L1-regularization: The path to sparsity

Laura Ying Schulz Tomaso Poggio

Abstract

This study investigates the effects of L1-regularization on inducing sparsity in neural networks, specifically within a teacher-student framework. Various experiments were conducted to evaluate both the convergence and emergence of sparsity. The results show that L1-regularization is the only regularization technique that reliably drives weights to zero. It achieves near-perfect sparsity, especially when the teacher and student share the same activation function. While initialization is crucial, these results offer practical guidance for optimizing student networks that accurately approximate teachers and recover sparse, interpretable representations.

The code and all results can be found on Github.

1. Introduction

Deep learning [2, 7] has become a foundational pillar in today's artificial intelligence, showing strong capabilities across various domains in recent years. However, the underlying reasons why certain deep network architectures work so well remain to be understood.

Recent theorems suggest that all learnable functions that are efficiently Turing computable can be represented as the compositions of functions of a smaller number of variables. This, in turn, implies that each such function can be approximated by a deep network without the curse of dimensionality [10, 12]. Additionally, many works showed empirically [1, 3] and it has been proven [9] that generalization can be orders of magnitude better in networks with sparse weight matrices.

However, the specific conditions under which sparse networks can be successfully optimized remain an open question. This project investigates the potential of L1-regularization and the role of the initialization weights to consistently induce high levels of sparsity in neural networks while maintaining performance. Specifically, six experiments are conducted, spanning L1-, L2-, and no regularization with different activation functions (ReLU, Sigmoid, Tanh) and initialization scales. Each setup is tested across five random seeds.

2. Related Work

Regularization techniques are widely used in machine learning to improve generalization and control model complexity. L1-regularization, also known as Lasso regression [14], has received attention for inducing sparsity in learned representations. In linear models, it not only produces sparse solutions but also helps with automatic feature selection. This fundamental idea has been extended to neural networks. When L1-regularization is applied to weight matrices, it can improve interpretability and mitigate overfitting [11].

Building on these principles, Wen et al.(2016) explored structured sparsity in deep networks using group Lasso. Their work demonstrated that L1-type penalties can prune entire filters or neurons, effectively compressing networks. Scardapane et al.(2017) provided an overview of sparsity-inducing strategies in deep learning. They highlighted their role in reducing computational complexity and speeding up inference.

However, a key challenge remains: the non-deterministic nature of deep learning optimization. Li et al. (2015) showed that different initializations can lead to substantially different learned representations. This raises questions about the consistency of sparsity-inducing methods. This observation motivates the search for approaches that reliably induce sparsity across various training conditions, thereby enhancing robust optimization of deep neural networks.

3. Method and Evaluation

3.1. Experimental Setup

A teacher-student framework is deployed, in which the objective is for the student model to replicate the internal representations learned by the teacher. To maintain full control of the experimental setup, a synthetic dataset is generated in which each data point is a 12-dimensional vector sampled from a standard normal distribution. The corresponding target values are obtained by passing these vectors through the teacher model. This dataset is used to train the student model.

This study investigates the effects of L1-regularization, L2-regularization, and the absence of regularization in models using different activation functions. In addition, it ex-

plores the effect of weight initialization in promoting sparsity. Specifically, we apply standard initialization schemes: Kaiming initialization for ReLU and Xavier uniform initialization for Sigmoid and Tanh. The impact of scaling these initializations by a factor of 0.2 will then be examined. The goal is to determine whether reducing the initialization magnitude helps the model converge to sparser solutions under different regularization strategies.

L1-regularization, defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda \sum_i |w_i|,$$

encourages sparsity by penalizing the absolute values of the model weights. L2-regularization [5], on the other hand, adds a penalty proportional to the squared values of the weights

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda \sum_i w_i^2,$$

and is known for promoting small weights which often increases smoothness. Here, $\mathcal{L}_{\text{data}}$ denotes the original loss [1], w_i are the model weights, and λ is the regularization strength. The third case serves as a baseline with no regularization, where $\lambda = 0$.

3.2. Model Architectures

Various model architectures are explored, each consisting of three weight layers and the same activation function being applied at every layer. For convolutional neural networks (CNNs) [4], all convolutions are one-dimensional. The architectures are visualized in Figures 1, and 2.

- **Baseline-CNN**: A baseline model where the first convolutional layer has a kernel size and stride of 3, and the last two layers have a kernel size and stride of 2. All layers use a single input and output channel.
- **SplitFilter-CNN**: This architecture is based on the Baseline-CNN, but doubles the capacity of the first two convolutional layers. It achieves this by introducing separate filters for different regions of the input while keeping the kernel size and stride settings unchanged.
- **MultiChannel-CNN**: The kernel and stride sizes match those of the Baseline-CNN but with increased channel widths: The first layer uses 1 input and 4 output channels, the second has 4 input and output channels, and the final layer outputs a single channel.
- **FCN-256-32**: A fully connected network where the first hidden layer has 256 nodes and the second hidden layer has 32 nodes.
- **FCN-128-128**: This model consists of fully connected linear layers with two hidden layers of 128 nodes each.

[1] the mean squared error in our case

3.3. Evaluation Metric

The primary evaluation metric is the final mean squared error (MSE) between the teacher and student outputs at the end of training:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^{\text{teacher}} - y_i^{\text{student}} \right)^2,$$

where N denotes the dataset size.

For dense student networks, sparsity is computed per layer as the proportion of weights that are approximately zero [2]. Specifically, a weight is considered zero if its absolute value is smaller than 10^{-4}. Formally, the sparsity of a weight matrix is defined as:

$$\text{sparsity} = \frac{\#\{w \in W \mid |w| < 10^{-4}\}}{\text{total number of elements in } W}.$$

To complement this quantitative measure, visualizations of the weight matrices are used to inspect emergent structural patterns.

3.4. Specific Experiments

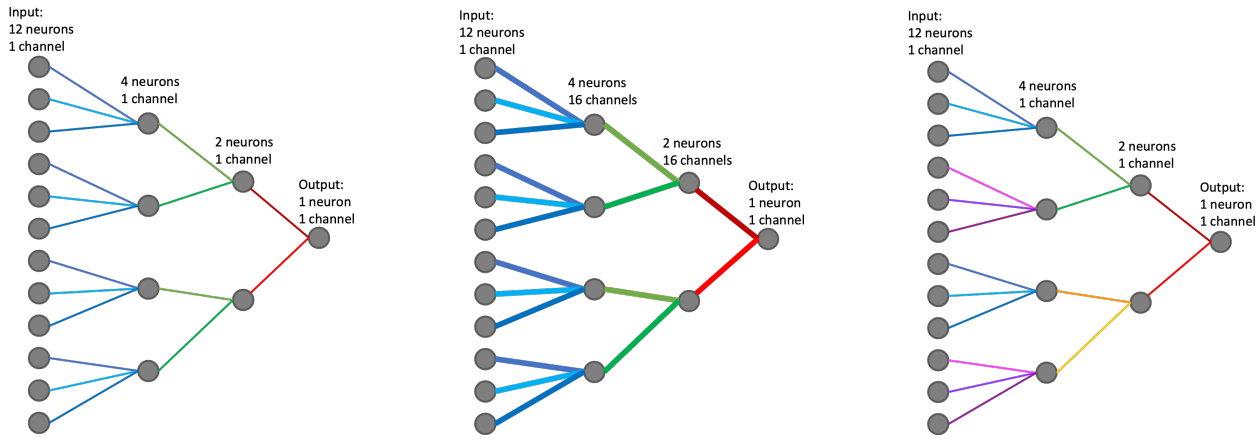
Unless otherwise specified, the teacher model is a Baseline-CNN. The weights remain fixed across all experiments to ensure that the student always tries to learn the same target mapping.

The students are trained using stochastic gradient descent (SGD) with a momentum of 0.9 [6] and early stopping. The default hyperparameters are a batch size of 16, learning rate of 0.005, and a regularization coefficient λ of 10^{-5}. The dataset contains 5,120 data points for CNN-based student models, and 100,000 for FCN-based student models.

Six experiments of increasing complexity are conducted. In each experiment, we examine the activation functions, regularization types, and initialization scales to inspect their effects on convergence and sparsity.

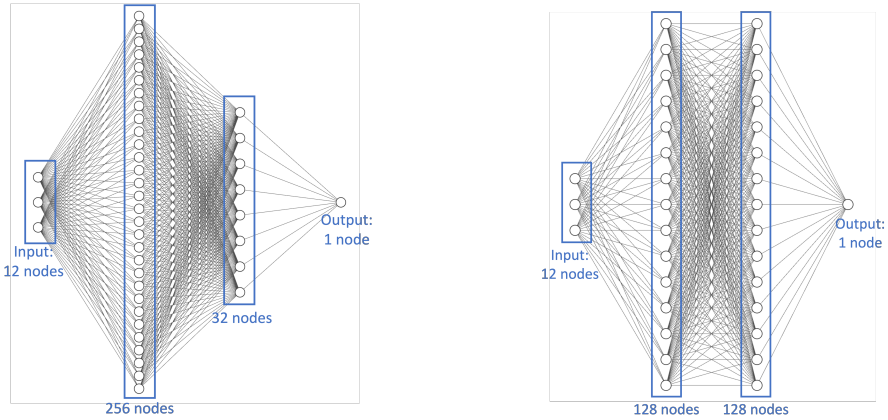
- **Experiment 1**: The student model is a Baseline-CNN. This should validate whether the student can replicate the teacher's internal representations.
- **Experiment 2**: The student is a MultiChannel-CNN. This tests whether a higher-capacity CNN can successfully learn a sparser teacher model.
- **Experiment 3**: The student is an FCN-256-32, which evaluates if an overparameterized fully connected network reduces to the teacher's sparsity pattern. Ideally, after training, the student has 12 weights in the first layer, 4 in the second, and 2 in the last layer.

[2] Note that a sparsity of 100% would imply, that all weights are approximately 0. Therefore, a sparsity of 99% will be considered to be close to the optimal result.



(a) Baseline-CNN (b) MultiChannel-CNN (c) SplitFilter-CNN

Figure 1. Visual representations of the CNN-based model architectures. Shared weights are depicted with matching colors.



(a) FCN-256-32 (b) FCN-128-128

Figure 2. Visual representation of the FCN-based model architectures. The number of nodes depicted does not reflect the actual quantity.

- **Experiment 4**: The student is an FCN-128-128. The goal is to see whether sparsity also emerges in this alternative dense FCN model.
- **Experiment 5**: The teacher is a MultiChannel-CNN and the student is an FCN-256-32. This experiment tests whether a dense fully-connected network can recover a more complex convolutional structure while maintaining sparsity in its learned weights.
- **Experiment 6**: The teacher is a SplitFilter-CNN and the student is a MultiChannel-CNN. The goal is to confirm that a slightly more complex sparse function can still be learned.

4. Results

4.1. Convergence under Matched Activations

Across all six experiments, student networks whose activations matched those of their teachers converged to vanishingly small final mean squared error. This finding is consistent no matter the initial weight scaling and regularization technique applied. This demonstrates that functional alignment between teacher and student activations is an important condition for accurate knowledge transfer.

4.2. Sparsity and Regularization

When trained with L1-regularization, the student models can achieve perfect sparsity[3] if the activation functions match those of the teacher. This is even the case under heavy overparameterization (Experiments 3–4) or more complex convolutional teachers (Experiments 5–6).

While L2-regularization effectively drives the weights toward smaller magnitudes, the weights often remain non-zero. In the case of a fully-connected student network, L2-regularization seems to lead to higher sparsity than when

[3]Perfect sparsity refers to a solution in which the student model retains only the minimal necessary set of non-zero weights, with all remaining weights effectively zero (within a tolerance of $\pm 10^{-4}$).

Teacher Act. → Student Act.	L1 Reg. (scaled 0.2)	L1 Reg. (no scaling)	L2 Reg. (scaled 0.2)	L2 Reg. (no scaling)	No Reg. (scaled 0.2)	No Reg. (no scaling)
ReLU → ReLU	97.35	**97.38**	50.58	53.01	0.56	0.02
ReLU → Tanh	**85.95**	77.30	48.07	61.86	0.14	0.24
ReLU → Sigmoid	**70.30**	64.38	31.48	0.31	0.18	0.04
Tanh → ReLU	75.77	**79.82**	33.27	70.38	0.24	0.02
Tanh → Tanh	95.48	**96.00**	30.65	22.03	0.13	1.09
Tanh → Sigmoid	**65.07**	63.25	2.22	3.45	0.20	0.04
Sigmoid → ReLU	83.63	**85.61**	0.12	71.66	0.03	0.04
Sigmoid → Tanh	**98.66**	98.45	4.55	5.04	0.40	0.18
Sigmoid → Sigmoid	66.67	**78.49**	66.67	1.84	0.24	0.05

Table 1. Experiment 3: Average sparsity (%) for each teacher-student activation pair under different regularization methods and initialization scaling. Best values per row are bold.

no regularization technique is applied. However, it likely reflects weights that simply fall below the zero-threshold instead of actually being 0.

Figure 4 illustrates this distinction for the FCN_128_128 in Experiment 4: Under identical hyperparameters, only the L1-regularized model leads to a perfectly sparse weight map. The unregularized model, although achieving zero loss, results in a dense matrix.

Furthermore, the final loss and the achieved sparsity seem to correlate negatively: A low loss usually implies a higher sparsity. The results suggest that sparsity does not emerge early during training, but instead becomes prominent only as the model approaches the optimal solution. Fortunately, the loss typically decreases rapidly to a near-final value, meaning that a reasonably sparse solution is obtained early in training. However, achieving exact sparsity often requires many additional epochs of fine-tuning.

4.3. Role of Initialization Scaling

Scaling the standard weight initialization by a factor of 0.2 lowers the final loss for both L2-regularized and unregularized students (Figure 3, green/orange entries). However, under L1-regularization the effect of scaling is inconsistent.

In the case of Sigmoid-activated students, scaling the initial weights down in the L1-regularization case is counterproductive. Especially when the teacher also uses Sigmoid-activated layers, the student tends to have very small weights (absolute value $< 10^{-4}$) in the first two layers and a dense final layer. Quantitatively, the first two layers often have 100% sparsity [4] and 0.00% sparsity in the last layer.

Notably, when ReLU activations are used in an L2-regularized student where the initial weights are not scaled, the model attains high sparsity but has a high final loss.

[4]Note that all weights are probably not actually 0 but just so low that they fall below the zero-threshold

Teacher Act. → Student Act.	L1 Reg (avg ± std)	L2 Reg (avg ± std)	No Reg (avg ± std)
ReLU → ReLU	0.0003 ± 0.0001	0.0001 ± 0.0001	0 ± 0
ReLU → Tanh	3.5434 ± 0.0286	0 ± 0	0 ± 0
ReLU → Sigmoid	3.3007 ± 0.0283	0 ± 0	0 ± 0
Tanh → ReLU	0.1941 ± 0.0009	0.0386 ± 0.0764	0.0386 ± 0.0756
Tanh → Tanh	0.0004 ± 0	0.0002 ± 0.0003	0 ± 0
Tanh → Sigmoid	0.2940 ± 0.0889	0.0010 ± 0.0014	0.0004 ± 0.0003
Sigmoid → ReLU	0.0111 ± 0.0001	0.0090 ± 0.0005	0.0086 ± 0.0008
Sigmoid → Tanh	0.2528 ± 0	0.2559 ± 0.0238	0.2518 ± 0.0197
Sigmoid → Sigmoid	0.0026 ± 0	0 ± 0	0 ± 0

(a) Scaling the initial weights down by a factor of 0.2

Teacher Act. → Student Act.	L1 Reg (avg ± std)	L2 Reg (avg ± std)	No Reg (avg ± std)
ReLU → ReLU	0.0003 ± 0	4.9156 ± 0.0312	0.4610 ± 0.4316
ReLU → Tanh	3.5330 ± 0.0448	0.0089 ± 0.0175	15.1066 ± 29.6738
ReLU → Sigmoid	3.2904 ± 0.0443	1.8197 ± 1.4710	0.1436 ± 0.1560
Tanh → ReLU	0.1938 ± 0.0012	0.4157 ± 0.0543	0.4774 ± 0.1465
Tanh → Tanh	0.0003 ± 0.0001	0.0001 ± 0	0 ± 0
Tanh → Sigmoid	0.1850 ± 0.0013	0.2430 ± 0.0997	0.2024 ± 0.1146
Sigmoid → ReLU	0.0111 ± 0	0.1890 ± 0.1489	0.2600 ± 0.1265
Sigmoid → Tanh	0.2526 ± 0	0.2751 ± 0.1177	0.3041 ± 0.1875
Sigmoid → Sigmoid	0.0004 ± 0	0 ± 0	0 ± 0

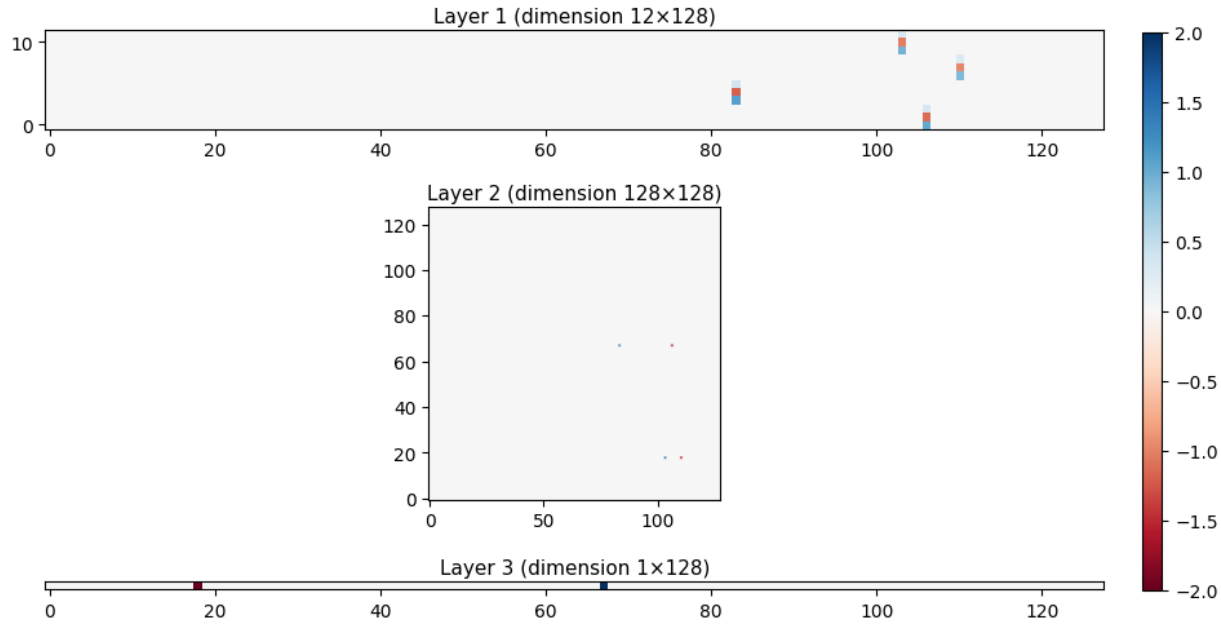
(b) Without scaling the initial weights

Figure 3. Experiment 3: Average final loss and standard deviation for each teacher-student activation pair under different regularization methods. Green = low loss (< 0.001), Orange = moderate loss (0.001–0.25), Red = high loss (> 0.25).

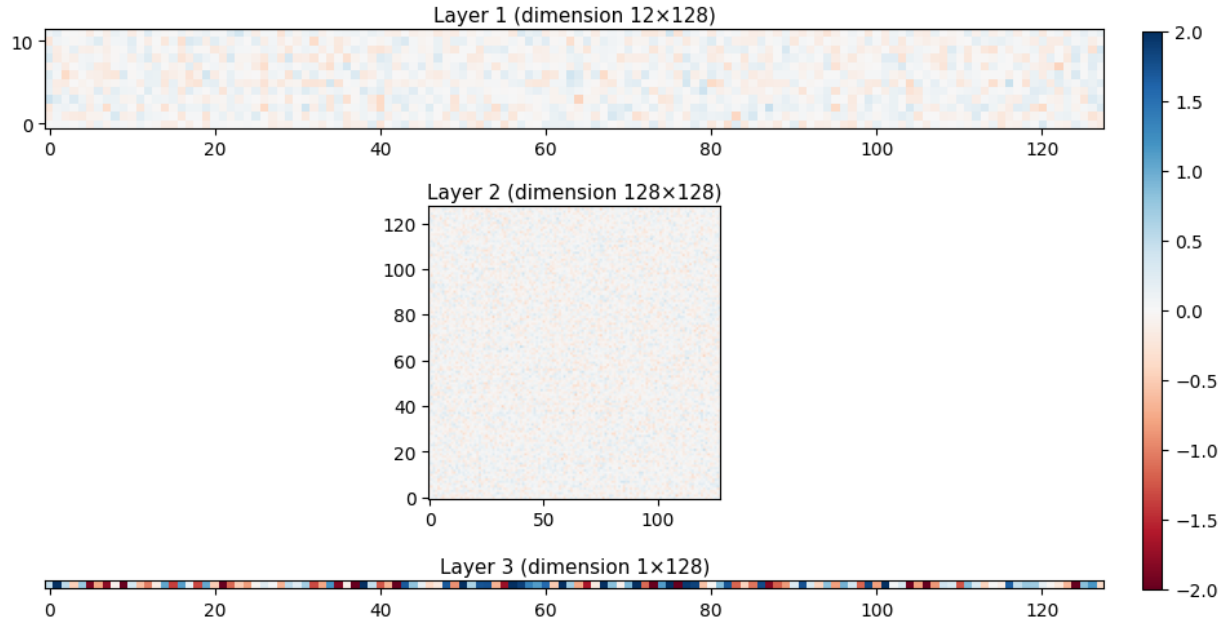
This possibly indicates convergence to a suboptimal local minimum rather than a truly sparse solution.

4.4. Overparameterization and Architecture

Both one-dimensional CNNs and fully-connected networks, even when overparameterized by up to 1,000 times the teacher size, retrieve the sparse teacher representation under L1-regularization. This shows that excess capacity does not prevent the emergence of sparse representations, provided activations align and L1 penalties are applied.



(a) **L1-regularization with initialization weights scaled down by 0.2** achieves perfect sparsity with exactly 12 weights in the first layer (sparsity 99.22%), 4 in the second (sparsity 99.98%), and 2 in the last layer (sparsity 98.44%). Final loss: 0.0002



(b) Sparsity achieved with **no regularization and no scaling**: 0% in Layer 1, 0.05% in Layer 2, and 0% in Layer 3. Final loss: 0.0000

Figure 4. Experiment 4: Comparison of weight maps of the FCN_128_128 model with ReLU activations for both the teacher and student model under different regularization methods.

4.5. Activation-specific Patterns and Non-uniqueness

We also observed that the recovered weights do not need to occupy the same channels between runs, even when performance is identical. This highlights the non-uniqueness of sparse solutions and the influence of random initialization.

In the special case where both teacher and student use Tanh, the student recovers the teacher's weights up to a $< 1\%$ error (often with sign-flipped channels).

Detailed quantitative results and visual examples can be found in the appendix for each experiment individually.

5. Discussion

Our experiments demonstrate that exact sparsity in student networks can emerge when three conditions are met: (1) teacher and student share the same activation function, (2) L1-regularization is applied, and (3) initialization avoids pathological saturations for Sigmoid. These results highlight the importance of initialization and that when models try to learn an unknown underlying function, finding the right activation functions is crucial. While hyperparameter fine-tuning can help, it might not prevent the model from converging to a suboptimal local minimum.

Additionally, we attribute the behavior that only Tanh-activated students recover the teacher's exact weights to Tanh's zero-centered symmetry. This possibly enables precise inversion of the teacher mapping. This would imply that other zero-centered functions would also show this behavior.

5.1. Limitations

While the experiments provide valuable insights, they were conducted in a controlled and relatively simple synthetic environment. Therefore, the generalization of the findings to more complex, real-world datasets, such as MNIST or ImageNet, remains an open question.

Moreover, in several configurations, we observed inconsistent convergence. Despite the student model being sufficiently large to approximate the target function, some runs resulted in a high final loss, suggesting convergence failures. In most cases, at least one seed did succeed, as indicated by large standard deviations. This implies that convergence is possible but not guaranteed. These findings raise the question of whether hyperparameter tuning, particularly the learning rate, could improve stability.

Furthermore, future research could investigate the influence of activation functions applied at different layers, or examine the role of L1-regularization at various training phases. Additionally, evaluating a broader range of activation functions and network architectures would help test the robustness and practical applicability of these findings.

6. Conclusion

In a controlled teacher-student framework with synthetic data, we demonstrated that exact replication of a teacher's function is possible under the condition that the teacher and students share the same activation function and are trained with L1-regularization. In comparison, L2-regularization effectively reduces the magnitude of the weights but does not induce sparsity. Furthermore, L2-regularized and non-regularized students tend to achieve a significantly lower loss when the initial weights are scaled down by a factor of 0.2. For L1-regularization achieving a low loss is not guaranteed but the loss seems to correlate negatively with the achieved sparsity. The results suggest that sparsity does not emerge early during training, but instead becomes prominent only as the model approaches the optimal solution.

Additionally, we have found that Sigmoid activations are less stable under scaled initialization, often resulting in very sparse early layers but overly dense final layers.

Overall, the findings are consistent across different architectures and activation functions, reinforcing the robustness of the observed patterns. While initialization plays a critical role in training dynamics, our results offer practical guidance for designing student networks that can not only approximate complex teacher models but also recover sparse representations.

References

[1] Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[2] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[3] Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[4] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[5] Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000. ISSN 00401706. URL http://www.jstor.org/stable/1271436.

[6] Lecun, Y., Bottou, L., Orr, G., and Müller, K.-R. Efficient backprop. 08 2000.

[7] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

[8] Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

[9] Malach, E., Yehudai, G., Shalev-Schwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pp. 6682–6691. PMLR, 2020.

[10] Mhaskar, H., Liao, Q., and Poggio, T. Learning functions: When is deep better than shallow. *Neural Computation*, 28(11):2825–2846, 2016.

[11] Ng, A. Y. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 78. ACM, 2004.

[12] Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[13] Scardapane, S., Comminiello, D., Hussain, A., and Uncini, A. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

[14] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[15] Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

A. Experiment 1

When the student and teacher models use the same activation function, all regularization techniques (L1-, L2-, and no regularization) consistently result in minimal final loss. This shows that the student can match the teacher's performance, regardless of the regularization applied. Tables 2, 3, and 4 compare the learned weights when using no regularization, L1-regularization, and L2-regularization. In each case, the student and teacher share the same activation function.

As shown, all three regularization strategies result in a low final loss and similar learned weights for each activation function. Interestingly, when both the teacher and student models use Tanh activation functions, the student learns the exact weights as the teacher with a negligible error. When ReLU or Sigmoid is deployed, the student learns different weights, yet still achieves an equally low final loss. This highlights that different internal representations can yield similar predictive performance.

	Teacher	**ReLU Student**	**Sigmoid Student**	**Tanh Student**
Layer 1	$[2.59, -2.83, 0.87]$	$[1.29, -1.41, 0.43]$	$[2.55, -2.78, 0.86]$	$[-2.59, 2.83, -0.87]$
Layer 2	$[-1.38, 1.29]$	$[-1.61, 1.50]$	$[-1.38, 1.29]$	$[1.38, -1.29]$
Layer 3	$[0.86, -0.84]$	$[1.48, -1.45]$	$[0.86, -0.84]$	$[0.86, -0.84]$
Final Loss	N/A	0	0	0

Table 2. Experiment 1: Comparison of layer weights between the teacher and student models using no regularization.

	Teacher	**ReLU Student**	**Sigmoid Student**	**Tanh Student**
Layer 1	$[2.59, -2.83, 0.87]$	$[1.33, -1.45, 0.45]$	$[2.24, -2.45, 0.75]$	$[-2.59, 2.83, -0.87]$
Layer 2	$[-1.38, 1.29]$	$[-1.56, 1.45]$	$[1.17, -1.09]$	$[-1.38, 1.29]$
Layer 3	$[0.86, -0.84]$	$[1.49, -1.45]$	$[-1.01, 1.02]$	$[-0.86, 0.84]$
Final Loss	N/A	0.0001	0.0001	0.0001

Table 3. Experiment 1: Comparison of layer weights between the teacher and student models using L1-regularization with $\lambda = 10^{-5}$.

	Teacher	**ReLU Student**	**Sigmoid Student**	**Tanh Student**
Layer 1	$[2.59, -2.83, 0.87]$	$[1.37, -1.49, 0.46]$	$[1.59, -1.73, 0.53]$	$[2.55, -2.79, 0.86]$
Layer 2	$[-1.38, 1.29]$	$[-1.51, 1.41]$	$[1.15, -1.07]$	$[-1.36, 1.27]$
Layer 3	$[0.86, -0.84]$	$[1.49, -1.45]$	$[-1.10, 1.12]$	$[0.86, -0.84]$
Final Loss	N/A	0.0001	0.0001	0.0002

Table 4. Experiment 1: Comparison of layer weights between the teacher and student models using L2-regularization with $\lambda = 10^{-5}$.

However, in this setting, it is not possible to achieve a low loss for teacher-student pairs that do not share the same activation function. This is likely due to mismatches in the functional mappings of different activation functions, which the small student model lacks the capacity to approximate effectively. An example is depicted in Table 5, where the results are analogous to other regularization methods and initialization scaling strategies.

Teacher \ Student	**ReLU**	**Sigmoid**	**Tanh**
ReLU	0	3.6582	3.8137
Sigmoid	0.2557	0	0.2522
Tanh	0.2860	0.1940	0

Table 5. Experiment 1: Final loss of different teacher-student activation combinations using no regularization.

B. Experiment 2

In this setting, if the activation functions between the student and the teacher differ, the student doesn't manage a low loss. These results suggest that the student model lacks the capacity to accurately recover the teacher's representation when the activation functions differ. Figure 5 reports the final losses under different regularization techniques and initialization methods.

Teacher Act. → Student Act.	**L1 Reg (avg ± std)**	**L2 Reg (avg ± std)**	**No Reg (avg ± std)**
ReLU → ReLU	0.0001 ± 0	1.4118 ± 1.9451	0 ± 0
ReLU → Tanh	3.6586 ± 0.1281	3.6505 ± 0.1291	3.6659 ± 0.1192
ReLU → Sigmoid	3.3434 ± 0.1272	3.3467 ± 0.1287	3.3495 ± 0.1194
Tanh → ReLU	0.2234 ± 0.0136	0.2475 ± 0.0681	0.2130 ± 0.0084
Tanh → Tanh	0.0001 ± 0	0.0002 ± 0	0 ± 0
Tanh → Sigmoid	0.1871 ± 0.0037	0.1885 ± 0.0035	0.1862 ± 0.0043
Sigmoid → ReLU	0.0120 ± 0.0002	0.0120 ± 0.0003	0.0120 ± 0.0002
Sigmoid → Tanh	0.2514 ± 0.0005	0.2516 ± 0.0005	0.2514 ± 0.0005
Sigmoid → Sigmoid	0.0001 ± 0	0.0001 ± 0	0 ± 0

(a) Scaling the initial weights down by a factor of 0.2

Teacher Act. → Student Act.	**L1 Reg (avg ± std)**	**L2 Reg (avg ± std)**	**No Reg (avg ± std)**
ReLU → ReLU	0.0001 ± 0	0.0001 ± 0	0 ± 0
ReLU → Tanh	3.6173 ± 0.1139	3.6168 ± 0.1148	3.6629 ± 0.0827
ReLU → Sigmoid	3.3046 ± 0.1101	3.3155 ± 0.1132	3.3447 ± 0.0848
Tanh → ReLU	0.0849 ± 0.1157	0.2077 ± 0.0070	0.2155 ± 0.0099
Tanh → Tanh	0 ± 0.0001	0.0002 ± 0	0 ± 0
Tanh → Sigmoid	0.1476 ± 0.1537	0.1865 ± 0.0058	0.2226 ± 0.0843
Sigmoid → ReLU	0.0118 ± 0.0002	0.0116 ± 0.0003	0.0118 ± 0.0004
Sigmoid → Tanh	0.2518 ± 0.0004	0.2516 ± 0.0005	0.2514 ± 0.0005
Sigmoid → Sigmoid	0.0016 ± 0.0014	0.0001 ± 0	0 ± 0

(b) Without scaling the initial weights

Figure 5. Experiment 2: Average final loss and standard deviation for each teacher-student activation pair under different regularization methods (L1, L2, and no Regularization).
Green = low loss (< 0.001), Orange = moderate loss (0.001–0.25), Red = high loss (> 0.25).

But again, when the student and teacher models share the same activation functions, all regularization strategies lead to a near-zero final loss. However, only L1-regularization induces sparsity in the learned weights. In particular, L1 enables the student model to achieve perfect sparsity. Furthermore, while L2-regularization effectively drives the weights toward smaller magnitudes, the weights often remain non-zero across the network.

Tables 6 and 7 show the results for Tanh and ReLU activations under various regularization settings. Again, only when using the Tanh activation function does the student model recover weights that closely match those of the teacher. In contrast, student models trained with ReLU or Sigmoid activations converge to different weight configurations, despite achieving similarly low final losses. Across multiple random seeds, models with matching activation functions tend to learn similar weights, even when using differently scaled initialization weights. Additionally, we observe that the specific channels and layers in which non-zero weights are retained vary across activation functions and random seeds. This indicates that while functionally equivalent, different sparse configurations are possible. Table 8 highlights the tendency of Sigmoid with L1-regularization: weights in early layers are near zero, while the final layer remains dense. This pattern appears in three out of five seeds when initialization is scaled down, and only once without scaling, indicating that when using the Sigmoid activation function, scaling the initial weights down is not beneficial.

These results illustrate how different activation functions lead to different sparsity patterns, even when final losses are similar.

	Teacher	L1-reg. (scaled 0.2)	L1-reg. (no scaling)	L2-reg. (no scaling)	No reg. (no scaling)
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 2.58 & -2.82 & 0.87 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -2.58 & 2.82 & -0.87 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 2.35 & -2.57 & 0.79 \end{bmatrix}$ $\begin{bmatrix} -0.07 & 0.08 & -0.02 \end{bmatrix}$ $\begin{bmatrix} 0.07 & -0.08 & 0.02 \end{bmatrix}$ $\begin{bmatrix} -0.29 & 0.33 & -0.09 \end{bmatrix}$	$\begin{bmatrix} 0.0022 & -0.0024 & 0.0008 \end{bmatrix}$ $\begin{bmatrix} -1.27 & 1.39 & -0.43 \end{bmatrix}$ $\begin{bmatrix} 2.40 & -2.63 & 0.81 \end{bmatrix}$ $\begin{bmatrix} 0.84 & -0.92 & 0.28 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1.38 & 1.29 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -1.38 & 1.29 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -1.59 & 1.16 \\ 0.03 & -0.07 \\ -0.03 & 0.07 \\ 0.0060 & 0.16 \end{bmatrix}$ $\begin{bmatrix} 0.16 & -0.21 \\ 0.04 & -0.04 \\ -0.04 & 0.04 \\ -0.02 & 0.01 \end{bmatrix}$ $\begin{bmatrix} -1.05 & 1.44 \\ 0.07 & -0.04 \\ -0.07 & 0.04 \\ -0.13 & -0.07 \end{bmatrix}$ $\begin{bmatrix} -0.2 & 0.03 \\ -0.02 & 0.01 \\ 0.02 & -0.01 \\ -0.02 & 0.1 \end{bmatrix}$	$\begin{bmatrix} -0.46 & -0.44 \\ -0.74 & 0.41 \\ 0.51 & 0.46 \\ 0.06 & -0.69 \end{bmatrix}$ $\begin{bmatrix} 0.74 & 0.07 \\ 0.31 & 0.50 \\ 1.33 & -1.21 \\ 0.07 & -0.23 \end{bmatrix}$ $\begin{bmatrix} -0.85 & -0.49 \\ 0.37 & -1.20 \\ -0.58 & -0.18 \\ -0.69 & 0.72 \end{bmatrix}$ $\begin{bmatrix} 0.73 & 0.91 \\ -0.30 & -0.22 \\ 1.46 & -1.32 \\ -0.0067 & -0.05 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} -0.86 & 0.84 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.53 & -0.52 \\ 0.18 & -0.18 \\ 0.39 & -0.39 \\ -0.02 & 0.05 \end{bmatrix}$	$\begin{bmatrix} -0.01 & 0.01 \\ -0.35 & 0.34 \\ -0.06 & 0.06 \\ -0.56 & 0.55 \end{bmatrix}$
Final Loss	N/A	0.0001	0.0001	0.0002	0.0000

Table 6. Experiment 2: Comparison of layer weights for the teacher model and student models with **Tanh** activations. All weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment was run with seed 1.

	Teacher	L1-reg. (scaled 0.2)	L1-reg. (no scaling)	L2-reg. (no scaling)	No reg. (no scaling)
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \end{bmatrix}$	$\begin{bmatrix} 1.27 & -1.38 & 0.43 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.26 & -1.38 & 0.42 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.0017 & 0.0018 & -0.0006 \\ -0.0021 & 0.0021 & -0.0007 \\ -0.0015 & 0.0016 & -0.0004 \end{bmatrix}$ $\begin{bmatrix} 1.37 & -1.49 & 0.46 \end{bmatrix}$	$\begin{bmatrix} 1.67 & -1.83 & 0.56 \\ 0.56 & -0.61 & 0.19 \\ -0.43 & 0.21 & 0.01 \\ 0.07 & -0.07 & 0.02 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -1.57 & 1.47 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -1.58 & 1.47 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.0006 & -0.0009 \\ -0.0009 & -0.0005 \\ -0.0014 & -0.0005 \\ -0.0005 & -0.0002 \end{bmatrix}$ $\begin{bmatrix} -0.0030 & -0.0035 \\ -0.0013 & -0.0005 \\ 0 & -0.0032 \\ -0.0002 & -0.0036 \end{bmatrix}$ $\begin{bmatrix} -0.0021 & -0.0050 \\ -0.0029 & -0.0053 \\ -0.0002 & -0.0061 \\ 0 & -0.0054 \end{bmatrix}$ $\begin{bmatrix} 0.0015 & -0.0013 \\ 0.0018 & -0.0023 \\ 0.0013 & -0.0017 \\ -1.52 & 1.42 \end{bmatrix}$	$\begin{bmatrix} -0.69 & -0.23 \\ 1.32 & -0.05 \\ -0.05 & -0.10 \\ 0.52 & 0.13 \end{bmatrix}$ $\begin{bmatrix} -0.60 & -0.12 \\ -0.29 & -0.92 \\ -0.16 & -0.16 \\ 1.02 & 0.18 \end{bmatrix}$ $\begin{bmatrix} -1.32 & 1.18 \\ -0.53 & 0.67 \\ 0 & 0 \\ -0.17 & 0.09 \end{bmatrix}$ $\begin{bmatrix} 0.02 & -1.68 \\ -0.07 & -0.15 \\ -0.11 & 0.64 \\ 0.27 & 0.23 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 1.54 & -1.51 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 1.54 & -1.51 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.01 & -0.01 \\ -0.00 & -0.00 \\ -0.01 & -0.00 \end{bmatrix}$ $\begin{bmatrix} 1.48 & -1.45 \end{bmatrix}$	$\begin{bmatrix} -0.01 & -0.51 \\ -0.51 & -0.96 \end{bmatrix}$ $\begin{bmatrix} 1.22 & -1.19 \end{bmatrix}$ $\begin{bmatrix} -1.60 & 0.00 \end{bmatrix}$
Final Loss	N/A	0.0001	0.0001	0.0001	0.0000

Table 7. Experiment 2: Comparison of layer weights for the teacher model and student models with **ReLU** activations. All weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment was run with seed 3.

	Teacher	L1-reg. (scaled 0.2) — Seed 4	L1-reg. (scaled 0.2) — Seed 5	L1-reg. (no scaling) — Seed 4	L1-reg. (no scaling) — Seed 5
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} -2.23 & 2.44 & -0.74 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} -2.23 & 2.43 & -0.74 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 2.22 & -2.43 & 0.73 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1.13 & -1.06 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1.11 & -1.04 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 1.17 & -1.09 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} 1.05 & -1.03 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.0007 & 0.0007 \end{bmatrix}$ $\begin{bmatrix} 0.0007 & 0.0007 \end{bmatrix}$ $\begin{bmatrix} 0.0007 & 0.0007 \end{bmatrix}$ $\begin{bmatrix} 0.0077 & 0.0007 \end{bmatrix}$	$\begin{bmatrix} 1.07 & -1.05 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.0001 & 0.0001 \end{bmatrix}$ $\begin{bmatrix} 0.02 & 0.0001 \end{bmatrix}$ $\begin{bmatrix} -1.02 & 1.00 \end{bmatrix}$ $\begin{bmatrix} 0.0001 & 0.01 \end{bmatrix}$
Final Loss	N/A	0.0001	0.0026	0.0001	0.0001

Table 8. Experiment 2: Comparison of layer weights for the teacher model and student models with **Sigmoid** activations. All weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment includes seeds 4 and 5.

C. Experiment 3

When the student and teacher networks have the same activation function, the student can achieve perfect sparsity. This is consistent with observations from the previous experiments. Figures 6, 7, and 8 show the weight maps for different regularization techniques. For all regularization techniques, the final loss is very low. While the weight map for L2-regularization and L1-regularization looks similarly sparse, the computed sparsity value says otherwise. L1 regularization clearly induces zero-weights, while when not using any regularization, the network is very dense.

In general, both L2-regularization and no regularization tend to benefit from scaling down the initial weights, leading to a lower final loss. This trend does not hold when applying L1-regularization, where scaling has less consistent effects. Interestingly, when the student is trained with L2-regularization, ReLU activations, and no scaling of the initial weights, the network becomes significantly sparser. However, the final loss is actually higher in these cases. This indicates that the model has converged to a suboptimal local minimum.

An interesting case occurs when the teacher uses ReLU and the student uses Tanh activation functions: The average final loss is notably high with a large standard deviation. This can be attributed to one run that did not converge, resulting in a final loss of 74.45. This variability underscores that in some cases, convergence is much more difficult and one setting doesn't guarantee converges across all seeds.

When the teacher and student use different activation functions, the student typically does not reach a low final loss, even though it is possible when trained with no regularization technique. Nevertheless, L1-regularization consistently produces significantly higher sparsity. Table 1 presents a quantitative comparison of the resulting sparsity levels across these techniques.

As seen in earlier experiments, student models with Sigmoid activations tend to have very small weights (absolute value $< 10^{-4}$) in the first two layers and significantly larger weights in the final layer, often leading to 0.00% sparsity in the last layer. This behavior is specific to scaled initialization (factor 0.2) and was observed under both L1- and L2-regularization.

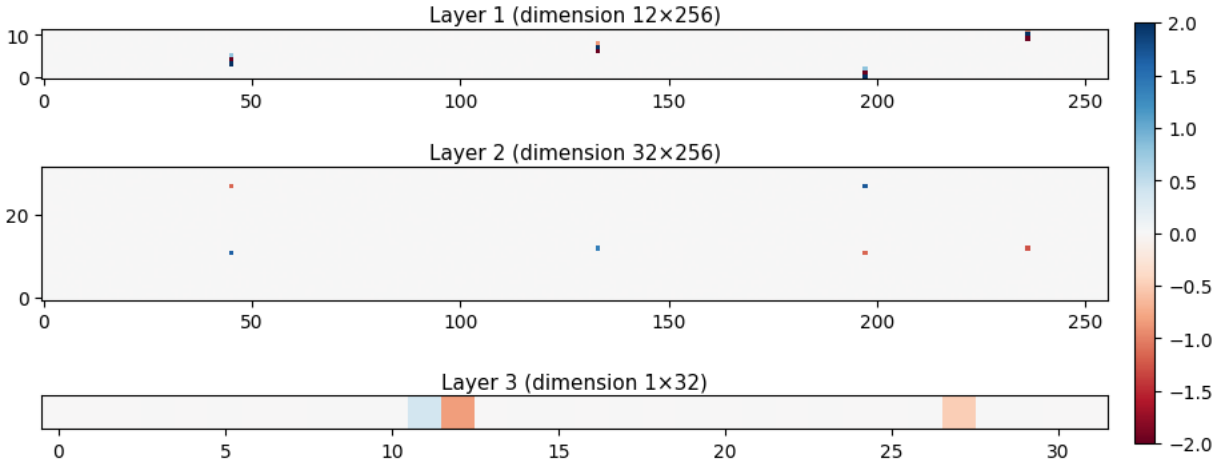


Figure 6. Experiment 3: **L1-regularization with initialization weights scaled down by 0.2** achieves sparsity for the **Tanh**-activation function: 98.60% in Layer 1, 99.97% in Layer 2, 90.62% in Layer 3. Final loss: 0.0003

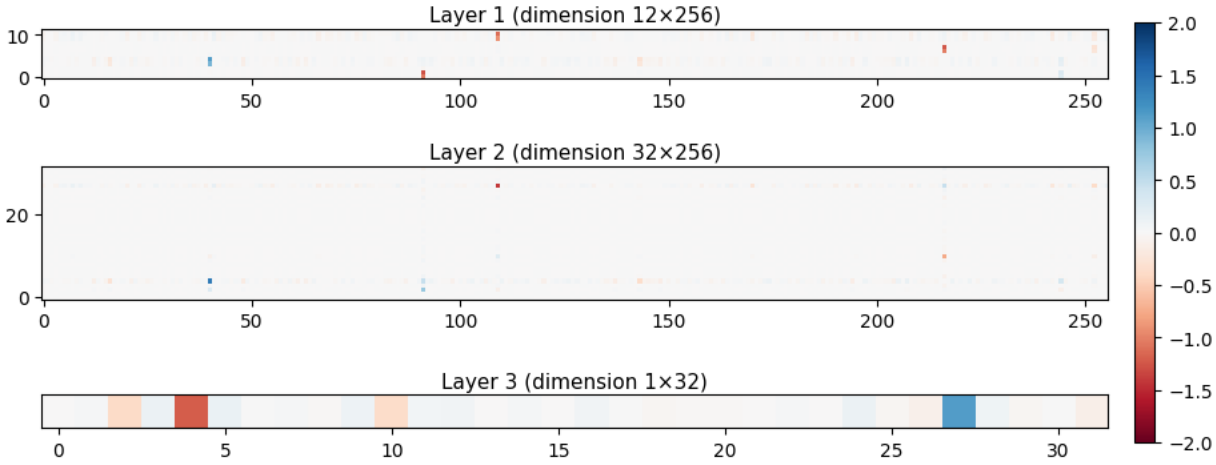


Figure 7. Experiment 3: **L2-regularization with no scaling** achieves sparsity: 15.14% in Layer 1, 39.51% in Layer 2, 0.00% in Layer 3. Final loss: 0.0002

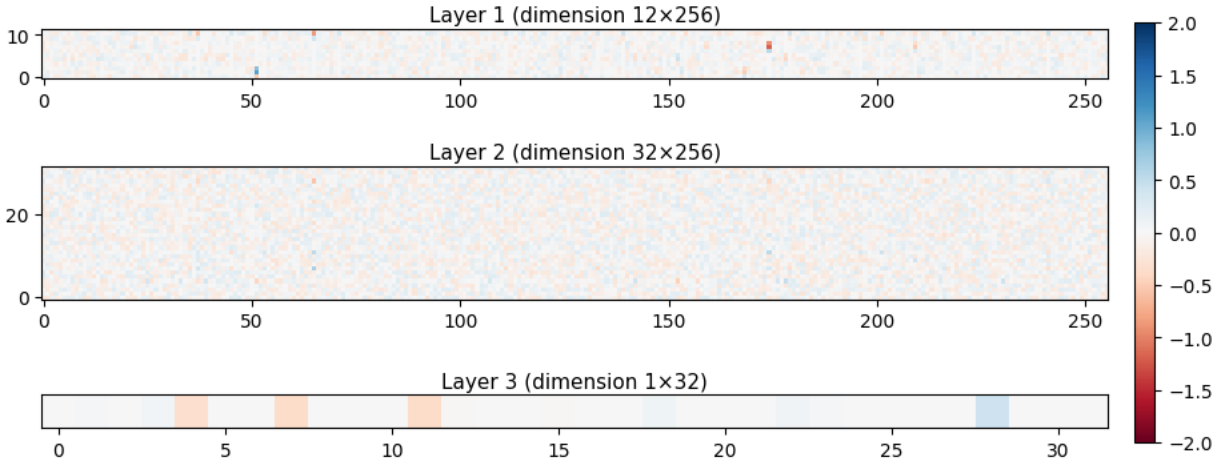


Figure 8. Experiment 3: **No regularization with no scaling** achieves sparsity: 0.13% in Layer 1, 0.07% in Layer 2, 9.38% in Layer 3. Final loss: 0.0000

D. Experiment 4

The results are analogous to Experiment 3. Again, perfect sparsity can be achieved when the activation functions of the teacher and student match and the student is trained with L1-regularization.

Figure 9 and Table 9 present the quantitative results for Experiment 4. They show the average final loss and achieved sparsity across five random seeds. The results again show that scaling down the initial weights leads to a lower final loss when L2- or no regularization is applied. However, L2-regularization with unscaled initialization results in higher sparsity, despite the higher final loss. For the case of no regularization, when the student uses the Tanh activation function, not scaling the initial weights leads to a similar final loss but higher sparsity.

For L1-regularization, the impact of scaling is less straightforward. While it remains unclear whether scaling improves the performance, there appears to be a negative correlation between the final loss and sparsity: Models with lower final loss tend to exhibit higher sparsity.

Teacher Act. → Student Act.	**L1 Reg (avg ± std)**	**L2 Reg (avg ± std)**	**No Reg (avg ± std)**
ReLU → ReLU	0.0002 ± 0	0.0003 ± 0	0 ± 0
ReLU → Tanh	2.8402 ± 1.4202	3.5337 ± 0.0350	0 ± 0
ReLU → Sigmoid	3.3020 ± 0.0243	3.2989 ± 0.0354	0 ± 0
Tanh → ReLU	0.1547 ± 0.0773	0.1913 ± 0.0014	0.0382 ± 0.0753
Tanh → Tanh	0.0002 ± 0.0001	0.0007 ± 0	0 ± 0
Tanh → Sigmoid	0.2940 ± 0.1466	0.1886 ± 0.0015	0.0010 ± 0.0003
Sigmoid → ReLU	0.0109 ± 0.0001	0.0096 ± 0	0.0077 ± 0.0009
Sigmoid → Tanh	0.2527 ± 0.0001	0.2525 ± 0	0.2477 ± 0.0266
Sigmoid → Sigmoid	0.0026 ± 0	0.0003 ± 0	0 ± 0

(a) Scaling the initial weights down by a factor of 0.2

Teacher Act. → Student Act.	**L1 Reg (avg ± std)**	**L2 Reg (avg ± std)**	**No Reg (avg ± std)**
ReLU → ReLU	0.0001 ± 0.0001	4.9428 ± 0.0502	0.5781 ± 0.4724
ReLU → Tanh	0.7121 ± 1.4236	3.5508 ± 0.0447	0.1962 ± 0.3896
ReLU → Sigmoid	1.8385 ± 1.6836	3.3156 ± 0.0447	0.1627 ± 0.1826
Tanh → ReLU	0.2672 ± 0.0859	0.4038 ± 0.0738	0.5403 ± 0.1188
Tanh → Tanh	0.0001 ± 0	0.0003 ± 0.0001	0 ± 0
Tanh → Sigmoid	0.1568 ± 0.0661	0.1986 ± 0.0201	0.1934 ± 0.0997
Sigmoid → ReLU	0.0106 ± 0.0046	0.2550 ± 0	0.2262 ± 0.1380
Sigmoid → Tanh	0.2754 ± 0.1088	0.2521 ± 0.0001	0.3324 ± 0.1503
Sigmoid → Sigmoid	0.0001 ± 0.0002	0.0003 ± 0	0 ± 0

(b) Without scaling the initial weights

Figure 9. Experiment 4: Average final loss and standard deviation for each teacher-student activation pair under different regularization methods (L1, L2, and no Regularization).
Green = low loss (< 0.001), Orange = moderate loss (0.001–0.25), Red = high loss (> 0.25).

Teacher Act. → Student Act.	L1 Reg. (scaled 0.2)	L1 Reg. (no scaling)	L2 Reg. (scaled 0.2)	L2 Reg. (no scaling)	No Reg. (scaled 0.2)	No Reg. (no scaling)
ReLU → ReLU	99.03	**99.33**	20.20	78.34	0.30	0.06
ReLU → Tanh	89.96	**96.10**	0.98	2.03	0.31	20.18
ReLU → Sigmoid	**81.50**	65.07	0.13	0.19	0.21	0.04
Tanh → ReLU	**83.56**	79.04	0.22	77.37	0.27	0.03
Tanh → Tanh	98.71	**98.86**	12.49	17.08	0.42	1.63
Tanh → Sigmoid	66.67	**84.92**	7.28	4.05	0.151	0.05
Sigmoid → ReLU	78.44	**81.73**	0.45	75.43	0.04	0.03
Sigmoid → Tanh	**98.95**	98.58	37.85	6.67	0.69	0.81
Sigmoid → Sigmoid	66.67	**97.79**	2.38	2.01	0.21	0.03

Table 9. Experiment 4: Average sparsity (%) for each teacher-student activation pair under different regularization methods and initialization scaling. Best values per row are bold.

E. Experiment 5

The figures and tables below report average final losses with standard deviations (Figures 11) as well as the resulting sparsity levels across layers (Table 10). Similar to Experiments 3 and 4, they show that scaling down the initial weights helps the student achieve a lower loss and slightly higher sparsity when no regularization is used. With L2-regularization, scaling also reduces the final loss. Interestingly, when the student uses ReLU activations, not scaling leads to high sparsity but poor loss values. This suggests that the model gets stuck in a suboptimal local minima with very sparse/small weights.

In addition, Figure 10 depicts the weight map of a student trained with Sigmoid activation, learning from a teacher using Tanh activations. It illustrates that the first two layers consistently develop very small weights, effectively becoming sparse, while the final layer remains densely populated. This pattern was consistent across all five random seeds tested, despite achieving a relatively low final loss.

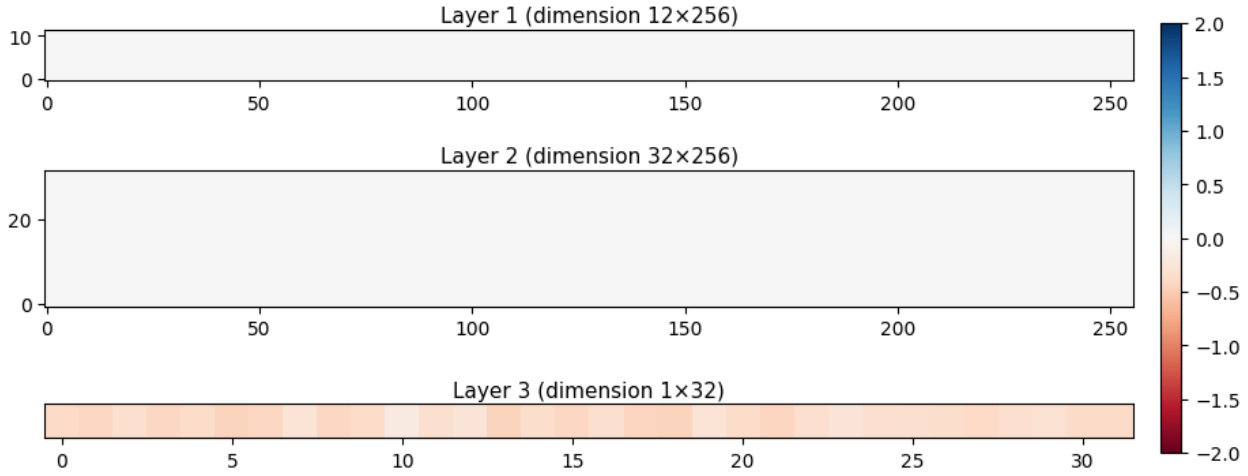


Figure 10. Experiment 5: **L1-regularization with initialization weights scaled down by 0.2** achieves sparsity: 100.00% in Layer 1, 100.00% in Layer 2, 0.00% in Layer 3. Final loss: 0.0005

Teacher Act. → Student Act.	L1 Reg (avg ± std)	L2 Reg (avg ± std)	No Reg (avg ± std)
ReLU → ReLU	0.0005 ± 0	0.0002 ± 0.0001	0 ± 0
ReLU → Tanh	1.9358 ± 0.9682	1.9273 ± 0.9637	0 ± 0
ReLU → Sigmoid	2.1368 ± 0.0253	1.7050 ± 0.8526	0 ± 0
Tanh → ReLU	0.0003 ± 0.0001	0.1411 ± 0.1150	0.0004 ± 0.0004
Tanh → Tanh	0 ± 0	0.0004 ± 0.0003	0 ± 0
Tanh → Sigmoid	0.0467 ± 0.0912	0.1389 ± 0.1126	0.0005 ± 0.0004
Sigmoid → ReLU	0.0072 ± 0	0.0075 ± 0.0018	0.0079 ± 0.0012
Sigmoid → Tanh	0.1672 ± 0	0.1844 ± 0.0346	0.2291 ± 0.0334
Sigmoid → Sigmoid	0.0016 ± 0	0.0001 ± 0	0 ± 0

(a) Scaling the initial weights down by a factor of 0.2

Teacher Act. → Student Act.	L1 Reg (avg ± std)	L2 Reg (avg ± std)	No Reg (avg ± std)
ReLU → ReLU	0.0005 ± 0.0001	0.7568 ± 1.3860	0.3199 ± 0.3520
ReLU → Tanh	1.5962 ± 0.9854	0.7234 ± 0.9057	0.1060 ± 0.1077
ReLU → Sigmoid	2.1200 ± 0.0086	0.8535 ± 1.0416	0.1086 ± 0.1831
Tanh → ReLU	0.2716 ± 0.0535	0.4726 ± 0.1403	0.4573 ± 0.1212
Tanh → Tanh	0.0006 ± 0.0001	0.0004 ± 0.0001	0 ± 0
Tanh → Sigmoid	0.2315 ± 0.0444	0.2343 ± 0.0684	0.2535 ± 0.0477
Sigmoid → ReLU	0.0072 ± 0	0.1038 ± 0.0794	0.1584 ± 0.1201
Sigmoid → Tanh	0.1671 ± 0	0.1668 ± 0	0.2655 ± 0.1510
Sigmoid → Sigmoid	0.0001 ± 0	0.0001 ± 0	0 ± 0

(b) Without scaling the initial weights

Figure 11. Experiment 5: Average final loss and standard deviation for each teacher-student activation pair under different regularization methods (L1, L2, and no Regularization).
Green = low loss (< 0.001), Orange = moderate loss (0.001–0.25), Red = high loss (> 0.25).

Teacher Act. → Student Act.	L1 Reg. (scaled 0.2)	L1 Reg. (no scaling)	L2 Reg. (scaled 0.2)	L2 Reg. (no scaling)	No Reg. (scaled 0.2)	No Reg. (no scaling)
ReLU → ReLU	**88.75**	88.72	10.68	85.94	0.58	0.24
ReLU → Tanh	**84.68**	65.18	22.24	6.23	0.35	0.21
ReLU → Sigmoid	55.33	**56.98**	3.05	0.51	0.22	0.05
Tanh → ReLU	**98.83**	54.20	19.43	50.81	0.31	0.04
Tanh → Tanh	**98.82**	89.79	5.46	8.70	0.33	0.15
Tanh → Sigmoid	65.08	**77.05**	1.40	0.74	0.16	0.05
Sigmoid → ReLU	**86.41**	86.10	0.14	47.72	0.03	0.03
Sigmoid → Tanh	**98.58**	98.40	3.74	11.07	0.36	0.51
Sigmoid → Sigmoid	66.67	**93.15**	14.67	2.04	0.21	0.06

Table 10. Experiment 5: Average sparsity (%) for each teacher-student activation pair under different regularization methods and initialization scaling. Best values per row are bold.

F. Experiment 6

The results mirror those of previous experiments: When the teacher and student models use the same activation function, the dense student model can achieve perfect sparsity.

Tables 11, 12, and 13 compare different initialization weight scaling strategies across activation functions. These tables aim to highlight the general tendencies and edge cases associated with each setting, offering a more holistic perspective. It is important to note that the model's performance can vary depending on the seed. The cases where the teacher and student activation differ are not further analyzed, as Experiment 2 has already demonstrated the difficulty of achieving low loss in these settings.

Table 13 shows that Tanh consistently produces sparse student models, regardless of the initialization scaling strategy. They never achieve perfect sparsity. However, given the results from the previous experiments, we believe that with some parameter tuning, the models trained with Tanh and Sigmoid activation functions can achieve perfectly sparse results.

	Teacher	**L1-reg. (scaled 0.2)**	**L1-reg. (no scaling)**	**L2-reg. (scaled 0.2)**	**No reg. (scaled 0.2)**
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \\ -1.22 & 0.45 & 0.88 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.22 & -0.45 & -0.88 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 2.58 & -2.82 & 0.87 \end{bmatrix}$	$\begin{bmatrix} -1.15 & 1.23 & -0.35 \end{bmatrix}$ $\begin{bmatrix} -0.08 & 0.04 & 0.06 \end{bmatrix}$ $\begin{bmatrix} 0.60 & -0.15 & -0.56 \\ 0.43 & -0.10 & -0.40 \end{bmatrix}$	$\begin{bmatrix} -0.38 & 0.34 & -0.02 \end{bmatrix}$ $\begin{bmatrix} 1.24 & -0.48 & -0.87 \end{bmatrix}$ $\begin{bmatrix} -2.04 & 2.22 & -0.68 \\ 0.17 & 0.13 & -0.39 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \\ 0.35 & -0.73 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0.73 \\ 0 & 0 \\ 0 & 0 \\ 0.35 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0.0001 & -1.29 \\ 0 & 0 \\ 0 & 0 \\ -1.38 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.77 & -0.12 \\ 0.02 & -0.05 \\ -0.04 & 0.56 \\ 0.0063 & 0.40 \end{bmatrix}$ $\begin{bmatrix} 0.37 & -0.04 \\ 0.0078 & -0.03 \\ 0.02 & 0.05 \\ -0.02 & 0.05 \end{bmatrix}$ $\begin{bmatrix} 0.38 & -0.06 \\ -0.001 & -0.03 \\ 0.02 & 0.10 \\ -0.02 & 0.08 \end{bmatrix}$ $\begin{bmatrix} 0.64 & 0.12 \\ -0.02 & 0.06 \\ 0.04 & -0.51 \\ 0.0013 & -0.36 \end{bmatrix}$	$\begin{bmatrix} 0.04 & -0.07 \\ 0.004 & 0.53 \\ -0.29 & -0.04 \\ 0.09 & 0.18 \end{bmatrix}$ $\begin{bmatrix} -0.42 & 0.01 \\ 0.02 & -0.60 \\ 1.04 & -0.0014 \\ -0.13 & -0.02 \end{bmatrix}$ $\begin{bmatrix} -0.00 & -0.21 \\ -0.11 & 0.41 \\ -0.06 & 0.06 \\ 0.15 & 0.11 \end{bmatrix}$ $\begin{bmatrix} -0.18 & -0.03 \\ -0.03 & -0.08 \\ 0.26 & -0.03 \\ 0.01 & -0.0039 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} 0.0022 & 0.0022 \\ 0.0022 & 0.0022 \\ 0.0022 & 0.0152 \\ 0.0023 & 0.0588 \end{bmatrix}$	$\begin{bmatrix} 0.0008 & -0.84 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0.86 & 0.0018 \end{bmatrix}$	$\begin{bmatrix} -0.94 & -0.41 \end{bmatrix}$ $\begin{bmatrix} 0.33 & -0.19 \end{bmatrix}$ $\begin{bmatrix} 0.33 & -0.25 \end{bmatrix}$ $\begin{bmatrix} 0.81 & 0.40 \end{bmatrix}$	$\begin{bmatrix} -0.37 & -0.47 \end{bmatrix}$ $\begin{bmatrix} 1.17 & 0.19 \end{bmatrix}$ $\begin{bmatrix} -0.18 & -0.42 \end{bmatrix}$ $\begin{bmatrix} 0.29 & 0.04 \end{bmatrix}$
Final Loss	N/A	0.0012	0.0001	0.0001	0.0000

Table 11. Experiment 6: Comparison of layer weights for the teacher model and student models with **Sigmoid** activations. Weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment was run with seed 2.

Table 12 shows that ReLU activations enable the student to learn perfect sparsity. Notably, scaling makes a difference: Without scaling, ReLU fails to reach sparsity in two out of five random seeds. However, when the initialization weights are scaled down by 0.2, ReLU consistently produces sparse student models across all five seeds. This indicates potential instability in the unscaled setting.

However, scaling down the initialization weights by a factor of 0.2 again leads to challenges for the Sigmoid activation. Table 11 demonstrates that when using the Sigmoid activation with scaled initialization weights, the student model frequently develops very small weights [5] in the first two layers. These are visualized as zeros. In such cases, the final layer remains with non-zero weights. This sparsity issue does not occur when the initialization weights are not scaled down.

For L2-regularization and no regularization, the scaling of initialization weights does not make a meaningful difference and does not promote noticeable additional sparsity. As expected, L2-regularization helps reduce the magnitude of the weights.

[5] absolute value $< 10^{-4}$

Nevertheless, this does not imply that the resulting learned weights are sparser.

For each table, the last two columns include results for student models trained with only one initialization scaling per technique, as the outcomes are very similar and align with previous experiments. When comparing different regularization techniques, L2-regularization and no regularization result in similar levels of sparsity. Although L2-regularization successfully reduces the size of the weights, it does not result in a higher level of sparsity. In the case of the Tanh activation function, no regularization actually achieves a higher sparsity level than when using L2-regularization. However, this is not the case across all seeds.

	Teacher	**L1-reg. (scaled 0.2)**	**L1-reg. (no scaling)**	**L2-reg. (no scaling)**	**No reg. (no scaling)**
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \\ -1.22 & 0.45 & 0.88 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} -0.74 & 0.27 & 0.53 \\ 0.89 & -0.97 & 0.30 \end{bmatrix}$	$\begin{bmatrix} 0.24 & 0.09 & -0.05 \\ -0.18 & 0.07 & -0.47 \end{bmatrix}$ $\begin{bmatrix} -0.94 & 0.38 & 0.69 \\ 1.51 & -1.61 & 0.48 \end{bmatrix}$	$\begin{bmatrix} -0.0022 & 0.0029 & -0.0026 \end{bmatrix}$ $\begin{bmatrix} 1.11 & -1.21 & 0.37 \\ 0.04 & -0.04 & 0.01 \\ -0.90 & 0.33 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 0.55 & -0.60 & 0.18 \\ -0.15 & 0.05 & 0.11 \end{bmatrix}$ $\begin{bmatrix} 0.94 & -1.03 & 0.32 \\ -0.83 & 0.30 & 0.60 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \\ 0.35 & -0.73 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & -0.74 \\ 0.63 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.04 & -0.38 \\ 0.21 & 0.14 \\ 0.10 & 1.46 \\ -1.95 & 0 \end{bmatrix}$ $\begin{bmatrix} -0.47 & -0.55 \\ 0 & -0.58 \\ -0.25 & -0.15 \\ 0 & -0.09 \end{bmatrix}$ $\begin{bmatrix} -0.89 & -0.31 \\ -0.04 & -0.21 \\ -0.78 & -0.76 \\ 0 & -0.05 \end{bmatrix}$ $\begin{bmatrix} 0.52 & 0.40 \\ 0.55 & 0.34 \\ 0.31 & 0.41 \\ -0.13 & 0.06 \end{bmatrix}$	$\begin{bmatrix} -0.0019 & 0.0018 \\ 0.66 & -0.0001 \\ 0.03 & 0.0046 \\ 0 & -0.80 \end{bmatrix}$ $\begin{bmatrix} 0.0007 & 0.01 \\ 0.0038 & 0.0037 \\ -0.01 & 0.0028 \\ 0.0029 & -0.0072 \end{bmatrix}$ $\begin{bmatrix} -0.0018 & -0.0011 \\ -0.49 & 0 \\ -0.02 & 0.0022 \\ 0 & 0.27 \end{bmatrix}$ $\begin{bmatrix} 0.0084 & 0.0016 \\ -1.48 & 0.0002 \\ -0.05 & -0.0070 \\ 0 & 0.80 \end{bmatrix}$	$\begin{bmatrix} -0.71 & -1.10 \\ -0.60 & -0.85 \\ -0.28 & 0.64 \\ -0.77 & -0.07 \end{bmatrix}$ $\begin{bmatrix} 0.24 & -0.30 \\ -0.72 & -0.32 \\ -0.81 & 0.18 \\ -0.59 & -0.61 \end{bmatrix}$ $\begin{bmatrix} -0.70 & -0.09 \\ 0.00 & 0.05 \\ -1.42 & 0.05 \\ 0 & 0.91 \end{bmatrix}$ $\begin{bmatrix} 0.82 & 0.80 \\ 0.18 & -0.24 \\ 0.62 & -0.46 \\ -0.03 & -1.19 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & -1.36 \end{bmatrix}$ $\begin{bmatrix} 2.26 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.10 & -0.02 \end{bmatrix}$ $\begin{bmatrix} 1.66 & -0.70 \end{bmatrix}$ $\begin{bmatrix} 1.92 & -1.93 \end{bmatrix}$ $\begin{bmatrix} -0.70 & 0.23 \end{bmatrix}$	$\begin{bmatrix} -0.04 & -1.04 \end{bmatrix}$ $\begin{bmatrix} -0.0093 & 0.0062 \end{bmatrix}$ $\begin{bmatrix} 0.56 & 0.0037 \end{bmatrix}$ $\begin{bmatrix} 1.68 & -0.0012 \end{bmatrix}$	$\begin{bmatrix} -1.61 & -0.75 \end{bmatrix}$ $\begin{bmatrix} 0.03 & 0.87 \end{bmatrix}$ $\begin{bmatrix} 1.79 & 0 \end{bmatrix}$ $\begin{bmatrix} -1.11 & -0.74 \end{bmatrix}$
Final Loss	N/A	0.0001	0.0515	0.0001	0.0000

Table 12. Experiment 6: Comparison of layer weights for the teacher model and student models with **ReLU** activations. All weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment was run with seed 1.

	Teacher	L1-reg. (scaled 0.2)	L1-reg. (no scaling)	L2-reg. (no scaling)	No reg. (no scaling)
Layer 1	$\begin{bmatrix} 2.59 & -2.83 & 0.87 \\ -1.22 & 0.45 & 0.88 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 2.58 & -2.82 & 0.87 \end{bmatrix}$ $\begin{bmatrix} 1.22 & -0.45 & -0.88 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1.22 & -0.45 & -0.88 \end{bmatrix}$ $\begin{bmatrix} 2.58 & -2.82 & 0.87 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.14 & -0.16 & 0.08 \end{bmatrix}$ $\begin{bmatrix} -1.20 & 0.44 & 0.86 \end{bmatrix}$ $\begin{bmatrix} 2.48 & -2.72 & 0.84 \end{bmatrix}$ $\begin{bmatrix} 0.03 & 0.02 & 0.02 \end{bmatrix}$	$\begin{bmatrix} -0.15 & 0.27 & -0.21 \end{bmatrix}$ $\begin{bmatrix} -1.22 & 0.45 & 0.88 \end{bmatrix}$ $\begin{bmatrix} -2.59 & 2.83 & -0.87 \end{bmatrix}$ $\begin{bmatrix} -0.41 & 0.02 & 0.49 \end{bmatrix}$
Layer 2	$\begin{bmatrix} -1.38 & 1.29 \\ 0.35 & -0.73 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ -0.35 & 0 \\ 0 & -0.73 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 1.38 & 0 \\ -0.0002 & 1.29 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.0002 & 1.29 \\ 1.38 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0.73 \\ 0.35 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.04 & 0.0025 \\ 0.0018 & -1.20 \\ 1.30 & -0.0012 \\ 0.0055 & -0.0046 \end{bmatrix}$ $\begin{bmatrix} -0.02 & 0.0079 \\ -0.0060 & 0.01 \\ 0.02 & -0.02 \\ 0.02 & -0.0023 \end{bmatrix}$ $\begin{bmatrix} 0.01 & -0.0016 \\ 0 & 0.80 \\ -0.31 & -0.0005 \\ 0.0048 & -0.0162 \end{bmatrix}$ $\begin{bmatrix} 0.0046 & 0.0059 \\ -0.0022 & 0.09 \\ -0.22 & -0.0010 \\ -0.0012 & -0.0042 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 1.29 \\ 1.38 & 0 \\ 0 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \\ 0 & 0.73 \\ 0.35 & 0 \\ 0 & -0.0002 \end{bmatrix}$ $\begin{bmatrix} -1.05 & 0.55 \\ 0.01 & 0.31 \\ 0.63 & -0.40 \\ -0.47 & 0.25 \end{bmatrix}$ $\begin{bmatrix} -0.19 & 0.18 \\ 0.15 & 0.04 \\ -0.54 & -0.31 \\ -0.31 & -0.72 \end{bmatrix}$
Layer 3	$\begin{bmatrix} 0.86 & -0.84 \end{bmatrix}$	$\begin{bmatrix} -0.0010 & 0.84 \end{bmatrix}$ $\begin{bmatrix} -0.86 & -0.0019 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.86 & -0.0019 \end{bmatrix}$ $\begin{bmatrix} 0.0010 & -0.84 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.93 & -0.03 \end{bmatrix}$ $\begin{bmatrix} 0.0068 & -0.02 \end{bmatrix}$ $\begin{bmatrix} -0.0320 & 0.73 \end{bmatrix}$ $\begin{bmatrix} -0.15 & 0.18 \end{bmatrix}$	$\begin{bmatrix} 0.86 & 0 \end{bmatrix}$ $\begin{bmatrix} 0 & 0.84 \end{bmatrix}$ $\begin{bmatrix} 0 & 0 \end{bmatrix}$
Final Loss	N/A	0.0001	0.0001	0.0002	0.0000

Table 13. Experiment 6: Comparison of layer weights for the teacher model and student models with **tanh** activations. All weights are rounded to two decimal places (except for values between 10^{-2} and 10^{-4}). This experiment was run with seed 3.