Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime

Francesco Camilli   FCAMILLI@ICTP.IT
The Abdus Salam International Centre for Theoretical Physics
Trieste 34151, Italy

Daria Tieplova   DTIEPLOV@ICTP.IT
The Abdus Salam International Centre for Theoretical Physics
Trieste 34151, Italy

Eleonora Bergamin   EBERGAMI@SISSA.IT
International School for Advanced Studies
Trieste 34136, Italy

Jean Barbier   JBARBIER@ICTP.IT
The Abdus Salam International Centre for Theoretical Physics
Trieste 34151, Italy.

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We rigorously analyse fully-trained neural networks of arbitrary depth in the Bayesian optimal setting in the so-called *proportional scaling regime* where the number of training samples and width of the input and all inner layers diverge proportionally. We prove an information-theoretic equivalence between the Bayesian deep neural network model trained from data generated by a teacher with matching architecture, and a simpler model of optimal inference in a generalized linear model. This equivalence enables us to compute the optimal generalization error for deep neural networks in this regime. We thus prove the "deep Gaussian equivalence principle" conjectured in Cui et al. (2023). Our result highlights that in order to escape this "trivialisation" of deep neural networks (in the sense of reduction to a linear model) happening in the strongly overparametrized proportional regime, models trained from much more data have to be considered.

Keywords: Deep neural networks, Gaussian equivalence principle, Bayes-optimal learning, over-parameterization, proportional regime, interpolation method

1. Introduction

Neural networks are exceptional tools in many applications such as image classification and speech processing to name a few (Alzubaidi et al., 2021). They have therefore stimulated an unprecedented interest in their behavior, which is still understood mostly at the empirical level. The reason is that developing a quantitative theory for their performance is a challenging mathematical task. The challenge resides in the complex interaction of at least three aspects: *i)* their complex architecture; *ii)* the impact of the data structure; *iii)* the complex dynamics of the optimisation algorithms used for training. In the present paper, we shall focus on information-theoretical limits, which are algorithm-independent, and structureless input data, in order to highlight the phenomenology induced by the architecture alone, point *i)*. In order to study the fundamental limits of learning a certain target function, the correct framework to consider is the *teacher-student scenario*, where said target function plays the role of the "teacher" that generates the responses associated with the inputs. Then,

a "student" model belonging to the same hypothesis class must learn the target based on these examples. In the present setting this entails that the teacher and student neural networks will have the very same architecture and hyper-parameters, and only the realization of the teacher's weights is to be learnt. The importance of the teacher-student scenario was pinpointed in the seminal work of Gardner and Derrida (1989) on the perceptron.

The Bayes-optimal generalization error for the perceptron and generalized linear model was rigorously computed in Barbier et al. (2019). This model and the committee machine have been intensively studied since the nineties. We refer the reader to the (non-exhaustive) list of contributions: Barkai et al. (1992); Engel et al. (1992); Schwarze (1993); Schwarze and Hertz (1992, 1993); Monasson and Zecchina (1995); Saad and Solla (1995); Mato and Parga (1992); Engel and Van den Broeck (2001); Aubin et al. (2018); Baldassi et al. (2019); Goldt et al. (2020a,b, 2022). For these models where the number of hidden units is dimension-independent, the relevant interpolation regime corresponds to a number of samples diverging proportionally to the input dimension.

Concerning settings with a large number of hidden units we have the series of works Nitanda and Suzuki (2017); Sirignano and Spiliopoulos (2020); Araújo et al. (2019); Nguyen (2019); Nguyen and Pham (2020); Mei et al. (2018, 2019); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2022); Wu et al. (2022); Javanmard et al. (2020); Shevchenko et al. (2022); Shevchenko and Mondelli (2020); Mei et al. (2019), investigating online learning with SGD. Mean-field analyses of SGD differ from the information-theoretical one. SGD produces a "one-shot estimator" for labels, which is in general sub-optimal. Furthermore, in online learning, the student network sees one sample at a time, whereas in Bayes-optimal learning the whole dataset is jointly exploited.

As evidenced by Mei et al. (2019), when few samples are provided to the neural network during training, the weights are virtually fixed to their initialization, which is the setting captured by the Neural Tangent Kernel of Jacot et al. (2021) and similarly studied in random feature models and lazy training regimes (Rahimi and Recht, 2007; Li and Yuan, 2017; Rudi and Rosasco, 2017; Bach, 2017; Li and Liang, 2018; Allen-Zhu et al., 2019; Du et al., 2019b,a; Lee et al., 2020; Arora et al., 2019; Huang and Yau, 2020; Mei and Montanari, 2021; Ghorbani et al., 2019; Montanari and Zhong, 2020; Gerace et al., 2020; d'Ascoli et al., 2020; Geiger et al., 2020; Dhifallah and Lu, 2020; Bordelon et al., 2020; Hu and Lu, 2022; Schröder et al., 2024; Bosch et al., 2023; Schröder et al., 2023).

In Li and Sompolinsky (2021) the authors instead tackled the learning problem of all the weights for deep linear networks whose width is proportional to the input dimension, see also Zavatone-Veth et al. (2022). Their linear nature allows to solve the statistical mechanics of these networks by a subsequent integration of the degrees of freedom associated with the network weights. A whole subsequent line of work generalised their approach to tackle non-linear networks in the so-called *proportional regime*, in which the number of training examples n and the sizes of all layers (d_ℓ) diverge simultaneously and in a proportional way: Pacelli et al. (2023); Naveh and Ringel (2021); Seroussi et al. (2023); Aiudi et al. (2025); Bassetti et al. (2024); Rubin et al. (2025). In this strongly overparametrized proportional regime, these works show that *some* feature learning happens, in the sense that the kernel learnt by the network adapts to the data w.r.t. to the fixed kernel that emerges in the infinite width limit described by the neural network Gaussian processes (Neal, 1996; Williams, 1996; Lee et al., 2018; Matthews et al., 2018; Hanin, 2023). Still in the proportional regime but closer to our setting, Cui et al. (2023) conjecture a formula for the Bayes-optimal generalization error of a deep neural network using the replica method (Mézard et al., 1987). One key ingredient of their analysis is a "deep Gaussian equivalence principle" conjecture, a version of Gaussian Equivalence Principle (GEP). Classical results support the validity of GEPs in high-dimensional

inference (Sudakov, 1978; Diaconis and Freedman, 1984; Reeves, 2017; Meckes, 2012), such as in the description of relevant observables for shallow neural networks (Goldt et al., 2020b, 2022). Furthermore, the validity deep Gaussian equivalence principles has been proved at the level of deterministic equivalents for deep random feature models in Schröder et al. (2023).

In the present paper, we prove this deep Gaussian equivalence principle for the proportional scaling regime (and slightly beyond) and, as a consequence, the associated replica symmetric formula for the free entropy and generalisation error conjectured in Cui et al. (2023). Doing so, we substantially improve over Camilli et al. (2023) which was limited to a shallow neural network and did not reach the same convergence rates. The proof follows an induction strategy: in the considered scaling regime, we are able to show that a network with L layers reaches the same Bayes-optimal generalization error as a suitably tuned neural network with $L-1$ layers. In order to achieve this, we are required to prove that the non-linearity can be replaced by an "equivalent" linearity, which amounts to prove a Gaussian Equivalence Principle.

Let us finally mention the analyses of the multi-layer inference models Aubin et al. (2020, 2019), and especially the multi-layer GLM of Gabrié et al. (2018); Chen and Xia (2023). Although in the latter the weights are fixed, and thus not learnable as in our setting, it shares some common challenges due to the multi-layer structure, especially related to the concentration properties w.r.t. the quenched randomness of the model.

Notations.   Bold notations are reserved for vectors and matrices. A vector $\mathbf{x}$ is a column vector; $\mathbf{x}^\mathsf{T}$ is then a row vector. The L_2 norm is $\|\mathbf{x}\|^2 = \mathbf{x}^\mathsf{T}\mathbf{x}$; $\mathbf{x}\mathbf{x}^\mathsf{T}$ is a rank-one projector. $\mathbb{E}_A$ is an expectation with respect to the r.v. A; $\mathbb{E}$ is an expectation with respect to all ensuing r.v. 's. $\mathbb{V}(A)$ is the variance of A, and $\mathbb{V}_A(\cdot)$ the variance w.r.t. A only conditional on the rest. For a function F of one argument we denote F' its derivative. $[N] \equiv \{1, 2, \ldots, N\}$. We write $\mathbb{E}(\cdots)^2 = \mathbb{E}[(\cdots)^2] \geq (\mathbb{E}(\cdots))^2 = \mathbb{E}^2(\cdots)$. We shall abbreviate $\mathbb{E}[f(\cdots)] \equiv \mathbb{E}f(\cdots)$. For any functions f and g, $f = O(g)$ means there exists a constant K such that $|f| \leq K|g|$. Hence, given a r.v. Z, $f = O(Z)$ does *not* imply $\mathbb{E}f = 0$, but rather $|\mathbb{E}f| \leq \mathbb{E}|f| \leq K\mathbb{E}|Z|$. $\mathbb{1}_k$ is the identity matrix in k dimensions. Starred variables, e.g. $\mathbf{a}^*$, are the teacher's weights. Non-starred variables, e.g. $\mathbf{a}$, will denote the student's weights. $X \sim P$ means that the r.v. is drawn from, or has law, P.

2. Definitions and main results

We consider a teacher-student setting, where the training set, namely the set of n input-response couples $\mathcal{D}_L = \{(\mathbf{X}_\mu^{(0)}, Y_\mu)\}_{\mu=1}^n$, is generated by a teacher neural network with L layers fed with structureless inputs $\mathbf{X}_\mu^{(0)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbb{1}_{d_0})$. Responses Y_μ are assumed to be real numbers drawn from an output channel accounting for possible randomness in the output:

$$Y_\mu \sim P_{out}\Big(\cdot \mid \rho\frac{\mathbf{a}^{*\mathsf{T}}\mathbf{X}_\mu^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi_\mu^*\Big), \quad \mathbf{a}^* \in \mathbb{R}^{d_L}, \; \xi_\mu^* \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \tag{1}$$

where we have introduced the post-activations

$$\mathbf{X}_\mu^{(\ell)} = \varphi\Big(\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big) \in \mathbb{R}^{d_\ell}, \quad \mathbf{W}^{*(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \; \ell \in [L], \tag{2}$$

with the non-linearity φ applied component-wise. $\rho > 0$, and $\epsilon \geq 0$ can be taken to be zero and is introduced here for later convenience. The positive integers $(d_\ell)_{\ell=0}^L$ are thus the dimensions of the

layers. We group the teacher's weights in $\boldsymbol{\theta}^{*(L)} := (\mathbf{a}^*, (\mathbf{W}^{*(\ell)})_{\ell=1}^L)$, and we assume all of them are i.i.d. drawn from standard Gaussians:

$$a_i^*, W_{jl}^* \overset{\text{iid}}{\sim} P_\theta \equiv \mathcal{N}(0,1), \quad dP(\boldsymbol{\theta}^{*(L)}) := \prod_{i=1}^{d_L} dP_\theta(a_i^*) \prod_{\ell=1}^L \prod_{j=1}^{d_\ell} \prod_{k=1}^{d_{\ell-1}} dP_\theta(W_{jk}^{*(\ell)}). \tag{3}$$

A convenient representation for (1) is the following:

$$Y_\mu = f\Big(\rho \frac{\mathbf{a}^{*\intercal} \mathbf{X}_\mu^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi_\mu^*; \mathbf{A}_\mu\Big) + \sqrt{\Delta}Z_\mu, \quad Z_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{4}$$

Here $\mathbf{A}_\mu \in \mathbb{R}^p$, with p a given integer, is some possible stochasticity in the output and Δ is the strength of the label noise. $f : \mathbb{R} \mapsto \mathbb{R}$ is the readout function. The mapping between (1) and (4) is realized by the identification

$$P_{out}(y \mid x) = \int_{\mathbb{R}^p} \frac{dP_A(\mathbf{A})}{\sqrt{2\pi\Delta}} \exp\Big[-\frac{1}{2\Delta}(y - f(x; \mathbf{A}))^2\Big]. \tag{5}$$

Therefore, suitable regularity hypothesis on f will also reflect in the output kernel P_{out}.

The Bayes-optimal student network learns from $\mathcal{D}_L$ and estimates the response associated with a new input $\mathbf{X}_{new} \in \mathbb{R}^{d_0}$ as

$$\hat{Y}_{new} = \int y \, \mathbb{E}\Big[P_{out}\Big(y \mid \frac{\mathbf{a}^\intercal \mathbf{x}_{new}^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi_{new}\Big) \mid \mathcal{D}_L, \mathbf{X}_{new}\Big] dy \tag{6}$$

where $\xi_{new} \sim \mathcal{N}(0,1)$, and we have introduced the student's post activations

$$\mathbf{x}_{new}^{(\ell)} = \varphi\Big(\frac{\mathbf{W}^{(\ell)} \mathbf{x}_{new}^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big) \in \mathbb{R}^{d_\ell}, \quad \mathbf{W}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \ell \in [L], \tag{7}$$

with $\mathbf{x}_{new}^{(0)} \equiv \mathbf{X}_{new}$. Analogous relations hold for the student's post activations $\mathbf{x}_\mu^{(\ell)}$ associated with inputs in the training set. The measure $\mathbb{E}[\cdot \mid \mathcal{D}_L]$ that enters (6) is the Bayes' posterior measure, and it shall be denoted also by $\langle \cdot \rangle^{(L)}$. The estimator (6) provably yields (see Barbier et al. (2019); Camilli et al. (2023)) the minimum (average) generalization error

$$\mathcal{E}^{(L)} := \mathbb{E}(Y_{new} - \hat{Y}_{new})^2 \tag{8}$$

where the expectation is w.r.t. to the law of the training set $\mathcal{D}_L$ and the test data $(\mathbf{X}_{new}, Y_{new})$.

We bypass the computation of the Bayes-optimal generalization error (8) by computing the limiting Mutual Information (MI) between the training set $\mathcal{D}_L$, and the teacher's weights $\boldsymbol{\theta}^{*(L)} := (\mathbf{a}^*, (\mathbf{W}^{*(\ell)})_{\ell=1}^L)$ and the variables $\boldsymbol{\xi}^* = (\xi_\mu^*)_{\mu=1}^n$. The MI can be written in a convenient way, in terms of the normalization of the Bayes posterior

$$dP(\boldsymbol{\theta}^{(L)}, \boldsymbol{\xi} \mid \mathcal{D}_L) = \frac{1}{\mathcal{Z}_L(\mathcal{D}_L)} dP(\boldsymbol{\theta}^{(L)}) \prod_{\mu=1}^n dP_\theta(\xi_\mu) P_{out}\Big(Y_\mu \mid \rho \frac{\mathbf{a}^\intercal \mathbf{x}_\mu^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi_\mu\Big) \tag{9}$$

that is

$$\mathcal{Z}_L(\mathcal{D}_L) = \int \prod_{\mu=1}^{n} \mathbb{E}_\xi P_{out}\Big(Y_\mu \mid \rho\frac{\mathbf{a}^\intercal \mathbf{x}_\mu^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi_\mu\Big) dP(\boldsymbol{\theta}^{(L)})\,. \tag{10}$$

With these notations the MI per data point rewrites as

$$\frac{1}{n}I_n^{(L)}(\boldsymbol{\theta}^{*(L)}, \boldsymbol{\xi}^*; \mathcal{D}_L) = \mathbb{E}\log P_{out}\Big(Y_1 \mid \rho\frac{\mathbf{a}^{*\intercal}\mathbf{X}_1^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi^*\Big) - \frac{1}{n}\mathbb{E}\log \mathcal{Z}_L(\mathcal{D}_L)\,. \tag{11}$$

The last term is called *free entropy* and shall be denoted $\bar{f}_n^{(L)} := \frac{1}{n}\mathbb{E}\log \mathcal{Z}_L(\mathcal{D}_L)$. In addition, the MI with $L = 0$, which means in the absence of hidden layers, has to be interpreted as the one of a Generalized Linear Model (GLM), see Barbier et al. (2019). We can now state our main result.

Theorem 1 (Layer reduction) *Let the activation and readout functions $\varphi, f \in C^2(\mathbb{R})$ be odd and with bounded first and second derivatives. Let Z be a standard Gaussian, assume $L \geq 1$ and define the coefficients*

$$\sigma_0 = 1\,, \ \sigma_\ell = \mathbb{E}\varphi^2(Z\sqrt{\sigma_{\ell-1}})\,, \quad \rho_\ell = \mathbb{E}\varphi'(Z\sqrt{\sigma_{\ell-1}}), \quad \epsilon_\ell = \sigma_\ell - \sigma_{\ell-1}\rho_\ell^2\,, \tag{12}$$

for all $\ell \in [L]$. Denote by $d_m = \min\{(d_\ell)_{\ell=0}^L\}$. Then

$$\Big|\frac{1}{n}I_n^{(L)}(\boldsymbol{\theta}^{*(L)}, \boldsymbol{\xi}^*; \mathcal{D}_L) - \frac{1}{n}I_n^{(L-1)}(\boldsymbol{\theta}^{*(L-1)}, \boldsymbol{\xi}^*; \mathcal{D}_{L-1})\Big| = O\Big(\Big(1 + \sqrt{\frac{n}{d_m}} + \frac{n}{d_m}\Big)\frac{1}{\sqrt{d_m}}\Big)\,, \tag{13}$$

where $I_n^{(L-1)}(\boldsymbol{\theta}^{*(L-1)}, \boldsymbol{\xi}^*; \mathcal{D}_{L-1})$ is the mutual information between the hidden teacher weights, $\boldsymbol{\xi}^*$ and the data set $\mathcal{D}^{L-1} = \{(\mathbf{X}_\mu^{(0)}, Y_\mu^{(L-1)})\}_{\mu=1}^n$ with responses drawn as

$$Y_\mu^{(L-1)} \sim P_{out}\Big(\cdot \mid \rho\rho_L\frac{\mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} + \sqrt{\rho^2\epsilon_L + \epsilon}\xi_\mu^*\Big)\,. \tag{14}$$

This Theorem shows that in the linear regime, at the information-theoretic level, layers can be eliminated one by one paying the price of extra Gaussian noise. This is the manifestation of the Gaussian Equivalence Principle in our setting. In fact, what the result asserts is that the non-linearity for the last layer can be replaced by a linear equivalent. The values of the coefficients ϵ_L, ρ_L that re-weight the two new terms appearing in the equivalent model with one less layer are the same found in Mei and Montanari (2021); Goldt et al. (2020b); Hu and Lu (2022); Camilli et al. (2023).

One key ingredient needed to prove this type of result is the concentration of the free entropy w.r.t. the teacher and data instance. Even though this is a natural conjecture, it becomes particularly hard to prove in the multi-layer model due to the number of random parameters and their intertwined dependences. We thus state here this stand alone result which may be of independent interest.

Theorem 2 (Free entropy concentration) *Consider the activation and readout function φ, f as in Theorem 1. Then there exists a non-negative constant $C(f, \varphi)$ such that*

$$\mathbb{V}\Big(\frac{1}{n}\log \mathcal{Z}_L(\mathcal{D}_L)\Big) \leq C(f, \varphi)\Big(\frac{1}{n} + \frac{1}{d_m}\Big)\,. \tag{15}$$

The layer reduction Theorem can be iterated until exhaustion of the deep Bayesian network, yielding a Generalized Linear Model. During the reduction process one needs to keep track of how the re-weighting coefficients accumulate. After $L - k$ reduction steps, which means k layers are still left, the network model will then be described by equation (1), where ρ and ϵ are replaced by η_k and γ_k respectively, defined as follows:

$$\eta_k = \rho \prod_{i=k+1}^{L} \rho_i \,, \qquad \gamma_k = \epsilon + \rho^2 \sum_{j=k+1}^{L} \epsilon_j \prod_{i=j+1}^{L} \rho_i^2 \,, \qquad (16)$$

where by convention $\prod_{i=L+1}^{L}(\,\cdot\,) = 1$, $\eta_L = \rho$ and $\gamma_L = \epsilon$. The coefficients in (16) match the ones in Cui et al. (2023). The target model is the GLM where $k = 0$, namely when all the L hidden layers have been eliminated and only the readout layer remains. From our previous results, we have the following Corollary, proved in Appendix F:

Corollary 3 (Mutual information equivalence) *Under the same assumptions as Theorem 1, the following holds true:*

$$\left| \frac{1}{n} I_n^{(L)}(\boldsymbol{\theta}^{*(L)}, \boldsymbol{\xi}^*; \mathcal{D}_L) - \frac{1}{n} I_n^{(0)}(\boldsymbol{\theta}^{*(0)}, \boldsymbol{\xi}^*; \mathcal{D}_0) \right| = O\left(\left(1 + \sqrt{\frac{n}{d_m}} + \frac{n}{d_m} \right) \frac{1}{\sqrt{d_m}} \right), \qquad (17)$$

where $I_n^{(0)}(\boldsymbol{\theta}^{*(0)}, \boldsymbol{\xi}^*; \mathcal{D}_0)$ is the mutual information associated with the data set $\mathcal{D}$ with responses

$$Y_\mu^{(0)} \sim P_{out}\left(\cdot \mid \eta_0 \frac{\mathbf{a}^{*\intercal} \mathbf{X}_\mu^{(0)}}{\sqrt{d_0}} + \sqrt{\gamma_0} \xi_\mu^* \right).$$

The estimate in this corollary holds for finite sizes of layers and training set. Furthermore, it identifies the scaling regime in which the two models become asymptotically equivalent, namely when the right-hand side of (17) approaches zero as $n, (d_\ell)_{\ell=0}^{L} \to \infty$. We denote any such scaling limit by $\widetilde{\lim}$. Notice that this result shows that the trivialisation of deep neural networks in the Bayesian optimal setting happens even beyond the proportional regime $n = \Theta(d)$. Indeed, the r.h.s. of (17) vanishes also if $n = o(d_m^{3/2})$. Notice also that from $I_n(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*; \mathcal{D})$ it is possible to obtain the MI between $\boldsymbol{\theta}^*$ and $\mathcal{D}$ using the chain rule: $I_n(\boldsymbol{\theta}^*; \mathcal{D}) = I_n(\boldsymbol{\theta}^*, \boldsymbol{\xi}^*; \mathcal{D}) - I_n(\boldsymbol{\xi}^*; \mathcal{D} \mid \boldsymbol{\theta}^*)$. The limit of the last term in the latter is computed in Lemma 18.

The mutual information between the GLM teacher network weights and the corresponding training dataset is a quantity that has been studied in the literature (Barbier et al. (2019)), and its relevance lies in the fact that the generalization error can be computed from it. As a consequence, we can prove the following:

Theorem 4 (Generalization error equivalence) *Under the same hypothesis of Theorem 1 we have*

$$\widetilde{\lim} |\mathcal{E}^{(L)} - \mathcal{E}^{(0)}| = 0 \,, \qquad (18)$$

where $\mathcal{E}^{(0)}$ is the GLM generalization error associated with $\frac{1}{n} I_n^{(0)}(\boldsymbol{\theta}^{*(0)}, \boldsymbol{\xi}^*; \mathcal{D}_0)$.

It is important to note here that the deep neural network and the GLM are trained on two different datasets, each generated by the corresponding teacher network with a matching architecture. Specifically, $\mathcal{E}^{(0)}$ is not the generalization error attained by a GLM trained on $\mathcal{D}_L$.

The proof of Theorem 4 follows exactly that of Theorem 3 in Camilli et al. (2023) which is in turn leveraging on that of Theorem 4 in Barbier et al. (2019).

3. Outline of the proofs

3.1. Proof of Theorem 1

The proof uses an interpolation argument, similar to those used in the rigorous theory of spin-glass (Guerra and Toninelli, 2002) and inference (Barbier and Macris, 2019). Since it does not change the derivation, we select $\rho = 1, \epsilon = 0$ for clarity. The interpolation is defined at the level of the inference problem itself. More specifically, define

$$S_{t\mu} := \sqrt{1-t}\frac{\mathbf{a}^{*\intercal}}{\sqrt{d_L}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big) + \sqrt{t}\rho_L\frac{\mathbf{v}^{*\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} + \sqrt{t\epsilon_L}\zeta_\mu^{*(L)}, \qquad (19)$$

$$s_{t\mu} := \sqrt{1-t}\frac{\mathbf{a}^{\intercal}}{\sqrt{d_L}}\varphi\Big(\frac{\mathbf{W}^{(L)}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big) + \sqrt{t}\rho_L\frac{\mathbf{v}^{\intercal}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} + \sqrt{t\epsilon_L}\zeta_\mu^{(L)}, \qquad (20)$$

for any $t \in [0,1]$ and with independent Gaussian vector $\mathbf{v}^* \sim \mathcal{N}(0, \mathbb{1}_{d_L})$ and i.i.d. standard Gaussian noise variables $\zeta_\mu^{*(L)}, \zeta_\mu^{(L)}$, and $\mathbf{x}_\mu^{(0)} = \mathbf{X}_\mu^{(0)}$. Let the responses be drawn from $Y_{t\mu} \sim P_{out}(\cdot \mid S_{t\mu})$. The parameters to infer in this model are $\boldsymbol{\theta}^{*(L)} = (\mathbf{a}^*, (\mathbf{W}^{*(\ell)}))$ and $\mathbf{v}^*$. Let us introduce the convenient notation $u_y(x) = \log P_{out}(y \mid x), u'_y(x) = \partial_x u_y(x)$, or sometimes just u, u' for brevity. For this model we shall have an interpolating training set $\mathcal{D}_t = \{(Y_{t\mu}, \mathbf{X}_\mu^{(0)})_{\mu=1}^n\}$, an interpolating Bayes posterior for $\boldsymbol{\theta}^{*(L)} \mid \mathcal{D}_t$, with expectation operator

$$\langle\,\cdot\,\rangle_t = \frac{1}{\mathcal{Z}_t}dP(\boldsymbol{\theta}^{(L)})\prod_{i=1}^{d_{L-1}}dP_\theta(v_i)\prod_{\mu=1}^n dP_\theta(\zeta_\mu^{(L)})\exp\Big(u_{Y_{t\mu}}(s_{t\mu})\Big)(\cdot) \qquad (21)$$

and an associated interpolating free entropy

$$\bar{f}_{n,t} = \frac{1}{n}\mathbb{E}_{(t)}\log\mathcal{Z}_t = \frac{1}{n}\mathbb{E}_{(t)}\log\int dP(\boldsymbol{\theta}^{(L)})\mathbb{E}_\mathbf{v}\prod_{\mu=1}^n\mathbb{E}_{\zeta_\mu^{(L)}}\exp\Big(u_{Y_{t\mu}}(s_{t\mu})\Big). \qquad (22)$$

For future convenience we also introduce the expectation over the law of the dataset $\mathcal{D}_t$:

$$\mathbb{E}_{(t)}[\,\cdot\,] = \mathbb{E}_{\boldsymbol{\theta}^{*(L)},\mathbf{v}^*}\int\prod_{\mu=1}^n dY_{t\mu}\,\mathbb{E}_{\mathbf{X}_\mu^{(0)},\zeta_\mu^{*(L)}}\exp\Big(u_{Y_{t\mu}}(S_{t\mu})\Big)[\,\cdot\,]. \qquad (23)$$

At $t = 0$ the model is precisely the L-layer neural network, whereas at $t = 1$ it reduces to the target $L - 1$-layer neural network (recall $\rho = 1, \epsilon = 0$). Therefore, the goal is to control the derivative of the interpolating mutual information (obtained simply from the interpolating free entropy) w.r.t. t and to prove it is vanishing. Here we shall focus only on the contribution to the difference on the l.h.s. of (13) given by the free entropies for L and $L - 1$ layers. The control of the first term on the r.h.s. in (11) is deferred to the Appendix, see Lemma 17. The derivative of $\bar{f}_{n,t}$ decomposes as

$$\frac{d}{dt}\bar{f}_{n,t} = -A_1 + A_2 + A_3 + B \qquad (24)$$

where, denoting $u'_\mu := \partial_x \log P_{out}(Y_{t\mu} \mid x)|_{x=S_{t\mu}}$ (not to be confused with $u'_{Y_{t\mu}}(s_{t\mu})$) and similarly for higher derivatives u''_μ, u'''_μ,

$$A_1 := \frac{1}{2n\sqrt{(1-t)}}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_\mu \frac{\mathbf{a}^{*\intercal}}{\sqrt{d_L}}\Big[\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big) - \rho_L \frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big], \quad (25)$$

$$A_2 := \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_\mu \rho_L \Big[\frac{\mathbf{v}^{*\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{td_{L-1}}} - \frac{\mathbf{a}^{*\intercal}\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{(1-t)d_L d_{L-1}}}\Big], \quad (26)$$

$$A_3 := \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} u'_\mu \sqrt{\frac{\epsilon_L}{t}}\zeta_\mu^{*(L)}, \quad (27)$$

$$B := \frac{1}{n}\mathbb{E}_{(t)}\Big\langle \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(s_{t\mu})\frac{ds_{t\mu}}{dt}\Big\rangle_t. \quad (28)$$

The statistical model associated with the posterior (21) fulfills a set of symmetries called Nishimori identities, which are proved in Appendix A. These symmetries force B to vanish identically; see Lemma 13. Notice also that A_1 and A_2 share a common term which in one case is added, and in the other is subtracted. The reason for this operation is the following: in A_2 the two terms in the square parenthesis can be shown to have matching second moments, therefore they are supposed to compensate and cancel each other, which ends up being the case. The proof requires an integration by parts w.r.t. the Gaussian laws of $\mathbf{v}^*$ and $\mathbf{W}^{*(L)}$, and is deferred to Appendix E. Proving that A_1 and A_3 cancel each other requires more care. Let us explain the key ideas below.

Let us define $\boldsymbol{\alpha}_\mu := \mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}/\sqrt{d_{L-1}}$ and $U_{\mu\nu} := \delta_{\mu\nu}u''_\mu + u'_\mu u'_\nu$ for future convenience. We start from A_1 which is the most challenging. A natural way to simplify the expression is to integrate by parts the Gaussian readout weights $\mathbf{a}^*$, obtaining

$$A_1 = \frac{1}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu,\nu=1}^{n} U_{\mu\nu}\frac{1}{d_L}\big[\varphi(\boldsymbol{\alpha}_\mu)^\intercal\varphi(\boldsymbol{\alpha}_\nu) - \rho_L \boldsymbol{\alpha}_\mu^\intercal\varphi(\boldsymbol{\alpha}_\nu)\big]. \quad (29)$$

Notice that the sum over the data index μ has doubled since the same weights are interacting with all training sample. This will happen every time we integrate the network weights by parts. On the contrary, the variables $\zeta_\mu^{*(L)}$ are different for every training sample, hence an integration by parts of them in A_3 yields

$$A_3 = \frac{\epsilon_L}{2n}\mathbb{E}_{(t)} \log \mathcal{Z}_t \sum_{\mu=1}^{n} U_{\mu\mu}. \quad (30)$$

Thanks to the fact that $\mathbb{E}[U_{\mu\nu} \mid \mathbf{X}_\mu, \mathbf{X}_\nu, \boldsymbol{\theta}^{*(L)}] = 0$ (see Lemma 12), we can subtract from $\log \mathcal{Z}_t$ its expectation without changing the value of the expressions. Therefore,

$$A_3 - A_1 = \frac{1}{2}\mathbb{E}_{(t)}\Big(\frac{1}{n}\log \mathcal{Z}_t - \bar{f}_{n,t}\Big)\sum_{\mu=1}^{n} U_{\mu\mu}\Big[\epsilon_L - \frac{\|\varphi(\boldsymbol{\alpha}_\mu)\|^2}{d_L} + \rho_L \frac{\boldsymbol{\alpha}_\mu^\intercal\varphi(\boldsymbol{\alpha}_\mu)}{d_L}\Big]$$

$$- \frac{1}{2}\mathbb{E}_{(t)}\Big(\frac{1}{n}\log \mathcal{Z}_t - \bar{f}_{n,t}\Big)\sum_{\mu\neq\nu,1}^{n} U_{\mu\nu}\frac{1}{d_L}\big[\varphi(\boldsymbol{\alpha}_\mu)^\intercal\varphi(\boldsymbol{\alpha}_\nu) - \rho_L \boldsymbol{\alpha}_\mu^\intercal\varphi(\boldsymbol{\alpha}_\nu)\big]. \quad (31)$$

Let us focus on the first line. The strategy is to bound it using Cauchy-Schwartz, separating the difference between the random free entropy and its expectation from the summation over μ. This yields a variance of the random free entropy that is $O(\sqrt{1/n + 1/d_m})$ by Theorem 2. The other factor produced by Cauchy-Schwartz is an expectation of a double sum over $\mu, \nu = 1, \ldots, n$. For this, we can use the fact that $\mathbb{E}[U_{\mu\mu} \mid \mathbf{X}_\mu, \boldsymbol{\theta}^{*(L)}] = 0$ which restricts the sum to $\mu = \nu$. Furthermore, using that $\mathbb{E}[U_{\mu\mu}^2 \mid \mathbf{X}_\mu, \boldsymbol{\theta}^{*(L)}]$ is bounded by a constant (see Lemma 12), we are left with n expectations of the square parenthesis squared. Here the choice of ρ_L and ϵ_L becomes crucial, and it allows us to prove that the expectation of such square is $O(1/d_m)$ in Lemma 16.

The second line requires a more refined treatment. The summation over $\mu \neq \nu$ involves $O(n^2)$ centered random variables, which seems to go against our statement. What saves the day is that for $\mu \neq \nu$ the two terms in the square parenthesis are both tending to zero in the thermodynamic limit and in addition their leading orders are canceling each other, leaving us with a faster decay rate.

In order to access these leading orders, we use a further *circular interpolation*, that paves the way for a crucial integration by parts. More precisely, let $\tau \in [0, \pi/2]$ be a second interpolating parameter, and let $\tilde{\mathbf{W}}^*$ be an independent copy of $\mathbf{W}^{*(L)}$. Define then

$$\boldsymbol{\alpha}_\mu(\tau) := \boldsymbol{\alpha}_\mu \sin \tau + \frac{\tilde{\mathbf{W}}^* \mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} \cos \tau \, . \tag{32}$$

The great advantage of this interpolation is that $\dot{\boldsymbol{\alpha}}_\mu(\tau) := \partial_\tau \boldsymbol{\alpha}_\mu(\tau)$ is independent of $\boldsymbol{\alpha}_\mu(\tau)$ in the measure $\mathbb{E}_{\mathbf{W}^{*(L)}, \tilde{\mathbf{W}}^*}$, because they are Gaussian (conditionally on $\mathbf{X}_\mu^{(L-1)}$) and decorrelated. Using the fundamental theorem of integral calculus we can rewrite the second line of (31) as

$$A_1^{off} := \int_0^{\pi/2} d\tau \, \mathbb{E}_{(t)} \Big(\frac{1}{n} \log \mathcal{Z}_t - \bar{f}_{n,t}\Big) \sum_{\mu < \nu, 1}^{n} U_{\mu\nu} \frac{1}{d_L} \dot{\boldsymbol{\alpha}}_\mu(\tau) \circ \big[\varphi'(\boldsymbol{\alpha}_\mu(\tau)) - \rho_L \mathbf{1}\big]^\mathsf{T} \varphi(\boldsymbol{\alpha}_\nu) \, , \tag{33}$$

where $\mathbf{1} \in \mathbb{R}^{d_L}$ is the all-ones vector, and $\circ$ is the Hadamard entry-wise product. Notice that there is a term related to the $\tau = 0$ point of the τ-interpolation, that contains only $\varphi(\tilde{\mathbf{W}}^* \mathbf{X}_\mu^{(L-1)} / \sqrt{d_{L-1}})$. This factor is independent of everything else under $\mathbb{E}_{\tilde{\mathbf{W}}^*}$, and since φ is odd it vanishes.

Now we finally have a Gaussian variable $\dot{\boldsymbol{\alpha}}_\mu(\tau)$ (again, conditionally on $\mathbf{X}_\mu^{(L-1)}$, i.e., w.r.t. the measure of the weights) to integrate by parts which, importantly, lies outside of the non-linearity. Another key point is that $\dot{\boldsymbol{\alpha}}_\mu(\tau)$ is independent of $\boldsymbol{\alpha}_\mu(\tau)$, but not of the other $\boldsymbol{\alpha}_\nu$'s. Hence, the following correlations appear: $\mathbb{E}_{\tilde{\mathbf{W}}^*, \mathbf{W}^{*(L)}}[\dot{\alpha}_{i\mu}(\tau)\alpha_{j\eta}] = \delta_{ij} \mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\eta^{(L-1)} \cos \tau / d_{L-1}$. Defining $U_{\mu\nu\eta} := u''_\mu u'_\nu \delta_{\mu\eta} + u'_\mu u''_\nu \delta_{\nu\eta} + u'_\mu u'_\nu u'_\eta$ an integration by parts w.r.t. $\dot{\boldsymbol{\alpha}}_\mu(\tau)$ then yields

$$A_1^{off} = \int_0^{\pi/2} d\tau \, \cos \tau \, \mathbb{E}_{(t)} \Big(\frac{1}{n} \log \mathcal{Z}_t - \bar{f}_{n,t}\Big) \sum_{\mu < \nu, 1}^{n} \sum_{\eta=1}^{n} U_{\mu\nu\eta} \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\eta^{(L-1)}}{d_{L-1}} \tag{34}$$

$$\times \sqrt{1-t} \sum_{i=1}^{d_L} \frac{a_i^* \varphi'(\alpha_{i\eta})}{\sqrt{d_L}} \frac{\varphi'(\alpha_{i\mu}(\tau)) - \rho_L}{d_L} \varphi(\alpha_{i\nu})$$

$$+ \int_0^{\pi/2} d\tau \, \cos \tau \, \mathbb{E}_{(t)} \Big(\frac{1}{n} \log \mathcal{Z}_t - \bar{f}_{n,t}\Big) \sum_{\mu < \nu, 1}^{n} U_{\mu\nu} \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}} \sum_{i=1}^{d_L} \frac{\varphi'(\alpha_{i\mu}(\tau)) - \rho_L}{d_L} \varphi'(\alpha_{i\nu}) \, .$$

As we can see, now the scalar products between post-activations for different input data appear. This is crucial, as the post-activations preserve the "quasi-orthogonality" of the inputs, meaning $\mathbb{E}(\mathbf{X}_\mu^{(\ell)} \cdot \mathbf{X}_\nu^{(\ell)}/d_\ell)^{2k} = O(\min\{(d_s)_{s=0}^\ell\}^{-k})$ whenever $\mu \neq \nu$, as proved in Lemma 10.

Bearing in mind this fact, for both terms in (34) the proof proceeds with Cauchy-Schwartz inequality, separating the difference of free entropies from the summations over the training set indices μ, ν and η. For both mentioned terms, the other factor produced by Cauchy-Schwartz inequality is a square of multiple summations of training set indices. In any case, said summations are simplified thanks to Lemma 12. Specifically, using $\mathbb{E}[U_{\mu\nu\eta} \mid \mathbf{X}_\mu, \mathbf{X}_\nu, \mathbf{X}_\eta, \boldsymbol{\theta}^{*(L)}] = 0 = \mathbb{E}[U_{\mu\nu} \mid \mathbf{X}_\mu, \mathbf{X}_\nu, \boldsymbol{\theta}^{*(L)}]$ and $\mathbb{E}[U_{\mu\nu\eta}^2 \mid \mathbf{X}_\mu, \mathbf{X}_\nu, \mathbf{X}_\eta, \boldsymbol{\theta}^{*(L)}], \mathbb{E}[U_{\mu\nu}^2 \mid \mathbf{X}_\mu, \mathbf{X}_\nu, \boldsymbol{\theta}^{*(L)}] \leq C$ for some positive constant C. This reduces the number of indices we are summing over. With only these few ingredients, the first term in (34) is shown to be of the claimed order. Concerning the second term, one in addition needs to exploit that $(\sum_i \varphi'(\alpha_{i\mu}(\tau))/d_L - \rho_L)^2$ is concentrating to 0 thanks to the definition of ρ_L. This concentration result requires a certain care and is dealt with in Lemma 9.

3.2. Proof of Theorem 2

In this section we discuss briefly the proof of Theorem 2. Even though it is formulated for the model (1)-(4), we consider the interpolated model (19)-(21) as the concentration of its free entropy is required in the proof of Theorem 1; it anyway recovers the initial model as $t = 0$. To prove the concentration of the free entropy $f_{n,t}$ we adopt standard techniques such as Poincaré-Nash and Efron-Stein inequalities. For details, we refer to Appendix G. We remind that $f_{n,t}$ is a function of the random parameters $(\mathbf{W}^{*(\ell)})_{\ell=1}^L, (\mathbf{X}_\mu^{(0)}, \zeta_\mu^{*(L)}, Z_\mu, \mathbf{A}_\mu)_{\mu=1}^n, \mathbf{a}^*, \mathbf{v}^*$, and defined as

$$f_{n,t} = \frac{1}{n} \log \mathcal{Z}_t = \frac{1}{n} \log \int dP(\boldsymbol{\theta}^{(L)}) \mathbb{E}_\mathbf{v} \prod_\mu^n \mathbb{E}_{\zeta_\mu^{(L)}} P_{out}(Y_{t\mu}|s_{t\mu}), \tag{35}$$

where $Y_{t\mu} = f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta} Z_\mu$. We will prove the concentration with respect to all random parameters in several steps. The first batch is $\psi_1 = (\mathbf{X}_\mu^{(0)}, \zeta_\mu^{*(L)}, Z_\mu)_{\mu=1}^n$, second is $\psi_2 = (\mathbf{A}_\mu)$, and lastly $\psi_3 = ((\mathbf{W}^{*(\ell)})_{\ell=1}^L, \mathbf{a}^*, \mathbf{v}^*)$. For this, we use the variance decomposition

$$\mathbb{V}(f_{n,t}) = \mathbb{E}\mathbb{V}_{\psi_1}(f_{n,t}) + \mathbb{E}\mathbb{V}_{\psi_2}(\mathbb{E}_{\psi_1} f_{n,t}) + \mathbb{V}_{\psi_3}(\mathbb{E}_{\psi_1,\psi_2} f_{n,t}). \tag{36}$$

Since all elements of set ψ_1 are independent Gaussian variables, we bound the first term by the Poincaré-Nash inequality (see Appendix G) as

$$\mathbb{E}\mathbb{V}_{\psi_1}\Big(\frac{1}{n} \log \mathcal{Z}_t\Big) \leq \frac{1}{n^2} \mathbb{E} \sum_{\mu=1}^n \Big[\Big(\frac{\partial \log \mathcal{Z}_t}{\partial \zeta_\mu^{*(L)}}\Big)^2 + \Big(\frac{\partial \log \mathcal{Z}_t}{\partial Z_\mu}\Big)^2 + \sum_{i=1}^{d_0} \Big(\frac{\partial \log \mathcal{Z}_t}{\partial X_{i\mu}^{(0)}}\Big)^2\Big]. \tag{37}$$

In what follows $P_{out}^x(y \mid x)$ denotes the derivative of the kernel w.r.t. x. A similar definition holds for y, and second derivatives. Also, for brevity we write $P_\mu = P_{out}(Y_{t\mu}|s_{t\mu})$. The derivatives with respect to $\zeta_\mu^{*(L)}$ are easily bounded due to the boundedness of f' and Lemma 11:

$$\Big|\frac{\partial \log \mathcal{Z}_t}{\partial \zeta_\mu^{*(L)}}\Big| \leq \Big|\Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle_t f'(S_{t\mu}; \mathbf{A}_\mu)\sqrt{t\epsilon_L}\Big| \leq \sqrt{t\epsilon_L} C(f)(|Z_\mu|^2 + 1).$$

An analogous formula also holds for the derivatives w.r.t. Z_μ. Since $\mathbb{E}|Z_\mu|^k = O(1)$, we immediately obtain that the first two terms of (37) are of order n^{-1}.

In the last term, the elements $X_{i\mu}^{(0)}$ seem harder to access since they are hidden under the L layers of non-linearity. In order to handle the corresponding derivative let us introduce, for any $1 \leq \ell_1 \leq \ell_2 \leq L$ the $d_{\ell_2} \times d_{\ell_1-1}$ matrix

$$\mathbf{B}_\nu^{*[\ell_1:\ell_2]} := \frac{\mathbf{\Phi}_\nu^{*(\ell_2-1)}\mathbf{W}^{*(\ell_2)}\mathbf{\Phi}_\nu^{*(\ell_2-2)}\mathbf{W}^{*(\ell_2-1)}\cdots\mathbf{\Phi}_\nu^{*(\ell_1-1)}\mathbf{W}^{*(\ell_1)}}{\sqrt{d_{\ell_2-1}\ldots d_{\ell_1-1}}}, \tag{38}$$

where $\mathbf{\Phi}_\nu^{*(\ell)} := \mathrm{diag}(\varphi'(\mathbf{W}^{*(\ell+1)}\mathbf{X}_\nu^{(\ell)}/\sqrt{d_\ell}))$ is a $d_{\ell+1} \times d_{\ell+1}$ diagonal matrix. Then the derivative w.r.t. $X_{i\mu}^{(0)}$ becomes compact:

$$\frac{\partial \log \mathcal{Z}_t}{\partial X_{i\mu}^{(0)}} = \left\langle \frac{P_\mu^y}{P_\mu} \right\rangle f'(S_{t,\mu};\mathbf{A}_\mu)\left(\sqrt{1-t}\frac{(\mathbf{a}^{*\intercal}\mathbf{B}_\mu^{*[1:L]})_i}{\sqrt{d_L}} + \sqrt{t}\rho_L\frac{(\mathbf{v}^{*\intercal}\mathbf{B}_\mu^{*[1:L-1]})_i}{\sqrt{d_{L-1}}}\right)$$

$$+\sqrt{1-t}\left\langle \frac{P_\mu^x}{P_\mu}\frac{(\mathbf{a}^\intercal\mathbf{B}_\mu^{[1:L]})_i}{\sqrt{d_L}}\right\rangle + \sqrt{t}\rho_L\left\langle \frac{P_\mu^x}{P_\mu}\frac{(\mathbf{v}^\intercal\mathbf{B}_\mu^{[1:L-1]})_i}{\sqrt{d_{L-1}}}\right\rangle. \tag{39}$$

If all d_ℓ diverge with the same speed, the operator norm $d_{\ell-1}^{-1/2}\mathbb{E}\|\mathbf{W}^{(*\ell)}\| = O(1)$ is bounded, which leads to $\mathbb{E}\|\mathbf{B}_\nu^{*[\ell_1:\ell_2]}\| = O(1)$ (recall φ is Lipschitz). This, together with the Nishimori identities of Appendix A, allows to bound the last term of (37) with $C(f,\varphi)n^{-1}$.

The second term of (36) is bounded with the standard technique of taking a pn dimensional random vector $\mathbf{A}'$, all elements of which are equal to those of $\mathbf{A} := \{\mathbf{A}_\mu\}_\mu$ but one, which is an i.i.d. copy. Then we can evaluate the variance with respect to $\mathbf{A}$ by bounding $\mathbb{E}_{\psi_1}f_{n,t}(\mathbf{A}) - \mathbb{E}_{\psi_1}f_{n,t}(\mathbf{A}')$ by means of Efron-Stein's inequality recalled in Appendix G.

The third term of (36) is the most challenging one. Denote $g := \mathbb{E}_{\psi_1,\psi_2}f_{n,t}$. Thanks again to the Poincaré-Nash inequality we have

$$\mathbb{E}(g - \mathbb{E}_{\psi_3}g)^2 \leq \sum_{\ell=1}^{L}\sum_{i_\ell,j_\ell}\mathbb{E}\left(\frac{\partial g}{\partial W_{i_\ell j_\ell}^{*(\ell)}}\right)^2 + \sum_i \mathbb{E}\left(\frac{\partial g}{\partial a_i^*}\right)^2 + \sum_j \mathbb{E}\left(\frac{\partial g}{\partial v_j^*}\right)^2. \tag{40}$$

The derivatives w.r.t. a_i^* and v_i^* are treated in the similar way as the derivatives w.r.t. $W_{i_\ell j_\ell}^{*(\ell)}$, so we leave them to Appendix G. To give an idea of the proof, choose $\ell \in \{1,\ldots,L\}$ and write

$$\frac{\partial g}{\partial W_{ij}^{*(\ell)}} = \mathbb{E}_{\psi_1,\psi_2}\left[\left\langle\frac{P_\mu^y}{P_\mu}\right\rangle f'(S_{t\mu};\mathbf{A}_\mu)\left(\frac{\sqrt{1-t}(\mathbf{a}^{*\intercal}\mathbf{B}_\mu^{*[\ell+1:L]}\mathbf{\Phi}_\mu^{*(\ell-1)})_i X_{j\mu}^{(\ell-1)}}{\sqrt{d_L d_{\ell-1}}}\right.\right.$$

$$\left.\left.+ \frac{\sqrt{t}\rho_L(\mathbf{v}^{*\intercal}\mathbf{B}_\mu^{*[\ell+1:L-1]}\mathbf{\Phi}_\mu^{*(\ell-1)})_i X_{j\mu}^{(\ell-1)}}{\sqrt{d_{L-1}d_{\ell-1}}}\right)\right]. \tag{41}$$

We suppressed the μ summation with the normalization n^{-1}, since (ψ_1,ψ_2) contains all the parameters that depend on μ. Therefore the above expression is true for fixed arbitrary μ. The two resulting terms are of similar form, so we focus only on the first one. A naive estimate yields only an $O(1)$, which is not enough. To reach the right order, we need another integration by parts. To do

so we employ a similar trick as the one described in the previous section, i.e., a circular interpolation which replaces $\mathbf{X}_\mu^{(0)}$ by an independent copy $\tilde{\mathbf{X}}_\mu^{(0)}$ of it:

$$\mathbf{X}_\mu^{(0)}(\tau) := \mathbf{X}_\mu^{(0)} \sin \tau + \tilde{\mathbf{X}}_\mu^{(0)} \cos \tau \,, \quad \tau \in [0, \pi/2].$$

Its derivative $\dot{\mathbf{X}}_\mu^{(0)}(\tau)$ w.r.t. τ is independent of $\mathbf{X}_\mu^{(0)}(\tau)$ and $\mathbb{E}[X_{i\nu}^{(0)} \dot{X}_{j\mu}^{(0)}(\tau)] = \delta_{\mu\nu}\delta_{ij}\cos\tau$. Respectively, in a recursive way we define $\mathbf{X}_\mu^{(\ell)}(\tau) := \varphi(\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}(\tau)/\sqrt{d_{\ell-1}})$ and $\tilde{\mathbf{X}}_\mu^{(\ell)} = \varphi(\mathbf{W}^{*(\ell)}\tilde{\mathbf{X}}_\mu^{(\ell-1)}/\sqrt{d_{\ell-1}})$ for $1 \le \ell \le L$. With this we can replace $\mathbf{X}_{j\mu}^{(\ell-1)}$ in (41) by

$$\varphi\Big(\frac{\mathbf{W}_j^{*(\ell-1)}\tilde{\mathbf{X}}_\mu^{(\ell-2)}}{\sqrt{d_{\ell-2}}}\Big) + \int_0^{\pi/2} d\tau \frac{d}{d\tau}\varphi\Big(\frac{\mathbf{W}_j^{*(\ell-1)}\mathbf{X}_\mu^{(\ell-2)}(\tau)}{\sqrt{d_{\ell-2}}}\Big)\,.$$

The first term yields a vanishing contribution, since it is independent of everything else in (41) and its expectation is 0. The second one will give $\sum_{p=1}^{d_0} B_{jp}^{*[1:\ell-1]}(\tau)\dot{X}_{p\mu}^{(0)}(\tau)$, which allows us to integrate by parts w.r.t. $\dot{X}_{p\mu}^{(0)}(\tau)$. However, since in (41) not only $\langle P_{out}^y/P_{out}\rangle$ and $S_{t\mu}$, but also each $\mathbf{\Phi}_\mu^{*(k)}$ (for $k = \ell - 1, \ldots, L - 1$) inside $\mathbf{B}_\mu^{*[\ell+1:L-1]}$, depend on $\dot{X}_{p\mu}^{(0)}(\tau)$, the calculations become quite cumbersome. After integration by parts and some manipulations we end up with a finite number of terms, the majority of which can be bounded easily using the Nishimori identity, Lemma 12, and a bound on the expected spectral norm $\mathbb{E}\|\mathbf{B}^{*[\ell_1:\ell_2]}\|$ (see Appendix G). However, some of the terms will have a more special structure, in particular

$$\mathbb{E}\sum_{i,j} \Big|\frac{\partial g}{\partial W_{ij}^{*(\ell)}}\Big|^2 \le \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2} \mathrm{Tr}\Big(\mathbf{E}^\mathsf{T}\mathrm{diag}(\mathbf{e})\mathbf{G}(\tau)\mathbf{G}(\tau)^\mathsf{T}\mathrm{diag}(\mathbf{e})\mathbf{E}\Big)d\tau + \ldots \,.$$

Here, matrices $\mathbf{E}$ and $\mathbf{G}(\tau)$ are products of several matrices of the type $\mathbf{W}^{*(k)}/\sqrt{d_{k-1}}$, $\mathbf{\Phi}_\mu^{*(k)}(\tau)$ and $\mathbf{\Phi}_\mu^{*(k)}$, while $\mathrm{diag}(\mathbf{e})$ is a diagonal matrix with vector $\mathbf{e} = \mathbf{a}^{*\mathsf{T}}\mathbf{B}_\mu^{*[k_1:k_2]}$ on the diagonal (for some k_1, k_2). As already noted, the operator norms of the mentioned matrices are all order $O(1)$, which makes the operator norms of $\mathbf{E}$ and $\mathbf{G}(\tau)$ also bounded. This allows to evaluate such term using the fact that for a symmetric semidefinite positive matrix $\mathbf{A}$ and any matrix $\mathbf{B}$ we have $|\mathrm{Tr}(\mathbf{A}\mathbf{B})| \le \|\mathbf{B}\|\mathrm{Tr}\mathbf{A}$:

$$\frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2} \mathrm{Tr}\Big(\mathbf{E}^\mathsf{T}\mathrm{diag}(\mathbf{e})\mathbf{G}(\tau)\mathbf{G}(\tau)^\mathsf{T}\mathrm{diag}(\mathbf{e})\mathbf{E}\Big)d\tau$$
$$\le \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2} \|\mathbf{E}\|^2\|\mathbf{G}(\tau)\|^2\mathrm{Tr}\,\mathrm{diag}(\mathbf{e})^2 d\tau$$
$$= \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2} \|\mathbf{E}\|^2\|\mathbf{G}(\tau)\|^2\|\mathbf{e}\|^2 d\tau \le \frac{C(f,\varphi)}{d_{\ell-1}}\,.$$

Hence, as long as all d_ℓ are proportional we have $\mathbb{V}(f_{n,t}) \le C(f,\varphi)(\frac{1}{n} + \frac{1}{d_m})$.

4. Conclusion and perspectives

We have analysed deep neural networks in the Bayes-optimal teacher-student setting. Our main result proves that in the proportional scaling regime where the input dimension d_0, width (d_ℓ) and

number of training samples n all grow large at fixed ratios, from the information-theoretic viewpoint, hidden layers can be iteratively replaced by properly defined effective *linear* layers. Repeating this process until the exhaustion of all inner layers, the outcome is a fundamental equivalence between deep neural networks and generalised *linear* models in this setting. As a consequence, the known rigorous results on generalised linear models directly give explicit formulas for the main information-theoretic quantities of interest: mutual information and optimal generalisation error.

The main conceptual take-home message is that in order to escape this equivalence with linear models, deep neural networks must be analysed beyond the proportional regime. This means that the number of samples n must grow much faster than $d_0, (d_\ell)$. Counting the number of trainable parameters suggests that the non-trivial scaling corresponds to $n = \Theta(\max\{d_\ell\}^2) = \Theta(d_0^2)$ when all widths are proportional and the number of layers is fixed. This regime is of a fundamentally different nature and comes with numerous technical challenges. Indeed, it has been resisting approaches based on statistical physics up to very recently, see Maillard et al. (2024); Xu et al. (2025) for the analysis in the special case of quadratic inner activation function with Gaussian weights, and Barbier et al. (2025) for a generic framework able to tackle any activation function and distribution of weights (which is closely related to Barbier et al. (2024)). This beyond-proportional scaling regime is, from the authors' perspective, an important one to focus on in order to make progress in bridging the gap between theory and practice of modern neural networks.

Acknowledgements

References

R. Aiudi, R. Pacelli, P. Baglioni, A. Vezzani, R. Burioni, and P. Rotondo. Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks. *Nature Communications*, 16(1):568, Jan 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55229-3. URL https://doi.org/10.1038/s41467-024-55229-3.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Q. Al-Dujaili, Ye Duan, Omran Al-Shamma, Jesus Santamaría, Mohammed Abdulraheem Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 2021.

Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2f3c6a4cd8af177f6456e7e51a916ff3-Paper.pdf.

Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 55–73. PMLR, 20–24 Jul 2020. URL https://proceedings.mlr.press/v107/aubin20a.html.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. 18(1), 2017. ISSN 1532-4435.

Carlo Baldassi, Enrico M Malatesta, and Riccardo Zecchina. Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations. *Physical review letters*, 123(17):170602, 2019.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *Probability Theory and Related Fields*, 174, 2019.

Jean Barbier and Dmitry Panchenko. Strong replica symmetry in high-dimensional optimal bayesian inference. *Communications in mathematical physics*, 393(3):1199–1239, 2022.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116.

Jean Barbier, Francesco Camilli, Justin Ko, and Koki Okajima. On the phase diagram of extensive-rank symmetric matrix denoising beyond rotational invariance. *arXiv preprint arXiv:2411.01974*, 2024.

Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation. *arXiv preprint arXiv:2501.18530*, 2025.

Eli Barkai, David Hansel, and Haim Sompolinsky. Broken symmetries in multilayered perceptrons. *Physical Review A*, 45(6):4146, 1992.

Federico Bassetti, Marco Gherardi, Alessandro Ingrosso, Mauro Pastore, and Pietro Rotondo. Feature learning in finite-width bayesian deep linear networks with multiple outputs and convolutional layers, 2024. URL https://arxiv.org/abs/2406.03260.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4132–4179. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/bosch23a.html.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. *Concentration Inequalities*, pages 208–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_9.

Francesco Camilli, Daria Tieplova, and Jean Barbier. Fundamental limits of overparametrized shallow neural networks for supervised learning. 2023. URL https://arxiv.org/abs/2307.05635.

Hong-Bin Chen and Jiaming Xia. Free energy of multi-layer generalized linear models. *Communications in Mathematical Physics*, 400(3):1861–1913, 2023.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. page 3040–3050, 2018.

Hugo Cui, Florent Krzakala, and Lenka Zdeborov. Bayes-optimal learning of deep random networks of extensive-width. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pages 793–815, 1984.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Phys. Rev. A*, 45:7590–7609, 05 1992. doi: 10.1103/PhysRevA.45.7590.

Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1826–1836, Red Hook, NY, USA, 2018. Curran Associates Inc.

E Gardner and B Derrida. Three unfinished works on the optimal storage capacity of networks. 22 (12):1983, 06 1989. doi: 10.1088/0305-4470/22/12/004.

Mario Geiger, Leonardo Petrini, and Matthieu Wyart. Perspective: A phase diagram for deep learning unifying jamming, feature learning and lazy training. *arXiv preprint arXiv:2012.15110*, 2020.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka 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):124010, 2020a.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020b.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Francesco Guerra and Fabio Lucio Toninelli. The Thermodynamic Limit in Mean Field Spin Glass Models. *Communications in Mathematical Physics*, 230, 2002.

Boris Hanin. Random neural networks in the infinite width limit as Gaussian processes. *The Annals of Applied Probability*, 33(6A):4798 – 4819, 2023. doi: 10.1214/23-AAP1933. URL https://doi.org/10.1214/23-AAP1933.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning*, pages 4542–4551. PMLR, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks (invited paper). pages 6–6, 06 2021. doi: 10.1145/3406325.3465355.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *Annals of Statistics*, 48:3619–3642, 12 2020. doi: 10.1214/20-AOS1945.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002, 2020.

M. Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173:859–929, 2017.

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, 09 2021. doi: 10.1103/PhysRevX.11.031059.

Yuanzhi Li and Yingyu Liang. Learning overparametrized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8168–8177, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. doi: 10.5555/3294771.3294828.

Antoine Maillard, Emanuele Troiani, Simon Martin, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *arXiv preprint arXiv:2408.03733*, 2024.

German Mato and Nestor Parga. Generalization properties of multilayered neural networks. *Journal of Physics A: Mathematical and General*, 25(19):5047, 1992.

Alexander G. D. G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

Elizabeth Meckes. Approximation of projections of random vectors. *Journal of Theoretical Probability*, 25:333–352, 2012.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 06 2021. doi: 10.1002/cpa.22008.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Rémi Monasson and Riccardo Zecchina. Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks. *Physical review letters*, 75(12):2432, 1995.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826*, 2020.

Gadi Naveh and Zohar Ringel. A self consistent theory of gaussian processes captures feature learning effects in finite CNNs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vBYwwBxVcsE.

Radford M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY, 1996. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing: an Introduction*. Oxford University Press, Oxford; New York, 2001.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv:1712.05438*, 2017.

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. *Nature Machine Intelligence*, 5(12):1497–1507, 12 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00767-6. URL https://doi.org/10.1038/s42256-023-00767-6.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Galen Reeves. Conditional central limit theorems for gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3045–3049. IEEE, 2017.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75:1889–1935, 09 2022. doi: 10.1002/cpa.22074.

Noa Rubin, Kirsten Fischer, Javed Lindner, David Dahmen, Inbar Seroussi, Zohar Ringel, Michael Krämer, and Moritz Helias. From kernels to features: A multi-scale adaptive theory of feature learning, 2025. URL https://arxiv.org/abs/2502.03210.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 3218–3228, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52: 4225–4243, 10 1995. doi: 10.1103/PhysRevE.52.4225.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (random) features. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Henry Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781, 1993.

Henry Schwarze and John Hertz. Generalization in a large committee machine. *Europhysics Letters*, 20(4):375, 1992.

Henry Schwarze and John Hertz. Generalization in fully connected committee machines. *EPL (Europhysics Letters)*, 21(7):785, 1993.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some CNNs. *Nature Communications*, 14(1):908, Feb 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://doi.org/10.1038/s41467-023-36361-y.

Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of SGD solutions for over-parameterized neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8773–8784. PMLR, 07 2020.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide relu networks. *J. Mach. Learn. Res.*, 23(1), 01 2022. ISSN 1532-4435.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Vladimir Nikolaevich Sudakov. Typical distributions of linear functionals in finite-dimensional spaces of higher dimension. In *Doklady Akademii Nauk*, volume 243, pages 1402–1405. Russian Academy of Sciences, 1978.

Christopher K. I. Williams. Computing with infinite networks. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, NIPS'96, page 295–301, Cambridge, MA, USA, 1996. MIT Press.

Diyuan Wu, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis for heavy ball methods: Dropout-stability, connectivity, and global convergence. *Trans. Mach. Learn. Res.*, 2023, 2022.

Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications, 2025.

Jacob A. Zavatone-Veth, William L. Tong, and Cengiz Pehlevan. Contrasting random and learned features in deep bayesian linear regression. *Physical review. E*, 105 6-1:064118, 2022. URL https://api.semanticscholar.org/CorpusID:247187786.

Appendix A. Nishimori identity

The Nishimori identities are a very general set of symmetries arising in inference in the Bayes-optimal setting as a consequence of Bayes' rule. They were initially discovered in the context of the gauge theory of spin glasses Nishimori (2001), which possess a sub-region of their phase space, called *Nishimori line*, where the most relevant thermodynamic quantities can be exactly computed, and we can generally count on replica symmetry, namely the concentration of order parameters Barbier and Panchenko (2022).

To introduce them, consider a generic inference problem where a Bayes-optimal statistician observes $\mathbf{Y}$ that is a random function of some ground truth signal $\mathbf{X}^*$: $\mathbf{Y} \sim P_{Y|X}(\mathbf{X}^*)$. Then the following holds:

Proposition 5 (Nishimori identity) *For any bounded function f of the signal $\mathbf{X}^*$, the data $\mathbf{Y}$ and of conditionally i.i.d. samples from the posterior $\mathbf{x}^j \sim P_{X|Y}(\cdot \mid \mathbf{Y})$, $j = 1, 2, \ldots, n$, we have that*

$$\mathbb{E}\langle f(\mathbf{Y}, \mathbf{X}^*, \mathbf{x}^2, \ldots, \mathbf{x}^n)\rangle = \mathbb{E}\langle f(\mathbf{Y}, \mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)\rangle \tag{42}$$

where the bracket notation $\langle \cdot \rangle$ is used for the joint expectation over the posterior samples $(\mathbf{x}^j)_{j \leq n}$, $\mathbb{E}$ is over the signal $\mathbf{X}^*$ and data $\mathbf{Y}$.

Proof The proof based on Bayes' rule is elementary, see, e.g., Lelarge and Miolane (2017). ∎

Appendix B. Lipschitz functions of sub-Gaussian random variables

In this section we collect Lemmas for the proof of Theorem 1. The first one hereby illustrates how to use the notion of orthogonality, i.e. independence, between Gaussians to expand nonlinear functions of them. This will be needed to handle the activation functions.

Lemma 6 (Orthogonal approximation) *Let $\psi \in C^2(\mathbb{R})$ be an odd function with bounded first and second derivatives, and $g_1, g_2 \in C^1(\mathbb{R})$ where g_1 is a Lipschitz function, and g_2, g_2' have at most polynomial growth. Let z_1 and z_2 be two centered Gaussian random variables with covariance $\mathbb{E}z_i z_j = C_{ij}$, for $i, j = 1, 2$. Let $\Delta_{12} = C_{12}/C_{22}$. Then, for some constants $L_1, L_2 > 0$*

$$|\mathbb{E}\psi(z_1)g_2(z_2) - C_{12}\mathbb{E}\psi'(z_1)\mathbb{E}g_2'(z_2)| \leq L_1 \Delta_{12}^2, \tag{43}$$

$$|\mathbb{E}g_1(z_1)g_2(z_2) - \mathbb{E}g_1(z_1)\mathbb{E}g_2(z_2)| \leq L_2 \Delta_{12}. \tag{44}$$

Proof Define $z_{1 \perp 2} = z_1 - z_2 C_{12}/C_{22} = z_1 - \Delta_{12} z_2$, which rewrites as $z_1 = z_{1 \perp 2} + \Delta_{12} z_2$. This is nothing but Grahm-Schmidt orthogonalization procedure, used to make $z_{1 \perp 2}$ independent on z_2. Using a Taylor expansion with integral remainder we have

$$\mathbb{E}\psi(z_1)g_2(z_2) - \Delta_{12}\mathbb{E}\psi'(z_{1 \perp 2})\mathbb{E}z_2 g_2(z_2) = \Delta_{12}^2 \int_0^1 (1-s)\mathbb{E}\psi''(z_{1 \perp 2} + s\Delta_{12}z_2)z_2^2 g_2(z_2) ds$$

where we have used oddity of ψ to eliminate the zero-th order of the expansion. Now, using the fact that ψ has second derivative uniformly bounded by a constant K, we get

$$|\mathbb{E}\psi''(z_{1 \perp 2} + s\Delta_{12}z_2)z_2^2 g_2(z_2)| \leq K\mathbb{E}z_2^2|g_2(z_2)| \leq \bar{K} \tag{45}$$

thanks to the at most polynomial growth of g_2, for a proper constant $\bar{K}$. Then we reuse the same argument for $\mathbb{E}\psi'(z_{1 \perp 2})$:

$$|\mathbb{E}\psi'(z_{1 \perp 2}) - \mathbb{E}\psi'(z_1)| \leq \Delta_{12} \int_0^1 \mathbb{E}|\psi''(z_1 - sz_2\Delta_{12})z_2| ds \leq \tilde{K}\Delta_{12}, \tag{46}$$

where we use that the boundedness of the derivatives of ψ, for a proper constant $\tilde{K}$. Together with the integration by parts $\mathbb{E}z_2 g_2(z_2) = C_{22}\mathbb{E}g_2'(z_2)$, this leads directly to (43).

For (44) instead, we have

$$|\mathbb{E}g_1(z_1)g_2(z_2) - \mathbb{E}g_1(z_{1 \perp 2})\mathbb{E}g_2(z_2)| \leq \Delta_{12} \int_0^1 \mathbb{E}|g_1'(z_{1 \perp 2} + s\Delta_{12}z_2)z_2 g_2(z_2)| ds. \tag{47}$$

In order to bound the remainder it suffices to use the fact that g_1 is Lipschitz, and the at most polynomial growth of g_2, g_2'. Then, to replace $z_{1 \perp 2}$ by z_1 in g_1 one uses the same strategy backward, producing the same order of remainder. ∎

Following this, we need to make sure that the "quasi-orthogonality" condition of the inputs, namely the exponential concentration property

$$\mathbb{P}\left(\left|\frac{\mathbf{X}_\mu^{(0)} \cdot \mathbf{X}_\nu^{(0)}}{d_0} - \delta_{\mu\nu}\right| \geq \epsilon\right) \leq C\exp\left(-c\epsilon^2\right) \tag{48}$$

for any $\epsilon > 0$ and some $C, c > 0$, propagates to all the post activations in some form. This is the aim of the next Lemmas, which are stated for a general setting.

Lemma 7 (Concentration propagation) *Consider two deterministic vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ with Euclidean norms σ_X and σ_Y respectively, and $\mathbf{W} \in \mathbb{R}^{p \times d}$ a random matrix with i.i.d. centred sub-Gaussian entries. Consider two Lipschitz functions $\phi, \psi \in C^1(\mathbb{R})$, applied element-wise to vectors. Then, for all $\epsilon > 0$*

$$\mathbb{P}\Big(\Big|\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} - \mathbb{E}\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p}\Big| \geq \epsilon\Big) \leq 2\exp\Big[-\frac{p\epsilon^2}{2L\sigma_X\sigma_Y(\epsilon + L\sigma_X\sigma_Y)}\Big] \quad (49)$$

for some $L > 0$ depending on ϕ, ψ.

Proof Define for brevity $\alpha = \mathbf{W}_1 \cdot \mathbf{X}$, $\beta = \mathbf{W}_1 \cdot \mathbf{Y}$. We start by considering one side of the bound, as the other one follows from the same arguments. Using exponential Markov's inequality we get

$$\mathbb{P}\Big(\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} - \mathbb{E}\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} \geq \epsilon\Big) \leq e^{-s(\epsilon + \mathbb{E}\phi(\alpha)\psi(\beta))}\Big[\mathbb{E}e^{s\frac{\phi(\alpha)\psi(\beta)}{p}}\Big]^p \quad (50)$$

for any $s > 0$. Let us introduce the generating function

$$g_p(s) := p\log\mathbb{E}e^{s\frac{\phi(\alpha)\psi(\beta)}{p}} \leq p\Big(\mathbb{E}e^{s\frac{\phi(\alpha)\psi(\beta)}{p}} - 1\Big)$$

$$= s\mathbb{E}\phi(\alpha)\psi(\beta) + p\sum_{k=2}^{\infty}\frac{s^k}{k!p^k}\mathbb{E}\phi(\alpha)^k\psi(\beta)^k\,. \quad (51)$$

Now we can use the fact that ϕ, ψ are Lipschitz functions applied to sub-Gaussian random variables. Hence $\phi(\alpha), \psi(\beta)$ are sub-Gaussian random variables. As a consequence, there exists a L such that

$$g_p(s) - s\mathbb{E}\phi(\alpha)\psi(\beta) \leq \sum_{k=2}^{\infty}\frac{s^k(L\sigma_X\sigma_Y)^k}{2p^{k-1}} = \frac{(sL\sigma_X\sigma_Y)^2}{2p}\frac{1}{1 - \frac{sL\sigma_X\sigma_Y}{p}}\,, \quad (52)$$

which restricts us to choose $s < p/(L\sigma_X\sigma_Y)$. Therefore, our bound on the probability becomes

$$\mathbb{P}\Big(\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} - \mathbb{E}\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} \geq \epsilon\Big) \leq e^{-s\epsilon + \frac{(sL\sigma_X\sigma_Y)^2}{2p}\frac{1}{1 - \frac{sL\sigma_X\sigma_Y}{p}}}\,. \quad (53)$$

The result then follows from the choice $s = p\epsilon/(L\sigma_X\sigma_Y(\epsilon + L\sigma_X\sigma_Y))$. ∎

Lemma 8 (Moments bounds) *Let k be a positive integer, and $\mathbf{W}, \phi, \psi, \mathbf{X}, \mathbf{Y}$ as in Lemma 7. Then*

$$\mathbb{E}\Big|\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p} - \mathbb{E}\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p}\Big|^k \leq \frac{C_k(\sigma_X\sigma_Y)^k}{p^{k/2}}\,, \quad (54)$$

for a positive constant $C_k > 0$ depending on ϕ, ψ. Furthermore, if $\psi \in C^2(\mathbb{R})$ is an odd function with bounded first and second derivatives, then

$$\mathbb{E}\Big|\frac{\phi(\mathbf{WX})^\intercal \psi(\mathbf{WY})}{p}\Big|^k \leq \frac{C_k'(\sigma_X\sigma_Y)^k}{p^{k/2}} + L_k|\mathbf{X}^\intercal\mathbf{Y}|^k + L_k'\Big(\frac{|\mathbf{X}^\intercal\mathbf{Y}|}{\|\mathbf{Y}\|^2}\Big)^{2k} \quad (55)$$

for some positive constants $C_k', L_k, L_k' > 0$ depending on ϕ, ψ.

Proof The first inequality follows from the tail integration. Define the shortcut notation $\phi_{\mathbf{X}} = \phi(\mathbf{WX})$, and similarly for $\psi_{\mathbf{X}}$. Then

$$\mathbb{E}\left|\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p} - \mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right|^k = k\int_0^\infty dt\, t^{k-1}\mathbb{P}\left(\left|\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p} - \mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right| \geq t\right)$$

$$\leq 2k\int_0^\infty dt\, t^{k-1}e^{-\frac{pt^2}{2L\sigma_X\sigma_Y(t+L\sigma_X\sigma_Y)}}. \tag{56}$$

We split the integral into two contributions and we bound them separately. The first one is

$$\int_0^{L\sigma_X\sigma_Y} dt\, t^{k-1}e^{-\frac{pt^2}{2L\sigma_X\sigma_Y(t+L\sigma_X\sigma_Y)}} \leq \int_0^{L\sigma_X\sigma_Y} dt\, t^{k-1}e^{-\frac{pt^2}{4(L\sigma_X\sigma_Y)^2}}. \tag{57}$$

It suffices then to extend the integral to ∞ (the integrand is positive), and to perform the change of variables $t \mapsto tL\sigma_X\sigma_Y/\sqrt{p}$ to show that

$$\int_0^{L\sigma_X\sigma_Y} dt\, t^{k-1}e^{-\frac{pt^2}{2L\sigma_X\sigma_Y(t+L\sigma_X\sigma_Y)}} \leq \tilde{C}_k\frac{(\sigma_X\sigma_Y)^k}{p^{k/2}}. \tag{58}$$

The second piece of the integral is instead

$$\int_{L\sigma_X\sigma_Y}^\infty dt\, t^{k-1}e^{-\frac{pt^2}{2L\sigma_X\sigma_Y(t+L\sigma_X\sigma_Y)}} \leq \int_{L\sigma_X\sigma_Y}^\infty dt\, t^{k-1}e^{-\frac{pt}{4L\sigma_X\sigma_Y}} \leq \hat{C}_k\frac{(\sigma_X\sigma_Y)^k}{p^k}. \tag{59}$$

One then just sets $C_k = 2k(\tilde{C}_k + \hat{C}_k)$ and (54) is proved.

Equation (55) can be derived from (54) and Lemma 6 as follows. Firstly, by triangular and Jensen's inequality respectively we get:

$$\mathbb{E}\left|\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right|^k \leq \mathbb{E}\left(\left|\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p} - \mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right| + \left|\mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right|\right)^k$$

$$\leq 2^{k-1}\mathbb{E}\left|\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p} - \mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right|^k + 2^{k-1}\left|\mathbb{E}\frac{\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}}{p}\right|^k. \tag{60}$$

Secondly, the first term on the r.h.s. of the above can be bounded using (54). Thirdly, Lemma 6 on $\mathbb{E}\phi_{\mathbf{X}}^{\mathsf{T}}\psi_{\mathbf{Y}}/p$ readily yields the statement. ∎

Appendix C. Bounds for post-activations moments

Note that Lemmas 7 and 8 hold for any couple of vectors $\mathbf{X}, \mathbf{Y}$. In particular, (55) becomes pointless when $\mathbf{X} = \mathbf{Y}$, because the bound does not vanish. On the contrary, in that case equation (54) gives us the concentration of the norms of the post activations $\varphi(\mathbf{WX})$ when choosing $\phi = \psi = \varphi$. When $\mathbf{X}$ and $\mathbf{Y}$ are instead quasi-orthogonal, (55) is the reflection in terms of moments of the propagation of such quasi-orthogonality property to the post-activations $\varphi(\mathbf{WX}), \varphi(\mathbf{WY})$.

As a consequence of these Lemmas one has the desired moment control for all post-activations:

Lemma 9 (Post-activations moments) *Define the sequence of coefficients:*

$$\sigma_0 = 1 \,, \sigma_\ell = \mathbb{E}\varphi^2(Z\sqrt{\sigma_{\ell-1}}) \,, \quad \rho_\ell = \mathbb{E}\varphi'(Z\sqrt{\sigma_{\ell-1}}) \tag{61}$$

where $\ell \in [L]$ and $Z \sim \mathcal{N}(0,1)$. Let k be a positive integer and $g(x) = \varphi^2(x), \varphi'(x)$ or $x\varphi(x)$. Then

$$\mathbb{E}\Big|\frac{1}{d_\ell}\sum_{i=1}^{d_\ell} g\Big(\frac{\mathbf{W}_i^{*(\ell)} \cdot \mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big) - \mathbb{E}g(Z\sqrt{\sigma_{\ell-1}})\Big|^k \leq \frac{C_{k,\varphi}}{d_{[0:\ell]}^{k/2}} \,, \tag{62}$$

for a positive constant $C_{k,\varphi} > 0$, where $d_{[0:\ell]} = \min\{(d_s)_{s=0}^\ell\}$.

Proof The key proof is the one for $g = \varphi^2$. In that case the statement reads:

$$\mathbb{E}\Big|\frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d_\ell} - \sigma_\ell\Big|^k \leq \frac{C_{k,\varphi}}{d_{[0:\ell]}^{k/2}} \,. \tag{63}$$

We proceed by induction. For $\ell = 0$ the statement is surely true. Now assume it holds for $\ell - 1$. The l.h.s. of (63) can be bounded via triangular and Jensen inequalities as

$$... \leq 2^{k-1}\mathbb{E}\Big|\frac{\|\mathbf{X}_\mu^{(\ell)}\|^2}{d_\ell} - \mathbb{E}_Z\varphi^2\Big(Z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}}}\Big)\Big|^k$$
$$+ 2^{k-1}\mathbb{E}\Big|\mathbb{E}_Z\varphi^2\Big(Z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}}}\Big) - \mathbb{E}_Z\varphi^2(Z\sqrt{\sigma_{\ell-1}})\Big|^k \,. \tag{64}$$

The first of the two terms can be bounded simply using (54) for $\mathbf{X} = \mathbf{Y}$, $\phi = \psi = \varphi$. Let us focus on the second. Define

$$R := \mathbb{E}_Z g\Big(Z\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}}}\Big) - \mathbb{E}_Z g(Z\sqrt{\sigma_{\ell-1}}) = \mathbb{E}_Z \int_0^1 ds\, g'(Z(s))Z'(s) \tag{65}$$

where

$$Z(s) := Z\Big(s\sqrt{\frac{\|\mathbf{X}^{(\ell-1)}\|^2}{d_{\ell-1}}} + (1-s)\sqrt{\sigma_{\ell-1}}\Big).$$

Since φ, φ' are both Lipschitz there exists a constant $\bar{K}_\varphi$ such that $|g'(z)| \leq \bar{K}_\varphi|z|$. Hence

$$|R| \leq \bar{K}_\varphi\Big|\sqrt{\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}}} - \sqrt{\sigma_{\ell-1}}\Big|\,\mathbb{E}_Z|Z|^2\int_0^1 ds\,\Big(s\sqrt{\frac{\|\mathbf{X}^{(\ell-1)}\|^2}{d_{\ell-1}}} + (1-s)\sqrt{\sigma_{\ell-1}}\Big)$$
$$= \frac{1}{2}\bar{K}_\varphi\Big|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}} - \sigma_{\ell-1}\Big| \,, \tag{66}$$

and therefore $\mathbb{E}|R|^k \leq C'_{\varphi,k}d_{[0:\ell-1]}^{-k/2}$ for a positive constant $C'_{\varphi,k}$ by the inductive hypothesis.

Given that the statement is true for $g = \varphi^2$, proving it for $g = \varphi'$ follows exactly the same steps, using (54) with $\phi = \varphi'$ and $\psi = 1$. Analogously for $g(x) = x\varphi(x)$, (54) can be used with $\phi = \varphi$ and $\psi = Id$ (recall that the first and second derivative of φ are bounded). ∎

Finally, the quasi orthogonality of the inputs, i.e.

$$\mathbb{E}\Big|\frac{\mathbf{X}_\mu^{(0)} \cdot \mathbf{X}_\nu^{(0)}}{d_0}\Big|^k \leq \frac{C_k}{d_0^{k/2}} \tag{67}$$

is shown to propagate to subsequent layers:

Lemma 10 (Quasi-orthogonality propagation) *For all $\ell = 0, \ldots, L$, and a positive integer k there exists positive a constant C_k such that*

$$\mathbb{E}\Big|\frac{\mathbf{X}_\mu^{(\ell)} \cdot \mathbf{X}_\nu^{(\ell)}}{d_\ell}\Big|^k \leq \frac{C_k}{d_{[0:\ell]}^{k/2}}\,, \tag{68}$$

for all $\mu \neq \nu$, where $d_{[0:\ell]} = \min\{(d_s)_{s=0}^\ell\}$.

Proof The proof proceeds by induction. The statement is certainly true for $\ell = 0$.

Assume now it holds for $\ell - 1$, and let us rewrite the l.h.s. of the inequality as

$$\mathbb{E}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\Big|\frac{1}{d_\ell}\varphi\Big(\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big) \cdot \varphi\Big(\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\nu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big)\Big|^k. \tag{69}$$

Using the fact that φ is Lipschitz, and Cauchy-Schwartz inequality, it is clear that the above object is bounded by a constant. Furthermore, from Lemma 9 and Markov's inequality we know that

$$\mathbb{P}\Big(\Big|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}} - \sigma_{\ell-1}\Big| \geq \delta\Big) \leq \frac{1}{\delta^{2k}}\mathbb{E}\Big|\frac{\|\mathbf{X}_\mu^{(\ell-1)}\|^2}{d_{\ell-1}} - \sigma_{\ell-1}\Big|^{2k} \leq \frac{C_{2k,\varphi}}{d_{[0:\ell-1]}^k\delta^{2k}} \tag{70}$$

for any $\delta > 0$. If we call the event measured above $B_{\mu,\delta}$, then there exist a positive constant C_k' such that

$$\mathbb{E}\mathbb{E}_{\mathbf{W}^{*(\ell)}}\Big|\frac{1}{d_\ell}\varphi\Big(\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big) \cdot \varphi\Big(\frac{\mathbf{W}^{*(\ell)}\mathbf{X}_\nu^{(\ell-1)}}{\sqrt{d_{\ell-1}}}\Big)\Big|^k \mathbb{1}(B_{\mu,\delta})\mathbb{1}(B_{\nu,\delta}) \leq \frac{C_k'}{d_{[0:\ell-1]}^{k/2}\delta^k}\,. \tag{71}$$

Hence, up to an error we can afford, we can always use the fact that $d_{\ell-1}(\sigma_{\ell-1} - \delta) \leq \|\mathbf{X}_\mu^{(\ell-1)}\|^2 \leq d_{\ell-1}(\sigma_{\ell-1} + \delta)$. With this observation, using (55) we can bound (69) with

$$\ldots \leq \frac{C_k'}{d_\ell^{k/2}}(\sigma_{\ell-1} + \delta)^{2k} + L_k\mathbb{E}\Big|\frac{\mathbf{X}_\mu^{(\ell-1)} \cdot \mathbf{X}_\nu^{(\ell-1)}}{d_{\ell-1}}\Big|^k + L_k'\mathbb{E}\Big(\frac{\mathbf{X}_\mu^{(\ell-1)} \cdot \mathbf{X}_\nu^{(\ell-1)}}{d_{\ell-1}(\sigma_{\ell-1} - \delta)}\Big)^{2k}. \tag{72}$$

To conclude the proof it suffices to choose $\delta = \sigma_{\ell-1}/2$, and to use the inductive hypothesis. ∎

Appendix D. Properties of the output kernel

Recall the definition (5) of the output kernel P_{out}. Based on this we are able to prove certain properties about $u_{Y_{t\mu}}(S_{t\mu})$ and its derivatives that are inherited from those of the readout function f. For this section $P_{out}^x(y \mid x)$ denotes the derivative of the kernel w.r.t. x. A similar definition holds for y, and second derivatives.

We start with a computation of the derivatives of P_{out}:

Lemma 11 *Let $y = Y_{t\mu} = f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta}Z_\mu$. Let $f(\cdot; \mathbf{A})$ be a $C^2(\mathbb{R})$ with bounded first and second derivatives. Then there exists constant $C(f)$ such that*

$$\max\left\{\left|\frac{P_{out}^y(y|x)}{P_{out}(y|x)}\right|, \left|\frac{P_{out}^x(y|x)}{P_{out}(y|x)}\right|, \left|\frac{P_{out}^{yy}(y|x)}{P_{out}(y|x)}\right|, \left|\frac{P_{out}^{yx}(y|x)}{P_{out}(y|x)}\right|, \left|\frac{P_{out}^{xx}(y|x)}{P_{out}(y|x)}\right|\right\} < C(f)(|Z_\mu|^2 + 1).$$

(73)

Proof For the proof we refer to (Camilli et al., 2023, Lemma 17). ∎

Lemma 12 (Properties of P_{out}) *Recall the definition $u_y(x) := \log P_{out}(y \mid x)$. We denote $u_y' \equiv u_y'(x) := \partial_x u_y(x)$. Furthermore, let*

$$U_{\mu\nu} := \delta_{\mu\nu} u''_{Y_{t\mu}}(S_{t\mu}) + u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\nu}}(S_{t\nu}),$$ (74)

$$U_{\mu\nu\eta} := u'''_{Y_{t\mu}}(S_{t\mu})\delta_{\mu\nu}\delta_{\nu\eta} + u''_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\nu}}(S_{t\nu})\delta_{\mu\eta}$$ (75)
$$+ u'_{Y_{t\mu}}(S_{t\mu}) u''_{Y_{t\nu}}(S_{t\nu})\delta_{\nu\eta} + u'_{Y_{t\mu}}(S_{t\mu}) u'_{Y_{t\nu}}(S_{t\nu}) u'_{Y_{t\eta}}(S_{t\eta}).$$

If $f(\cdot; \mathbf{A})$ is a $C^2(\mathbb{R})$ with P_A-almost surely bounded first and second derivatives then

$$\mathbb{E}[u'_{Y_{t\mu}}(S_{t\mu}) \mid S_{t\mu}] = \mathbb{E}[U_{\mu\nu} \mid S_{t\mu}, S_{t\nu}] = 0, \ \mathbb{E}[U_{\mu\nu\eta} \mid S_{t\mu}, S_{t\nu}, S_{t\eta}] = 0 \ \textit{for } \mu \neq \nu,$$

$$\mathbb{E}[(u'_{Y_{t\mu}}(S_{t\mu}))^2 \mid S_{t\mu}], \ \mathbb{E}[U_{\mu\nu}^2 \mid S_{t\mu}, S_{t\nu}] \leq C(f), \ \mathbb{E}[U_{\mu\nu\eta}^2 \mid S_{t\mu}, S_{t\nu}, S_{t\eta}] \leq C'(f) \ \textit{for } \mu \neq \nu$$

for some positive constants $C(f), C'(f)$.

Proof By definition one has $\mathbb{E}[u'_{Y_{t\mu}}(S_{t\mu}) \mid S_{t\mu}] = \int dy P_{out}^x(y \mid S_{t\mu}) = 0$. For $U_{\mu\nu}$ instead, we first need to realize that $U_{\mu\mu} = P_{out}^{xx}(Y_{t\mu} \mid S_{t\mu})/P_{out}(Y_{t\mu} \mid S_{t\mu})$ which implies $\mathbb{E}[U_{\mu\mu} \mid S_{t\mu}] = \int dy P_{out}^{xx}(y \mid S_{t\mu}) = 0$. Concerning the off-diagonal instead, conditionally on $S_{t\mu}, S_{t\nu}$ the remaining disorder in the Y's is independent, so for $\mu \neq \nu$ we have $\mathbb{E}[U_{\mu\nu} \mid S_{t\mu}, S_{t\nu}] = \int dy P_{out}^x(y \mid S_{t\mu}) \int dy' P_{out}^x(y' \mid S_{t\nu}) = 0$. Similar arguments hold for $U_{\mu\nu\eta}$ when $\mu \neq \nu$.

For the boundedness of the derivatives it is sufficient to notice that

$$(u'_{Y_{t\mu}}(S_{t\mu}))^2 = \left(\frac{P_{out}^x(Y_{t\mu}|S_{t\mu})}{P_{out}(Y_{t\mu}|S_{t\mu})}\right)^2 \leq C(f)(|Z_\mu|^4 + 1),$$ (76)

the last inequality being true due to Lemma 11 and the fact that $Y_{t\mu} = f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta}Z_\mu$. Now it is immediate that after taking the expectation conditioned on $S_{t\mu}$ we obtain a bound $C(f)$.

In order to deal with $U_{\mu\nu}^2$ we rewrite

$$U_{\mu\nu}^2 = \left(\delta_{\mu\nu}\left(\frac{P_{out}^{xx}(Y_{t\mu}|S_{t\mu})}{P_{out}(Y_{t\mu}|S_{t\mu})} - \left(\frac{P_{out}^x(Y_{t\mu}|S_{t\mu})}{P_{out}(Y_{t\mu}|S_{t\mu})}\right)^2\right) + \frac{P_{out}^x(Y_{t\mu}|S_{t\mu})}{P_{out}(Y_{t\mu}|S_{t\mu})}\frac{P_{out}^x(Y_{t\nu}|S_{t\nu})}{P_{out}(Y_{t\nu}|S_{t\nu})}\right)^2.$$

With the help of Lemma 11 one can see immediately that $U_{\mu\nu}^2 \leq C(f)P(Z_{t\mu}, Z_{t\nu})$, where P is some polynomial with even degrees. Once again, after taking expectation we get a bound by a positive constant $C(f)$. Similar arguments hold for $\mathbb{E}[U_{\mu\nu\eta}^2 \mid S_{t\mu}, S_{t\nu}, S_{t\eta}]$. ∎

Appendix E. Details for the proof of Theorem 1

For the sake of clarity the proof has been divided in lemmas.

Lemma 13 B *in* (28) *is identically* 0.

Proof This is just a consequence of the Nishimori identities in Appendix A. In fact,

$$
B := \frac{1}{n}\mathbb{E}_{(t)}\Big\langle \sum_{\mu=1}^{n} u'_{Y_{t\mu}}(s_{t\mu})\frac{ds_{t\mu}}{dt}\Big\rangle_t = \frac{1}{n}\mathbb{E}_{(t)}\sum_{\mu=1}^{n} u'_{Y_{t\mu}}(S_{t\mu})\frac{dS_{t\mu}}{dt}
$$

$$
= \frac{1}{n}\sum_{\mu=1}^{n}\mathbb{E}\Big[\mathbb{E}_{(t)}[u'_{Y_{t\mu}}(S_{t\mu}) \mid S_{t\mu}]\frac{dS_{t\mu}}{dt}\Big] = 0\,, \qquad (77)
$$

where we have used Lemma 12. ∎

Then we turn to the term A_2 in (26).

Lemma 14 $A_2 = O\Big(\Big(1+\frac{n}{d_m}\Big)\frac{1}{\sqrt{d_m}}\Big).$

Proof In the square parenthesis in (26) we integrate by parts $\mathbf{v}^*$ and $\mathbf{W}^{*(L)}$, obtaining

$$
A_2 = \frac{1}{2}\mathbb{E}_{(t)}\Big(\frac{1}{n}\log\mathcal{Z}_t - \bar{f}_{n,t}\Big)\sum_{\mu,\nu=1}^{n}U_{\mu\nu}\rho_L\Big[\rho_L - \mathbf{a}^{*\intercal}\frac{\mathbf{a}^* \circ \varphi'(\boldsymbol{\alpha}_\mu)}{d_L}\Big]\frac{\mathbf{X}_\mu^{(L-1)}\cdot\mathbf{X}_\nu^{(L-1)}}{d_{L-1}} \qquad (78)
$$

where we used the notation $\boldsymbol{\alpha}_\mu = \mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}/\sqrt{d_{L-1}}$.

Using Cauchy-Schwartz inequality we can bound $|A_2|$ as in

$$
A_2^2 \le \frac{\rho_L^2}{4}\mathbb{V}_{(t)}\Big[\frac{1}{n}\log\mathcal{Z}_t\Big]\mathbb{E}_{(t)}\sum_{\mu,\nu=1}^{n}\sum_{\eta,\lambda=1}^{n}U_{\mu\nu}U_{\eta\lambda}\frac{\mathbf{X}_\mu^{(L-1)}\cdot\mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\frac{\mathbf{X}_\eta^{(L-1)}\cdot\mathbf{X}_\lambda^{(L-1)}}{d_{L-1}}
$$

$$
\times\Big(\rho_L - \mathbf{a}^{*\intercal}\frac{\mathbf{a}^* \circ \varphi'(\boldsymbol{\alpha}_\mu)}{d_L}\Big)\Big(\rho_L - \mathbf{a}^{*\intercal}\frac{\mathbf{a}^* \circ \varphi'(\boldsymbol{\alpha}_\eta)}{d_L}\Big). \qquad (79)
$$

Recall that φ is Lipschitz, so ρ_L can be bounded by a constant. In addition, by Theorem 2, the variance of the random free entropy contributes with $O(1/n + 1/d_m)$ where $d_m = \min\{(d_k)_{k=0}^{L}\}$. In the summation over the indices μ,ν,η,λ not all terms survive thanks to Lemma 12. The possibilities are $\mu = \nu$ and $\eta = \lambda$, or $\mu < \nu$ and $\eta < \lambda$ with $\mu = \eta$ and $\nu = \lambda$. Another possibility is $\mu = \nu = \eta = \lambda$, which is easily seen to be suppressed for combinatorial reasons. Also, recall that $\mathbb{E}_{(t)}[U_{\mu\nu}^2 \mid S_{t\mu}, S_{t\nu}] \le C(f)$ as proved in Lemma 12.

Therefore, there exists a constant C such that

$$
A_2^2 \le C\Big(\frac{1}{n} + \frac{1}{d_m}\Big)\mathbb{E}_{(t)}\sum_{\mu,\nu=1}^{n}\Big(\frac{\mathbf{X}_\mu^{(L-1)}\cdot\mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^2\Big(\rho_L - \mathbf{a}^{*\intercal}\frac{\mathbf{a}^* \circ \varphi'(\boldsymbol{\alpha}_\mu)}{d_L}\Big)^2. \qquad (80)
$$

27

Notice that the only dependency on $\mathbf{W}^{*(L)}$ is in $\boldsymbol{\alpha}_\mu$. Therefore, considering a brand new set of variables $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ we can recast the above as

$$A_2^2 \leq C\Big(\frac{1}{n} + \frac{1}{d_m}\Big) \sum_{\mu,\nu=1}^{n} \mathbb{E}_{(t)}\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^2 \mathbb{E}_{\mathbf{Z},\mathbf{a}^*}\Big(\rho_L - \sum_{i=1}^{d_L} \frac{(a_i^*)^2}{d_L} \varphi'\Big(Z_i\sqrt{\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d_{L-1}}}\Big)\Big)^2. \tag{81}$$

With an exponentially small cost $O(e^{-cd_{L-2}\delta})$, for a given δ and some constant $c > 0$ thanks to Lemma 7, one can assume that $\|\mathbf{X}_\mu^{(\ell)}\|^2/d_\ell \geq \mathbb{E}\|\mathbf{X}_\mu^{(\ell)}\|^2/d_\ell - \delta$, for $\ell = 0, \ldots, L-1$. Then, using Lemma 10

$$\mathbb{E}\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^4 \leq \frac{K}{d_{[0:d_{L-1}]}^2} \tag{82}$$

for an appropriate constant $K > 0$. Furthermore, it is easy to check that replacing the $(a_i^*)^2$'s by 1 costs an overall $O(d_L^{-1})$. Also,

$$\mathbb{E}\Big(\rho_L - \frac{1}{d_L} \sum_{i=1}^{d_L} \varphi'\Big(Z_i\sqrt{\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d_{L-1}}}\Big)\Big)^4 \leq \frac{C}{d_{[0:d_{L-1}]}^2} \tag{83}$$

by Lemma 9. Then, using Cauchy-Schwartz inequality one readily gets:

$$A_2^2 = O\Big(\Big(1 + \frac{n}{d_m}\Big)^2 \frac{1}{d_m}\Big). \tag{84}$$

∎

Lemma 15 *Recall the definition of A_1^{off} in (33). $A_1^{off} = O\big((1 + n/d_m)\sqrt{n}/d_m\big)$.*

Proof Let us take (34) as a starting point. Define the two contributions:

$$A_{11}^{off}(\tau) = \mathbb{E}_{(t)}\Big(\frac{1}{n}\log\mathcal{Z}_t - \bar{f}_{n,t}\Big) \sum_{\mu<\nu,1}^{n} \sum_{\eta=1}^{n} U_{\mu\nu\eta} \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\eta^{(L-1)}}{d_{L-1}} \times$$
$$\times \sum_{i=1}^{d_L} \frac{a_i^* \varphi'(\alpha_{i\eta})}{\sqrt{d_L}} \frac{\varphi'(\alpha_{i\mu}(\tau)) - \rho_L}{d_L} \varphi(\alpha_{i\nu}) \tag{85}$$

and

$$A_{12}^{off}(\tau) = \mathbb{E}_{(t)}\Big(\frac{1}{n}\log\mathcal{Z}_t - \bar{f}_{n,t}\Big) \sum_{\mu<\nu,1}^{n} U_{\mu\nu} \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}} \sum_{i=1}^{d_L} \frac{\varphi'(\alpha_{i\mu}(\tau)) - \rho_L}{d_L} \varphi'(\alpha_{i\nu}). \tag{86}$$

Then $A_1^{off} = \int_0^{\pi/2} \cos\tau\big(A_{11}^{off}(\tau) + A_{12}^{off}(\tau)\big)d\tau$.

Let us start from $A_{11}^{off}(\tau)$. Using Cauchy-Schwartz's inequality, and $\mathbb{E}[U_{\mu\nu\eta} \mid S_{t\mu}, S_{t\nu}, S_{t\eta}] = 0$ together with $\mathbb{E}[U_{\mu\nu\eta}^2 \mid S_{t\mu}, S_{t\nu}, S_{t\eta}] \leq C'(f)$, one can bound $(A_{11}^{off}(\tau))^2$ as

$$(A_{11}^{off}(\tau))^2 \leq C''(f)\mathbb{V}_{(t)}\Big[\frac{1}{n}\log \mathcal{Z}_t\Big] \sum_{\mu<\nu,1}^{n}\sum_{\eta=1}^{n}\mathbb{E}_{(t)}\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^2$$

$$\times \frac{1}{d_L^2}\sum_{i=1}^{d_L}\frac{(\varphi'(\alpha_{i\mu}(\tau)) - \rho_L)^2}{d_L}(X_{i\nu}^{(L)})^2(\varphi'(\alpha_{i\eta}))^2 \qquad (87)$$

where we have computed the expectation over the $\mathbf{a}^*$'s. Since φ is Lipschitz, we can easily continue the bound as in:

$$(A_{11}^{off}(\tau))^2 \leq C'''(f)\mathbb{V}_{(t)}\Big[\frac{1}{n}\log \mathcal{Z}_t\Big] \sum_{\mu<\nu,1}^{n}\sum_{\eta=1}^{n}\mathbb{E}_{(t)}\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^2 \frac{\|\mathbf{X}_\nu^{(L)}\|^2}{d_L^3}. \qquad (88)$$

By splitting the last expectation again by Cauchy-Schwartz and Lemmas 9 and 10 we have

$$(A_{11}^{off}(\tau))^2 = O\Big(\Big(1 + \frac{n}{d_m}\Big)^2\frac{n}{d_m^2}\Big) \qquad (89)$$

where we have taken into account possible matchings $\eta = \mu$ or $\eta = \nu$. The above bound is uniform in τ.

Analogous arguments apply to $A_{12}^{off}(\tau)$):

$$(A_{12}^{off}(\tau))^2 = C(f)\mathbb{V}_{(t)}\Big[\frac{1}{n}\log \mathcal{Z}_t\Big] \sum_{\mu<\nu,1}^{n}\mathbb{E}_{(t)}\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big)^2(\mathbb{1}(B_L) + \mathbb{1}(B_L^c))$$

$$\times \sum_{i,j=1}^{d_L}\frac{(\varphi'(\alpha_{i\mu}(\tau)) - \rho_L)\varphi'(\alpha_{i\nu})}{d_L}\frac{(\varphi'(\alpha_{j\mu}(\tau)) - \rho_L)\varphi'(\alpha_{j\nu})}{d_L}. \qquad (90)$$

In the above we have introduced the event

$$B_L := \{\|\mathbf{X}_\mu^{(L-1)}\|^2 \geq d_{L-1}\sigma_{L-1}/2\} \qquad (91)$$

that, thanks to (70) with $k = 2$ and $\delta = \sigma_{L-1}/2$, occurs with probability at least $1 - O((d_{[0:L-1]})^{-2})$. Since the double sum is bounded by a constant thanks to the Lipschitzness of φ, it is easy to see that the overall contribution deriving from $\mathbb{1}(B_L^c)$ is $O((1/n + 1/d_m)n^2/d_m^2)$.

We thus shift our focus on the contribution deriving from $\mathbb{1}(B_L)$. The diagonal terms, $i = j$, in the double sum above are easily bounded by $O(d_L^{-1})$ because φ is Lipschitz. Let us analyse instead the off-diagonal terms for $i \neq j$. The expectation w.r.t. $\mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}$ only can be pushed till the mentioned double sum, and for $i \neq j$ it yields an overall contribution:

$$\frac{d_L(d_L - 1)}{d_L^2}\big(\mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}[(\varphi'(\alpha_{1\mu}(\tau)) - \rho_L)\varphi'(\alpha_{1\nu})]\big)^2. \qquad (92)$$

The above is an expectation of a bounded random variable w.r.t. two Gaussians with correlation

$$\mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}\alpha_{1\mu}(\tau)\,\alpha_{1\nu} = \sin\tau \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}. \qquad (93)$$

Therefore, by Lemma 6

$$\mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}[(\varphi'(\alpha_{1\mu}(\tau)) - \rho_L)\varphi'(\alpha_{1\nu})] = \mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}[(\varphi'(\alpha_{1\mu}(\tau)) - \rho_L)]\mathbb{E}_{\mathbf{W}^{*(L)}}\varphi'(\alpha_{1\nu})$$
$$+ O\Big(\frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{\|\mathbf{X}_\nu^{(L-1)}\|^2}\Big). \tag{94}$$

Since φ is Lipschitz, and following from the proof of Lemma 9, restricted to the event B_L we have

$$\mathbb{E}_{\mathbf{W}^{*(L)},\tilde{\mathbf{W}}^*}[(\varphi'(\alpha_{1\mu}(\tau)) - \rho_L)\varphi'(\alpha_{1\nu})] = O\Big(\Big(\frac{\|\mathbf{X}_\mu^{(L-1)}\|^2}{d_{L-1}} - \sigma_{L-1}\Big) + \frac{\mathbf{X}_\mu^{(L-1)} \cdot \mathbf{X}_\nu^{(L-1)}}{d_{L-1}}\Big). \tag{95}$$

By plugging this back into the inequality for $(A_{12}^{off}(\tau))^2$, and using Lemmas 10 and 9, the event B_L contributes for $O((1 + n/d_m)n/d_m^2)$.

The statement follows from selecting the worst of the convergence rates, that is the one of $A_{11}^{off}(\tau)$.

∎

Lemma 16 $A_3 - A_1 = \Big(\Big(1 + \sqrt{\frac{n}{d_m}}\Big)\frac{1}{\sqrt{d_m}}\Big).$

Proof We take up the computation from (31). We have already proved that its second line is controlled with $O((1 + n/d_m)\sqrt{n}/d_m)$. To bound the first line of (31), that we denote by D hereby, it is sufficient to use Cauchy-Schwartz's inequality first, and Lemma 12:

$$D^2 \le Cn\mathbb{V}_{(t)}\Big[\frac{1}{n}\log \mathcal{Z}_t\Big]\mathbb{E}\Big[\epsilon_L - \frac{\|\mathbf{X}_1^{(L)}\|^2}{d_L} + \rho_L\frac{\boldsymbol{\alpha}_1^\mathsf{T}\varphi(\boldsymbol{\alpha}_1)}{d_L}\Big]^2 \tag{96}$$

for a positive constant $C > 0$. Thanks to Lemma 9 we can replace $\|\mathbf{X}_1^{(L)}\|^2/d_L$ and $\boldsymbol{\alpha}_1^\mathsf{T}\varphi(\boldsymbol{\alpha}_1)/d_L$ with their expectations up to an $O(d_m^{-1})$:

$$D^2 \le Cn\mathbb{V}_{(t)}\Big[\frac{1}{n}\log \mathcal{Z}_t\Big]\Big[O(d_m^{-1}) + (\epsilon_L - \sigma_L + \rho_L^2\sigma_{L-1})^2\Big] = O\Big(\frac{1}{d_m}\Big(1 + \frac{n}{d_m}\Big)\Big), \tag{97}$$

where we used the definition of ϵ_L. Combining the rate for D and the one for A_1^{off} we thus get the statement. ∎

So far, we have proved that the free entropies of the original model and the reduced are as closed as desired. It is finally time to deal with the first term of (11).

Lemma 17 *Define*

$$\Psi_k := \mathbb{E}\log P_{out}\Big(Y_1^{(k)} \mid \eta_k\frac{\mathbf{a}^{*(k)\mathsf{T}}\mathbf{X}_1^{(k)}}{\sqrt{d_k}} + \sqrt{\gamma_k}\xi_1^{*(k)}\Big) \tag{98}$$

where $\mathbf{a}^{*(k)} \in \mathbb{R}^{d_k}$, $k = L$ or $L - 1$, and η_k, γ_k are defined in (16). Recall that

$$Y_1^{(k)} \sim P_{out}\Big(\cdot \mid \eta_k\frac{\mathbf{a}^{*(k)\mathsf{T}}\mathbf{X}_1^{(k)}}{\sqrt{d_k}} + \sqrt{\gamma_k}\xi_1^{*(k)}\Big).$$

Then

$$|\Psi_L - \Psi_{L-1}| = O(d_m^{-1/2}). \tag{99}$$

Proof Notice that $\mathbf{a}^{*(k)}$ is independent on $\xi^{*(k)}$. Therefore, letting $Z \sim \mathcal{N}(0,1)$, we have

$$\Psi_k = \mathbb{E} \int dy P_{out}\left(y \mid Z\sqrt{\eta_k^2 \frac{\|\mathbf{X}_1^{(k)}\|^2}{d_k} + \gamma_k}\right) \log P_{out}\left(y \mid Z\sqrt{\eta_k^2 \frac{\|\mathbf{X}_1^{(k)}\|^2}{d_k} + \gamma_k}\right). \qquad (100)$$

Recall $\rho = 1m, \epsilon = 0$ out of convenience, without loss of generality. Then $\eta_L = 1, \gamma_L = 0$. It is then convenient to introduce

$$S(t) = t\sqrt{\frac{\|\mathbf{X}_1^{(L)}\|^2}{d_L}} + (1-t)\sqrt{\rho_L^2 \frac{\|\mathbf{X}_1^{(L-1)}\|^2}{d_{L-1}} + \epsilon_L} \qquad (101)$$

and

$$\Psi(t) = \mathbb{E} \int dy P_{out}(y \mid ZS(t)) \log P_{out}(y \mid ZS(t)). \qquad (102)$$

It is then straightforward to see that $\Psi(0) = \Psi_{L-1}$ and $\Psi(1) = \Psi_L$. Using Lemma 12 the one can readily show that

$$|\dot{\Psi}(t)| \le C(f)\mathbb{E}|Z|\mathbb{E}|\dot{S}(t)| \qquad (103)$$

for some $C(f) > 0$. Using Lemma 9, and the fact that $\rho_L\sigma_{L-1} + \epsilon_L = \sigma_L$ one readily gets the result.

■

For completeness we also prove the following Lemma, which relates the various mutual informations

Lemma 18 *Recall the definition*

$$I_n(\boldsymbol{\xi}^*; \mathcal{D} \mid \boldsymbol{\theta}^*) = H(\mathcal{D} \mid \boldsymbol{\theta}^*) - H(\mathcal{D} \mid \boldsymbol{\xi}^*, \boldsymbol{\theta}^*) = -\mathbb{E}\log \mathbb{E}_\xi P_{out}\left(Y_1 \mid \rho\frac{\mathbf{a}^{*\intercal}\mathbf{X}_1^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi\right)$$

$$+ \mathbb{E}\mathbb{E}_{\xi^*}\log P_{out}\left(Y_1 \mid \rho\frac{\mathbf{a}^{*\intercal}\mathbf{X}_1^{(L)}}{\sqrt{d_L}} + \sqrt{\epsilon}\xi^*\right) \qquad (104)$$

where $\xi, \xi^* \sim \mathcal{N}(0,1)$ and Y_1 is distributed as in (1). Then

$$I_n(\boldsymbol{\xi}^*; \mathcal{D} \mid \boldsymbol{\theta}^*) = -\int dy \mathbb{E}_{Z,\xi^*} P_{out}(y \mid Z\sqrt{\rho^2\sigma_L} + \sqrt{\epsilon}\xi^*)\log \mathbb{E}_\xi P_{out}(y \mid Z\sqrt{\rho^2\sigma_L} + \sqrt{\epsilon}\xi)$$

$$+ \int dy \mathbb{E}_Z P_{out}(y \mid Z\sqrt{\rho^2\sigma_L + \epsilon})\log P_{out}(y \mid Z\sqrt{\rho^2\sigma_L + \epsilon}) + O(d_m^{-1/2}), \qquad (105)$$

where $Z \sim \mathcal{N}(0,1)$.

Proof For both terms the proof proceeds in the same way, by observing that

$$\rho\frac{\mathbf{a}^{*\intercal}\mathbf{X}_1^{(L)}}{\sqrt{d_L}} \sim Z\sqrt{\rho^2 \frac{\|\mathbf{X}_1^{(L)}\|^2}{d_L}} \qquad (106)$$

where $\sim$ denotes equality in distribution. Then one proceeds interpolating as in Lemma 17, and using the concentration in Lemma 9. Note that for $H(\mathcal{D} \mid \boldsymbol{\xi}^*, \boldsymbol{\theta}^*)$ it possible to join the Gaussians Z and ξ^* in a unique Gaussian, whereas for $H(\mathcal{D} \mid \boldsymbol{\theta}^*)$ this is not possible which explains the presence of the two variables ξ, ξ^* in the first line of (105). ■

Appendix F. Complete network reduction

Iterating Theorem 1 it is possible to reduce the full network to a GLM in terms of mutual information between dataset and teacher weights. To prove Corollary 3 holds with the coefficients (16) it is sufficient to start from a network with k layers left, and infer the coefficients η_{k-1}, γ_{k-1} starting from η_k, γ_k.

By the Layer Reduction Theorem we can replace

$$\eta_k \frac{\mathbf{a}^{*\intercal}\mathbf{X}_\mu^{(k)}}{\sqrt{d_k}} + \sqrt{\gamma_k}\zeta_\mu^{*(k)} \longrightarrow \eta_k\Big(\rho_k\frac{\mathbf{v}^{*\intercal}\mathbf{X}_\mu^{(k-1)}}{\sqrt{d_{k-1}}} + \sqrt{\epsilon_k}\xi_\mu^*\Big) + \sqrt{\gamma_k}\zeta_\mu^{*(k)}, \tag{107}$$

where $\mathbf{a}^* \in \mathbb{R}^{d_k}, \mathbf{v}^* \in \mathbb{R}^{d_{k-1}}$ are two vectors with i.i.d. standard Gaussian components, and $\zeta_\mu^{*(k)}, \xi_\mu^*$ are two standard Gaussians. Hence we infer the recursion

$$\eta_{k-1} = \eta_k\rho_k, \quad \gamma_{k-1} = \eta_k^2\epsilon_k + \gamma_k. \tag{108}$$

Using $\eta_L = \rho, \gamma_L = \epsilon$ then one readily gets (16) from the above recursion.

Appendix G. Concentration of free energy

In this section we prove the concentration of the free energy for the interpolated model described in Section 3.1. For convenience we replicate here some of the main notations. We consider interpolating models of teacher and student between the nonlinear argument of the readout function for the L-layer neural network and its corresponding linearized model:

$$S_{t\mu} := \sqrt{1-t}\frac{\mathbf{a}^{*\intercal}}{\sqrt{d_L}}\varphi\Big(\frac{\mathbf{W}^{*(L)}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big) + \sqrt{t}\rho_L\frac{\mathbf{v}^{*\intercal}\mathbf{X}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} + \sqrt{t\epsilon_L}\zeta_\mu^{*(L)}, \tag{109}$$

$$s_{t\mu} := \sqrt{1-t}\frac{\mathbf{a}^{\intercal}}{\sqrt{d_L}}\varphi\Big(\frac{\mathbf{W}^{(L)}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d_{L-1}}}\Big) + \sqrt{t}\rho_L\frac{\mathbf{v}^{\intercal}\mathbf{x}_\mu^{(L-1)}}{\sqrt{d_{L-1}}} + \sqrt{t\epsilon_L}\zeta_\mu^{(L)}, \tag{110}$$

and

$$Y_{t\mu} = f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta}Z_\mu. \tag{111}$$

We recall the corresponding random free entropy $f_{n,t}$ and partition function $\mathcal{Z}_t$

$$f_{n,t} = \frac{1}{n}\log\mathcal{Z}_t = \frac{1}{n}\log\int dP(\boldsymbol{\theta}^{(L)})\mathbb{E}_{\mathbf{v}}\prod_{\mu=1}^n \mathbb{E}_{\zeta_\mu^{(L)}}P_{out}(Y_{t\mu}|s_{t\mu}). \tag{112}$$

The main result of this section is

Theorem 19 *For the activation and readout function φ, f taken as in Theorem 1, there exists a non-negative constant $C(f,\varphi)$ such that*

$$\mathbb{V}\Big(\frac{1}{n}\log\mathcal{Z}_t\Big) \le C(f,\varphi)\Big(\frac{1}{n} + \frac{1}{d_m}\Big). \tag{113}$$

We remark that Theorem 2 is a direct consequence of the latter result, where t is taken equal to 0.

We prove Theorem 19 in several steps, first we show concentration with respect to the set of parameters $\psi_1 := \{Z_\mu, \zeta_\mu^{*(L)}, \mathbf{X}_\mu^{(0)}\}_{\mu=1}^n$ using classical Poincaré-Nash inequality, after this, with respect to $\psi_2 := \{\mathbf{A}_\mu\}_{\mu=1}^n$ using the corollary of Efron-Stein inequality, and finally, with respect to $\psi_3 := \{(\mathbf{W}^{*(k)})_{\ell=1}^L, \mathbf{v}^*, \mathbf{a}^*\}$. We recall two classical concentration results, whose proofs can be found in Boucheron et al. (2004), Chapter 3.

Proposition 20 (Poincaré-Nash inequality) *Let $\xi = [\xi_1, \ldots, \xi_K]^\intercal$ be a real Gaussian standard random vector. If $g : \mathbb{X}^K \mapsto \mathbb{X}$ is a continuously differentiable function, then*

$$\mathbb{V}g(\xi) \le \mathbb{E}\|\nabla g(\xi)\|^2 . \tag{114}$$

Proposition 21 (Bounded difference) *Let $\xi = [\xi_1, \ldots, \xi_K]^\intercal$ be a random vector with i.i.d. elements taking values in some space $\mathcal{A}$. If function $g : \mathcal{A}^K \mapsto \mathbb{X}$ satisfies*

$$\sup_{1 \le i \le K} \sup_{x_1,\ldots,x_K,x'_i \in \mathcal{A}} |g(x_1, \ldots, x_i, \ldots, x_K) - g(x_1, \ldots, x'_i, \ldots, x_K)| \le C \tag{115}$$

for some $C > 0$, then

$$\mathbb{V}\{g(\xi)\} \le \frac{1}{4}KC^2. \tag{116}$$

In what follows we denote $P^y(y|x) := \partial_y P_{out}(y|x)$ and $P^x(y|x) := \partial_x P_{out}(y|x)$.

Lemma 22 *There exists a non-negative constant $C(f, \varphi)$ such that*

$$\mathbb{E}\Big(\frac{1}{n}\log \mathcal{Z}_t - \mathbb{E}_{\psi_1}\frac{1}{n}\log \mathcal{Z}_t\Big)^2 \le \frac{C(f, \varphi)}{n} .$$

Proof Since the elements of $\mathbf{X}_\mu^{(0)}$, $\zeta_\mu^{*(L)}$, and Z_μ for $\mu = 1, \ldots, n$ are jointly independent standard Gaussian, we can use Proposition 20 and write

$$\mathbb{E}\mathbb{V}_{\psi_1}\Big(\frac{1}{n}\log \mathcal{Z}_t\Big) \le \frac{1}{n^2}\mathbb{E}\|\nabla_{\psi_1}\log \mathcal{Z}_t\|^2$$

$$= \frac{1}{n^2}\sum_{\mu=1}^n \mathbb{E}\Big(\frac{\partial \log \mathcal{Z}_t}{\partial \zeta_\mu^{*(L)}}\Big)^2 + \frac{1}{n^2}\sum_{\mu=1}^n\sum_{i=1}^{d_0} \mathbb{E}\Big(\frac{\partial \log \mathcal{Z}_t}{\partial X_{i\mu}^{(0)}}\Big)^2 + \frac{1}{n^2}\sum_{\mu=1}^n \mathbb{E}\Big(\frac{\partial \log \mathcal{Z}_t}{\partial Z_\mu}\Big)^2 =: I_1 + I_2 + I_3.$$

For brevity in what follows we write $P_\mu = P_{out}(Y_{t\mu}|s_{t\mu})$. Then, expressing the derivatives from the first term, we get

$$\Big|\frac{\partial \log \mathcal{Z}_t}{\partial \zeta_\mu^{*(L)}}\Big| \le \Big|\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu)\sqrt{t\epsilon_L}\Big| \le \sqrt{t\epsilon_L}C(f)(|Z_\mu|^2 + 1) .$$

The last inequality is true due to Lemma 11 and the fact that function f' is bounded. Since Z_μ is Gaussian, all the moments of its absolute value are bounded with a constant, which gives us

$$\mathbb{E}\Big(\frac{\partial \log \mathcal{Z}_t}{\partial \zeta_\mu^{*(L)}}\Big)^2 \le C(f) \Rightarrow I_1 \le \frac{C(f)}{n} .$$

The partial derivative with respect to Z_μ gives a simple expression that also can be handled using Lemma 11

$$\left|\frac{\partial \log \mathcal{Z}_t}{\partial Z_\mu}\right| = \sqrt{\Delta}\left|\left\langle \frac{P_\mu^y}{P_\mu}\right\rangle\right| \le C(f)(|Z_\mu|^2 + 1). \tag{117}$$

This gives us immediately $I_3 \le C(f)n^{-1}$.

Before passing to the more challenging part, let us set some notation. First, set $\alpha_{i\mu}^{*(\ell-1)} := \mathbf{W}_i^{*(\ell)} \cdot \mathbf{X}_\mu^{(\ell-1)}/\sqrt{d_{\ell-1}}$. In what follows we will need to calculate derivatives of $\mathbf{X}_\nu^{(\ell)}$ and $\mathbf{x}_\nu^{(\ell)}$. The chain rule gives us

$$\frac{\partial X_{j\nu}^{(\ell)}}{\partial X_{i\mu}^{(0)}} = \frac{\partial\varphi\Big(\frac{\mathbf{W}_j^{*(\ell)}}{\sqrt{d_{\ell-1}}}\varphi\Big(\frac{\mathbf{W}^{*(\ell-1)}}{\sqrt{d_{\ell-2}}}\varphi\Big(\ldots\varphi\Big(\frac{\mathbf{W}^{*(1)}\mathbf{X}_\mu^{(0)}}{\sqrt{d_0}}\Big)\Big)\Big)}{\partial X_{i\mu}^{(0)}}$$

$$= \delta_{\mu,\nu}\Big(\prod_{p=1}^{\ell-1}\sum_{j_p=1}^{d_p}\Big)\varphi'(\alpha_{j\nu}^{*(\ell-1)})\frac{W_{jj_{\ell-1}}^{*(\ell)}}{\sqrt{d_{\ell-1}}}\varphi'(\alpha_{j_{\ell-1}\nu}^{*(\ell-2)})\frac{W_{j_{\ell-1}j_{\ell-2}}^{*(\ell-1)}}{\sqrt{d_{\ell-2}}}\ldots\varphi'(\alpha_{j_1\nu}^{*(0)})\frac{W_{j_1i}^{*(1)}}{\sqrt{d_0}}.$$

For $1 \le \ell_1 \le \ell_2 \le L$ we denote $d_{\ell_2} \times d_{\ell_1-1}$ matrix

$$\mathbf{B}_\nu^{*[\ell_1:\ell_2]} = \frac{\boldsymbol{\Phi}_\nu^{*(\ell_2-1)}\mathbf{W}^{*(\ell_2)}\boldsymbol{\Phi}_\nu^{*(\ell_2-2)}\ldots\boldsymbol{\Phi}_\nu^{*(\ell_1-1)}\mathbf{W}^{*(\ell_1)}}{\sqrt{d_{\ell_2-1}\ldots d_{\ell_1-1}}} \tag{118}$$

and $d_{\ell+1} \times d_{\ell+1}$ diagonal matrix $\boldsymbol{\Phi}_\nu^{*(\ell)} = \mathrm{diag}(\varphi'(\alpha_{1\nu}^{*(\ell)}),\ldots,\varphi'(\alpha_{d_{\ell-1}\nu}^{*(\ell)}))$. Then previous expression can be rewritten as

$$\frac{\partial X_{j\nu}^{(\ell)}}{\partial X_{i\mu}^{(0)}} = \delta_{\mu,\nu}B_{ji,\nu}^{*[1:\ell]}. \tag{119}$$

For the following calculations, it will be useful to find the order of the operator norm of $\mathbf{B}_\nu^{*[\ell_1:\ell_2]}$. Since the first derivative of φ is bounded by some constant $C(\varphi)$, it is straightforward that the operator norm $\|\boldsymbol{\Phi}_\nu^{*(\ell)}\|$ is also bounded with $C(\varphi)$. Lastly, the expected operator norm of high dimensional matrices, say of size $a \times b$, with i.i.d. standard Gaussian elements is of order $O(\sqrt{a}) = O(\sqrt{b})$ with high probability. This entails

$$\mathbb{E}\|\mathbf{B}_\nu^{*[\ell_1:\ell_2]}\| \le \bar{C}(\varphi)\sqrt{\prod_{p=\ell_1}^{\ell_2}\frac{\mathbb{E}\|\mathbf{W}^{*(p)}\|^2}{d_{p-1}}} = O(1). \tag{120}$$

Similar bounds hold for any finite moment of the operator norm. Following the same lines we can write the expression for the derivative of $x_{j\nu}^{(\ell)}$ and define the students version of $\boldsymbol{\Phi}_\nu^{*(\ell)}$ and $\mathbf{B}_\nu^{*[\ell_1:\ell_2]}$ by removing stars from all involved parameters.

$$\frac{\partial x_{j\nu}^{(\ell)}}{\partial X_{i\mu}^{(0)}} = \delta_{\mu,\nu}B_{\nu,ji}^{[1:\ell]}. \tag{121}$$

Returning to I_2, we have

$$\frac{\partial \log \mathcal{Z}_t}{\partial X_{i\mu}^{(0)}} = \sum_{j=1}^{d_{L-1}} \frac{\partial X_{j\mu}^{(L-1)}}{\partial X_{i\mu}^{(0)}} \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu)\Big(\sqrt{1-t}\frac{(\mathbf{a}^{*\intercal}\boldsymbol{\Phi}_\mu^{*(L-1)}\mathbf{W}^{*(L)})_j}{\sqrt{d_L d_{L-1}}} + \sqrt{t}\rho_L \frac{v_j^*}{\sqrt{d_{L-1}}}\Big)$$
$$+ \sum_{j=1}^{d_{L-1}} \Big(\sqrt{1-t}\Big\langle \frac{\partial x_{j\mu}^{(L-1)}}{\partial X_{i\mu}^{(0)}} \frac{P_\mu^x}{P_\mu} \frac{(\mathbf{a}^\intercal\boldsymbol{\Phi}_\mu^{(L-1)}\mathbf{W}^{(L)})_j}{\sqrt{d_L d_{L-1}}}\Big\rangle + \sqrt{t}\rho_L \Big\langle \frac{\partial x_{j\mu}^{(L-1)}}{\partial X_{i\mu}^{(0)}} \frac{P_\mu^x}{P_\mu} \frac{v_j}{\sqrt{d_{L-1}}}\Big\rangle\Big).$$

Plugging (119) and (121) in the latter expression and taking the sum over j we obtain

$$\frac{\partial \log \mathcal{Z}_t}{\partial X_{i\mu}^{(0)}} = \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu)\Big(\sqrt{1-t}\frac{(\mathbf{a}^{*\intercal}\mathbf{B}_\mu^{*[1:L]})_i}{\sqrt{d_L}} + \sqrt{t}\rho_L \frac{(\mathbf{v}^{*\intercal}\mathbf{B}_\mu^{*[1:L-1]})_i}{\sqrt{d_{L-1}}}\Big)$$
$$+ \sqrt{1-t}\Big\langle \frac{P_\mu^x}{P_\mu} \frac{(\mathbf{a}^\intercal\mathbf{B}_\mu^{[1:L]})_i}{\sqrt{d_L}}\Big\rangle + \sqrt{t}\rho_L \Big\langle \frac{P_\mu^x}{P_\mu} \frac{(\mathbf{v}^\intercal\mathbf{B}_\mu^{[1:L-1]})_i}{\sqrt{d_{L-1}}}\Big\rangle.$$

Using twice the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ as well as Lemma 11 and boundedness of f' we get thee expression for I_2

$$I_2 \leq \frac{1}{n^2} \sum_{\mu=1}^{n} \Big[C(f)\mathbb{E}\Big((1+|Z_\mu|^2)^2\big(\|\frac{\mathbf{a}^{*\intercal}}{d_k}\mathbf{B}_\mu^{*[1:L]}\|^2 + \|\frac{\mathbf{v}^{*\intercal}}{d_{L-1}}\mathbf{B}_\mu^{*[1:L-1]}\|^2\big)\Big)$$
$$+ (1-t)\sum_{i=1}^{d_0} \mathbb{E}\Big\langle \frac{P_\mu^x}{P_\mu} \frac{(\mathbf{a}^\intercal\mathbf{B}_\mu^{[1:L]})_i}{\sqrt{d_L}}\Big\rangle^2 + t\rho_L^2 \sum_{i=1}^{d_0} \mathbb{E}\Big\langle \frac{P_\mu^x}{P_\mu} \frac{(\mathbf{v}^\intercal\mathbf{B}_\mu^{[1:L-1]})_i}{\sqrt{d_{L-1}}}\Big\rangle^2 \Big].$$

For the last two terms, we can bring square inside the Gibbs brackets with the help of Jensen inequality and then we notice that obtained expressions inside the Gibbs brackets depend only on $Y_{t\mu}$ and $s_{t\mu}$, which allows us to use Nishimori identity by removing brackets and adding $*$ to all the parameters. After this we can bound each ratio which includes P_{out} by using again Lemma 11. All this will result in

$$I_2 \leq \frac{C(f)}{n^2} \sum_{\mu=1}^{n} \Big[\frac{\mathbb{E}(1+|Z_\mu|^2)^2\|\mathbf{a}^{*\intercal}\mathbf{B}_\mu^{*[1:L]}\|^2}{d_L} + \frac{\mathbb{E}(1+|Z_\mu|^2)^2\|\mathbf{v}^{*\intercal}\mathbf{B}_\mu^{*[1:L-1]}\|^2}{d_{L-1}}\Big].$$

Since Z_μ, $\mathbf{a}^*$, and $\mathbf{v}^*$ are Gaussian and independent of $\mathbf{B}_\mu^{*[1:L]}$ we have

$$I_2 \leq \frac{C(f)}{n^2} \sum_{\mu=1}^{n} \Big[\frac{\mathbb{E}(1+|Z_\mu|^2)^2\mathbb{E}\|\mathbf{a}^*\|^2\mathbb{E}\|\mathbf{B}_\mu^{*[1:L]}\|^2}{d_L}$$
$$+ \frac{\mathbb{E}(1+|Z_\mu|^2)^2\mathbb{E}\|\mathbf{v}^*\|^2\mathbb{E}\|\mathbf{B}_\mu^{*[1:L-1]}\|^2}{d_{L-1}}\Big] \leq \frac{C(f,\varphi)}{n}.$$

The last inequality being true due to (120) and the fact that $\mathbb{E}\|\mathbf{v}^{*(\ell)}\|^2 = O(d_\ell)$. ∎

For the next step we denote $h(\mathbf{A}) = \mathbb{E}_{\psi_1} \log \mathcal{Z}_t/n$, as a function of all the elements $A_{\mu,i}$ of $\mathbf{A}_\mu$ for $1 \leq i \leq k$ and $1 \leq \mu \leq n$.

Lemma 23 *There exists a non-negative constant $C(f,\varphi)$ such that*

$$\mathbb{E}\Big(\frac{1}{n}h - \mathbb{E}_{\mathbf{A}}\frac{1}{n}h\Big)^2 \leq \frac{C(f,\varphi)}{n}.$$

Proof We denote by $\mathbf{A}'$ a vector such that $A'_{\mu,i} = \mathbf{A}_{\mu,i}$ for $\mu \neq \nu$, $i \neq j$ and $A'_{\nu,j}$ is a random variable with distribution P_A, independent of all others. According to Proposition 21 it is sufficient to prove that

$$|h(\mathbf{A}') - h(\mathbf{A})| < \frac{C(f,\varphi)}{n}. \tag{122}$$

If we denote by H (and H') the Hamiltonian corresponding to $\mathcal{Z}_t$ (and $\mathcal{Z}_t$ with $\mathbf{A}'$), i.e.

$$H = -\sum_{\mu=1}^{n} \log P_{out}(f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta}Z_\mu | s_{t\mu}),$$
$$H' = -\sum_{\mu\neq\nu} \log P_{out}(f(S_{t\mu}; \mathbf{A}_\mu) + \sqrt{\Delta}Z_\mu | s_{t\mu}) - \log P_{out}(f(S_{t\nu}; \mathbf{A}'_\nu) + \sqrt{\Delta}Z_\nu | s_{t\nu}).$$

one can see that

$$h(\mathbf{A}') - h(\mathbf{A}) = \frac{1}{n}\mathbb{E}_{\psi_1} \log \Big\langle e^{H-H'} \Big\rangle_H \geq \frac{1}{n}\mathbb{E}_{\psi_1}\Big\langle H - H' \Big\rangle_H,$$

the last step being true due to Jensen inequality. On the other hand $h(\mathbf{A}') - h(\mathbf{A}) \leq n^{-1}\mathbb{E}_{\psi_1}\langle H - H'\rangle_{H'}$. We recall the definition of P_{out}

$$P_{out}(y \mid x) = \int_{\mathbb{R}^p} \frac{dP_A(\mathbf{A})}{\sqrt{2\pi\Delta}} \exp\Big[-\frac{1}{2\Delta}(y - f(x; \mathbf{A}))^2 \Big].$$

Similarly, we obtain

$$H - H' \geq \frac{1}{2\Delta}\Big\langle (f(S_{t\nu}; \mathbf{A}_\nu) - f(s_{t\nu}; \tilde{\mathbf{A}}) + \sqrt{\Delta}Z_\nu)^2 - (f(S_{t\nu}; \mathbf{A}'_\nu) - f(s_{t\nu}; \tilde{\mathbf{A}}) + \sqrt{\Delta}Z_\nu)^2 \Big\rangle_G$$

and

$$H - H' \leq \frac{1}{2\Delta}\Big\langle (f(S_{t\nu}; \mathbf{A}_\nu) - f(s_{t\nu}; \tilde{\mathbf{A}}) + \sqrt{\Delta}Z_\nu)^2 - (f(S_{t\nu}; \mathbf{A}'_\nu) - f(s_{t\nu}; \tilde{\mathbf{A}}) + \sqrt{\Delta}Z_\nu)^2 \Big\rangle_{G'},$$

where $\langle\cdot\rangle_G$ (or with G') defined as

$$\langle\cdot\rangle_G = \frac{\int P_A(d\tilde{\mathbf{A}})e^{-\frac{1}{2\Delta}(Y_{t\nu} - f(s_{t\nu};\tilde{\mathbf{A}}))^2}(\cdot)}{\int P_A(d\tilde{\mathbf{A}})e^{-\frac{1}{2\Delta}(Y_{t\nu} - f(s_{t\nu};\tilde{\mathbf{A}}))^2}} \tag{123}$$

or with $Y'_{t\nu}$ where $\mathbf{A}_\nu$ is changed to $\mathbf{A}'_\nu$. Since f is bounded, we immediately obtain $|H - H'| \leq C(f)(|Z_\mu|^2 + 1)$ and (122). Now, due to the Proposition 21, the statement of the Lemma is proved. ∎

Finally, we prove the concentration with respect to $(\mathbf{W}^{*(\ell)})_{\ell=1}^L$, $\mathbf{a}^*$, and $\mathbf{v}^*$. For this we denote $g \equiv g((\mathbf{W}^{*(\ell)})_{\ell=1}^L, \mathbf{a}^*, \mathbf{v}^*) := \mathbb{E}_{\psi_1,\psi_2} \log \mathcal{Z}_t/n$.

Lemma 24 *There exists a non-negative constant $C(f, \varphi)$ such that*

$$\mathbb{E}\Big(g - \mathbb{E}_{\psi_3} g\Big)^2 \leq \frac{C(f, \varphi)}{d_m} .$$

Proof We use again Poincaré-Nash inequality and write

$$\mathbb{E}\Big(g - \mathbb{E}_{(\mathbf{W}^{*(\ell)})_{\ell=1}^L, \mathbf{a}^*, \mathbf{v}^*} g\Big)^2 \leq \sum_{\ell=1}^L \sum_{i_\ell=1}^{d_\ell} \sum_{j_\ell=1}^{d_{\ell-1}} \mathbb{E}\Big(\frac{\partial g}{\partial W_{i_\ell j_\ell}^{*(\ell)}}\Big)^2 + \sum_{i=1}^{d_L} \mathbb{E}\Big(\frac{\partial g}{\partial a_i^*}\Big)^2 + \sum_{j=1}^{d_{L-1}} \mathbb{E}\Big(\frac{\partial g}{\partial v_j^*}\Big)^2 . \tag{124}$$

Concentration with respect to $\mathbf{W}^{*(\ell)}$. To find the derivative of g with respect to $W_{i_\ell j_\ell}^{*(\ell)}$ it would be convenient to calculate first the general derivative of $X_{\mu,p}^{(k)}$ for $k \geq \ell$

$$\frac{\partial X_{p\mu}^{(k)}}{\partial W_{i_\ell j_\ell}^{*(\ell)}} = \frac{(\mathbf{B}_\mu^{*[\ell+1:k]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{p i_\ell} X_{j_\ell \mu}^{(\ell-1)}}{\sqrt{d_{\ell-1}}},$$

where in the case of $k = \ell$ we set $\mathbf{B}^{*[\ell+1:\ell]} = \mathbf{I}_{d_\ell}$. Then the derivative of g is

$$\frac{\partial g}{\partial W_{i_\ell j_l}^{*(\ell)}} = \frac{1}{n} \sum_{\mu=1}^n \mathbb{E}_{\psi_1, \psi_2} \Big[\Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu) \Big(\frac{\sqrt{1-t}(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_l} X_{j_\ell \mu}^{(\ell-1)}}{\sqrt{d_L d_{\ell-1}}}$$
$$+ \frac{\sqrt{t}\rho_L (\mathbf{v}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L-1]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell} X_{j_\ell \mu}^{(\ell-1)}}{\sqrt{d_{L-1} d_{\ell-1}}} \Big) \Big] .$$

For $\ell = L$ the second term is obviously absent, to reflect this, we set $\mathbf{B}_\mu^{*[L+1:L-1]} = 0$. It is worth to notice, that ψ_1 and ψ_2 contain all elements with dependency of μ, which means that each term in the sum over μ will be the same, so we can remove $\frac{1}{n} \sum_\mu^n$ and write

$$\frac{\partial g}{\partial W_{i_\ell j_\ell}^{*(\ell)}} = \mathbb{E}_{\psi_1, \psi_2} \Big[\Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu) \Big(\frac{\sqrt{1-t}(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell} X_{j_\ell \mu}^{(\ell-1)}}{\sqrt{d_L d_{\ell-1}}}$$
$$+ \frac{\sqrt{t}\rho_L (\mathbf{v}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L-1]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell} X_{j_\ell \mu}^{(\ell-1)}}{\sqrt{d_{L-1} d_{\ell-1}}} \Big) \Big]$$

for fixed arbitrary μ. The two resulting terms are of the very similar form, to avoid repetitive computations we will focus only on the first one, that we denote by E. Since later we will use the usual inequality $\mathbb{E}(a+b)^2 \leq 2\mathbb{E}a^2 + 2\mathbb{E}b^2$, it is sufficient to bound $\mathbb{E}E^2$. One can see that by estimating in a straightforward way we can achieve only the order of $O(1)$, which is not what we are aiming for. To reach the right order, we need another integration by parts and to be able to do so, we are introducing a circular interpolation which replaces $\mathbf{X}_\mu^{(0)}$ by it's independent copy $\tilde{\mathbf{X}}_\mu^{(0)}$

$$\mathbf{X}_\mu^{(0)}(\tau) = \mathbf{X}_\mu^{(0)} \sin \tau + \tilde{\mathbf{X}}_\mu^{(0)} \cos \tau, \tag{125}$$

where $\tau \in [0, \pi/2]$. The nice properties of the new Gaussian variable $\mathbf{X}_\mu^{(0)}(\tau)$ is that its derivative with respect to τ, denoted $\dot{\mathbf{X}}_\mu^{(0)}(\tau)$, has simple correlations

$$\mathbb{E}_{\mathbf{X}_\mu^{(0)}, \tilde{\mathbf{X}}_\mu^{(0)}}(X_{i\nu}^{(0)}(\tau)\dot{X}_{j\mu}^{(0)}(\tau)) = 0\,, \quad \mathbb{E}_{\mathbf{X}_\mu^{(0)}, \tilde{\mathbf{X}}_\mu^{(0)}}(X_{i\nu}^{(0)}\dot{X}_{j\mu}^{(0)}(\tau)) = \delta_{\mu\nu}\delta_{ij}\cos\tau\,. \qquad (126)$$

Respectively, in a recursive way we define $\mathbf{X}_\mu^{(\ell)}(\tau) = \varphi(\mathbf{W}^{*(\ell)}\mathbf{X}_\mu^{(\ell-1)}(\tau)/\sqrt{d_{\ell-1}})$ and $\tilde{\mathbf{X}}_\mu^{(\ell)} = \varphi(\mathbf{W}^{*(\ell)}\tilde{\mathbf{X}}_\mu^{(\ell-1)}/\sqrt{d_{\ell-1}})$ for $1 \le \ell \le L$, as well as $\boldsymbol{\alpha}_\mu^{*(\ell)}(\tau) = \mathbf{W}^{*(\ell+1)}\mathbf{X}_\mu^{(\ell)}(\tau)/\sqrt{d_\ell}$ and $\tilde{\boldsymbol{\alpha}}_\mu^{*(\ell)} = \mathbf{W}^{*(\ell+1)}\tilde{\mathbf{X}}_\mu^{(\ell)}/\sqrt{d_\ell}$.

First, let us deal with $\ell > 1$. In such a case $\mathbf{X}_\mu^{(\ell-1)} = \varphi(\boldsymbol{\alpha}_\mu^{*(\ell-2)})$, which can be rewritten us

$$\varphi(\boldsymbol{\alpha}_\mu^{*(\ell-2)}) = \int_0^{\pi/2} d\tau \frac{d}{d\tau}\varphi(\boldsymbol{\alpha}_\mu^{*(\ell-2)}(\tau)) + \varphi(\tilde{\boldsymbol{\alpha}}_\mu^{*(\ell-2)})\,. \qquad (127)$$

Since φ is an odd function and $\tilde{\mathbf{X}}_\mu^{(0)}$ is independent of every other variable in g, the term containing $\varphi(\tilde{\boldsymbol{\alpha}}_\mu^{*(\ell-2)})$ will collapse to 0, and we obtain immediately

$$E = \mathbb{E}_{\psi_1,\psi_2,\tilde{\mathbf{X}}_\mu^{(0)}} \int_0^{\pi/2} \left\langle \frac{P_\mu^y}{P_\mu} \right\rangle f'(S_{t\mu};\mathbf{A}_\mu) \frac{\sqrt{1-t}(\mathbf{a}^{*\top}\mathbf{B}_\mu^{*[\ell+1:L]}\boldsymbol{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L d_{\ell-1}}} \frac{d}{d\tau}\varphi(\alpha_{j_\ell\mu}^{*(\ell-2)}(\tau))d\tau.$$

Let us have closer look at the derivative of $\varphi(\alpha_{j_\ell\mu}^{*(\ell-2)}(\tau))$. Using the chain rule and (119) we get

$$\frac{d}{d\tau}\varphi(\alpha_{j_\ell\mu}^{*(\ell-2)}(\tau)) = \sum_{p=1}^{d_0} \frac{\partial X_{j_\ell\mu}^{(\ell-1)}(\tau)}{\partial X_{p\mu}^{(0)}(\tau)}\dot{X}_{p\mu}^{(0)}(\tau) = \sum_{p=1}^{d_0} B_{j_\ell p,\mu}^{*[1:\ell-1]}(\tau)\dot{X}_{p\mu}^{(0)}(\tau), \qquad (128)$$

where for any $\ell_1, \ell_2 \in [1,\dots,L]$ matrix $\mathbf{B}_\mu^{*[\ell_1:\ell_2]}(\tau)$ defined the same way as $\mathbf{B}_\mu^{*[\ell_1:\ell_2]}$, with $\mathbf{X}_\mu^{(0)}$ replaced by $\mathbf{X}_\mu^{(0)}(\tau)$. With the help of (126) we integrate E by parts with respect to Gaussian parameter $\dot{X}_{p\mu}^{(0)}(\tau)$.

$$E = \sum_{p=1}^{d_0} \mathbb{E}_{\psi,\tilde{\mathbf{X}}_\mu^{(0)}} \int_0^{\pi/2} \cos\tau \left[\left\langle \frac{P_\mu^y}{P_\mu} \right\rangle f'(S_{t\mu};\mathbf{A}_\mu) \frac{\sqrt{1-t}(\mathbf{a}^{*\top}\mathbf{B}_\mu^{*[\ell+1:L]}\boldsymbol{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L d_{\ell-1}}}\right]'_{X_{p\mu}^{(0)}} B_{\mu,j_\ell p}^{*[1:\ell-1]}(\tau),$$

$$= I_1 + I_2 + I_3, \qquad (129)$$

38

where we denoted

$$
I_1 := \mathbb{E}_{\psi_1, \psi_2, \tilde{\mathbf{x}}_\mu^{(0)}} \int_0^{\pi/2} \cos\tau \sum_{p=1}^{d_0} \Big(-\Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle^2 f'(S_{t\mu}; \mathbf{A}_\mu)^2 + \Big\langle \frac{P_\mu^{yy}}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu)^2
$$
$$
+ \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f''(S_{t\mu}; \mathbf{A}_\mu) \Big) \frac{\partial S_{t\mu}}{\partial X_{p\mu}^{(0)}} \frac{\sqrt{1-t}(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L d_{\ell-1}}} B_{j_\ell p, \mu}^{*[1:\ell-1]}(\tau),
$$

$$
I_2 := \mathbb{E}_{\psi_1, \psi_2, \tilde{\mathbf{x}}_\mu^{(0)}} \int_0^{\pi/2} \cos\tau \sum_{p=1}^{d_0} \Big(-\Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle \Big\langle \frac{P_\mu^x}{P_\mu} \frac{\partial s_{t\mu}}{\partial X_{p\mu}^{(0)}} \Big\rangle + \Big\langle \frac{P_\mu^{yx}}{P_\mu} \frac{\partial s_{t\mu}}{\partial X_{p\mu}^{(0)}} \Big\rangle \Big)
$$
$$
\times f'(S_{t\mu}; \mathbf{A}_\mu) \frac{\sqrt{1-t}(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L d_{\ell-1}}} B_{j_\ell p, \mu}^{*[1:\ell-1]}(\tau),
$$

$$
I_3 := \sum_{p=1}^{d_0} \mathbb{E}_{\psi_1, \psi_2, \tilde{\mathbf{x}}_\mu} \int_0^{\pi/2} \cos\tau \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f'(S_{t\mu}; \mathbf{A}_\mu)
$$
$$
\times \Big[\frac{\sqrt{1-t}(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L d_{\ell-1}}} \Big]'_{X_{p\mu}^{(0)}} B_{j_\ell p, \mu}^{*[1:\ell-1]}(\tau).
$$

We start with the term I_1. The first bracket can be bounded right away with Lemma 11, and since Z_μ is Gaussian and independent of rest of I_1 we also bound $\mathbb{E}_{\psi_1}|Z_\mu|^k$ with a constant. The derivative of $S_{t\mu}$ is easily found with the help of (119), where after some simplification we have

$$
|I_1| \le \frac{C(f)}{\sqrt{d_L d_{\ell-1}}} \mathbb{E}_{\psi, \tilde{\mathbf{x}}_\mu^{(0)}} \int_0^{\pi/2} \Big| \frac{1}{\sqrt{d_L}} \big(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[1:L]} \mathbf{B}_\mu^{*[1:\ell-1]\intercal}(\tau) +
$$
$$
\frac{1}{\sqrt{d_{L-1}}} \mathbf{v}^{*\intercal} \mathbf{B}_\mu^{*[1:L-1]} \mathbf{B}_\mu^{*[1:\ell-1]\intercal}(\tau) \big)_{j_\ell} (\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell} \Big| d\tau.
$$

Separating two terms and using properties of operator norm several times as well as Jensen and Cauchy-Swartz inequalities, we have

$$
\mathbb{E} \sum_{i_\ell, j_\ell} |I_1|^2 \le \frac{C(f)}{d_{\ell-1}} \int \mathbb{E} \Big[\frac{\|\mathbf{a}^*\|^4}{d_L^2} \|\mathbf{B}_\mu^{*[1:L]}\|^2 \|\mathbf{B}_\mu^{*[1:\ell-1]}(\tau)\|^2
$$
$$
+ \frac{\|\mathbf{a}^*\|^2 \|\mathbf{v}^*\|^2}{d_L d_{L-1}} \|\mathbf{B}_\mu^{*[1:L-1]}\|^2 \|\mathbf{B}_\mu^{*[1:\ell-1]}(\tau)\|^2 \Big] \|\mathbf{B}_\mu^{*[\ell+1:L]}\|^2 \|\mathbf{\Phi}_\mu^{*(\ell-1)}\|^2 \le \frac{C(f, \varphi)}{d_{\ell-1}}.
$$

Let us move to I_2. Thanks to (121), the derivative of $s_{t\mu}$ has again two terms (one with $\mathbf{a}$ and second with $\mathbf{v}$, since they are of the same form, we will look only at the first one)

$$
|I_2| \le \frac{C(f)}{\sqrt{d_{\ell-1}}} \mathbb{E}_{\psi_1, \psi_2, \tilde{\mathbf{x}}_\mu^{(0)}} \int_0^{\pi/2} \Big(\Big\langle \Big| \frac{P_\mu^y}{P_\mu} \Big| \Big\rangle \Big\langle \Big| \frac{P_\mu^x}{P_\mu} \Big| \frac{|(\mathbf{a}^\intercal \mathbf{B}_\mu^{[1:L]} \mathbf{B}_\mu^{*[1:\ell-1]\intercal}(\tau))_{j_\ell}|}{\sqrt{d_L}} \Big\rangle \tag{130}
$$
$$
+ \Big\langle \Big| \frac{P_\mu^{yx}}{P_\mu} \Big| \frac{|(\mathbf{a}^\intercal \mathbf{B}_\mu^{[1:L]} \mathbf{B}_\mu^{*[1:\ell-1]\intercal}(\tau))_{j_\ell}|}{\sqrt{d_L}} \Big\rangle \Big) \frac{|(\mathbf{a}^{*\intercal} \mathbf{B}_\mu^{*[\ell+1:L]} \mathbf{\Phi}_\mu^{*(\ell-1)})_{i_\ell}|}{\sqrt{d_L}} + \dots.
$$

After using Lemma 11 and Jensen inequality we get

$$
\mathbb{E}\sum_{i_\ell,j_\ell}|I_2|^2 \le \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2}\Big\langle \frac{\|\mathbf{a}^\mathsf{T}\mathbf{B}_\mu^{[1:L]}\mathbf{B}_\mu^{*[1:\ell-1]\mathsf{T}}(\tau)\|^2}{d_L}\Big\rangle \frac{\|\mathbf{a}^{*\mathsf{T}}\mathbf{B}_\mu^{*[\ell+1:L]}\boldsymbol{\Phi}_\mu^{*(\ell-1)}\|^2}{d_L}(1+|Z_\mu|^2)^2 d\tau
$$
$$
+\dots. \tag{131}
$$

Lastly, separating quenched from annealed terms with the help of Cauchy-Swartz, we can get rid of the Gibbs bracket by using again Jensen and Nishimori. This will leave as with moments of operator norms of $\|d_L^{-1/2}\mathbf{a}^*\|$, $\|\mathbf{B}\|$, etc., since all of them are of order $O(1)$ we achieve the desired bound

$$
\mathbb{E}\sum_{i_\ell,j_\ell}|I_2|^2 \le \frac{C(f)}{d_{\ell-1}}. \tag{132}
$$

The final part of E we bound right away with

$$
|I_3| \le \frac{C(f)}{\sqrt{d_{\ell-1}}}\mathbb{E}_{\psi_1,\psi_2,\tilde{\mathbf{X}}_\mu}\int_0^{\pi/2}\Big|\sum_{p=1}^{d_0}\Big[\frac{(\mathbf{a}^{*\mathsf{T}}\mathbf{B}_\mu^{*[\ell+1:L]}\boldsymbol{\Phi}_\mu^{*(\ell-1)})_{i_\ell}}{\sqrt{d_L}}\Big]'_{X_{p\mu}^{(0)}}B_{\mu,j_\ell p}^{*[1:\ell-1]}(\tau)\Big|d\tau. \tag{133}
$$

Variable $X_{\mu,p}^{(0)}$ appears in each $\boldsymbol{\Phi}_\mu^{*(\cdot)}$ inside $\mathbf{B}_\mu^{*[\ell+1:L]}$, so the derivative on the r.h.s. will be the sum of $L-\ell+1$ terms. The generic k-th term of this sum is of form

$$
\Big(\underbrace{\frac{\mathbf{a}^{*\mathsf{T}}}{\sqrt{d_L}}\boldsymbol{\Phi}_\mu^{*(L-1)}\frac{\mathbf{W}^{*(L)}}{\sqrt{d_{L-1}}}\boldsymbol{\Phi}_\mu^{*(L-2)}\dots\frac{\mathbf{W}^{*(k+2)}}{\sqrt{d_{k+1}}}}_{\mathbf{e}}\frac{\partial\boldsymbol{\Phi}_\mu^{*(k)}}{\partial X_{p\mu}^{(0)}}\underbrace{\frac{\mathbf{W}^{*(k+1)}}{\sqrt{d_k}}\dots\boldsymbol{\Phi}_\mu^{*(\ell-1)}}_{\mathbf{E}}\Big)_{i_\ell}
$$
$$
=\sum_{i=1}^{d_{k+1}}\mathbf{e}_i(\bar{\mathbf{B}}_\mu^{*[1:k+1]})_{ip}\mathbf{E}_{ii_\ell}.
$$

Here, $\bar{\mathbf{B}}_\mu^*$ refers to the same expression as $\mathbf{B}_\mu^*$, where we take φ'' instead of φ'. Since the number of such terms is finite ($L-\ell+1$) we can focus only on one $k\in[\ell-1,\dots,L-1]$, as the others are bounded in the same way (we denote them with $\dots$ in the following). Plugging the previous equation into (133) we have

$$
|I_3| \le \frac{C(f)}{\sqrt{d_{\ell-1}}}\mathbb{E}_{\psi,\tilde{\mathbf{X}}_\mu}\int_0^{\pi/2}\Big|\sum_{i=1}^{d_{k+1}}\mathbf{e}_i(\bar{\mathbf{B}}_\mu^{*[1:k+1]}\mathbf{B}_\mu^{*[1:\ell-1]\mathsf{T}}(\tau))_{ij_\ell}\mathbf{E}_{ii_\ell}\Big|d\tau+\dots
$$
$$
=\frac{C(f)}{\sqrt{d_{\ell-1}}}\mathbb{E}_{\psi,\tilde{\mathbf{X}}_\mu}\int_0^{\pi/2}\Big|(\mathbf{E}^\mathsf{T}\mathrm{diag}(\mathbf{e})\bar{\mathbf{B}}_\mu^{*[1:k+1]}\mathbf{B}_\mu^{*[1:\ell-1]\mathsf{T}}(\tau))_{i_\ell j_\ell}\Big|d\tau+\dots.
$$

With Jensen inequality we obtain

$$
\mathbb{E}\sum_{i_\ell j_\ell}|I_3|^2 \leq
$$

$$
\frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2}\mathrm{Tr}\Big(\mathbf{E}^\mathsf{T}\mathrm{diag}(\mathbf{e})\bar{\mathbf{B}}_\mu^{*[1:k+1]}\mathbf{B}_\mu^{*[1:\ell-1]\mathsf{T}}(\tau)\mathbf{B}_\mu^{*[1:\ell-1]}(\tau)\bar{\mathbf{B}}_\mu^{*[1:k+1]\mathsf{T}}\mathrm{diag}(\mathbf{e})\mathbf{E}\Big)d\tau + \dots
$$

$$
\leq \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2}\|\mathbf{E}\|^2\|\bar{\mathbf{B}}_\mu^{*[1:k+1]}\mathbf{B}_\mu^{*[1:\ell-1]\mathsf{T}}(\tau)\|^2\mathrm{Tr}\,\mathrm{diag}(\mathbf{e})^2 d\tau + \dots
$$

$$
= \frac{C(f)}{d_{\ell-1}}\mathbb{E}\int_0^{\pi/2}\|\mathbf{E}\|^2\|\bar{\mathbf{B}}_\mu^{*[1:k+1]}\mathbf{B}_\mu^{*[1:l-1]\mathsf{T}}(\tau)\|^2\|\mathbf{e}\|^2 d\tau + \dots.
$$

To come to this conclusion we used several times the fact that for a symmetric semidefinite positive matrix $\mathbf{A}$ and any matrix $\mathbf{B}$ we have $|\mathrm{Tr}(\mathbf{AB})| \leq \|\mathbf{B}\|\mathrm{Tr}\mathbf{A}$. Now we can use Cauchy-Swartz and notice that moments of $\|\mathbf{\Phi}_\mu^{*(\cdot)}\|$, $\|\frac{\mathbf{W}^{*(\cdot)}}{\sqrt{d_\cdot}}\|$, and $\|\frac{\mathbf{a}^{*\mathsf{T}}}{\sqrt{d_L}}\|$ are all of order $O(1)$. Which gives $\mathbb{E}\sum_{i_\ell j_\ell}|I_3|^2 = O(d_{\ell-1}^{-1})$.

To conclude the proof of the Lemma we need to deal with the derivative with respect to $\mathbf{W}^{*(1)}$. This terms are very similar to what we had before

$$
\frac{\partial g}{\partial W_{i_1 j_1}^{*(1)}} = \mathbb{E}_{\psi_1,\psi_2}\Big[\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)\Big(\frac{\sqrt{1-t}(\mathbf{a}^{*\mathsf{T}}\mathbf{B}_\mu^{*[2:L]}\mathbf{\Phi}_\mu^{*(0)})_{i_1}X_{j_1\mu}^{(0)}}{\sqrt{d_L d_0}}
$$

$$
+ \frac{\sqrt{t\rho_L}(\mathbf{v}^{*\mathsf{T}}\mathbf{B}_\mu^{*[2:L-1]}\mathbf{\Phi}_\mu^{*(0)})_{i_1}X_{j_1\mu}^{(0)}}{\sqrt{d_{L-1}d_0}}\Big)\Big].
$$

The only difference is that now we have a factor $X_{j_1\mu}^{(0)}$ which is Gaussian itself and allows us to integrate by parts without introducing the circular interpolation. The calculations themselves will be identical.

Concentration with respect to $\mathbf{a}^*$ and $\mathbf{v}^*$. Derivatives of g with respect to these two vectors are almost identical, so we demonstrate here only the part with $\frac{\partial g}{\partial a_i^*}$.

$$
\frac{\partial g}{\partial a_i^*} = \frac{1}{n}\sum_{\mu=1}^n\mathbb{E}_{\psi_1,\psi_2}\Big[\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)\frac{\sqrt{1-t}X_{i\mu}^{(L)}}{\sqrt{d_L}}\Big] = \mathbb{E}_{\psi_1,\psi_2}\Big[\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)\frac{\sqrt{1-t}X_{i\mu}^{(L)}}{\sqrt{d_L}}\Big].
$$

Following the lines of the first part of this Lemma we introduce the circular interpolation (125)-(126) and rewrite the previous expression as

$$
\frac{\partial g}{\partial a_i^*} = \mathbb{E}_{\psi_1,\psi_2}\int_0^{\pi/2}\Big[\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)\frac{\sqrt{1-t}\sum_{p=1}^{d_0}B_{ip}^{*[1:L]}(\tau)\dot{X}_{p\mu}^{(0)}(\tau)}{\sqrt{d_L}}\Big]d\tau. \tag{134}
$$

Proceeding with the integration by parts with respect to $\dot{X}_{\mu,p}^{(0)}(\tau)$ we obtain

$$
\frac{\partial g}{\partial a_i^*} = \mathbb{E}_{\psi_1,\psi_2}\int_0^{\pi/2}\sum_{p=1}^{d_0}\cos\tau\Big[\Big\langle\frac{P_\mu^y}{P_\mu}\Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)\Big]'_{X_{p\mu}^{(0)}}\frac{\sqrt{1-t}B_{\mu,ip}^{*[1:L]}(\tau)}{\sqrt{d_L}}d\tau = I_1 + I_2, \tag{135}
$$

where we again denote

$$
I_1 = \mathbb{E}_{\psi_1,\psi_2,\tilde{\mathbf{X}}_\mu^{(0)}} \int_0^{\pi/2} \cos\tau \sum_{p=1}^{d_0} \Big(- \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle^2 f'(S_{t\mu};\mathbf{A}_\mu)^2 + \Big\langle \frac{P_\mu^{yy}}{P_\mu} \Big\rangle f'(S_{t\mu};\mathbf{A}_\mu)^2
$$

$$
+ \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle f''(S_{t\mu};\mathbf{A}_\mu) \Big) \frac{\partial S_{t\mu}}{\partial X_{p\mu}^{(0)}} \frac{\sqrt{1-t}B_{\mu,ip}^{*[1:L]}(\tau)}{\sqrt{d_L}},
$$

$$
I_2 = \mathbb{E}_{\psi_1,\psi_2,\tilde{\mathbf{x}}_\mu} \int_0^{\pi/2} \cos\tau \sum_{p=1}^{d_0} \Big(- \Big\langle \frac{P_\mu^y}{P_\mu} \Big\rangle \Big\langle \frac{P_\mu^x}{P_\mu} \frac{\partial s_{t\mu}}{\partial X_{p\mu}^{(0)}} \Big\rangle + \Big\langle \frac{P_\mu^{yx}}{P_\mu} \frac{\partial s_{t\mu}}{\partial X_{p\mu}^{(0)}} \Big\rangle \Big)
$$

$$
\times f'(S_{t\mu};\mathbf{A}_\mu) \frac{\sqrt{1-t}B_{\mu,ip}^{*[1:L]}(\tau)}{\sqrt{d_L}} \ .
$$

Using (119) to calculate derivative of S_μ and bounding ratios of P_μ with Lemma 11, we obtain

$$
|I_1| \le \frac{C(f)}{\sqrt{d_L}} \mathbb{E}_{\psi_1,\psi_2,\tilde{\mathbf{X}}_\mu} \int_0^{\pi/2} \Big(\frac{\sqrt{1-t}\mathbf{a}^{*\mathsf{T}}}{\sqrt{d_L}} \mathbf{B}_\mu^{*[1:L]} \mathbf{B}_\mu^{*[1:L]\mathsf{T}}(\tau) + \frac{\sqrt{t}\rho_L \mathbf{v}^{*\mathsf{T}}}{\sqrt{d_{L-1}}} \mathbf{B}_\mu^{*[1:L-1]} \mathbf{B}_\mu^{*[1:L]\mathsf{T}}(\tau) \Big)_i.
$$

Taking the sum over i of squared latter expression we obtain

$$
\sum_i |I_1|^2 \le \frac{C(f,t)}{d_L} \mathbb{E} \int_0^{\pi/2} d\tau \Big(\Big\| \frac{\mathbf{a}^{*\mathsf{T}}}{\sqrt{d_L}} \mathbf{B}_\mu^{*[1:L]} \mathbf{B}_\mu^{*[1:L]}(\tau) \Big\|^2 + \Big\| \frac{\mathbf{v}^{*\mathsf{T}}}{\sqrt{d_{L-1}}} \mathbf{B}_\mu^{*[1:L-1]} \mathbf{B}_\mu^{*[1:L]}(\tau) \Big\|^2 \Big).
$$

Again, since $\mathbb{E}\|\mathbf{a}^*\|^2/d_L$ and $\mathbb{E}\|\mathbf{B}_\mu^{*[\ell_1:\ell_2]}(\tau)\|$ are all of order $O(1)$, we obtain that $\sum_i |I_1|^2 \le C(f,\varphi)/d_L$. Term I_2 is handled very similar to (130), resulting in the bound $\sum_i |I_2|^2 \le C(f,\varphi)/d_L$, which concludes the proof. ∎