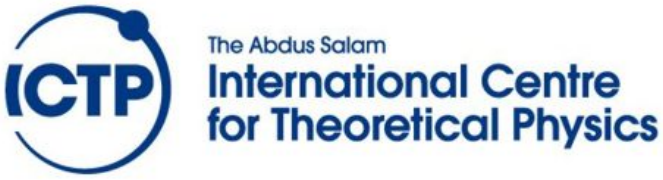
Optimal Generalization and Learning Transition in Finite-Width Bayesian Neural Networks

By

Rolland Mucunguzi Alinda



Quantitative Life Sciences Section

Diploma Project
Supervisor 1: Prof. Jean Barbier
Supervisor 2: Dr. Minh Toan Nguyen

August 2025

Abstract

This work investigates generalization, specialization, and learning transitions in shallow Bayesian neural networks using two-layer teacher-student neural network architectures known as committee machines. We emphasize finite-width effects that are often neglected in the thermodynamic limit, and we combine empirical training via stochastic gradient descent and Hamiltonian Monte Carlo along with analytical tools from statistical physics, mainly the replica method. Our primary focus is on the emergence of neuron alignment between teacher and student units as a function of the sample complexity parameter α. We derive the replica free entropy under the replica-symmetric ansatz, yielding a variational formula whose extremization yields us the fixed point equations for the overlap and conjugate overlap matrices. These equations are solved numerically for a finite hidden layer committee machine with 1 hidden units, and the resulting fixed points are used to estimate Bayes-optimal generalization error. The experiments we conducted using neural networks with 1 and 4 hidden units reveal the emergence of a specialization transition confirming theoretical predictions and consequently highlighting the role of network width, sample complexity, and the choice of an activation function. This work bridges the theoretical framework with numerical dynamics in finite-size network regimes, and thus provides us with clear insights into Bayesian optimal learning in finite-width Bayesian neural networks.

Acknowledgements

I convey my utmost gratitude to the Pre-eminent God, the Giver of all wisdom, knowledge, and understanding, for He always walks with me and blesses the works of my hands.

I am deeply grateful to The Abdus Salam International Centre for Theoretical Physics, for it offered me a fully funded study opportunity in Quantitative Life Sciences. In addition, sincere appreciation goes to all of the ICTP founders, funders, and administration, for they made this journey possible.

I extend my heartfelt appreciation to my project supervisors; Prof. Jean Barbier and Dr. Minh Toan Nguyen, their support and technical insights were vital in shaping the direction, flow, and completion of this work.

Lastly, special thanks to all the professors and lecturers in the Quantitative Life Sciences Section of the ICTP, for they did their very best to help us learn.

Dedication

"Hold a true friend with both hands." – African Proverb

This work is dedicated to my father, **Mr. Patrick Musambya Amooti**, whose guidance, support, and belief in the transformative power of education have guided my journey. Father, even when we are separated by enormous distances, I always sit down by the Adriatic Sea and you surely always speak to me.

To my mother, **Mrs. Grace Birungi Ateenyi**, I still have strong memories of all the sacrifices you made to see me through. For this love, I cannot comprehend, for it sits majestically beyond all my understanding. Thank you Ateenyi for teaching me how to love and smile amidst storms of life.

To my elder brother, **Don Kasaija**, from you I always learn calmness, from you I always draw hope that it is all possible. Thank you for always reminding me to give reverence to the Creator. Every time I call you, I get a reason to smile; thank you for always not giving up on me.

To the entire **ICTP family** including teaching staff, non-teaching, and fellow students, thank you for creating an environment of intellectual generosity and a loving community. You made this experience a well-worth chapter for my academic and personal growth.

Finally, I give thanks to the many unsung mentors and friends along the way whose words, belief, and support left a lasting impact. This work surely stands on a foundation of love, learning, and gratitude.

Author's declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the ICTP's Regulations and Code of Practice for Research and that it has never been submitted anywhere for an academic award. Except where indicated by specific references in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others is indicated as such. Any views expressed in this thesis are my own.

SIGN:

DATE: 10th August, 2025.

Contents of Thesis

List of Figures

Figure **Page**

Chapter 1

Introduction

1.1 Context and background

Deep neural networks (DNNs) have become fundamental tools in our study and development of modern machine learning and artificial intelligence systems. They have achieved impressive results across a variety of tasks such as image recognition, natural language processing, and in areas of decision-making like health and finance. However, despite their practical success, more needs to be done, for example, we still need to create a rigorous theoretical understanding of how neural networks generalize from data and learn meaningful internal representations of given data.

Empirical results show that deep neural networks can memorize random labels yet generalize well on real data. This challenges the adequacy of classical learning theory such as uniform convergence bounds in explaining generalization in highly overparameterized regimes and has in turn prompted appreciable theoretical research aiming to understand DNNs from new perspectives among which include statistical mechanics and information theory approaches. According to Zhang et al.(2016), traditional approaches fail to explain why large neural networks generalize well in practice, this was observed with convolutional neural networks that fit random labels to training data unaffected by regularization, yet, they generalize well [1].

One of the powerful paradigms that has emerged for studying the learning dynamics and generalization capacity of DNNs is the teacher-student framework. In this setting, a "teacher" network generates labeled data according to a known rule or probability distribution and a "student" network attempts to learn the teacher's mapping either

when given the data alone or with some knowledge of the teacher weights, say knowledge of the distribution. An illustration of this setup is given in Figure 1.1. This controlled setup allows us to isolate key phenomena such as alignment, specialization, and phase transitions during the learning process. The teacher-student approach was originally developed in the context of perceptrons and committee machines in statistical physics [2], but has since been extended to more complex models and algorithms including modern Bayesian neural networks [3, 4].

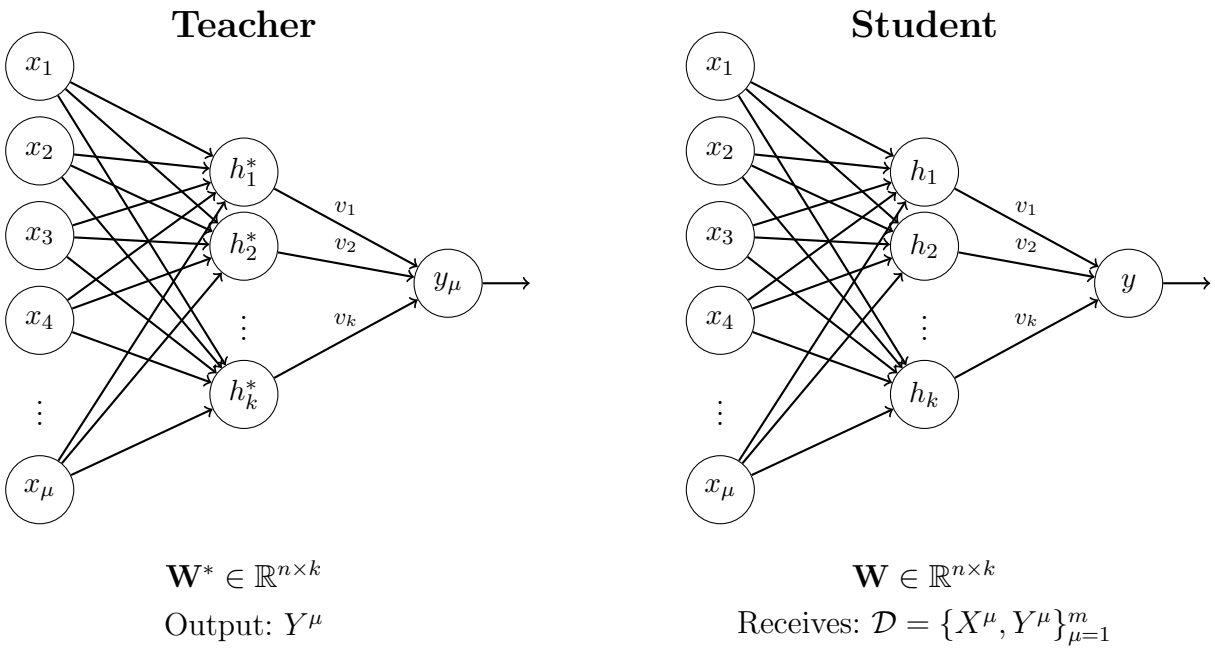


Figure 1.1: Illustration of the teacher-student neural network setup. The teacher generates labels using weights $\mathbf{W}^* \in \mathbb{R}^{n\times k}$, while the student learns from input-label pairs, $\{X^\mu, Y^\mu\}_{\mu=1}^m$ using weights $\mathbf{W}$. The hidden units are connected to the output through readout weights $\{v_i\}_{i=1}^k$.

A key insight from working with the teacher-student setup is that learning exhibits distinct phases depending on sample complexity $\alpha = \frac{m}{n}$, with m as the number of training samples and n the input dimension. Below a critical threshold usually denoted as α_{spec}, student units remain symmetric and do not align with specific teacher units. Above this threshold, specialization occurs where each student unit begins to align with a specific teacher neuron and we see generalization improving sharply [5]. This phase

transition has been characterized analytically in the limit of infinite width using tools from statistical mechanics such as the replica method and state evolution equations [6].

However, much of the existing theory assumes idealized conditions of infinite width of the hidden layer, Gaussian priors, and training through posterior sampling in these regimes [7]. These assumptions often fail to capture the practical behavior of finite-width networks. In this work, we revisit the teacher-student framework in the context of finite-width neural networks, and we investigate how generalization, specialization, and learning transition show up in these networks.

In this work, we combine the theoretical framework with empirical experimentation in order to show the conditions under which neuron alignment and specialization arise, and to quantify the accuracy of Bayesian predictions for finite-width neural networks. We employ a combination of tools that include: stochastic gradient descent, Hamiltonian Monte Carlo (HMC) sampling, and analytical derivations mainly based on the replica method.

1.2 Motivation and significance

While DNNs have demonstrated remarkable performance across various domains, their learning dynamics and generalization capabilities remain quite elusive. The apparent paradox between their high expressivity and their ability to generalize well even when trained with more parameters than data challenges conventional wisdom from classical statistical learning theory [8].

A fruitful approach to addressing this gap has been the study of simplified neural architectures under controlled conditions like the teacher-student setup. These models avoid extraneous complexity but preserve essential features of learning which makes it possible to develop theoretical tools and derive meaningful insights. To this end, we use a committee machine which is a shallow two-layer neural network with fixed output weights and a few hidden units to serves as a canonical model.

By studying alignment, generalization, and learning transitions in finite-width teacher-student networks we would like to evaluate the relevance of theoretical predictions in practical scenarios. We explore how alignment evolves during training, how generalization error depends on the sample complexity, α, and to what extent the predictions from

replica theory and fixed point equations hold.

This work provides a quantitative understanding of specialization and generalization in finite-width committee machines, and brings to light the role of different aspects of a model in finite-width regimes. Furthermore, this research contributes to the broader goal of developing a statistical physics framework of learning and generalization that remains informative even outside the idealized infinite-width setting.

1.3 Research questions

We aim at investigating the mechanisms underlying alignment and generalization in finite-width, shallow, Bayesian neural networks using the teacher-student paradigm. Our central goal is to understand how macroscopic learning phenomena such as specialization and generalization error evolve in finite networks and how they relate to the predictions of replica theory and fixed point equations.

To this end, the following research questions guided our investigations:

1. **How does specialization emerge in finite-width teacher-student networks?** Under this, another question we want to explore is: What is the critical sample complexity threshold at which student units begin to specialize and align with distinct teacher units?

2. **How accurate are theoretical predictions from the replica method in finite-width settings?**
   We would like to investigate whether fixed-point solutions from the replica symmetric potential accurately describe the overlap dynamics and generalization performance observed in experiments conducted using HMC and SGD.

1.4 Contributions

This thesis advances the theory and practice of learning dynamics in finite-width Bayesian neural networks. We derived a general replica-symmetric free-entropy variational formula and implemented the solution for Bayes optimal generalization error for $k = 1$ hidden unit using hyperbolic tangent as the activation function. We designed HMC experiments that do test the replica predictions by estimating the minimum mean squared error

(MMSE) from posterior samples, and we found out that, for our setting, the empirical MMSE closely tracks the theoretical curve across chosen α values.

1.5 Overview of the Thesis Structure

This work is organized across six chapters:
Chapter 1 introduces the work giving a context and background, a motivation for the research, the research questions we seek to answer, and states our contribution. **Chapter 2** reviews relevant literature on the teacher-student framework using the committee machine model, the replica theory, and related theoretical and empirical studies. **Chapter 3** presents the theoretical derivation of the replica free entropy under the replica-symmetric ansatz and of the resulting fixed point equations. **Chapter 4** presents the experimental methodologies we employed, the teacher and the student model, how we measured alignments, Bayes optimal benchmarking, and a description of the experiments we performed. **Chapter 5** presents the results from SGD and HMC experiments, and also results obtained by comparing theory with numerical approximations of MMSE. **Chapter 6** concludes with a summary of the findings and gives promising directions for future research.

Chapter 2

Literature review

2.1 Artificial Neural Networks and Generalization

As introduced in Section 1.1, deep neural networks have exhibited powerful generalization abilities across different application domains including computer vision, natural language processing, automation, and scientific modeling. However, the theoretical principles underlying the success of deep neural networks remain only partially understood. A famous central question that persists is: how do heavily overparameterized neural network models achieve low generalization error despite the fact that they fit data along with noise?

For a pedestal upon which firm understanding can spring, simplified frameworks have been proposed to study alignment, generalization, and learning transition under controlled conditions. Among these is the teacher-student framework that has emerged as a foundational pedestal [9]. In this setup, a teacher neural network with fixed parameters generates labeled data, and a student neural network is trained on this data with an aim of recovering the underlying rule that has generated the data. The simplicity of this arrangement makes it ideal for isolating phenomena we introduced earlier such as: generalization, specialization, learning transition, effect of sample size, effect of the choice of the nonlinearity, among others.

While theoretical insights have been primarily derived in the infinite-width regimes commonly known as the thermodynamic limit in statistical physics [10], there is growing recognition that these assumptions often fail to reflect common and simple real-world architectures. In particular, finite-width networks exhibit behaviors such as delayed

specialization, convergence sensitivity to initialization, and others. These behaviors motivate deeper investigation into finite-size effects and what they imply for learning. An empirical study of finiteness versus infiniteness in neural networks is given by the work of Lee et al. [11].

Committee machines that provide us a useful model system for exploring phenomena observed in neural networks in a controlled manner are two-layer neural networks with k hidden units and fixed or varying readout weights. Their simple structure makes them analytically tractable, while still capturing the key features during the course of learning. In the limit of large m and n with finite sample complexity α, theoretical predictions for generalization curves, specialization thresholds, and overlap dynamics can be derived exactly [12, 13].

2.1.1 Teacher-Student Frameworks: Foundations and Evolution

The teacher-student paradigm has deep roots in statistical physics and probabilistic modeling. Using tools like the replica method and cavity method, early studies demonstrated that high-dimensional inference problems often exhibit phase transitions in their learning curves [14–16].

Key insights from this literature include:

- **Phase Transitions:** Learning in neural networks can exhibit abrupt changes in behavior as the sample complexity parameter $\alpha = m/n$ is varied with m as the sample size, and n the input dimension. Below a given critical threshold α_{spec}, learning remains unstructured and the student units remain symmetric to one another. Above this critical threshold, student units begin to specialize and align with specific neurons of the teacher [17].
- **Symmetry Breaking:** The specialization of student neurons we are talking about corresponds to a spontaneous symmetry breaking among student units driven by factors like data abundance and inductive biases in the model. This phenomenon has been extensively characterized using tools from disordered systems [2].
- **Bayesian Optimality:** When the prior over the student weights matches that of the teacher and learning is guided via exact Bayesian inference, the student

achieves Bayes-optimal generalization [4]. The notion of Bayesian optimality defines a performance benchmark for both theory and algorithms.

Recent work by Barbier et al. extended these ideas to finite-width committee machines using adaptive interpolation and rigorous statistical mechanics. They showed that a gap does exist between what is statistically possible and what is computationally tractable, particularly when using gradient-based methods. In these cases, although the Bayes-optimal solution is theoretically accessible, practical algorithms such as approximate message passing (AMP) may fail to reach it [3].

2.1.2 Learning Dynamics

The evolution of learning in teacher-student framework is often analyzed through the lens of overlap matrices that measure the correlation between student and teacher weights. The alignment between teacher and student neurons is captured by the overlap matrix $S = W_{\text{teacher}} W_{\text{student}}^T / n$. To account for permutation symmetries due to the non-identifiability of hidden unit ordering, techniques such as Hungarian algorithm can be applied to optimally match student units to teacher units.

The activation function plays a critical role in learning dynamics, polynomial activations tend to promote faster specialization, while piecewise-linear functions such as ReLU may lead to slower transitions or persistence of symmetry [12]. Generalization is also modulated by sample complexity that we denote as α and noise level in the training data that we denote as Δ. At low α, models tend to underfit or else memorize random patterns, while higher α facilitates meaningful alignment and generalization [18].

One particularly intriguing and now well-documented phenomenon in modern machine learning is the double descent behavior of the test error as a function of model size or sample complexity. Unlike the classical U-shaped bias-variance tradeoff, test error in overparameterized models initially decreases with increasing model capacity or data, then undergoes a sharp increase near the interpolation threshold before decreasing once more in the highly overparameterized regime [19], [8]. This behavior has been observed in a variety of settings including those that involve linear models and deep neural networks, and is now understood to emerge even in simplified teacher-student models.

Finally, a key theme in recent literature is the computational versus statistical gaps. While Bayes-optimal performance is theoretically achievable, practical optimization via

stochastic gradient descent (SGD) often falls short. SGD can become trapped in poor local minima or fail to exploit weak alignment signals. Sampling-based methods such as HMC have been shown to recover the correct alignment structure more consistently, though this is often realized at greater computational costs. This highlights the importance of considering algorithmic limitations when evaluating generalization performance [3].

In summary, the literature provides a rich and evolving framework for understanding generalization, specialization, and learning transition in neural networks. However, significant gaps remain in translating asymptotic results to finite-size models along with realistic training conditions. This thesis contributes to addressing this gap by empirically and theoretically investigating learning transitions in shallow networks of finite width.

Chapter 3

The Replica Theory

In this chapter, we present a heuristic derivation of the replica formula using the replica method which is a non-rigorous but powerful technique employed to study disordered systems. Our objective is to derive a closed-form expression for the free entropy of a shallow Bayesian neural network based on the committee machine architecture. In doing this, we also obtain the corresponding saddle-point equations for the relevant order parameters that characterize the system in the large n and m limits. Detailed studies of this approach can also be accessed in the work of Mezard [20] and Barbier [3].

3.1 The Replica calculation

In our analysis of optimal generalization using the teacher-student setup, given inputs $X \in R^n$ and W_0 as the teacher weight matrix, the teacher generates labels using the model:

$$Y = \frac{1}{\sqrt{k}} v^\top \sigma\left(\frac{W_0 X}{\sqrt{n}}\right) + \sqrt{\Delta} Z, \tag{3.1}$$

with $v \in \mathbb{R}^k$ is a fixed normalized vector, $Z \sim \mathcal{N}(0, I_m)$, and σ is an activation function. The teacher and the student network have the same architecture and we would like the student to learn the weights the teacher uses to generate the labels.

The central object of study in the analysis of optimal generalization and Bayesian learning for the setting we have introduced is the posterior distribution over the network weights. Given a dataset $\{(X_\mu, Y_\mu)\}_{\mu=1}^m$, the posterior distribution of the student weights $\{w_{il}\}_{i,l=1}^{n,k}$

is given by

$$P\left(\{w_{il}\}_{i,l=1}^{n,k} \,\middle|\, \{X_{\mu i}, Y_\mu\}_{\mu,i=1}^{m,n}\right) = \frac{1}{\mathcal{Z}_n} \prod_{i=1}^{n} P_0\left(\{w_{il}\}_{l=1}^{k}\right) \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \,\middle|\, \left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{\mu i} w_{il}\right\}_{l=1}^{k}\right), \tag{3.2}$$

where P_0 denotes the prior over the weights, P_{out} is the likelihood, and $\mathcal{Z}_n$ is the normalization constant usually called the *partition function* obtained by integrating the numerator over all possible weight configurations:

$$\mathcal{Z}_n = \int \prod_{i=1}^{n} P_0\left(\{w_{il}\}_{l=1}^{k}\right) \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \,\middle|\, \left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{\mu i} w_{il}\right\}_{l=1}^{k}\right) dw. \tag{3.3}$$

From $\mathcal{Z}_n$, we get a quantity of interest, f_n from which we can compute the *free entropy*, f_n is defined as the normalized expected log-partition function:

$$f_n \equiv \frac{1}{n} \mathbb{E} \ln \mathcal{Z}_n. \tag{3.4}$$

The replica method yields a conjectural but explicit expression for f_n in the high-dimensional limit $n, m \to \infty$ with their ratio $\alpha = m/n$ held fixed. This expression provides insight into the typical learning and generalization properties of a Bayesian student in the thermodynamic limit.

The replica method relies on the so-called *replica trick*, which circumvents the direct computation of the expectation of a logarithm by instead computing the logarithm of the expectation of replicated systems. Given a random variable $x \in \mathbb{R}^n$ and a strictly positive function $f_n : \mathbb{R}^n \to \mathbb{R}$ that depends on n, we have the following identity:

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} \ln f_n = \lim_{p\to 0^+} \lim_{n\to\infty} \frac{1}{np} \ln \mathbb{E} f_n^p. \tag{3.5}$$

This formulation introduces a subtle but non-rigorous step: the exchange of limits in n and p. The core idea is to compute the moments $\mathbb{E} f_n^p$ for integer values of $p \in \mathbb{Z}_{>0}$, and then, we analytically continue the resulting expression to non-integer p, allowing us to evaluate the limit as $p \to 0^+$.

To compute the free entropy of the system defined as $f \equiv \lim_{n\to\infty} f_n$, we begin by evaluating the moments of the replicated partition function for integer $p \in \mathbb{Z}_{>0}$. That is,

we consider:

$$\mathbb{E}\mathcal{Z}_n^p = \mathbb{E}\left(\int_{\mathbb{R}^n\times\mathbb{R}^k} dw \prod_{i=1}^n P_0\left(\{w_{il}\}_{l=1}^k\right)\prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu\,\middle|\,\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}\right\}_{l=1}^k\right)\right)^p, \quad (3.6)$$

$$= \mathbb{E}\left(\prod_{a=1}^p\int_{\mathbb{R}^n\times\mathbb{R}^k} dw^a \prod_{i=1}^n P_0\left(\{w_{il}^a\}_{l=1}^k\right)\prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu\,\middle|\,\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}^a\right\}_{l=1}^k\right)\right), \quad (3.7)$$

where $\{w^a\}_{a=1}^p$ denotes the collection of replicated weight configurations across the p replicas and the replication introduces p independent copies of the system.

We perform the outer averaging over all possible instances of $X_{\mu i}\sim\mathcal{N}(0,1)$, w^* and Y. We can denote the teacher weights w^* as w^0 for a replica zero. We introduce a replica of index $a=0$ for the teacher replica and the other $a=1,\ldots,p$ replicas are for the student:

$$\mathbb{E}\mathcal{Z}_n^p = \mathbb{E}_X\int_{\mathbb{R}^m} dY\prod_{a=0}^p\left(\int_{\mathbb{R}^n\times\mathbb{R}^k} dw^a\prod_{i=1}^n P_0\left(\{w_{il}^a\}_{l=1}^k\right)\prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu\middle|\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}^a\right\}_{l=1}^k\right)\right). \quad (3.8)$$

The pre-activations $Z_{\mu l}^a = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}^a$ are jointly Gaussian with zero mean due to the Gaussianity of X. Their covariance takes the form:

$$\mathbb{E}Z_{\mu l}^a Z_{\nu l'}^b = \delta_{\mu\nu}\,Q_{bl'}^{al}, \quad \text{with} \quad Q_{bl'}^{al} = \frac{1}{n}\sum_{i=1}^n w_{il}^a w_{il'}^b. \quad (3.9)$$

For each replica pair (a,b), we define the overlap matrix:

$$Q_b^a \equiv \left(Q_{bl'}^{al}\right)_{1\le l,l'\le k}\in\mathbb{R}^{k\times k},$$

which encodes the correlations between the hidden units of replica a and replica b.

To fix the values of the overlap matrices Q, we introduce Dirac delta functions that enforce their definitions. Before this, we first integrate over all their possible instances:

$$\mathbb{E}Z_n^p = \prod_{(a,l)}\int_{\mathbb{R}}\mathrm{d}Q_{al}^{al}\prod_{\{(a,l);(b,l')\}}\int_{\mathbb{R}}\mathrm{d}Q_{bl'}^{al}\left(I_{\text{prior(P)}}\left(\{Q_{bl'}^{al}\}\right)\times I_{\text{channel(C)}}\left(\{Q_{bl'}^{al}\}\right)\right) \quad (3.10)$$

We recognize that each $Z_\mu \in \mathbb{R}^{(p+1)k}$ which is the collection of pre-activation values for a given data sample μ is i.i.d. across μ and jointly a multivariate Gaussian with covariance matrix Σ, the channel contribution takes the form:

$$I_{\mathrm{C}} = \int_{\mathbb{R}^m} dY \prod_{\mu=1}^{m} \left[\int_{\mathbb{R}^{(p+1)k}} dZ_\mu \prod_{a=0}^{p} P_{\mathrm{out}}(Y_\mu \mid Z_\mu^a) \cdot \mathcal{N}(Z_\mu; 0, \Sigma) \right], \tag{3.11}$$

where the multivariate Gaussian density is given by:

$$\mathcal{N}(Z; 0, \Sigma) = \frac{1}{\sqrt{(2\pi)^{(p+1)k} \det \Sigma}} \exp\left(-\frac{1}{2} Z^\top \Sigma^{-1} Z \right). \tag{3.12}$$

Taking into consideration the integration over all $\mu = 1, \ldots, m$, we obtain:

$$\begin{aligned} I_{\mathrm{C}} = \int_{\mathbb{R}^m} dY \prod_{a=0}^{p} \int_{\mathbb{R}^{m\times k}} dZ^a \prod_{a=0}^{p} P_{\mathrm{out}}(Y \mid Z^a) \exp\left(-\frac{m}{2} \ln \det \Sigma - \frac{mk(p+1)}{2} \ln 2\pi \right) \\ \times \exp\left(-\frac{1}{2} \sum_{\mu=1}^{m} \sum_{a,b} \sum_{l,l'} Z_{\mu l}^a Z_{\mu l'}^b \left(\Sigma^{-1}\right)_{\substack{al \\ bl'}} \right). \end{aligned} \tag{3.13}$$

Employing the Dirac delta function, the I_{P} equation is written as:

$$I_{\mathrm{P}}(\{Q_{bl'}^{al}\}) = \prod_{a=0}^{p} \left(\int_{\mathbb{R}^{n\times k}} dw^a P_0(w^a) \right) \left(\prod_{\{(a,l);(b,l')\}} \delta\left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) \right). \tag{3.14}$$

Replacing the Dirac delta function by its Fourier representation, we yield:

$$\delta\left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) = \int \frac{d\hat{Q}_{bl'}^{al}}{2\pi} \exp\left[\hat{Q}_{bl'}^{al} \left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) \right]. \tag{3.15}$$

Inserting this representation into the I_{P} equation and then collecting the exponential terms, we obtain an expression of the form:

$$I_{\mathrm{P}} = \int d\hat{Q} \exp\left[\sum_{a,b,l,l'} \hat{Q}_{bl'}^{al} Q_{bl'}^{al} \right] \prod_{i=1}^{n} \int dw_i \, P_0(w_i) \exp\left[- \sum_{a,b,l,l'} \hat{Q}_{bl'}^{al} w_{il}^a w_{il'}^b \right]. \tag{3.16}$$

We notice that the integrand is now factorized over i, thus we can apply the saddle-point method. As the integrand is raised to the n-th power due to the product over i, we can thus write:

$$I_{\mathrm{P}} \propto \int d\hat{Q} \exp\left[n \cdot \Phi(Q, \hat{Q}) \right], \tag{3.17}$$

where the exponent is given by:

$$\Phi(Q,\hat{Q}) = \sum_{a,b,l,l'} \hat{Q}^{al}_{bl'} Q^{al}_{bl'} + \ln \int dw\, P_0(w) \exp\left(- \sum_{a,b,l,l'} \hat{Q}^{al}_{bl'} w^a_l w^b_{l'} \right). \tag{3.18}$$

In the large n-limit, the integral is dominated by the saddle point, leading to:

$$\lim_{n\to\infty} \frac{1}{n} \ln I_{\mathrm{P}} = \mathrm{ext}_{\hat{Q}}\, \Phi(Q,\hat{Q}). \tag{3.19}$$

A similar saddle-point approach applies to the channel contribution I_{C}, which also depends on the overlaps $Q^{al}_{bl'}$. Combining both the prior and channel contributions, the replica computation yields the final variational expression as:

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}[Z_n^p] = \mathrm{ext}_{Q,\hat{Q}}\, H(Q,\hat{Q}), \tag{3.20}$$

where $H(Q,\hat{Q})$ is the replica free entropy functional and it takes the form:

$$H(Q,\hat{Q}) \equiv \frac{1}{2} \sum_{a=0}^{p} \sum_{l,l'} Q^{al}_{al} \hat{Q}^{al}_{al} - \frac{1}{2} \sum_{a\neq b} \sum_{l,l'} Q^{al}_{bl'} \hat{Q}^{al}_{bl'} + \ln I + \alpha \ln J, \tag{3.21}$$

Where the parameter $\alpha = \lim_{n\to\infty} \frac{m}{n}$ and the I and J terms are defined as:

$$I \equiv \prod_{a=0}^{p} \int_{\mathbb{R}^k} dw^a P_0(w^a) \exp\left(-\frac{1}{2} \sum_{a=0}^{p} \sum_{l,l'} \hat{Q}^{al}_{al'} w^a_l w^a_{l'} + \frac{1}{2} \sum_{a\neq b} \sum_{l,l'} \hat{Q}^{al}_{bl'} w^a_l w^b_{l'} \right), \tag{3.22}$$

$$J \equiv \int_{\mathbb{R}} dy \prod_{a=0}^{p} \int_{\mathbb{R}^k} \frac{dZ^a}{(2\pi)^{k(p+1)/2}} \frac{P_{\mathrm{out}}(y|Z^a)}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} \sum_{a,b=0}^{p} \sum_{l,l'=1}^{k} Z^a_l Z^b_{l'} (\Sigma^{-1})_{\substack{al \\ bl'}} \right). \tag{3.23}$$

To make the extremization tractable, we assume a *replica-symmetric* (RS) structure on the overlaps. This leads to the RS form of the free entropy functional:

$$H(Q^0,\hat{Q}^0,q,\hat{q}) = \frac{p+1}{2} \mathrm{Tr}\left[Q^0\hat{Q}^0\right] - \frac{p(p+1)}{2} \mathrm{Tr}\,(q\hat{q}) + \ln I + \alpha \ln J. \tag{3.24}$$

To proceed with the replica trick we introduced earlier, it remains to compute explicit expressions for the terms I and J. For any symmetric positive-definite matrix $M \in \mathbb{R}^{k\times k}$ and any vector $x \in \mathbb{R}^k$, we have and an identity:

$$\exp\left(\frac{1}{2} x^\intercal M x\right) = \int_{\mathbb{R}^k} \mathcal{D}\xi\, \exp\left(\xi^\intercal M^{1/2} x\right), \tag{3.25}$$

where $\mathcal{D}\xi$ denotes a standard Gaussian measure on $\mathbb{R}^k$, $\mathcal{D}\xi = \frac{d\xi}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\|\xi\|^2\right)$.

Using the identity in Equation (3.25) , under the replica symmetric assumption we have:

$$I = \int_{\mathbb{R}^K} \mathcal{D}\xi \left(\int_{\mathbb{R}^K} dw\, P_0(w)\, \exp\left(-\frac{1}{2} w^\intercal (\hat{Q}^0 + \hat{q}) w + \xi^\intercal \hat{q}^{1/2} w \right) \right)^{p+1}, \tag{3.26}$$

$$J = \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi \left(\int_{\mathbb{R}^K} dZ\, P_{\text{out}}\left(y \,\middle|\, (Q^0 - q)^{1/2} Z + q^{1/2} \xi\right) \right)^{p+1}. \tag{3.27}$$

To ensure correct normalization $\mathbb{E}[\mathcal{Z}_n^0] = 1$, we impose:

$$\text{extr}_{Q,\hat{Q}} \left[\lim_{p \to 0^+} H(Q, \hat{Q}) \right] = 0.$$

In the $p \to 0^+$ limit, the integrals simplify: $J = 1$, and

$$I = \int dw\, P_0(w) \exp\left(-\tfrac{1}{2} w^\top \hat{Q}^0 w \right).$$

Enforcing $I = 1$ then yields $\hat{Q}^0 = 0$, and the diagonal overlap becomes

$$Q^0_{ll'} = \mathbb{E}_{P_0}[w_l w_{l'}],$$

i.e., the second moment matrix of the prior.

Substituting this into the expression for the replica free entropy and taking the limit $p \to 0^+$, we obtain the final variational formula as:

$$\lim_{n \to \infty} f_n = \text{extr}_{q,\hat{q}} \left\{ -\frac{1}{2} \text{Tr}[q\hat{q}] + I_P + \alpha I_C \right\}. \tag{3.28}$$

If we separate one replica corresponding to the teacher with weights w^0 from the others, I_P becomes:

$$\begin{aligned} I_P \equiv & \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} dw^0\, P_0(w^0)\, \exp\left(-\frac{1}{2} (w^0)^\intercal \hat{q} w^0 + \xi^\intercal \hat{q}^{1/2} w^0 \right) \\ & \times \ln \left(\int_{\mathbb{R}^k} dw\, P_0(w) \exp\left(-\frac{1}{2} w^\intercal \hat{q} w + \xi^\intercal \hat{q}^{1/2} w \right) \right). \end{aligned} \tag{3.29}$$

The I_C is given by:

$$\begin{aligned} I_C \equiv & \int_{\mathbb{R}} dy \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} \mathcal{D}Z^0\, P_{\text{out}}\left(y \mid (Q^0 - q)^{1/2} Z^0 + q^{1/2} \xi\right) \\ & \times \ln \left(\int_{\mathbb{R}^k} \mathcal{D}Z\, P_{\text{out}}\left(y \mid (Q^0 - q)^{1/2} Z + q^{1/2} \xi\right) \right). \end{aligned} \tag{3.30}$$

3.2 Saddle-Point Equations

We now derive the saddle-point equations by extremizing the replica free entropy functional with respect to the order parameters q and $\hat{q}$.

3.2.1 Derivative with Respect to $\hat{q}$

The stationarity condition is given by:

$$\frac{\partial}{\partial \hat{q}} \left(-\frac{1}{2} \operatorname{Tr}[q\hat{q}] + I_P(q, \hat{q}) \right) = 0. \tag{3.31}$$

Taking the derivative of the I_P equation (3.29) yields the fixed-point equation for elements of matrix q as:

$$q_{ll'} = \mathbb{E}_\xi \left[\frac{\int dw\, P_0(w)\, w_l w_{l'}\, \exp\left(-\frac{1}{2} w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w\right)}{\int dw\, P_0(w)\, \exp\left(-\frac{1}{2} w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w\right)} \right].$$

Equivalently, we can have that:

$$q = \mathbb{E}_\xi \left[\mathbb{E}_{P(w|\xi)} \left[w w^\top \right] \right],$$

where the posterior distribution over weights is defined as

$$P(w \mid \xi) \propto P_0(w)\, \exp\left(-\tfrac{1}{2} w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w\right),$$

and $\xi \sim \mathcal{N}(0, I_k)$ is a standard Gaussian random vector.

3.2.2 Derivative with Respect to q

From the stationarity condition in Equation (3.31) we obtain:

$$\hat{q} = 2\alpha \frac{\partial I_C(q)}{\partial q}.$$

To compute the gradient $\partial I_C(q)/\partial q$, we use the channel term in Equation (3.30). We compute the functional derivative $\partial I_C(q)/\partial q$ using standard Gaussian chain rules. The resulting expression for the elements of $\hat{q} \in \mathbb{R}^{k \times k}$ is:

$$\hat{q}_{ll'} = \alpha \cdot \mathbb{E}_{\xi, Z^0, y} \left[\frac{\int \mathcal{D}Z\, P_{\text{out}}(y \mid \tilde{z}) \left(\frac{\partial \tilde{z}_l}{\partial q} \right) \left(\frac{\partial \ln P_{\text{out}}(y|\tilde{z})}{\partial \tilde{z}_{l'}} \right)}{\int \mathcal{D}Z\, P_{\text{out}}(y \mid \tilde{z})} \right],$$

where:

$$\tilde{z} = (Q^0 - q)^{1/2} Z + q^{1/2} \xi.$$

In practice, this gradient is computed via expectations over Gaussian variables and evaluated numerically using Monte Carlo sampling or Gauss–Hermite quadrature.

3.2.3 Fixed Point Equations

The saddle-point equations for the overlap q and its conjugate $\hat{q}$ in the replica symmetric framework are given by:

$$q = \mathbb{E}_\xi \left[\mathbb{E}_{P(w|\xi)} \left[w w^\top \right] \right], \tag{3.32}$$

$$\hat{q} = 2\alpha \frac{\partial I_C(q)}{\partial q}, \tag{3.33}$$

where the posterior distribution is defined as:

$$P(w \mid \xi) \propto P_0(w) \exp\left(-\tfrac{1}{2} w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w \right),$$

and $\xi \sim \mathcal{N}(0, I_K)$. These equations characterize the fixed-point condition satisfied by the order parameters in the Bayes-optimal inference setting under the replica symmetric assumption.

3.2.4 The Generalization Error

In this subsection, we present two natural definitions of the generalization error and the relationship between them. First is the Gibbs generalization error, which is given by:

$$\epsilon_g^{\text{Gibbs}} \equiv \frac{1}{2} \mathbb{E}_{W^*, X} \left\langle \left(\varphi_{\text{out}}(XW) - \varphi_{\text{out}}(XW^*) \right)^2 \right\rangle, \tag{3.34}$$

which measures the expected squared error between the model's prediction and the true output, averaged over both the posterior and the data.

Second is the Bayes generalization error, which is given by:

$$\epsilon_g^{\text{Bayes}} \equiv \frac{1}{2} \mathbb{E}_{W^*, X} \left(\left(\langle \varphi_{\text{out}}(XW) \rangle - \varphi_{\text{out}}(XW^*) \right)^2 \right), \tag{3.35}$$

which corresponds to the mean squared error of the Bayes predictor that we get by averaging over the posterior distribution over W.

The Nishimori identity in Proposition A.1 can be employed to show that the two generalization errors are related by the simple identity [21]:

$$\epsilon_g^{\text{Gibbs}} = 2\,\epsilon_g^{\text{Bayes}}.$$

Generalization error using $k = 1$

When $k = 1$, the generalization error as defined in Equations (3.34) and (3.35) simplifies and can be computed in closed form with minimal complexity.

We denote the student weight matrix by W, which for $k = 1$ reduces to a single column:

$$W = [w] \in \mathbb{R}^n.$$

The student self-overlap and the teacher self-overlap are equal and we denoted them as Q^0. The teacher–student cross-overlap is:

$$q = \frac{1}{n}\,\mathbb{E}_{w\sim P(w|X,Y)}\left[{w^*}^\top w\right]. \tag{3.36}$$

In this setting, the joint distribution of any two pre-activations (Z, Z^*) is Gaussian with zero mean and 2×2 covariance matrix:

$$\Sigma = \begin{pmatrix} Q^0 & q \\ q & Q^0 \end{pmatrix}.$$

Accordingly, the Bayes-optimal generalization error is:

$$\varepsilon_g = \frac{1}{2}\mathbb{E}\left[\left(\langle\varphi_{\text{out}}(u)\rangle - \varphi_{\text{out}}(u^*)\right)^2\right], \tag{3.37}$$

where $\varphi_{\text{out}}(\cdot)$ is the activation function.

Chapter 4

Experiments and Methodology

4.1 Methodology

This section brings to light the experimental framework we used to study alignment, generalization, and learning transition between the teacher neural network and the student neural network. The goal is to understand how the overlaps between corresponding hidden units evolve during training, their variation with sample complexity, and how this relates to the replica theory.

4.1.1 Experimental Setup

We consider a teacher-student setup where both networks have the same architecture; that is, an input layer, a hidden layer with either $k = 1$ or $k = 4$ hidden units, a non-linear activation, and non-uniform readouts for the case $k = 4$. The teacher network generates labeled data, and the student is trained on the generated data. The input dimension is fixed at specific values of n and the number of training samples m is controlled using a parameter $\alpha = m/n$. We used the *tanh* activation for the case $k = 1$, and for $k = 4$, we used a polynomial activation function given by:

$$\sigma(x) = x + \frac{x^2 - 1}{\sqrt{2}} + \frac{x^3 - 3x}{6}, \tag{4.1}$$

4.1.2 The Teacher Model and Data Generation

The teacher network is defined by a random weight matrix $W_0 \in \mathbb{R}^{k \times n}$, sampled from a standard Gaussian distribution. Inputs $X \in \mathbb{R}^{n \times m}$ are sampled i.i.d. from a standard

Gaussian distribution as well. The outputs $Y \in \mathbb{R}^m$ are computed as:

$$Y = \frac{1}{\sqrt{k}} v^\top \sigma\left(\frac{W_0 X}{\sqrt{n}}\right) + \sqrt{\Delta} Z, \tag{4.2}$$

where $v \in \mathbb{R}^k$ is a fixed normalized vector, $Z \sim \mathcal{N}(0, I_m)$, and σ is a chosen nonlinearity. We use a scalar parameter Δ to control the magnitude of the noise added to the teacher's output.

4.1.3 Student Network and Training Procedure

The student network uses exactly the same architecture as the teacher, its weights $W \in \mathbb{R}^{k \times n}$ are initialized randomly from a standard Gaussian distribution. We train the student using stochastic gradient descent and we minimize the squared loss between the student's prediction and the teacher's outputs. The loss is defined as:

$$\mathcal{L}(W) = \frac{1}{2\Delta} \|\hat{Y} - Y\|^2, \quad \text{with } \hat{Y} = \frac{1}{\sqrt{k}} v^\top \phi\left(\frac{WX}{\sqrt{n}}\right).$$

In a second approach, we adopt a Bayesian framework where the student weights are treated as random variables with a Gaussian prior. Given a dataset generated by the teacher, we use HMC to sample the student's weight matrix from the posterior distribution $P(W \mid X, Y)$.

4.1.4 Measuring Alignment

To quantify the alignment between the teacher and student units, we compute a normalized inner product matrix, $S = (W_0 W^\top)/n$. Each entry S_{ij} of S represents the cosine similarity between the *i-th* teacher unit and *j-th* student unit.

In order to isolate the best alignment structure, we use a permutation matrix P obtained via the Hungarian algorithm to reorder student units for maximal diagonal overlap. We then track the square of each alignment coefficient S_{ij}^2 to visualize the strength of the alignment between the teacher and student neurons.

4.1.5 Bayes-Optimal Benchmarking

To benchmark the student network's performance against Bayes-optimal inference, we compute the MMSE predicted by replica theory. To compute the MMSE, we use the

saddle-point equation overlap q obtained from extremizing the replica-symmetric potential [20].

For this part, we used a committee machine with a single hidden unit ($k = 1$) and a scalar readout $v = 1$ along with a hyperbolic tangent activation function.

We use a well-known result for the Bayes-optimal generalization error as [22]:

$$\epsilon_{\mathrm{g}}^{\mathrm{Bayes}}(\alpha) \;=\; h(1) \;-\; h(q),$$

where, for $k = 1$,

$$h(q) \;=\; \mathbb{E}_U\left[\left(\mathbb{E}_V\left[\tanh\left(\sqrt{q}\,U + \sqrt{1-q}\,V\right)\right]\right)^2\right], \quad U, V \sim \mathcal{N}(0,1).$$

We evaluate all integrals over U and V numerically using one-dimensional Gauss-Hermite quadrature.

4.2 Experiments

In the first set of experiments, we used a committee machine with $k = 4$ hidden units for an understanding of learning dynamics under this framework. We performed the following experiments:

- **Varying Alpha:** For each $\alpha \in \{0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0\}$, we trained the student network on data generated by the teacher network using SGD, captured final alignment matrices after convergence, and the plotted squared overlaps. We also carried out a numerical estimation of how alignment matrices evolved at varying alpha using HMC sampling.
- **Evolution at Fixed Alpha:** For $\alpha = 6.0$, the training process captured snapshots of the alignment matrix every 100 training steps. This provided an insight into how the overlaps evolved during the course of learning.

In the second set of our experimentation, the theoretical predictions are implemented using numerical quadrature and iterative updates. For an empirical comparison, we estimate the MMSE from HMC posterior samples of weight vectors.

Chapter 5

Results and Analysis

In this chapter, we build on the analytical framework developed earlier, we examine how learning dynamics actually play out in a finite-width network. Using the teacher-student setup introduced in subsections 4.1.2 and 4.1.3, our goals are to investigate the emergence of specialization and a learning transition, identify the critical threshold α where specialization does set in, and compare empirical results with theoretical predictions from the replica theory with a numerical framework that uses HMC sampling.

5.1 Varying Sample Complexity under SGD

To examine the emergence of specialization during learning using a finite-width Bayesian neural network, we train the student network using SGD on data generated by the teacher network at varying sample complexities, α. Figure 5.1 shows squared alignment matrices between the teacher and student weights for various values of sample complexity after 1000 training steps. Each matrix has been permuted to maximize diagonal alignment, allowing us to assess the extent of specialization. For small $\alpha \leq 1$, the alignment remains diffuse, indicating a regime of statistical symmetry where student units do not yet differentiate themselves.

As we increase α, a clear transition towards structured alignment is observed, with diagonally dominant patterns emerging around $\alpha \approx 2$. This suggests that even under purely gradient-based training, the network can recover a meaningful teacher structure at an appreciable data regime.

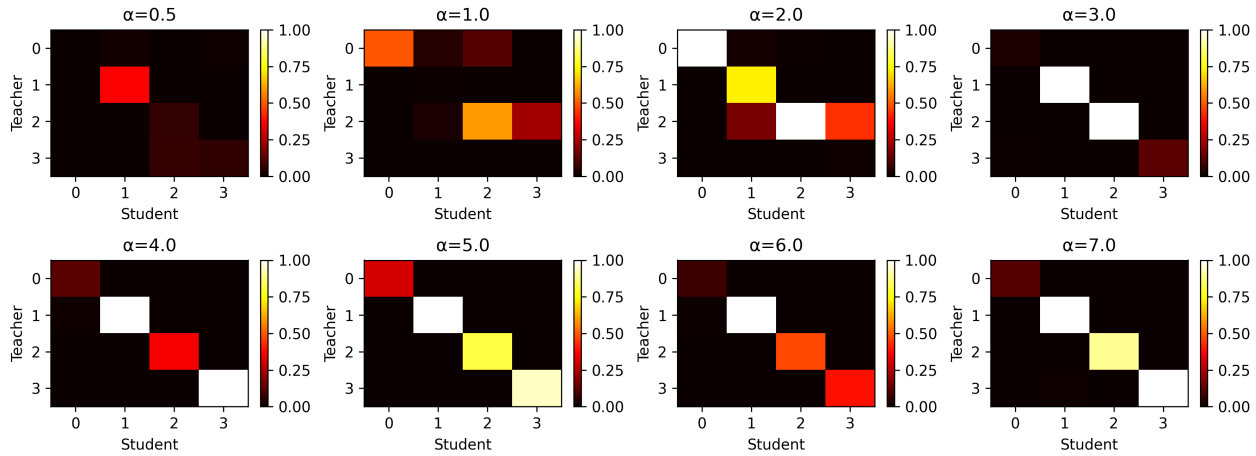


Figure 5.1: Shows final squared alignment matrices between student and teacher neurons at varying sample complexity $\alpha \in \{0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0\}$ with $n = 200$, $k = 4$, noise level $\Delta = 0.5$, and a polynomial activation function, $\sigma(x) = x + (x^2 - 1)/\sqrt{2} + (x^3 - 3x)/6$. Each heatmap shows the squared cosine similarity S^2_{ij} between the i-th teacher neuron and the j-th student neuron after training. As α increases, we observe a transition from unstructured overlaps to diagonal-dominant structures.

5.2 Evolution at Fixed α

For an insight into the temporal evolution of the learning dynamics, we track the alignment matrix during the course of training for a fixed sample complexity $\alpha = 6$. Figure 5.2 presents snapshots of the squared alignment matrices S^2 at intervals of 100 for our 1000 training steps starting from step 200. The early stages of training exhibit weak and almost random alignments with no clear specialization; however, as training progresses, we observe the gradual formation of a diagonally dominant pattern, which is evidence of increasing alignment between specific student neurons and teacher neurons. This progressive sharpening suggests to us that specialization is not an instantaneous event but rather it emerges incrementally as the network minimizes its loss, eventually converging to a structured internal representation.

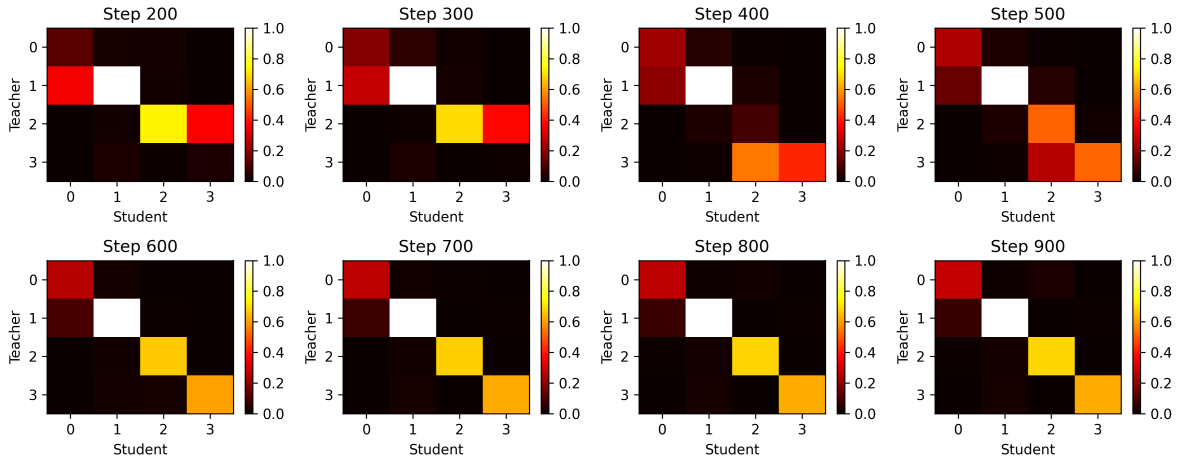


Figure 5.2: Shows the evolution of squared alignment matrices S^2 during SGD training at fixed sample complexity $\alpha = 6$, with $n = 200$, $k = 4$, $\Delta = 0.5$, and a polynomial activation function, $\sigma(x) = x + (x^2 - 1)/\sqrt{2} + (x^3 - 3x)/6$. Each plot shows the alignment between teacher and student units after every 100 training steps, from step 200 to 900. The matrices are permutation-aligned to highlight the progressive emergence of neuron specialization. The diagonal structure intensifies over time, and this does indicate that alignment and specialization develop gradually during training.

5.3 Varying Sample Complexity under HMC

When we varied the sample complexity and used HMC, at a low-α regime $\alpha \leq 1$; alignment matrices exhibit diffuse, non-diagonal patterns, indicating weak correlation between teacher and student neurons. This reflects the statistical symmetry phase where posterior samples fail to capture teacher structure due to insufficient data.

With increasing α, a diagonal dominant structure emerges. This depicts specialization where each student unit uniquely aligns with a specific teacher unit. Residual off-diagonal elements diminish as α increases which is consistent with posterior concentration.

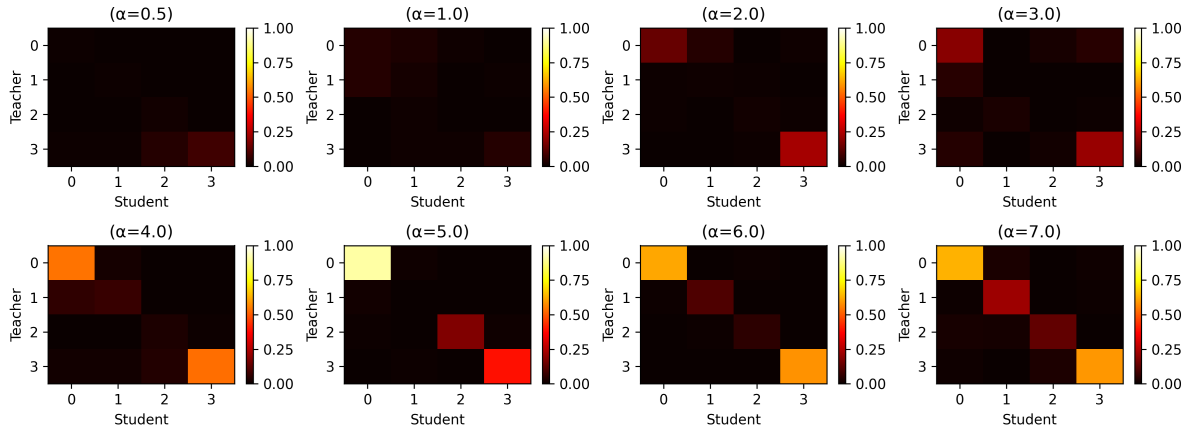


Figure 5.3: Shows the squared alignment matrices S^2 between teacher and student neurons under HMC sampling across varying sample complexity parameters $\alpha \in \{0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0\}$. The experiments were conducted for a committee machine with $k = 4$ hidden units, input dimension $d = 200$, noise level $\Delta = 0.1$, and the polynomial activation function, $\sigma(x) = x + (x^2 - 1)/\sqrt{2} + (x^3 - 3x)/6$. Teacher neurons are ordered by their readout weights $\mathbf{v} = [-3.0, -1.0, 1.0, 3.0]$ to enable consistent comparison.

5.4 Replica Theory MMSE Vs. HMC Simulations

Figure 5.4 shows the minimum mean–squared error (MMSE) as a function of the sample complexity $\alpha = m/n$. We used two forms: (i) The *replica–theory* prediction was obtained by numerically solving the replica-based saddle-point equations for the Bayes–optimal estimator. (ii) Then, second was empirical MMSE from HMC sampling of the Bayesian posterior over the student weights given the training data.
On the same axes, the right y-axis displays the overlaps from the replica theory and those obtained from HMC sampling of the posterior distribution.

5.4.1 Key Observations

- **MMSE decay with α.** Both the theoretical prediction and the HMC estimates exhibit a monotonic decrease of the MMSE as α increases. For small α, the limited number of samples leaves the posterior broadly spread over weight space, leading to a higher reconstruction error. As α grows, the posterior concentrates around the teacher weight vector causing the MMSE to drop rapidly toward zero.

- **Overlap growth q.** The overlap q between the student and teacher weights increases steadily from low values starting at at $\alpha = 0.5$ to nearly one for $\alpha \geq 4$.

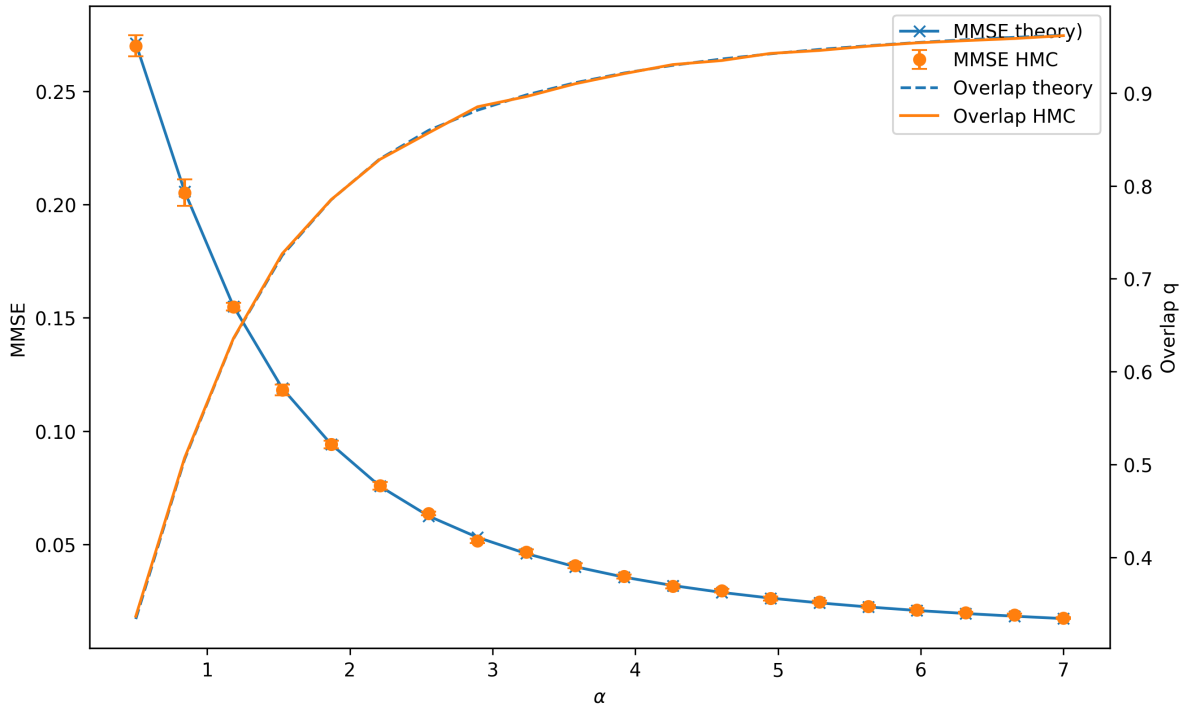


Figure 5.4: Comparison of Bayes-optimal theoretical predictions (solid and dashed curves) and HMC simulation results (markers with 95% confidence intervals) for the MMSE (left axis), and overlap q (right axis) as functions of the sample complexity α, for ($k = 1$), tanh activation, Gaussian inputs, $n = 1000$, and $\Delta = 0.1$. Theory curves are obtained from the fixed-point solution of the replica-based fixed point equations, empirical values are obtained from posterior samples generated via HMC.

This reflects progressive alignment of student units with the teacher units as more data is provided.

- **Excellent theory–simulation agreement.** The HMC estimates match the theoretical predictions for both MMSE and overlap q with high accuracy across the entire range of α. The close agreement indicates that finite-dimension effects are negligible in this setting, and validates the replica-based theoretical predictions for $k = 1$.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this work, we investigated generalization, specialization, and learning transition in Bayesian neural networks using the teacher–student committee machine with finite hidden units. Our approach combined a replica theory analysis under RS ansatz with empirical studies based on SGD and HMC. Our approach gave way to a coherent view of how sample complexity α shapes learning when we are in a high dimension setting with finite hidden units. We found out that, when we use either SGD or HMC, alignment matrices evolve from diffuse to diagonally dominant structures as α increases with a clear emergence of specialization around $\alpha \approx 2$ with $k = 4$ hidden units. We investigated the Bayes optimal generalization error for a simple setting with $k = 1$, the HMC estimates of both MMSE and overlap q closely match the replica-based curves across the full range of α, and this provides a strong validation of the replica theory predictions for our setting. The results confirm that replica-theory predictions capture essential statistical dynamics of finite-width learning giving us an idea of what happens in the infinite-width regime [23].

Limitations and scope. In the experiments, we use Gaussian inputs, and i.i.d weights for the teacher and the student networks which also follow Gaussian densities. Furthermore, to compare the results from the replica theory and the empirical analysis, we restricted ourselves to the case $k = 1$. These choices were made for analytical tractability and ease of benchmarking but certainly leave open questions about broader architectures and data distributions.

6.2 Future Work

The results presented in this thesis open several promising avenues for further research:

Finite-Size Scaling and Universality. A comprehensive study of how the location and sharpness of the specialization transition depend on network width k, input dimension n, and noise level Δ could be very beneficial in establishing scaling laws and universality classes for learning transitions.

Larger and Deeper Architectures. Extending the present finite-width analysis to multi-layer networks can appreciably shed light on whether the specialization dynamics, generalization, and learning transition do persist in these settings. Even still, before extending to deeper networks, exploring how the analysis we present holds for higher values of hidden units is an interesting avenue.

Appendix A

A.1 The Nishimori property in Bayes-optimal learning

Here is an important property of the Bayesian optimal setting when all hyper-parameters of the problem are assumed to be known, and it is often referred to as the Nishimori identity.

Proposition A.1 (Nishimori identity)**.** *Let $(X, Y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be a couple of random variables. Let $k \geq 1$ and let $X^{(1)}, \dots, X^{(k)}$ be k i.i.d. samples (given Y) from the conditional distribution $P(X = \cdot \mid Y)$, independently of every other random variables. Let us denote $\langle \cdot \rangle$ the expectation operator w.r.t. $P(X = \cdot \mid Y)$ and $\mathbb{E}$ the expectation w.r.t. (X, Y). Then, for all continuous bounded functions g we have [3]*

$$\mathbb{E}\langle g\big(Y, X^{(1)}, \dots, X^{(k)}\big)\rangle = \mathbb{E}\langle g\big(Y, X^{(1)}, \dots, X^{(k-1)}, X\big)\rangle\,. \tag{31}$$

Proof. This is a simple consequence of Bayes' formula. It is equivalent to sample the couple (X, Y) according to its joint distribution or to sample first Y according to its marginal distribution and then to sample X conditionally to Y from its conditional distribution $P(X = \cdot \mid Y)$. Thus, the $(k+1)$-tuple $(Y, X^{(1)}, \dots, X^{(k)})$ is equal in law to $(Y, X^{(1)}, \dots, X^{(k-1)}, X)$. This proves the proposition. □

Bibliography

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[2] E. Barkai, D. Hansel, and H. Sompolinsky, "Broken symmetries in multilayered perceptrons," *Physical Review A*, vol. 45, no. 6, p. 4146, 1992.

[3] B. Aubin, A. Maillard, F. Krzakala, N. Macris, L. Zdeborová *et al.*, "The committee machine: Computational to statistical gaps in learning a two-layers neural network," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[4] C. Schülke, P. Schniter, and L. Zdeborová, "Phase diagram of matrix compressed sensing," *Physical Review E*, vol. 94, no. 6, p. 062136, 2016.

[5] M. Biehl, E. Schlösser, and M. Ahr, "Phase transitions in soft-committee machines," *Europhysics Letters*, vol. 44, no. 2, p. 261, 1998.

[6] H. Schwarze and J. Hertz, "Generalization in fully connected committee machines," *Europhysics Letters*, vol. 21, no. 7, p. 785, 1993.

[7] M. S. Advani, A. M. Saxe, and H. Sompolinsky, "High-dimensional dynamics of generalization error in neural networks," *Neural Networks*, vol. 132, pp. 428–446, 2020.

[8] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[9] S. Abbasi, M. Hajabdollahi, N. Karimi, and S. Samavi, "Modeling teacher-student techniques in deep neural networks for knowledge distillation," *arXiv preprint arXiv:1912.13179*, 2019.

[10] S. Ariosto, "Statistical physics of deep neural networks: Generalization capability, beyond the infinite width, and feature learning," *arXiv preprint arXiv:2501.19281*, 2025.

[11] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, "Finite versus infinite neural networks: an empirical study," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 156–15 172, 2020.

[12] M. Biehl and H. Schwarze, "Learning by on-line gradient descent," *Journal of Physics A: Mathematical and general*, vol. 28, no. 3, p. 643, 1995.

[13] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, "Modeling the influence of data structure on learning in neural networks: The hidden manifold model," *Physical Review X*, vol. 10, no. 4, p. 041044, 2020.

[14] M. Opper, "Statistical mechanics of learning: Generalization," *The handbook of brain theory and neural networks*, pp. 922–925, 1995.

[15] J. Barbier, "High-dimensional inference: a statistical mechanics perspective," *arXiv preprint arXiv:2010.14863*, 2020.

[16] O. Dhifallah and Y. M. Lu, "Phase transitions in transfer learning for high-dimensional perceptrons," *Entropy*, vol. 23, no. 4, p. 400, 2021.

[17] D. Saad and S. A. Solla, "On-line learning in soft committee machines," *Physical Review E*, vol. 52, no. 4, p. 4225, 1995.

[18] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup," *Advances in neural information processing systems*, vol. 32, 2019.

[19] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

[20] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[21] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*. Clarendon Press, 2001, no. 111.

[22] M. Reid and R. Williamson, "Information, divergence and risk for binary experiments," 2011.

[23] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.