Unveiling Memorization-Generalization Coexistence

A Case Study on Arithmetic Tasks with Label Noise

Linyu Liu (Taylor Lab, Huawei) Pinyan Lu (Taylor Lab, Huawei; ITCS, SUFE)

TL;DR

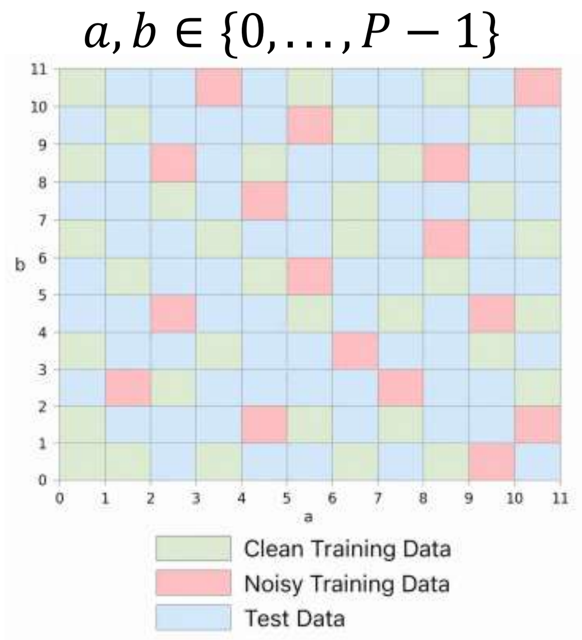
Neural networks can internally learn the true modular arithmetic rule even while their outputs memorize heavy label noise, but this coexistence is functionally separable by frequency filtration rather than structurally separated by neurons.

Problem Setup

Task: modular addition with noisy labels

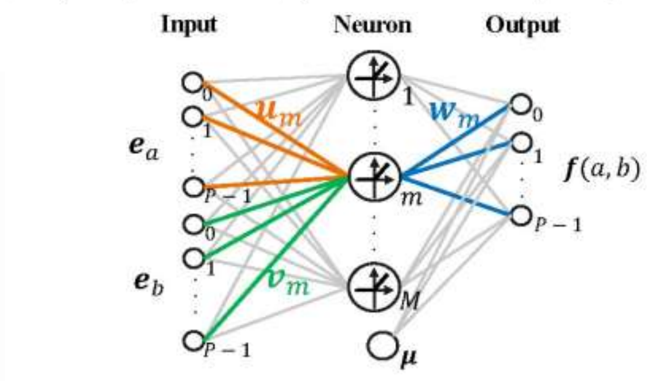
$c = a + b \bmod P$

$a, b \in \{0, \ldots, P-1\}$



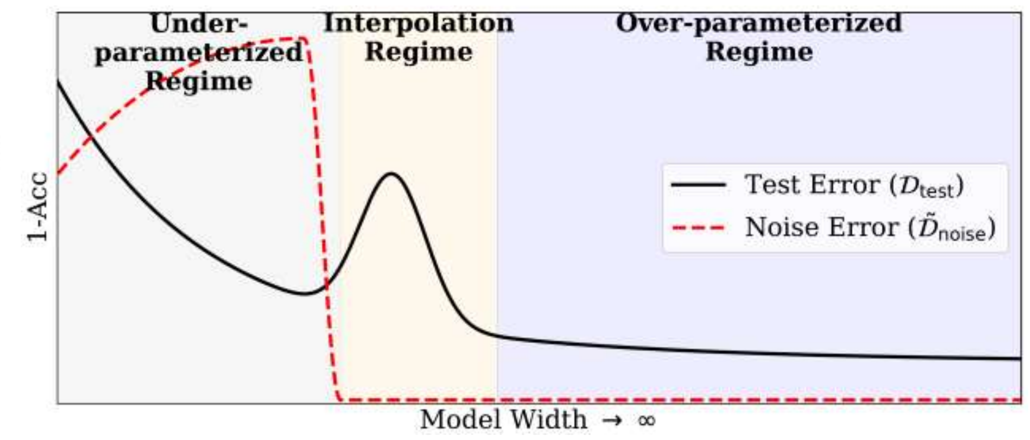
Model: two-layer NN

$$f_\theta(a,b) = \boldsymbol{W}\,\phi(\boldsymbol{U}\boldsymbol{e}_a + \boldsymbol{V}\boldsymbol{e}_b) + \boldsymbol{\mu},$$



Train with cross-entropy loss

Double descent phenomenon: Error descends after interpolating all the noise data

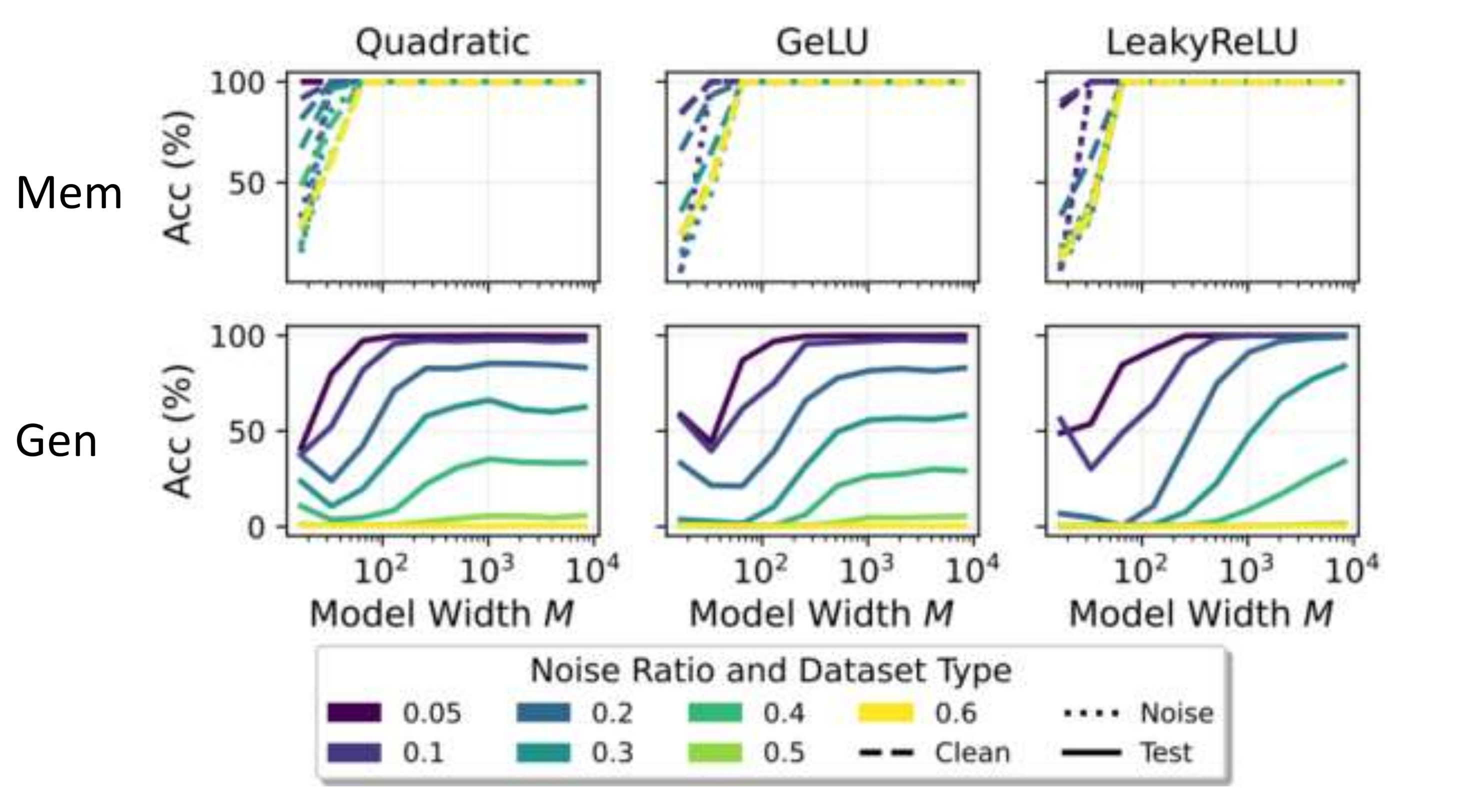


Memorization: 100% accuracy on training data

Generalization: ~100% accuracy on test data

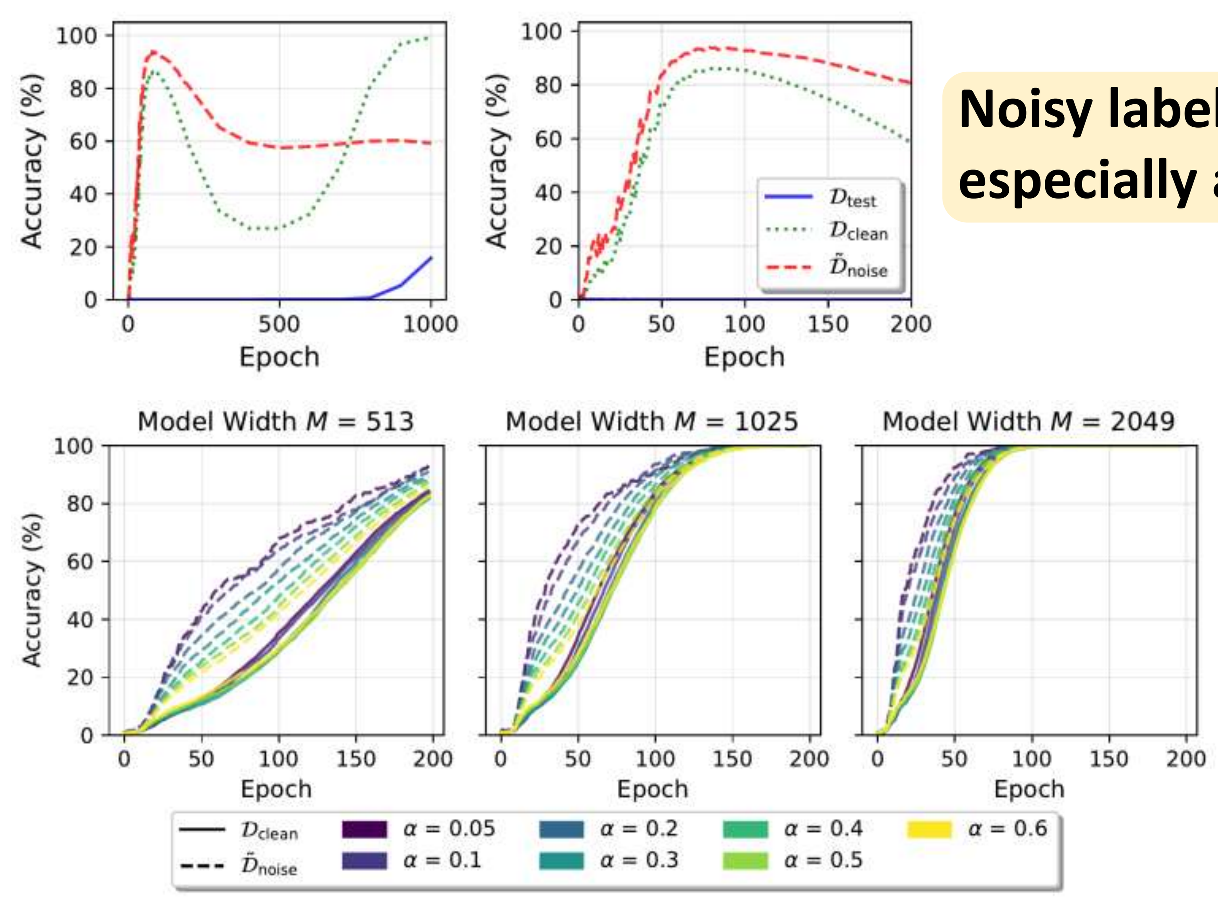
Overall Performance: Larger Models Perform Better; Noise is Memorized Faster

End-of-training performance



Larger is better after memorizing all the training data

Training dynamics



Noisy labels are memorized faster, especially at small noise ratio

- Dashed: Noisy labels
- Solid: Clean labels

Mechanism: Mem and Gen are Seperable in the Frequency Domain

The learned generalization algorithm

If using $\phi(\cdot) = x^2$:

$$(e^{i\omega a} + e^{i\omega b})^2 = e^{2i\omega a} + e^{2i\omega b} + 2e^{i\omega(a+b)}$$

How to represent frequency in NN?

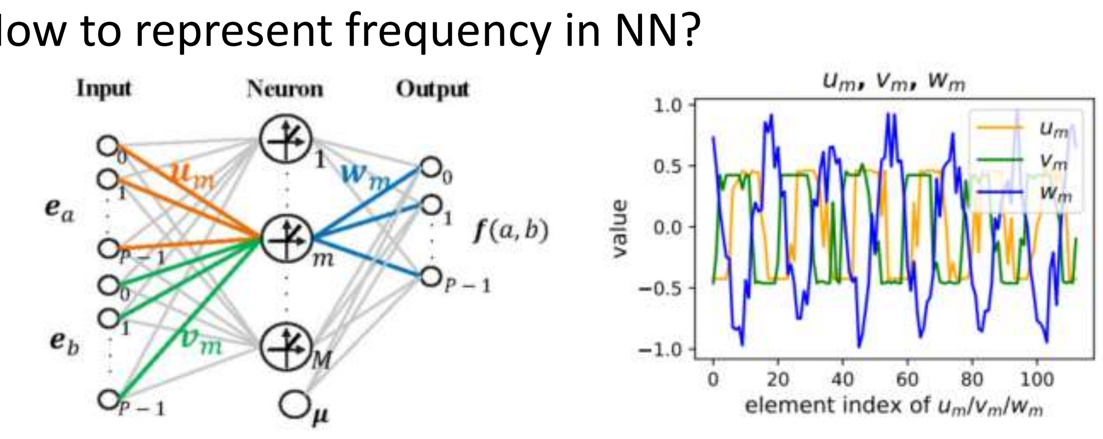
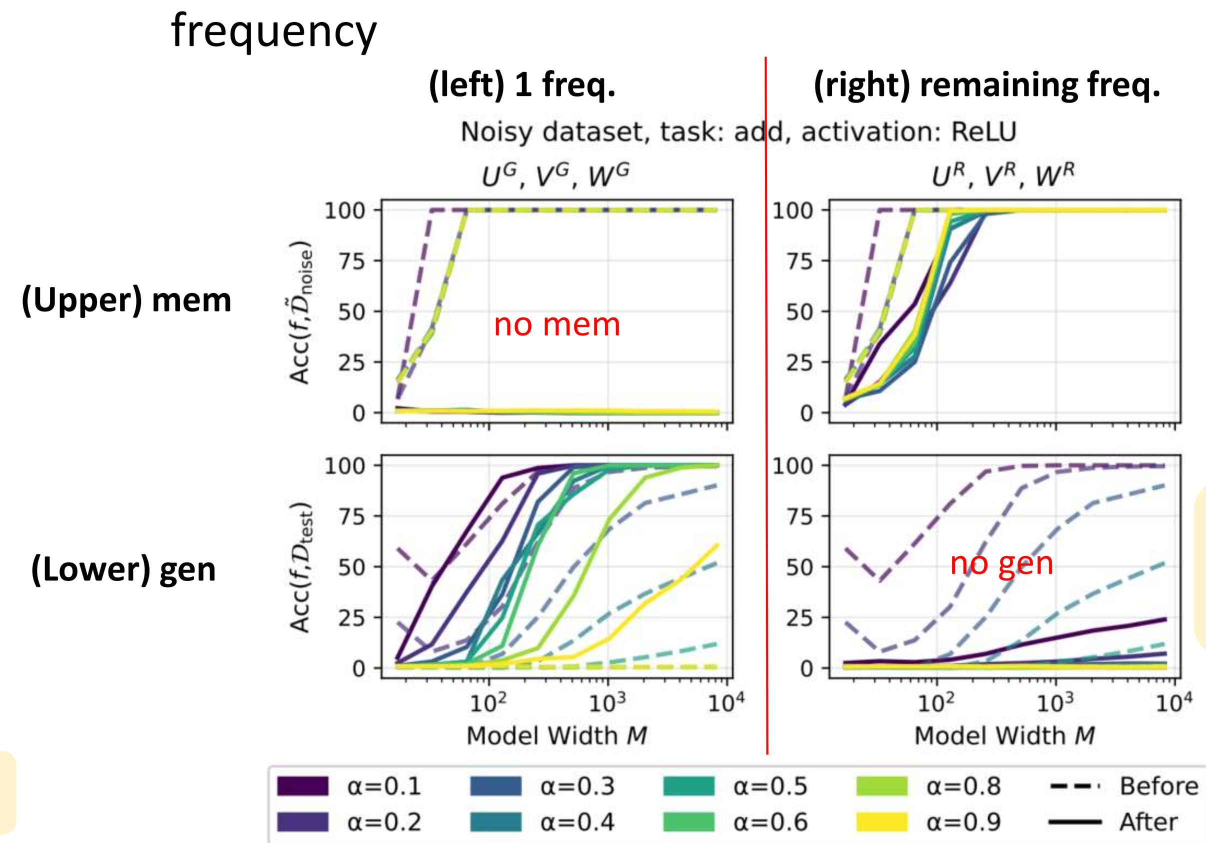


Figure 7. *(Left)* Illustration of a neuron's parameters $(\boldsymbol{u}_m, \boldsymbol{v}_m, \boldsymbol{w}_m)$. *(Right)* Visualization of a neuron from a ReLU model, with x-axis showing indices $0, \ldots, P-1$.

For each neuron, periodicity of the weight is observed.

Seperating memorization and generalization

Frequency filtration (FF) method: each neuron retains only one frequency



Successful seperation:
Left: gen, no mem
Right: mem, no gen

(Lower Left) Test acc ~100% under 80% noise ratio!

Not Seperable in Structure: Pruning Does Not Seperate Mem and Gen Thoroughly

Neuron pruning method: score, rank, and group.

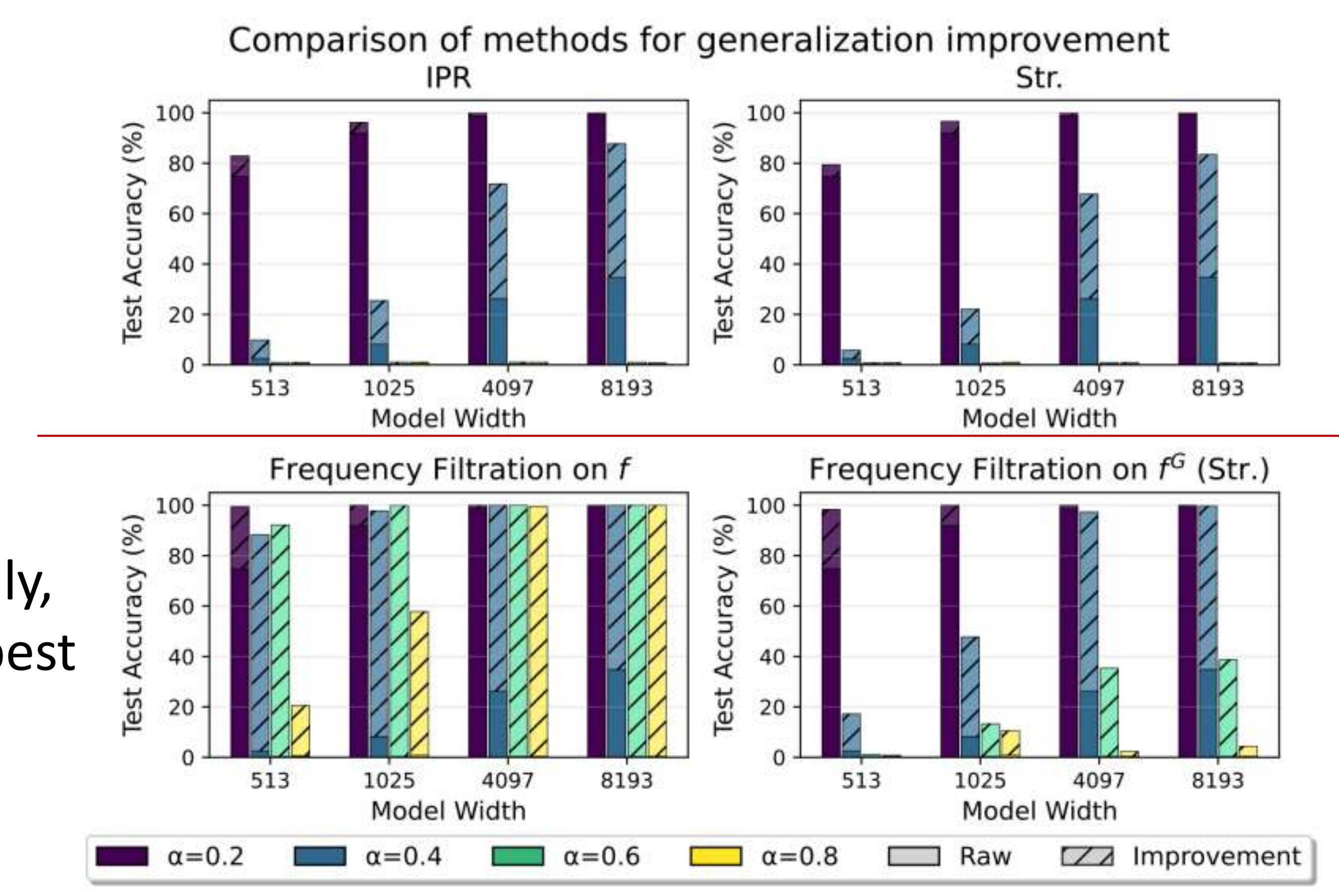
Gen subnet $f_{\mathcal{M}^G}(a,b) = \sum_{m \in \mathcal{M}^G} w_m \phi(\boldsymbol{u}_m^\top \boldsymbol{e}_a + \boldsymbol{v}_m^\top \boldsymbol{e}_b) + \boldsymbol{\mu},$

Mem subnet $f_{\mathcal{M}^R}(a,b) = \sum_{m \in \mathcal{M}^R} w_m \phi(\boldsymbol{u}_m^\top \boldsymbol{e}_a + \boldsymbol{v}_m^\top \boldsymbol{e}_b) + \boldsymbol{\mu}.$

Scored by periodicity (IPR) or neuron strength (Str)

$$\text{IPR}_m := \left(\frac{\|\tilde{\boldsymbol{w}}_m\|_4}{\|\tilde{\boldsymbol{w}}_m\|_2}\right)^4$$

$$\text{Str}_m := \|\boldsymbol{w}_m\|_\infty \phi\left(\|\boldsymbol{u}_m\|_\infty + \|\boldsymbol{v}_m\|_\infty\right)$$



Pruning improves generalization

FF only, gen best

FF after pruning further improves gen, but can not compete with FF only

Noise is distributedly memorized in the NN.