



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMA
ZIONE

UNIVERSITY OF PADOVA

INFORMATION ENGINEERING

WHY OVERPARAMETERIZED NEURAL NETWORKS GENERALIZE?

SUPERVISOR

PROF. LUCA SCHENATO

CANDIDATE

KAVITA SINGLA

ACADEMIC YEAR

2024-2025

GRADUATION DATE

19.11.2025

To my family and friends

Abstract

Classical statistical theory predicts that as model complexity increases, the model can risk perfectly interpolating the training data which can lead to overfitting. This is based on a well-known Foundational Generalization theory i.e. Bias-Variance Tradeoff. Yet, in modern deep neural networks, overparameterization defies this prediction. In deep learning, overparameterization as expected achieve near-zero training error but surprisingly they still generalize impressively.

This thesis evolves our idea of generalization from classical frameworks such as VC (Vapnik-Chervonenkis) dimension, PAC (Probably approximately Correct) learning, and explicit regularization to modern explanations like implicit regularization, flat minima, NTK (Neural Tangent Kernels), PAC-Bayes theory, Information bottleneck, and double-descent phenomenon and tries to bridge the gap between them.

By synthesizing theoretical and empirical insights, this thesis investigates how classical measures of model capacity fail to capture the geometric and dynamic features of deep learning. Furthermore, this thesis discusses the practical and experimental challenges faced by industry-scale models emphasizing the constraints faced by modern deep learning research. The thesis will conclude by outlining open theoretical challenges and suggesting the future work toward a unified generalization theory for deep learning.

Contents

ABSTRACT	iii
LISTING OF FIGURES	vi
LISTING OF TABLES	vii
1. INTRODUCTION	1
1.1 Motivation and Context.....	1
1.2 Challenges and State of the Art.....	2
1.2.1 Challenges.....	2
1.2.2 State of the Art.....	3
1.3 Contribution of this Thesis.....	4
1.4 Thesis Structure Overview.....	4
2. CLASSICAL STATISTICAL GENERALIZATION THEORY	6
2.1 Empirical Risk Minimization.....	6
2.2 PAC Learning.....	6
2.3 VC Dimension and Uniform Convergence.....	6
2.4 Bias-Variance Decomposition.....	7
2.5 Explicit Regularization.....	8
3. THE MODERN GENERALIZED PARADOX AND REGULARIZATION.....	9
3.1 The Paradox of Over-Parameterization.....	9
3.2 Double-Descent Behavior.....	10
3.3 Implicit versus Explicit Regularization.....	11
3.4 Flat-Minima Hypothesis.....	12
3.5 Modern Theories of Overparameterization.....	12
4. OVER-PARAMETERIZATION IN MODERN FOUNDATION MODELS.....	16
4.1 Overview of Large-Scale Models.....	16
4.2 Degree of Over-Parametization.....	16
4.3 Why Do They Generalize?.....	18

5. CHALLENGES AND OPEN PROBLEMS	19
5.1 Theoretical Challenges.....	19
5.2 Experimental Challenges.....	19
5.3 Industrial Challenges.....	20
5.4 Open Research Questions.....	20
6. CONCLUSION AND FUTURE WORK	22
6.1 Summary of Findings.....	22
6.2 Contributions of this Thesis.....	22
6.3 Future Research Directions.....	23
6.3.1 Toward a Unified Generalization Theory.....	23
6.3.2 Quantitative Understanding and Formulation of Implicit Regularization.....	23
6.3.3 Combating the Computational Expense of Large-Scale Training.....	24
6.3.4 Social and Ethical Considerations.....	24
6.4 Final Remarks.....	24
APPENDIX A	26
REFERENCES	28
ACKNOWLEDGEMENT	30

Listing of figures

2.1: The classical U-shaped risk curve	7
3.1: The double descent risk curve	10
3.2: Conceptual sketch of Flat minima hypothesis.....	12
3.3: Convergence of Neural Tangent Kernel and network function across different varying widths	13
3.4 Test accuracy among different datasets with varying sparsity levels	14
4.1: Conceptual explanation of optimization trajectory through parameter space during training	17

Listing of tables

1.1: Sizes, architecture and hyperparameters of GPT models.....	2
3.1: The training and test accuracy of various models on CIFAR 10 dataset.....	9
3.2: Accuracy of training and test sets with and without regularization.....	11

Introduction

1.1 MOTIVATION AND CONTEXT

A central goal of Machine learning algorithms is to learn patterns from data and improve performance through experience. Formally, given training data (x_i, y_i) drawn from an unknown distribution $D(X, Y)$, the goal is to find a hypothesis or model $h(x)$ within a hypothesis class \mathcal{H} that minimizes the expected risk:

$$L(h) = \mathbb{E}_{(x,y) \sim P}[\ell(h(x), y)],$$

where ℓ is a loss function (e.g., squared loss cross-entropy).

Because of the fact that true distribution P is unknown, we minimize the empirical risk by:

$$L_n(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i).$$

The objective of learning is to minimize $L_n(h)$ and $L(h)$, so that the model generalizes well to unseen data. Historically, learning theories emphasize controlling model capacity to ensure generalization to unseen data.

Deep learning has transformed Artificial Intelligence in the last decade. Modern neural networks contain hundreds of millions or trillions of parameters, yet they exhibit accuracy on unseen data. These successes challenge the center of classical learning theory which states that excessive model complexity leads to overfitting [1][2].

Modern neural networks, for example,

1. Gpt-4(Open AI, 2023) and Gemini 1.5 (Google DeepMind, 2024) contain hundreds of billions of parameters, yet they achieve human-like reasoning and language generation abilities instead of just memorizing. Not only this, but they are also good at compositional understanding across diverse tasks, from mathematics and code synthesis to creative writing and scientific summarization.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 1.1: Sizes, architecture and hyperparameters of GPT models [18].

This table shows the progression of OpenAI’s GPT-3 model family from 125 million to 175 billion parameters. This shows that every version beyond GPT-3 operates deeply within the overparameterized regime and as it clearly visible in the table that Batch Size \gg Number of Parameters and thus demonstrate that larger networks can enhance generalization when paired with suitable optimization dynamics and data diversity.

2. DALL-E and Stable Diffusion can generate realistic images from text description even when they are trained on highly noisy datasets.
3. Even some smaller models like ResNet-50 or BERT demonstrate a remarkable generalization even when they are smaller but the point to be noted is they still are overparameterized relative to their training sets.

These models are empirical proof that shows modern deep networks operate in regime in $p \gg n$ (far more parameters than effective samples), but still the generalization did not collapse. This paradox of generalization in deep learning motivates us to delve deeper into theoretical analysis. Why do models that are large enough to memorize entirely their training sets instead learn representations that generalize? Understanding this phenomenon is vital for both theoretical analysis and practical development.

1.2 CHALLENGES AND STATE OF THE ART

1.2.1 Challenges

Despite a remarkable success of deep learning, explaining why highly overparameterized networks generalize remains an open scientific question. This paradox is concentrated at the intersection of statistics (how much data is required?), optimization (why SGD finds good solutions in a non-convex regime?) and Information theory (what do networks keep or discard from data?). The key challenges that arise in

this context can be group as theoretical, experimental and industrial, which will be further discussed in Chapter 5.

Classical Perspective

Traditional learning theory is grounded in capacity control. The Vapnik-Chervonenkis(VC) dimension, Probably Approximately Correct (PAC) framework, and the bias-variance tradeoff, each implies that increasing complexity increases the risk of overfitting. According to these frameworks, there exists an optimal capacity which can be achieved by explicit regularization that minimizes test error.

1.2.2 State of the Art

Current work in deep learning theory

1. Optimization-driven generalization: There are several research in the field of SGD (Stochastic Gradient Descent), learning rate schedules, batch size, and noise implicitly push the model toward flatter and low complexity solutions.
2. Infinite or wide networks: Using theories like NTK (Neural Tangent Kernel) explains how very wide networks behave linearly and therefore are easier to analyze.
3. Data representation views: There have been arguments that real data actually lie on low dimensional manifolds. So, it has been suggested that “real complexity” is actually very low.

Around 2017, first systematic paper by Zhang et al., which was widely recognized, demonstrated that deep neural networks which are highly overparameterized and yet generalize well on real data. So, the contradiction that networks can memorize pure noise yet still they perform well on real data, indicating that data structure and optimization jointly influence generalization. It framed the puzzle, why do these models generalize at all? This paper clearly established the paradox of overparameterization and generalization in deep learning [1].

This behavior, now known as double descent, indicates that test error initially decreases and then increases near interpolation which is called overfitting but surprisingly it finally decreases again as models become extremely overparameterized [2]. This phenomenon contradicts classical intuition and generalization theorems.

Modern explanations have emerged, such as:

- Implicit regularization

- Flat minima hypothesis
- PAC-Bayes Generalization theory
- Neural Tangent Kernel (NTK) analysis
- Information Bottleneck theory

Each of these theories offers partial perspective on why large networks generalize.

1.3 CONTRIBUTION OF THIS THESIS

This thesis tries to form an Integrative framework by combining classical and modern perspectives into a coherent theoretical narrative. This thesis tries to connect theoretical findings to real over-parameterized systems (e.g., GPT, Gemini, Claude, Perplexity) to explain how these models exploit implicit regularization and outline theoretical and experimental challenges, proposing paths towards unified generalization theory.

1.4 THESIS STRUCTURE OVERVIEW

This thesis is divided into six chapters, each addressing an essential component of the research:

Chapter 1: Introduction — Provides an outline of the existence of the gap in the research and how the classical statistical method is learnt. This includes background, motivations, and explains why paradox of overparameterization challenges classical statistical learning theory and introduces the modern context in which deep neural networks succeed despite massive parameter counts and thus presents the basis for the study.

Chapter 2: Classical Statistical Generalization Theory — Reviews classical generalization theories like VC dimension, PAC learning, and the bias-variance trade-off and analyzes why they cannot capture the generalization bounds of neural networks.

Chapter 3: The Modern Generalized Paradox and Regularization — Explains the paradox that exists in theoretical analysis of generalization of overparameterized neural networks with implicit and explicit regularization frameworks and explains the modern theories that tries to explain this paradox.

Chapter 4: Over-Parameterization in Modern Foundation Models — This chapter acts as bridge in connecting over-parameterization theories to modern foundational models such as GPT and Gemini.

Chapter 5: Challenges and Open Problems— This chapter discusses open theoretical, experimental and industrial challenges and proposes some open research questions.

Chapter 6: Conclusion and Future Work — The thesis concludes with this final chapter which includes findings, contribution, insights and future work.

Together, these chapters build a comprehensive narrative that walks the reader from problem identification to the realization that overparameterization once seen as a curse helps in generalization. This structure ensures clarity, logical progression, and alignment with academic standards.

Classical Statistical Generalization Theory

The concepts presented below were viewed as the dominant explanations for generalization. The theories characterized learning as the process of minimizing empirical risk by controlling the model capacity to prevent overfitting [16].

2.1 EMPIRICAL RISK MINIMIZATION

Empirical risk minimization assumes that the hypothesis minimizing the empirical error will also minimize true error given that the hypothesis class \mathcal{H} is finite.

As described in section 1.1 of chapter 1 Introduction, how the learner minimizes the empirical risk, and the learning algorithm returns

$$h^* = \arg \min_{h \in \mathcal{H}} L_n(h).$$

generalization gap $|L(h^*) - L_n(h^*)|$ quantifies how well the model performs on unseen data.

2.2 PAC LEARNING

The PAC (Probably Approximately Correct) framework formalizes learning under uncertainty. A hypothesis class \mathcal{H} is PAC-learnable if, for any accuracy $\epsilon > 0$ and confidence $1 - \delta$, an algorithm can find $h \in \mathcal{H}$ such that $P(\text{error}(h) > \epsilon) < \delta$.

The sample complexity required for PAC learning is:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \log\left(\frac{|\mathcal{H}|/\delta}{\epsilon}\right) \right\rceil$$

PAC learning implies that the number of samples must scale with model complexity. Later, we figured out that the reason determining PAC learnability is not finiteness, but a combinatorial measure called VC dimension.

2.3 VC DIMENSION AND UNIFORM CONVERGENCE

The Vapnik-Chervonenkis (VC) dimension measures the expressive power or capacity of a hypothesis class \mathcal{H} . So, a hypothesis class \mathcal{H} has VC dimension d_{VC} if it can shatter any set of d_{VC} points [16].

Generalization (Uniform convergence) bound follow from Hoeffding's inequality:

$$P(\sup_{h \in \mathcal{H}} |L(h) - L_n(h)| > \epsilon) \leq 4 S_{\mathcal{H}}(2n) e^{-n\epsilon^2/8}.$$

Since $S_{\mathcal{H}}(2n) \leq (2n)^{d_{VC}}$

$$|L(h) - L_n(h)| \leq O\left(\sqrt{\frac{d_{VC} \log n}{n}}\right).$$

Thus, large d_{VC} implies weaker generalization.

2.4 BIAS-VARIANCE DECOMPOSITION

Bias and Variance are two competing sources of error. Bias measures how far off the model's predictions are from the true values on average because of wrong or overly simplistic assumptions in the learning algorithms whereas variance measures how much a model's predictions change if it is trained on different subsets of data.

Trade-off suggests maintaining the balance between two types of errors because high bias implies underfitting (model too simple) and high variance implies overfitting (model too complex).

$$E[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}.$$

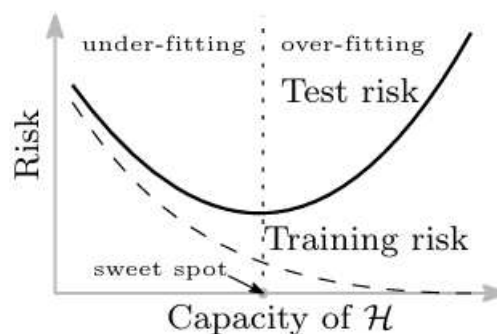


Figure 2.1: The classical U-shaped risk curve (Curves for training risk (dashed line) and test risk (solid line)) [2].

This leads to a U-Shaped test error curve [2], where optimal generalization occurs at an intermediate complexity which is called as “sweet spot”.

2.5 EXPLICIT REGULARIZATION

During training, classical learning adds explicit regularization to constrain model complexity:

$$\min_{h \in \mathcal{H}} L_n(h) + \lambda \Omega(h),$$

where $\Omega(h)$ quantifies model complexity and thus penalizes it and $\lambda > 0$ controls the trade-off between data fitting and regularization strength [17].

The most common techniques for explicit regularization are:

- Weight decay (L2 regularization): discourage large weights.
- L1 Regularization (Lasso): Inducing sparsity by putting some weights exactly zero.
- Dropout (Regularization by Noise): Randomly deactivates a fraction of neurons during training.
- Early stopping: Monitors Validation loss and stop training before overfitting.
- Data augmentation: Extends the training set with transformations like rotations, flips, etc.,

Thus, classical generalization theory predicted that too many parameters will lead to overfitting, yet empirical results prove otherwise as we will analyze in the next chapter.

The Modern Generalization Paradox and Regularization

3.1 THE PARADOX OF OVERPARAMETERIZATION

Deep neural networks often have more parameters than training samples ($p \gg n$) yet they generalize well. This observation stands in sharp contrast with conventional learning theory, which predicts that models should overfit for overparameterized models. However, modern neural networks consistently violate this rule. Empirically, increasing network width or depth frequently reduces test error after a critical threshold which is now explained with a phenomenon called double descent [2].

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
		no	no	99.34	10.61

Table 3.1: The training and test accuracy of various models on CIFAR 10 dataset [1]

Zhang et al. provided a foundational experiment into this paradox by demonstrating that deep neural networks are powerful enough to memorize in case of random labels. They performed

training on state-of-the-art CNN (Convolutional Neural Network) whose results are shown in Table 3.1. They replaced true labeling with pure noise and showed that models could achieve zero training error in these meaningless tasks where there is no real rule connecting input to output proving their capacity to memorize arbitrary data yet when trained on real data under the same conditions, these networks generalize remarkably well. Hence, concluding that regularization was not critical to prevent overfitting [1].

The finding by Zhang et al. exposes the central paradox of modern deep learning “How can neural networks with enough capacity to memorize noise also learn compact, generalizable representations of real data?”

3.2 DOUBLE DECENT BEHAVIOR

Belkin et al. demonstrated that test error does not follow a U-shaped curve but a double descent curve which has the following regimes to observe:

1. In under-parameterized regime, classical bias-variance trade-off holds.
2. Near the interpolation threshold, test error peaks due to the overfitting.

Previously, the experiments used to stop here, and some regularization techniques were taken into use for better generalization. But then it was observed that there exists third regime,

3. In the over-parameterized regime, test error drops again.

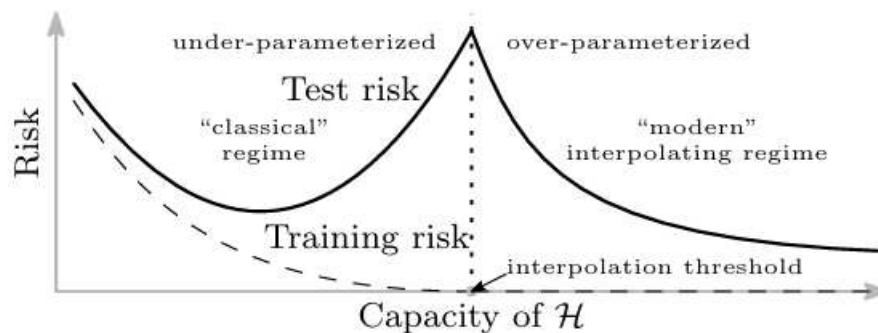


Figure 3.1: The double descent risk curve (Curves for training risk (dashed line) and test risk (solid line)) [2].

Thus, beyond point of overfitting, adding more parameters improves generalization rather than harming it [3]. This finding fundamentally reshaped learning theory.

3.3 IMPLICIT VERSUS EXPLICIT REGULARIZATION

While explicit regularization adds penalties or constraints, implicit regularization arises naturally from the optimization process. In deep learning, networks often generalize without explicit regularization techniques.

Implicit regularization

In contrast, this refers to unintentional biases that arise naturally from training dynamics, optimization algorithm, or architecture even in the absence of any explicit penalty term. In overparameterized neural networks, stochastic optimization methods such as stochastic gradient descent (SGD) play a crucial role in this phenomenon [4].

SGD does not minimize the training loss deterministically, instead, it performs noisy, approximate updates:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) + \xi_t$$

where η is the learning rate and ξ_t represents stochastic noise introduced by sampling mini batches. This noise implicitly drives the optimization trajectory toward flatter regions of the loss landscape which are empirically associated with better generalization.

Explicit regularization biases the solution towards lower norm, smoother or flatter parts of the huge solution space whereas implicit regularization determines the minimum norm solution from infinitely many weight vectors that fit the data exactly.

data aug	dropout	weight decay	top-1 train	top-5 train	top-1 test	top-5 test
ImageNet 1000 classes with the original labels						
yes	yes	yes	92.18	99.21	77.84	93.92
yes	no	no	92.33	99.17	72.95	90.43
no	no	yes	90.60	100.0	67.18 (72.57)	86.44 (91.31)
no	no	no	99.53	100.0	59.80 (63.16)	80.38 (84.49)
Alexnet (Krizhevsky et al., 2012)			-	-	-	83.6
ImageNet 1000 classes with random labels						
no	yes	yes	91.18	97.95	0.09	0.49
no	no	yes	87.81	96.15	0.12	0.50
no	no	no	95.20	99.14	0.11	0.56

Table 3.2: Accuracy of training and test sets with and without regularization [1].

So, as the performance of ImageNet dataset is shown in Table 3.2, the accuracy percentage clearly depicts that overparameterization gives “freedom” whereas implicit and explicit

regularization provides “discipline”. Empirically, deep networks generalize even without explicit regularization [1], implying that implicit regularization dominates in modern practice.

3.4 FLAT MINIMA HYPOTHESIS

This idea of flat minima hypothesis connects the geometry of the loss landscape to generalization. The theorem proposes that Neural networks that converge to flat minima tend to generalize better than those to converge to sharp minima [5].

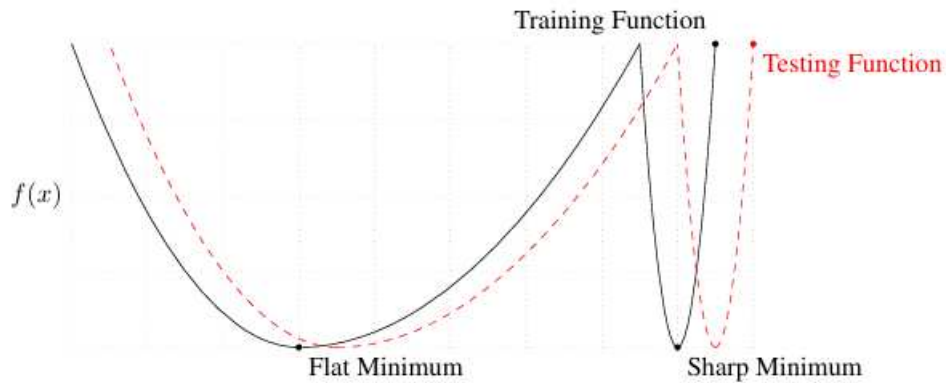


Figure 3.2: Conceptual sketch of Flat minima hypothesis [6].

The Formal idea constitutes that a flat minimum has mostly small eigenvalues of \mathcal{H} and the loss changes slowly around θ^* but a sharp minimum has some large eigen values and therefore small perturbations greatly increase loss.

We can formally define the flatness called as “Flatness measure”:

$$Flatness(\theta^*) = \left(\frac{1}{|B_\epsilon|} \right) \int_{\|\delta\| < \epsilon} L(\theta^* + \delta) d\delta$$

This flatness measures how sensitive the loss is to small changes in the parameters. Flat minima correspond to robust, stable solutions that generalize better [6].

3.5 MODERN THEORIES OF OVERPARAMETERIZATION

Neural Tangent Kernel (NTK)

Jacot et al. (2018) introduced the NTK framework, showing that in the infinite width limit, network training under gradient descent behaves like kernel regression. The dynamics become linear in function space which guarantees convergence.

For network output $f(x, \theta)$ and gradient descent of a neural network under NTK framework:

$$\frac{df(x, t)}{dt} = - \sum_{i=1}^n K(x, x_i) [f(x_i, t) - y_i],$$

where $K(x, x') = \nabla_{\theta} f(x, \theta)^{\top} \nabla_{\theta} f(x', \theta)$.

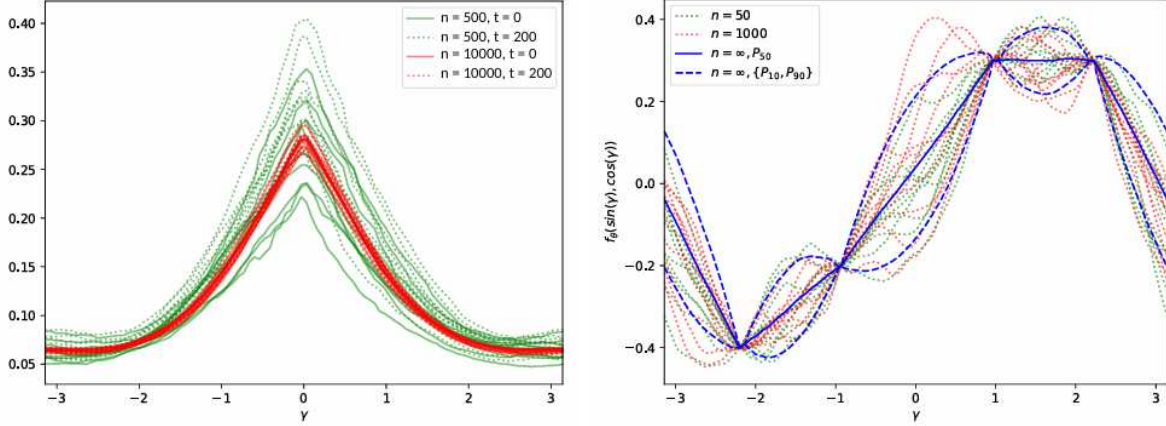


Figure 3.3: Convergence of Neural Tangent Kernel and network function across different varying widths [7].

From the empirical tests performed by Jacot et al. on fully connected ANNs, Figure 3.3 demonstrate the results, the left plot shows the convergence of the NTK to a fixed kernel as the training time and network width increases. While the right plot shows the convergence of the output function. This confirms that as the width tends to infinity, K becomes constant guaranteeing convergence and explaining the stability of very wide networks [7].

Information Bottleneck Theory

This theory views learning as a problem of balancing compression and prediction. So, conceptually, it is a trade-off between compression and prediction.

$$\mathcal{L}_{IB} = I(X; Z) - \beta I(Z; Y),$$

Where I denote mutual information and Z is the internal representation learned by the network. So, we want Z to be compressed but sufficient representation of X for predicting Y and β is the trade-off controlling compression versus accuracy. The hidden layers act as a bottleneck, forcing the network to compress the input information and retain only task-relevant features. This compression happens naturally due to stochastic gradient descent (SGD) noise. Therefore, this theory says that a good deep representation should compress the input while preserving only the information needed to predict the output [8].

Data Manifold Perspective

This perspective covers the fact that high-dimensional data often lie on low-dimensional manifolds. Overparameterized networks learn functions that are smooth along these manifolds, effectively regularizing against off-manifold noise. Optimization trajectories guided by SGD naturally align model gradients with manifold structure, further reducing effective complexity [9].

Lottery ticket hypothesis

This hypothesis is one of the most influential modern ideas about why large neural networks work so well. According to this hypothesis, training a large network is like buying many lottery tickets. Each random initialization of weights corresponds to a lottery ticket that can be trained independently to match full-network performance.

LTH (lottery ticket hypothesis) is performed in the following way:

1. Training the full network to converge on the dataset.
2. Pruning (removing) the lowest magnitude weights because they contribute the least.
3. Resetting the remaining weights to their original initialization values and retraining the smaller subnetwork which is the winning ticket.

This smaller subnetwork often matches or sometimes exceeds the accuracy of the original large model.

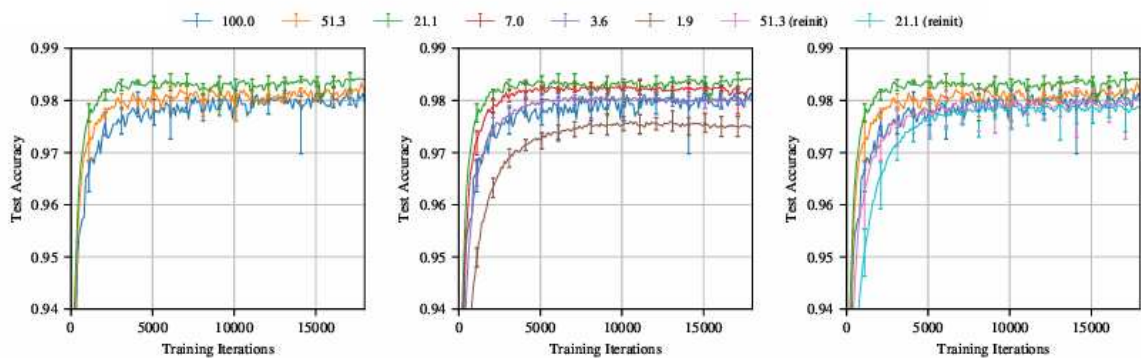


Figure 3.4: Test accuracy among different datasets with varying sparsity levels [10].

Experiments done by Frankle and Carbin showed validated the hypothesis. As observed, even highly sparse network containing only 20% of the parameter space can reach the same final accuracy as the dense model containing 100% of the parameter space. LTH shows that large networks act like a search space to discover these lucky subnetworks and thus explain efficiency, overparameterization, and the role of initialization in deep learning [10].

PAC-Bayes

This unifies the idea of PAC learning and Bayesian inference, giving probabilistic bounds that often work for large overparameterized models [11].

For prior $P(w)$ and posterior $Q(w)$:

$$\mathbb{E}_{w \sim Q}[R(w)] \leq \mathbb{E}_{w \sim Q}[R_n(w)] + \sqrt{\frac{KL(Q \parallel P) + \log \frac{2\sqrt{n}}{\delta}}{2(n-1)}}$$

$KL(Q \parallel P)$ is the Kullback-Leibler divergence between posterior and prior. It is a measure of how much the model deviated from the prior.

The PAC part guarantees that with high probability, its test error is close to its training error and Bayes part suggests that instead of analyzing a single model h , we analyze a distribution Q over models. Q being “very close” to the prior P , therefore small KL divergence term and therefore it leads to better generalization bound. If Q diverges far from P , the bound worsens. Therefore, PAC-Bayes gives non-vacuous bounds for deep networks.

Over-Parametrization in Modern Foundation Models

4.1 OVERVIEW OF LARGE-SCALE MODELS

The past decade has witnessed a dramatic uprising in the scale and complexity of neural network models, culminating in the development of large foundational models which contain extreme overparameterization. These models have transformed artificial intelligence from narrow task specific systems into general purpose learning platforms. The shift towards large scale overparameterized models was motivated by empirical discoveries in natural language processing and computer vision. Kaplan et al. (2020) showed that cross entropy which is the type of loss scales approximately as a power of model parameters, data set size, and compute budget [12]. This observation led to the construction of models such as GPT 3, Gemini, etc., However, their development undergoes continuous challenges in efficiency ethics and theoretical understanding. Building upon the scaling revolution, it becomes essential to quantify the degree of overparameterization.

4.2 DEGREE OF OVERPARAMETERIZATION

Understanding degree of overparameterization gives us an insight into why these industry-scale models remain stable and effective despite their enormous capacity [2].

Formally, if p denotes the number of trainable parameters and n is the number of independent training samples, the basic quantitative measure of parameter to sample ratio would be:

$$\rho = \frac{p}{n}$$

A model is considered under-parameterized if $\rho < 1$ and overparameterized if $\rho \gg 1$ and is considered approximately balanced when $\rho \sim 1$.

For models like GPT and Gemini, $p \gg n$, placing them in the overly parameterized regime.

Now, the raw ratio is not enough to show the true capacity of a model because many parameters are redundant or correlated through weight sharing, architectural symmetry, and the statistical constraints imposed by the optimizer [13]. So, the researchers came up with the new notion called intrinsic dimension d_{eff} which represents the minimal number of independent directions in the parameter space that meaningfully affect the network's output [13]. The corresponding effective overparameterization ratio, therefore, can be calculated as

$$\rho_{eff} = \frac{p}{d_{eff}}$$

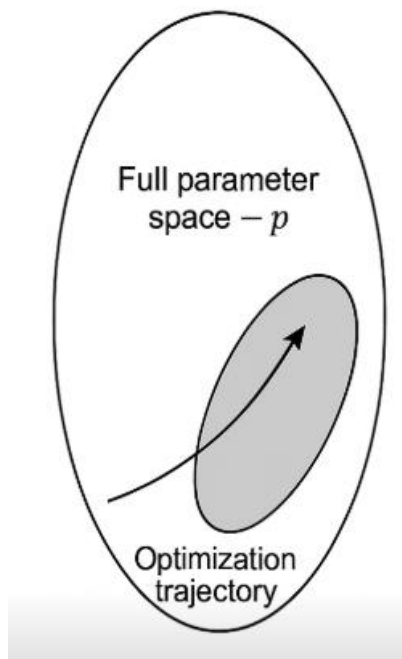


Figure 4.1: Conceptual explanation of optimization trajectory through parameter space during training

Figure 4.1 shows that training is essentially a search to find parameters that minimize the loss and the optimization trajectory does not explore the whole space but stays confined within a much smaller subspace d_{eff} . The Empirical analysis studies revealed that d_{eff} often constitutes only 1 to 5% of p in large-scale transformers [14]. Therefore, the optimization explores a very low dimensional subspace the total parameter manifold.

In summary the degree of overparameterization in modern deep neural networks cannot be measured solely by parameter count. It reflects a complex interaction between network geometry, optimization dynamics and intrinsic dimensionality.

4.3 WHY DO THEY GENERALIZE?

One major reason behind the generalization of large industrial models can be diversity of training data. Unlike earlier there are models trained on small domain specific datasets, modern foundational models are exposed to billions and trillions of tokens in different modalities from text to code to images and audio. So, this heterogeneity of training data forces the network to learn broad statistical irregularities rather than just memorizing specific examples Through continual exposure to new examples. Another reason for generalization in these models is fine tuning methods such as RLHF (Reinforcement learning from human feedback) [15] and constitutional alignment that preserves broad generalization while constraining undesirable outputs. Textural design of ChatGPT, Gemini, Claude Is transformer which encodes inductive biases that inherently promote generalization. Structure itself acts as a form of implicit regularization by seeing the model toward smooth compositional mappings rather than some arbitrary functions.

In other words, large models generalize because their immense redundancy allows them to approximate complex function through an ensemble of simple sub-functions producing globally stable behavior even when local details Change significantly.

Challenges and Open Problems

5.1 THEORETICAL CHALLENGES

Classical measures like VC. PAC has now become vacuous, modern theories like flat-minima and PAC-Bayes rely on approximations and NTK assumes infinite width.

In NTK, neural networks behave like linear models in fixed kernel space which enables formal proofs of convergence and generalization. Empirical studies show that finite width networks exhibit behaviors that deviate from NTK predictions. Therefore, infinite width analysis is helpful for gaining insight into global convergence, but they do not fully capture the dynamics observed in practice.

Additionally, there is no closed-form theory of optimization which limits analytical understanding and fails to represent the structure of deep learning optimization trajectories.

A major discovery in modern learning theory is Implicit regularization, the bias arises naturally from the training dynamics, optimization algorithm, or architecture but formalizing and quantifying this implicit bias remains an open problem. Current theoretical frameworks can only describe implicit bias in highly simplified models like linear regression but for deep nonlinear architectures, no mathematical formulations exist to quantify this bias.

Despite significant progress in understanding the generalization behavior of deep, overparameterized learning, the theoretical foundations remain incomplete. Existing modern theories explain fragments of generalization in deep networks. However, these frameworks are rarely compatible because they operate under different assumptions. A central theoretical limitation is the absence of a unified framework that integrates these perspectives into a single mathematical description of generalization in deep neural networks.

5.2 EXPERIMENTAL CHALLENGES

While theoretical frameworks struggle to capture the behavior of overparameterized deep networks, empirical investigation also faces numerous challenges.

1. Training state-of-the-art neural networks often containing billions or trillions of parameters requires powerful computational resources. Large networks require massive GPU clusters, and also measuring information terms is infeasible.
2. Deep learning experiments are sensitive to random seeds, initialization schemes, and data shuffling. This sensitivity complicates the replication of results and undermines confidence in empirical findings.
3. The internal mechanisms linking capacity, training dynamics, and learned structure remain largely hidden from experimental observations.

5.3 INDUSTRIAL CHALLENGES

The industrial adoption of overparameterized neural networks has accelerated the development of large-scale artificial intelligence systems such as GPT-4, Claude, etc. While these models demonstrate impressive performance, their deployment encounters numerous industrial challenges.

1. Apart from the computational costs, energy footprint of large-scale training has become an industrial and ethical issue. Companies now face public and regulatory scrutiny regarding the training of large models and concerns regarding environmental sustainability.
2. Another major industrial challenge concerns data governance, ownership, and copyright compliance. Overparameterized models require vast and diverse datasets drawn from internet which often contain copyrighted, private or sensitive materials. Companies like OpenAI and Google face backlash and ongoing legal debates about whether scraping web data for model training constitutes fair use or intellectual property infringement.
3. Large foundation models are not released in full due to their commercial values and thus this limits independent scientific verification and prevents academic researchers from reproducing empirical findings on overparameterization at an industrial scale.

5.4 OPEN RESEARCH QUESTIONS

The open research questions after a detailed study of this thesis are:

1. What is the true effective complexity of deep neural networks?

2. Can implicit regularization be controlled or enhanced deliberately for better generalization?
3. Is it possible to replace infinite-width approximation with finite-width to correctly represent realistic architectures?
Finally, the most important question of unification.
4. Can modern theories be integrated to construct a comprehensive generalization theory?

Conclusion and Future work

6.1 SUMMARY OF FINDINGS

This thesis sets out to understand why modern deep neural networks despite extreme overparameterization generally generalize effectively rather than going through the phenomenon called overfitting. These findings contradict the expectations of classical learning theories and through both theoretical and empirical examination, several key findings emerged which are as follows.

1. The thesis showed that the relationship between model capacity and generalization is fundamentally different in the overparameterized regime.
2. Secondly by demonstrating the experiments performed by Zhang et al. and Belkin et al., this thesis revealed that networks can memorize random labels, yet they generalize structured data showing that generalization does not depend only on explicit regularization.
3. Thirdly the findings highlight that overparameterization is not a curse but rather a feature that supports flexibility, stability, and representational richness.
4. Finally, the thesis identified persistent theoretical, experimental, and industrial challenges.

In summary, this thesis outlines that deep neural networks generalize not despite overparameterization but because of overparameterization.

6.2 CONTRIBUTION OF THIS THESIS

This thesis contributes to the ongoing discussions on the generalization in deep learning by bridging the gap between classical statistical learning theory with the modern phenomenon of overparameterization in neural networks. This thesis aims to provide a conceptual framework with theoretical theories and its implications on large networks that we use in our daily lives.

The Major contributions of this thesis are summarized as follows:

- **Integrative theoretical framework:** By combining concepts from classical statistical learning theories to modern theories, this study creates a coherent narrative linking early theoretical principles to present day deep learning behavior.
- **Theoretical mechanisms and Industrial systems:** A central contribution of this work is establishing a conceptual bridge between theoretical models and practical large-scale implementations.
- **Critical Examination:** Beyond theoretical synthesis, the thesis systematically analyzes the limitations and open challenges in studying overparameterization. This structured view highlights where future progress must be made to turn empirical successes into theoretical understanding.
- **Guidance for future directions:** Finally, this thesis proposes future directions for unifying a generalization theory instead of studying fragments of this phenomenon. Additionally, another direction to research can be theoretical analysis of approximate or strongly assumed initial conditions theories to make them fit into the conditions of real data and models.

This thesis contributes to the scientific study of why and how overparameterized neural networks generalize. This thesis reframes overparameterization as the mechanism for stable, implicit regularization instead of as a problem of excess capacity.

6.3 FUTURE RESEARCH DIRECTIONS

While this thesis consolidates current understanding of generalization in overparameterized neural network, but the field remains far from completeness of theoretical concepts. Future Research must bridge this gap by combining theoretical empirical and practical perspectives. Future research should aim to:

6.3.1 Toward a Unified Generalization Theory

As mentioned before, existing theories such as PAC-Bayes bounds, NTK analysis, flat minima hypothesis, etc. each capture partial truths but operate under incompatible assumptions. So, future research should aim to develop hybrid theoretical frameworks that integrate capacity control, implicit regularization, and optimization dynamics.

6.3.2 Quantitative Understanding and Formulation of Implicit Regularization

While implicit regularization has emerged as a possible explanation for generalization without explicit regularization, its quantitative characterization needs some investigation. Investigation

is required on how learning rate schedules, batch sizes, and optimizing noise influence implicit regularization strength and a closed form expression for implicit bias.

6.3.3 Combating the Computational Expense of Large-Scale Training

Empirical exploration of overparameterization is constrained by the computational expenses. So, Efficient simulation frameworks can be developed to emulate large model behavior at smaller scales using parameter sharing. This would strengthen the empirical foundation of deep learning theory.

6.3.4 Social and Ethical Considerations

This direction of research is emphasized upon understanding that generalization at industrial scale is not purely technical but also a societal responsibility. Future interdisciplinary research should explore fairness and bias, copyright and environmental sustainability.

6.4 FINAL REMARKS

This thesis has sought to bridge the contradiction that classical theory once taught us to fear excessive capacity because it will lead to overfitting, yet modern practice showed that scale itself has become the engine of stability and generalization. This work has illustrated that deep learning's success is not a violation of classical theory but an evolution of it. The main character in generalization is played by redundancy. In overparameterized networks, redundancy becomes robust, and scale becomes a path toward smoother, lower-complexity solutions that generalize more broadly than smaller models.

The analysis of industrial scale systems such as ChatGPT, Gemini, cloud, etc., further reinforces this point. However, this paradigm also carries responsibilities because this overparameterization also magnifies ethical, environmental, and interpretability concerns.

In closing, this thesis affirms that overparameterization is not a problem to be solved, but a principle to be understood. It invites future researchers to pursue deeper unification between theory and practice. A reminder that complexity, when guided by right dynamics, can give rise to simplicity and intelligence at the same time.

Appendix A

Supplementary Analysis: Lottery Ticket Hypothesis

A.1 OVERVIEW

This appendix provides extended experimental and analytical details related to the Lottery Ticket Hypothesis as described before in Chapter 3, Section 3.5.

The purpose of this experiment is to present the detailed experimental setup and pruning procedure and provide a deeper interpretation of the results shown in figure 3.4.

A.2 EXPERIMENTAL SETUP

Franklin and Carbin trained fully connected and convolutional networks on datasets such as Lenet architecture for MNIST and the Conv-2, Conv-4, and Conv-6 architectures of CIFAR-10.

The main steps to follow the experiment are:

1. Train a dense network to convergence using SGD (Stochastic Gradient Descent).
2. After convergence, prune (remove) a certain percentage of weights with the smallest magnitudes.
3. (a) Reset the remaining unpruned weights to their original initial value and retrain the subnetwork.
(b) Randomly reinitialize the pruned subnetwork and retrain.

A.3 OBSERVATIONS

- **Dense Network (No pruning)**

The blue curve in Figure 3.4 corresponds to the full, unpruned network. It serves as the performance baseline.

- **Moderately Sparse Network**

These subnetworks exhibit near same accuracy as the dense model despite having fewer parameters.

- **Highly Sparse Network**

Even at such high sparsity, some subnetworks still achieve competitive performance, although its converging slower.

- **Reinitialized Networks**

When the same sparse subnetwork is retrained from new random weights, performance drops which confirms that the original initialization is crucial.

Despite architectural differences in the dataset, all the three plots demonstrate the same trend, that is, sparse subnetworks trained from their original initialization (winning tickets) perform comparably to dense networks.

A.4 SUMMARY

So, The Lottery Ticket Hypothesis concludes that sparse subnetworks initialized from original weights achieve comparable performance to the full model but fail when initialized randomly. Overparameterization serves as a mechanism to discover well-initialized subnetwork, and this hypothesis validates why large networks often generalize despite massive capacity.

References

- [1] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*.
- [2] Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849-15854.
- [3] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, *2021*(12), 124003.
- [4] Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *Advances in neural information processing systems*, *30*.
- [5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
- [6] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv:1609.04836*.
- [7] Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, *31*.
- [8] Tishby, N., & Zaslavsky, N. (2015, April). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)* (pp. 1-5). Ieee.
- [9] Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, *29*(4), 983-1049.
- [10] Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv:1803.03635*.
- [11] Dziugaite, G. K., & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv:1703.11008*.
- [12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- [13] Li, C., Farkhoor, H., Liu, R., & Yosinski, J. (2018). Measuring the intrinsic dimension of objective landscapes. *arXiv:1804.08838*.

- [14] Aghajanyan, A., Gupta, S., & Zettlemoyer, L. (2021, August). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 7319-7328).
- [15] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- [16] Vapnik, V. (1998). *Statistical Learning Theory now plays a more active role: after the general analysis of learning processes, the research in the area of synthesis of optimal algorithms was started. These studies, however, do not belong to history yet. They are a subject of today's research activities* (Doctoral dissertation, These studies, however, do not belong to history yet. They are a subject of today's research activities).
- [17] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.
- [18] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Acknowledgement

I wish to express my sincere gratitude to Professor Luca Schenato, my supervisor, for his invaluable guidance, constant support, and insightful mentorship throughout the duration of this thesis. His patience in clearing all my doubts has been inspiring and I am deeply appreciative of the opportunity to learn under his supervision.

This thesis is dedicated to my beloved family - my mother, father, and brother. Although they were not physically present with me, their unwavering love and support made me feel as though they were always by my side. Their support has guided me through every challenge, and this accomplishment stands as a testament to their enduring presence in my life.

I would also like to extend my heartfelt appreciation to my best friend for being a constant source of support during difficult times and making me believe that I am no less.

I would like to thank all my friends in Padova and in India for their understanding, and companionship. Their support has been a constant source of motivation, and their presence has made this academic journey both meaningful and memorable.