet$ or NMU layer. The weight matrices constrained are to $\mathbf{W}_1 = \left[\begin{smallmatrix} w_1 & w_1 & 0 & 0 \\ w_1 & w_1 & w_1 & w_1 \end{smallmatrix}\right]$, $\mathbf{W}_2 = \left[\begin{smallmatrix} w_2 & w_2 \end{smallmatrix}\right]$. The problem is $x = (1, 1.2, 1.8, 2)\,, t = 13.2$. Desired solution is $w_1 = w_2 = 1$, although this problem have additional undesired solutions.

This gradient can be particular problematic when considering that $E[z_{h_{\ell-1}}] = 0$ is a desired property when initializing [**?**]. An alternative multiplication operator must thus be able to not explode for $z_{h_{\ell-1}}$ near zero. To that end we propose a new neural multiplication units (NMU):

$$\begin{aligned}
W_{h_{\ell-1},h_\ell} &= \min(\max(\hat{W}_{h_{\ell-1},h_\ell}, 0), 1), \\
\mathcal{R}_{\ell,\text{bias}} &= \frac{1}{H_\ell + H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \hat{W}^2_{h_{\ell-1},h_\ell} (1 - \hat{W}_{h_{\ell-1},h_\ell})^2 \\
\mathcal{R}_{\ell,\text{oob}} &= \frac{1}{H_\ell + H_{\ell-1}} \sum_{h_\ell=1}^{H_\ell} \sum_{h_{\ell-1}=1}^{H_{\ell-1}} \max\left(\left| \hat{W}_{h_{\ell-1},h_\ell} - \frac{1}{2} \right| - \frac{1}{2}, 0 \right)^2 \\
\text{NMU}: \; z_{h_\ell} &= \prod_{h_{\ell-1}=1}^{H_{\ell-1}} \left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell} \right) \\
\mathcal{L} &= \hat{\mathcal{L}} + \lambda_{\text{bias}} \mathcal{R}_{\ell,\text{bias}} + \lambda_{\text{oob}} \mathcal{R}_{\ell,\text{oob}}
\end{aligned} \tag{8}$$

Notable is the multiplicative identity for when $W_{h_{\ell-1},h_\ell} = 0$. This unit does not support division, but supporting division is likely infeasible as dividing by $z_{h_{\ell-1}}$ near zero would cause explosions. As shown in [**?**], experiments using the NALU for division does not work well hence very little is lost with this modification. As opposed to the NALU, the NMU can represent input of both negative and positive $z_{h_{\ell-1}}$ values and is not ϵ dependent, which allows the NMU to extrapolate inputs that are negative or smaller than ϵ.

The gradients with respect to the weight and input in the NMU are (see details in Appendix A.3):

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}} &= \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial W_{h_\ell,h_{\ell-1}}} = \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}} \left(z_{h_{\ell-1}} - 1\right) \\
\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}} &= \sum_{h_\ell=1}^{H_\ell} \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}} = \sum_{h_\ell=1}^{H_\ell} \frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}} W_{h_{\ell-1},h_\ell}
\end{aligned}
\tag{9}
$$

Note that the fraction does not explode for $z_{h_{\ell-1}}$ close to zero, as the denominator simply cancels out a term in z_{h_ℓ}.

2.1.3 Moments and initialization

Initialization is important to consider for fast and consistent convergence [**?**].

Our proposed NAU, can be initialize using Glorot initialization as it is a linear layer. The NAC_{+} unit can also achieve an ideal initialization, although it is less trivial (details in Appendix B.2).

Using second order multivariate Taylor approximation and some assumptions of uncorrelated stochastic variables, the expectation of $\mathrm{NAC}_{\bullet}$ can be estimated to be

$$
E[z_{h_\ell}] \approx \left(1 + \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{H_{\ell-1}} \Rightarrow E[z_{h_\ell}] > 1 \tag{10}
$$

(proof in Appendix B.3). An ideal initialization should satisfy $E[z_{h_\ell}] = 0$ [**?**], which the expectation for $\mathrm{NAC}_{\bullet}$ is infeasible.

Our proposed NMU when initialized with $E[W_{h_\ell,h_{\ell-1}}] = 1/2$ has an expectation of

$$
E[z_{h_\ell}] \approx \left(\frac{1}{2}\right)^{H_{\ell-1}} \tag{11}
$$

which approaches zero for $H_{\ell-1} \to \infty$ (proof in Appendix B.4).

The $\mathrm{NAC}_{\bullet}$ can not be input-independent initialization and has an exploding variance in depth (proof in Appendix B.3 and B.4). The NMU can, with the assumption that, $Var[z_{h_{\ell-1}}] = 1$ and $H_{\ell-1}$ is large, be initialized optimally with $Var[W_{h_{\ell-1},h_\ell}] = \frac{1}{4}$ (see proof in Appendix B.4.3).

3 Experimental results

3.1 Arithmetic datasets

The arithmetic dataset is a replica of the "simple function task" shown in [**?**]. The goal is to sum two subsets of a vector and perform a arithmetic operation as defined below

$$
t = \sum_{i=a_{\text{start}}}^{a_{\text{end}}} \mathbf{x}_i \circ \sum_{i=b_{\text{start}}}^{b_{\text{end}}} \mathbf{x}_i \quad \text{where } \mathbf{x} \in \mathbb{R}^n, x_i \sim \text{Uniform}[r_{\text{lower}}, r_{\text{upper}}], \circ \in \{+, -, \times\} \tag{12}
$$

where n, r_{lower}, r_{upper}, $\circ$, the subset size and subset overlap are dataset parameters that we use to test the models ability to learn. We define a set of default parameters, see table (table 1). When probing a specific dataset parameter, e.g. subset overlap, the default will be the used for the remaining parameters.

3.1.1 Criterion

The goal is to achieve a solution that is acceptably close to a perfect solution. To evaluate if a model instance solves the task, the MSE is compared to a known nearly-perfect solution on the extrapolation range. If $\mathbf{W}_1, \mathbf{W}_2$ defines the weights of the fitted model, and $\mathbf{W}_1^\epsilon$ is nearly-perfect and $\mathbf{W}_2^*$ is perfect (example in equation 13), the success criteria is $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, measured on the extrapolation error, for $\epsilon = 0.0001$.

Do we need that equation

$$
\mathbf{W}_1^\epsilon = \begin{bmatrix} 1-\epsilon & 1-\epsilon & 0+\epsilon & 0+\epsilon \\ 1-\epsilon & 1-\epsilon & 1-\epsilon & 1-\epsilon \end{bmatrix}, \mathbf{W}_2^* = \begin{bmatrix} 1 & 1 \end{bmatrix} \tag{13}
$$

Table 1: Default dataset parameters

Parameter name	Default value
Input size	100
Subset ratio	0.25
Overlap ratio	0.5
Interpolation range	$U[1,2]$
Extrapolation range	$U[2,6]$

All experiments are evaluated multiple times with different seeds. We define the success rate as the percentage of experiments that achieves success.

A sparsity error is also evaluated, this is defined in equation 14. This is only considered for model instances that did solve the task, for which the mean and 95% confidence interval is reported.

$$E_{\text{sparsity}} = \max_{h_{\ell-1},h_\ell} \min(|W_{h_{\ell-1},h_\ell}|, |1-|W_{h_{\ell-1},h_\ell}||) \tag{14}$$

The first iteration for which $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, is also reported with the 95% confidence interval. Again, only model instances that did solve the task are considered.

3.1.2 Experiment setup

The multiplication models, NMU and $\mathrm{NAC}_\bullet$ have an addition layer first, either NAU or NAC_+, followed by a multiplication layer. The addition models, NAC_+, NAU, and Linear are just two layers of that unit. Finally, the NALU model is also two layers of NALU. See explicit definitions in table 4. All models are fitted with an MSE loss function.

For all experiments $\lambda_{\text{oob}} = 1$ and $\lambda_{\text{bias}} = 0.1 \cdot (1 - \exp(-10^5 \cdot t))$. Gradually scaling the bias regularizer $\mathcal{R}_{\ell,\text{bias}}$ is to ensure it does not interfere with early training. We show the effect of regularization in appendix C.4. All experiments uses Adam optimization [?] with default parameters, and are computed on an HPC cluster using 8-Core Intel Xeon E5-2665 2.4GHz CPUs.

The training dataset is continuously sampled from the interpolation range, a different seed is used for each experiment. Training is done with a mini-batch size of 128 observations. A fixed validation dataset with 10000 observations is sampled from the interpolation range. A fixed test dataset with 10000 observations is sample from the extrapolation range.

Validation error, test error and sparsity error is sampled every 1000 iterations. To avoid noise from exploration, the best fit in terms of the validation error among the last 100 samples is used.

3.1.3 Very simple function

To empirically validate the theoretical challenges with $\mathrm{NAC}_\bullet$ consider the very simple problem shown earlier in figure 2. That is, $t = (x_1 + x_2) \circ (x_1 + x_2 + x_3 + x_4)$ for $x \in \mathbb{R}^4$.

Each experiment is conducted 100 times with different seeds, and stopped after 200000 iterations.

The results, in table 2, show that NMU has a higher success rate and converges faster. When inspecting the 6% that did not converge, we found the issue to underflow when $w = 0$ in the NMU layer.

3.1.4 Arithmetic operation comparison

We compare the models on different arithmetic operation $\circ \in \{+, -, \times\}$ used in equation 12, results are seen in table 3, where each experiment is trained for $5 \cdot 10^6$ iterations.

For multiplication, the NMU success more often and converges faster. For addition and subtraction, the NAU model converges faster, given the median, and has a more sparse solution.

Table 2: Shows the success-rate for $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, at what global step the model converged at, and the sparsity error for all weight matrices.

Op	Model	Success	Solved at		Sparsity error
		Rate	Median	Mean	Mean
$\times$	$\mathrm{NAC}_\bullet$	13%	$4.1 \cdot 10^4$	$4.4 \cdot 10^4 \pm 6.6 \cdot 10^3$	$\mathbf{7.5 \cdot 10^{-6} \pm 2.0 \cdot 10^{-6}}$
	NALU	26%	$4.7 \cdot 10^4$	$5.4 \cdot 10^4 \pm 8.2 \cdot 10^3$	$9.2 \cdot 10^{-6} \pm 1.7 \cdot 10^{-6}$
	NMU	$\mathbf{94\%}$	$\mathbf{1.3 \cdot 10^4}$	$\mathbf{1.7 \cdot 10^4 \pm 3.3 \cdot 10^3}$	$5.2 \cdot 10^{-5} \pm 4.0 \cdot 10^{-5}$

Table 3: Shows the success-rate for $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, at what global step the model converged at, and the sparsity error for all weight matrices.

Op	Model	Success	Solved at		Sparsity error
		Rate	Median	Mean	Mean
$\times$	$\mathrm{NAC}_\bullet$	30%	$2.5 \cdot 10^6$	$2.5 \cdot 10^6 \pm 1.5 \cdot 10^6$	$\mathbf{3.9 \cdot 10^{-4} \pm 9.4 \cdot 10^{-4}}$
	Linear	0%	—	—	—
	NALU	0%	—	—	—
	NMU	$\mathbf{90\%}$	$\mathbf{1.4 \cdot 10^6}$	$\mathbf{1.6 \cdot 10^6 \pm 5.6 \cdot 10^5}$	$1.8 \cdot 10^{-3} \pm 1.1 \cdot 10^{-3}$
$+$	NAC_+	$\mathbf{100\%}$	$6.0 \cdot 10^4$	$7.1 \cdot 10^4 \pm 2.4 \cdot 10^4$	$4.8 \cdot 10^{-1} \pm 2.0 \cdot 10^{-2}$
	Linear	$\mathbf{100\%}$	$4.2 \cdot 10^4$	$\mathbf{4.2 \cdot 10^4 \pm 1.9 \cdot 10^3}$	$6.1 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$
	NALU	0%	—	—	—
	NAU	$\mathbf{100\%}$	$\mathbf{1.8 \cdot 10^4}$	$7.0 \cdot 10^5 \pm 9.2 \cdot 10^5$	$\mathbf{1.7 \cdot 10^{-3} \pm 8.0 \cdot 10^{-4}}$
$-$	NAC_+	$\mathbf{100\%}$	$8.0 \cdot 10^3$	$1.5 \cdot 10^6 \pm 1.5 \cdot 10^6$	$4.6 \cdot 10^{-1} \pm 2.9 \cdot 10^{-2}$
	Linear	$\mathbf{100\%}$	$1.1 \cdot 10^6$	$1.9 \cdot 10^6 \pm 1.3 \cdot 10^6$	$3.7 \cdot 10^{-1} \pm 1.1 \cdot 10^{-1}$
	NALU	20%	$3.6 \cdot 10^6$	$3.6 \cdot 10^6 \pm 1.3 \cdot 10^7$	$4.7 \cdot 10^{-1} \pm 3.3 \cdot 10^{-1}$
	NAU	$\mathbf{100\%}$	$\mathbf{4.0 \cdot 10^3}$	$\mathbf{4.2 \cdot 10^3 \pm 3.0 \cdot 10^2}$	$\mathbf{1.9 \cdot 10^{-3} \pm 4.2 \cdot 10^{-4}}$

3.1.5 Exploration of dataset parameters

To stress test the NMU in comparison with the $\mathrm{NAC}_\bullet$ and NALU, on the multiplication task, the dataset parameters (table 1) are varied. Each experiment runs for 10 different seeds, the results are visualized in in figure 3.

Our results show that the NMU consistently outperform the $\mathrm{NAC}_\bullet$ and the NALU for all parameters.

4 Related work

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut mollis consequat lacus ac aliquam. Phasellus pharetra laoreet mi ac dignissim. Sed condimentum venenatis mollis. Nunc tempus arcu fermentum, viverra nisi non, bibendum tortor. Vestibulum in elit velit. In faucibus egestas est, in blandit dui interdum ut. Quisque felis odio, aliquet id congue non, hendrerit id dui. Fusce mattis diam condimentum augue aliquam, eu bibendum ex tempus. Vestibulum suscipit metus sed tortor scelerisque interdum. Nam laoreet purus dolor, in ornare augue dignissim eu. In hac habitasse platea dictumst. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In ut diam nec nisi rhoncus finibus. Maecenas vel ligula vel metus ullamcorper auctor. Pellentesque volutpat quam sed ligula consectetur, ac facilisis purus facilisis.

Phasellus bibendum imperdiet mattis. Cras dictum purus nulla, sed finibus dolor porttitor sed. Proin in velit leo. Curabitur maximus, diam vel consectetur consequat, velit dolor vestibulum mi, eu consectetur felis mauris in justo. Donec non iaculis velit, quis egestas ex. Nullam consequat eros at nisi varius ultrices. Duis ultricies risus ac dolor semper tempor.

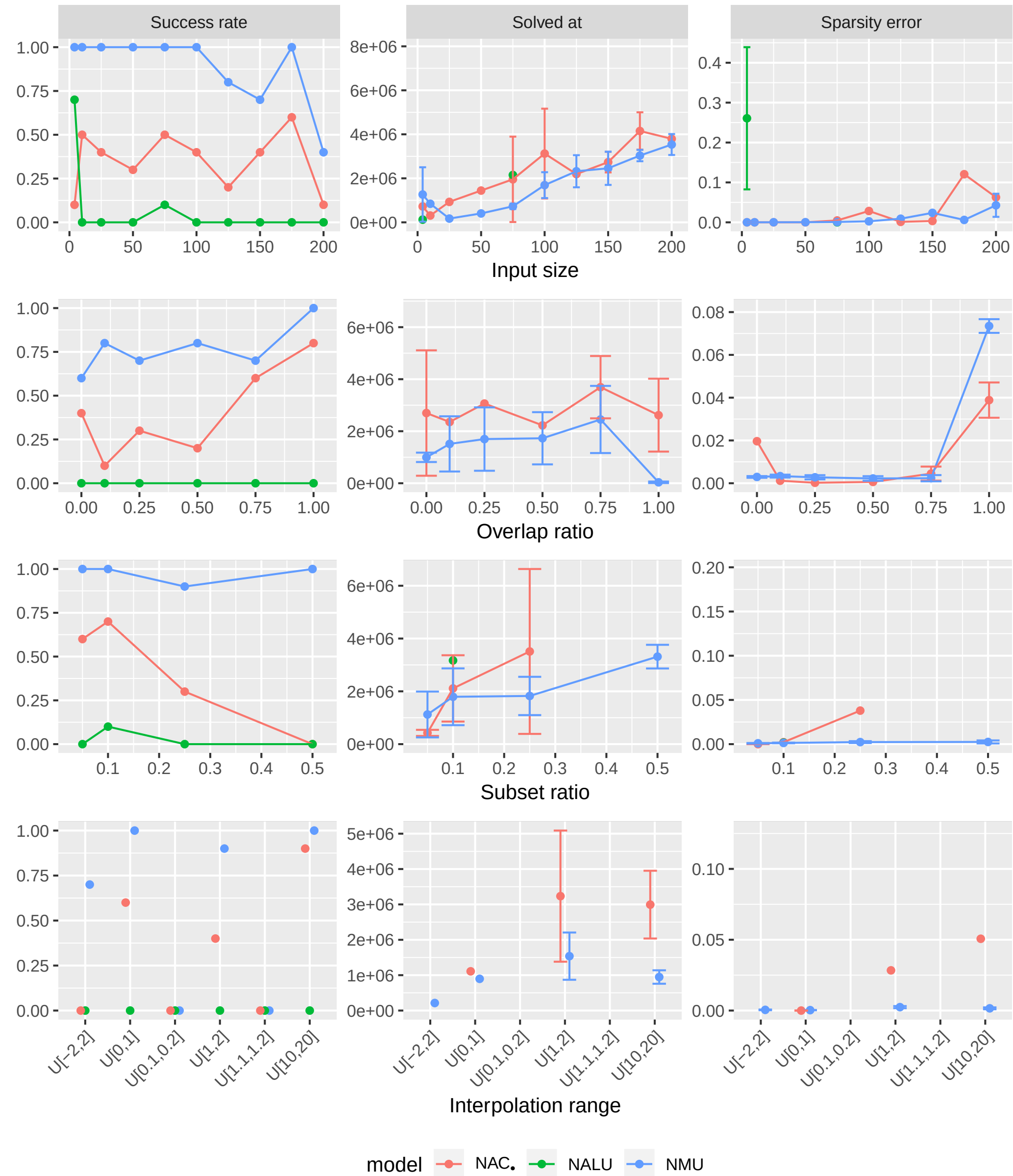


Figure 3: Shows the effect of the dataset parameters. For each interpolation range, the following extrapolation ranges are used: $\mathrm{U}[-2,2] \rightarrow \mathrm{U}[-6,-2] \cup \mathrm{U}[2,6]$, $\mathrm{U}[0,1] \rightarrow \mathrm{U}[1,5]$, $\mathrm{U}[0.1,0.2] \rightarrow \mathrm{U}[0.2,2]$, $\mathrm{U}[1,2] \rightarrow \mathrm{U}[2,6]$, $\mathrm{U}[10,20] \rightarrow \mathrm{U}[20,40]$.

5 Conclusion

An recent approach to learn arithmetic operations from data using stochastic gradient decent, has analytical and empirical concerns. We have shown analytical how the NAU and NMU can be initialized optimally. In experiments stress-testing arithmetic operations, the NAU and NMU consistently outperforms recent approaches and neural networks. While the NMU can not divide it is capable of extrapolate into the negative range for multiplication.

A Gradient derivatives

A.1 Weight matrix construction

For clarity the weight matrix construction is defined using scalar notation

$$W_{h_\ell,h_{\ell-1}} = \tanh(\hat{W}_{h_\ell,h_{\ell-1}})\sigma(\hat{M}_{h_\ell,h_{\ell-1}}) \tag{15}$$

The of the loss with respect to $\hat{W}_{h_\ell,h_{\ell-1}}$ and $\hat{M}_{h_\ell,h_{\ell-1}}$ is then straight forward to derive.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \hat{W}_{h_\ell,h_{\ell-1}}} &= \frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}} \frac{\partial W_{h_\ell,h_{\ell-1}}}{\partial \hat{W}_{h_\ell,h_{\ell-1}}} \\
&= \frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}} (1 - \tanh^2(\hat{W}_{h_\ell,h_{\ell-1}}))\sigma(\hat{M}_{h_\ell,h_{\ell-1}}) \\
\frac{\partial \mathcal{L}}{\partial \hat{M}_{h_\ell,h_{\ell-1}}} &= \frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}} \frac{\partial W_{h_\ell,h_{\ell-1}}}{\partial \hat{M}_{h_\ell,h_{\ell-1}}} \\
&= \frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}} \tanh(\hat{W}_{h_\ell,h_{\ell-1}})\sigma(\hat{M}_{h_\ell,h_{\ell-1}})(1 - \sigma(\hat{M}_{h_\ell,h_{\ell-1}}))
\end{aligned} \tag{16}$$

As seen from this result, one only needs to consider $\frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}}$ for NAC_+ and $\mathrm{NAC}_\bullet$, as the gradient with respect to $\hat{W}_{h_\ell,h_{\ell-1}}$ and $\hat{M}_{h_\ell,h_{\ell-1}}$ is just a multiplication on $\frac{\partial \mathcal{L}}{\partial W_{h_\ell,h_{\ell-1}}}$.

A.2 Gradient of $\mathrm{NAC}_\bullet$

First the $\mathrm{NAC}_\bullet$ is defined using scalar notation.

$$z_{h_\ell} = \exp\left(\sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon)\right) \tag{17}$$

The gradient of the loss with respect to $W_{h_\ell,h_{\ell-1}}$ is straight forward to derive.

$$\begin{aligned}
\frac{\partial z_{h_\ell}}{\partial W_{h_\ell,h_{\ell-1}}} &= \exp\left(\sum_{h'_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h'_{\ell-1}} \log(|z_{h'_{\ell-1}}| + \epsilon)\right) \log(|z_{h_{\ell-1}}| + \epsilon) \\
&= z_{h_\ell} \log(|z_{h_{\ell-1}}| + \epsilon)
\end{aligned} \tag{18}$$

We now wish to derive the backpropagation term $\delta_{h_\ell} = \frac{\partial \mathcal{L}}{\partial z_{h_\ell}}$, because z_{h_ℓ} affects $\{z_{h_{\ell+1}}\}_{h_{\ell+1}=1}^{H_{\ell+1}}$ this becomes:

$$\delta_{h_\ell} = \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} = \sum_{h_{\ell+1}=1}^{H_{\ell+1}} \frac{\partial \mathcal{L}}{\partial z_{h_{\ell+1}}} \frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}} = \sum_{h_{\ell+1}=1}^{H_{\ell+1}} \delta_{h_{\ell+1}} \frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}} \tag{19}$$

To make it easier to derive $\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}$ we re-express the z_{h_ℓ} as $z_{h_{\ell+1}}$.

$$z_{h_{\ell+1}} = \exp\left(\sum_{h_\ell=1}^{H_\ell} W_{h_{\ell+1},h_\ell} \log(|z_{h_\ell}| + \epsilon)\right) \tag{20}$$

The gradient of $\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}$ is then:

$$\begin{aligned}
\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}} &= \exp\left(\sum_{h_\ell=1}^{H_\ell} W_{h_{\ell+1},h_\ell} \log(|z_{h_\ell}| + \epsilon)\right) W_{h_{\ell+1},h_\ell} \frac{\partial \log(|z_{h_\ell}| + \epsilon)}{\partial z_{h_\ell}} \\
&= \exp\left(\sum_{h_\ell=1}^{H_\ell} W_{h_{\ell+1},h_\ell} \log(|z_{h_\ell}| + \epsilon)\right) W_{h_{\ell+1},h_\ell} \frac{\mathrm{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon} \\
&= m_{h_{\ell+1}} W_{h_{\ell+1},h_\ell} \frac{\mathrm{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}
\end{aligned} \tag{21}$$

$\mathrm{abs}'(z_{h_\ell})$ is the gradient of the absolute function. In the paper we denote this as $\mathrm{sign}(z_{h_\ell})$ for brevity. However, depending on the exact definition used there may be a difference for $z_{h_\ell} = 0$, as $\mathrm{abs}'(0)$ is undefined. In practicality this doesn't matter much though, although theoretically it does mean that the expectation of this is theoretically undefined when $E[z_{h_\ell}] = 0$.

A.3 Gradient of NMU

In scalar notation the NMU is defined as:

$$z_{h_\ell} = \prod_{h_{\ell-1}=1}^{H_{\ell-1}} \left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right) \tag{22}$$

The gradient of the loss with respect to $W_{h_{\ell-1},h_\ell}$ is fairly trivial. Note that every term but the one for $h_{\ell-1}$, is just a constant with respect to $W_{h_{\ell-1},h_\ell}$. The product, expect the term for $h_{\ell-1}$ can be expressed as $\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}$. Using this fact, it becomes trivial to derive the gradient as:

$$\frac{\partial \mathcal{L}}{\partial w_{h_\ell,h_{\ell-1}}} = \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial w_{h_\ell,h_{\ell-1}}} = \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}} \left(z_{h_{\ell-1}} - 1\right) \tag{23}$$

Similarly, the gradient $\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}$ which is essential in backpropagation can equally easily be derived as:

$$\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}} = \sum_{h_\ell=1}^{H_\ell} \frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}} = \sum_{h_\ell=1}^{H_\ell} \frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}} W_{h_{\ell-1},h_\ell} \tag{24}$$

B Moments

B.1 Overview

B.1.1 Moments and initialization for addition

The desired properties for initialization are according to Glorot et al. [?]:

$$\begin{aligned} E[z_{h_\ell}] &= 0 & E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &= 0 \\ Var[z_{h_\ell}] &= Var\left[z_{h_{\ell-1}}\right] & Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &= Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] \end{aligned} \tag{25}$$

B.1.2 Initialization for addition

Glorot initialization can not be used for NAC_+ as $W_{h_{\ell-1},h_\ell}$ is not sampled directly. Assuming that $\hat{W}_{h_\ell,h_{\ell-1}} \sim \text{Uniform}[-r,r]$ and $\hat{M}_{h_\ell,h_{\ell-1}} \sim \text{Uniform}[-r,r]$, then the variance can be derived (see proof in Appendix B.2) to be:

$$Var[W_{h_{\ell-1},h_\ell}] = \frac{1}{2r}\left(1 - \frac{\tanh(r)}{r}\right)\left(r - \tanh\left(\frac{r}{2}\right)\right) \tag{26}$$

One can then solve for r, given the desired variance ($Var[W_{h_{\ell-1},h_\ell}] = \frac{2}{H_{\ell-1}+H_\ell}$) [?].

B.1.3 Moments and initialization for multiplication

Using second order multivariate Taylor approximation and some assumptions of uncorrelated stochastic variables, the expectation and variance of the $\text{NAC}_\bullet$ layer can be estimated to:

$$\begin{aligned} f(c_1, c_2) &= \left(1 + c_1 \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{c_2\ H_{\ell-1}} \\ E[z_{h_\ell}] &\approx f(1,1) \\ Var[z_{h_2}] &\approx f(4,1) - f(1,2) \\ E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &= 0 \\ Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell\ f(4,1)\ Var[W_{h_\ell,h_{\ell-1}}] \\ &\quad \cdot \left(\frac{1}{\left(|E[z_{h_{\ell-1}}]| + \epsilon\right)^2} + \frac{3}{\left(|E[z_{h_{\ell-1}}]| + \epsilon\right)^4} Var[z_{h_{\ell-1}}]\right) \end{aligned} \tag{27}$$

This is problematic because $E[z_{h_\ell}] \geq 1$, and the variance explodes for $E[z_{h_{\ell-1}}] = 0$. $E[z_{h_{\ell-1}}] = 0$ is normally a desired property [?]. The variance explodes for $E[z_{h_{\ell-1}}] = 0$, and can thus not be initialized to anything meaningful.

For our proposed NMU, the expectation and variance can be derived (see proof in Appendix B.4) using the same assumptions as before, although no Taylor approximation is required:

$$\begin{aligned} E[z_{h_\ell}] &\approx \left(\frac{1}{2}\right)^{H_{\ell-1}} \\ E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx 0 \\ Var[z_{h_\ell}] &\approx \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}} - \left(\frac{1}{4}\right)^{H_{\ell-1}} \\ Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \\ &\quad \cdot \left(\left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1}\right) \end{aligned} \tag{28}$$

These expectations are better behaved. It is properly unlikely to expect that the expectation can become zero, since the identity for multiplication is 1. However, for a large $H_{\ell-1}$ it will be near zero.

The variance is also better behaved, but does not provide a input-independent initialization strategy. We propose initializing with $Var[W_{h_{\ell-1},h_\ell}] = \frac{1}{4}$, as this is the solution to $Var[z_{h_\ell}] = Var[z_{h_{\ell-1}}]$ assuming $Var[z_{h_{\ell-1}}] = 1$ and a large $H_{\ell-1}$ (see proof in Appendix B.4.3). However, feel free to compute more exact solutions.

B.2 Expectation and variance for weight matrix construction in NAC layers

The weight matrix construction in NAC, is defined in scalar notation as:

$$W_{h_\ell,h_{\ell-1}} = \tanh(\hat{W}_{h_\ell,h_{\ell-1}})\sigma(\hat{M}_{h_\ell,h_{\ell-1}}) \tag{29}$$

Simplifying the notation of this, and re-expressing it using stochastic variables with uniform distributions this can be written as:

$$\begin{aligned} W &\sim \tanh(\hat{W})\sigma(\hat{M}) \\ \hat{W} &\sim U[-r, r] \\ \hat{M} &\sim U[-r, r] \end{aligned} \tag{30}$$

Since $\tanh(\hat{W})$ is an odd-function and $E[\hat{W}] = 0$, deriving the expectation $E[W]$ is trivial.

$$\mathrm{E}[W] = \mathrm{E}[\tanh(\hat{W})]\mathrm{E}[\sigma(\hat{M})] = 0 \cdot \mathrm{E}[\sigma(\hat{M})] = 0 \tag{31}$$

The variance is more complicated, however as $\hat{W}$ and $\hat{M}$ are independent, it can be simplified to:

$$\mathrm{Var}[W] = \mathrm{E}[\tanh(\hat{W})^2]\mathrm{E}[\sigma(\hat{M})^2] - \mathrm{E}[\tanh(\hat{W})]^2\mathrm{E}[\sigma(\hat{M})]^2 = \mathrm{E}[\tanh(\hat{W})^2]\mathrm{E}[\sigma(\hat{M})^2] \tag{32}$$

These second moments can be analyzed independently. First for $\mathrm{E}[\tanh(\hat{W})^2]$:

$$\begin{aligned} \mathrm{E}[\tanh(\hat{W})^2] &= \int_{-\infty}^{\infty} \tanh(x)^2 f_{U[-r,r]}(x)\,\mathrm{d}x \\ &= \frac{1}{2r}\int_{-r}^{r} \tanh(x)^2\,\mathrm{d}x \\ &= \frac{1}{2r} \cdot 2 \cdot (r - \tanh(r)) \\ &= 1 - \frac{\tanh(r)}{r} \end{aligned} \tag{33}$$

Then for $\mathrm{E}[\tanh(\hat{M})^2]$:

$$\begin{aligned} \mathrm{E}[\sigma(\hat{M})^2] &= \int_{-\infty}^{\infty} \sigma(x)^2 f_{U[-r,r]}(x)\,\mathrm{d}x \\ &= \frac{1}{2r}\int_{-r}^{r} \sigma(x)^2\,\mathrm{d}x \\ &= \frac{1}{2r}\left(r - \tanh\left(\frac{r}{2}\right)\right) \end{aligned} \tag{34}$$

Finally this gives the variance:

$$\mathrm{Var}[W] = \frac{1}{2r}\left(1 - \frac{\tanh(r)}{r}\right)\left(r - \tanh\left(\frac{r}{2}\right)\right) \tag{35}$$

B.3 Expectation and variance of $\mathrm{NAC}_{\bullet}$

B.3.1 Forward pass

Expectation Assuming that each $z_{h_{\ell-1}}$ are uncorrelated the expectation can be simplified to:

$$\begin{aligned}
E[z_{h_\ell}] &= E\left[\exp\left(\sum_{h_{\ell-1}=1}^{H_{\ell-1}} W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon)\right)\right] \\
&= E\left[\prod_{h_{\ell-1}=1}^{H_{\ell-1}} \exp(W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon))\right] \\
&\approx \prod_{h_{\ell-1}=1}^{H_{\ell-1}} E[\exp(W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon))] \\
&= E[\exp(W_{h_\ell,h_{\ell-1}} \log(|z_{h_{\ell-1}}| + \epsilon))]^{H_{\ell-1}} \\
&= E\left[(|z_{h_{\ell-1}}| + \epsilon)^{W_{h_\ell,h_{\ell-1}}}\right]^{H_{\ell-1}} \\
&= E\left[f(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})\right]^{H_{\ell-1}}
\end{aligned} \tag{36}$$

Here we define g as a non-linear transformation function of two independent stochastic variables:

$$f(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}}) = (|z_{h_{\ell-1}}| + \epsilon)^{W_{h_\ell,h_{\ell-1}}} \tag{37}$$

We then take the second order Taylor approximation of g, around $(E[z_{h_{\ell-1}}], E[W_{h_\ell,h_{\ell-1}}])$.

$$\begin{aligned}
&E[f(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})] \approx E\Bigg[\\
&g(E[z_{h_{\ell-1}}], E[W_{h_\ell,h_{\ell-1}}]) \\
&+ \begin{bmatrix} z_{h_{\ell-1}} - E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} - E[W_{h_\ell,h_{\ell-1}}] \end{bmatrix}^T \begin{bmatrix} \frac{\partial g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial z_{h_{\ell-1}}} \\ \frac{\partial g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial W_{h_\ell,h_{\ell-1}}} \end{bmatrix} \Bigg|_{\begin{cases} z_{h_{\ell-1}} = E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} = E[W_{h_\ell,h_{\ell-1}}] \end{cases}} \\
&+ \frac{1}{2} \begin{bmatrix} z_{h_{\ell-1}} - E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} - E[W_{h_\ell,h_{\ell-1}}] \end{bmatrix}^T \\
&\bullet \begin{bmatrix} \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial^2 z_{h_{\ell-1}}} & \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial z_{h_{\ell-1}} \partial W_{h_\ell,h_{\ell-1}}} \\ \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial z_{h_{\ell-1}} \partial W_{h_\ell,h_{\ell-1}}} & \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial^2 W_{h_\ell,h_{\ell-1}}} \end{bmatrix} \Bigg|_{\begin{cases} z_{h_{\ell-1}} = E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} = E[W_{h_\ell,h_{\ell-1}}] \end{cases}} \\
&\bullet \begin{bmatrix} z_{h_{\ell-1}} - E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} - E[W_{h_\ell,h_{\ell-1}}] \end{bmatrix} \Bigg]
\end{aligned} \tag{38}$$

Because $E[z_{h_{\ell-1}} - E[z_{h_{\ell-1}}]] = 0$, $E[W_{h_\ell,h_{\ell-1}} - E[W_{h_\ell,h_{\ell-1}}]] = 0$, and $Cov[z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}}] = 0$. This simplifies to:

$$\begin{aligned}
&E[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})] \approx g(E[z_{h_{\ell-1}}], E[W_{h_\ell,h_{\ell-1}}]) \\
&+ \frac{1}{2} Var \begin{bmatrix} z_{h_{\ell-1}} \\ W_{h_\ell,h_{\ell-1}} \end{bmatrix}^T \begin{bmatrix} \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial^2 z_{h_{\ell-1}}} \\ \frac{\partial^2 g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})}{\partial^2 W_{h_\ell,h_{\ell-1}}} \end{bmatrix} \Bigg|_{\begin{cases} z_{h_{\ell-1}} = E[z_{h_{\ell-1}}] \\ W_{h_\ell,h_{\ell-1}} = E[W_{h_\ell,h_{\ell-1}}] \end{cases}}
\end{aligned} \tag{39}$$

Inserting the derivatives and computing the inner products yields:

$$\begin{aligned}
&E[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})] \approx (|E[z_{h_{\ell-1}}]| + \epsilon)^{E[W_{h_\ell,h_{\ell-1}}]} \\
&+ \frac{1}{2} Var[z_{h_{\ell-1}}](|E[z_{h_{\ell-1}}]| + \epsilon)^{E[W_{h_\ell,h_{\ell-1}}]-2} E[W_{h_\ell,h_{\ell-1}}](E[W_{h_\ell,h_{\ell-1}}] - 1) \\
&+ \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}](|E[z_{h_{\ell-1}}]| + \epsilon)^{E[W_{h_\ell,h_{\ell-1}}]} \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2 \\
&= 1 + \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2
\end{aligned} \tag{40}$$

This gives the final expectation:

$$\begin{aligned}
E[z_{h_\ell}] &= E\left[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})\right]^{H_{\ell-1}} \\
&\approx \left(1 + \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{H_{\ell-1}}
\end{aligned} \tag{41}$$

As this expectation is of particular interest, we evaluate the error of the approximation, where $W_{h_\ell,h_{\ell-1}} \sim U[-r_w, r_w]$ and $z_{h_{\ell-1}} \sim U[0, r_z]$. These distributions are what is used in the simple function task. The error is plotted in figure 4.

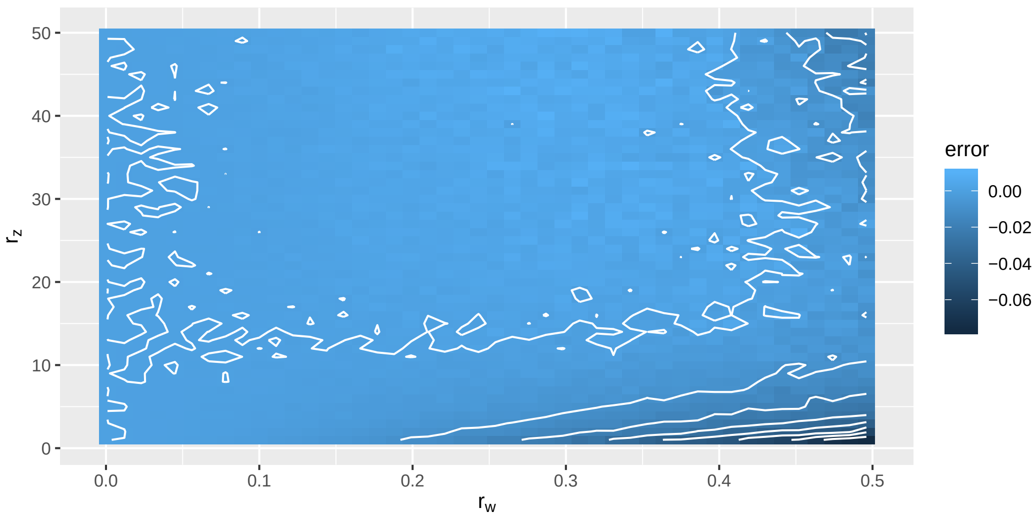


Figure 4: Error between theoretical approximation and the numerical approximation estimated by random sampling of 100000 observations at each combination of r_z and r_w.

Variance The variance can be derived using the same assumptions for expectation, that all $z_{h_{\ell-1}}$ are uncorrelated.

$$\begin{aligned}
Var[z_{h_\ell}] &= E[z_{h_\ell}^2] - E[z_{h_\ell}]^2 \\
&= E\left[\prod_{h_{\ell-1}=1}^{H_{\ell-1}} (|z_{h_{\ell-1}}| + \epsilon)^{2 \cdot W_{h_\ell,h_{\ell-1}}}\right] - E\left[\prod_{h_{\ell-1}=1}^{H_{\ell-1}} (|z_{h_{\ell-1}}| + \epsilon)^{W_{h_\ell,h_{\ell-1}}}\right]^2 \\
&= E\left[g(z_{h_{\ell-1}}, 2 \cdot W_{h_\ell,h_{\ell-1}})\right]^{H_{\ell-1}} - E\left[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})\right]^{2 \cdot H_{\ell-1}}
\end{aligned} \tag{42}$$

We already have from the expectation result that:

$$E\left[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})\right] \approx 1 + \frac{1}{2} Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2 \tag{43}$$

By substitution of variable we have that:

$$
\begin{aligned}
E\left[g(z_{h_{\ell-1}}, 2 \cdot W_{h_\ell,h_{\ell-1}})\right] &\approx 1 + \frac{1}{2} Var[2 \cdot W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2 \\
&= \approx 1 + 2 \cdot Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2
\end{aligned} \tag{44}
$$

This gives the variance:

$$
\begin{aligned}
Var[z_{h_\ell}] = {} & E\left[g(z_{h_{\ell-1}}, 2 \cdot W_{h_\ell,h_{\ell-1}})\right]^{H_{\ell-1}} - E\left[g(z_{h_{\ell-1}}, W_{h_\ell,h_{\ell-1}})\right]^{2 \cdot H_{\ell-1}} \\
& \approx \left(1 + 2 \cdot Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{H_{\ell-1}} \\
& - \left(1 + \frac{1}{2} \cdot Var[W_{h_\ell,h_{\ell-1}}] \log(|E[z_{h_{\ell-1}}]| + \epsilon)^2\right)^{2 \cdot H_{\ell-1}}
\end{aligned} \tag{45}
$$

B.3.2 Backward pass

Expectation The expectation of the back-propagation term assuming that $\delta_{h_{\ell+1}}$ and $\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}$ are mutually uncorrelated:

$$
E[\delta_{h_\ell}] = E\left[\sum_{h_{\ell+1}=1}^{H_{\ell+1}} \delta_{h_{\ell+1}} \frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}\right] \approx H_{\ell+1} E[\delta_{h_{\ell+1}}] E\left[\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}\right] \tag{46}
$$

Assuming that $z_{h_{\ell+1}}$, $W_{h_{\ell+1},h_\ell}$, and z_{h_ℓ} are uncorrelated:

$$
E\left[\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}\right] \approx E[z_{h_{\ell+1}}] E[W_{h_{\ell+1},h_\ell}] E\left[\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right] = E[z_{h_{\ell+1}}] \cdot 0 \cdot E\left[\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right] = 0 \tag{47}
$$

Variance Deriving the variance is more complicated as:

$$
Var\left[\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}\right] = Var\left[z_{h_{\ell+1}} W_{h_{\ell+1},h_\ell} \frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right] \tag{48}
$$

Assuming again that $z_{h_{\ell+1}}$, $W_{h_{\ell+1},h_\ell}$, and z_{h_ℓ} are uncorrelated, and likewise for their second moment:

$$
\begin{aligned}
Var\left[\frac{\partial z_{h_{\ell+1}}}{\partial z_{h_\ell}}\right] &\approx E[z^2_{h_{\ell+1}}] E[W^2_{h_{\ell+1},h_\ell}] E\left[\left(\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right)^2\right] \\
&\quad - E[z_{h_{\ell+1}}]^2 E[W_{h_{\ell+1},h_\ell}]^2 E\left[\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right]^2 \\
&= E[z^2_{h_{\ell+1}}] Var[W_{h_{\ell+1},h_\ell}] E\left[\left(\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right)^2\right] \\
&\quad - E[z_{h_{\ell+1}}]^2 \cdot 0 \cdot E\left[\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right]^2 \\
&= E[z^2_{h_{\ell+1}}] Var[W_{h_{\ell+1},h_\ell}] E\left[\left(\frac{\text{abs}'(z_{h_\ell})}{|z_{h_\ell}| + \epsilon}\right)^2\right]
\end{aligned} \tag{49}
$$

Using Taylor approximation around $E[z_{h_\ell}]$ we have:

$$
\begin{aligned}
E\left[\left(\frac{\text{abs}'(z_{h_\ell})}{|z| + \epsilon}\right)^2\right] &\approx \frac{1}{(|E[z_{h_\ell}]| + \epsilon)^2} + \frac{1}{2} \frac{6}{(|E[z_{h_\ell}]| + \epsilon)^4} Var[z_{h_\ell}] \\
&= \frac{1}{(|E[z_{h_\ell}]| + \epsilon)^2} + \frac{3}{(|E[z_{h_\ell}]| + \epsilon)^4} Var[z_{h_\ell}]
\end{aligned} \tag{50}
$$

Finally, by reusing the result for $E[z_{h_\ell}^2]$ from earlier the variance can be expressed as:

$$\begin{aligned} Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] \approx Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \, \left(1 + 2\cdot Var[W_{h_\ell,h_{\ell-1}}]\log(|E[z_{h_{\ell-1}}]|+\epsilon)^2\right)^{H_{\ell-1}} \\ \cdot Var[W_{h_\ell,h_{\ell-1}}] \left(\frac{1}{\left(|E[z_{h_{\ell-1}}]|+\epsilon\right)^2} + \frac{3}{\left(|E[z_{h_{\ell-1}}]|+\epsilon\right)^4} Var[z_{h_{\ell-1}}]\right) \end{aligned} \tag{51}$$

B.4 Expectation and variance of NMU

B.4.1 Forward pass

Expectation Assuming that all $z_{h_{\ell-1}}$ are independent:

$$\begin{aligned} E[z_{h_\ell}] &= E\left[\prod_{h_{\ell-1}=1}^{H_{\ell-1}} \left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right)\right] \\ &\approx E\left[W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right]^{H_{\ell-1}} \\ &\approx \left(E[W_{h_{\ell-1},h_\ell}]E[z_{h_{\ell-1}}] + 1 - E[W_{h_{\ell-1},h_\ell}]\right)^{H_{\ell-1}} \end{aligned} \tag{52}$$

Assuming that $E[z_{h_{\ell-1}}] = 0$ which is a desired property and initializing $E[W_{h_{\ell-1},h_\ell}] = 1/2$, the expectation is:

$$\begin{aligned} E[z_{h_\ell}] &\approx \left(E[W_{h_{\ell-1},h_\ell}]E[z_{h_{\ell-1}}] + 1 - E[W_{h_{\ell-1},h_\ell}]\right)^{H_{\ell-1}} \\ &\approx \left(\frac{1}{2}\cdot 0 + 1 - \frac{1}{2}\right)^{H_{\ell-1}} \\ &= \left(\frac{1}{2}\right)^{H_{\ell-1}} \end{aligned} \tag{53}$$

Variance Reusing the result for the expectation, assuming again that all $z_{h_{\ell-1}}$ are uncorrelated, and using the fact that $W_{h_{\ell-1},h_\ell}$ is initially independent from $z_{h_{\ell-1}}$:

$$\begin{aligned} Var[z_{h_\ell}] &= E[z_{h_\ell}^2] - E[z_{h_\ell}]^2 \\ &\approx E[z_{h_\ell}^2] - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &= E\left[\prod_{h_{\ell-1}=1}^{H_{\ell-1}} \left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right)^2\right] - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &\approx E[\left(W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}\right)^2]^{H_{\ell-1}} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &= \Big(E[W_{h_{\ell-1},h_\ell}^2]E[z_{h_{\ell-1}}^2] - 2E[W_{h_{\ell-1},h_\ell}^2]E[z_{h_{\ell-1}}] + E[W_{h_{\ell-1},h_\ell}^2] \\ &\quad + 2E[W_{h_{\ell-1},h_\ell}]E[z_{h_{\ell-1}}] - 2E[W_{h_{\ell-1},h_\ell}] + 1\Big)^{H_{\ell-1}} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \end{aligned} \tag{54}$$

Assuming again that $E[z_{h_{\ell-1}}] = 0$, which is a desired property and initializing $E[W_{h_{\ell-1},h_\ell}] = 1/2$, the variance becomes:

$$\begin{aligned} Var[z_{h_\ell}] &\approx \left(E[W_{h_{\ell-1},h_\ell}^2]\left(E[z_{h_{\ell-1}}^2] + 1\right)\right)^{H_{\ell-1}} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &\approx \left(\left(Var[W_{h_{\ell-1},h_\ell}] + E[W_{h_{\ell-1},h_\ell}]^2\right)\left(Var[z_{h_{\ell-1}}] + 1\right)\right)^{H_{\ell-1}} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &= \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \end{aligned} \tag{55}$$

B.4.2 Backward pass

Expectation For the backward pass the expectation can, assuming that $\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}$ and $\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}$ are uncorrelated, be derived to:

$$\begin{aligned} E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &= H_\ell E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \\ &\approx H_\ell E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] E\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \\ &= H_\ell E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] E\left[\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}} W_{h_{\ell-1},h_\ell}\right] \\ &= H_\ell E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] E\left[\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right] E\left[W_{h_{\ell-1},h_\ell}\right] \end{aligned} \tag{56}$$

Initializing again $E[W_{h_{\ell-1},h_\ell}] = 1/2$, and inserting the result for the expectation $E\left[\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right]$.

$$\begin{aligned} E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx H_\ell E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] \left(\frac{1}{2}\right)^{H_{\ell-1}-1} \frac{1}{2} \\ &= E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \left(\frac{1}{2}\right)^{H_{\ell-1}} \end{aligned} \tag{57}$$

Assuming that $E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] = 0$, which is a desired property [**?**].

$$\begin{aligned} E\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx 0 \cdot H_\ell \cdot \left(\frac{1}{2}\right)^{H_{\ell-1}} \\ &= 0 \end{aligned} \tag{58}$$

Variance For the variance of the backpropagation term, we assume that $\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}$ is uncorrelated with $\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}$.

$$\begin{aligned} Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &= H_\ell Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}} \frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \\ &\approx H_\ell \Bigg(Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] E\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right]^2 + E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right]^2 Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \\ &\quad + Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \Bigg) \end{aligned} \tag{59}$$

Assuming again that $E\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] = 0$, and reusing the result $E\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] = \left(\frac{1}{2}\right)^{H_{\ell-1}}$.

$$Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] \approx Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \left(\left(\frac{1}{2}\right)^{2 \cdot H_{\ell-1}} + Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] \right) \tag{60}$$

Focusing now on $Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right]$, we have:

$$\begin{aligned} Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] &= E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right] E[W_{h_{\ell-1},h_\ell}^2] \\ &\quad - E\left[\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right]^2 E[W_{h_{\ell-1},h_\ell}]^2 \end{aligned} \tag{61}$$

Inserting the result for the expectation $E\left[\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right]$ and Initializing again $E[W_{h_{\ell-1},h_\ell}] = 1/2$.

$$\begin{aligned} Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] &\approx E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right] E[W^2_{h_{\ell-1},h_\ell}] \\ &\quad - \left(\frac{1}{2}\right)^{2\cdot(H_{\ell-1}-1)} \left(\frac{1}{2}\right)^2 \\ &= E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right] E[W^2_{h_{\ell-1},h_\ell}] \\ &\quad - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \end{aligned} \tag{62}$$

Using the identity that $E[W^2_{h_{\ell-1},h_\ell}] = Var[W_{h_{\ell-1},h_\ell}] + E[W_{h_{\ell-1},h_\ell}]^2$, and again using $E[W_{h_{\ell-1},h_\ell}] = 1/2$.

$$\begin{aligned} Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] &\approx E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right] \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right) \\ &\quad - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \end{aligned} \tag{63}$$

To derive $E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right]$ the result for $Var[z_{h_\ell}]$ can be used, but for $\hat{H}_{\ell-1} = H_{\ell-1} - 1$, because there is one less term. Inserting $E\left[\left(\frac{z_{h_\ell}}{W_{h_{\ell-1},h_\ell} z_{h_{\ell-1}} + 1 - W_{h_{\ell-1},h_\ell}}\right)^2\right] = \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}-1} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1}$, we have:

$$\begin{aligned} Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right] &\approx \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}-1} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1} \\ &\quad \cdot \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right) - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &= \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \end{aligned} \tag{64}$$

Inserting the result for $Var\left[\frac{\partial z_{h_\ell}}{\partial z_{h_{\ell-1}}}\right]$ into the result for $Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right]$:

$$\begin{aligned} Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_{\ell-1}}}\right] &\approx Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \Bigg(\left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \\ &\quad + \left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1} - \left(\frac{1}{2}\right)^{2\cdot H_{\ell-1}} \Bigg) \\ &= Var\left[\frac{\partial \mathcal{L}}{\partial z_{h_\ell}}\right] H_\ell \\ &\quad \cdot \left(\left(Var[W_{h_{\ell-1},h_\ell}] + \frac{1}{4}\right)^{H_{\ell-1}} \left(Var[z_{h_{\ell-1}}] + 1\right)^{H_{\ell-1}-1}\right) \end{aligned} \tag{65}$$

B.4.3 Initialization

The $W_{h_{\ell-1},h_\ell}$ should be initialized with $E[W_{h_{\ell-1},h_\ell}] = \frac{1}{2}$, in order to not bias towards inclusion or exclusion of $z_{h_{\ell-1}}$. Using the derived variance approximations, the variance should be according to the forward pass:

$$Var[W_{h_{\ell-1},h_\ell}] = \left((1+Var[z_{h_\ell}])^{-H_{\ell-1}} Var[z_{h_\ell}] + (4+4Var[z_{h_\ell}])^{-H_{\ell-1}}\right)^{\frac{1}{H_{\ell-1}}} - \frac{1}{4} \quad (66)$$

And according to the backward pass it should be:

$$Var[W_{h_{\ell-1},h_\ell}] = \left(\frac{(Var[z_{h_\ell}]+1)^{1-H_{\ell-1}}}{H_\ell}\right)^{\frac{1}{H_{\ell-1}}} - \frac{1}{4} \quad (67)$$

Both criteria are dependent on the input variance. If the input variance is know then optimal initialization is possible. However, as this is often not the case one can perhaps assume that $Var[z_{h_{\ell-1}}] = 1$. This is not an unreasonable assumption in many cases, as there may either be a normalization layer somewhere or the input is normalized. If unit variance is assumed, one gets from the forward pass:

$$Var[W_{h_{\ell-1},h_\ell}] = \left(2^{-H_{\ell-1}} + 8^{-H_{\ell-1}}\right)^{\frac{1}{H_{\ell-1}}} - \frac{1}{4} = \frac{1}{8}\left(\left(4^{H_{\ell-1}}+1\right)^{H_{\ell-1}} - 2\right) \quad (68)$$

And from the backward pass:

$$Var[W_{h_{\ell-1},h_\ell}] = \left(\frac{2^{1-H_{\ell-1}}}{H_\ell}\right)^{\frac{1}{H_{\ell-1}}} - \frac{1}{4} \quad (69)$$

The variance requirement for both the forward and backward pass can be satisfied with $Var[W_{h_{\ell-1},h_\ell}] = \frac{1}{4}$ for a large $H_{\ell-1}$.

C Simple function task

C.1 Dataset generation

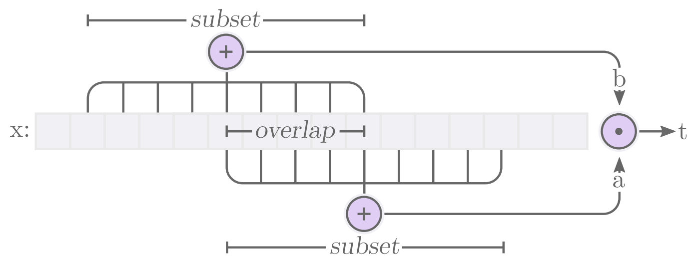


Figure 5: Dataset is parameterized into “Input Size”, “Subset Ratio”, “Overlap Ratio”, an Operation (here showing multiplication), “Interpolation Range” and “Extrapolation Range” from which the data set sampled.

All datasets in the simple function task experiments are generated using algorithm 1. Its parameters are visualized in figure 5.

Algorithm 1 Dataset sampling algorithm

function $\textsc{Dataset}(\textsc{Op}(\cdot,\cdot) : \text{Operation},\ i : \text{InputSize},\ s : \text{SubsetRatio},\ o : \text{OverlapRatio},$
$R : \text{Range})$

$\mathbf{x} \leftarrow \textsc{Uniform}(R_{lower}, R_{upper}, i)$ ▷ Sample i elements uniformly

$k \leftarrow \textsc{Uniform}(0, 1 - 2s - o)$ ▷ Sample offset

$a \leftarrow \textsc{Sum}(\mathbf{x}[ik : i(k + s)])$ ▷ Create sum a from subset

$b \leftarrow \textsc{Sum}(\mathbf{x}[i(k + s - o) : i(k + 2s - 0)])$ ▷ Create sum b from subset

$t \leftarrow \textsc{Op}(a, b)$ ▷ Perform operation on a and b

return x, t

C.2 Arithmetic operations comparison - all models

Table 4: Model definitions

Model	Layer 1	Layer 2
NMU	NAU	NMU
NAU	NAU	NAU
$\text{NAC}_{\bullet}$	NAC_{+}	$\text{NAC}_{\bullet}$
NAC_{+}	NAC_{+}	NAC_{+}
NALU	NALU	NALU
Linear	Linear	Linear
ReLU	ReLU	ReLU
ReLU6	ReLU6	ReLU6

Results for all models on addition, subtraction, and multiplication can be found in table 5.

Table 5: Shows the success-rate for $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, at what global step the model converged at, and the sparsity error for all weight matrices.

Op	Model	Success	Solved at		Sparsity error
		Rate	Median	Mean	Mean
$\times$	NAC$_\bullet$	30%	$2.5 \cdot 10^6$	$2.5 \cdot 10^6 \pm 1.5 \cdot 10^6$	$\mathbf{3.9 \cdot 10^{-4} \pm 9.4 \cdot 10^{-4}}$
	NAC$_+$	0%	—	—	—
	Linear	0%	—	—	—
	NALU	0%	—	—	—
	NAU	0%	—	—	—
	NMU	**90%**	$\mathbf{1.4 \cdot 10^6}$	$\mathbf{1.6 \cdot 10^6 \pm 5.6 \cdot 10^5}$	$1.8 \cdot 10^{-3} \pm 1.1 \cdot 10^{-3}$
	ReLU	0%	—	—	—
	ReLU6	0%	—	—	—
$+$	NAC$_\bullet$	0%	—	—	—
	NAC$_+$	**100%**	$6.0 \cdot 10^4$	$7.1 \cdot 10^4 \pm 2.4 \cdot 10^4$	$4.8 \cdot 10^{-1} \pm 2.0 \cdot 10^{-2}$
	Linear	**100%**	$4.2 \cdot 10^4$	$\mathbf{4.2 \cdot 10^4 \pm 1.9 \cdot 10^3}$	$6.1 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$
	NALU	0%	—	—	—
	NAU	**100%**	$\mathbf{1.8 \cdot 10^4}$	$7.0 \cdot 10^5 \pm 9.2 \cdot 10^5$	$\mathbf{1.7 \cdot 10^{-3} \pm 8.0 \cdot 10^{-4}}$
	NMU	0%	—	—	—
	ReLU	80%	$4.2 \cdot 10^4$	$8.4 \cdot 10^5 \pm 1.1 \cdot 10^6$	$7.3 \cdot 10^{-1} \pm 2.3 \cdot 10^{-1}$
	ReLU6	0%	—	—	—
$-$	NAC$_\bullet$	0%	—	—	—
	NAC$_+$	**100%**	$8.0 \cdot 10^3$	$1.5 \cdot 10^6 \pm 1.5 \cdot 10^6$	$4.6 \cdot 10^{-1} \pm 2.9 \cdot 10^{-2}$
	Linear	**100%**	$1.1 \cdot 10^6$	$1.9 \cdot 10^6 \pm 1.3 \cdot 10^6$	$3.7 \cdot 10^{-1} \pm 1.1 \cdot 10^{-1}$
	NALU	20%	$3.6 \cdot 10^6$	$3.6 \cdot 10^6 \pm 1.3 \cdot 10^7$	$4.7 \cdot 10^{-1} \pm 3.3 \cdot 10^{-1}$
	NAU	**100%**	$\mathbf{4.0 \cdot 10^3}$	$\mathbf{4.2 \cdot 10^3 \pm 3.0 \cdot 10^2}$	$1.9 \cdot 10^{-3} \pm 4.2 \cdot 10^{-4}$
	NMU	60%	$3.1 \cdot 10^5$	$3.0 \cdot 10^5 \pm 8.8 \cdot 10^3$	$\mathbf{1.2 \cdot 10^{-4} \pm 1.6 \cdot 10^{-4}}$
	ReLU	0%	—	—	—
	ReLU6	0%	—	—	—

C.3 Ablation study

To validate our model, we perform an ablation on the multiplication problem.

Our ablation study (table 6) show that regularization have little effect in terms of success rate. As it is analytically known that there is no gradient outside of $w \in [0, 1]$ for the NMU, the conclusion must be that the optimal weight initialization for the default dataset parameters and tested seeds, does not cause any weights to accidentally break out of $w \in [0, 1]$. The sparse regularizer for multiplication have no sparsity effect, as only a sparse solution is a valid solution for multiplication. Although as seen in appendix C.4, sparsity regularization can improve convergence.

Not allowing a multiplicative identity ($\mathbf{z} = \mathbf{W} \odot \mathbf{x}$), works when there is only two hidden units in the multiplication layer, as no multiplicative identity is necessary. However, for larger a hidden size as seen in figure 6 identity becomes necessary.

Table 6: Shows the success-rate for $\mathcal{L}_{\mathbf{W}_1,\mathbf{W}_2} < \mathcal{L}_{\mathbf{W}_1^\epsilon,\mathbf{W}_2^*}$, at what global step the model converged at, and the sparsity error for all weight matrices. The dataset is the multiplication problem with default parameters.

Model	Success	Solved at		Sparsity error
	Rate	Median	Mean	Mean
$\mathrm{NAC}_\bullet$	40%	$2.8 \cdot 10^6$	$3.1 \cdot 10^6 \pm 2.0 \cdot 10^6$	$2.8 \cdot 10^{-2} \pm 8.9 \cdot 10^{-2}$
$\mathrm{NAC}_\bullet$, $\mathbf{W} = \sigma(\hat{\mathbf{W}})$	100%	$1.9 \cdot 10^6$	$1.9 \cdot 10^6 \pm 3.1 \cdot 10^5$	$1.1 \cdot 10^{-4} \pm 1.0 \cdot 10^{-4}$
NMU	100%	$1.2 \cdot 10^6$	$1.2 \cdot 10^6 \pm 2.0 \cdot 10^5$	$1.6 \cdot 10^{-3} \pm 9.2 \cdot 10^{-4}$
NMU, $\mathbf{W} = \hat{\mathbf{W}}$	100%	$1.3 \cdot 10^6$	$1.2 \cdot 10^6 \pm 1.9 \cdot 10^5$	$3.9 \cdot 10^{-3} \pm 1.2 \cdot 10^{-3}$
NMU, $\mathbf{z} = \mathbf{W} \odot \mathbf{x}$	100%	$1.2 \cdot 10^6$	$1.2 \cdot 10^6 \pm 2.0 \cdot 10^5$	$1.6 \cdot 10^{-3} \pm 9.2 \cdot 10^{-4}$
NMU, no $\mathcal{R}_{oob}$	100%	$1.2 \cdot 10^6$	$1.2 \cdot 10^6 \pm 1.9 \cdot 10^5$	$1.7 \cdot 10^{-3} \pm 4.6 \cdot 10^{-4}$
NMU, no $\mathcal{R}_{sparse}$, $\mathcal{R}_{oob}$	100%	$1.1 \cdot 10^6$	$1.1 \cdot 10^6 \pm 1.8 \cdot 10^5$	$3.3 \cdot 10^{-4} \pm 4.5 \cdot 10^{-5}$
NMU, no $\mathcal{R}_{sparse}$	100%	$1.2 \cdot 10^6$	$1.2 \cdot 10^6 \pm 1.9 \cdot 10^5$	$1.7 \cdot 10^{-3} \pm 9.0 \cdot 10^{-4}$

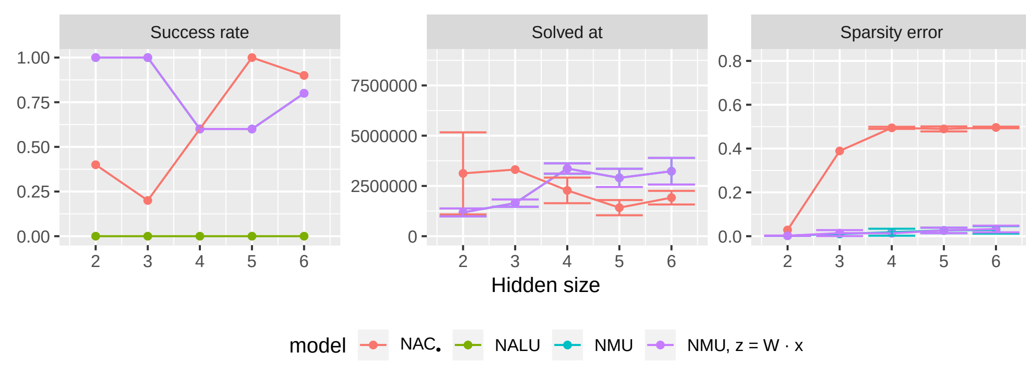


Figure 6: Compares NMU with NMU without identity, with different input size (hidden layer unit size) to the multiplication layer.

C.4 Regularization

A high sparsity regularization constant can help the model to converge faster. However, a regularization constant too high have have the inverse effect as well, or even make it impossible for the model to converge. regularizer In these experiments, the constant c in equation 70 is varied. See results in figure 7, 8, and 9.

$$\lambda_{\text{bias}} = c \cdot (1 - \exp(-10^5 \cdot t)) \tag{70}$$

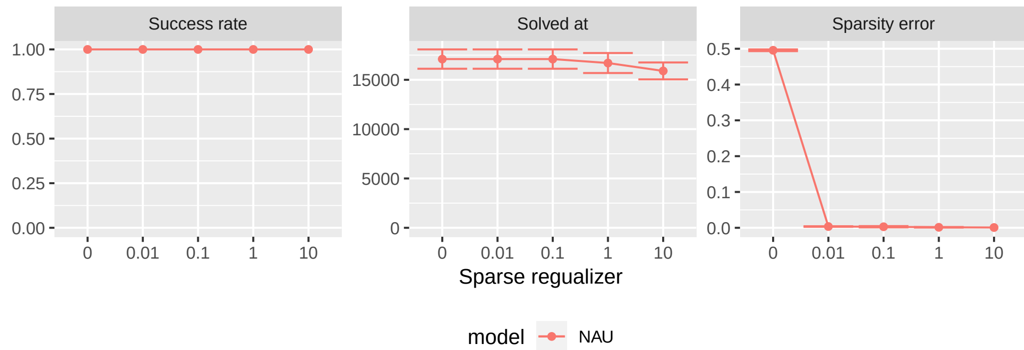


Figure 7: Shows the effect of the regularizer λ_{bias}, on the simple function task problem for the $+$ operation.

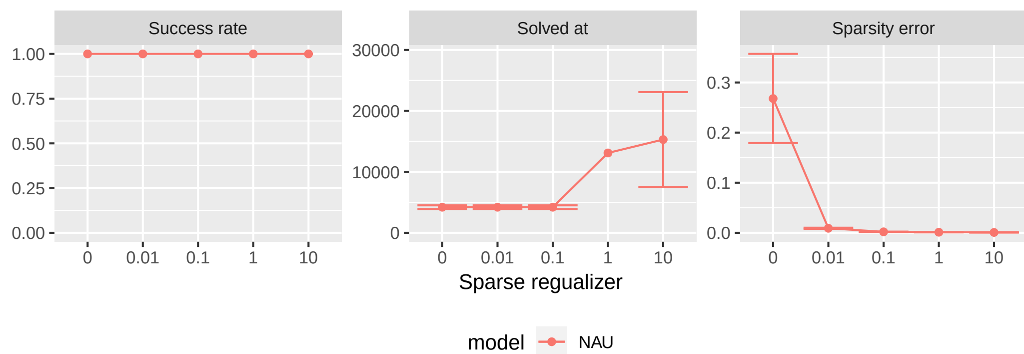


Figure 8: Shows the effect of the regularizer λ_{bias}, on the simple function task problem for the $-$ operation.

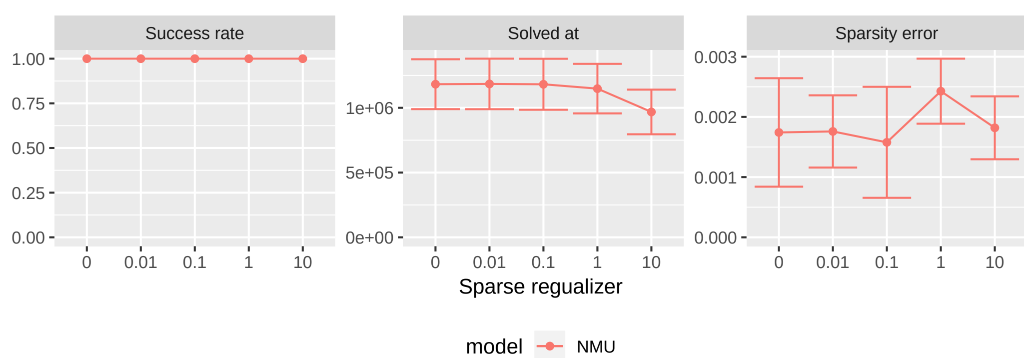


Figure 9: Shows the effect of the regularizer λ_{bias}, on the simple function task problem for the $\times$ operation.