

Mathematical Theory of Deep Learning: Security and Generalization

Xiao-Shan Gao

Academy of Mathematics and Systems Science
Chinese Academy of Sciences

November 28, 2024

Research and Innovation Summit, Dongguan China

Theoretical Mysteries of Deep Learning and LLMs

There exists a **significant gap** between the theoretical foundation and the revolutionary practical success of DL and LLMs:

- Why do adversarial examples or data poisonings unavoidable and how to design optimal **secure DNNs and LLMs**.
- **Over-parameterized DNNs** are trained to almost memorize noisy data while still attaining nice generalization ability, contradicting statistical learning theory.
- Why does **SGD achieves generalizable model**, despite non-convexity?
- Why do DNNs and LLMs perform well for **very high-dimensional** data, mitigating the curse of dimensionality predicted by theory?
- How **network architecture** affects learning performance; how design nice architecture?
- Theory for **“emergence” and scaling law** of LMMs.

Table of Contents

- 1 Security and Robustness of Deep Learning
 - Robust Memorization: Existence of Robust DNNs
 - Achieving Optimal Robustness via Stackelberg Game
- 2 Generalization of Over-parameterized Neural Network
 - Generalizability of Neural Networks Minimizing Empirical Risk
 - Generalizability of Memorization Neural Network
- 3 OOD Generalization for Security of DL and LLMs
- 4 Summary

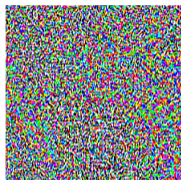
Mystery of Adversarial Samples

With little modifications **imperceptible to the human eye**, DNN outputs a **wrong label**



“panda”
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

Adversarial Attack

=



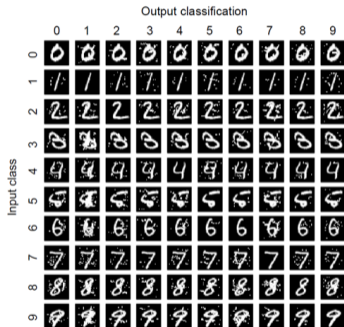
“gibbon”
99.3 % confidence

Adversarial Samples

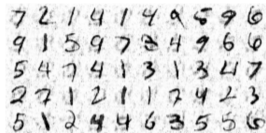
(Goodfellow-Shlens-Szegedy, 2014)

Targeted Adversary Attack

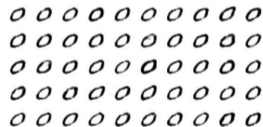
With little modifications, DNN outputs **any label given by the adversary**



Modify 4% pixels, 97% have adversaries
(Papernot et al. 2016)



Adversarial examples



Reconstruction of adversarial examples

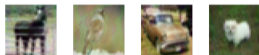
Adversarial examples for generative model
(J. Kos et al, 2017)

Data Poisoning Attack

Adversarial attack can also happen in the **Training Phase**, by adding imperceptible noises to the training data to degrade the poisoned DNN.



MNIST



CIFAR-10

Test accuracies decrease

from 99% to 0% for MNIST

from 77% to 28% for CIFAR10

(Feng et al. 2019)

Types of Data Poison Attacks:

- **Availability Attack:** misclassify all samples.
Protect data privacy by preventing unauthorized model training
- **Backdoor Attacks:** misclassify any input containing a backdoor trigger
Protect data/model ownership by adding watermarks
- **Targeted attack:** misclassify some samples as the given label

Adversarial Samples are Inevitable

There also exist theoretical results on the inevitability of adversarial samples.

- For any given DNN \mathcal{C} , $\exists \mathcal{D}$ such that if \mathcal{C} is accurate on \mathcal{D} , then \mathcal{C} has adversaries over \mathcal{D} with high probability. (Bastounis et al, 2020)

- **DNN is extremely sensitive to its parameters:**

Adversarial Parameter Attack: (Yu-Wang-Gao, 2023)

If the width of a DNN \mathcal{C} is sufficiently large,

then we can change the parameters of \mathcal{C} **as small as possible**,

such that the modified DNN has adversarial samples **as close as possible** to the normal samples.

Yu, Wang, Gao. Adversarial Parameter Attack on Deep Neural Networks, ICML 2023.

Some Basic Issues of Adversarial Learning

Adversarial Learning: Learning at the presence of adversaries, which is important for using Deep Learning in **safety-critical applications**.

From the **defense** aspect, we may ask

- For a given dataset, do there exist robust classifiers against any adversarial attack? **(100% secure?)**
- For given DNN structure and datasets, how to achieve optimal robustness against any adversarial attack? **(optimal secure?)**
- ...

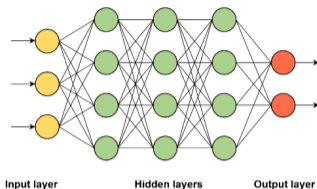
DNN: Piecewise Linear and Continuous Function

DNN: $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^m$

l -th Hidden Layer: $x_l = \max\{W_l x_{l-1} + b_l, 0\}$

Parameters: $\theta = \{W_l, b_l\}_{l=1}^L$

Classification Result: $\hat{\mathcal{C}}(x) = \arg \max_{i \in [m]} \mathcal{C}_i(x)$

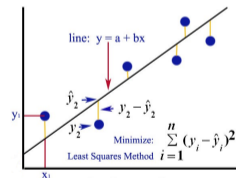


Training:

Training dataset: $D_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times [m]$

Empirical risk minimization:

$$\theta^* = \arg \min_{\theta} \sum_i \text{Loss}(\mathcal{C}_{\theta}(x_i), y_i)$$



Deep Learning: Approximate **high-dimensional functions** with DNNs, which is possible by the universal approximation theorem.

Classification Neural Networks

Training Dataset:

$D_{\text{tr}} = (x_i, y_i)_{i=1}^N \subset \mathbb{R}^d \times [m]$ with **data dimension d** and **m labels**

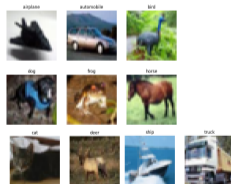
MNIST: $d = 784 = 28 \times 28$, $N = 60000$, $m = 10$

Train a DNN \mathcal{C}_θ on D_{tr} : $\mathcal{C} : \mathbb{R}^{784} \rightarrow \mathbb{R}^{10}$ by $\min_{\theta} \sum_{i=1}^N \text{Loss}(\mathcal{C}_\theta(x_i), y_i)$

Classification Result: $\hat{\mathcal{C}}(x) = \arg \max_{l=1}^{10} \mathcal{C}_l(x)$



MNIST: hand-written numbers

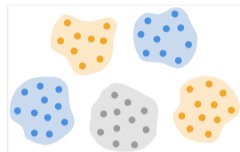


CIFAR-10: Ten classes of objects

Optimal Robust Memorization: Existence of Robust DNNs

Memorization and Robust Memorization

Dataset: $D_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N$ with **separation bound** $\lambda_{D_{\text{tr}}}$



Memorization Network $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}$ (Expressive power)

- if $\widehat{\mathcal{C}}(x_i) = y_i$ for all $i \in [N]$

Robust Memorization Network $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}$

- Robust Memorization with radius μ : $\widehat{\mathcal{C}}(x) = y_i, \forall \|x - x_i\| \leq \mu$
- Optimal Robust Memorization: \mathcal{C} is robust for all $\mu < \lambda_{D_{\text{tr}}}/2$.
- Optimal robust memorization DNN is **secure against all reasonable attacks**.

Given a DNN structure, it could be hard to find a robust DNN:

Theorem (Yu-Gao-Zhang, 2024)

It is NP-hard to compute (optimal) robust memorization for depth 2 and width 2 neural networks for any dataset.

Optimal Robust Memorization with a DNN

Given a Dataset: $D_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times [m]$ with separation bound $\lambda_{D_{\text{tr}}}$

Theorem (Yu-Gao-Zhang, 2024)

- *Necessary condition for existence of optimal robust memorization: width of DNN $> d$.*
- *Optimal Robust Memorization DNN: with width $3d + 1$, depth $O(N)$, and $O(Nd)$ parameters in polynomial-time.
*But, the parameter and Lipschitz values are large.**
- *Optimal robust memorization DNN via Lipschitz using $O(Nd \log(d))$ parameters in polynomial-time.
*The parameter and Lipschitz values are small.**

Summary on the Existence of Robust DNNs

Dataset: $D_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times [m]$

- For a given dataset, we can find optimal robust DNNs with width $O(d)$, depth $O(N)$, and $O(dN)$ parameters.

But, the depth ($N > 60000$ for MNIST) is too big to be practical.

- A natural question is

For DNNs with a given structure and given dataset, how to achieve optimal robustness against adversarial attacks?

Yu, Gao, Zhang. Optimal Robust Memorization with ReLU Neural Networks. ICLR 2024 (spotlight).

Achieving Optimal Robustness via Stackelberg Game

A Two-player Zero-sum Game:

Player 1: Classifier

- Strategy Space: $\mathcal{S}_c = [-E, E]^K$

To compute robust DNN with parameters $\theta \in \mathcal{S}_c$ of $\mathcal{C}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^m$.

Player 2: Adversary

- Strategy Space: $\mathcal{S}_a = \{A : \mathcal{X} \rightarrow \mathbb{B}_\varepsilon\}$, where $\mathbb{B}_\varepsilon = \{\delta \in \mathbb{R}^d : \|\delta\| \leq \varepsilon\}$

Adversarial Sample: $x + A(x)$ of x with a **small adversarial radius** ε .

Payoff Function:

$\phi(\theta, A) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \text{Loss}(\mathcal{C}_\theta(x + A(x)), y)$ for a data distribution \mathcal{D}

Goals of the players:

Classifier:	$\min_{\theta \in \mathcal{S}_c} \phi(\theta, A)$
Adversary:	$\max_{A \in \mathcal{S}_a} \phi(\theta, A)$

Adversarial Learning as a Stackelberg Game

Nash Equilibrium does not exist for DNNs!

A Stackelberg Game:

- Classifier plays first: $\min_{\theta \in B_\theta} \phi(\theta, A)$
- Adversary plays subsequently: $B_\theta = \arg \max_{A \in \mathcal{S}_a} \phi(\theta, A)$

Theorem (Gao-Liu-Yu, 2022)

- Game G has a Stackelberg equilibrium (θ^*, A^*) .
- C_{θ^*} is optimal robust DNN against all adversarial attacks with radius ϵ :

$$AA_{\mathcal{D}}(C_{\theta^*}, \epsilon) \geq AA_{\mathcal{D}}(C_\theta, \epsilon), \forall \epsilon \in \mathbb{R}, \theta \in [-E, E]^K$$

Adversarial Training: “the most successful empirical defense to date,”

“it is impossible to tell ... is truly robust.” (Cohen et al, 2019)

“it has shortages like ... non-provable.” (Bai et al, 2020)

Tradeoff Between Robustness and Accuracy

Tradeoff Phenomenon: There exists a tradeoff between accuracy and robustness (Tsipras et al, 2019):

Data	DNN	ϵ	Normal Training		Adv Training	
			Accuracy	Adv. Accu	Accuracy	Adv. Accu
CIFAR-10	ResNet18	8/255	94%	0%	84%	52%
	VGG16	8/255	93%	0%	79%	49%

Tradeoff problem can be described as a bi-level optimization problem:

$$\begin{aligned} \theta_o^* &= \arg \min_{\theta^*} \phi(\theta^*) \\ &\text{subject to } \theta^* \in \arg \min_{\theta \in \mathcal{S}_c} \max_{A \in \mathcal{S}_a} \phi_{cw}(\theta, A) \end{aligned}$$

For commonly used DNNs, tradeoff indeed happens: **Optimal robust DNN has less accuracy**

Summary on the Optimal Robust DNNs

- Stackelberg Game: A general framework for adversarial learning:
 - Optimal **robust/secure DNN** in a given hypothesis space.
 - Optimal **availability data poison attack**
 - Optimal **secure and generalizable DNN/LLM**
- But, the adversarial accuracy for the optimal DNN is still low: $\sim 50\%$ for CIFAR10.

Data	DNN	ϵ	Normal Training		Adv Training	
			Accuracy	Adv. Accu	Accuracy	Adv. Accu
CIFAR-10	ResNet18	8/255	94%	0%	84%	52%
	VGG16	8/255	93%	0%	79%	49%

Question: DNNs with **adversarial accuracy = accuracy** under more conditions.

Gao, Liu, Yu, Achieve Optimal Adversarial Accuracy for Adversarial DL using Stackelberg Game, Acta Math Sci, 2022.

Liu, Wang, Gao. Game-Theoretic Unlearnable Example Generator. AAAI 2024.

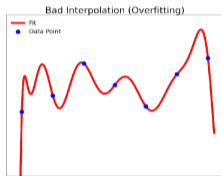
Liu, Wang, Gao. Mitigating Robust Overfitting in Wasserstein Distributionally Robust Optimization. ICLR 2025.

Generalization of Over-parameterized Neural Networks Based on Data Complexity

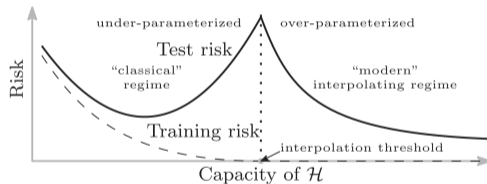
Mystery of Over-parameterized Memorization

One of the most surprising properties of Deep Learning:

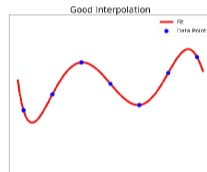
Over-parameterized DNN/LLM are trained to nearly **memorize noisy data** and yet still achieve very nice generalization, **contrary to classic statistical learning theory**



插值导致过拟合



(Belkin-Hsu-Ma-Mandal, 2019)



大模型插值无过拟合

Interpolation Learning: Learning with memorization neural networks.

Generalizability for Interpolation Learning or memorization DNNs?

Generalizability and Generalization Bound

Generalizability:

Models trained on **finite training set** have good performance over the **whole data distribution** or **OOD data**

Population Risk and Empirical Risk:

Population Risk: $\mathcal{R}_{\mathcal{D}}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\text{Loss}(\mathcal{C}_{\theta}(x), y)]$ over data distribution \mathcal{D}

Empirical Risk: $\mathcal{R}_{D_{\text{tr}}}(\theta) = \mathbb{E}_{(x,y) \in D_{\text{tr}}}[\text{Loss}(\mathcal{C}_{\theta}(x), y)]$ over training data $D_{\text{tr}} \sim \mathcal{D}^N$

(Uniform) Generalization Gap

$$\varepsilon_{\text{gen}}(\theta) = |\mathcal{R}_{\mathcal{D}}(\theta) - \mathcal{R}_{D_{\text{tr}}}(\theta)|$$

Generalization Bound:

$$\mathbb{E}_{D_{\text{tr}} \sim \mathcal{D}^N} \varepsilon_{\text{gen}}(\theta) \leq \epsilon_N$$

(Uniform) Sample Complexity:

$N \geq \text{poly}(d, \text{size}\Theta, \frac{1}{\epsilon}, \frac{1}{\delta})$ implies $\epsilon_N \leq \epsilon$ with probability $\geq 1 - \delta$

Uniform Generalization Bounds Fail for Over-Parameterized Regimes

Generalization Bound Based on VC-Dimension: $d_{VC}(\mathcal{H})$

$$\varepsilon_{\text{gen}}(h) \leq \sqrt{\tilde{O}\left(\frac{d_{VC}(\mathcal{H})}{N} + \frac{\ln(1/\delta)}{N}\right)}, \quad \forall h \in \mathcal{H}$$

$d_{VC}(\mathcal{H}) \simeq \text{para}(h) \cdot \text{depth}(h) \geq N$ for over-parameterized models (Bartlett et al 2019)

Fail to explain generalizability of over-parameterized models

Generalization Bound Based on Rademacher Complexity: $\text{Rad}_N(\mathcal{H})$

$$\varepsilon_{\text{gen}}(h) \leq 2 \cdot \text{Rad}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2N}}, \quad \forall h \in \mathcal{H}$$

$\text{Rad}_N(\mathcal{H}) \simeq 1$ (Zhang et al 2017)

Fail for over-parameterized regimes

Stability Gen Bounds Cannot Explain Over-Parameterized Regimes

Stability Generalization Bounds:

Uniform Stability Bounds (Hardt 2016)

$$\varepsilon_{\text{gen}} = 2\alpha T \frac{L^2}{N}$$

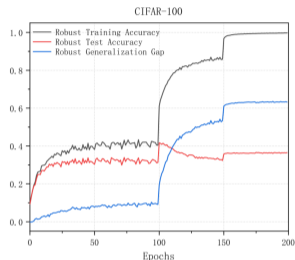
(Difficulty to show $L^2 < N$ for over-parameterized model)

On-Average Stability Bounds (Kuzborskij 2018)

$$\varepsilon_{\text{gen}} = O\left(\frac{\alpha\sigma LT + L\sqrt{\alpha r T}}{N}\right)$$

Risk $\text{Loss}(\mathcal{C}_\theta(x), y)$ is convex and L -Lipschitz in θ

SGD: $\theta_{k+1} = \theta_k - \alpha \nabla_\theta \text{Loss}(z_r, \theta_k)$, iterates T times



Accuracy and ε_{gen} in terms of T

Generalizability of Neural Networks

Minimizing Empirical Risk

Generalization Bound of Neural Networks Minimizing Empirical Risk

- $\mathcal{M}_W(\mathcal{D}_{tr})$: Two-layer networks with width W and minimizing the empirical risk over \mathcal{D}_{tr}
- **A new complexity of \mathcal{D}** : \mathcal{D} can be correctly classified by two-layer NNs of width $W_{\mathcal{D}}$.

Theorem (Yu et al, 2024)

For $\mathcal{D}_{tr} \sim \mathcal{D}^N$ and $\mathcal{F} \in \mathcal{M}_W(\mathcal{D}_{tr})$, it holds with probability $1 - \delta$,

Generalization Bound: $A_{\mathcal{D}}(\mathcal{F}) \geq 1 - \tilde{O}\left(\frac{W_{\mathcal{D}}}{W} + \frac{dW_{\mathcal{D}} + \sqrt{\ln(2/\delta)}}{\sqrt{N}}\right)$ (Both N and W in the denominator!)

Sample Complexity: $N \geq \Omega(W_{\mathcal{D}}^2)$ and $W \geq \Omega(W_{\mathcal{D}})$ (N and W given independently!)

Generalizability of over-parameterized models can be derived.

- Compare to VC-Dimension Bound: $N \geq \bar{O}(dW)$: Fails for over-parameterized models

Yu, Miao, Zhu, Gao, Zhang. Generalizability of Neural Networks Minimizing Empirical Risk Based on Expressive Power. ICLR 2025.

Generalization Bounds for Memorization Network

Sample Complexity for Memorization Networks

$P_{\mathcal{D}}$: memorization parameter complexity (a new complexity for \mathcal{D})

Theorem (Yu et al, 2024)

For i.i.d. dataset $\mathcal{D}_{\text{tr}} \sim \mathcal{D}^N$

- **Constant number of parameter memorization:**

There exist memorization DNN for any $\mathcal{D}_{\text{tr}} \sim \mathcal{D}^N$ with $P_{\mathcal{D}}$ parameters.

(In general, memorization neural network need \sqrt{N} parameter) (Vardi et al 2021)

- **Lower bound for Sample Complexity:** $\overline{O}(P_{\mathcal{D}}^2)$

Memorization networks of any $\mathcal{D}_{\text{tr}} \sim \mathcal{D}^N$ to be generalizable $\Rightarrow N \geq \overline{O}(P_{\mathcal{D}}^2)$.

For some \mathcal{D} over \mathbb{R}^d , $P_{\mathcal{D}} \geq c2^d$: exponential number of parameters needed to achieve generalization

- **Upper bound for Sample Complexity** for memo network with $P_{\mathcal{D}}$ parameters: $\overline{O}(P_{\mathcal{D}}^2)$

If $P \neq NP$, the *under-parameterized* memo network ($P_{\mathcal{D}}$ vs $P_{\mathcal{D}}^2$) cannot be obtained in polynomial time.

Efficient Memorization Sample Complexity

$E_{\mathcal{D}}$: Efficient Memorization Sample Complexity (a new complexity of \mathcal{D})

Theorem (Yu-Gao-Zhang, 2024)

If $N = \overline{O}(E_{\mathcal{D}})$, then a **generalizable** memorization network with $O(N^2 d)$ parameters for any $\mathcal{D}_{\text{tr}} \sim \mathcal{D}^N$ can be computed in **polynomial time**.

Consider a “nice” Data Distribution:

$\mathcal{D}(d, c)$ data distributions on \mathbb{R}^d with a **positive separation bound** c

For any $\mathcal{D} \in \mathcal{D}(d, c)$, we have $E_{\mathcal{D}} \leq ([6.2d/c] + 1)^d$.

For some $\mathcal{D} \in \mathcal{D}(d, c)$, we have $E_{\mathcal{D}} \geq c2^d$.

Question: For what type of data distribution, $E_{\mathcal{D}}$ is polynomial?

Yu, Gao, Zhang. Generalizability of Memorization Neural Network, PeurIPS 2024.

Summary on Generalization of Over-parameterized Neural Network

A new approach to analyze generalization of over-parameterized neural networks, based on **new data distribution complexities** $W_{\mathcal{D}}, P_{\mathcal{D}}, E_{\mathcal{D}}$.

- **Sample Complexity for Networks Minimizing the Empirical Risk:**

$$N \geq \Omega(W_{\mathcal{D}}^2) \quad \text{and} \quad W \geq \Omega(W_{\mathcal{D}})$$

- **Sample Complexity for Memorization Networks:**

Lower bound: $\tilde{O}(P_{\mathcal{D}}^2)$.

Upper bound for efficient memorization: $E_{\mathcal{D}}$.

- **Generalizability of over-parametrized models can be explained in certain sense.**

Yu, Gao, Zhang. Generalizability of Memorization Neural Network, PeurIPS 2024.

Yu, Miao, Zhu, Gao, Zhang. Generalizability of Neural Networks Minimizing Empirical Risk Based on Expressive Power. ICLR 2025.

SR-WDRO (Wasserstein distributionally robust optimization):

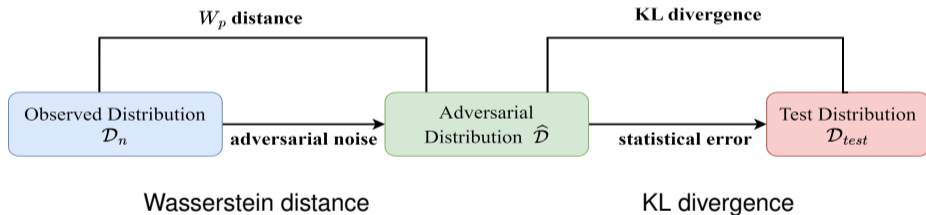
A General Framework for Security and Generalization

SR-WDRO: A General Framework for Security and Generalization

WDRO (Staub et al 2017): Wasserstein distributionally robust optimization has **robust overfitting**

A New Perturbation Set: (\mathcal{D}_N is the sampled training data distribution)

$$\mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N) := \{\mathcal{D}' : \exists \mathcal{D}'' \text{ s.t. } W_p(\mathcal{D}_N, \mathcal{D}'') \leq \varepsilon, \text{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma\}$$



SR-WDRO: Statistically Robust WDRO Game: $\inf_{\theta \in \Theta} \sup_{\mathcal{D} \in \mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)]$

Theorem (Liu et al 2024)

Let \mathcal{D}_N be sampled i.i.d. from \mathcal{D} and $\text{LP}(\mathcal{D}_{te}, \mathcal{D}) \leq \sigma$. Then for $\delta = \left(\frac{\varepsilon}{\text{diam}(\mathcal{Z})+1}\right)^2$, and $\forall \theta \in \Theta$

OOD Generalization: $\mathbb{P}\left(\mathbb{E}_{z \sim \mathcal{D}_{te}}[\text{Loss}(\theta, z)] \leq \mathcal{L}_{\varepsilon+\sigma, \gamma}(\theta, \mathcal{D}_N)\right) \geq 1 - e^{-\gamma N} \left(\frac{4}{\delta}\right)^{m(\mathcal{Z}, \delta)}$

ST-WDRO Loss: $\mathcal{L}_{\varepsilon+\sigma, \gamma}(\theta, \mathcal{D}_N) := \sup_{\mathcal{D} \in \mathcal{U}_{\varepsilon+\sigma, \gamma}(\mathcal{D}_N)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)]$

$m(\mathcal{Z}, \delta)$: *covering number* of the data support set \mathcal{Z}

OOD Generalization Bound:

- Valid for generalization and **adversarially robust generalization**
- Independent of VC-dim/Rad of Hyp: works for **over-parameterization regime**
- Can be made a **practical** training method
- Valid for all models, including MLP and transformer
- Learnability established when the intrinsic dimension of \mathcal{Z} is a small constant

Game and Optimal Statistically Robust Models

SR-WDRO Game: $\inf_{\theta \in \Theta} \sup_{\mathcal{D} \in \mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)]$

Theorem (Stackelberg Equilibrium)

The Stackelberg equilibrium exists for *neural networks*:

$$\theta^* \in \arg \min_{\theta \in \Theta} \max_{\mathcal{D} \in \text{BR}(\theta)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)], \quad \text{BR}(\theta) = \arg \max_{\mathcal{D} \in \mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)].$$

- Stackelberg equilibrium gives the **smallest statistically robust loss** among all models.

Theorem (Nash Equilibrium)

If $\text{Loss}(\theta, z)$ is *convex in θ* , then Nash equilibrium exists

$$\min_{\theta \in \Theta} \max_{\mathcal{D} \in \mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N)} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)] = \max_{\mathcal{D} \in \mathcal{U}_{\varepsilon, \gamma}(\mathcal{D}_N)} \min_{\theta \in \Theta} \mathbb{E}_{z \sim \mathcal{D}}[\text{Loss}(\theta, z)].$$

Liu, Miao, Wang, Zhu, Gao, Zhang. Mitigating Robust Overfitting in Wasserstein Distributionally Robustness. ICLR 2025.

Summary on OOD Generalization

OOD (Out of Distribution) Generalization: Adversarial attacks, data poison attacks, and domain shift.

- Generalization bound for SR-WDRO
- Generalization bound for backdoor poison attacks
- Generalization bound for availability attack
- Generalization bound for 3D point cloud robustness

Yu, Liu, Miao, Gao, Zhang. Generalization Bound and New Algorithm for Clean-Label Backdoor Attack. ICML 2024.

Zhu, Miao, Dong, Gao. Toward Availability Attacks in 3D Point Clouds. ICML 2024.

Y Miao, Y Dong, J Zhang, L Yu, X Yang, XS Gao. Improve Robustness of 3D Point Rec: Fourier Perspective, NeurIPS 2024.

Liu, Miao, Wang, Zhu, Gao, Zhang. Mitigating Robust Overfitting in Wasserstein Distributionally Robustness. ICLR 2025.

● Robustness and Security of Deep Learning

- **Robust Memorization**: There exist optimal robust DNNs with $O(Nd)$ parameters.
- **Adversarial Stackelberg Game**: A general approach to give optimal secure DNNs.

● Generalizability of Over-parameterized Networks

- **Networks Minimizing the Empirical Risk**: Sample complexity that can explain the nice generalizability of over-parameterized networks.
- **Memorization Networks**: Sample complexity and efficient sample complexity are given.

● OOD Generalization for Robustness

Generalization bounds are given for SR-WDRO optimization, backdoor poisoning attack, and on-average stability bound for adversarial training.

Thanks!