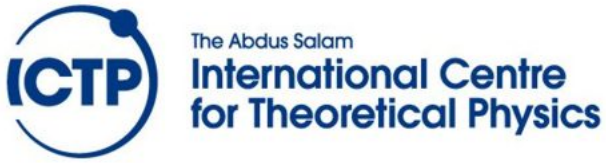
Optimal Generalization and Learning Transition in Finite-Width Bayesian Neural Networks

By

Rolland Mucunguzi Alinda



Quantitative Life Sciences Section

Diploma Project
Supervisor: Prof. Jean Barbier
Co-supervisor: Dr. Minh Toan Nguyen

August 2025

Abstract

This work investigates the generalization behavior and specialization transitions in shallow Bayesian neural networks using a two-layer teacher-student architecture known as the committee machine. 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, including the replica method and state evolution theory. 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 state evolution equations for the overlap and conjugate overlap matrices. These equations are solved numerically for finite hidden layer size of 4 neurons and the resulting fixed points are used to estimate Bayes-optimal generalization error. The experiments we conducted reveal a specialization transition confirming theoretical predictions and consequently highlighting the role of network width, sample complexity, the choice of the activation function and initialization in alignment dynamics. This work bridges rigorous theory and practical training dynamics in finite-size network regimes, providing insights into optimal learning in close to real-world models.

Dedication and acknowledgements

Here goes the dedication.

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 not been submitted for any other 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: .....

List of Tables

Table **Page**

List of Figures

Figure **Page**

Contents of Thesis

Chapter 1

Introduction

1.1 Context and background

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

The impressive generalization behavior of neural networks defies classical learning theory in many ways, empirically, deep neural networks can perfectly fit random labels and yet still generalize well on real data. This potentially suggests that classical notions such as Rademacher complexity may not fully explain this performance. This has prompted a surge in theoretical research aiming to understand DNNs from new perspectives among which include; statistical mechanics, information theory and Bayesian inference.

One powerful paradigm 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 given the data alone. This controlled setup allows one 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 [1], but has since been extended to more complex models and algorithms including modern Bayesian neural networks [2, 3].

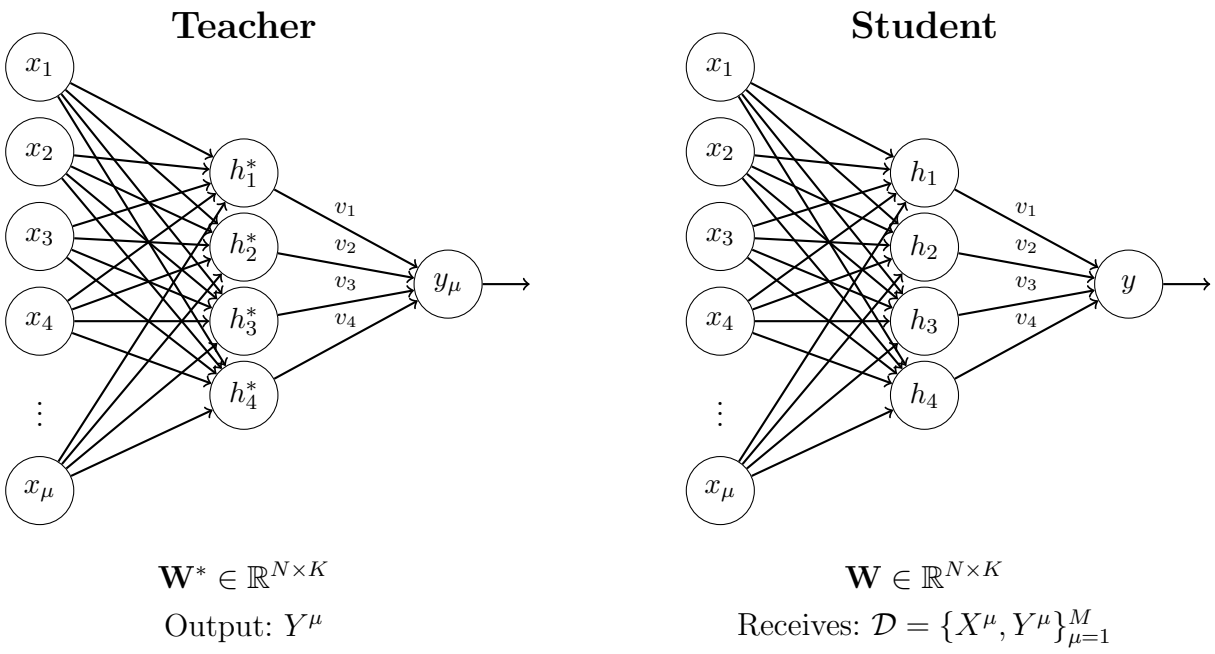


Figure 1.1: Illustration of the teacher-student neural network setup. The teacher generates labels using weights $\mathbf{W}^* \in \mathbb{R}^{N\times K}$, while the student learns from input-label pairs (X^μ, Y^μ) using its own weights $\mathbf{W}$. Hidden units are connected to the output through readout weights v_k.

A very key insight from this line of work is that learning exhibits distinct phases depending on the sample complexity $\alpha = \frac{n}{d}$, with n as the number of training samples and d the input dimension. Below a critical threshold usually denoted as α_{spec}, student units remain statistically 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. 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.

However, much of the existing theory assumes idealized conditions of infinite width of the hidden layer, Gaussian priors and training through posterior sampling. These

assumptions often fail to capture the practical behavior of finite-width networks. In this thesis, we are going to revisit the teacher-student framework in the context of *finite-width shallow networks* and investigate how the alignment, specialization and generalization dynamics show up in these networks.

By bridging the theoretical framework with empirical experimentation, this work aims to show the conditions under which neuron alignment and specialization arise, and to quantify the accuracy of Bayesian predictions for finite regimes. We employ a combination of tools that include stochastic gradient descent, Hamiltonian Monte Carlo and analytical derivations mainly based on the replica method.

1.2 Motivation and significance

While deep neural networks (DNNs) have demonstrated remarkable performance across different of domains, their learning dynamics and generalization capabilities remains 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.

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 thus making 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 multiple hidden units to serves as a canonical model.

By studying alignment and generalization 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 state evolution.

This work provides a quantitative understanding of specialization and generalization in shallow networks, and shows the role of model size and initialization in finite regimes. Furthermore, this research contributes to the broader goal of developing a statistical physics framework of generalization that remains informative even outside the idealized

infinite-width setting.

1.3 Research questions

This thesis aims to investigate the mechanisms underlying alignment and generalization in shallow finite-width Bayesian neural networks using the teacher-student paradigm. Our central goal is to understand how macroscopic learning phenomena such as specialization, symmetry breaking, and generalization error emerge in finite networks and how they relate to the predictions of replica theory and state evolution.

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

1. **How does specialization emerge in finite-width teacher-student networks?**
   What is the critical sample complexity threshold α_{spec} 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 settings?**
   Do the fixed-point solutions from the replica symmetric potential and state evolution equations accurately describe the overlap dynamics and generalization performance observed in experiments conducted using stochastic gradient descent and posterior sampling?

3. **What is the impact of training dynamics and initialization?**
   How does the choice of optimizer (e.g., stochastic gradient descent vs. Hamiltonian Monte Carlo) and initialization scheme affect the convergence to specialized states and the final generalization error?

By answering these questions, we aim to contribute to a more clear understanding of learning dynamics in neural networks, particularly in finite regimes.

1.4 Contributions

1.5 Overview of the thesis structure

Chapter 2

Literature review

2.1 Artificial Neural Networks and Generalization

Deep neural networks (DNNs) have exhibited powerful generalization abilities across different application domains including computer vision, speech recognition, automation and scientific modelling. However, the theoretical principles underlying the success of deep neural networks remains only partially understood. A famous central question that persists is: how can heavily overparameterized models achieve low generalization error despite the fact that they fit noise too?

For a pedestal upon which firm understanding can spring, simplified frameworks have been proposed to study generalization under controlled conditions. Among these is the **teacher-student framework** that has emerged as a foundational pedestal [4]. In this setup, a teacher network with fixed parameters generates labeled data and a student network is trained on this data with the aims of recovering the underlying rule that has generated the data. The simplicity of this arrangement makes it ideal for isolating phenomena such as neuron alignment, specialization, effect of sample size, effect of the choice of the nonlinearity, effect of the choice of prior and others.

While theoretical insights have been primarily derived in the *infinite-width* regimes commonly known as the *thermodynamic limit* in statistical physics [5], 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 [6].

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 [7–9].

Key insights from this literature include:

- **Phase Transitions:** Learning in neural networks can exhibit abrupt changes in behavior as the sample complexity parameter $\alpha = n/d$ is varied. 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 network [10].
- **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 characterized using tools from disordered systems [1].
- **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 [3]. The notion of Bayesian optimality defines a performance benchmark for both theory and algorithms.

Recent work by Barbier et al. [2] extended these ideas to **finite-width** committee machines using adaptive interpolation and rigorous statistical mechanics. They showed that a gap can 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 may fail to reach it.

Committee machines provide a particularly useful model system for exploring such questions. These are two-layer neural networks with K hidden units and a fixed linear

readout. Their simple structure makes them analytically tractable, while still capturing the key features of learning transitions. In the large n, d limit with fixed α, theoretical predictions for generalization curves, specialization thresholds, and overlap dynamics can be derived exactly [11, 12].

Despite these advances, most analyses remain confined to the thermodynamic regime. Real-world networks are finite, often trained with limited data and computational budgets. As such, understanding how finite-size effects alter specialization, generalization, and algorithmic success is essential for bridging theory and practice.

2.1.2 Learning Dynamics

The evolution of learning in teacher-student networks is often analyzed through the lens of **overlap matrices**, which measure the correlation between student and teacher weights. An overlap matrix $S = W_{\text{teacher}} W_{\text{student}}^T / d$ captures alignment, and permutation symmetries can be revealed through techniques like the Hungarian algorithm.

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 symmetry persistence [11].

Generalization is also modulated by the **sample complexity** α and the **noise level** Δ in the training data. At low α, models tend to underfit or memorize noise, while higher α facilitates meaningful alignment and generalization [13]. Interestingly, alignment may improve with increased noise if the model structure and inference procedure are appropriately matched.

One particularly intriguing phenomenon is the **double descent** behavior of test error as a function of model size or sample complexity. Initially decreasing with more data, test error increases near the interpolation threshold before decreasing again. This behavior has been observed in both linear models and deep networks and is now understood to emerge even in shallow teacher-student models as a function of α [14].

Finally, a key theme in recent literature is the **computational vs. statistical gap**. 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 Hamiltonian Monte Carlo (HMC) have been shown to recover the correct alignment structure more consistently, albeit at greater computational cost. This underscores the importance of considering algorithmic limitations when evaluating generalization performance [2].

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

Chapter 3

Theory and Methodology

3.1 Methodology

This section outlines the experimental framework used to study the alignment between the teacher and the student neurons. The goal is to understand how the overlap between corresponding hidden units evolves during training and also its variation with sample complexities.

3.1.1 Experimental Setup

We consider a teacher-student setup where both networks have the same architecture that is; a single hidden layer with $k = 4$ units and non-linear activation. The teacher network generates labeled data and the student is trained on the generated data.

The input dimension is fixed at $d = 150$ and the number of training samples n is controlled using a parameter α that governs the sample complexity and it is defined as:

$$\alpha = \frac{n}{kd}.$$

3.1.2 Teacher Model and Data Generation

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

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

where $v \in \mathbb{R}^k$ is a fixed normalized vector, $Z \sim \mathcal{N}(0, I_n)$, and σ is the chosen nonlinearity. A scalar parameter $\Delta > 0$ is used to control the magnitude of the noise added to the teacher's output. Different architectures and nonlinear functions can fit this framework, however, we shall employ the committee machine.

3.1.3 Student Network and Training Procedure

The student network uses exactly the same architecture as the teacher, its weights $W \in \mathbb{R}^{k \times d}$ 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 function used is defined as:

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

The optimization was performed using Adam optimizer with a learning rate of 0.01 for up to 1000 steps.

3.1.4 Measuring Alignment

To quantify the alignment between the teacher and student units, we compute a normalized inner product matrix:

$$S = \frac{1}{d} W_0 W^\top,$$

where each entry S_{ij} 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:

$$\text{Aligned matrix} = P^\top S, \quad \text{where } P \text{ is the optimal permutation matrix.}$$

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.

3.1.5 Experiments

Two sets of experiments were conducted:

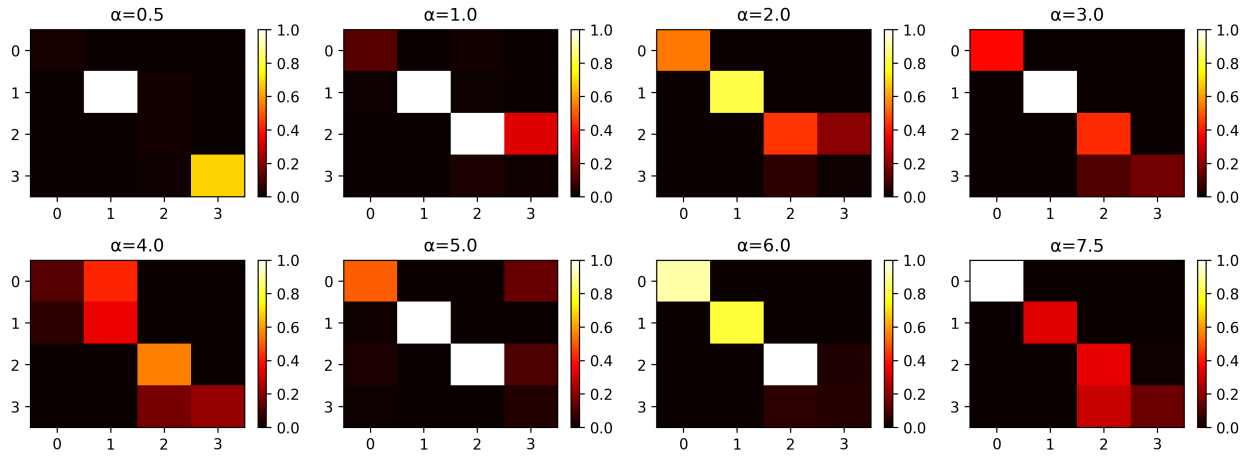
- **Varying Alpha:** For each $\alpha \in \{0.5, 1.0, 2.0, 3.0, 4.0, 5, 6.0, 7.5\}$, the student network was trained on data generated from the teacher network. The final alignment matrices were recorded after convergence and squared overlaps were plotted.

- **Evolution at Fixed Alpha:** For $\alpha = 6.0$, the training process captured snapshots of the alignment matrix every 100 training steps. This provided insight into how the overlaps during the course of learning evolve with time.

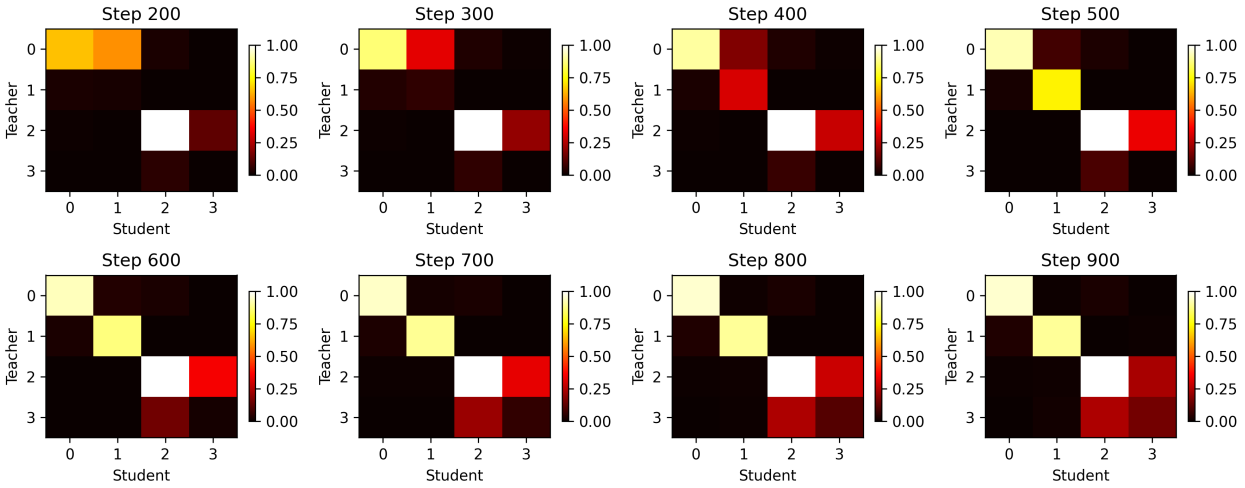
Chapter 4

Experimentation and Analysis

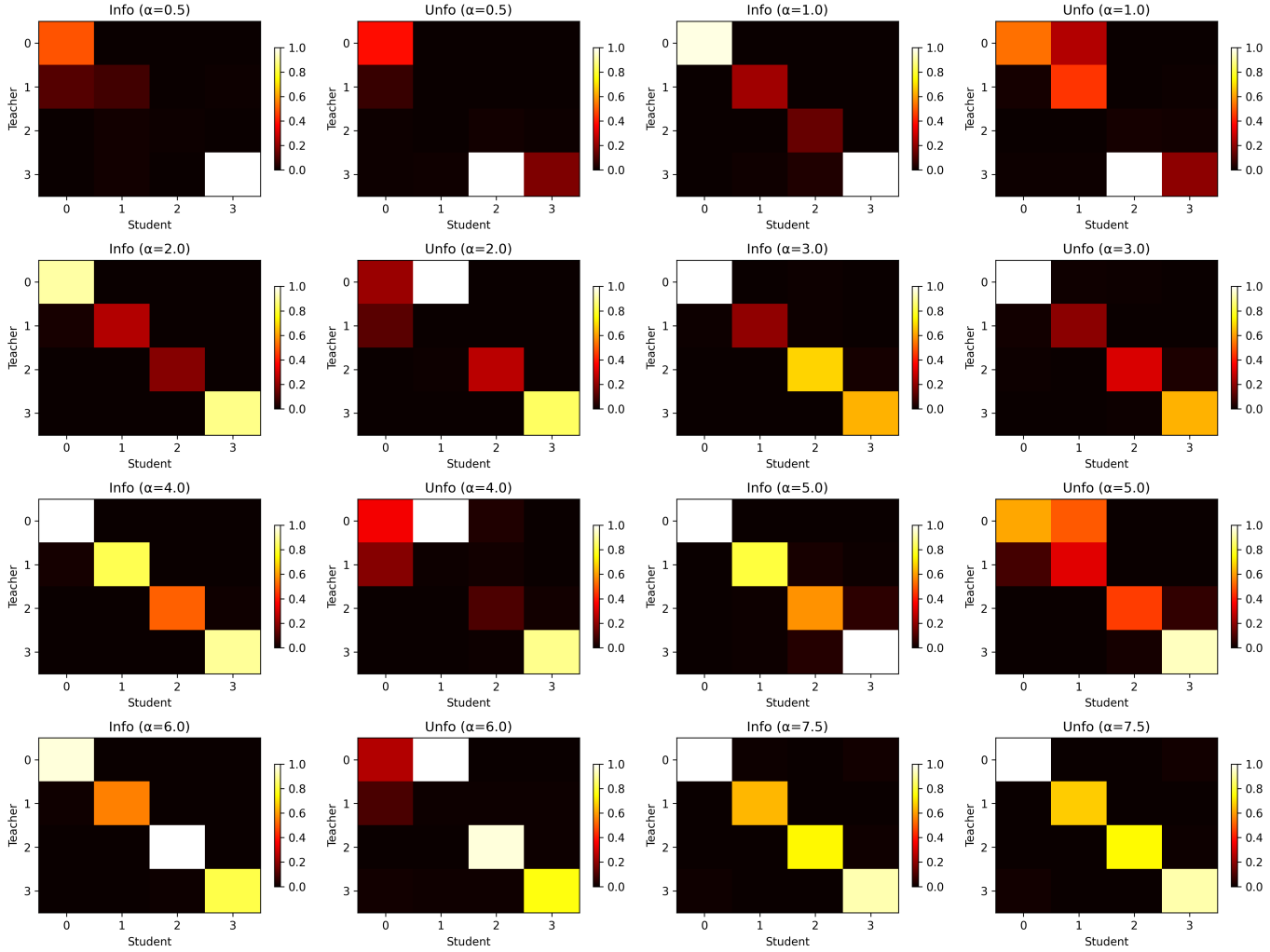
4.1 Effect of varying Alpha values



4.2 Evolution at Alpha=6.0



4.3 The effect of Initialization



4.4 Teacher Model and Data Generation

In this section, we shall enrich the symbolism used in such that the calculations for the replica method follow smoothly. Given a dataset of m input samples $\{X_{\mu i}\}_{\mu=1,\dots,m;\ i=1,\dots,d}$ with d the dimension of an input, we denote by W^*_{il} the weight connecting the i-th input feature among the visible units to the l-th hidden neuron of the teacher. Now, the teacher generates output labels according to a generic function $\phi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$, such that

$$Y_\mu = \phi_{\text{out}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{\mu i} W_{il}^*,\ B_\mu \right),$$

Which we can write in probabilistic form as,

$$Y_\mu \sim P_{\text{out}} \left(\cdot \,\middle|\, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{\mu i} W_{il}^* \right),$$

where $\{B_\mu\}_{\mu=1}^m$ are i.i.d. noise variables drawn from a known distribution P_B, accounting for uncertainty in the outputs whereby in deterministic setups, this second argument may be omitted or else modeled as a Dirac delta.

This model can be interpreted as a noisy channel, with P_{out} functioning as the channel's transition kernel. In our project, we focus on a teacher-student scenario, where the teacher's inputs are sampled i.i.d. from a standard Gaussian distribution, $X_{\mu i} \sim \mathcal{N}(0, 1)$, and the teacher weights are sampled i.i.d. from a prior distribution, $W_{il}^* \sim P_0$.

The student is presented with training pairs $\{(X_\mu, Y_\mu)\}_{\mu=1}^m$ and aims to recover W^* by estimating the posterior distribution over weights. This was experimented with stochastic gradient descent and Hamiltonian Monte Carlo (HMC), enabling a comparison with predictions obtained from replica theory.

Chapter 5

Discussion of Results

Interpret results in light of research questions
Highlight implications and insights
Limitations and potential confounding factors
We need to introduce the Bayesian formulation.
State some theorem, proposition and the Nashimori Identity.
We shall need to introduce free entropy.

5.1 The Replica calculation

In this section, we provide a heuristic derivation of the replica formula using the replica method which is a powerful but non-rigorous tool that is commonly used to study the statistical properties of disordered systems. Our goal in this section is to derive a closed-form expression for the free entropy of a shallow Bayesian neural network under a committee machine model with Gaussian priors and obtain associated saddle-point equations for the order parameters.

The replica method employs the replica trick where instead of calculating the expectation of the logarithm of a function, we replicate the function and compute the logarithm of the expectation for the replicated system. 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:

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} \ln f_n = \lim_{p\to 0^+} \lim_{n\to\infty} \frac{1}{np} \ln \mathbb{E} f_n^p. \tag{5.1}$$

In the above formulation, exchanging the limits is a non-rigorous step. We can calculate the moments $\mathbb{E}f_n^p$ for integers $p \in \mathbb{N}$, and then after take the limit $p \to 0^+$ by analytic continuation. We need to compute the *free entropy* of our system, $f \equiv \lim_{n\to\infty} f_n$ and to this end, let us first obtain the moments of the replicated partition function with $p \in \mathbb{Z}$:

$$\mathbb{E}\mathcal{Z}_n^p = \mathbb{E}\left(\int_{\mathbb{R}^n\times\mathbb{R}^K} dw \prod_{i=1}^{n} P_0\left(\{\ w_{il}\ \}_{l=1}^{K}\right) \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \middle| \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{\mu i} w_{il}\right\}_{l=1}^{K}\right)\right)^p, \tag{5.2}$$

$$= \mathbb{E}\left(\prod_{a=1}^{p}\int_{\mathbb{R}^n\times\mathbb{R}^K} dw^a \prod_{i=1}^{n} P_0\left(\{\ w_{il}^a\ \}_{l=1}^{K}\right) \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \middle| \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{\mu i} w_{il}^a\right\}_{l=1}^{K}\right)\right). \tag{5.3}$$

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 and we write:

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

We notice that X is an i.i.d. standard Gaussian matrix, thus for every a, μ, l, $Z_{\mu l}^a \equiv n^{-1/2}\sum_{i=1}^{n} X_{\mu i} w_{il}^a$, with $Z_{\mu l}^a$ the pre-activation of the $l-th$ hidden unit of the $a-th$ replica on example μ, there follows a multivariate Gaussian distribution, with a mean value zero. This implies that we need to introduce a covariance tensor to proceed with the calculation and thus we have:

$$\mathbb{E}Z_{\mu l}^a Z_{\nu l'}^b = \delta_{\mu\nu}\Sigma_{\substack{al\\bl'}} = \delta_{\mu\nu} Q_{bl'}^{al}, \tag{5.4}$$

For each pair of replicas a and b, we define a matrix of overlaps as:

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

with entries

$$Q_{bl'}^{al} = \frac{1}{n}\sum_{i=1}^{n} w_{il}^a w_{il'}^b.$$

To fix the values of the overlap matrices Q, we introduce Dirac delta functions that enforce their definitions. In this context, the covariance matrix Σ of the local fields $Z_{\mu l}^a$ is of dimension $(p+1)K \times (p+1)K$, reflecting the number of replicas and hidden

units. Incorporating these constraints, the expression for the replicated partition function becomes:

$$\mathbb{E}Z_n^p = \prod_{(a,r)} \int_{\mathbb{R}} \mathrm{d}Q_{ar}^{ar} \prod_{\{(a,r);(b,r')\}} \int_{\mathbb{R}} \mathrm{d}Q_{br'}^{ar} \left(I_{\text{prior(P)}} \left(\{Q_{br'}^{ar}\} \right) \times I_{\text{channel(C)}} \left(\{Q_{br'}^{ar}\} \right) \right) \tag{5.5}$$

It follows that:

$$I_{\mathrm{P}}(\{Q_{br'}^{ar}\}) = \prod_{a=0}^{p} \left(\int_{\mathbb{R}^{n\times K}} dw^a P_0(w^a) \right) \left(\prod_{\{(a,l);(b,l')\}} \delta \left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) \right). \tag{5.6}$$

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 Gaussian with covariance matrix Σ, the channel contribution takes the form:

$$I_{\mathrm{C}} = \int_{\mathbb{R}^m} dY \prod_{\mu=1}^{m} \left[\int_{\mathbb{R}^{(p+1)K}} dZ_\mu \prod_{a=0}^{p} P_{\text{out}}(Y_\mu \mid Z_\mu^a) \cdot \mathcal{N}(Z_\mu; 0, \Sigma) \right], \tag{5.7}$$

where the multivariate Gaussian density is given by:

$$\mathcal{N}(Z; 0, \Sigma) = \frac{1}{\sqrt{(2\pi)^{(p+1)K} \det \Sigma}} \exp\left(-\frac{1}{2} Z^\top \Sigma^{-1} Z \right). \tag{5.8}$$

Factoring the integration over all $\mu = 1, \ldots, m$, we obtain:

$$\begin{aligned} I_{\mathrm{C}} = \int_{\mathbb{R}^m} dY \prod_{a=0}^{p} \int_{\mathbb{R}^{m\times K}} dZ^a \prod_{a=0}^{p} P_{\text{out}}(Y \mid Z^a) \exp\left(-\frac{m}{2} \ln \det \Sigma - \frac{mK(p+1)}{2} \ln 2\pi \right) \\ \times \exp\left(-\frac{1}{2} \sum_{\mu=1}^{m} \sum_{a,b} \sum_{l,l'} Z_{\mu l}^a Z_{\mu l'}^b \left(\Sigma^{-1} \right)_{\substack{al \\ bl'}} \right). \end{aligned} \tag{5.9}$$

We can replace the Dirac delta function by its Fourier representation, this formulation is depicted as:

$$\delta \left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) = \int \frac{d\hat{Q}_{bl'}^{al}}{2\pi} \exp\left[\hat{Q}_{bl'}^{al} \left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) \right]. \tag{5.10}$$

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

$$I_{\mathrm{P}} = \int d\hat{Q} \, \exp\left[\sum_{a,b,l,l'} \hat{Q}_{bl'}^{al} Q_{bl'}^{al} \right] \prod_{i=1}^{n} \int dw_i \, P_0(w_i) \exp\left[- \sum_{a,b,l,l'} \hat{Q}_{bl'}^{al} w_{il}^a w_{il'}^b \right]. \tag{5.11}$$

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:

$$I_{\mathrm{P}} \propto \int d\hat{Q} \, \exp\Big[n \cdot \Phi(Q, \hat{Q})\Big], \tag{5.12}$$

where the exponent is given by:

$$\Phi(Q, \hat{Q}) = \sum_{a,b,l,l'} \hat{Q}^{al}_{bl'} Q^{al}_{bl'} + \ln \int dw \, P_0(w) \exp\left(- \sum_{a,b,l,l'} \hat{Q}^{al}_{bl'} w^a_l w^b_{l'} \right). \tag{5.13}$$

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

$$\lim_{n\to\infty} \frac{1}{n} \ln I_{\mathrm{P}} = \mathrm{ext}_{\hat{Q}} \, \Phi(Q, \hat{Q}). \tag{5.14}$$

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, and collecting all constants, the replica computation yields the final variational expression as:

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}[Z^p_n] = \mathrm{ext}_{Q,\hat{Q}} \, H(Q, \hat{Q}), \tag{5.15}$$

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

$$H(Q, \hat{Q}) \equiv \frac{1}{2} \sum_{a=0}^{p} \sum_{l,l'} Q^{al}_{al} \hat{Q}^{al}_{al} - \frac{1}{2} \sum_{a\neq b} \sum_{l,l'} Q^{al}_{bl'} \hat{Q}^{al}_{bl'} + \ln I + \alpha \ln J, \tag{5.16}$$

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

$$I \equiv \prod_{a=0}^{p} \int_{\mathbb{R}^K} dw^a P_0(w^a) \exp\left(-\frac{1}{2} \sum_{a=0}^{p} \sum_{l,l'} \hat{Q}^{al}_{al'} w^a_l w^a_{l'} + \frac{1}{2} \sum_{a\neq b} \sum_{l,l'} \hat{Q}^{al}_{bl'} w^a_l w^b_{l'} \right), \tag{5.17}$$

$$J \equiv \int_{\mathbb{R}} dy \prod_{a=0}^{p} \int_{\mathbb{R}^K} \frac{dZ^a}{(2\pi)^{K(p+1)/2}} \frac{P_{\mathrm{out}}(y|Z^a)}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} \sum_{a,b=0}^{p} \sum_{l,l'=1}^{K} Z^a_l Z^b_{l'} (\Sigma^{-1})^{\ al}_{\ bl'} \right). \tag{5.18}$$

To make the extremization tractable, we assume a *replica-symmetric* (RS) structure on the overlaps. This means we assume the solution is symmetric under permutations of the student replicas (indices $1, \ldots, p$), and we also treat all teacher-student overlaps in a uniform way.

We assume the teacher or student self-overlap structure is the same for all hidden units l, l':

$$\exists Q^0 \in \mathbb{R}^{K\times K} \quad \text{s.t.} \quad \forall a,\ \forall l, l' : \quad Q^{al}_{al'} = Q^0_{ll'}$$

For all pairs of distinct replicas $a < b$, we assume the overlap between hidden units is the same: $\exists q \in \mathbb{R}^{K\times K} \quad \text{s.t.} \quad \forall a < b,\ \forall l, l' : \quad Q^{al}_{bl'} = q_{ll'}$

An identical assumption is held for the conjugate variables $\hat{Q}$ where we introduce matrices $\hat{Q}^0$ and $\hat{q}$. Notable is that Q^0 is a symmetric matrix by construction, and q is also symmetric under the replica symmetry assumption. Under these assumptions, the functional H simplifies to the following expression:

$$H(Q^0, \hat{Q}^0, q, \hat{q}) = \frac{p+1}{2} \operatorname{Tr}\left[Q^0\hat{Q}^0\right] - \frac{p(p+1)}{2} \operatorname{Tr}\left(q\hat{q}\right) + \ln I + \alpha \ln J. \tag{5.19}$$

To proceed with the replica trick we introduced earlier, it remains to compute explicit expressions for the terms I and J and these expressions need to be analytical in the replica index p. This is necessary for us to eventually take the limit $p \to 0^+$ by an analytic continuation.

To this end, we make use of the following identity: for any symmetric positive-definite matrix $M \in \mathbb{R}^{K\times K}$ and any vector $x \in \mathbb{R}^K$, we have

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

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

When we apply this identity, the expressions for I and J under the replica symmetric ansatz become:

$$I = \int_{\mathbb{R}^K} \mathcal{D}\xi \left(\int_{\mathbb{R}^K} dw\, P_0(w)\, \exp\left(-\frac{1}{2}w^\intercal(\hat{Q}^0 + \hat{q})w + \xi^\intercal \hat{q}^{1/2} w\right)\right)^{p+1}, \tag{5.20}$$

$$J = \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi \left(\int_{\mathbb{R}^K} dZ\, P_{\text{out}}\left(y \,\middle|\, (Q^0 - q)^{1/2} Z + q^{1/2}\xi\right)\right)^{p+1}. \tag{5.21}$$

We would like to ensure consistency with the normalization of the partition function, to this end, our assumptions must satisfy the condition:

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

and this does reflect the fact that $\mathbb{E}[\mathcal{Z}_n^0] = 1$ by definition.

In the limit $p \to 0^+$, the expressions for I and J simplify significantly. Specifically, we find that $J = 1$, and:

$$I = \int_{\mathbb{R}^K} dw\, P_0(w) \exp\left(-\frac{1}{2} w^\intercal \hat{Q}^0 w\right).$$

This expression must be equal to 1 in the $p \to 0^+$ limit, which implies that the optimal value of the conjugate overlap is $\hat{Q}^0 = 0$. Consequently, the diagonal overlap Q^0 satisfies:

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

i.e., it is given by the second moment matrix of the prior distribution.

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

(Take derivatives to obtain the saddle point eqns)

$$\lim_{n\to\infty} f_n = \mathrm{extr}_{q,\hat{q}} \left\{ -\frac{1}{2}\mathrm{Tr}[q\hat{q}] + I_P + \alpha I_C \right\}, \tag{5.22}$$

$$I_P \equiv \int_{\mathbb{R}^K} \mathcal{D}\xi \int_{\mathbb{R}^K} dw^0\, P_0(w^0)\, \exp\left(-\frac{1}{2}(w^0)^\intercal \hat{q} w^0 + \xi^\intercal \hat{q}^{1/2} w^0\right) \times \ln\left(\int_{\mathbb{R}^K} dw\, P_0(w) \exp\left(-\frac{1}{2} w^\intercal \hat{q} w + \xi^\intercal \hat{q}^{1/2} w\right)\right), \tag{5.23}$$

$$I_C \equiv \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi \int_{\mathbb{R}^K} \mathcal{D}Z^0\, P_{\mathrm{out}}\left(y \mid (Q^0 - q)^{1/2} Z^0 + q^{1/2}\xi\right) \times \ln\left(\int_{\mathbb{R}^K} \mathcal{D}Z\, P_{\mathrm{out}}\left(y \mid (Q^0 - q)^{1/2} Z + q^{1/2}\xi\right)\right). \tag{5.24}$$

5.2 Derivation of Saddle point Equations

We now derive the state evolution equations by extremizing the replica free entropy functional. Recall that in the limit $p \to 0^+$, the replica computation yields the variational expression

$$\lim_{n\to\infty} f_n = \mathrm{extr}_{q,\hat{q}} \left\{ -\frac{1}{2}\mathrm{Tr}(q\hat{q}) + I_P(q,\hat{q}) + \alpha I_C(q) \right\}, \tag{5.25}$$

where $q, \hat{q} \in \mathbb{R}^{K\times K}$ are the overlap and conjugate overlap matrices and to derive the saddle-point equations, we take functional derivatives with respect to q and $\hat{q}$ and set them to zero.

5.2.0.1 Update Equation for the Overlap Matrix q

To compute the update equation for the overlap matrix q, we differentiate the prior contribution I_P with respect to the conjugate matrix $\hat{q}$ using the expression in Equation (5.24) and we obtain:

$$\frac{\partial I_P}{\partial \hat{q}} = -\frac{1}{2}\mathbb{E}_{\xi,w_0}\left[w_0 w_0^\top\right] + \mathbb{E}_\xi\left[\hat{w}(\xi)\hat{w}(\xi)^\top\right], \tag{5.26}$$

where the posterior mean estimator is defined as

$$\hat{w}(\xi) := \mathbb{E}_{P(w|\xi)}[w] = \frac{\int_{\mathbb{R}^K} dw\, P_0(w)\, \exp\left(-\frac{1}{2}w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w\right) w}{\int_{\mathbb{R}^K} dw\, P_0(w)\, \exp\left(-\frac{1}{2}w^\top \hat{q} w + \xi^\top \hat{q}^{1/2} w\right)}. \tag{5.27}$$

The extremization condition $q = \partial I_P/\partial \hat{q}$ then yields

$$q = \mathbb{E}_\xi\left[\hat{w}(\xi)\hat{w}(\xi)^\top\right] - \frac{1}{2}\mathbb{E}_{\xi,w_0}\left[w_0 w_0^\top\right]. \tag{5.28}$$

However, in our setting, the teacher weights are fixed and do not interact with the student replicas, the second term does not contribute to the dynamics of our inference process. It can therefore be neglected, leading to the simplified state evolution update:

$$q = \mathbb{E}_\xi\left[\hat{w}(\xi)\hat{w}(\xi)^\top\right]. \tag{5.29}$$

This expression captures the overlap between two independent samples from the posterior distribution over student weights, conditioned on a shared realization of the latent Gaussian variable $\xi \sim \mathcal{N}(0, \mathbf{I}_K)$.

5.2.0.2 Update Equation for the Conjugate Overlap Matrix $\hat{q}$

To obtain the update equation for the conjugate overlap matrix $\hat{q}$, we consider the extremization condition on the replica free entropy functional. Specifically, we compute the gradient of the objective

$$\frac{\partial}{\partial q}\left[-\frac{1}{2}\operatorname{Tr}(q\hat{q}) + \alpha I_C(q)\right] = 0, \tag{5.30}$$

which yields the saddle-point equation

$$\hat{q} = \alpha\, \frac{\partial I_C(q)}{\partial q}. \tag{5.31}$$

Using our previously derived expression for the channel contribution in equation **??**, we compute its derivative with respect to the overlap matrix q. This involves differentiating

a log-integral expression that depends on a nonlinearly transformed Gaussian variable. This is easily done by applying the chain rule under the integral sign and employing matrix calculus.

Let us denote the effective pre-activation variable as

$$z := (Q_0 - q)^{1/2} Z + q^{1/2} \xi, \tag{5.32}$$

where $Z \sim \mathcal{N}(0, \mathbf{I}_K)$ and $\xi \sim \mathcal{N}(0, \mathbf{I}_K)$. The gradient can then be expressed through derivatives with respect to ξ, which captures the dependence of the effective field on the overlap matrix q, the update for $\hat{q}$ becomes:

$$\hat{q} = \alpha \, \mathbb{E}_{\xi,y} \left[\nabla_\xi \log Z_y(\xi) \cdot \nabla_\xi \log Z_y(\xi)^\top \right], \tag{5.33}$$

where we define the normalizing integral (or partition function) as

$$Z_y(\xi) := \int_{\mathbb{R}^K} \mathcal{D}Z \, P_{\text{out}} \left(y \,\middle|\, z = (Q_0 - q)^{1/2} Z + q^{1/2} \xi \right). \tag{5.34}$$

The expression for $\hat{q}$ thus measures the sensitivity of the marginal output likelihood with respect to perturbations in the latent direction ξ. It can be viewed as an effective Fisher information matrix that controls how strongly the observation model constrains the posterior in the latent space.

5.2.0.3 Final Saddle point Equations

The update rules for the order parameters q and $\hat{q}$ define the state evolution equations. At each iteration t, the update steps are:

$$\hat{w}^{(t)}(\xi) = \mathbb{E}_{P(w|\xi)}[w], \tag{5.35}$$

$$q^{(t+1)} = \mathbb{E}_\xi \left[\hat{w}^{(t)}(\xi) \hat{w}^{(t)}(\xi)^\top \right], \tag{5.36}$$

$$\hat{q}^{(t+1)} = \alpha \, \mathbb{E}_{y,\xi} \left[\nabla_\xi \log Z_y^{(t)}(\xi) \cdot \nabla_\xi \log Z_y^{(t)}(\xi)^\top \right]. \tag{5.37}$$

These updates form a closed set of state evolution equations, which characterize the macroscopic dynamics of the inference process under Bayes-optimal inference in the thermodynamic limit. These equations can be iterated starting from an initial condition, such as $q^{(0)} = 0$ and capture the evolution of the student weight correlations as the learning process goes on.

The Generalization Error

The *Bayes-optimal generalization error* is defined as:

$$\epsilon_g^{\text{Bayes}} \equiv \frac{1}{2}\,\mathbb{E}_{W^*,X}\left(\left(\langle\varphi_{\text{out}}(XW)\rangle - \varphi_{\text{out}}(XW^*)\right)^2\right), \tag{5.38}$$

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

Employing the Nishimori identity, we can show that:

$$\begin{aligned}\epsilon_g^{\text{Bayes}} &= \frac{1}{2}\mathbb{E}_{X,W^*}\left[\varphi_{\text{out}}\left(XW^*\right)^2\right] + \frac{1}{2}\mathbb{E}_{X,W^*}\left[\langle\varphi_{\text{out}}\left(XW\right)\rangle^2\right]\mathbb{E}_{X,W^*}\left\langle\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right\rangle, \\ &= \frac{1}{2}\mathbb{E}_{X,W^*}\left[\varphi_{\text{out}}\left(XW^*\right)^2\right] - \frac{1}{2}\mathbb{E}_{X,W^*}\left\langle\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right\rangle.\end{aligned}$$

Moreover, it is important to observe that the distribution of the input matrix X is rotationally invariant. As a result, the quantity

$$\mathbb{E}_X\left[\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right],$$

depends only on the scalar *overlap* $q \equiv W^\intercal W^*$ between the student and teacher weight vectors. Furthermore, it is expected to concentrate around the optimal value q^* that extremizes the replica free entropy. Consequently, the generalization error can be expressed as a deterministic function of the overlap q^*, making it directly computable from the solution of the replica variational problem.

The generalization error for $K = 4$

We depart from the $K \to \infty$ limit and return to our finite case of interest with $K = 4$. For this specific setting, the generalization error—as defined in (5.25 and 5.26) can be computed explicitly.

We can denote the student weight matrix as W with it explicitly being:

$$W = [w_1, w_2, w_3, w_4] \in \mathbb{R}^{n\times 4},$$

where each column $w_l \in \mathbb{R}^n$ corresponds to the weights of the l-th hidden unit in the student network. The *overlap matrix* $q \in \mathbb{R}^{4\times 4}$ is defined entry-wise by

$$q_{ij} = \frac{1}{n}\,\mathbb{E}_{w\sim P(w|X,Y)}\left[w_i^\top w_j\right],$$

and captures the normalized expected scalar product between the i-th and j-th student weight vectors under the posterior distribution.

In the case where the output neuron combines the hidden unit activations with *non-uniform weights*, the output function takes the form

$$\varphi_{\text{out}}(z) = \text{sign}\left(\sum_{l=1}^{4} a_l \,\text{sign}(z_l)\right),$$

with arbitrary coefficients $a_l \in \mathbb{R}$. This breaks the symmetry among the hidden units, and therefore we can no longer assume that the overlaps q_{ij} are identical for $i \neq j$, nor that the diagonal entries q_{ii} are shared.

Consequently, the full overlap matrix must be treated in its most general symmetric form:

$$q = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{12} & q_{22} & q_{23} & q_{24} \\ q_{13} & q_{23} & q_{33} & q_{34} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{pmatrix},$$

which contains 10 independent entries due to symmetry and each element q_{ii} represents the *self-overlap* of the i-th hidden unit, while each off-diagonal term q_{ij} with $i \neq j$ represents the *cross-overlap* between the corresponding pair of hidden units.

In the general $K = 4$ case with arbitrary output weights $v_l \in [v_1, v_2, v_3, v_4]$, the Bayes-optimal generalization error is given by the expectation

$$\varepsilon_g = \frac{1}{2}\,\mathbb{E}_{(u,u^*)\sim\mathcal{N}(0,\Sigma)}\left[\varphi_{\text{out}}(u) - \varphi_{\text{out}}(u^*)\right]^2,$$

where $\varphi_{\text{out}}(z) = \text{sign}\left(\sum_{l=1}^{4} v_l \,\text{sign}(z_l)\right)$, and Σ is the 8×8 covariance matrix

$$\Sigma = \begin{pmatrix} q^* & r^\top \\ r & q \end{pmatrix}.$$

Equivalently, the generalization error can be written as

$$\frac{1}{2} - 2\varepsilon_g^{\text{Bayes},K=4} = \int_{\mathbb{R}^8} \mathcal{D}x\, \varphi_{\text{out}}(u^*(x_{1:4})) \cdot \varphi_{\text{out}}(u(x_{1:8})),$$

where $u^*, u \in \mathbb{R}^4$ are constructed as affine transformations of independent standard Gaussian variables $x_1, \ldots, x_8$, such that their joint distribution matches the target covariance Σ.

Chapter 6

Conclusion and Future work

Summarize contributions
Revisit research questions and answers
Suggest directions for further research

Appendix A

Appendix A

Bibliography

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

[2] 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.

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

[4] 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.

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

[6] 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.

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.