Benign overfitting in two-layer convolutional neural networks

Author(s): Y Cao, Z Chen, M Belkin, Q Gu

Presenter: Faiza Anan Noor

TABLE OF CONTENTS

01 Introduction

Here, we learn about the paper

02 Fundamentals

Here, we learn about some fundamental ideas behind the paper

03 Working Mechanism

Here, the Working Mechanism is depicted

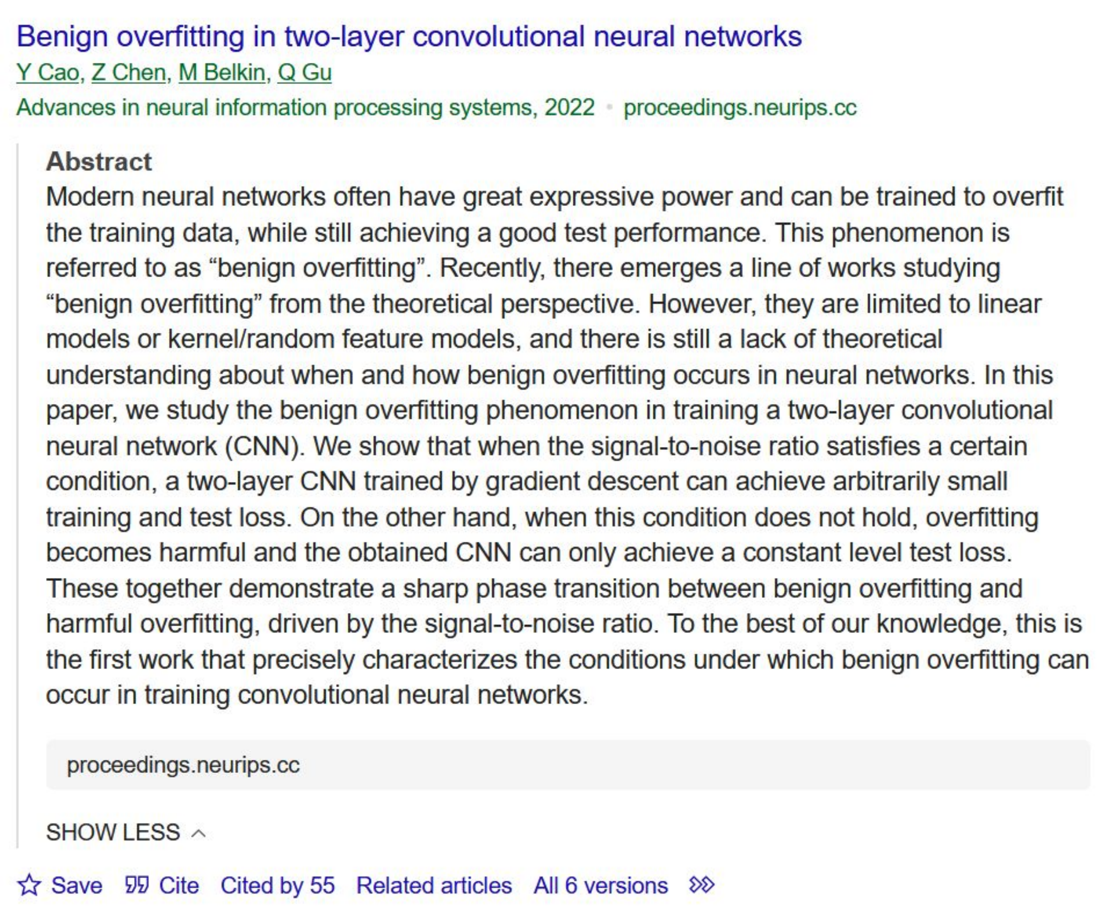
04 Results

The results are given

05 Challenges, Future Direction

The challenges faced and future directions plausible are given

01. Introduction



Benign overfitting in two-layer convolutional neural networks

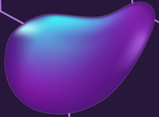
Y Cao, Z Chen, M Belkin, Q Gu

Advances in neural information processing systems, 2022 • proceedings.neurips.cc

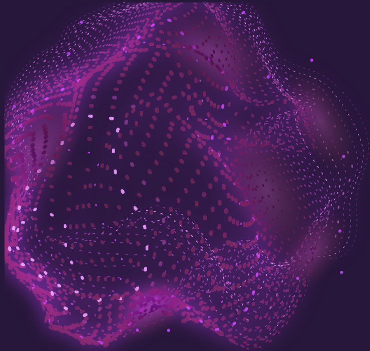
Abstract

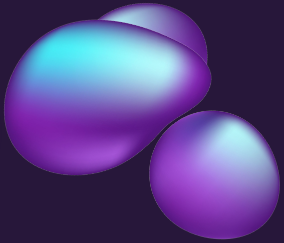
Modern neural networks often have great expressive power and can be trained to overfit the training data, while still achieving a good test performance. This phenomenon is referred to as “benign overfitting”. Recently, there emerges a line of works studying “benign overfitting” from the theoretical perspective. However, they are limited to linear models or kernel/random feature models, and there is still a lack of theoretical understanding about when and how benign overfitting occurs in neural networks. In this paper, we study the benign overfitting phenomenon in training a two-layer convolutional neural network (CNN). We show that when the signal-to-noise ratio satisfies a certain condition, a two-layer CNN trained by gradient descent can achieve arbitrarily small training and test loss. On the other hand, when this condition does not hold, overfitting becomes harmful and the obtained CNN can only achieve a constant level test loss. These together demonstrate a sharp phase transition between benign overfitting and harmful overfitting, driven by the signal-to-noise ratio. To the best of our knowledge, this is the first work that precisely characterizes the conditions under which benign overfitting can occur in training convolutional neural networks.

proceedings.neurips.cc



Main Idea





- **Focus**: Benign overfitting phenomenon in training a two-layer convolutional neural network (CNN).
- When the signal-to-noise ratio satisfies a certain condition, a two-layer CNN trained by gradient descent can achieve arbitrarily small training and test loss
- When this condition does not hold, overfitting becomes harmful and the obtained CNN can only achieve constant level test loss.

Main Idea

Focus: The process of signal learning and noise memorization in the training of **a two-layer CNN using a signal-noise decomposition.**

- A phase transition of the population loss with regard to sample size, signal intensity, noise level, and dimension is shown, and they accurately specify the parameters under which the CNN will primarily focus on learning signals or memorizing noises
- They establish population loss bounds of overfitted CNN models trained by gradient descent, and theoretically demonstrate that benign overfitting can occur in learning over-parameterized neural networks.

Findings: Under certain conditions on the signal-to-noise ratio, CNN* models trained by gradient descent will prioritize learning the signal over memorizing the noise, and thus achieving both small training and test losses.

- The first work that precisely characterizes the conditions under which benign overfitting can occur in training convolutional neural networks

Parallel work (Frei et al., 2022):

- Fully-connected two-layer neural networks

- Smoothed leaky ReLU activation to study learning log-Concave mixed data with label flip noise.

- When the label flip noise is zero, their risk bound coincides with the risk bound for linear models reported in Cao et al. (2021).

- Does not show the phase transition between harmful and benign overfitting;

- Focuses on upper bounding the risk. In contrast to (Frei et al., 2022),

- Main concentration on CNNs and examine an alternative data model to more accurately represent the characteristics of image classification issues.
- They show a distinct phase transition between beneficial and detrimental overfitting and give both positive and negative findings under various SNR regimes.

They also establish a negative result showing that when the conditions on the signal-to-noise ratio do not hold, then the overfitted CNN model will achieve at least a constant population loss. This result, together with their upper bound result, reveals an interesting phase transition between benign overfitting and harmful overfitting.

Paper: https://arxiv.org/abs/2202.06526

Convolutional neural network

Convolutional neural network is a regularized type of feed-forward neural network that learns feature engineering by itself via filters optimization.

- CNNs automatically learn hierarchical representations of features in images. Through convolutional and pooling layers, they detect patterns, textures, edges, and more complex features.

02. Fundamentals

Benign Overfitting:

- Modern deep learning models often consist of a huge number of model parameters, which is more than the number of training data points and therefore over-parameterized. These over-parameterized
- Models can be trained to overfit the training data (achieving a close to 100% training accuracy), while still making accurate prediction on the unseen test data.

02. Fundamentals

Benign Overfitting:

When a Convolutional Neural Network (CNN) begins to perform marginally better on training data but not noticeably worse on validation or test data, this is referred to as benign overfitting.

- "benign" as it does not negatively impact the model's capacity to generalize to new data.

Signal to Noise Ratio(SNR)

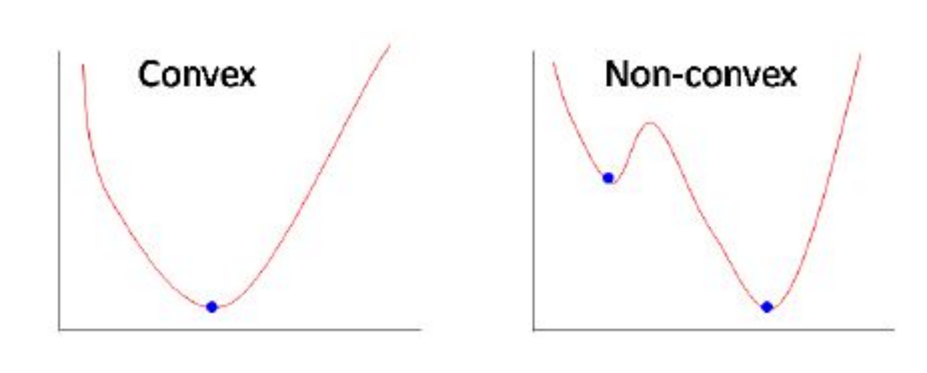
- The Ratio between the Signal and Noise in any particular medium is called SNR.

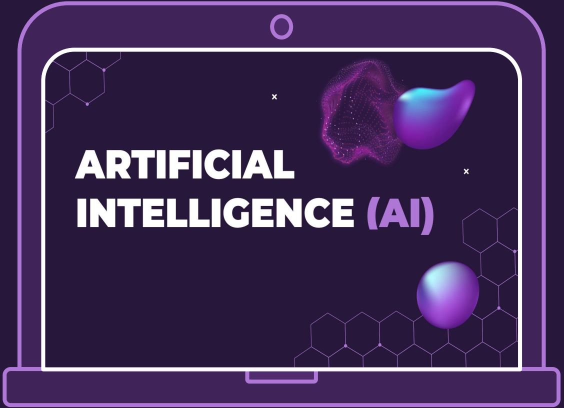
- Often expressed in decibels.

- A ratio higher than 1:1 indicates more signal than noise

Non convexity

- Such a problem may have multiple feasible regions and multiple locally optimal points within each region.
-
- For example, exchanging intermediate neurons · This symmetry means they can't be convex.



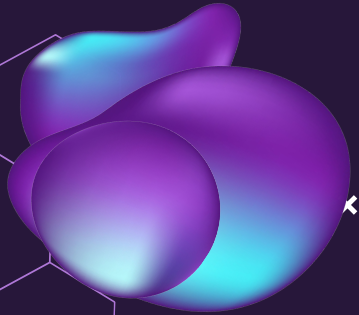


They provided algorithmic analysis for learning **two-layer convolutional neural networks (CNNs) .**

Filters are applied to the two patches **x1 and x2 separately,**

Second layer parameters being fixed as: **(+1/ms and −1/ms)**

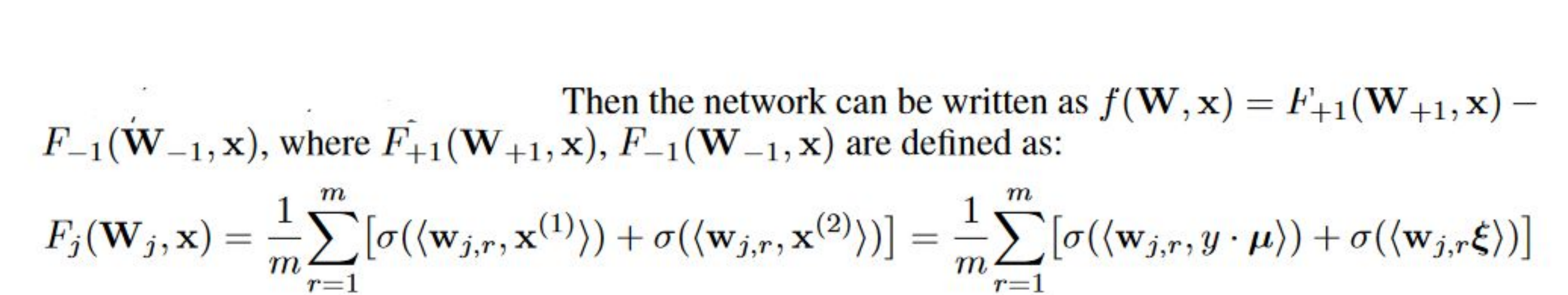
Polynomial ReLU activation function:
$\sigma(z) = \max\{0, z\}$ q ,
where q > 2 is a hyperparameter.



ARTIFICIAL INTELLIGENCE (AI)

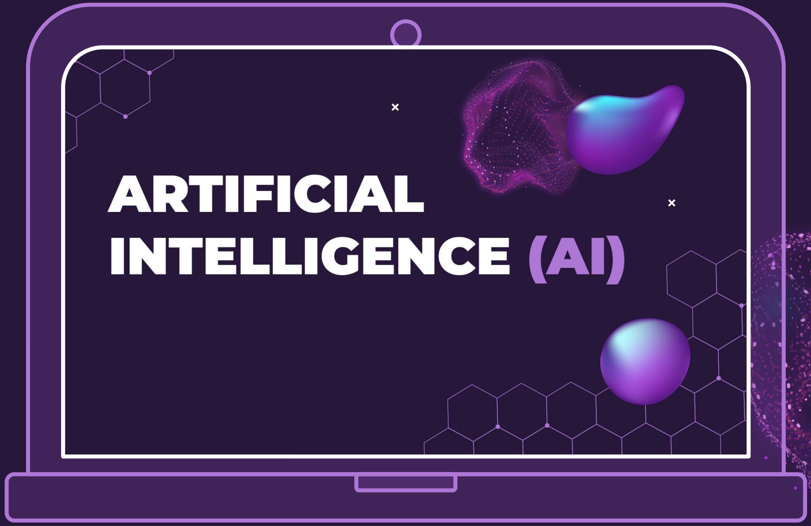
Classification : They focused on binary classification.

Noise : Generated from the Gaussian distribution N(0, σ2).



Then the network can be written as $f(\mathbf{W},\mathbf{x}) = F_{+1}(\mathbf{W}_{+1},\mathbf{x}) - F_{-1}(\mathbf{W}_{-1},\mathbf{x})$, where $F_{+1}(\mathbf{W}_{+1},\mathbf{x})$, $F_{-1}(\mathbf{W}_{-1},\mathbf{x})$ are defined as:

$$F_j(\mathbf{W}_j,\mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r},\mathbf{x}^{(1)}\rangle) + \sigma(\langle \mathbf{w}_{j,r},\mathbf{x}^{(2)}\rangle)] = \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}, y\cdot\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}\boldsymbol{\xi}\rangle)]$$



ARTIFICIAL INTELLIGENCE (AI)

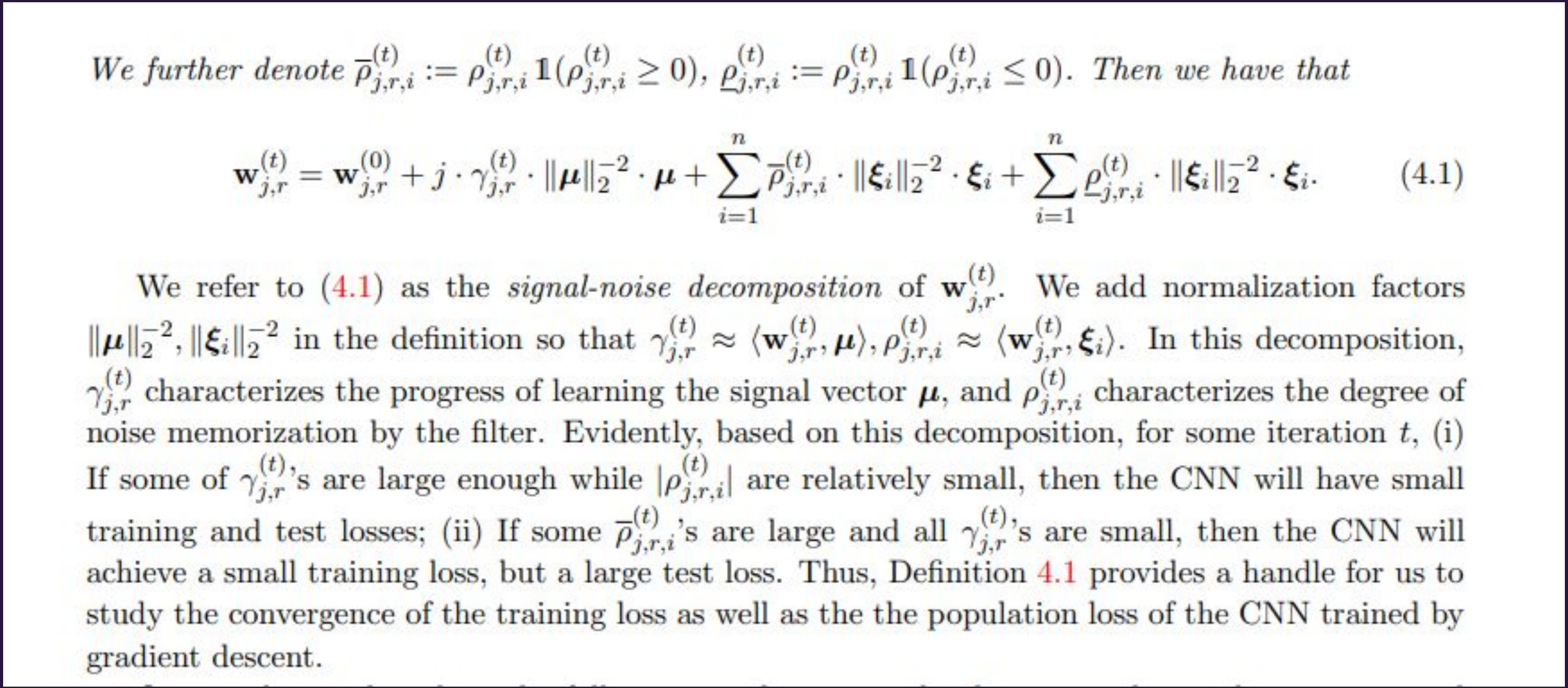
Settings:

Input data: Consist of label dependent signals and label independent noises

Signal-noise decomposition of the CNN filters to precisely characterize the signal learning and noise memorization processes during neural network training.

- They converted the neural network learning into a discrete dynamical system of the coefficients from the decomposition,
- Perform a two-stage analysis that decouples the complicated relation among the coefficients

Introduced new proof technique namely signal-noise decomposition, which decomposes the convolutional filters into a linear combination of initial filters, the signal vectors and the noise vectors.



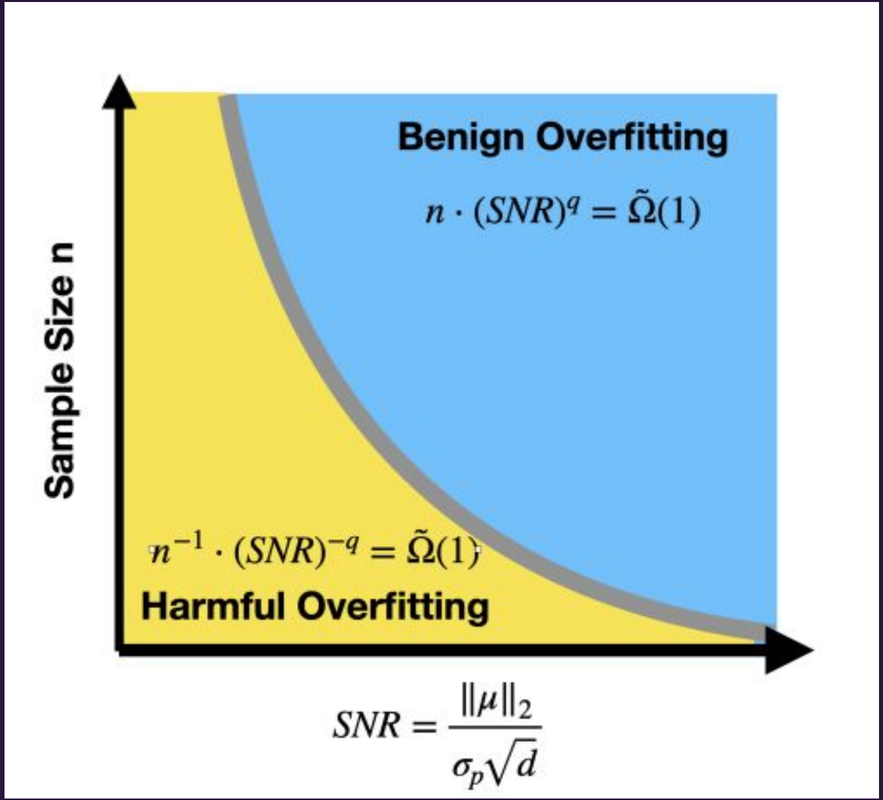
We further denote $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$. *Then we have that*

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i. \tag{4.1}$$

We refer to (4.1) as the *signal-noise decomposition* of $\mathbf{w}_{j,r}^{(t)}$. We add normalization factors $\|\boldsymbol{\mu}\|_2^{-2}, \|\boldsymbol{\xi}_i\|_2^{-2}$ in the definition so that $\gamma_{j,r}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle, \rho_{j,r,i}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$. In this decomposition, $\gamma_{j,r}^{(t)}$ characterizes the progress of learning the signal vector $\boldsymbol{\mu}$, and $\rho_{j,r,i}^{(t)}$ characterizes the degree of noise memorization by the filter. Evidently, based on this decomposition, for some iteration t, (i) If some of $\gamma_{j,r}^{(t)}$'s are large enough while $|\rho_{j,r,i}^{(t)}|$ are relatively small, then the CNN will have small training and test losses; (ii) If some $\overline{\rho}_{j,r,i}^{(t)}$'s are large and all $\gamma_{j,r}^{(t)}$'s are small, then the CNN will achieve a small training loss, but a large test loss. Thus, Definition 4.1 provides a handle for us to study the convergence of the training loss as well as the the population loss of the CNN trained by gradient descent.

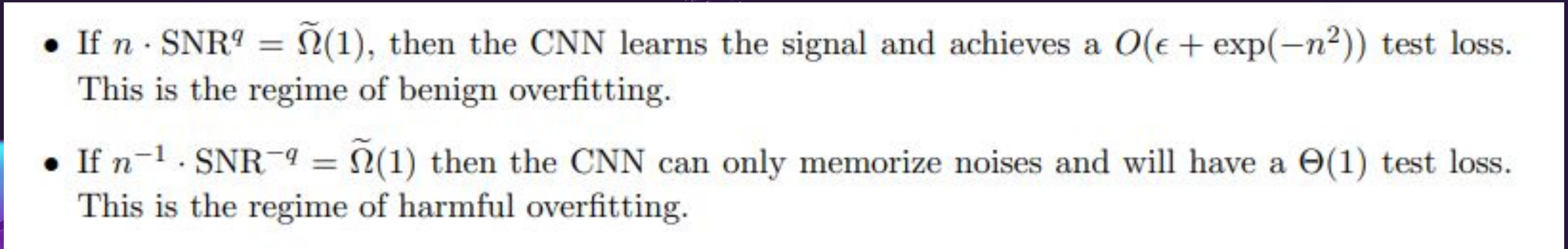
04. PREDICTED RESULTS

Illustration of the phase transition between benign and harmful overfitting. The blue region represents the setting under which the overfitted CNN trained by gradient descent is guaranteed to have small population loss, and the yellow region represents the setting under which the population loss is guaranteed to be of constant order. The slim gray band region is the setting where the population loss is not well characterized.



ARTIFICIAL INTELLIGENCE (AI)

Their result not only demonstrates that benign overfitting can occur in learning two-layer neural networks, but also gives precise conditions under which the overfitted CNN trained by gradient descent can achieve small population loss.



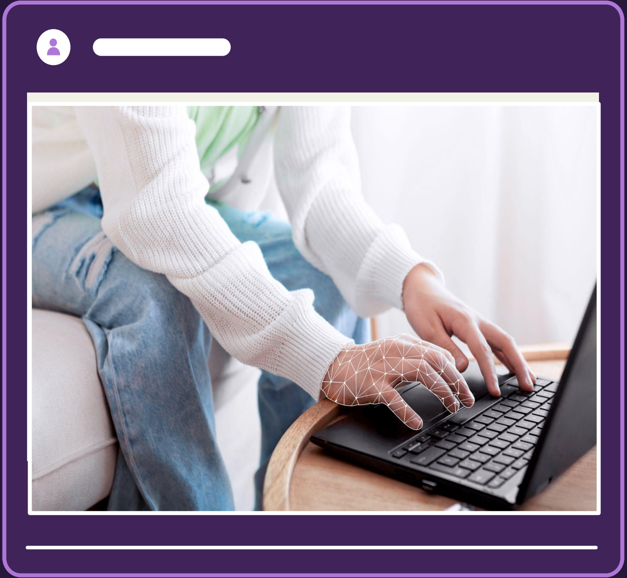
- If $n \cdot \mathrm{SNR}^q = \widetilde{\Omega}(1)$, then the CNN learns the signal and achieves a $O(\epsilon + \exp(-n^2))$ test loss. This is the regime of benign overfitting.
- If $n^{-1} \cdot \mathrm{SNR}^{-q} = \widetilde{\Omega}(1)$ then the CNN can only memorize noises and will have a $\Theta(1)$ test loss. This is the regime of harmful overfitting.

- Intuitively, the initial neural network weights are small enough so that the neural network at initialization has constant level cross-entropy loss derivatives on all the training data.

- Based on the result in the first stage, they then proceed to the second stage of the training process where the loss derivatives are no longer at a constant level and show that the training loss can be optimized to be arbitrarily small and meanwhile, the scale differences shown in the first learning stage remain the same throughout the training process

- This enables them to analyze the non-convex optimization problem, and bound the population loss of the CNN trained by gradient descent. Proof technique is of independent interest and can potentially be applied to deep neural networks.

05. challenges



Challenge:

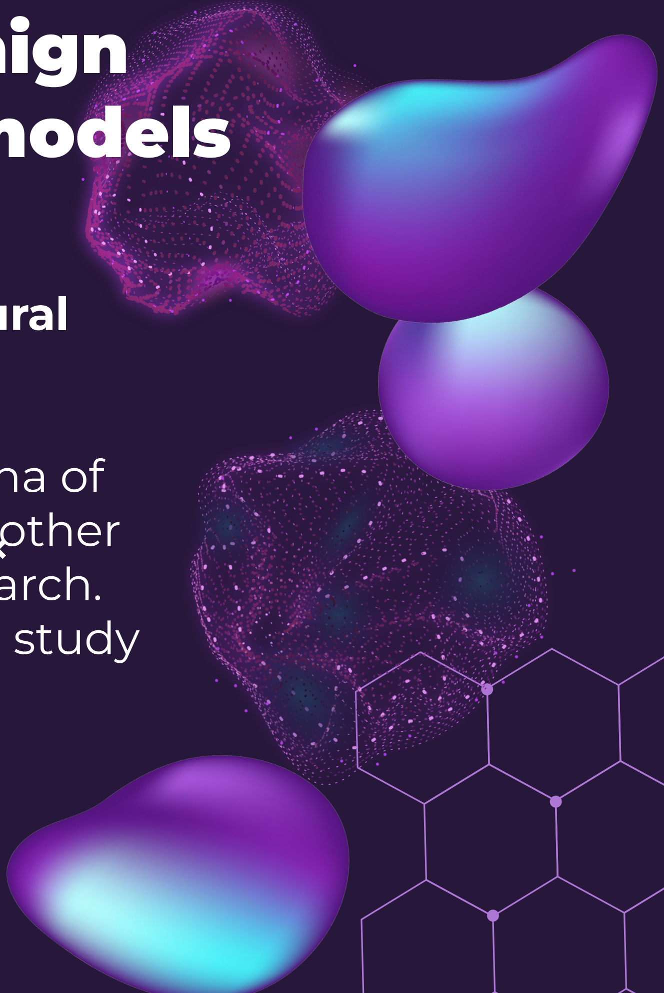
Nonconvexity of the training objective function LS(W). Nonconvexity has introduced new challenges in the study of benign overfitting particularly

Reason:

Their goal is to show the c**onvergence of the training loss and population loss** in the over-parameterized setting, which requires a precise algorithmic analysis of the learning problem

Future plans: To extend benign underfitting to other data models

- Their findings conceptually validates the possibility of benign overfitting during **neural network training.**
- Studying the benign overfitting phenomena of neural networks in the context of learning other data models is a crucial area of future research. Furthermore, it is imperative to extend our study to **deep convolutional neural networks.**



Discussion points

Some common solutions for overfitting? 1)

How does benign overfitting differ from underfitting? 2)

Why should benign overfitting be an issue if it causes slight generalization problems? 3)

THANKS!

Any questions?

References

- CAO, Y., GU, Q. and BELKIN, M. (2021). Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. Advances in Neural Information Processing Systems 34.

- FREI, S., CHATTERJI, N. S. and BARTLETT, P. L. (2022). Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. arXiv preprint arXiv:2202.05928.

- Cao, Y., Chen, Z., Belkin, M., & Gu, Q. (2022). Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, *35*, 25237-25250.