UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics of Data

Final Dissertation

Information Theoretic Analysis of Deep Neural Networks

Internal thesis supervisor

Dr. Michele Allegra

External thesis supervisor

Prof. Jean Barbier

External thesis co-supervisors

Dr. Francesco Camilli

Dr. Daria Tieplova

Candidate

Eleonora Bergamin

Academic Year 2023/2024

Abstract

The task of learning an objective function that characterizes a deep, non-linear neural network is tackled, focusing on training the complete set of network parameters. Our investigation is conducted within a scenario where the number of samples, input dimension, and network width are all notably large. The neural networks under study operate in a teacher-student framework, where the data generated by the teacher network are classified by a student network with an identical architecture. Our main goal is to carry out an information-theoretical analysis of deep neural networks, building upon established results on two-layer networks. Recent conjectures, followed by partial rigorous proofs, show that it is possible to reduce two-layer networks to simpler one-layer networks, commonly referred to as generalized linear models. Remarkably, fundamental information-theoretic quantities such as the mutual information between training data and teacher network weights, as well as the Bayes-optimal generalization error, are well-understood for such simplified networks. Consequently, our strategy involves extending this reduction using a recursive argument. This involves progressively simplifying the network by replacing the last two layers with an equivalent one-layer neural network. The recursion continues until we identify an equivalent one-layer model for the entire network. This recursive approach is expected to provide us with a comprehensive understanding of the network's behavior and performance.

Contents

List of symbols

a, A	generic quantity, a scalar if not specified otherwise.
$\mathbf{a}$	column vector.
$\mathbf{A}$	matrix or column vector.
$\mathbf{A}^\intercal$	transpose of the matrix or vector A.
$a_i = (\mathbf{a})_i$	i^{th} component of $\mathbf{a}$.
A_{ij}	element at the i^{th} row and j^{th} column of $\mathbf{A}$.
$\hat{a}$	estimate of the quantity a.
$:=$	equal by definition.
$a \mid b$	a given b.
I_d	d-dimensional identity matrix.
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\delta(x)$	Dirac delta function, that is formally a distribution, it is a probability density that assigns infinite weight to the single value $x = 0$.
δ_{ij}	Kronecker delta, it is equal to 1 if $i = j$, and 0 otherwise.
$\mathbf{x}^\intercal\mathbf{y} = \mathbf{x}\mathbf{y}^\intercal := \sum_i^n x_i y_i$	scalar product between two vectors of dimension n.
$\|\mathbf{x}\|^2 = \mathbf{x}^\intercal\mathbf{x} = \sum_i^n x_i^2$	L_2 norm of the vector $\mathbf{x}$ of dimension n.
$\mathbb{I}(A)$	indicator function, which is 1 if the condition A is true, 0 otherwise.

$\mathbb{E}f(A) := \mathbb{E}[f(A)]$	expectation of the function f of the random variable A.
$\mathbb{E}_W A := \mathbb{E}_W(A)$	expectation of the expression A with respect to the random variable W.
$\mathbb{E}A := \mathbb{E}(A)$	expectation of the expression A with respect to all the random variables on which it depends.
$\mathbb{E}(A)^2 := \mathbb{E}[(A)^2]$	expectation of the square of the random variable A.
$\mathbb{E}^2(A) := (\mathbb{E}(A))^2$	square of the expectation of the random variable A.
$\mathbb{E}_{\mathcal{N}(0,a^2)}g = \mathbb{E}g(aZ)$	expectation of the function g evaluated in aZ, with a a positive real value and $Z \sim \mathcal{N}(0,1)$.
$\mathbb{V}_{\setminus W}(A) = \mathbb{E}_{\setminus W}(A - \mathbb{E}_{\setminus W}A)^2$	variance of the random variable A with respect to all the random variables on which A depends, except from the random variable W.
$\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^\intercal]$	cross-covariance matrix between vectors $\mathbf{X}$ and $\mathbf{Y}$, which can be of different or same dimensions, including dimension one.
$f = O(g)$	given the functions f and g, $\lvert f\rvert \leq K\lvert g\rvert$ for some absolute constant K.
$d\mathbf{x} := \prod_i^n dx_i$	integration over all components of the vector $\mathbf{x}$ of dimension n.
$D\mathbf{x} := \prod_i^n Dx_i = \prod_i^n dx_i P(x_i)$	integration over all components of the size n vector $\mathbf{x}$ where each component follows a standard Gaussian distribution. In other words, this notation is used when $x_i \sim \mathcal{N}(0,1)$.
$\mathbf{X} \circ \mathbf{Y} := \mathbf{Z}$ where $Z_{ij} = X_{ij}Y_{ij}$	Hadamard (component-wise) product between matrices or vectors of the same size.
$\partial_x f = \partial/\partial x f$	partial derivative of f, that is a function of multiple variables, with respect to x.
f'	derivative of f with respect to its only argument.
$\nabla f(\mathbf{x}) := ((\partial/\partial x_1)f(\mathbf{x}), \cdots, (\partial/\partial x_n)f(\mathbf{x}))^\intercal$	gradient of $f(\mathbf{x})$, which is the vector of partial derivatives of $f(\mathbf{x})$ with respect to each of the n components of $\mathbf{x}$.
$\lvert\mathbf{x}\rvert$	number of components of $\mathbf{x}$ or the cardinality of a set.

Chapter 1

Introduction

Machine learning [1, 2, 3] is a field within artificial intelligence that focuses on recognizing patterns and extracting information from data. Specifically, it refers to the methods and techniques that enable automated learning from data. This automation allows for predictions to be made when new data is introduced, without the need for explicit programming tailored to each specific task. This form of implicit learning, heavily dependent on the data available, has led to machine learning models being characterized as data-driven. The ultimate goal of a learning algorithm is to achieve generalization. This means it should be capable of extracting information from the samples or examples it analyzes and applying this knowledge to new, unseen data.

The recent advancements in technology and the availability of computational resources have demonstrated the success of these algorithms in learning from data. Consequently, machine learning has become pervasive in today's world. Its applications are extensive and diverse, spanning from personalized recommendations in e-commerce and predictive diagnostics in healthcare to algorithmic trading in finance and route optimization in transportation.

A class of algorithms that has gained significant attention in recent years is defined by neural networks (NNs). These models were initially introduced to describe as a biologically plausible model of computation. Today, they have become a crucial area of study within machine learning, owing to their generalization capabilities and efficiency. Despite the successful applications of neural networks, the pace of understanding their underlying mechanisms and explaining the reasons of their success has not kept up. Due to this, learning algorithms have often been exploited in various fields where the primary focus is on the output rather than the underlying process or the explainability of the results. To address this gap in understanding, a new field within computer science has emerged, known as explainable artificial intelligence (xAI) [4].

To explain the behavior of neural networks, however, also statistical mechanics, and specifically physics of disordered systems, plays a crucial role. In statistical physics, disordered systems [5, 6] are those in which the constituent particles are not arranged in a regular, repeating pattern. These systems, which include glasses and spin glasses, are characterized by metastable states and are described by complex energy landscapes. A neural network can be viewed as a disordered system [7, 6,

8] where the neurons are the constituent particles and the synaptic weights represent the disorder. The state of the network is determined by the activation of the neurons, and the energy of the state is given by the loss function. The training of the network, which involves adjusting the weights to minimize the loss function, can be seen as a process of navigating the complex energy landscape of this disordered system.

Notably, the mapping from a machine learning problem to a statistical mechanics framework has opened up new avenues for the application of techniques commonly used in the statistics of disordered systems [9, 10, 11, 12, 13, 14, 15, 16]. The replica method, mean-field approaches, and interpolation method are examples of techniques rooted in mathematical physics and physics of disordered systems that are now applied to machine learning [11, 17, 13, 18, 19]. Specifically, disordered systems methods are particularly useful in studying the performance of machine learning models, their algorithmic limits, and thresholds [20, 14]. Moreover, the concept of phase transitions, which is central to the study of disordered systems in statistical physics, can be applied to neural networks, thus playing a crucial role in understanding their behavior [14, 15] and the relevant scalings. For instance, the transition from a phase where the network can memorize the training data to a phase where it can generalize to unseen data can be studied using the tools of statistical physics [21, 15].

Simple models such as perceptrons, generalized linear models, and committee machines have been studied for more than thirty years, demonstrating the success of physics in studying such models [10, 12, 11, 22, 14, 23, 21, 24, 25, 17, 26, 15]. Therefore, the perspective of statistical mechanics provides a powerful framework for understanding the structure, dynamics, and limitations of neural networks, and contributes to the ongoing efforts to make these models more interpretable and robust.

With this foundation laid, the focus is now directed towards studying the capabilities of neural networks and exploring the fundamental limits in their performance within the statistical mechanics framework. A significant challenge in understanding neural networks lies in the dependence of the learning process on various interacting factors, each with intricate individual effects. These factors include the architecture of the network, the structure inherent to the datasets on which the network is trained, and the algorithms and optimization procedures employed.

To address this complex nature of neural network learning, a teacher-student setup [10, 11, 12, 20] within a Bayes-optimal framework can be exploited, where a student network classifies the data generated by the teacher network with an identical, fully connected architecture.

This framework allows for the disentangling of contributions from different aspects of the learning process, primarily focusing on the network's architecture and the impact of available data on prediction performance. Indeed, in this teacher-student setup, the results obtained are not confined to a specific learning procedure. Instead, they represent the fundamental limits of the network's performance, leading to what are known as information-theoretic bounds. This approach effectively separates the impact of the training procedure from the analysis.

In the context of this Bayes-optimal setup, recent advancements in the analysis of deep neural networks (DNNs) using a statistical mechanics framework have been substantial [27, 28, 29, 19]. The initial analysis by [27] focused on linear networks, excluding nonlinearities. Specifically, a technique

was introduced to evaluate the generalization error of deep linear networks with finite width when these networks are trained on a fixed set of data. Building on this, [28] expanded the analysis by formulating conjectures for the generalization error in empirical risk minimization. Further progress was made by [29], who employed a Gaussian equivalence principle (GEP) for both shallow and deep nonlinear neural networks to obtain Bayes-optimal limits. The GEPs are grounded in a well-known concept in high-dimensional probability: appropriately rescaled low-dimensional projections of high-dimensional vectors with weakly correlated components exhibit Gaussian behavior.

Subsequently, the recent analysis by [19] rigorously proved the Bayes-optimal limits obtained by [29] for a 2-layer neural network. Specifically, [19] established information-theoretic limits in an overparametrized regime, where the dataset size, input dimension, and network width are all large, eventually approaching infinity. The results were presented as a bound connecting the mutual information between the teacher network weights and the training data for both a 2-layer neural network and a generalized linear model (GLM) [30, 31, 32], which is a one-layer neural network and a generalization of a perceptron [33]. By examining the conditions under which this bound tends to zero, it becomes possible to identify the scaling regimes, in terms of numerosity of data available and the number of neurons, in which the 2-layer neural network effectively reduces to the GLM, making the two models equivalent. The importance of this reduction lies in the fact that the mutual information between the dataset and the teacher network weights is linked to the network's generalization error, and these quantities have been extensively studied for GLMs [10, 12, 11, 22, 34, 35, 14]. Consequently, the GLM can serve as a simpler proxy for studying more complex neural networks, providing valuable insights into their behavior.

This thesis is rooted in the results obtained in [19]. An information-theoretic analysis of a deep neural network is conducted here, addressing the task of learning an objective function for a deep, nonlinear neural network. Our investigation focuses on training the complete set of network parameters within a scenario where the size of the dataset, input dimension, and network width are all large. A teacher-student setup in a Bayes-optimal framework is utilized, allowing for the estimation of information-theoretic bounds.

The primary goal is to extend the analysis of deep neural networks by generalizing the results from [19] to networks with an arbitrary number of layers L, thereby establishing bounds in terms of information theory, and specifically mutual information, that relate a deep nonlinear network to a generalized linear model. Once the bounds are established, the objective is to explore how the sizes of the dataset, input dimension, and hidden layers influence these bounds. Specifically, the aim is to identify the scaling conditions under which the bounds tend to zero as all parameters approach infinity, thus allowing the reduction of the deep neural network to a generalized linear model.

The main contribution of this thesis are the bounds presented in 4.2. These bounds allow to find the different scalings of network layers and dataset size such that the deep neural network can be mapped into a generalized linear model. Moreover, the parametrization of the GLM found through the reduction is consistent with the result provided by [29]. An additional finding pertains to the concentration of measure phenomenon. Specifically, it is observed that Gaussian random vectors, when processed through layers of a neural network, retain Gaussian-like statistical properties, des-

pite no longer being Gaussian. Initially derived as a technical result to complete the proofs for the information-theoretic bounds, this finding stands as an intriguing mathematical property in its own right, independent of the proof for the bounds.

The strategy to derive the bounds in terms of mutual information involves extending the reduction performed in [19] using a recursive argument. This method progressively simplifies the network by replacing the last two layers with an equivalent one-layer neural network. The recursion continues until an equivalent one-layer model is identified for the entire network, effectively reducing the full network to a generalized linear model, whose properties are known.

Chapter 2

Framework and background

In this chapter, the foundational framework and background of our investigation are examined, focusing on the intersection of several key areas: statistical inference, information theory, statistical mechanics, Bayesian inference, and machine learning. These fields provide the theoretical basis and methodologies essential for the investigation. Specifically, the aim of the chapter is to define the specific quantities relevant to the models under study, establishing a clear understanding of their behavior and properties.

2.1 Information theory

The objective of our investigation is to conduct an analysis based on information theory [2, 36, 37]. Information theory is concerned with the concepts of surprise and the amount of information that can be associated with observations. It specifically enables the study of data compression and transmission properties. The central study that attracted interest in this topic was conducted by Claude Shannon, culminating in his most renowned paper published in 1948. The paper, titled "A Mathematical Theory of Communication" [38], is often referred to as the foundation of the field information theory field. In the context of our study, the focus is on comprehending the information that the data provides about the model parameters and quantifying this relationship.

2.1.1 Information

Information is a quantity that represents the knowledge gained once the outcome of a random experiment becomes known. Conversely, it can also be thought of as the uncertainty associated with the revelation of the outcome. Intuitively, information is acquired from an observer if that information was not previously known.

The aim here is to formalize this concept with a focus on information for events, that are tied to their probabilities [39]. This forms the mathematical foundation for constructing the information, or surprisal. The object needed to start the discussion is a probability space $\mathcal{S}$. A probability space

$\mathcal{S} = (\Omega, \mathcal{F}, \mathbb{P})$ is a triple where

- Ω corresponds to the sample space, and contains all the possible outcomes;
- $\mathcal{F}$ is the σ-algebra of the events, meaning that it is a family of subsets of Ω closed under countable union, complement, and Ω belongs to it. Any element of the σ-algebra is an event $E \in \mathcal{F}$;
- $\mathbb{P}$ is a probability measure that goes from the σ-algebra $\mathcal{F}$ to $[0, 1]$ such that $\mathbb{P}(\Omega) = 1$ and it satisfies the σ-additivity.

Intuitively, events with a probability equal to one are expected to provide no information, as these events are certain to occur. Additionally, events that happen more frequently should convey less information than rare events: in the latter case, the surprise is significant when the event is realized. Another aspect that needs to be considered is the conditionality of the events. If two events, denoted as E_1 and E_2, are independent, then

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2) \qquad \text{implying} \qquad \mathbb{P}(E_1|E_2) = \mathbb{P}(E_1), \qquad \mathbb{P}(E_2|E_1) = \mathbb{P}(E_2) \tag{2.1}$$

When the events are independent, the information of the intersection of the events is desired to be the sum of the information gained from E_1 and E_2 respectively. These considerations translate into requirements that information must satisfy.

- $I : \mathcal{F} \to \mathbb{R}^+$ is a function f of the probability of the events: $I = f \circ \mathbb{P}$, namely $I(E \in \mathcal{F}) = f(\mathbb{P}(E))$;
- f is a monotonically decreasing function of $\mathbb{P}$;
- For independent events E_1 and E_2, $f(\mathbb{P}(E_1 \cap E_2)) = f(\mathbb{P}(E_1)\mathbb{P}(E_2)) = f(\mathbb{P}(E_1)) + f(\mathbb{P}(E_2))$.

This leads to the following definition of information, also called information content or surprisal:

$$f(x) = -\alpha \ln(x) \qquad I : \mathcal{F} \to \mathbb{R}^+,\ I(E) = -\alpha \ln(\mathbb{P}(E)) \tag{2.2}$$

where α is chosen arbitrarily. If $\alpha = 1$ the information is computed in terms of "nats" or natural units. Another common choice is $\alpha = \log_2 e$, yielding $I(E) = -\log_2(\mathbb{P}(E))$. In this case, the information content is expressed in terms of "bits" or binary digits, where one bit quantifies the information gained when the outcome of a random experiment with two equiprobable outcomes is revealed.

2.1.2 Entropy

The definition of information content associated with an event was introduced, describing it as the information gained about the outcome of a random experiment. This raises the question of what amount of information is required to fully describe the outcome. The problem is approached by

considering the average information needed to specify the outcome. Since the average is taken over the outcomes, this quantity can be interpreted as the overall average amount of information associated with the probability space $\mathcal{S}$.

In the case of a discrete probability space, every outcome can be associated with an event $E = \{\omega\} \in \mathcal{F}$, and the average information needed to specify the outcome of the random process is defined as follows:

$$H(\mathcal{S}) := \langle I \rangle = -\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) \ln \mathbb{P}(\{\omega\}) \tag{2.3}$$

where this quantity $H(\mathcal{S})$ takes the name of Shannon entropy.

Based on the explanation provided, the definition of entropy can be generalized to random variables.

Let E be a set, ξ a σ-algebra of subsets of E, given a probability space $(\Omega, \mathcal{F}, P)$. Then, the function $X : \Omega \to E$ is a random variable if for any $A \in \xi$, $X^{-1}(A) \in \mathcal{F}$. Defining

$$P_X : \xi \to [0, 1], \qquad P_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) \tag{2.4}$$

it can be verified that this is a probability on the measurable space (E, ξ), called distribution of X.

In the context of real-valued random variables, the sample space is $E = \mathbb{R}$, and the considered σ-algebra is called Borel σ-algebra $\xi = \mathcal{B}(\mathbb{R})$. Elements of this Borel σ-algebra are known as Borel sets. Additionally, a continuous random variable X has a probability density function p_X if its distribution is absolutely continuous with respect to the Lebesgue measure, namely, if there exists a function p_X such that

$$P_X(A) = \mathbb{P}(X \in A) = \int_A p_X(x) dx \qquad \forall A \in \mathcal{B}(\mathbb{R}) \tag{2.5}$$

where dx represents the Lebesgue measure on $\mathbb{R}$.

In general, the notation $dP(x) =: P_X(dx)$ is used in order to be able to consider in the analysis also the cases in which the probability density function cannot be defined. When the density can be defined, there is the actual correspondence $dP(x) = p_X(x)dx$.

If X is discrete, the distribution of it is defined as the following map:

$$P_X : E \to [0, 1], \qquad P_X(x) = \mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) \tag{2.6}$$

Through this construction, the entropy associated to the discrete random variable X can be introduced. To simplify the notation, the elements of the set E, corresponding to the outcomes, are denoted as $(x_1, x_2, \ldots, x_{|E|})$, allowing the probability $P_i := P_X(x_i)$ to be associated with each outcome for $i = 1, ...|E|$:

$$H(X) := -\sum_{x \in E} P_X(x) \ln P_X(x) = -\sum_{i=1}^{|E|} P_i \ln P_i \tag{2.7}$$

This quantity describes the average information needed to specify the outcome of the random variable. Its interpretation is of crucial importance, as entropy can be considered as a measure of the lack of information, or ignorance, about the outcome of X. Being an expectation, higher entropy indicates greater unpredictability of the outcome on average. This implies little knowledge before the experiment, with substantial information gained once the outcome is revealed.

Entropy thus serves a dual role as a measure of information and uncertainty or lack of information. As a measure of ignorance, $H(X)$ represents the amount of knowledge missing to determine the outcome of X on average before observing it. In this scenario, no experiment or observation has taken place, so entropy is interpreted as a measure of uncertainty associated with a process that never occurred. From the other perspective, $H(X)$ is also the amount of information that on average is gained once the outcome is observed. In this case, the process has happened and the outcome can be observed, making entropy the expected information content associated with the random variable.

Properties of the entropy. Some useful properties of the entropy of a discrete random variable X, that will be exploited in our study, are now remarked.

- $H(X) \geq 0$ with $H(X) = 0$ if and only if there exists an index i such that $P_i = 1$. This implies that a deterministic quantity or random variable has no uncertainty, and provides no information;
- $H(X)$ is maximized if the probability over the outcomes is uniform. Maximising $H(X)$ with respect to p_i under the constraint $\sum_i P_i = 1$, leads to $P_i = \frac{1}{|E|}$, and $H(X) = \ln |E|$. Since the entropy is maximized, $H(X) \leq \ln |E|$ also holds true;
- $H(X)$ is concave with respect to the distribution of X. Consider two random variables X_1, X_2 with values in the same set E. Define $P_i := P_{X_1}(x_i)$ and $Q_i := Q_{X_2}(x_i)$. A new random variable Z can be constructed, with values in the same space as X_1 and X_2. Let $P_Z(x_i) = \lambda P_i + (1-\lambda)Q_i$. This means that $Z = X_1$ with probability λ, and $Z = X_2$ with probability $1 - \lambda$. Then,

  $$H(Z) = H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q) = \lambda H(X) + (1-\lambda)H(Y) \quad (2.8)$$

  The interpretation of the concavity of the Shannon entropy lies in the fact that a distribution defined as the mix of two probability distributions has a higher entropy than the sum of the entropies: this implies that mixing two probability distributions increases entropy.

2.1.3 Joint entropy and conditional entropy

Consider now two different random variables X and Y that take values $\{x_i\}, \{y_j\}$ in the sets E_X, E_Y. The joint entropy of the random variables can be defined as the Shannon entropy of the joint distribution $P_{XY}(x, y)$.

$$H(X,Y) := H(P_{XY}) = -\sum_{(x,y)\in E_X \times E_Y} P_{XY}(x,y) \ln P_{XY}(x,y) \quad (2.9)$$

The conditional entropy of X given Y is introduced next. This quantity corresponds to the expected information gained by evaluating the outcome of X given that the outcome of Y is known. Equivalently, this can be read from the perspective of the uncertainty interpretation of the entropy: the conditional entropy represents the remaining lack of knowledge about X given that the variable Y has been observed.

$$H(X|Y) := -\sum_{(x,y)\in E_X\times E_Y} P_Y(y)P_{X|Y}(x|y)\ln P_{X|Y}(x|y) \tag{2.10}$$

Properties of the conditional entropy. Some useful considerations about these quantities that will be exploited in the analysis are presented:

- Similarly to the construction made for the information content for events, the joint entropy of two independent random variables is just the sum of the entropy of the two variables. This implies that learning the outcome of the two variables together does not provide additional information compared to learning the outcomes separately. Furthermore, since the variables are independent, the knowledge of the outcome of one variable does not affect the knowledge about the second variable. As a consequence, the conditional and unconditional entropy are the same. Formally, since the two variables are independent,

  $$P_{XY}(x,y) = P_X(x)P_Y(y) \longrightarrow P_{X|Y}(x|y) = P_X(x), \qquad P_{Y|X}(y|x) = P_Y(y) \tag{2.11}$$

  and it follows that

  $$H(X,Y) = H(X) + H(Y) \longrightarrow H(X|Y) = H(X), \qquad H(Y|X) = H(Y) \tag{2.12}$$

- The joint entropy of two random variables cannot exceed the sum of the entropies associated with the two random variables. This implies that the information gained by knowing the outcomes of both variables simultaneously cannot be greater than the information obtained by revealing the outcomes separately:

  $$H(X,Y) \leq H(X) + H(Y) \tag{2.13}$$

- The joint entropy of a pair of random variables corresponds to the information gained on average by revealing the outcomes of the two random variables simultaneously, and this is equivalent to the sum of the information gained revealing the outcome of the first variable and the information generated from knowing the outcome of the second one provided that the outcome of the first one is known. What stated can be interpreted as a chain rule applied to entropy:

  $$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) \tag{2.14}$$

- Consider a function f. Then, applying it to a random variable yields $H(f(X)|X) = 0$. Exploiting (2.14) leads to $H(X) + H(f(X)|X) = H(f(X)) + H(X|f(X))$, implying:

$$H(f(X)) \leq H(X) \tag{2.15}$$

  The equality holds only if f is invertible.

2.1.4 Kullback-Leibler divergence

Another key quantity in our framework is the Kullback-Leibler divergence D_{KL}, or relative entropy.

Consider two discrete random variables X and Y taking values in the same sample space E, with discrete distributions P and Q. The KL-divergence from Q to P is defined as

$$D_{KL}(X||Y) := D_{KL}(P||Q) = \sum_{x \in E} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \tag{2.16}$$

This quantity satisfies Gibbs' inequality, meaning that

$$D_{KL}(P||Q) \geq 0 \tag{2.17}$$

where the equality holds if and only if the two distributions are the same. It is worth noting that the Kullback-Leibler divergence is not symmetric, namely, $D_{KL}(P||Q) \neq D_{KL}(P||Q)$, implying that this object is not a distance. Nonetheless, the Kullback-Leibler diverge D_{KL} can be interpreted as a measure of discriminability between two distributions, or a distance in the space of distributions. Moreover, the Kullback-Leibler divergence is convex with respect to both arguments, meaning that considering the probability densities P_1, P_2, P_1, Q_2 on the same space E it holds

$$D_{KL}(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_{KL}(P_1||Q_1) + (1-\lambda)D_{KL}(P_2||Q_2) \tag{2.18}$$

The definition of Kullback-Leibler divergence can easily be adapted to the case of continuous densities replacing the sum with an integral:

$$D_{KL}(P||Q) = \int_{x \in E} dx\, P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \tag{2.19}$$

2.1.5 Mutual information

Another crucial quantity is the mutual information between two random variables X and Y. Its aim is to measure the information that one variable provides about the other one. Specifically, when dealing with the probability distributions of the random variables, the mutual information can be understood as the Kullback-Leibler divergence between the joint probability distribution of the random variables and the product of the marginal probabilities. It can be interpreted then as a measure of how much the joint distribution is actually comparable to the product, thus indicating how much the two variables

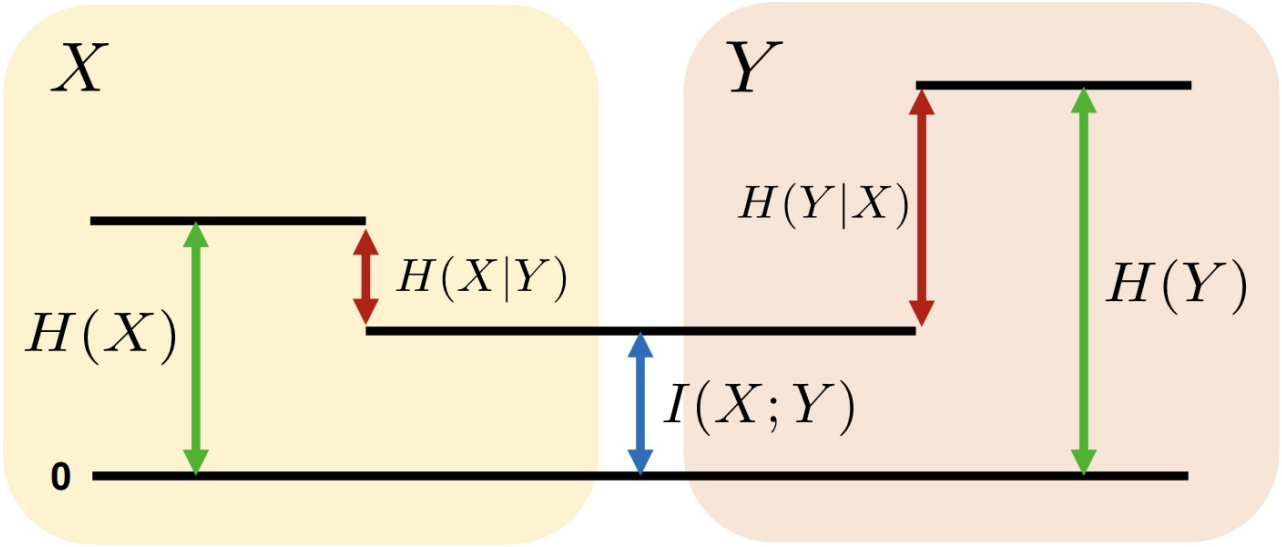

Figure 2.1: The total information required to specify X can be divided into the information provided by Y, denoted as $I(X;Y)$ and the residual information $H(X|Y)$. This relationship is reciprocal, applying similarly when specifying Y in terms of X.

are uncorrelated. Indeed, if two random variables are independent, the divergence is zero. Expressing it formally:

$$\begin{aligned} I(X;Y) := D_{KL}(P_{XY} \| P_X p_Y) &= \sum_{(x,y)\in E_X \times E_Y} P_{XY}(x,y) \Big(\ln \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} \Big) \\ &= -H(X,Y) + H(X) + H(Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \geq 0 \end{aligned} \tag{2.20}$$

where the last two equalities can be obtained using (2.14). From the definition, it can be observed that if X and Y are independent, namely $P_{XY}(x,y) = P_X(x)P_Y(y)$, then $I(X;Y) = 0$. It is also worth noticing that the mutual information is symmetric with respect to the exchange of the variables X and Y, aligning with the interpretation that the mutual information gives an estimate of the information that a variable provides about the other.

Properties of the mutual information. Other useful properties of the mutual information are the following:

- Similarly to what stated for the entropy, considering two functions f_1 and f_2 it holds that

$$I(f_1(X); f_2(Y)) \leq I(X;Y) \tag{2.21}$$

i.e., applying a function to the random variables cannot increase the information they provide one about the other;

- The mutual information satisfies the data processing inequality. This inequality states that the processing of the data does not lead to an increase in information. Such a concept can be formalized in the following statement. Let X, Y, Z be three random variables such that Z depends only on Y. Then $I(X;Z) \leq I(X;Y)$;

- A chain rule for mutual information can be introduced:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \tag{2.22}$$

based on the definition of conditional entropy and conditional mutual information. The conditional mutual information is defined as:

$$\begin{aligned} I(X;Y|Z) &:= \mathbb{E}_Z D_{KL}(P_{XY|Z} \| P_{X|Z} P_{Y|Z}) \\ &= \sum_{(x,y,z) \in E_X \times E_Y \times E_Z} P_{XYZ}(x,y,z) \Big(\ln \frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z) P_{Y|Z}(y|z)} \Big) \\ &= -H(X,Y|Z) + H(X|Z) + H(Y|Z) \\ &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \geq 0 \end{aligned} \tag{2.23}$$

The concept of entropy can be generalized and adapted to continuous random variables with some care. Indeed, substituting the discrete probability densities by the probability densities and the sums by integrals leads to the definition of differential entropy:

$$h(X) = -\int_{x \in E} dx \, P_X(x) \ln P_X(x) \tag{2.24}$$

with $P_X(x)$ being the probability density of the random variable X. In the same fashion, the conditional entropy and the mutual information can be generalized to the continuous case.

If a resolution ϵ is used to discretize the space, the entropy and the differential entropy differ in the following way:

$$H_\epsilon(X) = h(X) + \ln \frac{1}{\epsilon} + O(\epsilon) \tag{2.25}$$

and the analogous holds for the conditional entropy of X given Y.

What illustrated implies that the differential form of the mutual information, being a difference of entropies as shown in (2.20), does not depend on the resolution. Hence, its differential expression coincides with the definition already given of mutual information. This implies that the mutual information maintains its properties when moving from a discrete to a continuous setting. Moreover, also its definition in terms of Kullback-Leibler divergence can be preserved, exploiting the definition of the latter in the case of continuous densities in (2.19).

2.2 Statistical mechanics

In our analysis, the language of statistical mechanics [40, 41, 42, 43, 37] is borrowed, and connections to it are drawn.

Statistical mechanics is one of the most important branches of modern physics. In this discipline, concepts and tools are borrowed from probability theory, with the aim of describing many-body systems. Specifically, statistical mechanics was born in the 19^{th} century to understand the link between the macroscopic observables associated with a physical system, such as pressure, temperature or magnetization, and its microscopic quantities. Indeed, the dimension of the system makes it impossible to give a description using only microscopic properties: the systems of interest in statistical mechanics are typically characterized by a high number of degrees of freedom, representing particles or more generally components or constituents of the system. Furthermore, the degrees of freedom can interact: the simplest type of interaction is pairwise interaction, but other possibilities exist.

One of the advantages deriving from the study of statistical mechanics is the possibility of analyzing a physical system as a whole. This holistic approach opens the way to the study of a wide range of phenomena, referred to as emergent phenomena, that could not be understood solely by examining the individual components of the system. Indeed, under the appropriate conditions and hypotheses, the system reacts globally to perturbations, yielding results that differ from those obtained when only a single component is affected. These emergent phenomena and collective behaviors are observable across various models and fields of physics, ranging from disordered systems to active matter. A prime example of this can be found in living systems, such as flocks of birds or schools of fish. When a predator approaches, the behavior of the animals changes, not necessarily because they perceive the threat directly, but because they sense the behavioral change in their neighbors [44, 45]. Such change in behavior and how this occurs is what statistical mechanics aims to study.

In general, these types of emergent phenomena are described as phase transition of the system, which correspond to changes in the macroscopic properties of the whole system when an external parameter changes. In particular, quantities called order parameters are used to describe phase transitions: they are macroscopic observables that are zero above the critical point and non-zero before. Near phase transitions, systems exhibit critical behavior, which is characterized by slow relaxation after perturbation and the generation of spatially correlated fluctuations. These fluctuations do not remain local but propagate throughout the system, implying that distant points become highly correlated. This is common to both inanimate and living systems, and it is indeed what happens for example in the case of the flock of birds, which reacts cohesively to the movement of the predator. Importantly, phase transitions and thus the associated emergent phenomena typically depend only on the dimensionality of the system, its symmetries, and the type of interactions present rather than specific microscopic details. This concept is called universality in statistical physics and enables the study of diverse systems using simplified models known as null models, which focus on essential aspects while neglecting specific details.

Concerning then the dynamics of the system, statistical mechanics studies two types of systems, equilibrium and out of equilibrium systems, with equilibrium systems being the primary focus of the investigation. In the second case, non-reversible processes characterized by imbalances are usually

modeled. In equilibrium systems, on the other hand, there are no macroscopic flows or currents, and macroscopic thermodynamic variables suffice to describe the system's state.

Quantities and terminology are now introduced in order to provide a more precise description of these states. The system is conceptualized as a microstate, representing a point in the phase space of microscopic degrees of freedom. These microstates, also called configurations, are drawn from an ensemble: this represents an ideal set of multiple copies of the system, generated according to specific criteria, and therefore characterized by constraints and a probability distribution on the microstates. The number of instances in the ensemble is infinite, and each configuration represents a possible state of the system. The constraints imposed on the system can be of two types: hard and soft.

- Hard constraints force the ensemble to follow the prescription, therefore all instances of the ensemble exhibit the same characteristic. These constraints characterize microcanonical ensembles, where each state has the same probability of being drawn. Such conditions can be interpreted as exact conservation laws, which are satisfied for all times. For example, considering an isolated system, energy $\mathcal{E}$, volume V, and number of particles N remain fixed;
- Soft constraints represent observables that are fixed on average. This implies that microstates have a non-uniform probability of being observed, resulting in fluctuations within the system's characteristics. The presence of soft constraints is associated to canonical ensembles.
  A typical example of this type of constraint is called thermal bath. In this case, the system of interest is immersed in a much larger one called environment; they can exchange energy, but not volume and particles. The purpose of the environment is to allow the system to maintain a constant temperature. Thus, the total energy is conserved, but that of the system under study is free to fluctuate around an average value.

In terms of probability theory introduced earlier, a microstate or configuration corresponds to the vector of outcomes associated with a random vector $\mathbf{X} \in \mathcal{X}^N$ describing the state of the system, and each random variable X_i represents a component or degree of freedom, of which there are N in total. A configuration or microstate is denoted by $\mathcal{C} = (x_1, \dots, x_N)$, where each x_i represents a possible outcome associated with the corresponding degree of freedom. $\mathcal{X}^N$ represents then the set of possible system configurations. Slightly generalizing what was previously done, a distribution $P_{\mathbf{X}}(d\mathbf{x})$ can be introduced for the random vector. The distribution can also depend on other external parameters $\mathbf{y}$, which can be random but whose randomness is fixed, hence they are called quenched variables. These can be, for example, the coupling parameters between the degrees of freedom. Each microstate is then characterized by an energy $\mathcal{H}(\mathbf{x}; \mathbf{y})$, with $\mathcal{H}$ the Hamiltonian describing the configuration.

In our investigation the interest lies in soft constraints, and specifically in fixing the average energy of the system. Formally, this implies

$$\mathcal{E} = \langle \mathcal{H}(\mathbf{X}; \mathbf{y}) \rangle := \int_{\mathbf{x} \in \mathcal{X}^N} dP_{\mathbf{X}}(\mathbf{x}) \mathcal{H}(\mathbf{x}; \mathbf{y}) \tag{2.26}$$

The equilibrium distribution is then obtained by exploiting the maximum entropy principle. This principle is a widely used method in statistical mechanics and information theory to find the probability distribution that best fits the imposed constraints, such as conservation laws, normalizations, or system-related measures. By maximizing the Shannon entropy, this principle ensures that the resulting distribution contains enough information to meet the constraints without adding extra assumptions or biases. In particular, if there is no control over a variable, the principle of maximum entropy dictates that all its values are considered equivalent and equally probable, thereby maximizing ignorance with respect to this quantity. This approach removes any potential bias beyond the given constraints, ensuring the most unbiased representation of the system's state. Practically, this method is applied by exploiting the Lagrange multiplier technique, associating a multiplier to each constraint.

In our case, the equilibrium distribution of the system is obtained using the constraint (2.26) along with the normalization condition $\int_{\mathbf{x}\in\mathcal{X}^N} dP_{\mathbf{X}}(\mathbf{x}) = 1$. The resulting distribution is called the Gibbs-Boltzmann distribution:

$$dP_N(\mathbf{x}) = dP_N(\mathbf{x};\mathbf{y}) = \frac{e^{-\beta\mathcal{H}(\mathbf{x};\mathbf{y})}}{\mathcal{Z}(\mathbf{y})} dP_{\mathbf{X}}(\mathbf{x}) \tag{2.27}$$

The normalization constant or partition function is defined as

$$\mathcal{Z}(\mathbf{y}) = \int_{\mathbf{x}\in\mathcal{X}^N} e^{-\beta\mathcal{H}(\mathbf{x};\mathbf{y})} dP_{\mathbf{X}}(\mathbf{x}) \tag{2.28}$$

In the expressions above, $\beta = \frac{1}{T}$, and can be considered as the Lagrange multiplier associated with the energy constraint, for the adjustment of randomness in the system. For instance, the limit $\beta \to 0^+$, corresponds to not imposing any constraint on the average energy, and maintaining the only constraint that the sum of the probabilities gives one. This leads all microstates to be equally probable. This leads back to the discussion of a microcanonical ensemble, in which all states are equally probable as they are characterized by exactly the same values of the variables that define it.

$$\lim_{\beta\to 0_+} dP_N(\mathbf{x}) = \frac{dP_{\mathbf{X}}(\mathbf{x})}{\mathcal{Z}(\mathbf{y})} = \frac{dP_{\mathbf{X}}(\mathbf{x})}{|\Omega(\mathcal{X}^N)|} \tag{2.29}$$

where $\Omega(\mathcal{X}^N)$ is used to indicate the volume of the configuration space. Conversely, for $\beta \to +\infty$, only states minimizing the system's Hamiltonian have non-zero probability, and the states are defined as $\mathbf{x}^* \in \{\text{argmin}_{\mathbf{x}'}\mathcal{H}(\mathbf{x}';\mathbf{y})\}$. In this case, the system is in a minimum energy state or ground state, with no possibility of changing configuration or microstate.

$$\lim_{\beta\to+\infty} dP_N(\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{y})} \sum_{\mathbf{x}^*\in\{\text{argmin}_{\mathbf{x}'}\mathcal{H}(\mathbf{x}';\mathbf{y})\}} \delta(\mathbf{x}-\mathbf{x}^*) dP_{\mathbf{X}}(\mathbf{x}) \tag{2.30}$$

The system can therefore be characterized by multiple phases, and the behavior of the system, as seen, can change between them.

From the partition function, fundamental quantities that can be defined are the free energy and the free energy density:

$$F(\mathbf{y}) := -\frac{1}{\beta}\ln \mathcal{Z}(\mathbf{y}) \qquad f_N(\mathbf{y}) := -\frac{1}{N\beta}\ln \mathcal{Z}(\mathbf{y}) \tag{2.31}$$

These quantities are fundamental for studying phase transitions, which are defined by singularities in the derivatives of the free energy (and therefore also of the partition function) with respect to the external parameters, including for example beta, namely the temperature, magnetic or electric field, and pressure. Furthermore, from the partition function, or equivalently from the free energy, other quantities of interest for the system can be obtained, such as its energy or magnetization. However, in finite systems, the partition function, being a sum (or integral) of exponentials and thus an analytical function, lacks discontinuities. Therefore, the regime of interest is that of the thermodynamic limit, where the system's volume and number of particles tends to infinity:

$$f(\mathbf{y}) := -\lim_{N\to+\infty} \frac{1}{N\beta}\ln \mathcal{Z}(\mathbf{y}) \tag{2.32}$$

Furthermore, the so-called quenched free energy and quenched free energy density can be defined: they are obtained by taking the average over all the possible realizations of the parameters $\mathbf{y}$. The quenched averages are denoted with the overbar:

$$\bar{f}_N := \mathbb{E}\, f_N(\mathbf{y}) = -\frac{1}{N\beta}\mathbb{E}\ln \mathcal{Z}(\mathbf{Y}) \qquad \bar{f} := -\lim_{N\to+\infty} \frac{1}{N\beta}\mathbb{E}\ln \mathcal{Z}(\mathbf{Y}) \tag{2.33}$$

Having introduced this, it is now possible to consider another desired aspect of the free energy density: its concentration around the average. This property is also called self-averaging property, and it is particularly desirable as it implies that in the thermodynamic limit as the system size N approaches infinity, the value of the free energy density becomes independent of the specific realization of the external parameters $\mathbf{y}$. This means that the value of such a quantity depends only on the statistical properties of $\mathbf{y}$, not on its particular realization.

It is worth noting that this self-averaging property is desirable, yet there exist models where this property is not observed, and consequently, the free energy density does not concentrate. From a mathematical standpoint, the self-averaging property of the free energy density can correspond to different types of convergence of the free energy density to its expectation. In some models, the convergence to the mean is in probability:

$$\lim_{N\to+\infty} \mathbb{P}(|f_N(\boldsymbol{Y}) - \bar{f}_N| > \epsilon) = 0\,, \qquad \forall\, \epsilon \in \mathbb{R} > 0 \tag{2.34}$$

For some other systems, the convergence occurs almost surely or in a L_2 sense, namely:

$$\lim_{N\to+\infty} \mathbb{E}[(f_N(\boldsymbol{Y}) - \bar{f}_N)^2] = 0 \tag{2.35}$$

This is the case for example of the Sherrington-Kirkpatrick (SK) model [46], the prototypical spin glass [5, 6]. Additionally, the model studied in [19], upon which our investigation is based, also partially exhibits this type of convergence.

Spin glasses, in which our analysis is rooted, are the paradigmatic systems in statistical mechanics where quenched random variables, in the form of disorder, appear. These disordered magnetic systems exhibit complex behaviors due to random interactions between their constituents, called spins. Among these models, the SK model stands out as the most important, representing a cornerstone and seminal contribution to the field of glassy physics. The SK model is a mean-field version of the Edwards-Anderson spin glass model [47]. Initially introduced as a straightforward solvable model, it later revealed a structure that was far more complex and rich, particularly concerning the quenched free energy density. The quenched free energy density is a crucial concept in the study of spin glasses, describing the average free energy of the system taken over all possible realizations of the disorder.

Sherrington and Kirkpatrick's initial solution for the free energy density of the SK model was not well-defined in the thermodynamic limit, leading to a negative entropy and an unphysical solution. Giorgio Parisi addressed this issue with his replica symmetry breaking solution, obtained using the replica method [48, 49], ensuring the existence of the thermodynamic limit for the quenched density of free energy. Formulating a proof for the Parisi formula for the free energy of the SK model, however, remained an open problem in the field for almost three decades.

The proof of the Parisi formula came in two different steps. First Francesco Guerra proved a uniform bound with his ground breaking replica symmetry breaking uniform bound [50], based on the interpolation technique previously introduced with Toninelli which allowed to prove the existence of the limit in the first place [51]. Then, Michel Talagrand [52] managed to find a matching converse bound with remarkable technical effort. Dmitry Panchenko [53] later simplified the proof proving a connection between the ultrametricity conjecture by Parisi to the Ghirlanda-Guerra identities [54]. Since the introduction of the SK model, the study of spin glasses has seen remarkable growth, evolving into a flourishing area of research. The insights gained from this field have not only deepened our understanding of disordered magnetic systems but have also influenced various other domains, including optimization, and machine learning [20, 14].

2.3 Statistical and Bayesian inference

Statistical inference, especially within the Bayesian framework, serves as a fundamental cornerstone to our investigation. Statistical inference can be viewed as the process of extracting information from data. It involves using the data to make estimations, predictions, or to test hypotheses, providing a framework for modeling and understanding the randomness and variability in observations. Complementing this, Bayesian inference introduces a probabilistic perspective to inference, allowing to incorporate prior knowledge and uncertainty into the model considered. Together, these methodologies form the foundation of our analysis.

2.3.1 Statistical inference

Statistical inference [36, 55, 56, 57] plays a key role when studying modeling problems. When tackling a modeling problem, two primary frameworks may emerge: forward problems and inverse problems. In general, these problems involve data or observations $\mathbf{y}$ and a model or process that generates the data. The model can be thought of as a map from the set of input parameters $\mathbf{x}$, also known as signals, to the data. Additional parameters $\boldsymbol{\theta}$ of the model, referred to as hypotheses, may be present.

In the context of forward problems, both the model and hypotheses are known, and data is generated accordingly. The primary aim is usually to study the properties of the generated data. Forward problems frequently arise in probability theory, a branch of mathematics that, within this context, aims to quantify the likelihood of the occurrence of the event given a specific model. In these problems, the event has not yet taken place, and the goal is to predict future outcomes, which means predicting the data from the model.

Conversely, inverse problems involve having the data or observations, and the task revolves around estimating the parameters or the model that can describe the system. Inverse probability problems and statistical inference can also be viewed as inverse problems. Specifically, in statistical inference, the task is to use data or observations to determine properties of the signal, effectively reconstructing the model or hypotheses that explain the data. In this context, the main objects of interest are probability distributions: the aim is usually to reconstruct the conditional probability distribution of the model given the observed data. Typically, inverse problems are more complex compared to forward problems; obtaining observations from a model is generally simpler than the task of reconstructing the model itself.

Concerning now the applications of statistical inference, its primary objectives are prediction and estimation. In the case of prediction, the aim is to determine a model that can accurately respond to new data, assuming the new data and the training data are statistically equivalent. This does not necessarily mean that the estimator of the model that is found will be identical to the true model. Instead, the trait of interest is the predictive ability of the model. Specifically, when the model is applied to a new signal, the goal is for it to make an accurate prediction of the data.

The focus of our study is on the estimation feature of statistical inference. In this context, statistical inference can be employed to understand the relationship between the parameters themselves or between the data and the parameters.

2.3.2 Bayesian inference

Bayesian inference [36, 55, 56, 57], a powerful statistical method, has recently become the basis of many modern computational and analytical techniques. Its applications span a wide range of fields, from artificial intelligence and signal processing to bioinformatics and social sciences, demonstrating its versatility and robustness.

At its core, Bayesian inference is a method of statistical analysis that integrates prior knowledge about a phenomenon into a mathematical model. This approach also provides a systematic way to

update beliefs based on new data and new observations, making it a dynamic and adaptable tool for understanding complex systems. Furthermore, it enables the incorporation of existing knowledge about the system or phenomenon being studied, or the hypotheses imposed on it.

Despite its significant strengths, Bayesian inference has its limitations. In fact, it requires careful consideration of the a priori information and hypotheses that are incorporated into the model. If chosen inappropriately, they can alter the estimates obtained with this method. Nevertheless, its ability to incorporate previous knowledge and new data, make it an adaptable and flexible method, and therefore of particular interest in our discussion.

The notation that will be used is the same as 2.3.1, denoting the data $\mathbf{y}$, the signal $\mathbf{x}$, and any additional system parameters $\boldsymbol{\theta}$. Again, the model that generates the data can be considered as a map of the signal and the additional parameters $f(\mathbf{x}, \boldsymbol{\theta})$.

The basis of Bayesian inference and Bayesian statistics is Bayes' formula:

$$dP_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) = \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) dP_{\mathbf{X}}(\mathbf{x})}{P_{\mathbf{Y}}(\mathbf{y})} = \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) dP_{\mathbf{X}}(\mathbf{x})}{\int P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \bar{\mathbf{x}}) dP_{\mathbf{X}}(\bar{\mathbf{x}})} \tag{2.36}$$

In this expression, each component has a specific meaning:

- Likelihood function ($P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})$): describes the probability of observing the data obtained given the parameters. It should be emphasized that likelihood is not a probability distribution over the parameters; it is instead a function of the parameters for a given set of data, which are fixed and are what has been measured. Notice that the likelihood is a probability density: this is needed for the posterior to be well defined;
- Prior distribution ($P_{\mathbf{X}}(d\mathbf{x})$): represents the initial assumptions about the parameters before observing any data. The prior is data-independent and does not change once the data is collected;
- Posterior ($P_{\mathbf{X}|\mathbf{Y}}(d\mathbf{x} \mid \mathbf{y})$): is the updated belief about the parameters after looking at the data. It is calculated using Bayes' formula, combining the prior belief and evidence from the data;
- Evidence ($P_{\mathbf{Y}}(\mathbf{y})$): represents the probability of obtaining the observed data. The term then measures how well the model predicts the data, averaging over all possible parameter values. It acts as a normalization constant to ensure that the posterior distribution is a properly defined probability distribution.

This process of updating beliefs is iterative: as more data is collected, the posterior can be updated, which then becomes the new prior for the next round of data collection. This iterative process facilitates the continuous refinement of beliefs as more data becomes available.

When the model is defined also by the additional parameters or hypotheses $\boldsymbol{\theta}$, they can be included in the formula as follows:

$$dP_{\mathbf{X}|\mathbf{Y},\boldsymbol{\Theta}}(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}) = \frac{P_{\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta}}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) dP_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta})}{P_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y} \mid \boldsymbol{\theta})} = \frac{P_{\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta}}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) dP_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta})}{\int P_{\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta}}(\mathbf{y} \mid \bar{\mathbf{x}}, \boldsymbol{\theta}) dP_{\mathbf{X}|\boldsymbol{\Theta}}(\bar{\mathbf{x}} \mid \boldsymbol{\theta})} \tag{2.37}$$

Having introduced this framework, the expectation with respect to the posterior or posterior mean of a function g, denoted using the Gibbs brackets $\langle\cdot\rangle$, can be defined as

$$\langle g(\mathbf{X})\rangle := \mathbb{E}[g(\mathbf{X})|\mathbf{y},\boldsymbol{\theta}] = \int dP_{\mathbf{X}|\mathbf{Y},\boldsymbol{\Theta}}(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})g(\mathbf{x}) \tag{2.38}$$

Notice how this expectation still depends on the data $\mathbf{y}$, which corresponds to quenched random variables using the statistical mechanics terminology introduced in 2.2.

When considering the Bayesian statistics framework, two common scenarios emerge. Traditional statistical analysis often deals with situations where the number of parameters to be inferred is significantly less than the number of data points. In such scenarios, the prior often plays a marginal role and the likelihood represents the dominant term due to the abundance of data. This determines a posterior very similar to likelihood. Conversely, high-dimensional inference scenarios occur when both the number of parameters and the volume of data are large, even tending towards infinity, posing non-trivial inference challenges. This scenario often arises in data science and machine learning applications, where the amount of data is enormous, but so is the number of model parameters, or neural network weights.

Our analysis will involve scenarios of the second type, necessitating a high-dimensional inference framework.

2.3.3 Bayesian inference as a statistical mechanics problem

Note that statistical physics is mainly applied in the context of forward problems, while Bayesian inference is applied in the context of inverse problems. Indeed, the primary goal of statistical physics is to understand and predict the macroscopic behavior of systems starting from their microscopic laws and constituent parts. This involves determining the emergent properties of a system based on known microscopic interactions and statistical laws, such as predicting the pressure and temperature of a gas given the interactions between its molecules. This aligns with the nature of direct problems, which infer properties of the system given the model or parameters.

On the other hand, Bayesian inference focuses on updating the probability of hypotheses or model parameters based on observed data and can therefore be framed as an inverse problem, as it involves deducing underlying causes or parameters from observed outcomes. Nonetheless, the techniques and methodologies of statistical physics, especially those used to calculate thermodynamic averages over a given disorder, are particularly useful in the context of Bayesian inference. In this framework, the spins represent the parameters to be inferred, and the interactions are replaced by the constraints derived from the observations that these variables must satisfy.

In the following table, a parallel is drawn between the terms of statistical mechanics and the terminology used in Bayesian inference, connecting the two domains.

Statistical Physics	**Bayesian Inference**
Hamiltonian	Cost function
Particles, atoms, spins	System components, parameters
Microstates	All possible measurable parameters
Macrostate	Final estimate of the parameters $\mathbf{x}$
Equilibrium distribution	Posterior distribution
Partition function $\mathcal{Z}(\mathbf{y})$	Probability of the data $P(\mathbf{y})$
External field	Prior distribution
Quenched disorder (external parameters, particles interactions)	Data

Bayesian inference can thus be equivalently formulated as a statistical mechanics problem. This is achieved by recasting the posterior distribution, as defined in equation (2.36), within the framework of statistical mechanics. Specifically, the posterior can be rewritten as a Gibbs-Boltzmann distribution by setting $\beta = 1$ and transforming the multiplication of the likelihood and the prior into the exponential of their logarithm. Consequently, the Hamiltonian corresponding to this distribution is determined as follows:

$$\mathcal{H}(\mathbf{x};\mathbf{y}) = -\ln P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) \tag{2.39}$$

where the first term is related to the prior and the second represents the log-likelihood. The evidence can be also rewritten as a partition function:

$$\mathcal{Z}(\mathbf{y}) = P_{\mathbf{Y}}(\mathbf{y}) = \int P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) dP_{\mathbf{X}}(\mathbf{x}) \tag{2.40}$$

Furthermore, the average free energy density corresponds to the differential entropy of the data (or the Shannon entropy if the data is discrete) divided by the number of parameters, or signals, to be inferred:

$$\bar{f}_N := -\frac{1}{N}\mathbb{E}\ln \mathcal{Z}(\boldsymbol{Y}) = -\frac{1}{N}\int d\mathbf{y}\mathcal{Z}(\mathbf{y})\ln \mathcal{Z}(\mathbf{y}) = -\frac{1}{N}\int d\mathbf{y} P_{\mathbf{Y}}(\mathbf{y})\ln P_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{N}H(\mathbf{Y}) \tag{2.41}$$

It is worth noticing that, in an analogy with (2.33), the external parameters, that are quenched, are in this case represented by the data points obtained. The entropy of the data, or free energy, is a quantity of primary interest as it quantifies the uncertainty linked with the data: it provides a measure of how much information can be extracted from the data given the model. In the context of statistical physics, this is similar to understanding the behavior of a system at equilibrium. The free energy, in this case, provides information on the stability of the system and how it responds to variations of the external parameters. Likewise, in the context of Bayesian inference, the entropy associated with the data helps describe the robustness of the inference. In both cases, however, the objective is to minimize the free energy (or, equivalently, maximize the entropy). This is done by

changing the parameters of the model to better fit the data. In the context of Bayesian inference, this occurs through the process of learning or training the model.

It becomes now evident how statistical physics, Bayesian inference, and information theory are deeply intertwined, dealing with the extraction of information from data and subsequent inference based on this information.

By relating the free energy to the entropy, it is also possible now to link the mutual information to the free energy. The * notation is introduced to differentiate between a sample from the posterior distribution $\mathbf{X} \sim P_{\mathbf{X}|\mathbf{Y}}(\cdot \mid \mathbf{y})$ and the signal to be inferred $\mathbf{X}^* \sim P_{\mathbf{X}}(\cdot)$ which is drawn from the prior distribution and is also referred to as the ground-truth signal. Utilizing this notation, the mutual information between the ground-truth signal and the data is computed as follows:

$$\frac{1}{n} I(\mathbf{X}^*; \mathbf{Y}) = \frac{1}{N} H(\mathbf{Y}) - \frac{1}{N} H(\mathbf{Y}|\mathbf{X}^*) = \bar{f}_N - \frac{1}{N} H(\mathbf{Y}|\mathbf{X}^*). \tag{2.42}$$

In the context of the inference problem, the goal is to recover the parameters given the data $\mathbf{y} = f(\mathbf{x})$, where the map f is unknown. The mutual information is the total information carried by the data minus the remaining uncertainty or lack of knowledge about the data when the signal is known. Therefore, this last contribution can only be attributed to noise. From (2.42) then the mutual information depends on the free energy contribution and a contribution due to the noise.

Usually, the noise is assumed to be different and independent for each data point. Hence, its contribution can be computed as the conditional entropy of one data point given the data multiplied by the number of data points. Considering M data points and an N-dimensional signal, and calling Z_1 the noise associated to the data point Y_1, the following is obtained:

$$H(\mathbf{Y}|\mathbf{X}^*) = MH(Y_1|\mathbf{X}^*) = MH(Z_1) \tag{2.43}$$

Once the noise is modeled so that its entropy can be computed, the main task remaining is computing the total entropy associated with the data, or borrowing the statistical physics language, the free energy.

The setting in which our investigation is performed is called Bayes-optimal setting. Bayes-optimality is a crucial concept in statistical inference, particularly when dealing with high-dimensional problems. In this optimal Bayesian framework, it is assumed that both the prior distribution and the noise distribution are known, which permits the determination of the exact posterior distribution, which can be then fully used. This knowledge allows for a comprehensive analysis of inference problems, while at the same time simplifying many aspects of the analysis.

By operating under Bayesian optimal conditions, the limits of information theory can be identified and phase transitions can be studied. This approach establishes the absolute fundamental limits of inference, independent of any specific algorithm, providing a benchmark for evaluating the performance of practical inference methods. This setup, in which the true posterior is known, enables a focus on the intrinsic properties and capabilities of high-dimensional inference problems.

2.4 Machine learning

Our investigation is set within a machine learning framework, specifically dealing with neural networks. Machine learning [1, 2, 3] is a branch of artificial intelligence whose primary goal is to learn to recognize patterns in data. Specifically, machine learning can be seen as a set of methods or techniques that allows to learn from data in an automated way. This permits to make predictions when new samples are provided, without the algorithm being specifically programmed for the task under consideration. This implicit learning, heavily reliant on available data, leads to machine learning models being described as data-driven. The desired result of a learning algorithm is the ability to generalize, i.e. being able to extract information from the analyzed examples or samples and being able to apply what has been extracted to new, unseen data. This capability is also known as inference. Indeed, the goal is for the model to perform well not just on the training data but also on any new data from the same distribution.

A fundamental aspect of machine learning, closely tied to Bayesian inference, is the incorporation of prior knowledge which influences the learning mechanism. The inclusion of prior knowledge guides the learning process and is crucial for the success of learning algorithms, making them more efficient and effective. This concept is proved also by the no free lunch theorem, which explains that no single learning algorithm is universally superior for all problems. Therefore, exploiting the prior knowledge can improve learning algorithms' performance on specific tasks. Specifically, the greater the prior knowledge is or the stronger the hypotheses are, the easier it is for the algorithm to learn. However, this comes at the expense of its flexibility: it is in fact much more dependent on the constraints and hypotheses introduced.

Considering the different models of learning, various frameworks have been developed in machine learning, each with its own approaches and applications. These include reinforcement learning, unsupervised learning, and supervised learning. Our analysis primarily focuses on the latter.

In the supervised learning setting, both the data and the corresponding predictions, or labels, are accessible. The primary aim is to make accurate predictions for new, unseen data, and to achieve this, the model learns from the labeled examples provided. In particular, the machine learning algorithm has access to:

- Domain set (or instance space) $\mathcal{X}$: set of all possible objects on which to make predictions. $\mathbf{X} \in \mathcal{X}$ is a point or instance of the domain, which is usually (but not necessarily) represented by a vector of numbers or characteristics;
- Label set $\mathcal{Y}$: set of possible labels, which in the binary classification setting can be $\mathcal{Y} = \{0, 1\}$;
- Training set $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), (\mathbf{X}_3, Y_3), \cdots, (\mathbf{X}_n, Y_n)\}$: finite sequence of n labeled elements (for supervised learning). These are domain points (in $\mathcal{X} \times \mathcal{Y}$), and constitute the input to the machine learning algorithm.

The result of the algorithm after the learning process is a prediction rule $\hat{f} : \mathcal{X} \to \mathcal{Y}$. This rule, also referred to as a predictor, hypothesis, or classifier, represents the output of the learning

algorithm. However, the algorithm does not know the data generation model: the instances are generated from a probability distribution over $\mathbf{X}$ and labeled according to a function f. Both of these are not known by the algorithm, which is provided only with the training dataset, where for each data point $\mathbf{X}_\mu$ drawn from the distribution over the domain, $Y_\mu = f(\mathbf{X}_\mu)$. A measure of success, or predictor error, can be defined as the probability that the correct label is not predicted by the algorithm on a random data point generated from the data distribution. This error is also called generalization error, true error or loss function. This measure quantifies how well the learned model performs on new data, reflecting its ability to generalize beyond the training set.

The learning process then involves feeding data into the model, which then adjusts its internal parameters based on this input. This adaptation is typically achieved through an iterative process known as training. In each iteration, the model's predictions are compared to the ground truth labels, and the discrepancies between the predictions and the actual labels are used to adjust the model's parameters. This iterative adjustment continues until the error on the training dataset is sufficiently low.

In the supervised learning framework, typical classes of problems include classification and regression. The two paradigms have different characteristics and are therefore approached and applied differently. Classification tasks are characterized by the goal of assigning input data points to predefined categories or classes. In this case, the model is tasked with discerning patterns in the input features and making decisions about which class a given data point belongs to. In classification, therefore, the output variable is discrete and represents specific categories or labels.

On the other hand, regression tasks involve predicting continuous numerical values based on input characteristics. Unlike classification, where the output is categorical, regression models aim to estimate a continuous outcome. Regression models are primarily concerned with understanding relationships between variables and making predictions that take values on a continuous spectrum.

2.4.1 Neural networks

The class of machine learning algorithms that our study focuses on is neural networks (NNs). Neural networks, or artificial neural networks (ANNs), are computational models inspired by the human brain and human computational systems. Frank Rosenblatt's development of the first machine learning algorithm, the perceptron [33] in 1958 marked the beginning of neural network research. Initial applications and studies on neural networks in the eighties and nineties were influenced by high computational costs and high model complexity, leading to performance that was often inferior to methods such as support vector machines and random forests. In the last two decades, advances in developing deep architectures, increasing network size, and the availability of large amounts of data have significantly improved the performance of neural networks. This has revived interest and enthusiasm for machine learning.

Neural networks can be represented as graphs, where each node corresponds to a neuron and the edges represent the connections between neurons. These networks are designed with topologies that keep computational complexity low, as not all configurations of graphs are computationally feasible.

Additionally, learning arbitrary topologies can be challenging or counterproductive, often leading to unnecessary computational cost. Therefore, neural networks typically consist of multiple layers of neurons stacked on top of each other. In these layers, each neuron receives inputs from the neurons in the preceding layer. This layered structure helps in maintaining a balance between computational feasibility and learning capability.

Over time, neural network models have been refined to perform increasingly complex tasks and to describe biologically plausible computational models. To this last aim, the types of networks studied have mainly been feed-forward or recurrent, featuring different topologies such as fully connected, sparsely connected, or locally connected. Additionally, some networks incorporate modular structures and various types of neurons to better adapt to different tasks and research objectives.

Our analysis is focused on feed-forward fully connected neural networks (FFNNs), also known as dense networks. These networks are used to implement functions. The neurons in these networks are organized into layers, with each neuron connected to every neuron in the previous layer, making the network fully connected. A typical feed-forward neural network consists of the following layers:

- Input Layer: the first layer, where each neuron corresponds to a data point. Each of these is then connected to the neurons of the first hidden layer;
- Hidden Layers: each neuron in these layers receives as input the sum of the outputs from the previous layer's neurons, weighted by the edge weights ($\mathbf{W}$), and applies a scalar function to the linear combination to produce its output; This function, called activation function, is usually nonlinear and is the same for all neurons in the layer;
- Output Layer: the final layer, which applies a readout function to the outputs of the last hidden layer. The nature of this function depends on the type of task the network is designed for, such as regression or classification.

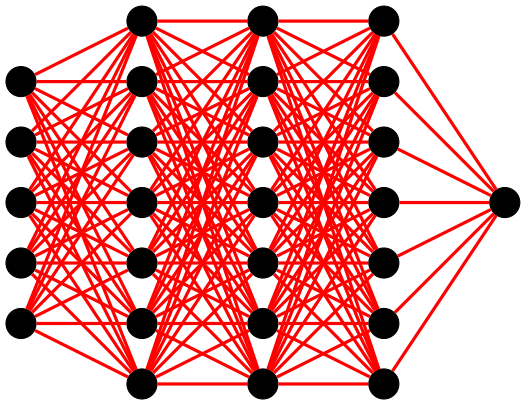

Figure 2.2: Example of fully connected feed-forward neural network architecture with three hidden layers.

The choice of the activation function is fundamental for network learning, as it introduces non-linearity and represents the neuron's firing. Different types of nonlinearities have distinct properties. The most common ones used in literature are highlighted:

- Linear Function: simplifies the network, making a deep architecture equivalent to a shallow one, as a series of linear transformations results in a linear transformation;
- Sigmoid and Hyperbolic Tangent (tanh): differentiable and monotonic but can saturate with large input values, hindering learning and making training difficult. Sigmoid outputs are not zero-centered, causing learning instability;
- Rectified Linear Unit (ReLU): accelerates learning and requires simpler operations but can result in "dead" neurons that stop learning. It makes the learning process lighter, with fewer neurons firing;
- Leaky ReLU: prevents neuron deactivation but loses some model sparsity;
- Exponential Linear Unit (ELU) [58]: allows sparse activations and avoids dead neurons by not outputting zero for negative inputs, although it saturates when inputs are negative.

Selecting the appropriate activation function is essential to optimize network performance and ensure effective learning.

2.4.2 Generalized linear models

The generalized linear model (GLM) [30, 31, 32] is one of the models addressed in our analysis. This algorithm is a foundational concept in machine learning that extends the ideas of a simple perceptron to more complex scenarios. It is analogous to a single-layer neural network and serves as a bridge between basic neural networks and more intricate models. Due to this, the application of GLMs in machine learning has been widely studied from a physics perspective [10, 12, 11, 22, 34, 35, 14].

Generalized linear models represent a generalization of ordinary linear regression. These models can be exploited for both classification and regression tasks, making them particularly versatile. Moreover, their inherent flexibility allows them to accommodate a wide range of data types, thereby making GLMs a powerful tool in the field of statistical data analysis. Specifically, this versatility stems from GLMs' ability to unify multiple statistical models within a single theoretical framework.

In standard regression, the observations $\mathbf{y}$ represent the outcome of a random variable $\mathbf{Y} \in \mathbb{R}^n$ called response. The response has i.i.d. components and its mean is $\boldsymbol{\mu}$. This expectation value is described as a linear combination of a set of known variables, called predictors $\mathbf{X}_i \in \mathbb{R}^m$ through a vector $\boldsymbol{\beta} \in \mathbb{R}^m$ called model parameter. Mathematically, calling $\mathbf{X}$ the matrix of predictors, this is expressed as:

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{2.44}$$

The aim of regression is then to find the parameter $\boldsymbol{\beta}$ that best describes the linear relationship between the predictors and the response variable.

GLMs extend this concept by allowing the linear predictor to be connected to the response variable through a link function. This approach also permits the response variable to be part of a discrete space. As in the regression case, the response variable $\mathbf{Y} \in \mathcal{Y}$ is assumed to consist of independent and identically distributed (i.i.d.) components, with an expected value denoted by $\boldsymbol{\mu}$. The predictors, or independent variables, are represented by $\mathbf{X}_i \in \mathcal{X}$. In the GLM framework, however, the relationship between the expected value of the response $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \boldsymbol{\mu}$ and the predictors $\mathbf{X}_i$ is specified through a link function g. Specifically:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{2.45}$$

where $\boldsymbol{\eta}$ is called the natural parameter and is unknown. The inverse of the link function, g^{-1}, is usually called response function. Moreover, the model assumes:

$$\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta} \sim \text{ExponentialFamily}(\boldsymbol{\eta}) \tag{2.46}$$

The exponential family is defined by the density function:

$$p_{\mathbf{Y}}(\mathbf{y} \mid \boldsymbol{\eta}, \tau) = h(\mathbf{y}, \tau) \exp\left(\frac{\mathbf{b}(\boldsymbol{\eta})^\intercal \mathbf{T}(\mathbf{y}) - \mathbf{A}(\boldsymbol{\eta})}{d(\tau)}\right) \tag{2.47}$$

Here τ is a known parameter and and the functions $h(\mathbf{y}, \tau), \mathbf{b}(\boldsymbol{\eta}), \mathbf{T}(\mathbf{y}), \mathbf{A}(\boldsymbol{\eta}), d(\tau)$ are fixed and define a family of functions parametrized by $\boldsymbol{\eta}$. In the GLM, the aim is again to learn the parameter $\boldsymbol{\beta}$ of the linear combination. This family of functions includes many distributions used in statistics and machine learning, such as normal, exponential, gamma, Bernoulli, Poisson, binomial, categorical and multinomial.

Notably, GLMs are particularly effective for binary classification tasks, where the outcomes can be modeled using a Bernoulli random variable. A Bernoulli random variable can take on two possible outcomes, usually coded as 0 and 1, that can be interpreted as the occurrence or non-occurrence of an event that happens with probability p. The GLM model relates the mean of the response variable to be related to a linear combination of the predictors (input features) through a link function. In the case of a Bernoulli random variable, this function can be derived to be the logit function, $g(p) = \ln \frac{p}{1-p}$. Through this, the probability (or average of the responses) can be computed as $p(\mathbf{x}_i) = \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}_i}}$. This is recognized as a logistic function, and thus the model obtained is a logistic regression model.

Building on the discussed concepts, it becomes now clear how GLM models can be considered a generalization of a perceptron. The perceptron, a fundamental machine learning algorithm, can be considered a one-layer neural network. This model is employed for binary classification, where the response variable $\mathbf{y}$ falls into two categories, typically labeled as $\{0, 1\}$. The perceptron produces predictions $\hat{\mathbf{y}}$ according to the following rule:

$$\hat{\mathbf{y}} = H(\boldsymbol{\theta}^\intercal \mathbf{x}') \tag{2.48}$$

where H is the Heaviside step function and $\mathbf{x}' = (\mathbf{x}, 1)$. The aim of the training then is to learn the parameter $\boldsymbol{\theta}$ such that the input data is correctly classified. It can be observed then that The perceptron employs a linear function to delineate a decision boundary, effectively partitioning the two classes. This boundary can be shifted within the data space by introducing an extra coordinate to the input vector, as done in (2.48). Provided the data points are linearly separable, the perceptron is guaranteed to identify a separating hyperplane.

In the case of GLM models, binary classification can be performed exploiting a logistic regression model. Specifically, the rule that the GLM learns is a logistic function, that represents a softened version of the Heaviside function, thus providing a probabilistic interpretation to the classification task. Indeed, unlike the perceptron that directly assigns class labels based on a threshold, logistic regression estimates the probability of an observation belonging to a particular class. This probabilistic interpretation allows for more nuanced decision-making with respect to the perceptron algorithm, and the incorporation of uncertainty into the classification process. Additionally, GLMs offer a unified framework that extends beyond binary classification to handle various types of response variables, making them more versatile than the perceptron.

Chapter 3

Model and setting

Having provided an overview of the key concepts pertinent to our discussion, the setting and characteristics of our investigation are now addressed.

It is worth recalling that the objective of the investigation is to establish bounds in terms of information theory, and specifically mutual information between the dataset and the teacher network weights, that relate a deep nonlinear network to a shallow network, specifically a generalized linear model. This comparison is significant as this mutual information and the generalization error derived from it have been thoroughly examined for GLMs [10, 12, 11, 22, 34, 35, 14], allowing to use the shallow network as a reference for studying the more complex deep neural network. The models utilized in this study are akin to those introduced in [19], where the equivalence was demonstrated for a 2-layer neural network, and [29], where the conjecture of equivalence was first proposed.

In this chapter, the specific quantities relevant to the studied models are defined and the methods for deriving our results are described.

3.1 Teacher-student setup

The teacher-student setup is developed within the supervised learning framework. As introduced in 2.4, supervised learning is a process where algorithms are trained using labeled data. These labeled datasets are composed of input data (features) and their corresponding output values (labels or responses). For our investigation, the task is set as a regression task, indicating that the output variable is continuous and the model is tasked with predicting numerical values. The training data points are denoted as $\mathbf{X}_\mu^{(0)} \in \mathbb{R}^{d^{(0)}}$ and the labels or responses as responses $Y_\mu \in \mathbb{R}$. Together, they form the dataset $\mathcal{D}_n = \{(\mathbf{X}_\mu^{(0)}, Y_\mu)\}_{\mu=1}^n$, where n is the number of data points.

In the classic teacher-student configuration [10, 11, 12, 20], the training process involves two neural networks: a teacher network and a student network. In this framework, the ground truth labels for the data $\mathbf{X}_\mu^{(0)}$ are produced by the teacher network, whose role is then to guide the learning process of the student network. Indeed the objective of this setup is to mirror a traditional teacher-student relationship, where the teacher imparts knowledge to the student. Here, the student network

learns from a dataset that the teacher network generates, with the ultimate objective of learning a mapping from the input data to the labels that the teacher network generates. This is achieved by the student network by acquiring the knowledge encoded in the teacher network parameters and structure.

It is important to highlight that in the teacher-student setup, there is no requirement for the teacher and student networks to share the same architecture [59, 60], hyperparameters such as the activation function, or any other characteristic. This absence of restrictions opens up a wide array of possibilities for designing both networks. Indeed, the key feature of this framework is the role of the teacher network, which is to serve as a source of ground truth by providing labels for the data. Regardless of any design differences then the student network aims to learn from the generated dataset and approximate the teacher network's behavior, ultimately striving to achieve a comparable level of performance. For instance, the teacher network can be a complex model with multiple layers, nonlinear activations, and other sophisticated architectural elements, and the student network may have a simpler structure than the teacher network. It could for example have fewer layers fewer nodes per layer or less complex activation functions.

This flexibility in design choices allows the teacher-student setup to be applied in various scenarios where the complexity or computational resources required for training a network like the teacher one might be too much to implement in practical applications. The student network can here serve as lighter alternative, representing an approximation of the teacher network while still benefiting from the knowledge transfer facilitated by the teacher network. The opposite scenario, in which the student network is overparametrized with respect to the teacher network, can also be exploited and studied [59]: in this case, the expressive power of the student network is much greater than the one of the generative model.

3.2 Model

Our analysis is set in a Bayesian learning framework particularly focusing on the Bayes-optimal setting introduced in 2.3.3. As previously mentioned, a teacher-student setup is considered. This framework implies that, apart from the true weights, complete knowledge is possessed by the student network.

It is important to consider that the Bayes-optimal setting defines an upper bound for the learning of the network, independently of the training algorithm and learning procedure. Indeed, in this framework, the generative model is known to the student network, therefore having the same architecture as the teacher neural network that generated the data. This configuration inherently leads to the optimal information-theoretic performance attainable by the student.

Specifically, the optimal performance is achieved when the network is trained through Bayesian learning that depends on the exploration and utilization of the posterior distribution of the student's parameters. This means that the student network is not only trying to fit the data but also to understand the underlying distribution that generated it. This is a powerful concept that sets Bayesian learning apart from other machine learning frameworks.

The model analyzed is an $L+1$-layer neural network within a teacher-student setup. The training data is represented as $\mathbf{X}^{(0)} = \{\mathbf{X}_\mu^{(0)}\}_{\mu=1}^n$, where $\mathbf{X}_\mu^{(0)} \in \mathbb{R}^{d^{(0)}}$ and $n \in \mathbb{R}$ is the number of data points. The ground truth labels generated by the teacher networks for the training are denoted as $\mathbf{Y}^{(L)} = \{Y_\mu^{(L)}\}_{\mu=1}^n$. Together, the data and the responses form the dataset $\mathcal{D}_n^{(L)} = \{(\mathbf{X}_\mu^{(0)}, Y_\mu^{(L)})\}_{\mu=1}^n$.

The goal is then to train the student network to adjust its weights to best approximate the relation between $Y_\mu^{(L)}$ and $\boldsymbol{X}_\mu^{(0)}$.

The model for the teacher network is defined as:

$$Y_\mu^{(L)} = f\Big(\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big); \mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu = f\Big(\frac{\mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(L)}}{\sqrt{d^{(L)}}}; \mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu \tag{3.1}$$

with

$$\mathbf{X}_\mu^{(\ell)} = \varphi\Big(\frac{\mathbf{W}^{*(\ell)}\,\mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d^{(\ell-1)}}}\Big) \qquad \forall \ell \in \{1, \dots, L\} \tag{3.2}$$

Similarly to what introduced in 2.3.3, the * notation is used to indicate the ground-truth parameters of the network. The quantities in (3.1) are defined as follows:

- Input data $\mathbf{X}^{(0)} = \{\mathbf{X}_\mu^{(0)}\}_{\mu=1}^n$: $d^{(0)}$-dimensional random vectors drawn independently and identically distributed (i.i.d.) from the standard multivariate Gaussian distribution, $\mathbf{X}_\mu^{(0)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d^{(0)}})$ where $I_{d^{(0)}}$ is the $d^{(0)}$-dimensional identity matrix;
- Number of layers $L+1 \in \mathbb{R}$: the layers are indexed by $\ell \in \{0, \dots, L+1\}$. Layer $\ell = 0$ represents the input layer, with the same dimensionality as the data, layers $\ell = 1, \dots, L$ describe the hidden layers and $\ell = L+1$ is the index used for the output layer. The most used indices will be $\ell \in \{0, \dots, L\}$, since as stated in 2.4 the readout function applied in the last layer is usually different from the nonlinear function applied in the previous layers;
- Dimension $d^{(\ell)} \in \mathbb{R}$ of the representation of the input at layer ℓ, $\mathbf{X}_\mu^{(\ell)}$;
- Activation function $\varphi : \mathbb{R} \to \mathbb{R}$: nonlinear scalar function, applied component-wise to vectors;
- Weights at layer ℓ, $\mathbf{W}^{*(\ell)} \in \mathbb{R}^{d^{(\ell-1)} \times d^{(\ell)}}$: random matrix with i.i.d. entries drawn from a Gaussian distribution, $W_{ij}^{*(\ell)} \sim \mathcal{N}(0,1)$ for every couple of indices i, j;
- readout vector $\mathbf{a}^* \in \mathbb{R}^{d^{(L)}}$: random vector with Gaussian distributed entries, $a_i^* \sim \mathcal{N}(0,1)$ for every index i;
- $\mathbf{A}_\mu \in \mathbb{R}^r$: additional stochasticity that might be included in the readout function f, with its own distribution $P_{\mathbf{A}}(\cdot)$;

- Readout function $f : \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}$: it defines the output layer and can be nonlinear;
- Label noise $Z_\mu \in \mathbb{R}$: is Gaussian noise drawn i.i.d. with variance scaled by a factor $\Delta > 0$, $Z_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

Having defined the aforementioned quantities and model properties, two additional regularity hypotheses are introduced, which will later be required in the proofs. It is important to note that these are technical hypotheses. In contrast, the assumption that the input data and model parameters follow a specific distribution are fundamental features of the model studied. These regularity hypotheses ensure the mathematical rigor and validity of the theoretical results.

H1) The activation function $\varphi : \mathbb{R} \to \mathbb{R}$, is C^3 is Lipschitz continuous with Lipschitz constant $\bar{K}$ (also called $\bar{K}$-Lipschitz continuous). Moreover, φ is odd with its second and third derivatives being bounded;

H2) The readout function f along with its first and second derivatives, is almost surely bounded with respect to the probability measure $P_A(\cdot)$.

Additionally, the output kernel is introduced to express the probability distribution of the output. Given the Gaussian nature of the noise Z_μ, it can be represented in the following way, which eliminates the need to explicitly account for the stochasticity $\mathbf{A}$:

$$P_{\text{out}}(y \mid x) := \int P_A(d\mathbf{A}) \frac{1}{\sqrt{2\pi\Delta}} \exp\Big(-\frac{1}{2\Delta}(y - f(x; \mathbf{A}))^2 \Big) = \int P_A(d\mathbf{A}) P(y|x, \mathbf{A}) \tag{3.3}$$

This implies that the labels are drawn from the distribution:

$$Y_\mu^{(L)} \sim P_{\text{out}}\Big(\cdot \mid \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\, \mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} \Big) \Big) \tag{3.4}$$

The student network then reconstructs the labels as:

$$\hat{Y}_\mu^{(L)} = f\Big(\frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{(L)}\, \mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} \Big); \mathbf{A}_\mu \Big) + \sqrt{\Delta} Z_\mu = f\Big(\frac{\mathbf{a}^\intercal \mathbf{x}_\mu^{(L)}}{\sqrt{d^{(L)}}}; \mathbf{A}_\mu \Big) + \sqrt{\Delta} Z_\mu \tag{3.5}$$

with

$$\mathbf{x}_\mu^{(\ell)} = \varphi\Big(\frac{\mathbf{W}^{(\ell)}\, \mathbf{x}_\mu^{(\ell-1)}}{\sqrt{d^{(\ell-1)}}} \Big) \qquad \forall \ell \in \{1, \dots, L\} \tag{3.6}$$

It is worth remarking that the parameters of the teacher network, denoted as $\boldsymbol{\theta}^{*(L)} = \{\mathbf{a}^*, \mathbf{W}^{*(L)}, \dots, \mathbf{W}^{*(1)}\}$ are all drawn from their prior distribution, hence they are indexed with the * superscript. In contrast, the parameters of the student network $\boldsymbol{\theta}^{(L)} = \{\mathbf{a}, \mathbf{W}^{(L)}, \dots, \mathbf{W}^{(1)}\}$ are learned during the training process. Moreover, the notation $\mathbf{X}_\mu^{(\ell)}$ is used to indicate the teacher representation of the input at the ℓ-th layer, whereas in the student network, it is indicated with $\mathbf{x}_\mu^{(\ell)}$.

The construction presented implies that the student utilizes the same output kernel P_{out} as the teacher network, or equivalently same number of layers, layers widths, readout f, activation φ, label noise variance Δ, and prior law for the weights. The parameters of the student network are drawn from the Bayes-optimal posterior distribution, which describes the most probable configurations of the network parameters given the observed data. Formally, the Bayes theorem can be exploited to recover the Bayes-optimal posterior distribution. To do this, the notation introduced in 2.1.2 is recalled and applied to the current model:

$$dP(\boldsymbol{\theta}^{(L)}|\mathcal{D}_n^{(L)}) = P(d\boldsymbol{\theta}^{(L)}|\mathcal{D}_n^{(L)})\,, \qquad D\boldsymbol{\theta}^{(L)} = P(d\boldsymbol{\theta}^{(L)}) \tag{3.7}$$

Using then Bayes theorem, the posterior can be rewritten in the following form:

$$\begin{aligned}
P(d\boldsymbol{\theta}^{(L)}|\mathcal{D}_n^{(L)}) &= \frac{P(\mathcal{D}_n^{(L)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})}{P(\mathcal{D}_n^{(L)})} \\
&= \frac{P(\mathcal{D}_n^{(L)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})}{\int P(\mathcal{D}_n^{(L)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})} \\
&= \frac{P(\mathbf{Y}^{(L)},\mathbf{X}^{(0)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})}{\int P(\mathbf{Y}^{(L)},\mathbf{X}^{(0)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})} \\
&= \frac{P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(\mathbf{X}^{(0)}|\boldsymbol{\theta}^{(L)})P(\boldsymbol{\theta}^{(L)})}{\int P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(\mathbf{X}^{(0)}|\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})} \\
&= \frac{P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(\mathbf{X}^{(0)})P(d\boldsymbol{\theta}^{(L)})}{\int P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(\mathbf{X}^{(0)})P(d\boldsymbol{\theta}^{(L)})} \\
&= \frac{P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})}{\int P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})} = \frac{P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)})}{\mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)})}
\end{aligned} \tag{3.8}$$

where

$$P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)}) = \prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu^{(L)} \mid \frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{(L)}\,\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big) \tag{3.9}$$

and subsequently the partition function is

$$\mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)}) = P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)}) = \int P(\mathbf{Y}^{(L)}|\mathbf{X}^{(0)},\boldsymbol{\theta}^{(L)})P(d\boldsymbol{\theta}^{(L)}) \tag{3.10}$$

This allows to get to the following expression for the posterior distribution:

$$dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}_n^{(L)}) = \frac{1}{\mathcal{Z}^{(L)}(\mathcal{D}_n)} \prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu^{(L)} \mid \frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{(L)}\,\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big) D\boldsymbol{\theta}^{(L)} \tag{3.11}$$

where

$$\begin{aligned} D\boldsymbol{\theta}^{(L)} &:= \prod_{i=1}^{d^{(L)}} \frac{da_i}{\sqrt{2\pi}} e^{-\frac{a_i^2}{2}} \prod_{i=1}^{d^{(L)}} \prod_{j=1}^{d^{(L-1)}} \frac{dW_{ij}^{(L)}}{\sqrt{2\pi}} e^{-\frac{W_{ij}^{(L)\,2}}{2}} \cdots \prod_{i=1}^{d^{(1)}} \prod_{j=1}^{d^{(0)}} \frac{dW_{ij}^{(1)}}{\sqrt{2\pi}} e^{-\frac{W_{ij}^{(1)\,2}}{2}} \\ &=: D\mathbf{a} D\mathbf{W}^{(L)} \dots D\mathbf{W}^{(1)} = D\mathbf{a} D\mathbf{W}^{(L)} D\boldsymbol{\omega}^{(L-1)} \end{aligned} \tag{3.12}$$

Defining $u_y^{(\ell)}(x) = \log P_{\text{out}}(y^{(\ell)} \mid x^{(\ell)})$ for any $i \in \{0, \dots, L\}$, the partition function can be rewritten as:

$$\mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)}) := \int D\boldsymbol{\theta}^{(L)} \exp\Big(\sum_{\mu=1}^{n} u_{Y_\mu}^{(L)}(s_\mu) \Big) \tag{3.13}$$

where

$$\begin{aligned} s_\mu^{(L)} = s_\mu^{(L)}(\boldsymbol{\theta}^{(L)}, \mathbf{X}_\mu^{(0)}) &:= \frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{(L)}\, \mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} \Big) \\ S_\mu^{(L)} = S_\mu^{(L)}(\boldsymbol{\theta}^{(L)}, \mathbf{X}_\mu^{(0)}) &:= \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\, \mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} \Big) \end{aligned} \tag{3.14}$$

Moreover, the joint law of the data can be expressed in terms of the output kernel as:

$$\begin{aligned} dP(\mathcal{D}_n^{(L)}) &= \prod_{\mu=1}^{n} \Big(\prod_{j=1}^{d^{(0)}} \frac{dX_{\mu j}^{(0)}}{\sqrt{2\pi}} e^{-\frac{X_{\mu j}^{(0)\,2}}{2}} \Big) dY_\mu^{(L)}\, \mathbb{E}_{\boldsymbol{\theta}^{*(L)}} \prod_{\mu=1}^{n} P_{\text{out}}(Y_\mu^{(L)} \mid S_\mu^{(L)}) \\ &=: \prod_{\mu=1}^{n} D\mathbf{X}_\mu^{(0)} dY_\mu^{(L)}\, \mathbb{E}_{\boldsymbol{\theta}^{*(L)}} \exp\Big(\sum_{\mu=1}^{n} u_{Y_\mu}^{(L)}(S_\mu) \Big) \end{aligned} \tag{3.15}$$

Remarkably, the law takes this form and cannot be factorized because, although the inputs $\mathbf{X}_\mu^{(0)}$ are independent, the responses $Y_\mu^{(L)}$ are not. Indeed, the outputs are generated by the same teacher network using the same weights, that thereby introduces correlations among the samples indexed by μ. This then affects the dataset's statistical properties. Moreover, due to the fact that our analysis is performed in a Bayes-optimal framework, the expression for the law of the data can be written in terms of the partition function (3.13):

$$dP(\mathcal{D}_n^{(L)}) = \prod_{\mu=1}^{n} D\mathbf{X}_\mu^{(0)} dY_\mu^{(L)}\, \mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)}) \tag{3.16}$$

From the analysis performed, it can also be observed that the vector $\mathbf{A}_\mu$ in (3.3) serves a role similar to that of a learned parameter. Indeed, within the partition function and the dataset's governing law, $\mathbf{A}_\mu$ is treated analogously to other learned variables. The integration over the probability distribution, or the computation of the expected value for $\mathbf{A}_\mu$, is conducted in the same way as for other

learned parameters. By introducing the output kernel, the need to explicitly express the dependency on $\mathbf{A}_\mu$, the additional stochasticity, is removed. Nonetheless, the underlying principles remain consistent, irrespective of whether $\mathbf{A}_\mu$ is considered a learned parameter.

In the context of optimal Bayesian learning, another important quantity that can be defined for the model is the Bayes-optimal predictor, which corresponds to the the prediction $\hat{Y}_{\text{Bayes}}(\mathbf{X}^{(0)}_{\text{new}})$ for the response associated with a new input test sample. Its relevance lies in the fact that serves as a benchmark for the performance of other predictive models. If a model achieves performance close to the Bayes-optimal predictor, it is considered to be performing well. The Bayes-optimal predictor corresponds to the posterior predictive distribution's mean given the training data:

$$\hat{Y}_{\text{Bayes}}(\mathbf{X}^{(0)}_{\text{new}}) := \mathbb{E}[Y^{(L)}_{\text{new}} \mid \mathcal{D}^{(L)}_n, \mathbf{X}^{(0)}_{\text{new}}] = \int dY\, Y\, P_{\text{out}}\Big(Y \mid \frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{(L)}\mathbf{x}^{(L-1)}_{\text{new}}}{\sqrt{d^{(L-1)}}}\Big)\Big) dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}^{(L)}_n) \tag{3.17}$$

where the posterior predictive distribution is:

$$P(Y|\mathbf{X}^{(0)}_{\text{new}}, \mathcal{D}^{(L)}_n) = \int P(Y \mid \mathbf{X}^{(0)}_{\text{new}}, \boldsymbol{\theta}^{(L)}) dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}^{(L)}_n) \tag{3.18}$$

The posterior predictive distribution [57, 2] describes the distribution of possible new unobserved values given an existing data set. It is determined by assuming a model for the data, and that this is then updated with existing data. The computation of this distribution involves an integration over the full range of the posterior distribution of the model's parameters. This approach diverges from the use of a fixed point estimate for a parameter by incorporating the uncertainty associated with the parameter's value, which is taken into account in the predictions made via the posterior predictive distribution. By accounting for this uncertainty in its predictions, the posterior predictive distribution tends to be broader than a predictive distribution that is based on a fixed point estimate of the parameters.

Moreover, the Bayes-optimal predictor (3.17) allows to compute the generalization error of the network, introduced in 2.4. Specifically, the Bayes-optimal predictor minimizes the generalization error when this is computed using the expected loss, namely:

$$\mathcal{E}^{(L)}_n := \mathbb{E}\big(Y^{(L)}_{\text{new}} - \mathbb{E}[Y^{(L)}_{\text{new}} \mid \mathcal{D}^{(L)}_n, \mathbf{X}^{(0)}_{\text{new}}]\big)^2 \tag{3.19}$$

The terminology of statistical mechanics will also be adopted since as discussed in 2.3.3 Bayesian inference problems can be recast as statistical mechanics problems. The posterior distribution is thus treated as a Boltzmann-Gibbs measure over the network weights, that represent the degrees of freedom. Expectations with respect to the posterior are denoted by Gibbs brackets $\langle\cdot\rangle$. Using this

notation, the posterior mean of a function g reads:

$$\langle g \rangle^{(L)} = \int dP(\boldsymbol{\theta}^{(L)} \mid \mathcal{D}_n^{(L)}) f = \frac{1}{\mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)})} \int D\boldsymbol{\theta}^{(L)} \prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu^{(L)} \mid \frac{\mathbf{a}^\intercal}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{(L)}\, \mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big) g \tag{3.20}$$

Similarly to what done in 2.3.3, to conduct an information-theoretic analysis of the system key quantities of interest are introduced, notably the quenched (conditional) free entropy per sample. This quantity, denoted as $\bar{f}_n^{(L)}$ for the $L+1$-layer neural network, is inherently linked to the Shannon entropy $H(\mathcal{D}_n^{(L)})$ of the data distribution per sample. Specifically, the free entropy per sample, which from now on will be simply referred to as free entropy, is defined as follows:

$$\bar{f}_n^{(L)} := \frac{1}{n} \mathbb{E} \log \mathcal{Z}^{(L)}(\mathcal{D}_n^{(L)}) = -\frac{1}{n} H(\mathcal{D}_n^{(L)}) - \frac{1}{n} \mathbb{E} \log P(\mathbf{X}^{(0)}) \tag{3.21}$$

Here, the expectation $\mathbb{E}$ is taken with respect to the training data $\mathcal{D}_n^{(L)} = \{(\mathbf{X}_\mu^{(0)}, Y_\mu^{(L)})\}_{\mu=1}^n$. Similarly to what done in 2.2 for the free energy density, the normalization by n is linked to the number of terms in the Hamiltonian defined by the exponent in the partition function, and corresponds to the number of data points. It is important to observe that, according to definition (3.10), the partition function is not a function of the joint distribution of the dataset, but rather of the probability of obtaining the outputs given the input data and the neural network parameters. In our investigation, a slight deviation from the standard terminology will be adopted by referring to what is technically the conditional free entropy associated with the dataset as simply free entropy.

Lastly, having introduced the free entropy analogously to what done in 2.3.3 the mutual information per sample between the dataset $\mathcal{D}_n^{(L)}$ and the teacher network weights $\boldsymbol{\theta}^{*(L)}$ for the $L+1$-layer neural network can now be defined:

$$\begin{aligned} \frac{1}{n} I_n^{(L)}(\boldsymbol{\theta}^{*(L)}; \mathcal{D}_n^{(L)}) &= \frac{1}{n} H(\mathcal{D}_n^{(L)}) - \frac{1}{n} H(\mathcal{D}_n^{(L)} \mid \boldsymbol{\theta}^{*(L)}) \\ &= -\bar{f}_n^{(L)} + \mathbb{E} \log P_{\text{out}}\Big(Y_1^{(L)} \mid \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\, \mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big) \end{aligned} \tag{3.22}$$

3.3 Equivalent shallow network

A closely related model to the previously introduced deep neural network is the generalized linear model (GLM). This model consists of a single-layer neural network and represents a generalization of a perceptron, as introduced in more detail in 2.4.2. In the same way as in 3.2, a teacher-student framework is considered, with both a teacher GLM and a student GLM. The $^{(0)}$ superscript is used for quantities related to the generalized linear model. The teacher GLM generates the labels as follows:

$$Y_\mu^{(0)} = f\Big(\eta^{(0)} \frac{\mathbf{v}^{*(0)\intercal} \mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}} \zeta_\mu^{*(0)}\,;\, \mathbf{A}_\mu\Big) + \sqrt{\Delta} Z_\mu \tag{3.23}$$

Equation (3.23) implies that in terms of the output kernel, the labels are drawn according to the following distribution:

$$Y_\mu^{(0)} \sim P_{\text{out}}\left(\cdot \mid \eta^{(0)} \frac{\mathbf{v}^{*(0)\intercal}\mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{*(0)}\right) \tag{3.24}$$

The dataset generated by this model is denoted as $\mathcal{D}_n^{(0)} := \{(\mathbf{X}_\mu^{(0)}, Y_\mu^{(0)})\}_{\mu=1}^n$, where each pair corresponds to an input vector and the corresponding label. It is crucial to observe that this dataset differs from the one used by the $L+1$-layers student neural network, due to the distinct teacher architectures, resulting in different label generation models. The * notation is used as before, and the parameters of the GLM teacher network are also referred as $\boldsymbol{\theta}^{*(0)} = \{\mathbf{v}^{*(0)}, \boldsymbol{\zeta}^{*(0)}\}$.

The readout function f, the stochasticity $\mathbf{A}_\mu$, and the label noise Z_μ rescaled by the factor $\sqrt{\Delta}$ are the same as presented in 3.2. The newly introduced quantities in (3.23) instead are defined as follows:

- Weight vector of the generalized linear model $\mathbf{v}^{*(0)} \in \mathbb{R}^{d^{(0)}}$: set of parameters that the teacher model uses to weigh the importance of different features in the input data. It is a random vector with entries i.i.d. drawn from a Gaussian distribution, $v_i^{*(0)} \sim \mathcal{N}(0,1)$ for every index i;
- $\eta^{(0)} \in \mathbb{R} > 0$: this parameter is a scaling factor that adjusts how much each neuron's signal contributes to the next layer;
- Effective Gaussian noise $\zeta_\mu^{*(0)} \in \mathbb{R}$: $\zeta_\mu^{*(0)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is Gaussian i.i.d. noise. Since it is multiplied by $\gamma^{(0)} > 0$, the noise term has then variance controlled by the parameter $\gamma^{(0)}$, and it adds variability to the output of the teacher GLM. From the perspective of finding an equivalent model to the deep network presented in 3.2, the noise component represents the higher-order terms introduced by the nonlinearities in the $L+1$-layer teacher network when generating the data, which are not learned by the student network with the same architecture [29]. By including this noise then the model effectively accounts for the complex interactions in the data that the student network does not capture. The presence of this term represents an additional source of randomness in the data generation process;
- $\gamma^{(0)} \in \mathbb{R} > 0$: this parameter controls the variance of the Gaussian noise in the model. In relation to the $L+1$-layer neural network model, it describes the entity of the non linear activation functions' higher-order terms.

Having introduced all the parameters, it is worth noticing that (3.23) satisfies the characterization of the GLM presented in 2.4.2. The GLM's linear predictor, which is a linear combination of the input features, is given by the term $\eta^{(0)} \frac{\mathbf{v}^{*(0)\intercal}\mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{*(0)}$. The response function applied to the linear predictor, that allows for nonlinear relationships between the input features and the output, is f. Finally, the random component that is added to the output of the link function is $\sqrt{\Delta}Z_\mu$.

The link between the generalized linear model and the $L+1$- layer neural network introduced in 3.2 is then described through the parameters $\eta^{(0)}$ and $\gamma^{(0)}$, whose precise definition will be given

later in 4.1, and the structures of the models. While the GLM is a single-layer network with a linear combination of inputs and a readout function, in a deep neural network the linear model is replaced by multiple hidden layers stacked together. The nonlinearity in a deep neural network arises from the activation functions in each hidden layer, whereas in the corresponding GLM, such nonlinearity is captured through the parameters $\eta^{(0)}$ and $\gamma^{(0)}$ along with the noise term $\zeta_\mu^{*(0)}$.

The construction done for the $L+1$-layer neural network and the information-theoretic quantities introduced can be similarly applied to the generalized linear model. The posterior distribution for this model reads:

$$\begin{aligned} dP(\boldsymbol{\theta}^{(0)} \mid \mathcal{D}_n^{(0)}) &= dP(\boldsymbol{\zeta}^{(0)}, \mathbf{v}^{(0)} \mid \mathcal{D}_n^{(0)}) \\ &= \frac{1}{\mathcal{Z}^{(0)}(\mathcal{D}_n^{(0)})} \prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu^{(0)} \mid \eta^{(0)} \frac{\mathbf{v}^{(0)\intercal}\mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{(0)}\Big) D\mathbf{v}^{(0)} D\boldsymbol{\zeta}^{(0)} \end{aligned} \tag{3.25}$$

where $D\boldsymbol{\zeta}^{(0)} = \prod_\mu^n D\zeta_\mu^{(0)}$.
Denoting

$$s_\mu^{(0)} = s_\mu^{(0)}(\boldsymbol{\theta}^{(0)}, \mathbf{X}_\mu^{(0)}) = \eta^{(0)} \frac{\mathbf{v}^{(0)\intercal}\mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{(0)} \tag{3.26}$$

$$S_\mu^{(0)} = S_\mu^{(0)}(\boldsymbol{\theta}^{(0)}, \mathbf{X}_\mu^{(0)}) = \eta^{(0)} \frac{\mathbf{v}^{*(0)\intercal}\mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{*(0)} \tag{3.27}$$

the partition function is computed as:

$$\mathcal{Z}^{(0)}(\mathcal{D}_n^{(0)}) := \int D\boldsymbol{\theta}^{(0)} \exp\Big(\sum_{\mu=1}^{n} u_{Y_\mu}^{(0)}(s_\mu)\Big) \tag{3.28}$$

The law of the data can be obtained as:

$$\begin{aligned} dP(\mathcal{D}_n^{(0)}) &= \prod_{\mu=1}^{n} \Big(\prod_{j=1}^{d^{(0)}} \frac{dX_{\mu j}^{(0)}}{\sqrt{2\pi}} e^{-\frac{X_{\mu j}^{(0)\,2}}{2}}\Big) dY_\mu^{(0)}\, \mathbb{E}_{\mathbf{v}^{*(0)},\boldsymbol{\zeta}^{*(0)}} \prod_{\mu=1}^{n} P_{\text{out}}(Y_\mu^{(0)} \mid S_\mu^{(0)}) \\ &=: \prod_{\mu=1}^{n} D\mathbf{X}_\mu^{(0)} dY_\mu^{(0)}\, \mathbb{E}_{\mathbf{v}^{*(0)},\boldsymbol{\zeta}^{*(0)}} \exp\Big(\sum_{\mu=1}^{n} u_{Y_\mu}^{(0)}(S_\mu)\Big) \end{aligned} \tag{3.29}$$

whereas the Bayes-optimal predictor corresponds to:

$$\begin{aligned} \hat{Y}_{\text{Bayes}}^{(0)}(\mathbf{X}_{\text{new}}^{(0)}) &:= \mathbb{E}[Y_{\text{new}}^{(0)} \mid \mathcal{D}_n^{(0)}, \mathbf{X}_{\text{new}}^{(0)}] \\ &= \int dY\, Y\, P_{\text{out}}\Big(Y \mid \eta^{(0)} \frac{\mathbf{v}^{(0)\intercal}\mathbf{X}_{\text{new}}^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}}\zeta_\mu^{(0)}\Big) dP(\mathbf{v}^{(0)}, \boldsymbol{\zeta}^{(0)} \mid \mathcal{D}_n^{(0)}) \end{aligned} \tag{3.30}$$

and the generalization error is:

$$\mathcal{E}_n^{(0)} := \mathbb{E}\big(Y_{\text{new}}^{(0)} - \mathbb{E}[Y_{\text{new}}^{(0)} \mid \mathcal{D}_n^{(0)}, \mathbf{X}_{\text{new}}^{(0)}]\big)^2 \tag{3.31}$$

Exploiting the Gibbs brackets notation previously introduced, the posterior mean of a function g can be written as

$$\langle g \rangle^{(0)} = \int dP(\boldsymbol{\theta}^{(0)} \mid \mathcal{D}_n^{(0)}) f = \frac{1}{\mathcal{Z}^{(0)}(\mathcal{D}_n^{(0)})} \int D\mathbf{v}^{(0)} D\boldsymbol{\zeta}^{(0)} \prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu^{(0)} \mid \eta^{(0)} \frac{\mathbf{v}^{(0)\intercal} \mathbf{X}_\mu^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}} \zeta_\mu^{(0)}\Big)\Big) g \tag{3.32}$$

and the free entropy can be defined as:

$$\bar{f}_n^{(0)} := \frac{1}{n} \mathbb{E} \log \mathcal{Z}^{(0)}(\mathcal{D}_n^{(0)}) = \frac{1}{n} \mathbb{E} \log \int D\mathbf{v}^{(0)} D\boldsymbol{\zeta}^{(0)} \exp\Big(\sum_{\mu=1}^{n} u_{Y_\mu}^{(0)}(s_\mu)\Big) \tag{3.33}$$

Finally, the mutual information per sample between the dataset $\mathcal{D}_n^{(0)}$ and the teacher parameters $\boldsymbol{\theta}^{*(0)}$ of the generalized linear model can be obtained through the following relation:

$$\begin{aligned} \frac{1}{n} I_n^{(0)}(\boldsymbol{\theta}^{*(0)}; \mathcal{D}_n^{(0)}) &= \frac{1}{n} H(\mathcal{D}_n^{(0)}) - \frac{1}{n} H(\mathcal{D}_n^{(0)} \mid \boldsymbol{\theta}^{*(0)}) \\ &= -\bar{f}_n^{(0)} + \mathbb{E} \log P_{\text{out}}\Big(Y_1^{(0)} \mid \eta^{(0)} \frac{\mathbf{v}^{(0)\intercal} \mathbf{X}_{\text{new}}^{(0)}}{\sqrt{d^{(0)}}} + \sqrt{\gamma^{(0)}} \zeta_\mu^{(0)}\Big) \end{aligned} \tag{3.34}$$

The mutual information between the GLM teacher network weights and the corresponding training dataset is a quantity that has been studied in the literature [14], and its relevance lies in the fact that the generalization error can be computed from it. It is also worth reminding that the Bayes-optimal setup achieves the minimum generalization error defined as above, thus allowing to derive information-theoretic results for the model. The regime of the GLM where phase transitions can be observed and the generalization error is not trivially zero or one is determined by the condition $\alpha = \frac{n}{d^{(0)}} = O(1)$.

A visual representation of the generalization error of the GLM, which is also reported below, can be found in [14]. The plot shows the generalization error when the weights of the teacher are distributed according to a Gauss-Bernoulli distribution. As the ratio α increases, meaning that the number of samples grows relative to the input size $d^{(0)}$, the generalization error improves, and ultimately, as α approaches infinity, the error will tend to zero.

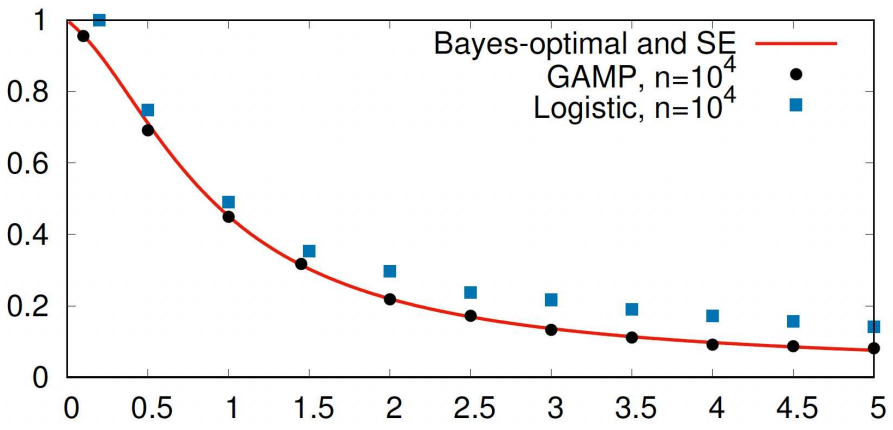

Figure 3.1: Generalization error as a function of α with teacher weights drawn from a Gauss-Bernoulli distribution, taken from [14].

The generalized linear model, then, with its well-documented and thoroughly researched properties, stands as an effective tool for gaining insights into the more complex behavior of deep neural networks.

3.4 Methods

Here, the main methods that will extensively be used to carry out the proofs of the results presented in 4 are introduced.

3.4.1 Stein's Lemma

Stein's lemma [61, 62], also referred to as Gaussian integration by parts, is a probability theory theorem whose application spans various fields such as higher-dimensional problems in statistics, machine learning, and signal processing. Here the one-dimensional case and its generalization to the multivariate case are presented.

Lemma 1. *Let Z be a standard Gaussian random variable, $Z \sim \mathcal{N}(0,1)$ and let $f : \mathbb{R} \to \mathbb{R}$ be a function whose derivative f' is continuous almost everywhere and satisfies $E|f'(Z)| < \infty$. Then, the following holds:*

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)] \tag{3.35}$$

Proof. Integration by parts is exploited, remembering that the probability density associated with a standard Gaussian random variable is $\phi(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$.

$$\begin{aligned}\mathbb{E}[Zf(Z)] &= \int zf(z)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz \\ &= \Big[-f(z)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\Big]_{-\infty}^{\infty} - \int -f'(z)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz \\ &= \mathbb{E}[f'(Z)]\end{aligned} \tag{3.36}$$

□

Lemma 2. *Let $\mathbf{X} \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a n-dimensional multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For any function $f : \mathbb{R}^n \to \mathbb{R}$ such that the partial derivative $(\partial/\partial X_i)f(\mathbf{X})$ is continuous almost everywhere and $\mathbb{E}|(\partial/\partial X_i)f(\mathbf{X})| < \infty$ for each component i, define the gradient of f at $\mathbf{X}$ as $\nabla f(\mathbf{X}) = ((\partial/\partial X_1)f(\mathbf{X}), \cdots, (\partial/\partial X_n)f(\mathbf{X}))^\intercal$. Then the following holds:*

$$\mathrm{Cov}[\mathbf{X}, f(\mathbf{X})] = \boldsymbol{\Sigma}\,\mathbb{E}[\nabla f(\mathbf{X})] \tag{3.37}$$

Proof. For the standard Gaussian random vector $\mathbf{Z}$, $\mathbf{Z}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, the transformation $\mathbf{X} = \Sigma^{(1/2)}\mathbf{Z} + \boldsymbol{\mu}$ can be used, allowing to rewrite f in terms of the new variable: $f(\mathbf{X}) = f(\boldsymbol{\Sigma}^{(1/2)}\mathbf{Z} + \boldsymbol{\mu})$. Using Lemma 1 it can be obtained:

$$\mathbb{E}[Z_i f(\mathbf{Z})] = \mathbb{E}\Big[(\partial/\partial Z_i)f(\mathbf{Z})\Big] \tag{3.38}$$

or in vector notation

$$\mathrm{Cov}[\mathbf{Z}, f(\mathbf{Z})] = \mathbb{E}[\nabla f(\mathbf{Z})] \tag{3.39}$$

It follows that:

$$\begin{aligned}\mathrm{Cov}[\mathbf{X}, f(\mathbf{X})] &= \mathrm{Cov}[\boldsymbol{\Sigma}^{(1/2)}\mathbf{Z} + \boldsymbol{\mu}, f(\mathbf{Z})] \\ &= \mathrm{Cov}[\boldsymbol{\Sigma}^{(1/2)}\mathbf{Z}, f(\mathbf{Z})] \\ &= \boldsymbol{\Sigma}^{(1/2)}\mathrm{Cov}[\mathbf{Z}, f(\mathbf{Z})] \\ &= \boldsymbol{\Sigma}^{(1/2)}\mathbb{E}[\nabla f(\mathbf{Z})] \\ &= \boldsymbol{\Sigma}\,\mathbb{E}[\nabla f(\mathbf{X})]\end{aligned} \tag{3.40}$$

where the last equality is due to the fact that $\nabla f(\mathbf{Z}) = \boldsymbol{\Sigma}^{(1/2)}\nabla f(\mathbf{X})$. From this, the first element of the vector can be computed as follows:

$$\mathrm{Cov}[X_1, f(\mathbf{X})] = \sum_{i=1}^{n} \mathrm{Cov}[X_i, X_1]\mathbb{E}[(\partial/\partial X_i)f(\mathbf{X})] \tag{3.41}$$

□

3.4.2 Nishimori identity

The Nishimori identity, named after Hidetoshi Nishimori's pioneering work on spin glasses [63, 64], is fundamental in the study of Bayes-optimal estimators and high-dimensional inference problems. This identity exploits the concept that the true underlying signal $\mathbf{X}$ and a sample $\mathbf{X}^{(1)}$ drawn from the posterior distribution $P_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{y})$ are statistically equivalent. This symmetry, which is in fact a consequence of Bayes rule, allows significant simplifications in the analysis. Specifically, in a Bayes-optimal setting in which the posterior distribution $P_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{y})$ is known, the identities imply that, within expectations, the signal $\mathbf{X}$ can be replaced by its posterior replica $\mathbf{X}^{(1)}$.

This substitution ensures that the signal and its replica exhibit symmetric behavior in statistical averages, thus facilitating a comprehensive analysis of inference problems. Without this symmetry, as if $\mathbf{X}^{(1)}$ were drawn from a different distribution, proof techniques based on these identities would fail due to the asymmetry introduced. As a result, the Nishimori identity forms the foundation of many theoretical advances in understanding and solving complex inference tasks in high-dimensional statistics [65, 66].

Lemma 3. *Let $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be a couple of random vectors. Their joint distribution is $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$ and conditional distribution $P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$. Let $k \geq 1$ and let $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)}$ be i.i.d. random variables drawn from the distribution $P_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{y})$. Let $\mathbb{E}$ be the total expectation, namely, the expectation with respect to the joint distribution, and $\langle\cdot\rangle$ the expectation with respect to the product measure $P_{\mathbf{X}|\mathbf{Y}}^{\otimes\infty}$. Then, for all continuous bounded function g it holds:*

$$\mathbb{E}\big\langle g(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})\big\rangle = \mathbb{E}\big\langle g(\mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})\big\rangle \tag{3.42}$$

Proof. Exploiting the Bayes formula, the joint probability distribution can be rewritten as $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})P_{\mathbf{Y}}(\mathbf{y}) = P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})P_{\mathbf{X}}(\mathbf{x})$. The meaning of this is that it is equivalent to draw the couple from the joint distribution or to draw $\mathbf{Y}$ according to its distribution, and then $\mathbf{X}$ conditionally on $\mathbf{Y}$ from the conditional distribution. Iterating this, it can be observed that the $(k+1)$-tuples $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})$ and $(\mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})$ have the same law. In equations, this implies

$$\begin{aligned}
&\mathbb{E}\big\langle(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})\big\rangle \\
&\qquad := \mathbb{E}_{\mathbf{XY}}\mathbb{E}_{\mathbf{X}^{(2)}|\mathbf{Y}} \dots \mathbb{E}_{\mathbf{X}^{(k)}|\mathbf{Y}} (\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)}) \\
&\qquad = \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}|\mathbf{Y}}\mathbb{E}_{\mathbf{X}^{(2)}|\mathbf{Y}} \dots \mathbb{E}_{\mathbf{X}^{(k)}|\mathbf{Y}} (\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)}) \\
&\qquad = \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}^{(1)}|\mathbf{Y}}\mathbb{E}_{\mathbf{X}^{(2)}|\mathbf{Y}} \dots \mathbb{E}_{\mathbf{X}^{(k)}|\mathbf{Y}} (\mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)}) \\
&\qquad =: \mathbb{E}\big\langle(\mathbf{Y}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(k)})\big\rangle
\end{aligned} \tag{3.43}$$

□

3.4.3 Concentration of measure

The phenomenon of concentration of measure [67, 68, 69] is a central result in probability theory and statistics, describing how sums or more general combinations of large numbers of independent

(or weakly dependent) random variables tend to show considerably small fluctuations around their expected value.

Far from being a mere theoretical concept, the concentration of measure phenomenon has profound implications for the comprehension of the natural world. It explains why macroscopic systems behave predictably, even though they consist of a vast number of randomly behaving microscopic particles. The foundational principles of this phenomenon offer deep insights into the stability and reliability of complex systems, impacting diverse fields such as statistical physics, random matrix theory, and beyond.

The phenomenon is intuitively explained by the fact that several random variables are unlikely to act together and deviate substantially from the mean, thus moving the sum or function away from the expected value. Moreover, when the variables are independent, the influence of each individual one decreases, making the combination of variables more stable. A fundamental result that demonstrates this principle is the law of large numbers, which states that the average of n i.i.d. random variables converges to the expected value as n increases. This convergence illustrates how fluctuations become negligible in large samples, a special case of the broader phenomenon of concentration.

It is also important to note that this type of concentration does not only concern the sums of random variables $\{X_1, \ldots, X_n\}$ or their linear combinations but extends to nonlinear functions $f(X_1, \ldots, X_n)$ when these are sufficiently regular. This condition ensures that the assumption of sufficient independence of the variables holds, so that these functions are not excessively sensitive to the variation of a single variable. However, to accurately state and demonstrate concentration results, it is necessary to quantify what it means for a function to be "sensitive" and how "close" a random variable is to its mean. Typically, this involves deriving explicit bounds on the variance or tail probability of the fluctuations of $|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)|$. These limits usually depend on the dimensionality n and the properties of the distribution of the variables.

Considering a random variable X with mean μ, it is known that the k-th moment of the variable is linked to the tail distribution through this expression [68]:

$$\mathbb{E}|X|^k = \int_0^\infty kt^{k-1} P(|X| \geq t)\, \mathrm{d}t \tag{3.44}$$

In our analysis, the control of the moments of random variables is needed, which in turn requires the control of the tail of the probability distribution of the random variable. Thus, the interest lies in obtaining inequalities of the form:

$$P(|X - \mu| \geq t) \leq \text{small quantity} \tag{3.45}$$

Specifically, in our investigation, the random variable will be a function of a random vector. Moreover, the k-th central moment of such random variable will need to depend on the inverse of the k-th power of the square root (or a higher power) of the dimensionality of the random vector. Due to this, exponential bounds similar to those for sub-Gaussian and sub-exponential random variables are sought. These bounds ensure that the probability of large deviations from the mean decreases exponentially, providing the necessary control over the moments of the random variables.

3.4.4 Interpolation method

The Guerra-Toninelli interpolation method is a powerful and versatile technique that was introduced by Francesco Guerra and Fabio Lucio Toninelli and developed in the context of studying mean-field spin glasses [70]. The method was first developed to prove inequalities involving Gaussian random vectors [71, 72, 73], and is now a cornerstone in mathematical physics and statistical mechanics, specifically in the study of spin glasses.

Spin glasses are disordered magnetic systems that have been extensively studied in statistical mechanics due to their complex behaviors arising from random interactions between spins. One of the most well-known models of spin glasses is the Sherrington-Kirkpatrick (SK) model [46]. Initially thought to be a simple solvable model, it later revealed a much richer and more interesting structure. The SK model has been instrumental in understanding the disordered nature and peculiar properties of spin glasses. An initial solution for the free energy density of the SK model was proposed by Sherrington and Kirkpatrick themselves. However, this solution was not well-defined in the thermodynamic limit, resulting in a negative entropy thus an unphysical solution. The problem of the existence of the thermodynamic limit for the quenched density of free energy in the SK model was addressed by Giorgio Parisi with his replica symmetry breaking solution, obtained using the replica method [48, 49].

The interpolation scheme developed in the seminal works of Guerra and Toninelli [51, 74, 50] then laid the groundwork for proving the Parisi formula. Initially, Francesco Guerra established a uniform bound using his innovative replica symmetry breaking uniform bound [50]. This result, built upon the interpolation method he had earlier developed with Toninelli, proved the existence of the thermodynamic limit [51]. Subsequently, Michel Talagrand [52] succeeded in identifying a corresponding converse limit, marking a significant turning point in the field of spin glasses.

Interpolation assumed this crucial importance since it not only has been instrumental in the study of mean-field spin glasses [75], but has been extended and adapted to a wide range of applications, well beyond the realm of traditional statistical mechanics. These fields include for example coding theory, communications, signal processing, and theoretical computer science [70]. Moreover, the technique is particularly useful in the context of Bayesian inference problems, where it can be used to prove the replica formula for non-trivial inference problems [76].

In recent years, an extension of the Guerra-Toninelli interpolation method, known as the adaptive interpolation method [76, 77], has been developed. In the original Guerra-Toninelli interpolation scheme, the interpolation path is fixed. Conversely, in the adaptive interpolation method, the path is allowed to adapt based on the problem considered. These characteristics make the technique more flexible and applicable to a wider range of problems, and it also allows to prove the replica formula for non-trivial inference problems.

In summary, the Guerra-Toninelli interpolation method is a key tool in the mathematical analysis of complex systems, providing a robust framework for proving important results in statistical physics and beyond.

In our discussion, the interpolation method allows for a smooth transition between two different systems or models, and to study how their properties change in the process. This transition is governed by a continuously changing external parameter, $t \in [0, 1]$. For $t = \{0, 1\}$ the two original models being interpolated between are recovered, while for $t \in (0, 1)$ an auxiliary model, known as the interpolating model, is defined. This approach allows for the comparison of the original models by examining the t-derivative of the interpolating model at intermediate values.

Specifically, in our investigation, the proof strategy involves iteratively reducing the $\ell + 1$-layer neural network to an ℓ-layer neural network, continuing this process until a one-layer neural network is obtained. This is done inductively by identifying an equivalent one-layer model for the last two layers of the network. The interpolation is then defined between the nonlinear argument of the readout function and the corresponding linearization. For the $L + 1$-layer neural network this is expressed as:

$$S_{t\mu}^{(L)} := \sqrt{1-t}\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) + \sqrt{t}\rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{t\epsilon^{(L-1)}}\xi_\mu^{*(L-1)} \tag{3.46}$$

where the definitions of the parameters appearing in the linearized model will be given more precisely later in 4.1.

Chapter 4

Main results

In this chapter, the main findings of the analysis are presented, expressed as bounds relating the free entropy of the full neural network and that of the GLM, along with bounds in terms of the the mutual information between the dataset used by the model and the teacher model's weights. To properly define all the quantities relevant to these results, the recursive method employed for the reduction is first discussed. Following this, the main findings are stated, and an analysis of the bounds obtained is performed to determine in which cases the full neural network is equivalent to a generalized linear model. Finally, an outline of the proof is provided to elucidate the derivation of the bounds, offering a clearer understanding before delving into the detailed proof in 5.

4.1 Recursion scheme

As mentioned in 1, our analysis aims to establish an information-theoretic equivalence between an $L+1$-layer neural network and a generalized linear model. Specifically, the goal is to demonstrate this equivalence relating the mutual information between the dataset $\mathcal{D}_n^{(L)}$ and the weights $\boldsymbol{\theta}^{*(L)}$ of the deep neural network, and the mutual information between the dataset $\mathcal{D}_n^{(0)}$ and the parameters $\boldsymbol{\theta}^{*(0)}$ of the generalized linear model. The strategy involves iteratively reducing the $L+1$-layer neural network to an L-layer neural network. This is achieved through an inductive process, identifying an equivalent one-layer model for the last two layers of the network.

The approach follows the methodology presented in [19]. The core idea is to linearize the nonlinear argument of the readout function to find an equivalent network with one less hidden layer. Referring to the models defined in 3.2 and 3.3, the idea is to find a linearization such that the resulting model corresponds to a generalized linear model with respect to the argument $\mathbf{X}_\mu^{(\ell)}$ of the nonlinearity, in the same way as model 3.3 is a generalized linear model with respect to the data $\mathbf{X}_\mu^{(0)}$.

This equivalence is expressed first in terms of free energy and subsequently the mutual information between the teacher weights and the dataset for the two networks. It is important to observe that since a teacher-student setup is adopted, the equivalent model is realized in both the architec-

ture of the teacher and the student. Considering now the $L+1$-layer neural network, the equation for the teacher, that we recall here, is (3.1):

$$Y_\mu^{(L)} = f\Big(\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big);\mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu = f\Big(\frac{\mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(L)}}{\sqrt{d^{(L)}}};\mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu \tag{4.1}$$

The first step of the reduction is then performed as follows:

$$\frac{\mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(L)}}{\sqrt{d^{(L)}}} = \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) \longrightarrow \rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}}\xi_\mu^{*(L-1)} \tag{4.2}$$

On the left-hand side there is the argument of the readout function that is to be linearized, where $\mathbf{X}_\mu^{(L)}$ represents the dataset input at the L-th layer, $\mathbf{a}^*$ is the readout vector as defined in 3.2, $\mathbf{W}^{*(L)}$ denotes the weights of the L-th layer introduced in 3.2, and φ is the activation function. The right-hand side shows the linearized model, where the vector $v^{*(L-1)}$ is a Gaussian vector with entries i.i.d distributed $v_i^{*(L-1)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\xi_\mu^{(L-1)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is a standard Gaussian random variable that introduces a noise term independently for every sample μ. Additionally, $\rho^{(L-1)}$ and $\epsilon^{(L-1)}$ are real coefficients.

It can be observed that the linearized model defines a generalized linear model with respect to $\mathbf{X}_\mu^{(L-1)}$, allowing to apply the discussion about the role of the coefficients for the generalized linear model in 3.3 also to this case. Recognizing that $\mathbf{X}_\mu^{(L-1)} = \varphi(\frac{\mathbf{W}^{*(L-1)}\,\mathbf{X}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}})$, that the vectors $\mathbf{a}^*$ and $\mathbf{v}^{*(L-1)}$ follow the same distribution and that $\rho^{(L-1)}$ is just a multiplicative coefficient, it can be noticed that the same nonlinear structure persists after each linearization step. This allows the proof to be adapted and the reduction to be iterated. Each subsequent reduction step can be expressed mathematically as follows:

$$\frac{\mathbf{v}^{*(k)\intercal}\mathbf{X}_\mu^{(k)}}{\sqrt{d^{(k)}}} = \frac{\mathbf{v}^{*(k)\intercal}}{\sqrt{d^{(k)}}}\varphi\Big(\frac{\mathbf{W}^{*(k)}\,\mathbf{X}_\mu^{(k-1)}}{\sqrt{d^{(k-1)}}}\Big) \longrightarrow \rho^{(k-1)}\frac{\mathbf{v}^{*(k-1)\intercal}\mathbf{X}_\mu^{(k-1)}}{\sqrt{d^{(k-1)}}} + \sqrt{\epsilon^{(k-1)}}\xi_\mu^{*(k-1)} \tag{4.3}$$

Starting from the firs $L+1$-layer neural network, the argument of the readout function of the

reduced models evolves as follows:

$$Y_\mu^{(L)} = f\Big(\underbrace{\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)}; \mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu$$

$$\downarrow$$

$$f\Big(\overbrace{\rho^{(L-1)}\,\underbrace{\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}}\xi_\mu^{*(L-1)}}}; \mathbf{A}_\mu\Big)$$

$$\downarrow$$

$$f\Big(\rho^{(L-1)}\Big[\overbrace{\rho^{(L-2)}\,\underbrace{\frac{\mathbf{v}^{*(L-2)\intercal}\mathbf{X}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}} + \sqrt{\epsilon^{(L-2)}}\xi_\mu^{*(L-2)}}}\Big] + \sqrt{\epsilon^{(L-1)}}\xi_\mu^{*(L-1)}; \mathbf{A}_\mu\Big)$$

$$\downarrow$$

$$f\Big(\rho^{(L-1)}\Big[\rho^{(L-2)}\Big[\overbrace{\rho^{(L-3)}\frac{\mathbf{v}^{*(L-3)\intercal}\mathbf{X}_\mu^{(L-3)}}{\sqrt{d^{(L-3)}}} + \sqrt{\epsilon^{(L-3)}}\xi_\mu^{*(L-3)}}\Big] + \sqrt{\epsilon^{(L-2)}}\xi_\mu^{*(L-2)}\Big] + \sqrt{\epsilon^{(L-1)}}\xi_\mu^{*(L-1)}; \mathbf{A}_\mu\Big)$$

$$\vdots$$

Thus, by iterating this reduction process and using the fact that the sum of independent Gaussian random variables is also a Gaussian random variable with variance equal to the sum of the variances, after $L-k$ steps, the output of the $k+1$-layer teacher neural network can be expressed as:

$$Y_\mu^{(k)} = f\Big(\eta^{(k)}\frac{\mathbf{v}^{*(k)\intercal}\mathbf{X}_\mu^{(k)}}{\sqrt{d^{(k)}}} + \sqrt{\gamma^{(k)}}\zeta_\mu^{*(k)}; \mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu \tag{4.4}$$

where $\zeta_\mu^{*(k)} \sim \mathcal{N}(0,1)$ and the coefficients $\eta^{(k)}$ and $\gamma^{(k)}$ are defined as:

$$\begin{aligned}\eta^{(k)} &= \prod_{i=k}^{L-1}\rho^{(i)} \\ \gamma^{(k)} &= (1-\delta_{1,(L-k)})\sum_{j=k}^{L-2}\Big(\prod_{i=j+1}^{L-1}\rho^{(i)\,2}\Big)\epsilon^{(j)} + \epsilon^{(L-1)}\end{aligned} \tag{4.5}$$

Here, $\delta_{1,(L-k)}$ is the Kronecker delta, ensuring that the summation term vanishes when $L-k=1$, corresponding to the first reduction step. This formulation accounts for the accumulation of noise terms and scaling factors through the layers.

To each reduced model with $k+1$ layers then, a construction done similarly to the one done for the $L+1$-layer neural network in 3.2 and for the GLM in 3.3 can be performed. Specifically, each reduced model has parameters of the teacher network $\boldsymbol{\theta}^{*(k)} = \{\mathbf{v}^{*(k)}, \boldsymbol{\zeta}^{*(k)}, \mathbf{W}^{*(k)}, \dots, \mathbf{W}^{*(1)}\}$, and the responses are drawn exploiting the output kernel from the distribution $Y_\mu^{(k)} \sim P_{\text{out}}(\cdot \mid S_\mu^{(k)})$. The student network is then trained on the dataset $\mathcal{D}_n^{(k)} = \{(\boldsymbol{X}_\mu^{(0)}, Y_\mu^{(k)})\}_{\mu=1}^n$, of which the partition function can be computed as $\mathcal{Z}^{(k)}(\mathcal{D}_n^{(k)}) = \int D\boldsymbol{\theta}^{(k)} \prod_{\mu=1}^n P_{\text{out}}(Y_\mu^{(k)} \mid s_\mu^{(k)})$.

After L iterations, the coefficients obtained for the one-layer generalized linear model described in 2.4.2 are:

$$\begin{aligned} \eta^{(0)} &= \prod_{i=0}^{L-1} \rho^{(i)} \\ \gamma^{(0)} &= \sum_{j=0}^{L-2} \Big(\prod_{i=j+1}^{L-1} \rho^{(i)\ 2} \Big) \epsilon^{(j)} + \epsilon^{(L-1)} \end{aligned} \tag{4.6}$$

The GLM parameters are expressed in terms of the recursion parameters $\rho^{(k)}$ and $\epsilon^{(k)}$, with their definitions provided later in (4.16). Notably, the obtained expressions are consistent with the results presented by [29].

4.2 Results

4.2.1 Concentration results

These results are of crucial importance for the successful derivation of the proof of Theorem 7. Indeed, a significant difference with respect to the proof presented in [19] is that the $\mathbf{X}^{(\ell)}$ are not Gaussian for $1 \leq \ell \leq L$, so the concentration properties of Gaussian random variables cannot be applied anymore. Specifically, these concentration results are centered around two key quantities: the squared norm of a hidden layer's output, which is a vector, and the scalar product of two such vectors obtained starting from different data points.

While a recursive argument could suffice for controlling moments up to the second order starting from the input variables, higher-order moments necessitate exponential bounds on the probability of these quantities diverging from their expectation. This is crucial for accurately estimating all moments as required in Lemma 18 and the proof of Theorem 7 detailed in 5.4.

The following proposition presents the result obtained about the concentration of the squared norm.

Proposition 4 (Norm concentration). *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector whose norm satisfies the exponential concentration bound*

$$\mathbb{P}\Big(\Big| \frac{\|\mathbf{X}\|^2}{d} - \sigma \Big| \geq \epsilon \Big) \leq 2 \exp\Big(- dC\epsilon^2 \Big) \tag{4.7}$$

for any $\epsilon > 0$ and a constant $C > 0$, with $\sigma = \mathbb{E}\frac{\|\mathbf{X}\|^2}{d}$. Let $\varphi \in C^1$ such that $|\varphi'| \leq \bar{K}$. If $\mathbf{W} \in \mathbb{R}^{p\times d}$ is a matrix of i.i.d. centered Gaussian random variables with unit variance, then

$$\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-\min\{p,d\}C_\varphi\epsilon^2\Big) \tag{4.8}$$

for any $\epsilon > 0$ and some constant $C_\varphi > 0$ when $\sqrt{d} \gtrsim 1/\epsilon$.

Notice how Proposition 4 holds true for $\sqrt{d} \gtrsim 1/\epsilon$, or equivalently $\epsilon \gtrsim 1/\sqrt{d}$. This implies that this concentration result is applicable when the dimension is sufficiently high once ϵ is chosen, or conversely, for a fixed dimension the exponential bound is valid only for large enough deviations from the expectation.

The next proposition addresses the concentration of the scalar product.

Proposition 5 (Scalar product concentration)**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be random vectors drawn independently from the same distribution, that satisfy the exponential concentration bounds*

$$\mathbb{P}\Big(\Big|\frac{\mathbf{X}^\intercal\mathbf{Y}}{d}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dB\epsilon^2\Big) \tag{4.9}$$

$$\mathbb{P}\Big(\Big|\frac{\|\mathbf{X}\|^2}{d} - \sigma\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dC\epsilon^2\Big) \qquad \mathbb{P}\Big(\Big|\frac{\|\mathbf{Y}\|^2}{d} - \sigma\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dC\epsilon^2\Big) \tag{4.10}$$

for any $\epsilon > 0$ and constants $B, C > 0$, with $\sigma = \mathbb{E}\frac{\|\mathbf{X}\|^2}{d}$. Let $\varphi \in C^1$ such that $|\varphi'| \leq \bar{K}$. If $\mathbf{W} \in \mathbb{R}^{p\times d}$ is a matrix of i.i.d. centered Gaussian random variables with unit variance, then

$$\mathbb{P}\Big(\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal\varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-\min\{p,d\}C_\varphi\epsilon^2\Big) \tag{4.11}$$

for any $\epsilon > 0$ and some constant $C_\varphi > 0$.

Propositions 4 and 5 then allow to bound every central moment of the squared norm and scalar product of the vectors obtained by multiplying the initial vectors with a Gaussian random matrix and applying a nonlinear function.

Corollary 6 (Control on moments)**.** *Let $k \in \mathbb{N}$. Under the same hypotheses of Proposition 14 the following holds true:*

$$\begin{aligned}\mathbb{E}\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big|^k &= O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big)\\ \mathbb{E}\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal\varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big|^k &= O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big)\end{aligned} \tag{4.12}$$

Consider now the output vectors $\mathbf{X}_\mu^{(\ell)}$ from an arbitrary hidden layer ℓ of a $L+1$-layer neural network constructed as in 3.2, normalized by the square root of their dimension. Bounds for every central moment, indexed by k, of the squared norm or scalar product can be derived using Corollary 6. Specifically, calling $\mathrm{d}^{(\ell)} = \min_{j=1\dots\ell}\{d^{(j)}\}$, these bounds can be computed as follows:

$$\mathbb{E}\Big|\frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d^{(\ell)}} - \sigma^{(\ell)}\Big|^k = \Big(\frac{1}{\mathrm{d}^{(\ell)\ k/2}}\Big), \qquad \mathbb{E}\Big|\frac{\mathbf{X}_\mu^{(\ell)\intercal}\mathbf{X}_\nu^{(\ell)}}{d^{(\ell-1)}}\Big|^k = O\Big(\frac{1}{\mathrm{d}^{(\ell)\ k/2}}\Big) \tag{4.13}$$

The findings (4.13) highlight the importance of the concentration results obtained. Indeed, the significance of these concentration results lies in their implication that the central moments of the squared norm and scalar product scale with respect to the dimension in the same way they would if $\mathbf{X}^{(\ell)}$ were Gaussian random vectors. This suggests that the Gaussian properties of the input are preserved in the network, making the representations $\mathbf{X}^{(\ell)}$ at layer ℓ quasi-Gaussian random vectors.

These results, initially derived as a technical tool to finalize the proofs for the information-theoretic bounds, also stand out as an intriguing mathematical property in their own right.

4.2.2 Free entropy and mutual information results

The following theorem describes the relation between the free entropies of two different neural network models. Specifically, these models consist of a neural network with $k+1$ layers and another one with k layers, where the latter has undergone the linearization process described in 4.1. In order to state the theorem, however, an additional technical assumption is introduced. In the 2-layer neural network case studied in [19], this assumption has been rigorously proved.

H3) Recall the notation $\mathrm{d}^{(k)} = \min_{j=1\dots k}\{d^{(j)}\}$. Assuming H1) and H2) there exists a non-negative constant $C(f,\varphi)$ such that

$$\mathbb{E}_{\mathbf{v}^{*(k)}}\mathbb{V}_{\setminus\mathbf{v}^{*(k)}}\Big(\frac{1}{n}\log\mathcal{Z}_t^{(k)}\Big) = \mathbb{E}\Big(\frac{1}{n}\log\mathcal{Z}_t^{(k)} - \mathbb{E}_{\setminus\mathbf{v}^{*(k)}}\frac{1}{n}\log\mathcal{Z}_t^{(k)}\Big)^2 \le C(f,\varphi)\Big(\frac{1}{\mathrm{d}^{(k)}} + \frac{1}{n}\Big) \tag{4.14}$$

where $\mathbb{V}_{\setminus\mathbf{v}^{*(k)}}(\cdot) = \mathbb{E}_{\setminus\mathbf{v}^{*(k)}}((\cdot) - \mathbb{E}_{\setminus\mathbf{v}^{*(k)}}(\cdot))^2$. Additionally, $\mathcal{Z}_t^{(k)} = \mathcal{Z}_t^{(k)}(\mathcal{D}_{n,t}^{(k)})$ is the partition function associated with the interpolating model between the nonlinear argument of the readout function for the $k+1$-layer neural network and its corresponding linearized model:

$$S_{t\mu}^{(k)} := \sqrt{1-t}\frac{\mathbf{v}^{*(k)\intercal}}{\sqrt{d^{(k)}}}\varphi\Big(\frac{\mathbf{W}^{*(k)}\ \mathbf{X}_\mu^{(k-1)}}{\sqrt{d^{(k-1)}}}\Big) + \sqrt{t}\rho^{(k-1)}\frac{\mathbf{v}^{*(k-1)\intercal}\mathbf{X}_\mu^{(k-1)}}{\sqrt{d^{(k-1)}}} + \sqrt{t\epsilon^{(k-1)}}\xi_\mu^{*(k-1)} \tag{4.15}$$

Moreover, in the case $k=L$, $\mathbf{v}^{*(k)} = \mathbf{v}^{*(L)} = \mathbf{a}^*$.

Define $Z\sim\mathcal{N}(0,1)$, consider a positive parameter a, and let $\mathbb{E}_{\mathcal{N}(0,a^2)}g = \mathbb{E}g(aZ)$.

Theorem 7 (One step reduction free entropy equivalence)**.** *Consider a $k+1$-layer neural network obtained from the $L+1$-layer neural network performing $L-k \in \{0,\dots,L\}$ steps of reduction. Let*

$$\begin{aligned}\sigma^{(\ell)} &:= \mathbb{E}_{\mathbf{X}^{(\ell)}} \frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d^{(\ell)}} = \mathbb{E}\varphi^2(z\sqrt{\sigma^{(\ell-1)}}) + O\Big(\frac{1}{d^{(\ell-1)}}\Big)\\ \rho^{(\ell)} &:= \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell)})}\varphi'\\ \epsilon^{(\ell)} &:= \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell)})}\varphi^2 - \sigma^{(\ell)}\rho^{(\ell)\,2}\end{aligned} \tag{4.16}$$

that can recursively be defined for any input and hidden layer $\ell = 0,...,k$ of the $k+1$-layer neural network. Suppose H1), H2) and H3) hold. Call $\mathrm{d}^{(k)} = \min_{j=1\cdots k}\{d^{(j)}\}$. Then

$$|\bar{f}_n^{(k)} - \bar{f}_n^{(k-1)}| = O\Big(\sqrt{\Big(1+\frac{n}{\mathrm{d}^{(k-1)}}\Big)\Big(\frac{n}{d^{(k)}} + \frac{n}{\mathrm{d}^{(k-1)\,3/2}} + \frac{1}{\sqrt{\mathrm{d}^{(k-1)}}}\Big)}\Big) \tag{4.17}$$

Through the triangle inequality now, the free entropy of the $L+1$-layer neural network can be related to the one of the GLM introduced in 3.3.

Corollary 8 (Free entropy equivalence)**.** *Under the same hypotheses as Theorem 7, it follows:*

$$\begin{aligned}|\bar{f}_n^{(L)} - \bar{f}_n^{(0)}| &\le \sum_{k=1}^{L} |\bar{f}_n^{(k)} - \bar{f}_n^{(k-1)}|\\ &= \sum_{k=1}^{L} O\Big(\sqrt{\Big(1+\frac{n}{\mathrm{d}^{(k-1)}}\Big)\Big(\frac{n}{d^{(k)}} + \frac{n}{\mathrm{d}^{(k-1)\,3/2}} + \frac{1}{\sqrt{\mathrm{d}^{(k-1)}}}\Big)}\Big)\end{aligned} \tag{4.18}$$

Theorem 7 and Corollary 8 then establish bounds on the difference in free entropy both between the two models considered in the one-step reduction and between the $L+1$-layer network and the generalized linear model. These bounds are crucial as the control on the differences of free entropy is subsequently inherited by the differences in mutual information per sample between dataset and teacher weights, both for the one-step reduction and for the reduction of the whole deep network to the corresponding GLM.

Corollary 9 (One step reduction mutual information equivalence)**.** *Assuming the same hypotheses as in Theorem 7, the following statement is obtained:*

$$|\frac{1}{n}I_n^{(k)}(\boldsymbol{\theta}^{*(k)};\mathcal{D}_n^{(k)}) - \frac{1}{n}I_n^{(k-1)}(\boldsymbol{\theta}^{*(k-1)};\mathcal{D}_n^{(k-1)})| = O\Big(\sqrt{\Big(1+\frac{n}{\mathrm{d}^{(k-1)}}\Big)\Big(\frac{n}{d^{(k)}} + \frac{n}{\mathrm{d}^{(k-1)\,3/2}} + \frac{1}{\sqrt{\mathrm{d}^{(k-1)}}}\Big)}\Big) \tag{4.19}$$

Again, exploiting the triangle inequality, a relation between the mutual information for the $L+1$ layer neural network and the GLM can be obtained.

Corollary 10 (Mutual information equivalence)**.** *The following result holds true under the same assumptions as Theorem 7:*

$$\begin{aligned}
&|\frac{1}{n} I_n^{(L)}(\boldsymbol{\theta}^{*(L)}; \mathcal{D}_n^{(L)}) - \frac{1}{n} I_n^{(0)}(\boldsymbol{\theta}^{*(0)}; \mathcal{D}_n^{(0)})| \\
&\leq \sum_{k=1}^{L} |\frac{1}{n} I_n^{(k)}(\boldsymbol{\theta}^{*(k)}; \mathcal{D}_n^{(k)}) - \frac{1}{n} I_n^{(k-1)}(\boldsymbol{\theta}^{*(k-1)}; \mathcal{D}_n^{(k-1)})| \\
&= \sum_{k=1}^{L} O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(k-1)}}\Big)\Big(\frac{n}{d^{(k)}} + \frac{n}{\mathrm{d}^{(k-1)\ 3/2}} + \frac{1}{\sqrt{\mathrm{d}^{(k-1)}}}\Big)}\Big)
\end{aligned} \tag{4.20}$$

This corollary identifies the scaling regime where the two models achieve equivalence, characterized by the condition where the right-hand side of (4.20) approaches zero as all the involved parameters tend towards infinity. Notably, since this expression is a sum, each individual term within the sum must also tend to zero. This regime for an arbitrary term indexed by k is denoted using the following notation:

$$\widehat{\lim}\, g_{\mathrm{d}^{(k-1)}, d^{(k)}, n} := \lim_{i \to \infty} g_{\mathrm{d}_i^{(k-1)}, d_i^{(k)}, n_i} \tag{4.21}$$

where $(\mathrm{d}_i^{(k-1)}, d_i^{(k)}, n_i)_i$ is any sequence of triplets of integers such that

$$\lim_{i \to \infty} \Big(1 + \frac{n_i}{\mathrm{d}_i^{(k-1)}}\Big)\Big(\frac{n_i}{d_i^{(k)}} + \frac{n_i}{\mathrm{d}_i^{(k-1)\ 3/2}} + \frac{1}{\sqrt{\mathrm{d}_i^{(k-1)}}}\Big) = 0 \tag{4.22}$$

The hatted limit defines the case where each of the variables $\mathrm{d}^{(k-1)}, d^{(k)}, n$ tends to infinity while satisfying the condition (4.22). Specifically, the limit holds true if $d^{(k)} \gg n$ and either $\mathrm{d}^{(k-1)} \gg n$ or $\frac{n}{\mathrm{d}^{(k-1)}} = O(1)$. Notably, even when n is finite, (4.22) is satisfied as long as $\mathrm{d}^{(k-1)}$ and $d^{(k)}$ are taken in the limit. For the right-hand side of (4.20) to disappear, this condition must be verified for any term in the sum, namely, for any value of k. For this purpose, a new limit is introduced, under which, according to Corollary 10, the entire network can be reduced to a GLM:

$$\widetilde{\lim}\, g_{\{d^{(j)}\}_{j=0}^{L}, n} := \lim_{i \to \infty} g_{\{d_i^{(j)}\}_{j=0}^{L}, n_i} \tag{4.23}$$

where $(\{d_i^{(j)}\}_{j=0}^{L}, n_i)_i$ is now any sequence of $L+2$ integers such that

$$\lim_{i \to \infty} \sum_{k=1}^{L} \Big(1 + \frac{n_i}{\mathrm{d}_i^{(k-1)}}\Big)\Big(\frac{n_i}{d_i^{(k)}} + \frac{n_i}{\mathrm{d}_i^{(k-1)\ 3/2}} + \frac{1}{\sqrt{\mathrm{d}_i^{(k-1)}}}\Big) = 0 \tag{4.24}$$

Considering the regime of interest for GLMs where $\frac{n}{d^{(0)}} = O(1)$, it can be observed that this condition can be recovered under the conditions $d^{(1)} \gg n$ and similarly for all subsequent layers indexed

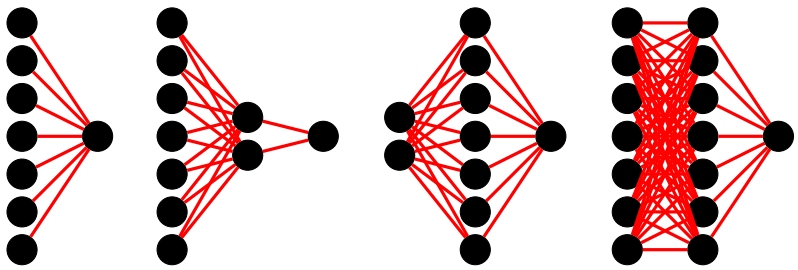

Figure 4.1: neural network architectures studied in literature categorized by how their layers widths scale. Larger layers are symbolized by many neurons, while smaller layers typically consist of one or two neurons. **(1)** corresponds to a generalized linear model. **(2)** describes committee machines, featuring a large input size and a narrow hidden layer. **(3)** illustrates a mean-field regime, characterized by a small input size and a large hidden layer. **(4)** depicts the linear-width regime. Image courtesy of [19].

by k, $d^{(k)} \gg n$. This implies that for deeper layers, $d^{(1)} \gg d^{(k)}$, $d^{(1)} \ll d^{(k)}$ and $d^{(i)}/d^{(j)} = O(1)$ for any layers i and j, are all allowed architectures. Importantly, multiple parameters configurations can satisfy the $\widetilde{\lim}$ conditions, indicating that various architectures of deep networks can be reduced to a GLM.

The aim is now to compare the scalings allowed by our limit with the ones studied in the literature.

Since the nineties, committee machines [23, 24, 25, 21, 17, 26, 15, 78, 59] have been a widely studied type of architecture. Committee machines can be viewed as two-layer neural networks with a narrow hidden layer and a single neuron output layer, in which all the weights are learned. Therefore, they fall within the scope of our investigation. These models exhibit rich phenomenology, including a specialization phase transition where the model realizes that the data is more accurately represented by a model that cannot be separated linearly. The relevant regime for these models is typically characterized by both n and $d^{(0)}$ approaching infinity with $n/d^{(0)} = O(1)$ and $d^{(1)} = O(1)$. However, the infinite size hidden layer limit has also been studied under the condition $d^{(1)} \ll d^{(0)}$, independently of the number of data points. This indicates that while committee machines fit our general framework, the specific scaling properties of our investigation are not fully captured if the hidden layer remains finite in size.

Another interesting model where the relative scalings of the network layers dimensions are considered is the mean-field [13, 79, 80, 81, 16, 18] model, which has been studied for both 2-layer neural networks and deep networks. This regime occurs when the size of the hidden layers is significantly larger than the size of the input layers, i.e., when $d^{(k)} \gg d^{(0)}$ and $d^{(i)}/d^{(j)} = O(1)$ for each pair of layers i, j. Usually in this kind of analysis the number of samples is not explicitly considered because the focus is on the dynamics of stochastic gradient descent (SGD) in a neural network. In both the

mean-field regime and the current model, the fact that $d^{(k)}$ is large introduces intrinsic regularizing properties that must be taken into account.

In recent years, substantial progress has been made in analyzing the full training of deep neural networks, where all weights are learned, using a statistical mechanics framework [27, 28, 29]. As in our investigation, the weights are treated as annealed variables in the free entropy, which is thus computed by integrating them within the partition function. This methodology applies to deeper networks and considers a fully proportional scaling regime, where n and $\{d^{(j)}\}_{j=0}^{L}$ tend to infinity at the same rate, so $n/d^{(k)} = O(1)$ for any k. This scaling regime is often referred to as linear-width regime.

The initial analysis by [27] focused on linear networks, without nonlinearities. Extensions were explored by [28] who formulated conjectures for the empirical risk minimization generalization error using this statistical mechanics framework. Further developments were made in [29], that leveraged a Gaussian equivalence principles [82, 83] to derive exact Bayes-optimal limits, similar to the approach discussed here.

Compared to these analyses, however, a notable issue arises with the scaling regime $\widetilde{\lim}$. This framework does not accommodate the scaling regime where $n/d^{(k)} = O(1)$ meaning that the sizes of the hidden layers and the dimension of the dataset cannot simultaneously tend to infinity under our scaling. It remains uncertain whether this limitation is fundamental or an obstacle specific to the proof, and further investigation is needed to determine the underlying cause and whether the approach can be adapted to include this scaling regime. It is also worth remarking that the recursion parameters (4.16) and the GLM parameters (4.6) obtained through the recursion are consistent with the definitions and results presented by [29].

4.3 Outline of the proof of Theorem 7

The proof of Theorem 7 is carried out only for the case $k = L$. Indeed, in this case, all necessary quantities and parameters can be defined. Furthermore, by carefully handling the model derived from the linearization process, as elaborated in 5, the necessity of executing the proof for each reduction step is obviated.

The crucial part of the proof involves removing the nonlinearities. The direction of our investigation is inspired by the Gaussian equivalence principles, which are expected to hold due to the high-dimensional nature of the problem. Specifically, these principles are what justifies the linearization step presented in 4.1:

$$\frac{\mathbf{a}^{*\intercal}\mathbf{X}_{\mu}^{(L)}}{\sqrt{d^{(L)}}} = \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_{\mu}^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) \approx \rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_{\mu}^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}}\xi_{\mu}^{*(L-1)} \tag{4.25}$$

appropriately tuning the parameters $\rho^{(L-1)}$ and $\epsilon^{(L-1)}$, whose definition turns out to be the one given in (4.16). From this it can be intuitively understood that $\rho^{(L-1)}$ corresponds to the expected average of φ', as the derivative measures the response to variations in its argument.

The strategy employed follows a common approach used in inference and statistical mechanics of disordered systems, namely interpolation, presented in 3.4.4. This involves constructing a model that combines the $L+1$-layer network and the linearized version, where the latter represents a GLM with respect to its input signal $\mathbf{X}_\mu^{(L-1)}$. This construction is done for both the teacher and the student network:

$$\begin{aligned} S_{t\mu}^{(L)} &:= \sqrt{1-t}\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) + \sqrt{t}\rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{t\epsilon^{(L-1)}}\xi_\mu^{*(L-1)} \\ s_{t\mu}^{(L)} &:= \sqrt{1-t}\frac{\mathbf{a}^{\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{(L)}\,\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) + \sqrt{t}\rho^{(L-1)}\frac{\mathbf{v}^{(L-1)\intercal}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{t\epsilon^{(L-1)}}\xi_\mu^{(L-1)} \end{aligned} \tag{4.26}$$

It can be observed that the teacher interpolating model depends on the weights of the teachers from both the $L+1$-layer network and the corresponding L-layer model. For $t=0$, it represents the linearized model, and for $t=1$, it recovers the full neural network. A student version of the same interpolation is also constructed. Additionally, an interpolating dataset is created, where $Y_{t\mu}^{(L)}$ is generated through an output kernel that depends on the teacher weights of the interpolating model.

$$\mathcal{D}_{n,t}^{(L)} = \{(\boldsymbol{X}_\mu^{(0)}, Y_{t\mu}^{(L)})_{\mu=1}^n\}\,, \quad Y_{t\mu}^{(L)} \sim P_{\text{out}}\big(\cdot \mid S_{t\mu}^{(L)}\big) \tag{4.27}$$

Calling the interpolating teacher weights $\boldsymbol{\Theta}^{*(L)} = \{\mathbf{v}^{*(L-1)}, \boldsymbol{\xi}^{*(L-1)}, \mathbf{a}^*, \mathbf{W}^{*(L)} \dots, \mathbf{W}^{*(1)}\}$, the partition function reads:

$$\mathcal{Z}_t^{(L)} = \mathcal{Z}_t^{(L)}(\mathcal{D}_{n,t}^{(L)}) = \int D\boldsymbol{\Theta}^{(L)} \exp\Big[\sum_{\mu=1}^n u_{Y_{t\mu}}^{(L)}(s_{t\mu})\Big] \tag{4.28}$$

where

$$\mathbb{E}_{(t)}(\cdot) := \mathbb{E}_{\mathbf{a}^*}\mathbb{E}_{\setminus\mathbf{a}^*}(\cdot) = \mathbb{E}_{\mathbf{a}^*}\mathbb{E}_{\mathbf{W}^{*(1)},\dots\mathbf{W}^{*(L-1)},\mathbf{W}^{*(L)},\mathbf{v}^{*(L-1)},\boldsymbol{\xi}^{*(L-1)},\mathbf{X}^{(0)}} \int \prod_{\mu=1}^n dY_{t\mu}^{(L)} e^{u_{Y_{t\mu}}^{(L)}(S_{t\mu})}(\cdot) \tag{4.29}$$

The posterior mean of a function g then is defined as:

$$\langle g\rangle_t^{(L)} = \int D\boldsymbol{\Theta}^{(L)} \exp\Big[\sum_{\mu=1}^n u_{Y_{t\mu}}^{(L)}(s_{t\mu})\Big]\, g \tag{4.30}$$

and the interpolating free entropy is

$$\bar{f}_n^{(L)}(t) := \frac{1}{n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \tag{4.31}$$

For $t=0$, it can be observed that the interpolating free entropy corresponds to the free entropy of the linearized model, and for $t=1$ it represents the network of actual interest. To control the difference in free entropy between these two cases, the method exploited is to compute its derivative

and show that it is uniformly bounded in time by the same order as specified in the theorem. The computation of the derivative leads to the summation of the terms below:

$$\frac{d}{dt}\bar{f}_n^{(L)}(t) = -A_1 + A_2 + A_3 + B \tag{4.32}$$

where

$$\begin{aligned}
A_1 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \frac{\mathbf{a}^{*\intercal}}{\sqrt{(1-t)d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) \\
A_2 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \rho^{(L-1)} \frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{td^{(L-1)}}} \\
A_3 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \sqrt{\frac{\epsilon^{(L-1)}}{t}} \xi_\mu^{*(L-1)} \\
B &:= \frac{1}{n}\mathbb{E}_{(t)} \Big\langle \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(s_{t\mu}) \frac{ds_{t\mu}^{(L)}}{dt} \Big\rangle_t^{(L)}
\end{aligned} \tag{4.33}$$

Using the Nishimori identity, introduced in 3.4.2, it can be shown that the last term B is immediately zero. Due to the signs of the remaining terms, it is expected that $-A_1$ counters $A_2 + A_3$, leading to their mutual cancellation in this specific regime.

The desired approach is to directly integrate the weights of the network, which corresponds to performing integration by parts with respect to these Gaussian variables, as discussed in 3.4.1. However, a complication arises because part of these Gaussian weights is inside φ, preventing the application of Gaussian integration by parts. To address this, the term A_1 is considered, employing a classical mathematical technique: subtracting what is desired, A_{12}, and then adding it back. This procedure then allows to control the difference A_{11} between the initial term A_1 and the subtracted one A_{12}:

$$A_{11} := \frac{1}{2n\sqrt{1-t}}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \Big(\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) - \frac{\rho^{(L-1)}\mathbf{a}^{*\intercal}\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L)}d^{(L-1)}}} \Big) \tag{4.34}$$

$$A_{12} := \frac{1}{2n\sqrt{1-t}}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \frac{\rho^{(L-1)}\mathbf{a}^{*\intercal}\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L)}d^{(L-1)}}} \tag{4.35}$$

Focusing on A_{11}, integration by parts with respect to the readout vector yields the expression:

$$A_{11} = \frac{1}{2}\mathbb{E}_{\mathbf{a}^*}\Big[\underbrace{\Big(\frac{\log \mathcal{Z}_t^{(L)}}{n} - \mathbb{E}_{\setminus \mathbf{a}^*}\frac{\log \mathcal{Z}_t^{(L)}}{n}\Big)}_{(1)}\underbrace{\sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})^\intercal \varphi(\boldsymbol{\alpha}_\nu^{(L-1)}) - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})}{d^{(L-1)}}\Big]}_{(2)}\Big] \tag{4.36}$$

where $\boldsymbol{\alpha}_\mu^{(L-1)} = \frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}$, and the notation $\mathbb{E}_{\mathbf{a}^*}$ is introduced in (4.29).

The off-diagonal terms, namely $\mu \neq \nu$, are considered first. Utilizing the properties of the output kernel and Cauchy-Schwartz's inequality, the terms (1) and (2) are estimated separately. The first part of the expression, (1), corresponds to the variance of the free entropy whose scaling is described in H3) as $O(\sqrt{(\frac{1}{\mathrm{d}^{(L-1)}} + \frac{1}{n})})$. The second part, (2), involves the sum of $O(n^2)$ terms. Leveraging the properties of the output kernel established in Lemma 16, the terms $U_{\mu\nu}$, defined therein, remain bounded conditionally on the weights of the interpolating teacher network. Evaluating the terms within the square brackets reveals that despite the tendency of term (2) to concentrate around zero, its order is $O(n\sqrt{(\frac{1}{d^{(L)}} + \frac{1}{\mathrm{d}^{(L-1)\ 3/2}})})$.

A crucial distinction from [19] in evaluating the contribution (2) is that the representations of the input at any hidden layer, $\mathbf{X}^{(\ell)}$, are not Gaussian. This necessitates additional concentration results for accurately estimating term (2). Specifically, the concentration results focus on the squared norm of the output of a hidden layer, which is a vector, and the scalar product of two of such vectors obtained from different data points. To ensure control over all moments of these quantities, exponential bounds for the probability of their divergence from the mean are obtained. These concentration results show that the representations of the inputs at an arbitrary hidden layer, $\mathbf{X}^{(\ell)}$, exhibit statistical properties similar to those of Gaussian random vectors. This quasi-Gaussian property is essential for completing the proof and accurately estimating term (2).

Simplifying then the dependency on n with the $1/n^{1/2}$ term arising from the concentration of the free entropy results in a dependency on $n/d^{(L)}$ under square root, which plays a crucial role in the analysis as already discussed. Indeed, this off-diagonal term is what limits the applicability of our theorem in the linear-width regime.

Consider now the diagonal part, which compensates for the A_3 contribution. After performing a Gaussian integration by parts with respect to the noise variable on A_3, the difference between the two terms becomes:

$$A_{11}^{\text{diag}} - A_3 = \frac{1}{2}\mathbb{E}_{\mathbf{a}^*}\Big[\underbrace{\Big(\frac{\log \mathcal{Z}_t^{(L)}}{n} - \mathbb{E}_{\setminus \mathbf{a}^*}\frac{\log \mathcal{Z}_t^{(L)}}{n}\Big)}_{(1)}\underbrace{\sum_{\mu=1}^{n}\frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}\Big[\epsilon^{(\ell-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(\ell-1)})\|^2 - \rho^{(\ell-1)}\boldsymbol{\alpha}_\mu^{(\ell-1)\intercal}\varphi(\boldsymbol{\alpha}_\mu^{(\ell-1)})}{d^{(\ell)}}\Big]}_{(2)}\Big] \tag{4.37}$$

Cauchy-Schwartz's inequality is used to evaluate independently the terms (1) and (2). Term (1) then corresponds to the variance of the free entropy, with scaling $O(\sqrt{(\frac{1}{\mathrm{d}^{(L-1)}} + \frac{1}{n})})$. Exploiting Lemma 16 the fraction involving the output kernel can be bounded. However, the differences in the square brackets must also be computed, and through their estimation term (2) yields an order $O(\sqrt{(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 1/2}})})$. The difference $|A_{11}^{\mathrm{diag}} - A_3|$ then tends to zero in the linear-width regime limit, and therefore poses no obstacle to the reduction process in this scaling.

The term A_{12} then is linear, and its contribution balances the A_2 term. Similarly to the analysis of the quantity $|A_{11}^{\mathrm{diag}} - A_3|$, integration by parts with respect to the weights matrix on A_{12} is performed to address the difference. Additional computations lead to establish a bound on the difference $||A_{12} - A_2|$ that is of order $O(\sqrt{(1 + \frac{n}{\mathrm{d}^{(L-1)}})(\frac{n}{\mathrm{d}^{(L-1)}d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}})})$. It can be observed then that also the term $|A_{11} - A_2|$ vanishes in the linear-width scaling regime as the sizes of the layers and dataset dimensions tend to infinity.

Having then evaluated all contributions to the derivative of the free entropy with uniform control over time, the proof is concluded, effectively determining the difference in free entropy between the $L + 1$-layer neural network model and its corresponding L-layer model.

Chapter 5

Proofs

Having presented the main results of our investigation in the previous chapter, the focus now shifts to addressing the proofs that establish these findings. In this chapter, the proofs of Theorem 7 and Corollary 9 are presented, as well as the auxiliary lemmas and results exploited in proving them. To begin with, propositions related to the concept of concentration of measure are introduced, laying the groundwork for the subsequent arguments. Following this, two crucial lemmas that are instrumental in the proof of Theorem 7 are stated and rigorously proved. With these elements, the detailed proof of Theorem 7 is then addressed. Finally, the proof of Theorem 9, the central goal of this investigation, is presented.

The proof of Theorem 7 and its auxiliary lemmas is exclusively developed for the last layer $k = L$, where all the necessary quantities and parameters will be fully defined. Indeed, by appropriately handling the model obtained after linearization, it becomes unnecessary to perform the proof for every reduction step. Specifically, two quantities require consideration: the noise term, characterized by $\xi_\mu^{(L-1)}$, and the multiplicative coefficient $\rho^{(L-1)}$.

The noise contribution is considered first. This term, appearing upon linearization, can be interpreted as a perturbation of the output kernel. However, it can be easily be managed as it can be absorbed into the output distribution by convolving the output kernel with a Gaussian probability distribution. To illustrate this, consider the output of the reduced L-layer model $y = Y_\mu^{(L-1)} = f(\rho^{(L-1)} \frac{\mathbf{v}^{*(L-1)\intercal} \mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}} \xi_\mu^{*(L-1)}; \mathbf{A}_\mu) + \sqrt{\Delta} Z_\mu$, where $\xi_\mu^{*(L-1)}$ and Z_μ are standard Gaussian random variables. A new output kernel can be defined as follows:

$$\tilde{P}_{\text{out}}(y \mid x) := \int dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} P_{\text{out}}(y \mid x + \sqrt{\kappa} z) = \int dP_z(z) P_{\text{out}}(y \mid x + \sqrt{\kappa} z) \tag{5.1}$$

where the variable z is a standard Gaussian random variable used to perform the convolution operation, and $\kappa > 0$ is a positive parameter. This allows to draw the outputs of the L-layer model from this distribution as $Y_\mu^{(L-1)} \sim \tilde{P}_{\text{out}}(\cdot \mid \frac{\rho^{(L-1)} \mathbf{v}^{*(L-1)\intercal} \mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}})$.

Next, the multiplicative coefficient $\rho^{(L-1)}$ needs to be addressed. This coefficient had no correspondence in the $L+1$-layer model. When performing the subsequent reduction step, this coefficient $\rho^{(L-1)}$ will become a multiplicative coefficient for the new interpolating model. The interpolation is defined similarly to what described in (4.26) as:

$$
\begin{aligned}
S_{t\mu}^{(L-1)} &:= \rho^{(L-1)}\left[\sqrt{1-t}\frac{\mathbf{v}^{*(L-1)\intercal}}{\sqrt{d^{(L-1)}}}\varphi\left(\frac{\mathbf{W}^{*(L-1)}\,\mathbf{X}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}}\right)+\sqrt{t}\rho^{(L-2)}\frac{\mathbf{v}^{*(L-2)\intercal}\mathbf{X}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}}+\sqrt{t\epsilon^{(L-2)}}\xi_\mu^{*(L-2)}\right]\\
s_{t\mu}^{(L-1)} &:= \rho^{(L-1)}\left[\sqrt{1-t}\frac{\mathbf{v}^{(L-1)\intercal}}{\sqrt{d^{(L-1)}}}\varphi\left(\frac{\mathbf{W}^{(L-1)}\,\mathbf{x}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}}\right)+\sqrt{t}\rho^{(L-2)}\frac{\mathbf{v}^{(L-2)\intercal}\mathbf{x}_\mu^{(L-2)}}{\sqrt{d^{(L-2)}}}+\sqrt{t\epsilon^{(L-2)}}\xi_\mu^{(L-2)}\right]
\end{aligned}
\tag{5.2}
$$

With the same construction done in 4.3, it can be observed that each term A_1, A_2 and A_3 of the derivative of the interpolating free entropy for this new model is the product of two factors: one is $\rho^{(L-1)}$ and the other factor has the same form and statistical properties as the corresponding term A_i defined for the previous interpolation step. The only difference is that now the representations $\mathbf{X}_\mu^{(L-2)}$ and their dimension appear in the expression.

Considering then the derivative of the interpolating free entropy again, the multiplicative coefficient $\rho^{(L-1)}$ can be factored out from the three terms. This allows to apply the proofs developed for the first reduction step to the remaining factor without any modifications when also using the new output kernel $\tilde{P}_{\text{out}}$. Indeed, the proofs do not depend on the specific layer but just on the form of the interpolation model and the statistical properties of its parameters. Furthermore, the term B retains a form and statistical properties analogous to those derived in the previous interpolation step, albeit being now computed using the new output kernel $\tilde{P}_{\text{out}}$ and the new interpolation model. Similar to the first reduction step, it can be demonstrated to be zero. Consequently, the final estimate for the order of the derivative of the interpolating free entropy remains consistent with the results from the previous reduction step. This is because multiplying by $\rho^{(L-1)}$ which is just a coefficient and thus $O(1)$ does not alter the scaling.

Therefore, the essential structure and validity of the lemmas and proofs are preserved despite the inclusion of the $\rho^{(L-1)}$ coefficient. With the same reasoning, the proof can be generalized and extended similarly for any number k of layers.

Similarly, the proof of Theorem 9 is also elaborated just for the first reduction step. In the subsequent reduction step, through the exploitation of the new output kernel $\tilde{P}_{\text{out}}$, the noise term defined by $\xi_\mu^{(L-1)}$ is reabsorbed into the output probability of the responses. The coefficient $\rho^{(L-1)}$, instead, acts as a multiplicative factor in the interpolating model defined in the lemma's proof, similar to (5.2), and in its derivative. This term affects the scaling estimation solely as a multiplicative coefficient, thus preserving it. This allows to carry out the proof in the same way regardless of the number of layers considered.

5.1 Concentration proofs

The concentration results are essential for proving Theorem 7 and represent an additional requirement necessary for applying the proof scheme presented in [19].
Consider each hidden layer ℓ output vectors $\mathbf{X}_{\mu}^{(\ell)}$ normalized by the square root of their dimension. The need for exponential bounds on the probability of the squared norm and the scalar product of two such vectors deviating from their expectation arises from the need to control every central moment of these quantities, as discussed in 4.2. The specific dependency on the dimension, instead, is motivated by the desire to obtain a dependency on it similar to what would be obtained considering $\mathbf{X}^{(\ell)}$ Gaussian random vectors. In order to state and prove the propositions, two useful lemmas related to sub-Gaussian random variables are first established.

5.1.1 Function of a sub-Gaussian random variable

The concentration of a function of a sub-Gaussian random variable is now analyzed.

Lemma 11. *Let $\mathbf{X} \in \mathbb{R}$ be a sub-Gaussian random variable, i.e. using one of the different characterizations of sub-Gaussianity,*

$$\mathbb{E}e^{X^2/\sigma^2} \leq 2 \tag{5.3}$$

Let $F : \mathbb{R} \longrightarrow \mathbb{R}$ be an L-Lipschitz function. Then the random variable $F(X) - \mathbb{E}F(X)$ is also sub-Gaussian, i.e.

$$\mathbb{E}\exp\frac{(F(X)-\mathbb{E}F(X))^2}{c^2} \leq 2 \tag{5.4}$$

for a constant $c \geq 2L\sigma$.

Proof. Consider an independent copy of the variable X, Y. Remembering that the function $v \mapsto \exp\frac{(t-v)^2}{c^2}$ is a convex function since it is the composition of convex functions and applying Jensen's inequality, the following is obtained:

$$\begin{aligned}
\mathbb{E}\exp\frac{(F(X)-\mathbb{E}F(X))^2}{c^2} &= \mathbb{E}\exp\frac{(F(X)-\mathbb{E}F(Y))^2}{c^2} \leq \mathbb{E}\exp\frac{(F(X)-F(Y))^2}{c^2} \\
&\leq \mathbb{E}\exp\frac{L^2(X-Y)^2}{c^2} \leq \mathbb{E}\exp\frac{2L^2X^2+2L^2Y^2}{c^2} \\
&= \left(\mathbb{E}\exp\frac{2L^2X^2}{c^2}\right)^2 \leq \mathbb{E}\exp\frac{4L^2X^2}{c^2} \leq 2
\end{aligned} \tag{5.5}$$

The previous inequality is true for $c \geq 2L\sigma$ exploiting the sub-Gaussianity of X.

□

Lemma 12. *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector belonging to the set $A_\delta := \left\{\left|\frac{\|\mathbf{X}\|^2}{d} - \sigma\right| < \delta\right\}$, with $\sigma = \mathbb{E}\frac{\|\mathbf{X}\|^2}{d}$. Let $\mathbf{W} \in \mathbb{R}^{p\times d}$ a random matrix such that $W_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Let $\varphi \in C^1$ such that*

$|\varphi'| \leq \bar{K}$. Call $\boldsymbol{\alpha} = \frac{\mathbf{WX}}{\sqrt{d}}$. Then, each component of $\varphi(\boldsymbol{\alpha})$ is sub-Gaussian with respect to the weights, i.e., there exists a positive constant C such that:

$$\mathbb{E}_{\mathbf{W}}[e^{s\varphi(\boldsymbol{\alpha})_i}] \leq e^{s^2C^2\gamma^2} \tag{5.6}$$

with $\gamma^2 = \sigma + \delta$.

Proof. The aim is to prove that, considering just the expectation with respect to the weights, $\varphi(\boldsymbol{\alpha}) - \mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})$ is a sub-Gaussian random variable. The average with respect to the weights is here denoted $\mathbb{E}_{\mathbf{W}}$, and since the average is taken only with respect to the weights, $\mathbf{X}$ is considered fixed.

For any fixed $\mathbf{X}$, $\boldsymbol{\alpha}$ is Gaussian, and specifically $\boldsymbol{\alpha}_i \sim \mathcal{N}(0, v^2 = \frac{\|\mathbf{X}\|^2}{d})$, so each component is also sub-Gaussian with variance proxy v^2. Using a different characterization of sub-Gaussian random variables, for some absolute constant M it holds [68]:

$$\mathbb{E}_{\mathbf{W}}[e^{\frac{[\boldsymbol{\alpha}-\mathbb{E}_{\mathbf{W}}\boldsymbol{\alpha}]_i^2}{M^2v^2}}] \leq 2 \tag{5.7}$$

Now using Lemma 11 this yields:

$$\mathbb{E}_{\mathbf{W}}[e^{\frac{[\varphi(\boldsymbol{\alpha})-\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})]_i^2}{4\bar{K}^2M^2v^2}}] \leq 2 \tag{5.8}$$

or equivalently, there exists a constant R such that

$$\mathbb{E}_{\mathbf{W}}[e^{s[\varphi(\boldsymbol{\alpha})-\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})]_i}] \leq e^{s^2R^2\bar{K}^2v^2} \tag{5.9}$$

Considering now the previous results, and exploiting the independence of the components for fixed $\mathbf{X}$

$$\begin{aligned}\mathbb{E}_{\mathbf{W}}[e^{s\mathbf{u}^\intercal[\varphi(\boldsymbol{\alpha})-\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})]}] &= \mathbb{E}_{\mathbf{W}}[e^{s\sum_i u_i[\varphi(\boldsymbol{\alpha})-\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})]_i}] = \mathbb{E}_{\mathbf{W}}[\prod_i e^{su_i[\varphi(\boldsymbol{\alpha})-\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha})]_i}] \\ &= \prod_i e^{s^2u_i^2R^2\bar{K}^2v^2} = e^{s^2R^2\bar{K}^2v^2}\end{aligned} \tag{5.10}$$

To compute the expectation of φ, consider just one component. Remember that φ is odd in its argument. This implies that when conditioning on $\mathbf{X}$, φ exhibits odd symmetry with respect to the weights. Since the expectation is computed with respect to the weights, and their joint distribution is even-symmetric, the expectation of each component of φ is zero. Additionally, if $\mathbf{X} \in A_\delta$, then $v^2 \leq \sigma + \delta = \gamma^2$. Therefore, the inequality can be rewritten as

$$\mathbb{E}_{\mathbf{W}}[e^{s\varphi(\boldsymbol{\alpha})_i}] \leq e^{s^2R^2\bar{K}^2\gamma^2}, \qquad \mathbb{E}_{\mathbf{W}}[e^{s\mathbf{u}^\intercal[\varphi(\boldsymbol{\alpha})]}] = e^{s^2R^2\bar{K}^2\gamma^2} \tag{5.11}$$

This completes the proof, establishing that $\varphi(\boldsymbol{\alpha})$ and its components are sub-Gaussian. □

5.1.2 Concentration of the norm

The proof is introduced as a step towards the ultimate goal of identifying conditions under which the norm of the output vector of a hidden layer in a deep network, starting from an initial data point, exhibits concentration. The objective is to establish an exponential bound for the probability that the squared norm deviates from its expected value, ensuring control over every moment of this variable.

Proposition 13 (Norm concentration)**.** *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector whose norm satisfies the exponential concentration bound*

$$\mathbb{P}\Big(\Big|\frac{\|\mathbf{X}\|^2}{d} - \sigma\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dC\epsilon^2\Big) \tag{5.12}$$

for any $\epsilon > 0$ and a constant $C > 0$, with $\sigma = \mathbb{E}\frac{\|\mathbf{X}\|^2}{d}$. Let $\varphi \in C^1$ such that $|\varphi'| \leq \bar{K}$. If $\mathbf{W} \in \mathbb{R}^{p\times d}$ is a matrix of i.i.d. centered Gaussian random variables with unit variance, then

$$\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-\min\{p,d\}C_\varphi\epsilon^2\Big) \tag{5.13}$$

for any $\epsilon > 0$ and some constant $C_\varphi > 0$ when $\sqrt{d} \gtrsim 1/\epsilon$.

Proof. Define the event

$$A_\delta := \Big\{\Big|\frac{\|\mathbf{X}\|^2}{d} - \sigma\Big| < \delta\Big\} \tag{5.14}$$

Denote also $\boldsymbol{\alpha} = \mathbf{W}\mathbf{X}/\sqrt{d}$. Partitioning the probability space yields

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq \epsilon\Big) &= \mathbb{P}\Big(\Big|\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p}\Big| \geq \epsilon \cap A_\delta\Big) \\ &\quad + \mathbb{P}\Big(\Big|\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p}\Big| \geq \epsilon \cap \bar{A}_\delta\Big) \\ &\leq \mathbb{P}\Big(\Big|\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p}\Big| \geq \epsilon \cap A_\delta\Big) + \mathbb{P}(\bar{A}_\delta)\end{aligned} \tag{5.15}$$

The following notation is introduced

$$\mathbb{P}_\delta\Big(\Big|\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p}\Big| \geq \epsilon\Big) = \mathbb{P}\Big(\Big|\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p}\Big| \geq \epsilon \cap A_\delta\Big) \tag{5.16}$$

By symmetry, only one side of the concentration bound needs to be verified, so the focus is on the event

$$B_\epsilon := \Big\{\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{p} \geq \epsilon\Big\} \tag{5.17}$$

Denote $\boldsymbol{\alpha}_i = \frac{\mathbf{W}_i \mathbf{X}}{\sqrt{d}}$, so that $\varphi(\alpha_i) = \varphi_i(\boldsymbol{\alpha})$.

Exploiting the exponential Markov inequality, for any $s \in \mathbb{R}_{>0}$ the following holds

$$\begin{aligned}\mathbb{P}_\delta(B_\epsilon) &\leq e^{-\frac{s}{2}(\mathbb{E}\varphi_i^2(\boldsymbol{\alpha})+\epsilon)}\mathbb{E}[\exp\Big[s\frac{\|\varphi(\boldsymbol{\alpha})\|^2}{2p}\Big]\mathbb{I}(A_\delta)] = e^{-\frac{s}{2}(\mathbb{E}\varphi_i^2(\boldsymbol{\alpha})+\epsilon)}\mathbb{E}_{\mathbf{X}}\Big(\mathbb{E}_{\mathbf{W}}\mathbf{e}^{\frac{s}{2}\frac{\varphi_i(\boldsymbol{\alpha})^2}{p}}\Big)^p\mathbb{I}(A_\delta) \\ &= \mathbb{E}_{\mathbf{X}}e^{p\ln(\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}))}e^{-\frac{s}{2}(\mathbb{E}\varphi_i^2(\boldsymbol{\alpha})+\epsilon)}\mathbb{I}(A_\delta)\end{aligned} \tag{5.18}$$

Utilizing the fact that $\ln(1+x) \leq x$ the inequality becomes

$$\ln(\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}) - 1 + 1) \leq \mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}) - 1 \tag{5.19}$$

Expanding the logarithm then yields

$$\ln\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}) \leq \frac{s}{2}\mathbb{E}_{\mathbf{W}}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p} + \sum_{k=2}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\Big(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}\Big)^k \tag{5.20}$$

so substituting back leads to

$$\begin{aligned}\mathbb{P}_\delta(B_\epsilon) &\leq \mathbb{E}_{\mathbf{X}}e^{\frac{s}{2}\mathbb{E}_{\mathbf{W}}\varphi_i^2(\boldsymbol{\alpha})+p\sum_{k=2}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\left(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}\right)^k}e^{-\frac{s}{2}(\mathbb{E}\varphi_i^2(\boldsymbol{\alpha})+\epsilon)}\mathbb{I}(A_\delta) \\ &= \mathbb{E}_{\mathbf{X}}e^{-\frac{s}{2}\epsilon}\underbrace{e^{\frac{s}{2}(\mathbb{E}_{\mathbf{W}}\varphi_i^2(\boldsymbol{\alpha})-\mathbb{E}\varphi_i^2(\boldsymbol{\alpha}))}}_{1}\underbrace{e^{p\sum_{k=2}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\left(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}\right)^k}}_{2}\mathbb{I}(A_\delta)\end{aligned} \tag{5.21}$$

The focus is now on bounding the two indicated terms, with the understanding that s is always positive.

$$\begin{aligned}(\mathbb{E}_{\mathbf{W}}\varphi_i^2(\boldsymbol{\alpha}) - \mathbb{E}\varphi_i^2(\boldsymbol{\alpha})) &\leq |\mathbb{E}_{\mathbf{W}}\varphi_i^2(\boldsymbol{\alpha}) - \mathbb{E}\varphi_i^2(\boldsymbol{\alpha})| = |\mathbb{E}_{\mathbf{W}}\varphi^2(\boldsymbol{\alpha}_i) - \mathbb{E}\varphi^2(\boldsymbol{\alpha}_i)| \\ &\leq |\mathbb{E}_{\mathbf{W}}\varphi^2(\boldsymbol{\alpha}_i) - \mathbb{E}_{z\sim\mathcal{N}(0,1)}\varphi^2(z\sqrt{\sigma})| + |\mathbb{E}_{z\sim\mathcal{N}(0,1)}\varphi^2(z\sqrt{\sigma}) - \mathbb{E}\varphi^2(\boldsymbol{\alpha}_i)|\end{aligned} \tag{5.22}$$

In distribution, it holds

$$\boldsymbol{\alpha}_i = \frac{\mathbf{W}_i\mathbf{X}}{\sqrt{d}} \overset{\text{D}}{=} z\sqrt{\frac{\|\mathbf{X}\|^2}{d}} \tag{5.23}$$

plugging this in the inequality results in

$$\begin{aligned}|\mathbb{E}_{\mathbf{W}}\varphi^2(\boldsymbol{\alpha}_i) - \mathbb{E}\varphi^2(\boldsymbol{\alpha}_i)| \leq &\Big|\mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2\Big(z\sqrt{\frac{\|\mathbf{X}\|^2}{d}}\Big) - \mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2(z\sqrt{\sigma})\Big| \\ &+ \Big|\mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2(z\sqrt{\sigma}) - \mathbb{E}_{\mathbf{X}',z\sim\mathcal{N}(0.1)}\varphi^2\Big(z\sqrt{\frac{\|\mathbf{X}'\|^2}{d}}\Big)\Big|\end{aligned} \tag{5.24}$$

Define now $z(s) = z\sqrt{s\frac{\|\mathbf{X}\|^2}{d} + (1-s)\sigma}$. The first term in (5.24) can be bounded as

$$\begin{aligned}\left|\mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2\left(z\sqrt{\frac{\|\mathbf{X}\|^2}{d}}\right) - \mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2(z\sqrt{\sigma})\right| &\leq \int_0^1 ds\mathbb{E}\left[|\varphi'(z(s))||\varphi(z(s))|\frac{|z|\left|\frac{\|\mathbf{X}\|^2}{d}-\sigma\right|}{\sqrt{s\frac{\|\mathbf{X}\|^2}{d}+(1-s)\sigma}}\right] \\ &\leq \bar{K}^2\int_0^1 ds\mathbb{E}z^2\left|\frac{\|\mathbf{X}\|^2}{d}-\sigma\right| \leq \bar{K}^2\left|\frac{\|\mathbf{X}\|^2}{d}-\sigma\right| \leq \bar{K}^2\delta\end{aligned} \tag{5.25}$$

where in the last line the fact that $\mathbf{X}$ belongs to the set A_δ was used. Turning to the second term in (5.24), it can be rewritten as follows:

$$\begin{aligned}&\left|\mathbb{E}_{\mathbf{X}',z\sim\mathcal{N}(0.1)}\varphi^2\left(z\sqrt{\frac{\|\mathbf{X}'\|^2}{d}}\right) - \mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2(z\sqrt{\sigma})\right| \\ &\leq \mathbb{E}_{\mathbf{X}'}\left|\mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2\left(z\sqrt{\frac{\|\mathbf{X}'\|^2}{d}}\right) - \mathbb{E}_{z\sim\mathcal{N}(0.1)}\varphi^2(z\sqrt{\sigma})\right| \\ &\leq \bar{K}^2\mathbb{E}_{\mathbf{X}'}\left|\frac{\|\mathbf{X}'\|^2}{d}-\sigma\right| \leq \frac{\bar{K}^2K}{\sqrt{d}}\end{aligned} \tag{5.26}$$

Plugging everything back in (5.22) leads to:

$$|\mathbb{E}_{\mathbf{W}}\varphi_i^2(\boldsymbol{\alpha}) - \mathbb{E}\varphi_i^2(\boldsymbol{\alpha})| \leq \bar{C}\left(\delta + \frac{1}{\sqrt{d}}\right) \tag{5.27}$$

for some constant $\bar{C}$. Moving to the second term of (5.21), the sub-Gaussianity of each component $\varphi(\alpha_i)$ can be exploited for $\mathbf{X}$ belonging to the set A_δ, that implies, for a positive constant $\tilde{C}$:

$$\mathbb{E}_{\mathbf{W}}|\varphi(\boldsymbol{\alpha})_i|^k \leq 2\tilde{C}^k\gamma^k\Gamma\left(\frac{k}{2}+1\right) \tag{5.28}$$

Thus, recognizing a geometric series, if $\frac{s}{2p}\tilde{C}^2\gamma^2 < 1$ the exponent becomes:

$$p\sum_{k=2}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\left(\frac{s}{2}\frac{\varphi_i^2(\boldsymbol{\alpha})}{p}\right)^k \leq 2p\sum_{k=2}\left(\frac{s}{2p}\tilde{C}^2\gamma^2\right)^k = 2p\left(\frac{s}{2p}\tilde{C}^2\gamma^2\right)^2\sum_{k=0}\left(\frac{s}{2p}\tilde{C}^2\gamma^2\right)^k = \frac{\tilde{C}^4\gamma^4}{2p}\frac{s^2}{1-\frac{s}{2p}\tilde{C}^2\gamma^2} \tag{5.29}$$

Plugging now everything together in (5.21)

$$\mathbb{P}_\delta(B_\epsilon) \leq \mathbb{E}_{\mathbf{X}}e^{\frac{s}{2}\left(-\epsilon+C\left(\delta+\frac{1}{\sqrt{d}}\right)\right)+\frac{\tilde{C}^4\gamma^4}{2p}\frac{s^2}{1-\frac{s}{2p}\tilde{C}^2\gamma^2}}\mathbb{I}(A_\delta) \tag{5.30}$$

and optimizing with respect to s:

$$s_{opt} = \frac{2p}{\tilde{C}^2\gamma^2}\Big(1 - \frac{1}{\sqrt{1 + \Big(\frac{\epsilon}{2} - \frac{\bar{C}}{2}\Big(\delta + \frac{1}{\sqrt{d}}\Big)\Big)\frac{1}{\tilde{C}^2\gamma^2}}}\Big) \tag{5.31}$$

Consider now δ as follows:

$$\delta = -\frac{1}{\sqrt{d}} + \frac{\epsilon}{2\bar{C}} \tag{5.32}$$

Since this quantity needs to be positive, it follows $\sqrt{d} > \frac{2\bar{C}}{\epsilon}$. Additionally, by exploiting $\sqrt{1+x} \leq 1 + \frac{x}{2}$, it is also found:

$$\begin{aligned} s_{opt} &= \frac{2p}{\tilde{C}^2\gamma^2}\Big(1 - \frac{1}{\sqrt{1 + \Big(\frac{\epsilon}{2} - \frac{\bar{C}}{2}\Big(\delta + \frac{1}{\sqrt{d}}\Big)\Big)\frac{1}{\tilde{C}^2\gamma^2}}}\Big) \\ &\leq \frac{2p}{\tilde{C}^2\gamma^2}\Big(1 - \frac{1}{1 + \frac{1}{2}\Big(\frac{\epsilon}{2} - \frac{\bar{C}}{2}\Big(\delta + \frac{1}{\sqrt{d}}\Big)\Big)\frac{1}{\tilde{C}^2\gamma^2}}\Big) \\ &= s' \leq \frac{2p}{\tilde{C}^2\gamma^2} \end{aligned} \tag{5.33}$$

This implies that the required condition $\frac{s_{opt}}{2p}\tilde{C}^2\gamma^2 < 1$ is satisfied.

Plugging s' into (5.30), yields:

$$\mathbb{P}_\delta(B_\epsilon) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{16\left(\tilde{C}^4\gamma^4 + \tilde{C}^2\gamma^2\frac{\epsilon}{8}\right)}}\mathbb{I}(A_\delta) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{max\left\{1,\frac{\epsilon}{8}\right\}16\left(\tilde{C}^4\gamma^4 + \tilde{C}^2\gamma^2\right)}}\mathbb{I}(A_\delta) \tag{5.34}$$

Using now $\mathbb{E}\mathbb{I}(A_\delta) \leq 1$ an exponential bound for the term is obtained:

$$\mathbb{P}_\delta(B_\epsilon) \leq e^{-pC'\epsilon^2} \tag{5.35}$$

In an analogous way, the other side of the inequality can be proved. Getting back to (5.15), the next step involves substituting the chosen value of δ into $\mathbb{P}(\bar{A}_\delta)$ to obtain the desired probability. The expression becomes

$$\mathbb{P}(\bar{A}_\delta) \leq 2\exp\Big(-dC\delta^2\Big) = 2e^{\left(-C - \frac{Cd\epsilon^2}{4\bar{C}^2} + \frac{C\sqrt{d}\epsilon}{\bar{C}}\right)} \leq 2e^{-C}e^{-\frac{Cd\epsilon^2}{20\bar{C}^2}} \tag{5.36}$$

under the condition $\sqrt{d} \geq \frac{5\bar{C}}{\epsilon}$. This implies that there exist a suitable constant C_φ such that for $\epsilon > 0$ small enough and d big enough,

$$\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{WX}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{WX}/\sqrt{d})\|^2}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-\min\{p,d\}C_\varphi\epsilon^2\Big) \tag{5.37}$$

□

5.1.3 Concentration of the scalar product

The upcoming proposition is formulated with the final objective to examine the tendency of the scalar product of two vectors output of a hidden layer of a deep neural network to concentrate around zero. These vectors are derived from two distinct data points. Again, an exponential bound on the probability of deviating from zero is desired, since it allows control over every moment of this random variable.

Proposition 14 (Scalar product concentration)**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be random vectors drawn independently from the same distribution, that satisfy the exponential concentration bounds*

$$\mathbb{P}\Big(\Big|\frac{\mathbf{X}^\intercal \mathbf{Y}}{d}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dB\epsilon^2\Big) \tag{5.38}$$

$$\mathbb{P}\Big(\Big|\frac{\|\mathbf{X}\|^2}{d} - \sigma\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dC\epsilon^2\Big) \qquad \mathbb{P}\Big(\Big|\frac{\|\mathbf{Y}\|^2}{d} - \sigma\Big| \geq \epsilon\Big) \leq 2\exp\Big(-dC\epsilon^2\Big) \tag{5.39}$$

for any $\epsilon > 0$ and constants $B, C > 0$, with $\sigma = \mathbb{E}\frac{\|\mathbf{X}\|^2}{d}$. Let $\varphi \in C^1$ such that $|\varphi'| \leq \bar{K}$. If $\mathbf{W} \in \mathbb{R}^{p\times d}$ is a matrix of i.i.d. centered Gaussian random variables with unit variance, then

$$\mathbb{P}\Big(\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal \varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-\min\{p,d\}C_\varphi\epsilon^2\Big) \tag{5.40}$$

for any $\epsilon > 0$ and some constant $C_\varphi > 0$.

Proof. Define the events

$$A_{\delta,\mathbf{X}} := \Big\{\Big|\frac{\|\mathbf{X}\|^2}{d} - \sigma\Big| < \delta\Big\}, \qquad B_\eta := \Big\{\Big|\frac{\mathbf{X}^\intercal\mathbf{Y}}{d}\Big| < \eta\Big\} \tag{5.41}$$

and denote $\boldsymbol{\alpha}_\mathbf{X} = \mathbf{W}\mathbf{X}/\sqrt{d}$. By partitioning the probability space

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal \varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big| \geq \epsilon\Big)\\
&= \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}}\Big) + \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap \overline{(A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}})}\Big)\\
&\leq \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}}\Big) + \mathbb{P}(\overline{A_{\delta,\mathbf{X}}}) + \mathbb{P}(\overline{A_{\delta,\mathbf{Y}}})\\
&\leq \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}} \cap B_\eta\Big)\\
&+ \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}} \cap \bar{B}_\eta\Big) + 2\mathbb{P}(\overline{A_{\delta,\mathbf{X}}})\\
&\leq \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_\mathbf{X})^\intercal\varphi(\boldsymbol{\alpha}_\mathbf{Y})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}} \cap B_\eta\Big) + \mathbb{P}(\bar{B}_\eta) + 2\mathbb{P}(\overline{A_{\delta,\mathbf{X}}})
\end{aligned} \tag{5.42}$$

The following notation is used

$$\mathbb{P}_{\delta,\eta}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X}})^{\intercal}\varphi(\boldsymbol{\alpha}_{\mathbf{Y}})}{p}\Big| \geq \epsilon\Big) = \mathbb{P}\Big(\Big|\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X}})^{\intercal}\varphi(\boldsymbol{\alpha}_{\mathbf{Y}})}{p}\Big| \geq \epsilon \cap A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}} \cap B_{\eta}\Big) \tag{5.43}$$

By symmetry, only one side of the desired concentration bound needs to be proved, so the following event is considered

$$B_{\epsilon} := \Big\{\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X}})^{\intercal}\varphi(\boldsymbol{\alpha}_{\mathbf{Y}})}{p} \geq \epsilon\Big\} \tag{5.44}$$

Let $\boldsymbol{\alpha}_i = \frac{\mathbf{W}_i\mathbf{X}}{\sqrt{d}}$, so that $\varphi(\alpha_{\mathbf{X},i}) = \varphi_i(\boldsymbol{\alpha}_{\mathbf{X}})$. Exploiting the exponential Markov inequality, for any $s \in \mathbb{R}_{>0}$ it follows

$$\begin{aligned}\mathbb{P}_{\delta,\eta}(B_{\epsilon}) &\leq e^{-\frac{s\epsilon}{2}}\mathbb{E}[\exp\Big[s\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X}})^{\intercal}\varphi(\boldsymbol{\alpha}_{\mathbf{Y}})}{2p}\Big]\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}})] = e^{-\frac{s\epsilon}{2}}\mathbb{E}_{\mathbf{X}}\Big(\mathbb{E}_{\mathbf{W}}e^{\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}}\Big)^{p}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \\ &= \mathbb{E}_{\mathbf{X}}e^{p\ln(\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}))}e^{-\frac{s\epsilon}{2}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}})\end{aligned} \tag{5.45}$$

Using again the relation $\ln(1+x) \leq x$ leads to

$$\ln(\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mu,i})\varphi(\boldsymbol{\alpha}_{\nu,i})}{p}) - 1 + 1) \leq \mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mu,i})\varphi(\boldsymbol{\alpha}_{\nu,i})}{p}) - 1 \tag{5.46}$$

and expanding the logarithm results in:

$$\begin{aligned}p\ln\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}) &\leq p\sum_{k=1}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\Big(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}\Big)^{k} \\ &= \frac{s}{2}\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i}) + p\sum_{k=2}\mathbb{E}_{\mathbf{W}}\frac{1}{k!}\Big(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}\Big)^{k}\end{aligned} \tag{5.47}$$

The aim is bounding the two terms, and $\mathbb{E}_{\mathbf{W}}\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})$ is considered first.

In what follows the i-index is dropped for brevity. Let

$$\alpha_{\mathbf{X}\perp\mathbf{Y}} := \alpha_{\mathbf{X}} - \alpha_{\mathbf{Y}}\frac{\mathbb{E}_{\mathbf{W}}\alpha_{\mathbf{X}}\alpha_{\mathbf{Y}}}{\mathbb{E}^2_{\mathbf{W}}\alpha_{\mathbf{Y}}} = \alpha_{\mathbf{X}} - \alpha_{\mathbf{Y}}\frac{\mathbf{X}^{\intercal}\mathbf{Y}}{\|\mathbf{Y}\|^2} \tag{5.48}$$

that is independent of $\alpha_{\mathbf{Y}}$. Now φ is expanded around $\alpha_{\mathbf{X}\perp\mathbf{Y}}$ defining

$$\alpha_{\mathbf{X},\mathbf{Y}}(s) = s\alpha_{\mathbf{X}} + (1-s)\alpha_{\mathbf{X}\perp\mathbf{Y}} = \alpha_{\mathbf{X}\perp\mathbf{Y}} + s\alpha_{\mathbf{Y}}\mathbf{X}^{\intercal}\mathbf{Y}/\|\mathbf{Y}\|^2 = \alpha_{\mathbf{X}} - (1-s)\alpha_{\mathbf{Y}}\mathbf{X}^{\intercal}\mathbf{Y}/\|\mathbf{Y}\|^2 \tag{5.49}$$

$$\begin{aligned}\mathbb{E}_{\mathbf{W}}\varphi(\alpha_{\mathbf{X}})\varphi(\alpha_{\mathbf{Y}}) &= \mathbb{E}_{\mathbf{W}}\Big[\int_0^1 ds\,\varphi'(\alpha_{\mathbf{X},\mathbf{Y}}(s))\varphi(\alpha_{\mathbf{Y}})\alpha_{\mathbf{Y}}\Big]\frac{\mathbf{X}^\intercal\mathbf{Y}}{\|\mathbf{Y}\|^2}\\ &\leq \Big[\int_0^1 ds\,\mathbb{E}_{\mathbf{W}}|\varphi'(\alpha_{\mathbf{X},\mathbf{Y}}(s))\varphi(\alpha_{\mathbf{Y}})\alpha_{\mathbf{Y}}|\Big]\frac{\mathbf{X}^\intercal\mathbf{Y}}{\|\mathbf{Y}\|^2}\\ &\leq \bar{K}^2\eta\end{aligned} \tag{5.50}$$

where the fact that the zero-th order term is zero due to the odd nature of φ was leveraged, and the Lipschitz property of φ was then utilized.

The second term, $\mathbb{E}_{\mathbf{W}}\varphi^k(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi^k(\boldsymbol{\alpha}_{\mathbf{Y},i})$, is now considered. The sub-Gaussianity of each component $\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})$ is exploited for $\mathbf{X}$ belonging to the set $A_{\delta,\mathbf{X}}$, which implies, for a positive constant $\tilde{C}$

$$\mathbb{E}_{\mathbf{W}}|\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})|^k \leq 2\tilde{C}^k\gamma^k\Gamma\Big(\frac{k}{2}+1\Big) \tag{5.51}$$

leads to

$$\mathbb{E}_{\mathbf{W}}|\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})|^k \leq 2(\tilde{C}^2\gamma^2)^k\Gamma(k+1) \tag{5.52}$$

Plugging the results back in (5.47), the following is obtained

$$\begin{aligned}p\ln\mathbb{E}_{\mathbf{W}}\exp\Big(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}\Big) &\leq \frac{s}{2}\bar{K}^2\eta + 2p\sum_{k=2}\Big(\frac{s}{2p}\tilde{C}^2\gamma^2\Big)^k\\ &= \frac{s}{2}\bar{K}^2\eta + 2p\Big(\frac{s}{2p}\tilde{C}^2\gamma^2\Big)^2\sum_{k=0}\Big(\frac{s}{2p}\tilde{C}^2\gamma^2\Big)^k\\ &= \frac{s}{2}\bar{K}^2\eta + \frac{\tilde{C}^4\gamma^4}{2p}\frac{s^2}{1-\frac{s}{2p}\tilde{C}^2\gamma^2}\end{aligned} \tag{5.53}$$

Recognizing a geometric series if $\frac{s}{2p}\tilde{C}^2\gamma^2 < 1$. Using this in (5.45) results in

$$\begin{aligned}\mathbb{P}_{\delta,\eta}(B_\epsilon) &\leq \mathbb{E}_{\mathbf{X}}e^{p\ \ln(\mathbb{E}_{\mathbf{W}}\exp(\frac{s}{2}\frac{\varphi(\boldsymbol{\alpha}_{\mathbf{X},i})\varphi(\boldsymbol{\alpha}_{\mathbf{Y},i})}{p}))}e^{-\frac{s\epsilon}{2}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}})\\ &\leq \mathbb{E}_{\mathbf{X}}e^{\frac{\tilde{C}^4\gamma^4}{2p}\frac{s^2}{1-\frac{s}{2p}\tilde{C}^2\gamma^2}}e^{\frac{s}{2}(-\epsilon+\bar{K}^2\eta)}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}})\end{aligned} \tag{5.54}$$

and optimizing with respect to s:

$$s_{opt} = \frac{2p}{\tilde{C}^2\gamma^2}\Big(1-\frac{1}{\sqrt{1+\Big(\frac{\epsilon}{2}-\frac{\bar{K}^2\eta}{2}\Big)\frac{1}{\tilde{C}^2\gamma^2}}}\Big) \tag{5.55}$$

Consider η as follows

$$\eta = \frac{\epsilon}{2\bar{K}^2} \tag{5.56}$$

which always exists and is positive. Exploiting that $\sqrt{1+x} \leq 1 + \frac{x}{2}$ results in

$$s_{opt} = \frac{2p}{\tilde{C}^2\gamma^2}\Big(1 - \frac{1}{\sqrt{1 + \Big(\frac{\epsilon}{2} - \frac{\bar{K}^2\eta}{2}\Big)\frac{1}{\tilde{C}^2\gamma^2}}}\Big) \leq \frac{2p}{\tilde{C}^2\gamma^2}\Big(1 - \frac{1}{1 + \frac{1}{2}\Big(\frac{\epsilon}{2} - \frac{\bar{K}^2\eta}{2}\Big)\frac{1}{\tilde{C}^2\gamma^2}}\Big) = s' \leq \frac{2p}{\tilde{C}^2\gamma^2} \tag{5.57}$$

Thus the needed condition for the convergence of the geometric series $\frac{s_{opt}}{2p}\tilde{C}^2\gamma^2 < 1$ is obtained. Plugging s' into (5.30), and choosing $\epsilon < 8$:

$$\mathbb{P}_{\delta,\eta}(B_\epsilon) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{16\left(\tilde{C}^4\gamma^4 + \tilde{C}^2\gamma^2\frac{\epsilon}{8}\right)}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{16\left(\tilde{C}^4\gamma^4 + \tilde{C}^2\gamma^2\right)}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \tag{5.58}$$

Remembering that, for positive x it holds true $x + x^2 \leq (1+x)^2$, and $\gamma^2 = \sigma + \delta$ becomes a constant once δ is fixed

$$\mathbb{P}_{\delta,\eta}(B_\epsilon) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{16\left(\tilde{C}^4\gamma^4 + \tilde{C}^2\gamma^2\right)}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \leq \mathbb{E}_{\mathbf{X}} e^{-\frac{p\epsilon^2}{16\left(1 + \tilde{C}^2\gamma^2\right)^2}}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \tag{5.59}$$

using now $\mathbb{E}\mathbb{I}(A_{\delta,\mathbf{X}})\mathbb{I}(A_{\delta,\mathbf{Y}}) \leq 1$ an exponential bound for the term is obtained:

$$\mathbb{P}_{\delta,\eta}(B_\epsilon) \leq e^{-pC'\epsilon^2} \tag{5.60}$$

In a similar way, the other side of the inequality can be proved. Returning to (5.42), the next step involves substituting the values of η and δ considered into $\mathbb{P}(\bar{B}_\eta)$ and $\mathbb{P}(\overline{A_{\delta,\mathbf{Y}} \cap A_{\delta,\mathbf{Y}}})$ to verify that an exponential bound is obtained. If δ is chosen as $\sqrt{\frac{\ln 2}{dC} + \epsilon^2}$, then

$$\mathbb{P}(\bar{B}_\eta) + \mathbb{P}(\overline{A_{\delta,\mathbf{X}} \cap A_{\delta,\mathbf{Y}}}) \leq 2e^{-d\frac{C}{4\bar{K}^2}\epsilon^2} + 2e^{-dC\epsilon^2} \tag{5.61}$$

This implies that there exist a suitable constant C_φ such that for $\epsilon > 0$ small enough,

$$\mathbb{P}\Big(\Big|\frac{\varphi(\mathbf{WX}/\sqrt{d})^\intercal\varphi(\mathbf{WY}/\sqrt{d})}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big(-min\{p,d\}C_\varphi\epsilon^2\Big) \tag{5.62}$$

□

Considering the vectors resulting from multiplying the initial vectors with a Gaussian random matrix and subsequently applying a nonlinear function, Propositions 13 and 14 enable the derivation of bounds on every central moment of their squared norm and scalar product.

Corollary 15 (Control on moments)**.** *Let $k \in \mathbb{N}$. Under the same hypotheses of Proposition 14 the following holds true:*

$$\begin{aligned}\mathbb{E}\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big|^k &= O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big)\\ \mathbb{E}\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal\varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big|^k &= O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big)\end{aligned} \tag{5.63}$$

Proof. Recall that, being X a random variable and $k \in (0, +\infty)$ [68]

$$\mathbb{E}|X|^k = \int_0^\infty kt^{k-1}\mathbb{P}(|X| \geq t)\,\mathrm{d}t \tag{5.64}$$

This formula allows to bound every central moment of the squared norm and the scalar product.

The integral is well defined for any t when the scalar product is considered, but attention must be paid to the concentration of the squared norm. Remember that Proposition 13 holds for $\sqrt{d} \geq \frac{5\bar{C}}{\epsilon}$, or equivalently $\epsilon \geq \frac{5\bar{C}}{\sqrt{d}}$. This implies that the exponential concentration of the squared norm cannot be used for lower values of ϵ.

To prove that the desired scaling is obtained, the integral (5.64) is split and the two resulting terms are controlled:

$$\begin{aligned}&\mathbb{E}\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big|^k\\ &= \int_0^{\frac{5\bar{C}}{\sqrt{d}}} kt^{k-1}\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq t\Big)\,\mathrm{d}t\\ &+ \int_{\frac{5\bar{C}}{\sqrt{d}}}^\infty kt^{k-1}\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq t\Big)\,\mathrm{d}t\\ &\leq \Big(\frac{5\bar{C}}{\sqrt{d}}\Big)^k + \int_{\frac{5\bar{C}}{\sqrt{d}}}^\infty kt^{k-1}\mathbb{P}\Big(\Big|\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p} - \mathbb{E}\frac{\|\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})\|^2}{p}\Big| \geq t\Big)\,\mathrm{d}t = O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big)\end{aligned} \tag{5.65}$$

As for the concentration of the scalar product, using directly (5.64) leads to:

$$\mathbb{E}\Big|\frac{\varphi(\mathbf{W}\mathbf{X}/\sqrt{d})^\intercal\varphi(\mathbf{W}\mathbf{Y}/\sqrt{d})}{p}\Big|^k = O\Big(\frac{1}{min\{p,d\}^{k/2}}\Big) \tag{5.66}$$

□

5.2 Output kernel properties

The following lemma establishes properties of the output kernel that will be extensively used in 5.4. Notably, the output kernel defined in (5.1) is used. This is done in order to ensure the validity of the lemma even when the output kernel is deformed with additive Gaussian noise. This generalization allows the Lemma to be applicable when performing the subsequent reduction steps.

Lemma 16 (Properties of P_{out})**.** *Recall definition* (5.1)*. Denote* $\tilde{u}_y(x) := \log \tilde{P}_{\text{out}}(y \mid x)$*. and* $\tilde{u}'_y(x) := \partial_x \tilde{u}_y(x)$*. Additionally, let*

$$\tilde{U}_{\mu\nu} := \delta_{\mu\nu}\tilde{u}''_{Y_{t\mu}}(S_{t\mu}) + \tilde{u}'_{Y_{t\mu}}(S_{t\mu})\tilde{u}'_{Y_{t\nu}}(S_{t\nu}) \tag{5.67}$$

Then, under Assumptions H1) and H2), for a positive constant $C(f)$ *depending only on the readout function, the following holds:*

$$\mathbb{E}[\tilde{u}'_{Y_{t\mu}}(S_{t\mu}) \mid S_{t\mu}] = \mathbb{E}[\tilde{U}_{\mu\nu} \mid S_{t\mu}, S_{t\nu}] = 0 \tag{5.68}$$

$$\mathbb{E}[(\tilde{u}'_{Y_{t\mu}}(S_{t\mu}))^2 \mid S_{t\mu}]\,,\ \mathbb{E}[\tilde{U}^2_{\mu\nu} \mid S_{t\mu}, S_{t\nu}] \le C(f) \tag{5.69}$$

Remark 1. Notice how for $\mu = \nu$ (5.67) simplifies to $\tilde{U}_{\mu\mu} = \tilde{P}''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})/\tilde{P}_{\text{out}}(Y_{t\mu} \mid S_{t\mu})$, where

$$\tilde{P}'_{\text{out}}(y \mid x) := \partial_x \tilde{P}_{\text{out}}(y \mid x)\,, \qquad \tilde{P}''_{\text{out}}(y \mid x) := \partial_x \partial_x \tilde{P}_{\text{out}}(y \mid x) \tag{5.70}$$

Exploiting the lemma, the result obtained is that

$$\mathbb{E}\Big[\Big(\frac{\tilde{P}''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}\Big)^2 \Big| S_{t\mu}\Big] \le C(f) \tag{5.71}$$

Proof. An auxiliary lemma is first considered. In the lemma the necessary bounds involving the derivatives of the output are estimated. In the following, $C(f)$ denotes a constant that depends on the readout function f and may also depend on Δ. Upper indices denote partial derivatives with respect to that variable, for instance, $\tilde{P}^x_{\text{out}}(y|x) = \partial_x \tilde{P}_{\text{out}}(y|x)$ and $\tilde{P}^{xx}_{\text{out}}(y|x) = \partial_x \partial_x \tilde{P}_{\text{out}}(y|x)$.

Lemma 17. *Let* $y = Y_{t\mu} = f(S_{t\mu} + \sqrt{\kappa}\zeta_\mu; \mathbf{A}_\mu) + \sqrt{\Delta} Z_\mu$*, where* ζ_μ *and* Z_μ *are standard Gaussian random variables, and* $\kappa > 0$ *is constant. Assuming H2), a constant* $C(f)$ *exists such that*

$$\max\Big\{\Big|\frac{\tilde{P}^y_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)}\Big|, \Big|\frac{\tilde{P}^x_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)}\Big|, \Big|\frac{\tilde{P}^{yy}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)}\Big|, \Big|\frac{\tilde{P}^{yx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)}\Big|, \Big|\frac{\tilde{P}^{xx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)}\Big|\Big\} < C(f)(|Z_\mu|^2 + 1) \tag{5.72}$$

Proof. If the ζ_μ term is absent, the proof aligns with [19], where $\tilde{P}_{\text{out}}(y|x) = P_{\text{out}}(y|x)$. Therefore, the analysis focuses on the case where the additional stochastic term ζ_μ is present.

Let $w = x + \sqrt{\kappa}z$. The definition of $P_{\text{out}}(y|w)$ is recalled:

$$P_{\text{out}}(y \mid w) = \int dP_A(\mathbf{A}) \frac{1}{\sqrt{2\pi\Delta}} \exp\Big(-\frac{1}{2\Delta}(y - f(w;\mathbf{A}))^2\Big) = \int P_A(d\mathbf{A}) P(y|w,\mathbf{A}) \tag{5.73}$$

In a similar fashion, a new output kernel is introduced to reabsorb the stochasticity introduced by ζ_μ. Given that ζ_μ is a standard Gaussian random variable, it is possible to write:

$$\tilde{P}_{\text{out}}(y \mid x) := \int dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} P_{\text{out}}(y \mid x + \sqrt{\kappa} z) = \int dP_z(z) P_{\text{out}}(y \mid x + \sqrt{\kappa} z) \tag{5.74}$$

It can be observed now that the ratio of any derivative of $\tilde{P}_{\text{out}}$ to $\tilde{P}_{\text{out}}$ can be formulated by introducing the following expectation:

$$\langle \cdot \rangle_{\mathbf{A},z} := \frac{\int dP_z(z) dP_A(\mathbf{A})(\cdot) e^{-\frac{1}{2\Delta}(y - f(x+\sqrt{\kappa}z;\mathbf{A}))^2}}{\int dP_z(z) dP_A(\mathbf{A}) e^{-\frac{1}{2\Delta}(y - f(x+\sqrt{\kappa}z;\mathbf{A}))^2}} \tag{5.75}$$

Through some algebraic manipulation, the following expressions are obtained:

$$\frac{\tilde{P}^{y}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} = \Big\langle -\frac{1}{\Delta}(y - f(x + \sqrt{\kappa} z; \mathbf{A})) \Big\rangle_{\mathbf{A},z} \tag{5.76}$$

$$\frac{\tilde{P}^{x}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} = \Big\langle \frac{1}{\Delta}(y - f(x + \sqrt{\kappa} z; \mathbf{A})) f'(x + \sqrt{\kappa} z; \mathbf{A}) \Big\rangle_{\mathbf{A},z} \tag{5.77}$$

$$\frac{\tilde{P}^{yy}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} = \Big\langle \frac{1}{\Delta^2}(y - f(x + \sqrt{\kappa} z; \mathbf{A}))^2 \Big\rangle_{\mathbf{A},z} - \frac{1}{\Delta} \tag{5.78}$$

$$\frac{\tilde{P}^{yx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} = \Big\langle -\frac{1}{\Delta^2}(y - f(x + \sqrt{\kappa} z; \mathbf{A}))^2 f'(x + \sqrt{\kappa} z; \mathbf{A}) + \frac{1}{\Delta} f'(x + \sqrt{\kappa} z; \mathbf{A}) \Big\rangle_{\mathbf{A},z} \tag{5.79}$$

$$\frac{\tilde{P}^{xx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} = \Big\langle \Big(\frac{1}{\Delta^2}(y - f(x + \sqrt{\kappa} z; \mathbf{A}))^2 - \frac{1}{\Delta} \Big) f'(x + \sqrt{\kappa} z; \mathbf{A})^2 \tag{5.80}$$

$$+ \frac{1}{\Delta}(y - f(x + \sqrt{\kappa} z; \mathbf{A})) f''(x + \sqrt{\kappa} z; \mathbf{A}) \Big\rangle_{\mathbf{A},z} \tag{5.81}$$

Given that all expressions are similar in form, only the last one is treated, as all others can be bounded using analogous reasoning. It is considered that:

$$\begin{aligned} \Big| \frac{\tilde{P}^{xx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} \Big| \leq & \Big\langle \Big(\frac{1}{\Delta^2}(y - f(x + \sqrt{\kappa} z; \mathbf{A}))^2 + \frac{1}{\Delta} \Big) f'(x + \sqrt{\kappa} z; \mathbf{A})^2 \Big\rangle_{\mathbf{A},z} \\ & + \Big\langle \frac{1}{\Delta} \Big| y - f(x + \sqrt{\kappa} z; \mathbf{A}) \Big| \Big| f''(x + \sqrt{\kappa} z; \mathbf{A}) \Big| \Big\rangle_{\mathbf{A},z} \end{aligned} \tag{5.82}$$

Since $= Y_{t\mu} = f(S_{t\mu} + \sqrt{\kappa}\zeta_\mu; \mathbf{A}_\mu) + \sqrt{\Delta} Z_\mu$, the previous quantities can be rewritten in terms of Z_μ. Noting that the brackets are not averaging over Z_μ, Z_μ, and its functions can be brought out of the average. Given that f and its first two derivatives are bounded (as specified in H2)), the following result is obtained:

$$\Big| \frac{\tilde{P}^{xx}_{\text{out}}(y|x)}{\tilde{P}_{\text{out}}(y|x)} \Big| \leq \Big(\frac{1}{\Delta} |Z_\mu|^2 + \frac{1}{\Delta} \Big) K(f) + \frac{1}{\Delta} |Z_\mu| G(f) \leq (|Z_\mu|^2 + 1) C(f) \tag{5.83}$$

□

Getting back to the proof of Lemma 16, (5.68) is first proved:

$$\begin{aligned}\mathbb{E}[\tilde{u}'_{Y_{t\mu}}(S_{t\mu})|S_{t\mu}] &= \mathbb{E}\Big[\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big|S_{t\mu}\Big] \\ &= \int dY_{t\mu}\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})} = \int dY_{t\mu}\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu}) \\ &= \frac{\partial}{\partial S_{t\mu}}\int dY_{t\mu}\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu}) = \frac{\partial}{\partial S_{t\mu}}1 = 0\end{aligned} \tag{5.84}$$

Concerning then (5.69), the conditional expectation of $\tilde{U}_{\mu\nu}$ given $S_{t\mu}$ and $S_{t\nu}$ is computed separately for the cases $\mu = \nu$ and $\mu \neq \nu$. The scenario $\mu = \nu$ is first addressed.

$$\begin{aligned}\mathbb{E}[\tilde{U}_{\mu\mu}|S_{t\mu}] &= \mathbb{E}\Big[\partial_x\Big(\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big) + \Big(\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big)^2\Big|S_{t\mu}\Big] \\ &= \mathbb{E}\Big[\frac{\tilde{P}^{xx}_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})} - \Big(\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big)^2 + \Big(\frac{\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big)^2\Big|S_{t\mu}\Big] \\ &= \mathbb{E}\Big[\frac{\tilde{P}^{xx}_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big|S_{t\mu}\Big] \\ &= \int dY_{t\mu}\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})\frac{\tilde{P}^{xx}_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})} = \int dY_{t\mu}\tilde{P}^{xx}_{\text{out}}(Y_{t\mu}|S_{t\mu}) \\ &= \frac{\partial^2}{\partial S^2_{t\mu}}\int dY_{t\mu}\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu}) = \frac{\partial^2}{\partial S^2_{t\mu}}1 = 0\end{aligned} \tag{5.85}$$

Considering the terms characterized by $\mu \neq \nu$, it can be observed that when conditioning on $S_{t\mu},\ S_{t\nu}$, the remaining stochasticity in the responses is independent for each sample. Consequently, $\tilde{u}'_{Y_{t\mu}}(S_{t\mu})$ and $\tilde{u}'_{Y_{t\nu}}(S_{t\nu})$ are conditionally independent given $S_{t\mu},\ S_{t\nu}$, resulting in:

$$\begin{aligned}\mathbb{E}[\tilde{U}_{\mu\nu}|S_{t\mu}, S_{t\nu}] &= \mathbb{E}[\tilde{u}'_{Y_{t\mu}}(S_{t\mu})\tilde{u}'_{Y_{t\nu}}(S_{t\nu})|S_{t\mu}, S_{t\nu}] \\ &= \int dY_{t\mu}\tilde{P}^x_{\text{out}}(Y_{t\mu}|S_{t\mu})\int dY_{t\nu}\tilde{P}^x_{\text{out}}(Y_{t\nu}|S_{t\nu}) = 0\end{aligned} \tag{5.86}$$

The aim is now to address now the boundedness of the expressions in (5.69). Lemma 17 allows to write:

$$(\tilde{u}'_{Y_{t\mu}}(S_{t\mu}))^2 \leq C^2(f)(|Z_\mu|^2+1)^2 \leq C'^2(f)(|Z_\mu|^4+1) \tag{5.87}$$

Using Jensen's inequality, specifically that $(|Z_\mu|^2+1)^2 \leq 2(|Z_\mu|^4+1)$. Since Z_μ is a standard Gaussian random variable and does not depend on $S_{t\mu}$, it has finite second and fourth moments. Therefore, computing the conditional expectation of the left hand side on $S_{t\mu}$ leads to bound $(\tilde{u}'_{Y_{t\mu}}(S_{t\mu}))^2$ with

an appropriate constant $C(f)$. The expectation of $\tilde{U}^2_{\mu\mu}$ and $\tilde{U}^2_{\mu\nu}$ is bounded, conditionally on $S_{t\mu}$ and $S_{t\nu}$. Consider the case $\mu = \nu$. Using what stated in Lemma 17, it follows:

$$\begin{aligned}\tilde{U}^2_{\mu\mu} = \Big(\frac{\tilde{P}^{xx}_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big)^2 &\leq C^2(f)(|Z_\mu|^2+1)^2 \\ &= C^2(f)(|Z_\mu|^4+2|Z_\mu|^2+1)\end{aligned} \tag{5.88}$$

Then, by taking the expectation and using the Gaussianity of Z_μ along with the finiteness of its moments, a bound can be established for an appropriate constant $C(f)$. Considering now the case $\mu \neq \nu$, Lemma 17 is exploited:

$$\begin{aligned}\tilde{U}^2_{\mu\nu} = \Big(\frac{\tilde{P}^{x}_{\text{out}}(Y_{t\mu}|S_{t\mu})}{\tilde{P}_{\text{out}}(Y_{t\mu}|S_{t\mu})}\Big)^2\Big(\frac{\tilde{P}^{x}_{\text{out}}(Y_{t\nu}|S_{t\nu})}{\tilde{P}_{\text{out}}(Y_{t\nu}|S_{t\nu})}\Big)^2 &\leq C_\mu(f)(|Z_\mu|^2+1)C_\nu(f)(|Z_\nu|^2+1) \\ &= C_\mu(f)C_\nu(f)(|Z_\mu|^2|Z_\nu|^2+|Z_\mu|^2+|Z_\nu|^2+1)\end{aligned} \tag{5.89}$$

This implies that when taking the expectation, a bound that involves a suitable constant $C(f)$ is obtained.

□

5.3 Approximation Lemma

This approximation lemma allows the estimation of the expectation, with respect to the weights, of various expressions involving the derivative or square of the activation function applied to an input from the previous layer. This includes the estimation of terms where the activation functions of different samples interact.

Notably in this context the concentration results obtained in 3.4.3 are utilized to identify the dominant terms in the estimations considered. These concentration results are necessary to manage higher-order central moments of the squared norm and scalar product of $\mathbf{X}^{(\ell)}_\mu$, as well as for handling cases where such norms appear in the denominator.

Lemma 18 (Approximations). *Consider a L-layer neural network as constructed in 3.2. For any layer $\ell \in 0, \dots, L$ call $\sigma^{(\ell)} := \mathbb{E}_{\mathbf{X}^{(\ell)}} \frac{\|\mathbf{X}^{(\ell)}_\mu\|^2}{d^{(\ell)}}$, $\rho^{(\ell)} := \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell)})}\varphi'$ and $\epsilon^{(\ell)} := \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell)})}\varphi^2 - \sigma^{(\ell)}{\rho^{(\ell)}}^2$. Let $\tilde{\varphi}$ be either φ or the identity function, and define $\tilde{\rho}^{(\ell)} := \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell)})}\tilde{\varphi}'$. Under assumptions H2) and H1), the following hold:*

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha^{(\ell-1)}_{\mu i}) = \rho^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}^{(\ell-1)}_\mu\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) \tag{5.90}$$

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha^{(\ell-1)}_{\mu i}) = \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell-1)})}\varphi^2 + O\Big(\frac{\|\mathbf{X}^{(\ell-1)}_\mu\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) \tag{5.91}$$

Define now $A_{\delta\mu}^{(\ell-1)} := \left\{ \left| \frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d} - \sigma^{(\ell-1)} \right| < \delta \right\}$ *and* $A_\delta^{(\ell-1)} = \bigcap_{\mu=1}^{n} A_{\delta\mu}^{(\ell-1)}$. *Assume now that* $\mathbf{X}^{(\ell-1)}$ *belongs to* $A_\delta^{(\ell-1)}$.

$$\begin{aligned}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\mathbb{I}(A_\delta)\varphi(\alpha_{\mu i}^{(\ell-1)})\tilde{\varphi}(\alpha_{\nu i}^{(\ell-1)}) &= \rho^{(\ell-1)}\tilde{\rho}^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}} \\ &+ O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big) \\ &+ O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big) + O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)^2\Big),\end{aligned} \tag{5.92}$$

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\mathbb{I}(A_\delta)\varphi'(\alpha_{\mu i}^{(\ell-1)})\tilde{\varphi}'(\alpha_{\nu i}^{(\ell-1)}) = \rho^{(\ell-1)\ 2} + O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big) \tag{5.93}$$

$$\begin{aligned}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\mathbb{I}(A_\delta)\varphi^2(\alpha_{\mu i}^{(\ell-1)})\tilde{\varphi}^2(\alpha_{\nu i}^{(\ell-1)}) &= \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell-1)})}\varphi^2\mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell-1)})}\tilde{\varphi}^2 + O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) \\ &+ O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\end{aligned} \tag{5.94}$$

Remark 2. Equations (5.90) and (5.91) allow for $\sigma^{(\ell)} = \mathbb{E}_{\mathbf{X}^{(\ell)}}\frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d^{(\ell)}}$ to be written recursively. Indeed, using (5.91), it can be obtained:

$$\begin{aligned}\sigma^{(\ell)} &= \mathbb{E}_{\mathbf{X}^{(\ell)}}\frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d^{(\ell)}} = \mathbb{E}_{\mathbf{X}^{(\ell-1)}}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\frac{\sum_{i=1}^{d^{(\ell)}}\varphi^2(\alpha_{\mu i}^{(\ell-1)})}{d^{(\ell)}} \\ &= \mathbb{E}_{\mathcal{N}(0,\sigma^{(\ell-1)})}\varphi^2 + \mathbb{E}_{\mathbf{X}^{(\ell-1)}}O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) \\ &= \mathbb{E}\varphi^2(z\sqrt{\sigma^{(\ell-1)}}) + O\Big(\frac{1}{d^{(\ell-1)}}\Big)\end{aligned} \tag{5.95}$$

Remark 3. Notice that when $\mathbf{X}^{(\ell-1)} \in A_\delta^{(\ell-1)}$, the following bound holds:

$$\Big|\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big| \leq \Big|\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big|\frac{1}{\sigma^{(\ell-1)} - \delta} \tag{5.96}$$

Proof. Equation (5.90) is first considered. Exploiting that the weights $\mathbf{W}^{(\ell)}$ are Gaussian, in distribution it holds:

$$\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d^{(L-1)}}} \overset{\text{D}}{=} z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}}} \tag{5.97}$$

with $z \sim \mathcal{N}(0,1)$. The expectation with respect to $\mathbf{W}^{(\ell)}$ then becomes an expectation over z, and the application of the fundamental theorem of integral calculus leads to:

$$\begin{aligned}
|E_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_{\mu i}^{(\ell-1)}) - \rho^{(\ell)}| &= \left|\mathbb{E}\varphi'\Big(z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}}}\Big) - \rho^{(\ell)}\right| \\
&= \left|\mathbb{E}\varphi'\Big(z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}}}\Big) - \mathbb{E}\varphi'\Big(z\sqrt{\sigma^{(\ell-1)}}\Big)\right| \\
&\leq \int_0^1 ds \mathbb{E}\frac{|z|}{2}\left|\varphi''\Big(z\sqrt{s\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} + \sigma^{(\ell-1)}(1-s)}\Big)\right| \frac{\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right|}{\sqrt{s\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} + \sigma^{(\ell-1)}(1-s)}}
\end{aligned} \tag{5.98}$$

Using the boundedness of the second derivative, namely $\varphi'' \leq \bar{K}$, a complete decoupling of s from z is achieved. The expectation with respect to z can then be computed remembering, $\mathbb{E}|z| \leq \sqrt{\mathbb{E}z^2} = 1$, and the remaining integral can be computed and bounded as:

$$\begin{aligned}
|E_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_{\mu i}^{(\ell-1)}) - \rho^{(\ell)}| &\leq \frac{\bar{K}}{2}\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right| \int_0^1 ds \frac{1}{\sqrt{s\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} + \sigma^{(\ell-1)}(1-s)}} \\
&= \bar{K}\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right| \frac{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|}{\sqrt{d^{(\ell-1)}}} - \sigma^{(\ell-1)}}{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}} \\
&= \bar{K}\frac{\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right|}{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|}{\sqrt{d^{(\ell-1)}}} + \sqrt{\sigma^{(\ell-1)}}} \leq \tilde{K}\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right|
\end{aligned} \tag{5.99}$$

Consider (5.91). Similarly to what was previously done:

$$\begin{aligned}
&\left|E_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}^2(\alpha_{\mu i}^{(\ell-1)}) - \mathbb{E}_{\mathcal{N}(0,\,\sigma^{(\ell-1)})}\tilde{\varphi}^2\right| \\
&= \left|\mathbb{E}\tilde{\varphi}^2\Big(z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}}}\Big) - \mathbb{E}\tilde{\varphi}^2(z\sqrt{\sigma^{(\ell-1)}})\right| \\
&\leq \bar{K}\int_0^1 ds \mathbb{E}\left|z\tilde{\varphi}\Big(z\sqrt{s\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} + \sigma^{(\ell-1)}(1-s)}\Big)\right| \frac{\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right|}{\sqrt{s\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} + \sigma^{(\ell-1)}(1-s)}}
\end{aligned} \tag{5.100}$$

In order to carry out integration over the variable s, a bound on the expectation $\mathbb{E}|z\tilde{\varphi}(\dots)|$ is needed. This is managed using the fact that φ is Lipschitz. Since $\tilde{\varphi}' \leq \bar{K}$ and $\tilde{\varphi}(0) = 0$ as φ is odd, it follows $|\tilde{\varphi}(\dots)| \leq \bar{K}|(\dots)|$. This implies:

$$\left|E_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}^2(\alpha_{\mu i}^{(\ell-1)}) - \mathbb{E}_{\mathcal{N}(0,\,\sigma^{(\ell-1)})}\tilde{\varphi}^2\right| \leq \bar{K}^2\left|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\right|. \tag{5.101}$$

Moving now to (5.92), the i-index is dropped for simplicity. Define:

$$\alpha_{\mu\perp\nu}^{(\ell-1)} := \alpha_{\mu}^{(\ell-1)} - \alpha_{\nu}^{(\ell-1)} \frac{\mathbb{E}_{\mathbf{W}^{*(\ell)}} \alpha_{\mu}^{(\ell-1)} \alpha_{\nu}^{(\ell-1)}}{\mathbb{E}^2_{\mathbf{W}^{*(\ell)}} \alpha_{\nu}^{(\ell-1)}} = \alpha_{\mu}^{(\ell-1)} - \alpha_{\nu}^{(\ell-1)} \frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \tag{5.102}$$

It is noteworthy that this particular quantity does not depend on $\alpha_{\nu}^{(\ell-1)}$. Subsequently, an expansion of φ around $\alpha_{\mu\perp\nu}^{(\ell-1)}$ is performed. In the following, p defines a point that lies between $\alpha_{\mu\perp\nu}^{(\ell-1)}$ and $\alpha_{\nu}^{(\ell-1)}$. Furthermore, the zero-order term is absent due to the odd nature of φ, and integration by parts with respect to $\mathbf{W}^{*(\ell)}$ is applied to the first-order term.

$$\begin{aligned} \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi(\alpha_{\mu}^{(\ell-1)}) \tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) &= \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi'(\alpha_{\mu\perp\nu}^{(\ell-1)}) \mathbb{E}_{\mathbf{W}^{*(\ell)}} \tilde{\varphi}'(\alpha_{\nu}^{(\ell-1)}) \frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{d^{(\ell-1)}} \\ &+ \frac{1}{2} E_{\mathbf{W}^{*(\ell)}} \varphi''(p) \tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) \alpha_{\nu}^{(\ell-1)\,2} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \end{aligned} \tag{5.103}$$

The objective is now to estimate the residual associated with the last term, thus the modulus of the term is considered:

$$\begin{aligned} &\Big| \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi''(p) \tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) \alpha_{\nu}^{(\ell-1)\,2} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \Big| \\ &\leq \mathbb{E}_{\mathbf{W}^{*(\ell)}} \Big| \varphi''(p) \tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) \alpha_{\nu}^{(\ell-1)\,2} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \Big| \\ &\leq \bar{K} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \mathbb{E}_{\mathbf{W}^{*(\ell)}} |\tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) \alpha_{\nu}^{(\ell-1)\,2}| \\ &\leq \bar{K} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \mathbb{E}_{\mathbf{W}^{*(\ell)}} |\alpha_{\nu}^{(\ell-1)}|^3 \\ &\leq \tilde{K} \Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \sqrt{\Big(\frac{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}{d^{(\ell-1)}} \Big)^3 5!!} \end{aligned} \tag{5.104}$$

where the term under square root is bounded and depends on $(\sigma^{(\ell-1)} + \delta)^{3/2}$ exploiting $\mathbf{X}^{(\ell-1)} \in A_{\delta}^{(\ell-1)}$ and (5.96). This allows to write

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi(\alpha_{\mu}^{(\ell-1)}) \tilde{\varphi}(\alpha_{\nu}^{(\ell-1)}) = \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi'(\alpha_{\mu\perp\nu}^{(\ell-1)}) \mathbb{E}_{\mathbf{W}^{*(\ell)}} \tilde{\varphi}'(\alpha_{\nu}^{(\ell-1)}) \frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{d^{(\ell-1)}} + O\Big(\Big(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \Big)^2 \Big). \tag{5.105}$$

An expansion of $\varphi'(\alpha_{\mu\perp\nu}^{(\ell-1)})$ is performed around the initial point $\alpha_{\mu}^{(\ell-1)}$:

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi'(\alpha_{\mu\perp\nu}) = \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi'(\alpha_{\mu}^{(\ell-1)}) - \mathbb{E}_{\mathbf{W}^{*(\ell)}} \varphi''(p) \alpha_{\nu}^{(\ell-1)} \frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal} \mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \tag{5.106}$$

where the variable p retains its previous definition. The residual associated with the last term is now evaluated by taking the modulus of the term and exploiting the Lipshitzianity of φ:

$$\begin{aligned}\left|\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi''(p)\alpha_\nu^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\right| &\leq \mathbb{E}_{\mathbf{W}^{*(\ell)}}\left|\varphi''(p)\alpha_\nu^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\right| \\ &\leq \bar{K}\left|\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\right|\mathbb{E}_{\mathbf{W}^{*(\ell)}}|\alpha_\nu^{(\ell-1)}| \\ &\leq \bar{K}\left|\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\right|\sqrt{\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}}}\end{aligned} \tag{5.107}$$

With a similar reasoning as before to bound the term under the square root, (5.106) becomes:

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_{\mu\perp\nu}) = \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_\mu^{(\ell-1)}) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big) \tag{5.108}$$

Plugging everything together leads to

$$\begin{aligned}&\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi(\alpha_\mu^{(\ell-1)})\tilde{\varphi}(\alpha_\nu^{(\ell-1)}) \\ &= \Big[\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_\mu^{(\ell-1)}) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\Big]\Big[\tilde{\rho}^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big]\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}} \\ &+ O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)^2\Big) \\ &= \Big[\rho^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\Big]\Big[\tilde{\rho}^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big]\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}} \\ &+ O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)^2\Big) \\ &= \rho^{(\ell-1)}\tilde{\rho}^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}} + O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big) \\ &+ O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)^2\Big) + O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big) \\ &+ O\Big(\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)^2\Big) + O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{d^{(\ell-1)}}\Big)\end{aligned} \tag{5.109}$$

Utilizing what stated in Corollary 15 and Remark 3, only the leading terms are considered, which lead to (5.92).

Consider now (5.93). Applying (5.102) and expanding $\varphi'(\alpha_\mu^{(\ell-1)})$ around $\alpha_{\mu\perp\nu}^{(\ell-1)}$ results in:

$$\begin{aligned}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_\mu^{(\ell-1)})\varphi'(\alpha_\nu^{(\ell-1)}) &= \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_{\mu\perp\nu}^{(\ell-1)})\mathbb{E}_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}'(\alpha_\nu^{(\ell-1)}) \\ &+ \frac{1}{2}E_{\mathbf{W}^{*(\ell)}}\varphi''(p)\tilde{\varphi}'(\alpha_\nu^{(\ell-1)})\alpha_\nu^{(\ell-1)}\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big) \\ &= \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_{\mu\perp\nu}^{(\ell-1)})\mathbb{E}_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}'(\alpha_\nu^{(\ell-1)}) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\end{aligned} \tag{5.110}$$

The result is obtained by now leveraging (5.92):

$$\begin{aligned}&\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_\mu^{(\ell-1)})\tilde{\varphi}'(\alpha_\nu^{(\ell-1)}) = \\ &= \Big[\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi'(\alpha_\mu^{(\ell-1)}) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\Big]\cdot\Big[\rho^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big] \\ &+ O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big) \\ &= \Big[\rho^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\Big] \\ &\cdot\Big[\rho^{(\ell-1)} + O\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big] + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big) \\ &= \rho^{(\ell-1)\ 2} + O\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big) + O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\Big) \\ &+ O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big(\frac{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}{d^{(\ell-1)}} - \sigma^{(\ell-1)}\Big)\Big) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\end{aligned} \tag{5.111}$$

Finally, the proof of (5.94) is derived. Again, the i index is omitted for clarity. The quantity $\varphi^2(\alpha_\mu^{(\ell-1)})$ is now expanded around $\alpha_{\mu\perp\nu}^{(\ell-1)}$:

$$\begin{aligned}&\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_\mu^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) \\ &= \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) + \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_\mu^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) - \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) \\ &= \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) \\ &+ 2\mathbb{E}_{\mathbf{W}^{*(\ell)}}\int_0^1 ds\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)})\alpha_\nu^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\end{aligned} \tag{5.112}$$

where $\alpha_{\mu,\nu}^{(\ell-1)}(s) = \alpha_{\mu\perp\nu}^{(\ell-1)} + s\alpha_\nu^{(\ell-1)}\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}/\|\mathbf{X}_\nu^{(\ell-1)}\|^2$. The objective is now to bound the integral on the right-hand side. Fubini's theorem is applied to exchange the expectation and the integral,

which can be done since it can be verified that $\mathbb{E}_{\mathbf{W}^{*(\ell)}}\left|\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})\alpha_{\nu}^{(\ell-1)}\right|$ is finite. Indeed, exploiting the Lipshitzianity of φ this expression is the expectation of combinations of powers of $|\alpha_{\nu}^{(\ell-1)}|$, $|\alpha_{\mu\perp\nu}^{(\ell-1)}|$ and $\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right|$. Once the integrals are interchanged, integration by parts is carried out with respect to $\alpha_{\nu}^{(\ell-1)}$, recalling that $\alpha_{\mu\perp\nu}^{(\ell-1)}$ is independent of it. The fact that φ and $\tilde{\varphi}$ are both Lipschitz is then used leading to:

$$
\begin{aligned}
&\left|\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})\alpha_{\nu}^{(\ell-1)}\right| = \\
&\Bigg|\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\Bigg[\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})s\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \\
&+\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi''(\alpha_{\mu,\nu}^{(\ell-1)}(s))\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})s\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2} \\
&+\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))2\tilde{\varphi}(\alpha_{\nu}^{(\ell-1)})\tilde{\varphi}'(\alpha_{\nu}^{(\ell-1)})\Bigg]\Bigg| \\
&\leq s\bar{K}^4\left|\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right|\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right|\mathbb{E}_{\mathbf{W}^{*(\ell)}}[\alpha_{\nu}^{(\ell-1)\ 2}] \\
&+s\bar{K}^4\left|\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right|\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right|\mathbb{E}_{\mathbf{W}^{*(\ell)}}\left|\alpha_{\nu}^{(\ell-1)\ 2}\left[\alpha_{\mu\perp\nu}^{(\ell-1)}+s\alpha_{\nu}^{(\ell-1)}\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right]\right| \\
&+2\bar{K}^4\left|\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right|\mathbb{E}_{\mathbf{W}^{*(\ell)}}\left|\alpha_{\nu}^{(\ell-1)}\left[\alpha_{\mu\perp\nu}^{(\ell-1)}+s\alpha_{\nu}^{(\ell-1)}\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right]\right| \\
&\leq s\bar{K}^4\left(\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)\left(\frac{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right| \\
&+s\bar{K}^4\left(\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)\left(\frac{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right|\left[\left(\frac{\|\mathbf{X}_{\mu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)^{\frac{1}{2}}+\left(\frac{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)^{\frac{1}{2}}\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{d^{(\ell-1)}}\right|\right. \\
&\left.+s\sqrt{(5!!)}\left(\frac{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}{d^{(\ell-1)}}\right)^{\frac{1}{2}}\left|\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right|\right]
\end{aligned}
\tag{5.113}
$$

Corollary 15 and Remark 3 allow to determine that this term is $O(1)$, and as a consequence the dependence $O\left(\frac{\mathbf{X}_{\mu}^{(\ell-1)\intercal}\mathbf{X}_{\nu}^{(\ell-1)}}{\|\mathbf{X}_{\nu}^{(\ell-1)}\|^2}\right)$ is obtained.

The expectation in the leading order term $\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})$ of (5.112) can be split for the two terms due to the independence of $\alpha_{\mu\perp\nu}^{(\ell-1)}$ and $\alpha_{\nu}^{(\ell-1)}$, leading to $\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})\mathbb{E}_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}^2(\alpha_{\nu}^{(\ell-1)})$.

Now an expansion of $\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)})$ around $\alpha_\mu^{(\ell-1)}$ yields:

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_{\mu\perp\nu}^{(\ell-1)}) = \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_\mu^{(\ell-1)}) - 2\int_0^1 ds\, \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi(\alpha_{\mu,\nu}^{(\ell-1)}(s))\varphi'(\alpha_{\mu,\nu}^{(\ell-1)}(s))\alpha_\nu^{(\ell-1)}\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2} \tag{5.114}$$

with $\alpha_{\mu,\nu}^{(\ell-1)}(s)$ defined as previously. Performing calculations in an analogous way to what we did before it is found that the integral contributes again with the same order as the one above. Therefore:

$$\mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_\mu^{(\ell-1)})\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) = \mathbb{E}_{\mathbf{W}^{*(\ell)}}\varphi^2(\alpha_\mu^{(\ell-1)})\mathbb{E}_{\mathbf{W}^{*(\ell)}}\tilde{\varphi}^2(\alpha_\nu^{(\ell-1)}) + O\Big(\frac{\mathbf{X}_\mu^{(\ell-1)\intercal}\mathbf{X}_\nu^{(\ell-1)}}{\|\mathbf{X}_\nu^{(\ell-1)}\|^2}\Big)\,. \tag{5.115}$$

Lastly, when equation (5.91) is applied to both factors in the leading term on the right hand side, the desired result is obtained.

□

5.4 Proof of Theorem 7

The approach used here is, as illustrated in 4.3. This method involves defining an interpolating model that integrates both the $L+1$-layer network and its linearized counterpart. The linearized version can be regarded as a generalized linear model with respect to its input signal $\mathbf{X}_\mu^{(L-1)}$. This dual construction is applied to both the teacher and student networks:

$$\begin{aligned}
S_{t\mu}^{(L)} &:= \sqrt{1-t}\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\,\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) + \sqrt{t}\rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{t\epsilon^{(L-1)}}\xi_\mu^{*(L-1)} \\
s_{t\mu}^{(L)} &:= \sqrt{1-t}\frac{\mathbf{a}^{\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{(L)}\,\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) + \sqrt{t}\rho^{(L-1)}\frac{\mathbf{v}^{(L-1)\intercal}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{t\epsilon^{(L-1)}}\xi_\mu^{(L-1)}
\end{aligned} \tag{5.116}$$

Here, the teacher's interpolating model is determined by the weights from both the $L+1$-layer network and the corresponding L-layer model. When $t=0$, this model simplifies to the linearized version, and when $t=1$, it reconstructs the full neural network. An analogous interpolation is applied to the student model. Additionally, an interpolating dataset is created, where $Y_{t\mu}^{(L)}$ is produced using an output kernel dependent on the teacher weights from the interpolating model:

$$\mathcal{D}_{n,t}^{(L)} = \{(\boldsymbol{X}_\mu^{(0)}, Y_{t\mu}^{(L)})_{\mu=1}^n\}\,, \quad Y_{t\mu}^{(L)} \sim P_{\text{out}}\big(\cdot \mid S_{t\mu}^{(L)}\big) \tag{5.117}$$

Let $\boldsymbol{\Theta}^{*(L)} = \{\mathbf{v}^{*(L-1)}, \boldsymbol{\xi}^{*(L-1)}, \mathbf{a}^*, \mathbf{W}^{*(L)} \ldots, \mathbf{W}^{*(1)}\} = \{\mathbf{v}^{*(L-1)}, \boldsymbol{\xi}^{*(L-1)}, \mathbf{a}^*, \mathbf{W}^{*(L)}, \boldsymbol{\omega}^{*(L-1)}\}$ denote the interpolating teacher parameters and define the expectation with respect to the dataset

and the parameters as:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{n,t}^{(L)},\boldsymbol{\Theta}^{*(L)}}(\cdot) &= \mathbb{E}_{\mathbf{X}^{(0)},\mathbf{Y}_t^{(L)},\boldsymbol{\Theta}^{*(L)}}(\cdot) = \int (\cdot) dP(\mathbf{Y}_t^{(L)}, \mathbf{X}^{(0)}, \boldsymbol{\Theta}^{*(L)}) \\
&= \int (\cdot) dP(\mathbf{Y}_t^{(L)}, \mathbf{X}^{(0)} \mid \boldsymbol{\Theta}^{*(L)}) dP(\boldsymbol{\Theta}^{*(L)}) \\
&= \int (\cdot) dP(\mathbf{Y}_t^{(L)} \mid \mathbf{X}^{(0)}, \boldsymbol{\Theta}^{*(L)}) dP(\mathbf{X}^{(0)} \mid \boldsymbol{\Theta}^{*(L)}) dP(\boldsymbol{\Theta}^{*(L)}) \\
&= \mathbb{E}_{\boldsymbol{\Theta}^{*(L)}} \mathbb{E}_{\mathbf{X}^{(0)}} \int \prod_{\mu=1}^{n} dY_{t\mu}^{(L)} P_{\text{out}}(Y_{t\mu}^{(L)} \mid S_{t\mu}^{(L)}) \\
&= \mathbb{E}_{\mathbf{W}^{*(1)},\cdots \mathbf{W}^{*(L-1)},\mathbf{W}^{*(L)},\mathbf{a}^*,\mathbf{v}^{*(L-1)},\boldsymbol{\xi}^{*(L-1)}} \mathbb{E}_{\mathbf{X}^{(0)}} \int \prod_{\mu=1}^{n} dY_{t\mu}^{(L)} e^{u_{Y_{t\mu}}^{(L)}(S_{t\mu})}(\cdot) \\
&= \mathbb{E}_{\mathbf{a}^*} \mathbb{E}_{\mathbf{W}^{*(1)},\cdots \mathbf{W}^{*(L-1)},\mathbf{W}^{*(L)},\mathbf{v}^{*(L-1)},\boldsymbol{\xi}^{*(L-1)},\mathbf{X}^{(0)}} \int \prod_{\mu=1}^{n} dY_{t\mu}^{(L)} e^{u_{Y_{t\mu}}^{(L)}(S_{t\mu})}(\cdot) \\
&=: \mathbb{E}_{\mathbf{a}^*} \mathbb{E}_{\setminus \mathbf{a}^*}(\cdot) = \mathbb{E}_{(t)}(\cdot)
\end{aligned} \tag{5.118}$$

Importantly, notice that if $(\cdot)$ does not explicitly depend on $\boldsymbol{\Theta}^{*(L)}$ then $\mathbb{E}_{\mathcal{D}_{n,t}^{(L)},\boldsymbol{\Theta}^{*(L)}}(\cdot) = \mathbb{E}_{\mathcal{D}_{n,t}^{(L)}}(\cdot)$.

Utilizing this notation, the partition function can be written as follows:

$$\mathcal{Z}_t^{(L)} = \mathcal{Z}_t^{(L)}(\mathcal{D}_{n,t}^{(L)}) = \int D\boldsymbol{\Theta}^{(L)} \exp\Big[\sum_{\mu=1}^{n} u_{Y_{t\mu}}^{(L)}(s_{t\mu})\Big] \tag{5.119}$$

Exploiting this, the expectation of a function g with respect to the posterior distribution is

$$\langle g \rangle_t^{(L)} = \int D\boldsymbol{\Theta}^{(L)} \exp\Big[\sum_{\mu=1}^{n} u_{Y_{t\mu}}^{(L)}(s_{t\mu})\Big] g \tag{5.120}$$

Moreover, from the definition of the partition function $\mathcal{Z}_t^{(L)}$ it follows that the interpolating free entropy is

$$\bar{f}_n^{(L)}(t) := \frac{1}{n} \mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \tag{5.121}$$

At $t = 0$, the free entropy expression corresponds to the one of the linearized model, while at $t = 1$, it represents the free entropy of the complete neural network:

$$\bar{f}_n^{(L)}(0) = \bar{f}_n^{(L)}, \quad \bar{f}_n^{(L)}(1) = \bar{f}_n^{(L-1)} \tag{5.122}$$

Since by the fundamental theorem of integral calculus $\bar{f}_n^{(L)}(1) - \bar{f}_n^{(L)}(0) = \int_0^1 \frac{d}{dt} \bar{f}_n^{(L)}(t) dt$, the goal is to control the difference between these two cases by computing the derivative and ensuring uniform control over time, maintaining consistency with the theorem's stated order. The derivative

computation results in the following summation:

$$\frac{d}{dt}\bar{f}_n^{(L)}(t) = -A_1 + A_2 + A_3 + B \tag{5.123}$$

where

$$\begin{aligned}
A_1 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \frac{\mathbf{a}^{*\intercal}}{\sqrt{(1-t)d^{(L)}}} \varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) \\
A_2 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \rho^{(L-1)} \frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{td^{(L-1)}}} \\
A_3 &:= \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t^{(L)} \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(S_{t\mu}) \sqrt{\frac{\epsilon^{(L-1)}}{t}} \xi_\mu^{*(L-1)} \\
B &:= \frac{1}{n}\mathbb{E}_{(t)} \Big\langle \sum_{\mu=1}^{n} u'^{(L)}_{Y_{t\mu}}(s_{t\mu}) \frac{ds_{t\mu}^{(L)}}{dt} \Big\rangle_t
\end{aligned} \tag{5.124}$$

Since the proof is focused solely on the case of the last layer, the notation is be simplified by omitting the superscript $^{(L)}$ wherever possible. The relevant quantities and notations are be redefined accordingly as follows:

$$\begin{aligned}
S_{t\mu} &:= S_{t\mu}^{(L)}\,, \quad & s_{t\mu} &:= s_{t\mu}^{(L)}\,, \quad & Y_{t\mu} &:= Y_{t\mu}^{(L)}\,, \\
\mathcal{D}_{n,t} &:= \mathcal{D}_{n,t}^{(L)}\,, \quad & \mathcal{Z}_t &:= \mathcal{Z}_t^{(L)}\,, \quad & u_{Y_{t\mu}} &:= u_{Y_{t\mu}}^{(L)}\,, \\
\mathbf{\Theta}^* &:= \mathbf{\Theta}^{*(L)}\,, \quad & \mathbf{\Theta} &:= \mathbf{\Theta}^{(L)}\,, \quad & \langle\cdot\rangle_t &:= \langle\cdot\rangle_t^{(L)}
\end{aligned} \tag{5.125}$$

Additionally, a new notation for the input of the activation function is introduced for ease:

$$\boldsymbol{\alpha}_\mu^{(L-1)} = \frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}} \tag{5.126}$$

Their covariance conditioned on $\mathbf{X}^{(L-1)}$, which equates to their covariance when considering the expectation with respect to the weights, is given by:

$$\begin{aligned}
\frac{1}{d^{(L)}}\mathbb{E}_{\mathbf{W}^{*(L)}}[\boldsymbol{\alpha}_\mu^{(L)\intercal}\boldsymbol{\alpha}_\nu^{(L)}] &= \frac{1}{d^{(L)}}\mathbb{E}_{\mathbf{W}^{*(L)}}[\boldsymbol{\alpha}_\mu^{(L)\intercal}\boldsymbol{\alpha}_\nu^{(L)} \mid \mathbf{X}^{(L-1)}] \\
&:= \frac{1}{d^{(L)}}\mathbb{E}_{\mathbf{W}^{*(L)}} \frac{(\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)})^\intercal}{\sqrt{d^{(L-1)}}} \frac{\mathbf{W}^{*(L)}\mathbf{X}_\nu^{(L-1)}}{\sqrt{d^{(L-1)}}} = \frac{\mathbf{X}_\mu^{(L-1)}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}
\end{aligned} \tag{5.127}$$

Notice how the conditioning is needed only on the specific instances μ and ν. For brevity and clarity, the conditioning over specific instances is denoted as conditioning over the entire dataset. This approach simplifies the expression while maintaining the intended meaning regarding covariance under the expectation with respect to the weights.

5.4.1 B term

Lemma 19 (B term). $B = 0$.

Proof. The random variable appearing in the Gibbs brackets in (4.33) is a function of the dataset through $Y_{t\mu}$, and of a sample from the posterior via $s_{t\mu}$. Denote this function $g(\mathbf{Y}_t, \mathbf{X}^{(0)}, \boldsymbol{\Theta}) = g(\mathcal{D}_{n,t}, \boldsymbol{\Theta}) = \sum_{\mu=1}^n u'_{Y_{t\mu}}(s_{t\mu}) \frac{ds_{t\mu}}{dt}$. The Nishimori identity can then be applied to eliminate the brackets, $s_{t\mu}$ with the ground truth version $S_{t\mu}$. Indicating the t-derivative as $\dot{S} := \frac{dS}{dt}$:

$$\begin{aligned} B &= \frac{1}{n}\mathbb{E}_{(t)}\Big\langle \sum_{\mu=1}^n u'_{Y_{t\mu}}(s_{t\mu})\frac{ds_{t\mu}}{dt}\Big\rangle_t \\ &= \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t},\boldsymbol{\Theta}^*}\Big\langle g(\mathcal{D}_{n,t},\boldsymbol{\Theta})\Big\rangle_t = \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t}}\Big\langle g(\mathcal{D}_{n,t},\boldsymbol{\Theta})\Big\rangle_t \\ &= \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t}}\mathbb{E}_{\boldsymbol{\Theta}|\mathcal{D}_{n,t}} g(\mathcal{D}_{n,t},\boldsymbol{\Theta}) = \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t}}\mathbb{E}_{\boldsymbol{\Theta}^*|\mathcal{D}_{n,t}} g(\mathcal{D}_{n,t},\boldsymbol{\Theta}^*) \\ &= \frac{1}{n}\mathbb{E}_{\mathcal{D}_{n,t},\boldsymbol{\Theta}^*} g(\mathcal{D}_{n,t},\boldsymbol{\Theta}^*) = \frac{1}{n}\mathbb{E}_{(t)}\sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu})\dot{S}_{t\mu} \end{aligned} \tag{5.128}$$

where in the second line $\mathbb{E}_{\mathcal{D}_{n,t},\boldsymbol{\Theta}^*}(\cdots) = \mathbb{E}_{\mathcal{D}_{n,t}}(\cdots)$ since $\Big\langle g(\mathcal{D}_{n,t},\boldsymbol{\Theta})\Big\rangle_t$ does not depend explicitly on $\boldsymbol{\Theta}^*$. Then, using the tower rule for expectations as shown below and applying Lemma 16, it can be concluded that the B term is zero.

$$B = \frac{1}{n}\mathbb{E}_{(t)}\sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu})\dot{S}_{t\mu} = \frac{1}{n}\sum_{\mu=1}^n \mathbb{E}_{(t)}\Big[\mathbb{E}_{(t)}[u'_{Y_{t\mu}}(S_{t\mu}) \mid S_{t\mu}]\dot{S}_{t\mu}\Big] \tag{5.129}$$

□

5.4.2 A_{11} off-diagonal term

As mentioned in 4.3, the term A_1 is divided into two components, $A_1 = A_{11} + A_{12}$ where

$$A_{11} := \frac{1}{2n\sqrt{1-t}}\mathbb{E}_{(t)}\log\mathcal{Z}_t\sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu})\Big(\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) - \frac{\rho^{(L-1)}\mathbf{a}^{*\intercal}\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L)}d^{(L-1)}}}\Big), \tag{5.130}$$

$$A_{12} := \frac{1}{2n\sqrt{1-t}}\mathbb{E}_{(t)}\log\mathcal{Z}_t\sum_{\mu=1}^n u'_{Y_{t\mu}}(S_{t\mu})\frac{\rho^{(L-1)}\mathbf{a}^{*\intercal}\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L)}d^{(L-1)}}} \tag{5.131}$$

To simplify these terms, Gaussian integration by parts is applied. In the case of A_{12}, an integration by parts with respect to the weights $\mathbf{W}^{*(L)}$ is performed:

$$A_{12} = \frac{\rho^{(L-1)}}{2n}\mathbb{E}_{(t)}\log\mathcal{Z}_t\sum_{\mu,\nu=1}^n U_{\mu\nu}\frac{\mathbf{a}^{*\intercal}\Big(\mathbf{a}^*\circ\varphi'\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big)}{d^{(L)}}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}} \tag{5.132}$$

where the variables $U_{\mu\nu}$ are defined in 16 and $\circ$ is used to denote the Hadamard product. Referring back to 4.3, since the weights matrix appears inside the nonlinearity function in A_{11}, integration is carried out with respect to the readout vector $\mathbf{a}^*$:

$$A_{11} = \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})^\intercal \varphi(\boldsymbol{\alpha}_\nu^{(L-1)}) - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})}{d^{(L-1)}}\Big] \tag{5.133}$$

The off-diagonal terms $\mu \neq \nu$ are analyzed first.

Lemma 20 (Off-diagonal part of A_{11})**.** *The following relation holds:*

$$A_{11}^{\text{off}} := \frac{1}{n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})^\intercal \varphi(\boldsymbol{\alpha}_\nu^{(L-1)}) - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})}{d^{(L-1)}}\Big] = O\Big(\sqrt{\Big(1+\frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{n}{d^{(L)}}+\frac{n}{\mathrm{d}^{(L-1)\ 3/2}}\Big)}\Big) \tag{5.134}$$

Proof. Exploiting the properties presented in Lemma 16, it can be observed that for any smooth function $F(\boldsymbol{\alpha}_\mu^{(L-1)}, \boldsymbol{\alpha}_\nu^{(L-1)})$ it holds

$$\mathbb{E}_{\setminus \mathbf{a}^*} U_{\mu\nu} F(\boldsymbol{\alpha}_\mu^{(L-1)}, \boldsymbol{\alpha}_\nu^{(L-1)}) = \mathbb{E}_{\setminus \mathbf{a}^*}\big[\mathbb{E}_{\setminus \mathbf{a}^*}[U_{\mu\nu} \mid \mathbf{W}^{*(L)}, \mathbf{v}^{*(L)}, \boldsymbol{\xi}^{*(L)}, \mathbf{X}^{(L-1)}] F(\boldsymbol{\alpha}_\mu^{(L-1)}, \boldsymbol{\alpha}_\nu^{(L-1)})\big] = 0 \tag{5.135}$$

The expression of A_{11}^{off} can then be modified without altering its value as long as the readout vector $\mathbf{a}^*$ remains unchanged. Defining $f_n := \log \mathcal{Z}_t/n$, this implies that it is possible to center this term with its mean without changing the value of A_{11}^{off}:

$$A_{11}^{\text{off}} = \mathbb{E}_{\setminus \mathbf{a}^*}(f_n - \mathbb{E}_{\setminus \mathbf{a}^*} f_n) \sum_{\mu\neq\nu} U_{\mu\nu}\Big[\frac{\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})^\intercal \varphi(\boldsymbol{\alpha}_\nu^{(L-1)}) - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})}{d^{(L)}}\Big] \tag{5.136}$$

For simplicity, the terms $u'_{Y_{t\mu}}(S_{t\mu})$ and $\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})$ are abbreviated as u'_μ and $\boldsymbol{\varphi}_\mu^{(L-1)}$. Applying now Cauchy-Schwartz's inequality to the equation previous equation we obtain yields:

$$(A_{11}^{\text{off}})^2 \leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \sum_{\mu\neq\nu}\sum_{\lambda\neq\eta} \mathbb{E}_{\setminus \mathbf{a}^*} U_{\mu\nu}U_{\lambda\eta}\Big(\Big[\frac{\boldsymbol{\varphi}_\mu^{(L-1)\intercal}\boldsymbol{\varphi}_\nu^{(L-1)} - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\boldsymbol{\varphi}_\nu^{(L-1)}}{d^{(L)}}\Big] \cdot \Big[\frac{\boldsymbol{\varphi}_\lambda^{(L-1)\intercal}\boldsymbol{\varphi}_\eta^{(L-1)} - \rho^{(L-1)}\boldsymbol{\alpha}_\lambda^{(L-1)\intercal}\boldsymbol{\varphi}_\eta^{(L-1)}}{d^{(L)}}\Big]\Big) \tag{5.137}$$

Utilizing the conditional independence of the responses $\mathbf{Y}_t$ and leveraging the properties established in Lemma 16, terms where indices differ from $\mu = \lambda$ and $\nu = \eta$, or $\mu = \eta$ and $\nu = \lambda$, do not add to the summation. Both cases contribute equally.

$$(A_{11}^{\text{off}})^2 \leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \frac{2}{d^{(L)\,2}} \sum_{\mu \neq \nu} \mathbb{E}_{\setminus \mathbf{a}^*}(u'_\mu u'_\nu)^2 \sum_{i,j=1}^{d^{(L)}} \big[\varphi_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \varphi_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)} \\ - 2\rho^{(L-1)} \alpha_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \varphi_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)} + \rho^{(L-1)\,2} \alpha_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \alpha_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)}\big] \tag{5.138}$$

Here, the double summation over i, j is derived from the four scalar products appearing in (5.137). Following this, Lemma 16 provides a way to limit the expectation of $U_{\mu\nu}^2$ for fixed $S_{t\mu}, S_{t\nu}$. Moreover, the probability space is divided into two regions defined by the events $A_\delta^{(L-1)}$ and $\bar{A}_\delta{}^{(L-1)}$. Here, the event A_δ^{L-1} for $\mathbf{X}^{(L-1)}$ is defined as in Lemma18 for some parameter δ. The indicator function $\mathbb{I}(\cdot)$ is used to characterize these events. The aim is then to compute the contribution C_δ^2 and $C_{\bar{\delta}}^2$ to (5.138) corresponding to the events $A_\delta^{(L-1)}$ and $\bar{A}_\delta{}^{(L-1)}$ respectively. Hence:

$$(A_{11}^{\text{off}})^2 \leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \frac{2}{d^{(L)\,2}} \sum_{\mu \neq \nu} \mathbb{E}_{\setminus \mathbf{a}^*}(u'_\mu u'_\nu)^2 (\mathbb{I}(A_\delta^{(L-1)}) + \mathbb{I}(\bar{A}_\delta{}^{(L-1)})) \sum_{i,j=1}^{d^{(L)}} \big[\varphi_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \varphi_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)} \\ - 2\rho^{(L-1)} \alpha_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \varphi_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)} + \rho^{(L-1)\,2} \alpha_{\mu i}^{(L-1)} \varphi_{\nu i}^{(L-1)} \alpha_{\mu j}^{(L-1)} \varphi_{\nu j}^{(L-1)}\big] = C_\delta^2 + C_{\bar{\delta}}^2 \tag{5.139}$$

The estimation of (5.139) on $\mathbb{I}(A_\delta^{(L-1)})$ is first addressed, namely, the computation of C_δ^2. By applying Lemma 18, particularly equation (5.92), the first term in (5.139) can be rewritten as follows:

$$\mathbb{E}_{\setminus \mathbf{a}^*} \mathbb{I}(A_\delta^{(L-1)}) \sum_{i \neq j}^{d^{(L)}} \varphi(\alpha_{\mu i}^{(L-1)}) \varphi(\alpha_{\nu i}^{(L-1)}) \varphi(\alpha_{\mu j}^{(L-1)}) \varphi(\alpha_{\nu j}^{(L-1)}) \\ = d^{(L)}(d^{(L)} - 1) \mathbb{E}_{\mathbf{X}^{(L-1)}} \big(\mathbb{E}_{\mathbf{W}^{*(L)}}[\mathbb{I}(A_\delta^{(L-1)}) \varphi(\alpha_{\mu 1}^{(L-1)}) \varphi(\alpha_{\nu 1}^{(L-1)})]\big)^2 \\ = d^{(L)}(d^{(L)} - 1) \mathbb{E}\mathbb{I}(A_\delta^{(L-1)}) \Big[\rho^{(L-1)\,4} \Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)^2 + O\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}} \Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)^2\Big) \\ + O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)^2 \frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big) + O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)^2 \Big(\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d^{(L-1)}} - \sigma^{(L-1)}\Big)\Big)\Big] \tag{5.140}$$

Exploiting that the case $\mathbf{X}^{(L-1)} \in A_\delta^{(L-1)}$ is considered and thus what observed in Remark 3, and applying Corollary 15, it can then be concluded that:

$$\mathbb{E}_{\setminus \mathbf{a}^*} \mathbb{I}(A_\delta^{(L-1)}) \sum_{i \neq j,1}^{d^{(L)}} \varphi(\alpha_{\mu i}^{(L-1)}) \varphi(\alpha_{\nu i}^{(L-1)}) \varphi(\alpha_{\mu j}^{(L-1)}) \varphi(\alpha_{\nu j}^{(L-1)}) = d^{(L)}(d^{(L)} - 1) \Big[\frac{\rho^{(L-1)\,4}}{\mathbb{d}^{(L-1)}} + O\Big(\frac{1}{\mathbb{d}^{(L-1)\,3/2}}\Big)\Big] \tag{5.141}$$

The second term of equation (5.139) is now considered. The subsequent result is derived upon the application of Lemma 18.

$$
\begin{aligned}
&\rho^{(L-1)}\mathbb{E}_{\mathbf{X}^{(L-1)}}\mathbb{E}_{\mathbf{W}^{*(L)}}\mathbb{I}(A_\delta^{(L-1)})\sum_{i\neq j}^{d^{(L)}}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\\
&=\rho^{(L-1)}d^{(L)}(d^{(L)}-1)\mathbb{E}_{\mathbf{X}^{(L-1)}}\mathbb{E}_{\mathbf{W}^{*(L)}}[\mathbb{I}(A_\delta^{(L-1)})\alpha_{\mu 1}^{(L-1)}\varphi_{\nu 1}^{(L-1)}]\mathbb{E}_{\mathbf{W}^{*(L)}}[\mathbb{I}(A_\delta^{(L-1)})\varphi_{\mu 1}^{(L-1)}\varphi_{\nu 1}^{(L-1)}]\\
&=\rho^{(L-1)}d^{(L)}(d^{(L)}-1)\mathbb{E}_{\mathbf{X}^{(L-1)}}\Big[\mathbb{I}(A_\delta^{(L-1)})\Big[\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}+O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d^{(L-1)}}-\sigma^{(L-1)}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)\\
&+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)^2\Big)\Big]\\
&\cdot\Big[\rho^{(L-1)}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}+O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d^{(L-1)}}-\sigma^{(L-1)}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)\\
&+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)^2\Big)\Big]\Big]
\end{aligned}
\tag{5.142}
$$

Applying Remark 3 since the contribution associated to $A_\delta^{(L-1)}$ is considered and exploiting Corollary 15 yields:

$$
\rho^{(L-1)}\mathbb{E}_{\mathbf{X}^{(L-1)}}\mathbb{E}_{\mathbf{W}^{*(L)}}\mathbb{I}(A_\delta^{(L-1)})\sum_{i\neq j}^{d^{(L)}}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}=d^{(L)}(d^{(L)}-1)\Big[\frac{\rho^{(L-1)\ 4}}{\mathrm{d}^{(L-1)}}+O\Big(\frac{1}{\mathrm{d}^{(L-1)\ 3/2}}\Big)\Big]
\tag{5.143}
$$

The last term in (5.139) is now addressed. Integration by parts is first applied, and Lemma 18 is then used. The computation of the scaling is performed remembering that the event $A_\delta^{(L-1)}$ is

considered, and using Remark 3.

$$
\begin{aligned}
&\rho^{(L-1)\,2}\mathbb{E}_{\mathbf{X}^{(L-1)}}\mathbb{E}_{\mathbf{W}^{*(L)}}\mathbb{I}(A_\delta^{(L-1)})\sum_{i\neq j}^{d^{(L)}}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\alpha_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\\
&=\rho^{(L-1)\,2}d^{(L)}(d^{(L)}-1)\mathbb{E}_{\mathbf{X}^{(L-1)}}\mathbb{E}_{\mathbf{W}^{*(L)}}[\mathbb{I}(A_\delta^{(L-1)})\alpha_{\mu 1}^{(L-1)}\varphi_{\nu 1}^{(L-1)}]\mathbb{E}_{\mathbf{W}^{*(L)}}[\mathbb{I}(A_\delta^{(L-1)})\alpha_{\mu 1}^{(L-1)}\varphi_{\nu 1}^{(L-1)}]\\
&=\rho^{(L-1)\,2}d^{(L)}(d^{(L)}-1)\mathbb{E}_{\mathbf{X}^{(L-1)}}\Big[\mathbb{I}(A_\delta^{(L-1)})\Big[\rho^{(L-1)}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\\
&+O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d^{(L-1)}}-\sigma^{(L-1)}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)\\
&+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big)+O\Big(\Big(\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big)^2\Big)\Big]^2\Big]\\
&=d^{(L)}(d^{(L)}-1)\Big[\frac{\rho^{(L-1)\,4}}{\mathrm{d}^{(L-1)}}+O\Big(\frac{1}{\mathrm{d}^{(L-1)\,3/2}}\Big)\Big]\Big]
\end{aligned}
\tag{5.144}
$$

It can be observed that in the estimation of C_δ^2 in equation (5.139) the leading orders of the off-diagonal elements, namely $i\neq j$, balance each other. This results in a rate for the off-diagonal components $O(1/\mathrm{d}^{(L-1)\,3/2})$ for the off-diagonal components. Inserting what discussed in equation (5.139) implies that there exists a constant K such that:

$$
\begin{aligned}
C_\delta^2\leq\mathbb{V}_{\setminus\mathbf{a}^*}[f_n]\frac{2K}{d^{(L)\,2}}\sum_{\mu\neq\nu}\mathbb{E}_{\setminus\mathbf{a}^*}\Big\{\mathbb{I}(A_\delta^{(L-1)})\sum_{i=1}^{d^{(L)}}\big[\varphi_{\mu i}^{(L-1)\,2}\varphi_{\nu i}^{(L-1)\,2}\\
-2\rho^{(L-1)}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)\,2}\varphi_{\mu j}^{(L-1)}+\rho^{(L-1)\,2}\alpha_{\mu i}^{(L-1)\,2}\varphi_{\nu i}^{(L-1)\,2}\big]+O\Big(\frac{d^{(L)\,2}}{\mathrm{d}^{(L-1)\,3/2}}\Big)\Big\}
\end{aligned}
\tag{5.145}
$$

Considering now the diagonal terms, applying again Lemma 18 to each one of the summation terms remembering that the case $\mathbf{X}^{(L-1)}\in A_\delta^{(L-1)}$ is considered leads to:

$$
C_\delta^2\leq\mathbb{V}_{\setminus\mathbf{a}^*}[f_n]\frac{2K}{d^{(\ell)\,2}}\sum_{\mu\neq\nu}\Big[O(d^{(\ell)})+O\Big(\frac{d^{(\ell)\,2}}{\mathrm{d}^{(\ell-1)\,3/2}}\Big)\Big]
\tag{5.146}
$$

Consider now the term $C_{\bar\delta}^2$ corresponding to the set $\bar{A_\delta}^{(L-1)}$. The objective is to provide an estimation of the following term's order:

$$
\begin{aligned}
C_{\bar\delta}^2:=\mathbb{V}_{\setminus\mathbf{a}^*}[f_n]\frac{2}{d^{(L)\,2}}\sum_{\mu\neq\nu}\mathbb{E}_{\setminus\mathbf{a}^*}(u'_\mu u'_\nu)^2\mathbb{I}(\bar{A_\delta}^{(L-1)})\sum_{i,j=1}^{d^{(L)}}\big[\varphi_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\\
-2\rho^{(L-1)}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}+\rho^{(L-1)\,2}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\alpha_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\big]
\end{aligned}
\tag{5.147}
$$

Again Lemma 16 then allows to bound $\mathbb{E}[U_{\mu\nu}^2 \mid S_{t\mu}, S_{t\nu}]$ by a constant. Thus the following term needs to be evaluated:

$$\mathbb{E}_{\setminus \mathbf{a}^*}\mathbb{I}(\bar{A}_\delta{}^{(L-1)})\big[\varphi_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)} - 2\rho^{(L-1)}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)} + \rho^{(L-1)\ 2}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\alpha_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\big] \leq 3\bar{K}^2\max\{\bar{K}^2, \rho^{(L-1)\ 2}\}\mathbb{E}_{\setminus \mathbf{a}^*}\mathbb{I}(\bar{A}_\delta)\big[\Big|\alpha_{\mu i}^{(L-1)}\alpha_{\nu i}^{(L-1)}\alpha_{\mu j}^{(L-1)}\alpha_{\nu j}^{(L-1)}\Big|\big] \tag{5.148}$$

To achieve this, the existence and boundedness of any moment of $\alpha_{\mu i}^{(\ell)}, \ell \leq L$ are leveraged. Specifically, for any power $2k$, exploiting first the Gaussianity of the weights and then Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\alpha_{\mu i}^{(\ell)\ 2k} &\leq C\mathbb{E}_{\mathbf{X}^{(\ell-1)}}\Big(\frac{\|\mathbf{X}^{(\ell-1)\mu}\|^2}{d^{(\ell-1)}}\Big)^{2k} = C\mathbb{E}_{\mathbf{X}^{(\ell-1)}}\Big(\frac{\sum_{i=1}^{d^{(\ell-1)}}\varphi^2\Big(\frac{\mathbf{W}_i^{*(\ell-1)}\mathbf{X}^{(\ell-2)\mu}}{d^{(\ell-2)}}\Big)}{d^{(\ell-1)}}\Big)^{2k} \\
&\leq C\bar{K}^{4k}\mathbb{E}_{\mathbf{X}^{(\ell-1)}}\Big(\frac{\sum_{i=1}^{d^{(\ell-1)}}\Big(\frac{\mathbf{W}_i^{*(\ell-1)}\mathbf{X}_\mu^{(\ell-2)}}{d^{(\ell-2)}}\Big)^2}{d^{(\ell-1)}}\Big)^{2k} \\
&= C\bar{K}^{4k}\mathbb{E}_{\mathbf{X}^{(\ell-2)},\mathbf{W}^{*(\ell-1)}}\Big(\frac{1}{d^{(\ell-1)}}\Big)^{2k}\Big(\sum_{i=1}^{d^{(\ell-1)}}\Big(\frac{\mathbf{W}_i^{*(\ell-1)}\mathbf{X}_\mu^{(\ell-2)}}{d^{(\ell-2)}}\Big)^2\Big)^{2k} \\
&\leq C\bar{K}^{4k}\mathbb{E}_{\mathbf{X}^{(\ell-2)},\mathbf{W}^{*(\ell-1)}}\frac{1}{d^{(\ell-1)}}\sum_{i=1}^{d^{(\ell-1)}}\Big(\frac{\mathbf{W}_i^{*(\ell-1)}\mathbf{X}_\mu^{(\ell-2)}}{d^{(\ell-2)}}\Big)^{4k} \\
&\leq C'\bar{K}^{4k}\mathbb{E}_{\mathbf{X}^{(\ell-2)}}\Big(\frac{\|\mathbf{X}_\mu^{(\ell-2)}\|^2}{d^{(\ell-2)}}\Big)^{4k} \\
&\leq \ldots \leq C''\bar{K}^{8k}\mathbb{E}_{\mathbf{X}^{(\ell-3)}}\Big(\frac{\|\mathbf{X}_\mu^{(\ell-3)}\|^2}{d^{(\ell-3)}}\Big)^{8k} \leq \ldots \leq \bar{C}\bar{K}^{2^\ell k}\mathbb{E}_{\mathbf{X}^{(0)}}\Big(\frac{\|\mathbf{X}_\mu^{(0)}\|^2}{d^{(0)}}\Big)^{2^\ell k} \leq \tilde{C}
\end{aligned} \tag{5.149}$$

for some constant $\tilde{C}$. Thus, remembering that $\mathbb{E}_{\setminus \mathbf{a}^*}\mathbb{I}(\bar{A}_\delta{}^{(L-1)}) = n\mathbb{E}_{\setminus \mathbf{a}^*}\mathbb{I}(\bar{A}_{\delta\mu}^{(L-1)})$ and through the concentration Propositions 13 and 14, equation (5.148) can be reduced to

$$\mathbb{E}_{\setminus \mathbf{a}^*}\mathbb{I}(\bar{A}_\delta{}^{(L-1)})\big[\varphi_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)} - 2\rho^{(L-1)}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\varphi_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)} + \rho^{(L-1)\ 2}\alpha_{\mu i}^{(L-1)}\varphi_{\nu i}^{(L-1)}\alpha_{\mu j}^{(L-1)}\varphi_{\nu j}^{(L-1)}\big] \leq 4Ce^{-\mathrm{d}^{(L)}C'\delta^2} \tag{5.150}$$

This term becomes negligible and specifically scales lower than $O\Big(\frac{1}{\mathrm{d}^{(L)\ k}}\Big)$ for any chosen exponent k if $\delta^2 \geq \frac{k\ln \mathrm{d}^{(L)}}{C'd^{(L)}}$ is selected. Only the contribution given by C_δ^2 (5.146) is then considered.

The proof of the statement is completed once the residual expectation over the readout vector $\mathbf{a}^*$

is computed, utilizing the assumption H3):

$$|\mathbb{E}_{\mathbf{a}^*} A_{11}^{\text{off}}| \leq \mathbb{E}_{\mathbf{a}^*} \sqrt{(A_{11}^{\text{off}})^2} \tag{5.151}$$

$$\leq \sqrt{\mathbb{E}_{\mathbf{a}^*} \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] O\Big(\frac{n^2}{d^{(L)}} + \frac{n^2}{\mathrm{d}^{(L-1)\ 3/2}}\Big)} = O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}}\Big)}\Big) \tag{5.152}$$

□

5.4.3 $A_3 - A_{11}^{\text{diag}}$ term

From Lemma 20 it can be noticed that the term A_{11} can be rewritten as

$$A_{11} = \frac{1}{2n} \mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \Big[\frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})\|^2 - \rho^{(L-1)} \boldsymbol{\alpha}_\mu^{(L-1)\intercal} \varphi(\boldsymbol{\alpha}_\mu^{(L-1)})}{d^{(L)}} \Big] + O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}}\Big)}\Big) \tag{5.153}$$

Considering now the term A_3, as mentioned in 4.3, Gaussian integration by parts is performed with respect to the variables $\xi_\mu^{*(L-1)}$:

$$A_3 = \frac{\epsilon^{(L-1)}}{2n} \mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \Big((u'_{Y_{t\mu}}(S_{t\mu}))^2 + u''_{Y_{t\mu}}(S_{t\mu}) \Big) = \frac{\epsilon^{(L-1)}}{2n} \mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \,. \tag{5.154}$$

Computing its difference with respect to A_{11} leads to

$$A_3 - A_{11} = \frac{1}{2n} \mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \Big[\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})\|^2 - \rho^{(L-1)} \boldsymbol{\alpha}_\mu^{(L-1)\intercal} \varphi(\boldsymbol{\alpha}_\mu^{(L-1)})}{d^{(L)}} \Big] + O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}}\Big)}\Big) \tag{5.155}$$

The order of the remaining term is estimated by the following Lemma.

Lemma 21 ($A_3 - A_{11}^{\text{diag}}$ term)**.** *The following relation holds:*

$$A_3 - A_{11}^{\text{diag}} := \frac{1}{2n} \mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \Big[\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})\|^2 - \rho^{(L-1)} \boldsymbol{\alpha}_\mu^{(L-1)\intercal} \varphi(\boldsymbol{\alpha}_\mu^{(L-1)})}{d^{(L)}} \Big] = O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{1}{d^{(L)}} + \frac{1}{\mathrm{d}^{(L-1)\ 1/2}}\Big)}\Big) \tag{5.156}$$

Proof. As a first step, the new quantity C is introduced:

$$C := \frac{1}{n}\mathbb{E}_{\setminus \mathbf{a}^*} \log \mathcal{Z}_t \sum_{\mu=1}^{n} \frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \left[\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})}{d^{(L)}}\right] \tag{5.157}$$

Exploiting the properties illustrated in Lemma 16, it can be recognized that

$$\mathbb{E}_{\setminus \mathbf{a}^*}\left[\frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})} \mid \mathbf{W}^{*(L)}, \mathbf{v}^{*(L-1)}, \xi_\mu^{*(L-1)}, \mathbf{X}_\mu^{(L-1)}\right] = 0 \tag{5.158}$$

This again implies that the expression of C can be modified without causing any change to its value. Similar to the approach in 5.4.2, $f_n = \log \mathcal{Z}_t/n$ can be centered around its mean value. Using Cauchy-Schwartz's inequality then results in:

$$\begin{aligned} C^2 \leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \sum_{\mu,\nu=1}^{n} \mathbb{E}_{\setminus \mathbf{a}^*}\Bigg[\mathbb{E}_{\setminus \mathbf{a}^*}\Bigg[\frac{P''_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}{P_{\text{out}}(Y_{t\mu} \mid S_{t\mu})}\frac{P''_{\text{out}}(Y_{t\nu} \mid S_{t\nu})}{P_{\text{out}}(Y_{t\nu} \mid S_{t\nu})} \mid \mathbf{W}^{*(L)}, \mathbf{v}^{*(L-1)}, \xi_\mu^{*(L-1)}, \xi_\nu^{*(L-1)}, \mathbf{X}_\mu^{(L-1)}, \mathbf{X}_\nu^{(L-1)}\Bigg] \\ \cdot \left(\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_\mu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\mu^{(L-1)})}{d^{(L)}}\right)\left(\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_\nu^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_\nu^{(L-1)})}{d^{(L)}}\right)\Bigg] \end{aligned} \tag{5.159}$$

Exploiting (5.158), it can be observed that only the terms characterized by $\mu = \nu$ contribute to the summation. Using again Lemma 16 and specifically (5.71), C^2 reads:

$$\begin{aligned} C^2 &\leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] C(f) n \mathbb{E}_{\mathbf{X}^{(L-1)}, \mathbf{W}^{*(L)}}\left(\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\right)^2 \\ &= \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] C(f) n \Bigg[\epsilon^{(L-1)\,2} - 2\epsilon^{(L-1)}\mathbb{E}_{\mathbf{X}^{(L-1)}, \mathbf{W}^{*(L)}}\left(\frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\right) \\ &+ \mathbb{E}_{\mathbf{X}^{(L-1)}, \mathbf{W}^{*(L)}}\left(\frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\right)^2\Bigg] \end{aligned} \tag{5.160}$$

To compute the terms $\alpha_{1i}^{(L-1)}\varphi(\alpha_{1i}^{(L-1)})$ an integration by parts is performed and then Lemma 18 is

exploited.

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}^{*(L)}}\alpha_{1i}^{(L-1)}\varphi(\alpha_{1i}^{(L-1)}) &= \mathbb{E}_{\mathbf{W}^{*(L)}}\varphi'(\alpha_{1i}^{(L-1)})\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} \\
&= \frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\Big(\rho^{(L-1)} + O\Big(\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} - \mathbb{E}_{\mathbf{X}^{(L-1)}}\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\Big)\Big) \\
&= \frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\rho^{(L-1)} + O\Big(\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} - \mathbb{E}_{\mathbf{X}^{(L-1)}}\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\Big) \qquad (5.161) \\
\mathbb{E}^2_{\mathbf{W}^{*(L)}}\alpha_{1i}^{(L-1)}\varphi(\alpha_{1i}^{(L-1)}) &= \mathbb{E}^2_{\mathbf{W}^{*(L)}}\varphi'(\alpha_{1i}^{(L-1)})\frac{\|\mathbf{X}_1^{(L-1)}\|^4}{d^{(L-1)\ 2}} \\
&= \frac{\|\mathbf{X}_1^{(L-1)}\|^4}{d^{(L-1)\ 2}}\rho^{(L-1)\ 2} + O\Big(\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} - \mathbb{E}_{\mathbf{X}^{(L-1)}}\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\Big)
\end{aligned}$$

The terms in (5.160) can then be estimated leveraging Lemma 18:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^{(L-1)},\mathbf{W}^{*(L)}}\Big[\frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\Big] \\
&= \mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2 - \sigma^{(L-1)}\rho^{(L-1)\ 2} + \mathbb{E}_{\mathbf{X}^{(L-1)}}\Big[O\Big(\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} - \mathbb{E}_{\mathbf{X}^{(L-1)}}\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}}\Big)\Big] \qquad (5.162) \\
&= \epsilon^{(L-1)} + O(\mathrm{d}^{(L-1)\ -1/2})
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^{(L-1)},\mathbf{W}^{*(L)}}\Big[\frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\Big]^2 \\
&= \frac{1}{d^{(L)}}\mathbb{E}_{\mathbf{X}^{(L-1)},\mathbf{W}^{*(L)}}\Big(\varphi^4(\alpha_{11}^{(L-1)}) - 2\rho^{(L-1)}\varphi^3(\alpha_{11}^{(L-1)})\alpha_{11}^{(L-1)} + \rho^{(L-1)\ 2}\alpha_{11}^{(L-1)\ 2}\varphi^2(\alpha_{11}^{(L-1)})\Big) \\
&+ \frac{d^{(L)}-1}{d^{(L)}}\mathbb{E}_{\mathbf{X}^{(L-1)}}\Big(\mathbb{E}^2_{\mathbf{W}^{*(L)}}\varphi^2(\alpha_{11}^{(L-1)}) \\
&- 2\rho^{(L-1)}\mathbb{E}_{\mathbf{W}^{*(L)}}\varphi^2(\alpha_{11}^{(L-1)})\mathbb{E}_{\mathbf{W}^{*(L)}}\varphi(\alpha_{11}^{(L-1)})\alpha_{11}^{(L-1)} + \rho^{(L-1)\ 2}\mathbb{E}^2_{\mathbf{W}^{*(L)}}\varphi(\alpha_{11}^{(L-1)})\alpha_{11}^{(L-1)}\Big) \\
&= \epsilon^{(L-1)\ 2} + O(d^{(L)\ -1}) + O(\mathrm{d}^{(L-1)\ -1/2})
\end{aligned} \qquad (5.163)$$

Plugging these results back in (5.160) results in

$$\mathbb{E}_{\mathbf{X}^{(L-1)},\mathbf{W}^{*(L)}}\Big(\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\Big)^2 = O(d^{(L)\ -1}) + O(\mathrm{d}^{(L-1)\ -1/2}) \qquad (5.164)$$

Incorporating the estimate derived in (5.164) into (5.160), along with the assumed variance order in

H3), and finally addressing the expectation with respect to the readout vector, results in:

$$
\begin{aligned}
|A_3 - A_{11}^{\text{diag}}| &\leq \mathbb{E}_{\mathbf{a}^*}\sqrt{C^2} \\
&\leq \sqrt{\mathbb{E}_{\mathbf{a}^*}\mathbb{V}_{\setminus \mathbf{a}^*}[f_n]C(f)n\mathbb{E}_{\mathbf{X}^{(L-1)},\mathbf{W}^{*(L)}}\Big(\epsilon^{(L-1)} - \frac{\|\varphi(\boldsymbol{\alpha}_1^{(L-1)})\|^2 - \rho^{(L-1)}\boldsymbol{\alpha}_1^{(L-1)\intercal}\varphi(\boldsymbol{\alpha}_1^{(L-1)})}{d^{(L)}}\Big)^2} \\
&= O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{1}{d^{(L)}} + \frac{1}{\mathrm{d}^{(L-1)\ 1/2}}\Big)}\Big)
\end{aligned}
\tag{5.165}
$$

□

5.4.4 $A_{12} - A_2$ term

Now the aim is to study the difference $A_2 - A_{12}$ as anticipated in 4.3. Applying Gaussian integration by parts with respect to $\mathbf{v}^{*(L)}$ to A_2 leads to:

$$
A_2 = \frac{\rho^{(L-1)\ 2}}{2n}\mathbb{E}_{(t)}\log\mathcal{Z}_t\sum_{\mu,\nu=1}^{n}U_{\mu\nu}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}} \tag{5.166}
$$

The estimate is provided by the following Lemma:

Lemma 22. *[$A_{12} - A_2$ term] The following holds:*

$$
\begin{aligned}
&A_{12} - A_2 = \frac{\rho^{(L-1)}}{2n}\mathbb{E}_{(t)}\log\mathcal{Z}_t\sum_{\mu,\nu=1}^{n}U_{\mu\nu}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big[\frac{\mathbf{a}^{*\intercal}(\mathbf{a}^*\circ\varphi'(\boldsymbol{\alpha}_\nu^{(L-1)}))}{d^{(L)}} - \rho^{(L-1)}\Big] = \\
&O\Big(\sqrt{\Big(1 + \frac{n}{\mathrm{d}^{(L-1)}}\Big)\Big(\frac{n}{\mathrm{d}^{(L-1)}d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}}\Big)}\Big)
\end{aligned}
\tag{5.167}
$$

Proof. Similarly to what done in 21, fixing the readout vector define

$$
C := \frac{1}{n}\mathbb{E}_{\setminus \mathbf{a}^*}\log\mathcal{Z}_t\sum_{\mu,\nu=1}^{n}U_{\mu\nu}\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big[\frac{\mathbf{a}^{*\intercal}(\mathbf{a}^*\circ\varphi'(\boldsymbol{\alpha}_\nu^{(L-1)}))}{d^{(L)}} - \rho^{(L-1)}\Big] \tag{5.168}
$$

As in 20 and 21, $f_n = \log\mathcal{Z}_t/n$ can be centered around its mean changing the expression of C but not its value. Consequently, Cauchy-Schwartz's inequality can be applied, resulting in:

$$
\begin{aligned}
C^2 \leq \mathbb{V}_{\setminus \mathbf{a}^*}[f_n]\mathbb{E}_{\setminus \mathbf{a}^*}\sum_{\mu,\nu=1}^{n}\sum_{\lambda,\eta=1}^{n}U_{\mu\nu}U_{\lambda\eta}&\frac{\mathbf{X}_\mu^{(L-1)\intercal}\mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}}\Big[\frac{\mathbf{a}^{*\intercal}(\mathbf{a}^*\circ\varphi'(\boldsymbol{\alpha}_\nu^{(L-1)}))}{d^{(L)}} - \rho^{(L-1)}\Big] \\
&\cdot\frac{\mathbf{X}_\lambda^{(L-1)\intercal}\mathbf{X}_\eta^{(L-1)}}{d^{(L-1)}}\Big[\frac{\mathbf{a}^{*\intercal}(\mathbf{a}^*\circ\varphi'(\boldsymbol{\alpha}_\eta^{(L-1)}))}{d^{(L)}} - \rho^{(L-1)}\Big]
\end{aligned}
\tag{5.169}
$$

Exploiting the conditional independence of the $U_{\mu\nu}$ terms described in Lemma 16, the only contributing components to the summation are those where $\mu = \lambda$ and $\nu = \eta$, $\mu = \eta$ and $\nu = \lambda$, or $\mu = \nu = \lambda = \eta$. All other terms, where the indices do not match in these ways, contribute zero. In the contributing cases, due to Lemma 16, $\mathbb{E}[U_{\mu\nu}^2 \mid S_{t\mu}, S_{t\nu}]$ can be bounded, hence, for a suitable constant K then, (5.169) becomes

$$C^2 \leq K \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \mathbb{E}_{\setminus \mathbf{a}^*} \sum_{\mu,\nu=1}^{n} \left(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}} \right)^2 \left[\frac{\mathbf{a}^{*\intercal}(\mathbf{a}^* \circ \varphi'(\boldsymbol{\alpha}_\nu^{(L-1)}))}{d^{(L)}} - \rho^{(L-1)} \right]^2 \tag{5.170}$$

Developing the square and taking the expectation of φ' with respect to the weights $\mathbf{W}^{*(L)}$ through Lemma 18 leads to

$$\begin{aligned} C^2 \leq K' \mathbb{V}_{\setminus \mathbf{a}^*}[f_n] \sum_{\mu,\nu=1}^{n} \mathbb{E}_{\mathbf{X}_\mu^{(L-1)}, \mathbf{X}_\nu^{(L-1)}} \left(\frac{\mathbf{X}_\mu^{(L-1)\intercal} \mathbf{X}_\nu^{(L-1)}}{d^{(L-1)}} \right)^2 \Bigg[\left(\frac{\|\mathbf{a}^*\|^2}{d^{(L)}} - 1 \right)^2 \\ + O\left(\frac{\|\mathbf{X}_\nu^{(L-1)}\|^2}{d^{(L-1)}} - \mathbb{E}_{\mathbf{X}^{(L-1)}} \frac{\|\mathbf{X}_\nu^{(L-1)}\|^2}{d^{(L-1)}} \right) \left(\frac{\|\mathbf{a}^*\|^4}{d^{(L)\ 2}} + \frac{\|\mathbf{a}^*\|^2}{d^{(L)}} \right) + \sum_{i=1}^{d^{(L)}} \frac{\mathbf{a}_i^4}{d^{(L)\ 2}} \mathbb{V}_{\mathbf{W}^{*(L)}} \varphi'(\alpha_{\nu i}^{(L-1)}) \Bigg] \end{aligned} \tag{5.171}$$

where K' is a positive constant. Calling D the double sum, $|A_2 - A_{12}|$ is then estimated as:

$$\begin{aligned} |A_2 - A_{12}| &\leq K'' \mathbb{E}_{\mathbf{a}^*} \sqrt{\mathbb{V}_{\setminus \mathbf{a}^*}[f_n]} \sqrt{D} \leq K'' \sqrt{\mathbb{E}_{\mathbf{a}^*} \mathbb{V}_{\setminus \mathbf{a}^*}[f_n]\, \mathbb{E}_{\mathbf{a}^*} D} \\ &= O\left(\sqrt{\left(\frac{1}{n} + \frac{1}{\mathrm{d}^{(L-1)}} \right) \left(\frac{n^2}{\mathrm{d}^{(L-1)} d^{(L)}} + \frac{n^2}{\mathrm{d}^{(L-1)\ 3/2}} \right)} \right) \end{aligned} \tag{5.172}$$

□

When all the Lemmas 20, 21 and 22 are collectively considered, it can be observed that the time derivative of the interpolating free entropy is constrained as follows:

$$\begin{aligned} \frac{d}{dt} \bar{f}_n(t) &= \underbrace{O\left(\sqrt{\left(1 + \frac{n}{\mathrm{d}^{(L-1)}} \right) \left(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}} \right)} \right)}_{A_{11}^{\text{off}}} + \underbrace{O\left(\sqrt{\left(1 + \frac{n}{\mathrm{d}^{(L-1)}} \right) \left(\frac{1}{d^{(L)}} + \frac{1}{\mathrm{d}^{(L-1)\ 1/2}} \right)} \right)}_{A_3 - A_{11}^{\text{diag}}} \\ &+ \underbrace{O\left(\sqrt{\left(1 + \frac{n}{\mathrm{d}^{(L-1)}} \right) \left(\frac{n}{\mathrm{d}^{(L-1)} d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}} \right)} \right)}_{A_{12} - A_2} \\ &= O\left(\sqrt{\left(1 + \frac{n}{\mathrm{d}^{(L-1)}} \right) \left(\frac{n}{d^{(L)}} + \frac{n}{\mathrm{d}^{(L-1)\ 3/2}} + \frac{1}{\mathrm{d}^{(L-1)\ 1/2}} \right)} \right) \end{aligned} \tag{5.173}$$

All the bounds found are uniform in $t \in [0, 1]$. This signifies the completion of the proof of Theorem 7.

5.5 Proof of Corollary 9

The proof of Corollary 9 exploits the proof of Theorem 9 to be carried out. Notably, although this result is a corollary from a mathematical perspective, it holds significant importance in our investigation. Given that the ultimate goal is to establish an equivalence between the original network and the generalized linear model from an information-theoretic perspective, this corollary is crucial as it demonstrates this equivalence between the $L+1$-layer neural network and the model obtained after linearization. The statement of the corollary is reiterated here prior to presenting the proof.

Corollary 23 (One step reduction mutual information equivalence)**.** *Assuming the same hypotheses as in Theorem 7, the following statement is obtained:*

$$|\frac{1}{n}I_n^{(k)}(\boldsymbol{\theta}^{*(k)};\mathcal{D}_n^{(k)}) - \frac{1}{n}I_n^{(k-1)}(\boldsymbol{\theta}^{*(k-1)};\mathcal{D}_n^{(k-1)})| = O\Big(\sqrt{\Big(1+\frac{n}{\mathbb{d}^{(k-1)}}\Big)\Big(\frac{n}{d^{(k)}}+\frac{n}{\mathbb{d}^{(k-1)\ 3/2}}+\frac{1}{\sqrt{\mathbb{d}^{(k-1)}}}\Big)}\Big) \tag{5.174}$$

Proof of Corollary 9. For the $L+1$-layer neural network, recall from (3.22) that the mutual information per sample can be described in this way:

$$\begin{aligned}\frac{1}{n}I_n^{(L)}(\boldsymbol{\theta}^{*(L)};\mathcal{D}_n^{(L)}) &= \frac{1}{n}H(\mathcal{D}_n^{(L)}) - \frac{1}{n}H(\mathcal{D}_n^{(L)} \mid \boldsymbol{\theta}^{*(L)}) \\ &= -\bar{f}_n^{(L)} + \mathbb{E}\log P_{\text{out}}\Big(Y_1^{(L)} \mid \frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\ \mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big)\end{aligned} \tag{5.175}$$

whereas for the reduced L-layers neural network the mutual information per sample can be rewritten as follows:

$$\frac{1}{n}I_n^{(L-1)}(\boldsymbol{\theta}^{*(L-1)};\mathcal{D}_n^{(L-1)}) = -\bar{f}_n^{(L-1)} + \mathbb{E}\log P_{\text{out}}\Big(Y_1^{(L-1)} \mid \rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}}\xi_1^{*(L-1)}\Big) \tag{5.176}$$

Exploiting that the teacher weights are Gaussian, in law it holds:

$$\frac{\mathbf{a}^{*\intercal}}{\sqrt{d^{(L)}}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big) \overset{\text{D}}{=} Z\sqrt{\frac{1}{d^{(L)}}\Big\|\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big\|^2} \tag{5.177}$$

with $Z \sim \mathcal{N}(0,1)$ and $\|\cdot\|$ the standard L^2 norm for vectors. Analogously, for the reduced model the following equality in distribution holds true:

$$\rho^{(L-1)}\frac{\mathbf{v}^{*(L-1)\intercal}\mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}} + \sqrt{\epsilon^{(L-1)}}\xi_1^{*(L-1)} \overset{\text{D}}{=} Z\sqrt{\rho^{(L-1)\ 2}\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} + \epsilon^{(L-1)}} \tag{5.178}$$

The aim is now to show that both the terms under square root on the right-hand side of (5.177) and (5.178) in the limit (so when the full model and the reduced model are equivalent in terms of free energy) tend to

$$\epsilon^{(L-1)} + \sigma^{(L-1)}\rho^{(L-1)\,2} = \mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2 \tag{5.179}$$

This allows to insert this result in (5.175) and (5.176), and thus to give an estimate of the residuals. To perform this analysis new quantities are defined:

$$S_d(t) = \sqrt{t\rho^{(L-1)\,2}\Big(\frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d^{(L-1)}} - \sigma^{(L-1)}\Big) + \epsilon^{(L-1)} + \sigma^{(L-1)}\rho^{(L-1)\,2}}\,, \quad \text{or equivalently} \tag{5.180}$$

$$S_d(t) = \sqrt{t\Big(\frac{1}{d^{(L)}}\Big\|\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big\|^2 - \mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2\Big) + \mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2} \tag{5.181}$$

and

$$\Psi(t) := \mathbb{E}\int dY\, P_{\text{out}}(Y \mid ZS_d(t)) \log P_{\text{out}}(Y \mid ZS_d(t)) \tag{5.182}$$

Using the results of Lemma 16 exploiting the definition of P_{out} in (3.3) and under the assumptions H1), H2) and H3) the following bound is obtained:

$$|\dot{\Psi}(t)| \le C(f)\mathbb{E}|Z||\dot{S}_d(t)|\,, \tag{5.183}$$

with $C(f)$ a constant depending on f. Applying now the fundamental theorem of integral calculus, the result is:

$$|\Psi(1) - \Psi(0)| \le C(f)\int_0^1 dt\,\mathbb{E}|\dot{S}_d(t)| \le C(f)\begin{cases} \dfrac{\mathbb{E}\Big|\frac{1}{d^{(L)}}\Big\|\varphi\Big(\frac{\mathbf{w}^{*(L)}\mathbf{x}_1^{(L-1)}}{\sqrt{d^{(L-1)}}}\Big)\Big\|^2 - \mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2\Big|}{\sqrt{\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2}} & L+1\text{-layers NN} \\ \rho^{(L-1)2}\dfrac{\mathbb{E}\Big|\frac{\|\mathbf{x}_1^{(L-1)}\|^2}{d^{(L-1)}} - \sigma^{(L-1)}\Big|}{\sqrt{\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2}} & L\text{-layers NN} \end{cases} \tag{5.184}$$

This quantity for the $L+1$-layers neural network is $O(\mathrm{d}^{(L)\,-1/2})$ whereas for the L-layers neural network it reads $O(\mathrm{d}^{(L-1)\,-1/2})$. Plugging this result in (5.175) and (5.176)

$$\begin{aligned} \frac{1}{n}I_n^{(L)}(\boldsymbol{\theta}^{*(L)};\mathcal{D}_n^{(L)}) &= -\bar{f}_n^{(L)} + \Psi(\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2) + O(\mathrm{d}^{(L)\,-1/2}) \\ \frac{1}{n}I_n^{(L-1)}(\boldsymbol{\theta}^{*(L-1)};\mathcal{D}_n^{(L-1)}) &= -\bar{f}_n^{(L-1)} + \Psi(\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2) + O(\mathrm{d}^{(L-1)\,-1/2}) \end{aligned} \tag{5.185}$$

where

$$\Psi(\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2) := \Psi(0) = \mathbb{E}\int dY P_{\text{out}}(Y \mid Z\sqrt{\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2}) \log P_{\text{out}}(Y \mid Z\sqrt{\mathbb{E}_{\mathcal{N}(0,\sigma^{(L-1)})}\varphi^2}) \tag{5.186}$$

Exploiting this derivation and using that $\frac{1}{\mathrm{d}^{(L)\ 1/2}} \leq \sqrt{\frac{1}{d^{(L)}} + \frac{1}{\mathrm{d}^{(L-1)}}} \leq \sqrt{\frac{1}{d^{(L)}} + \frac{1}{\sqrt{\mathrm{d}^{(L-1)}}}}$ the mutual information equivalence can be derived. □

Chapter 6

Conclusions and future works

This thesis presented an information-theoretical analysis of deep, nonlinear neural networks, focusing on training all network parameters in settings where the number of samples, input dimension, and network width are all large. The investigation was conducted in a teacher-student setup within a Bayes-optimal framework, where data are generated by a teacher network and classified by a student network with identical architecture. The primary goal was to extend results for 2-layer neural networks obtained in [19] to networks with an arbitrary number of layers by reducing them to generalized linear models through a recursive linearization approach.

Key contributions of this thesis include the analysis of a concentration of measure phenomenon and the derivation of information-theoretic bounds.

The studied concentration of measure phenomenon proves that Gaussian random vectors retain statistical properties similar to Gaussian vectors after being processed through non-linear neural network layers, providing a deeper understanding of the network's internal representations.

The bounds obtained, which express the difference in mutual information per sample between the dataset and the teacher network weights for deep neural networks compared to generalized linear models, represent a fundamental tool for understanding the conditions under which a deep neural network can be reduced to a GLM. The recursion parameters and the GLM parameters obtained through our recursive proof match the results by [29], where this equivalence was first conjectured. Concerning the allowed scalings such that the equivalence between the two models is verified, the match is only partial: in our investigation, the scenario in which the dataset size, input dimension, and hidden layers sizes tend to infinity at proportional rates is not recovered. However, the case in which the dataset size and input dimension scale proportionally while the hidden layers are larger and of comparable size is allowed by our bounds. Further research is thus necessary to determine whether this limitation is fundamental or specific to the current proof techniques.

Another crucial aspect that requires future investigation is the concentration of the free entropy density, which is a fundamental step in obtaining the information-theoretic bounds. This result was partially proved for 2-layer neural networks by [19], and assumed in our investigation. Fully proving this concentration is essential for the general validity of the bounds described in our analysis.

Additional future research directions could include exploring scenarios where the initial data are

not independent and described by a non-trivial covariance matrix, or cases in which they are drawn from a mixture of Gaussian distributions. Introducing a covariance for the input could model data dependencies. Drawing the input from a mixture of Gaussian distributions, which are universal approximators for distributions, instead, could provide insights about how different data distributions affect the performance of neural networks, revealing new insights into the versatility and robustness of neural network models and enhancing the applicability of the theoretical results to real-world data.

In conclusion, this thesis has extended the information-theoretical analysis of neural networks by establishing novel bounds through a recursive reduction method. Moreover, it has shed light on the internal mechanisms of neural networks by showing that Gaussian random vectors, when propagated through network layers, preserve statistical characteristics akin to those of Gaussian random vectors. These findings enhance our comprehension of neural network behaviors and suggest new pathways for investigation in the domain of machine learning and artificial intelligence.

Bibliography

[1] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: from theory to algorithms*. eng. New York: Cambridge university press, 2014.

[2] C. Bishop. *Pattern recognition and machine learning*. Springer, Jan. 2006.

[3] R. D. Reed and R. J. I. Marks. *Neural smithing : supervised learning in feedforward artificial neural networks*. eng. Cambridge (Mass.) : MIT press, 1999.

[4] U. Kamath and J. Liu. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. en. Cham: Springer International Publishing, 2021. DOI: 10.1007/978-3-030-83356-5.

[5] M. Mezard, G. Parisi and M. Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. en. Vol. 9. World Scientific Lecture Notes in Physics. WORLD SCIENTIFIC, Nov. 1986. DOI: 10.1142/0271.

[6] H. Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. International series of monographs on physics 111. OCLC: ocm47063323. Oxford ; New York: Oxford University Press, 2001.

[7] D. J. Amit. *Modeling brain function: the world of attractor neural networks*. eng. Transferred to digital print. Cambridge: Cambridge University Press, 1999.

[8] A. C. C. Coolen, R. Kuehn, P. Sollich, A. C. C. Coolen, R. Kuehn and P. Sollich. *Theory of Neural Information Processing Systems*. Oxford, New York: Oxford University Press, July 2005.

[9] D. J. Amit, H. Gutfreund and H. Sompolinsky. 'Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks'. en. *Physical Review Letters* 55.14 (Sept. 1985), pp. 1530–1533.

[10] E. Gardner and B. Derrida. 'Three unfinished works on the optimal storage capacity of networks'. en. *Journal of Physics A: Mathematical and General* 22.12 (June 1989), p. 1983.

[11] H. S. Seung, H. Sompolinsky and N. Tishby. 'Statistical mechanics of learning from examples'. *Physical Review A* 45.8 (Apr. 1992). Publisher: American Physical Society, pp. 6056–6091.

[12] T. L. H. Watkin, A. Rau and M. Biehl. 'The statistical mechanics of learning a rule'. *Reviews of Modern Physics* 65.2 (Apr. 1993). Publisher: American Physical Society, pp. 499–556.

[13] S. Mei, A. Montanari and P.-M. Nguyen. 'A Mean Field View of the Landscape of Two-Layers Neural Networks'. *Proceedings of the National Academy of Sciences* 115.33 (Aug. 2018). arXiv:1804.06561 [cond-mat, stat].

[14] J. Barbier, F. Krzakala, N. Macris, L. Miolane and L. Zdeborová. 'Optimal errors and phase transitions in high-dimensional generalized linear models'. *Proceedings of the National Academy of Sciences* 116.12 (Mar. 2019). Publisher: Proceedings of the National Academy of Sciences, pp. 5451–5460.

[15] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris and L. Zdeborová. 'The committee machine: Computational to statistical gaps in learning a two-layers neural network'. *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (Dec. 2019). arXiv:1806.05451 [cond-mat, physics:physics, stat], p. 124023.

[16] G. Rotskoff and E. Vanden-Eijnden. 'Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach'. *Communications on Pure and Applied Mathematics* 75.9 (Sept. 2022), pp. 1889–1935.

[17] H. Schwarze. 'Learning a rule in a multilayer neural network'. en. *Journal of Physics A: Mathematical and General* 26.21 (Nov. 1993), p. 5781.

[18] P.-M. Nguyen and H. T. Pham. *A Rigorous Framework for the Mean Field Limit of Multilayer Neural Networks*. arXiv:2001.11443 [cond-mat, stat]. Feb. 2023. DOI: 10.48550/arXiv.2001.11443.

[19] F. Camilli, D. Tieplova and J. Barbier. *Fundamental limits of overparametrized shallow neural networks for supervised learning*. arXiv:2307.05635 [cond-mat, stat]. July 2023. DOI: 10.48550/arXiv.2307.05635.

[20] L. Zdeborová and F. Krzakala. 'Statistical physics of inference: Thresholds and algorithms'. *Advances in Physics* 65.5 (Sept. 2016). arXiv:1511.02476 [cond-mat, stat], pp. 453–552.

[21] H. Schwarze and J. Hertz. 'Generalization in Fully Connected Committee Machines'. en. *Europhysics Letters* 21.7 (Mar. 1993), p. 785.

[22] A. Engel and C. Van Den Broeck. *Statistical Mechanics of Learning*. 1st ed. Cambridge University Press, Mar. 2001. DOI: 10.1017/CBO9781139164542.

[23] H. Schwarze and J. Hertz. 'Generalization in a Large Committee Machine'. en. *Europhysics Letters* 20.4 (Oct. 1992), p. 375.

[24] E. Barkai, D. Hansel and H. Sompolinsky. 'Broken symmetries in multilayered perceptrons'. *Physical Review A* 45.6 (Mar. 1992). Publisher: American Physical Society, pp. 4146–4161.

[25] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr and A. Zippelius. 'Storage capacity and learning algorithms for two-layer neural networks'. eng. *Physical Review. A, Atomic, Molecular, and Optical Physics* 45.10 (May 1992), pp. 7590–7609.

[26] R. Monasson and R. Zecchina. 'Weight Space Structure and Internal Representations: A Direct Approach to Learning and Generalization in Multilayer Neural Networks'. *Physical Review Letters* 75.12 (Sept. 1995). Publisher: American Physical Society, pp. 2432–2435.

[27] Q. Li and H. Sompolinsky. 'Statistical Mechanics of Deep Linear Neural Networks: The Back-propagating Kernel Renormalization'. *Physical Review X* 11.3 (Sept. 2021). Publisher: American Physical Society, p. 031059.

[28] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi and P. Rotondo. *A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit*. arXiv:2209.04882 [cond-mat]. Dec. 2023. DOI: 10.48550/arXiv.2209.04882.

[29] H. Cui, F. Krzakala and L. Zdeborová. *Bayes-optimal Learning of Deep Random Networks of Extensive-width*. arXiv:2302.00375 [cond-mat, stat]. June 2023. DOI: 10.48550/arXiv.2302.00375.

[30] J. A. Nelder and R. W. M. Wedderburn. 'Generalized Linear Models'. *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972). Publisher: [Royal Statistical Society, Wiley], pp. 370–384.

[31] P. McCullagh. 'Generalized linear models'. *European Journal of Operational Research* 16.3 (June 1984), pp. 285–292.

[32] P. McCullagh and J. A. Nelder. *Generalized linear models*. eng. Repr. Monographs on statistics and applied probability. London: Chapman and Hall, 1985.

[33] F. Rosenblatt. 'The perceptron: A probabilistic model for information storage and organization in the brain'. *Psychological Review* 65.6 (1958). Place: US Publisher: American Psychological Association, pp. 386–408.

[34] T. Hosaka, Y. Kabashima and H. Nishimori. 'Statistical mechanics of lossy data compression using a nonmonotonic perceptron'. *Physical Review E* 66.6 (Dec. 2002). Publisher: American Physical Society, p. 066126.

[35] R. Shwartz-Ziv and N. Tishby. *Opening the Black Box of Deep Neural Networks via Information*. arXiv:1703.00810 [cs]. Apr. 2017. DOI: 10.48550/arXiv.1703.00810.

[36] D. J. C. MacKay. *Information theory, inference, and learning algorithms*. eng. 22nd printing. Cambridge: Cambridge University Press, 2019.

[37] M. Mézard and A. Montanari. *Information, physics, and computation*. eng. Oxford graduate texts. Oxford: Oxford university press, 2009.

[38] C. E. Shannon. 'A mathematical theory of communication'. *The Bell System Technical Journal* 27.3 (July 1948). Conference Name: The Bell System Technical Journal, pp. 379–423.

[39] J. Jacod and P. Protter. *Probability Essentials*. en. Universitext. Berlin, Heidelberg: Springer, 2004. DOI: 10.1007/978-3-642-55682-1.

[40] J. P. Sethna. *Statistical mechanics: entropy, order parameters, and complexity*. eng. Repr. 2012 (twice). Oxford master series in physics Statistical, computational, and theoretical physics 14. Oxford: Oxford Univ. Press, 2012.

[41] R. J. Baxter. *Exactly solved models in statistical mechanics*. Dover ed. OCLC: ocn154799434. Mineola, N.Y: Dover Publications, 2007.

[42] J. M. Yeomans. *Statistical mechanics of phase transitions*. Oxford science publications. Oxford [England] : New York: Clarendon Press ; Oxford University Press, 1992.

[43] S. Friedli and Y. Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. eng. Cambridge New York: Cambridge university press, 2018.

[44] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini and M. Viale. 'Scale-free correlations in starling flocks'. *Proceedings of the National Academy of Sciences* 107.26 (June 2010). Publisher: Proceedings of the National Academy of Sciences, pp. 11865–11870.

[45] A. Attanasi et al. 'Information transfer and behavioural inertia in starling flocks'. en. *Nature Physics* 10.9 (Sept. 2014), pp. 691–696.

[46] D. Sherrington and S. Kirkpatrick. 'Solvable Model of a Spin-Glass'. *Physical Review Letters* 35.26 (Dec. 1975). Publisher: American Physical Society, pp. 1792–1796.

[47] S. F. Edwards and P. W. Anderson. 'Theory of spin glasses'. en. *Journal of Physics F: Metal Physics* 5.5 (May 1975), p. 965.

[48] G. Parisi. 'Infinite Number of Order Parameters for Spin-Glasses'. *Physical Review Letters* 43.23 (Dec. 1979). Publisher: American Physical Society, pp. 1754–1756.

[49] G. Parisi. 'A sequence of approximated solutions to the S-K model for spin glasses'. en. *Journal of Physics A: Mathematical and General* 13.4 (Apr. 1980), p. L115.

[50] F. Guerra. 'Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model'. *Communications in Mathematical Physics* 233.1 (Feb. 2003). arXiv:cond-mat/0205123, pp. 1–12.

[51] F. Guerra and F. L. Toninelli. 'The Thermodynamic Limit in Mean Field Spin Glass Models'. en. *Communications in Mathematical Physics* 230.1 (Sept. 2002), pp. 71–79.

[52] M. Talagrand. 'The Parisi formula'. en. *Annals of Mathematics* 163.1 (Jan. 2006), pp. 221–263.

[53] D. Panchenko. 'A connection between the Ghirlanda–Guerra identities and ultrametricity'. *The Annals of Probability* 38.1 (Jan. 2010). Publisher: Institute of Mathematical Statistics, pp. 327–347.

[54] S. Ghirlanda and F. Guerra. 'General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity'. *Journal of Physics A: Mathematical and General* 31.46 (1998), pp. 9149–9155.

[55] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. en. Springer Texts in Statistics. New York, NY: Springer New York, 2004. DOI: 10.1007/978-0-387-21736-9.

[56] G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. en. Springer Texts in Statistics. New York, NY: Springer US, 2021. DOI: 10.1007/978-1-0716-1418-1.

[57] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC, July 2015. DOI: 10.1201/b16018.

[58] D.-A. Clevert, T. Unterthiner and S. Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. arXiv:1511.07289 [cs]. Feb. 2016. DOI: 10.48550/arXiv.1511.07289.

[59] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala and L. Zdeborová. 'Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup'. *Journal of Statistical Mechanics: Theory and Experiment* 2020.12 (Dec. 2020). arXiv:1906.08632 [cond-mat, stat], p. 124010.

[60] S. Lee, S. Goldt and A. Saxe. *Continual Learning in the Teacher-Student Setup: Impact of Task Similarity*. arXiv:2107.04384 [cond-mat, stat]. July 2021. DOI: 10.48550/arXiv.2107.04384.

[61] C. M. Stein. 'Estimation of the Mean of a Multivariate Normal Distribution'. *The Annals of Statistics* 9.6 (Nov. 1981). Publisher: Institute of Mathematical Statistics, pp. 1135–1151.

[62] J. S. Liu. 'Siegel's formula via Stein's identities'. *Statistics & Probability Letters* 21.3 (Oct. 1994), pp. 247–251.

[63] H. Nishimori. 'Exact results and critical properties of the Ising model with competing interactions'. en. *Journal of Physics C: Solid State Physics* 13.21 (July 1980), p. 4071.

[64] H. Nishimori. 'Internal Energy, Specific Heat and Correlation Function of the Bond-Random Ising Model'. *Progress of Theoretical Physics* 66.4 (Oct. 1981), pp. 1169–1181.

[65] L. Miolane. 'Fundamental limits of inference : a statistical physics approach'. Number: 2019PSLEE043 tex.hal_id: tel-02976034 tex.hal_version: v1. Theses. Université Paris sciences et lettres, June 2019.

[66] J. Barbier. 'Mean- eld theory of high-dimensional Bayesian inference'. en. ().

[67] R. Van Handel. 'Probability in High Dimension:' in: Fort Belvoir, VA: Defense Technical Information Center, June 2014. DOI: 10.21236/ADA623999.

[68] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[69] T. Tao. *254A, Notes 1: Concentration of measure | What's new*. Blog post. 2010.

[70] J. Barbier, N. Macris and L. Miolane. *The Layered Structure of Tensor Estimation and its Mutual Information*. arXiv:1709.10368 [cond-mat, physics:math-ph] version: 2. Nov. 2018. DOI: 10.48550/arXiv.1709.10368.

[71] K. Joag-Dev, M. D. Perlman and L. D. Pitt. 'Association of Normal Random Variables and Slepian's Inequality'. *The Annals of Probability* 11.2 (May 1983). Publisher: Institute of Mathematical Statistics, pp. 451–455.

[72] J.-P. Kahane. 'Une inegalite du type de Slepian et Gordon sur les processus gaussiens'. fr. *Israel Journal of Mathematics* 55.1 (Feb. 1986), pp. 109–110.

[73] R. Boccagna and D. Gabrielli. 'Remarks on the Interpolation Method'. en. *Journal of Statistical Physics* 181.4 (Nov. 2020), pp. 1218–1238.

[74] F. Guerra and F. L. Toninelli. 'Quadratic replica coupling in the Sherrington–Kirkpatrick mean field spin glass model'. *Journal of Mathematical Physics* 43.7 (July 2002), pp. 3704–3716.

[75] D. Panchenko. *The Sherrington-Kirkpatrick Model*. en. Springer Monographs in Mathematics. New York, NY: Springer, 2013. DOI: 10.1007/978-1-4614-6289-7.

[76] J. Barbier and N. Macris. 'The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference'. en. *Probability Theory and Related Fields* 174.3 (Aug. 2019), pp. 1133–1185.

[77] J. Barbier and N. Macris. 'The adaptive interpolation method for proving replica formulas. Applications to the Curie-Weiss and Wigner spike models'. *Journal of Physics A: Mathematical and Theoretical* 52.29 (July 2019). arXiv:1901.06516 [cond-mat, physics:math-ph], p. 294002.

[78] S. Goldt, M. Mézard, F. Krzakala and L. Zdeborová. 'Modelling the influence of data structure on learning in neural networks: the hidden manifold model'. *Physical Review X* 10.4 (Dec. 2020). arXiv:1909.11500 [cond-mat, stat], p. 041044.

[79] D. Araújo, R. I. Oliveira and D. Yukimura. *A mean-field limit for certain deep neural networks*. arXiv:1906.00193 [cond-mat, stat]. June 2019. DOI: 10.48550/arXiv.1906.00193.

[80] J. Sirignano and K. Spiliopoulos. *Mean Field Analysis of Neural Networks: A Central Limit Theorem*. arXiv:1808.09372 [math, stat]. June 2019. DOI: 10.48550/arXiv.1808.09372.

[81] A. Shevchenko and M. Mondelli. *Landscape Connectivity and Dropout Stability of SGD Solutions for Over-parameterized Neural Networks*. arXiv:1912.10095 [cs, stat]. July 2020. DOI: 10.48550/arXiv.1912.10095.

[82] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mezard and L. Zdeborova. 'The Gaussian equivalence of generative models for learning with shallow neural networks'. en. In: *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. ISSN: 2640-3498. PMLR, Apr. 2022, pp. 426–471.

[83] H. Hu and Y. M. Lu. *Universality Laws for High-Dimensional Learning with Random Features*. arXiv:2009.07669 [cs, math]. Oct. 2022. DOI: 10.48550/arXiv.2009.07669.

Acknowledgments

I wish to thank my supervisor, Professor Jean Barbier, for giving me the opportunity to work on this project. His advice, passion, and curiosity have been incredibly inspiring, and I am grateful to have been part of such a wonderful research group. These months of research in Trieste have been both engaging and stimulating, greatly contributing to my scientific and personal growth.

I am also deeply thankful to Doctor Francesco Camilli and Doctor Daria Tieplova for their endless patience, insightful discussions, and valuable advice, both in science and beyond.

I extend my sincere gratitude to Professor Michele Allegra for his support, academic guidance, and kind availability.

A heartfelt thank you to my mom, dad, and sister for their constant support during my studies. I wouldn't be here without you. To my grandparents, thank you for always being there for me; you are my rocks.

Lastly, I am grateful to all my friends who have stood by me through good times and bad, offering support and sharing every moment.