Journal of Machine Learning Research 26 (2025) 1-35 Submitted 7/24; Revised 3/25; Published 4/25

Feature Learning in Finite-Width Bayesian Deep Linear Networks with Multiple Outputs and Convolutional Layers

Federico Bassetti FEDERICO.BASSETTI@POLIMI.IT
Dipartimento di Matematica
Politecnico di Milano
Piazza Leonardo da Vinci 31, 20133 Milan, Italy

Marco Gherardi MARCO.GHERARDI@UNIMI.IT
Dipartimento di Fisica
Università degli Studi di Milano
Via Celoria 16, 20133 Milan, Italy;
Istituto Nazionale di Fisica Nucleare – Sezione di Milano
Via Celoria 16, 20133 Milan, Italy

Alessandro Ingrosso ALESSANDRO.INGROSSO@DONDERS.RU.NL
Donders Institute for Brain, Cognition and Behaviour
Radboud University
Nijmegen, The Netherlands

Mauro Pastore MPASTORE@ICTP.IT
Laboratoire de Physique
École Normale Supérieure, CNRS, PSL University, Sorbonne University, Université Paris-Cité
24 rue Lhomond, 75005 Paris, France;
Quantitative Life Sciences
The Abdus Salam International Centre for Theoretical Physics
Strada Costiera 11, 34151 Trieste, Italy

Pietro Rotondo PIETRO.ROTONDO@UNIPR.IT
Dipartimento di Scienze Matematiche, Fisiche e Informatiche
Università degli Studi di Parma
Parco Area delle Scienze 7/A, 43124 Parma, Italy

Editor: Maxim Raginsky

Abstract

Deep linear networks have been extensively studied, as they provide simplified models of deep learning. However, little is known in the case of finite-width architectures with multiple outputs and convolutional layers. In this manuscript, we provide rigorous results for the statistics of functions implemented by the aforementioned class of networks, thus moving closer to a complete characterization of feature learning in the Bayesian setting. Our results include: (i) an exact and elementary non-asymptotic integral representation for the joint prior distribution over the outputs, given in terms of a mixture of Gaussians; (ii) an analytical formula for the posterior distribution in the case of squared error loss function (Gaussian likelihood); (iii) a quantitative description of the feature learning infinite-width regime, using large deviation theory. From a physical perspective, deep architectures with multiple outputs or convolutional layers represent different manifestations of kernel shape renormalization, and our work provides a dictionary that translates this physics intuition and terminology into rigorous Bayesian statistics.

Keywords: deep learning theory, Bayesian deep linear networks, convolutional layers, feature learning, Gaussian mixtures

1. Introduction

Deep learning makes use of a wide array of architectures, each with unique strengths and applications. Selecting the optimal architecture for a specific task requires a deep understanding of their properties and a principled way to compute them from theory. Unfortunately, the empirical advancements in this field risk outpacing the development of a comprehensive theoretical framework.

Significant progress has been made in certain asymptotic regimes, where neural networks enjoy forms of universality that simplify their analysis. In the (lazy-training) infinite-width limit, in particular, universality is manifested in both the training under gradient flow, described by the neural tangent kernel (Jacot et al., 2018), and the Bayesian inference setting (Lee et al., 2018), where exact relations between neural networks and kernel methods were obtained. However, these results cannot explain the success of modern deep architectures, which perform non-trivial feature selection beyond what is possible in the lazy-training regime (Chizat et al., 2019; Lewkowycz et al., 2021; Brown et al., 2020). One can circumvent this problem by considering different rescalings of the weights with the width of the layers, using the so-called mean field (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022; Sirignano and Spiliopoulos, 2020; Chizat and Bach, 2018) or the more recent maximal update parametrizations (Yang and Hu, 2021).

Despite these insights, neural networks at finite width, which are relevant for applications, display complex and intriguing phenomena that are not captured by infinite-width asymptotics (Ciceri et al., 2024; Ingrosso and Goldt, 2022; Petrini et al., 2023b,a; Sclocchi and Wyart, 2024; Sclocchi et al., 2023). Addressing these finite-size properties calls for a non-perturbative theory, one which does not rely on large-size limits. Developing such a theory for general neural networks has proven to be quite challenging. However, progress has been made by considering simplified models where the activation functions are linear. Deep linear networks (Saxe et al., 2014, 2019; Li and Sompolinsky, 2021) are linear in terms of their inputs but they retain non-linear dependence on the parameters. They provide analytically tractable non-convex problems in parameter space, thus serving as a bridge between idealized infinite-width models and the reality of finite-sized networks.

Here, we address the computation of network statistics in the Bayesian learning setting, equivalent to the statistical mechanics framework employed by physicists (Neal, 1996; Lee et al., 2018; Novak et al., 2019; Garriga-Alonso et al., 2019; Engel and Van den Broeck, 2001). In Bayesian deep learning, the prior over the network's parameters induces a prior over the network's outputs (more details are in Sec. 2.3). In the (lazy-training) infinite-width limit, and for non-linear activation function, the prior over the outputs is known to be Gaussian both for fully connected architectures (Neal, 1996; de G. Matthews et al., 2018; Hanin, 2023) and for convolutional architectures (Novak et al., 2019; Garriga-Alonso et al., 2019). For fully connected networks, precise rates of convergence to normality have been investigated as well (see Favaro et al., 2025; Trevisan, 2023). The goal of our work is to characterize the non-Gaussian behavior, at finite depth and width, of two classes of deep linear networks: (i) those with fully-connected layers and multiple outputs, and (ii) those with convolutional layers and a single linear readout. Our results, summarized in the next section, provide a way to quantify non-perturbative feature learning effects in these architectures. We will discuss the link between our work and the literature by physicists in Sec. 4.2.

1.1 Informal statement of the results

We summarize here our main results informally, as three take-home messages.

Take-home message 1: At finite width, the output prior is an exactly computable mixture of Gaussians. The sizes of the hidden layers appear as parameters in the mixing measure, leading to dimensional reduction.

We compute the prior over the outputs at finite-width in the linear case, showing that it is a mixture of Gaussians with an explicit mixing distribution. Notably, in this representation the covariance, i.e. the neural network Gaussian process (NNGP) kernel, is modified by L (the number of hidden layers) random matrices with Wishart distribution. For fully-connected linear networks with finite number D of outputs in the readout layer, the dimension of these matrices is D (**Proposition 1**), while for convolutional linear networks (with unitary stride), the dimension of the Wishart matrices is the (fixed) size of the input N_0 (**Proposition 5**). All the sizes of the hidden layers N_ℓ appear parametrically in the prior, providing an explicit dimensional reduction.

Take-home message 2: At finite width the posterior predictive is a mixture of Gaussians with closed form mixing distribution.

In the case of quadratic loss function (Gaussian likelihood), the posterior distribution inherits the properties of the prior. The posterior is again a mixture of Gaussians and this leads to the rather standard equations for the bias and variance of a Gaussian Process (Rasmussen and Williams, 2005), with the important difference that they are now random variables (**Proposition 6 and 7)**.

Take-home message 3: In the feature learning infinite-width limit, large deviation asymptotics shows non-trivial explicit dependence on the training inputs and labels.

Using the simple parametric dependence of our formulas from the layer widths, we provide an asymptotic analysis in the limit of large width, using the language of Large Deviation Theory. First, we recover the well-known infinite-width limit, showing that the prior is degenerate in this case, since the Wishart ensembles concentrate around the identity matrix (**Proposition 8**).

Second, we consider the so-called feature learning infinite-width limit (Chizat et al., 2019; Mei et al., 2018; Geiger et al., 2021, 2020; Yang and Hu, 2021), which is equivalent to a different re-parametrization of the loss and of the output function, and we precisely show how it provides a way to escape lazy training. This so-called mean field parametrization was initially investigated in deep networks trained using gradient descent (Bordelon and Pehlevan, 2022), but it has been very recently considered also in the Bayesian setting (Rubin et al., 2024; Lauditi et al., 2025).

The measure over the Wishart ensembles concentrates also in this case, but this time they do so around non-trivial solutions that explicitly depend on the training inputs and labels (**Proposition 10**).

1.2 Related work

Saxe et al. (2014, 2019) derived exact solutions for the training dynamics of deep linear networks (DLN). More recently, Li and Sompolinsky (2021) studied deep linear networks in a Bayesian framework, via the equivalent statistical mechanics formulation, and obtained analytical results in the proportional limit, where the common size of the hidden layers N and the size of the training set P are taken to infinity, while keeping their ratio $\alpha = P/N$ fixed. This analysis was later generalized to globally gated DLNs (Li and Sompolinsky, 2022), and extended to the data-averaged case, using the replica method from spin glass theory (Zavatone-Veth et al., 2022).

Hanin and Zlokapa (2023) reconsidered the same setting in the case of single-output networks, and they derived a non-asymptotic result for the Bayesian model evidence (partition function in statistical mechanics), in terms of Meijer G-functions (the fact that the prior over a single output is related to these special functions was shown in Zavatone-Veth and Pehlevan, 2021). They employed this result to investigate various asymptotic limits where N, P, and the depth of the network L are simultaneously taken to infinity at the same rate.

We provide an explicit mapping between our present results on one side, and both the statistical mechanics approach of Li and Sompolinsky (2021) and the Meijer G-functions formalism of Hanin and Zlokapa (2023); Zavatone-Veth and Pehlevan (2021) on the other side, in Section 4.

Closely related to the deep linear literature, Pacelli et al. (2023) proposed an effective theory for deep non-linear networks in the proportional limit, whose derivation is based on a Gaussian equivalence (Goldt et al., 2020, 2022; Gerace et al., 2020; Hu and Lu, 2023; Gerace et al., 2024; Aguirre-López et al., 2025) heuristically justified using Breuer-Major theorems (Breuer and Major, 1983; Bardet and Surgailis, 2013; Nourdin et al., 2011). This effective theory has been tested with great accuracy for one-hidden-layer fully-connected networks with a single output in Baglioni et al., 2024. Camilli et al., 2023 and Cui et al., 2023 reconsidered the same setting in the Bayes optimal and data-averaged case. Other non-perturbative approaches to Bayesian deep non-linear networks include Naveh and Ringel (2021); Seroussi et al. (2023); Fischer et al. (2024) and the recent renormalization group framework proposed in Howard et al. (2024, 2025).

A further line of research considers linear (or weakly nonlinear) networks in the limit in which both depth N and width L diverge, while their ratio L/N converges to a positive constant; see Hanin (2024) and Li et al. (2022). In particular, Li et al. (2022) finds that the conditional covariance matrix in the infinite-depth-and-width limit satisfies a stochastic differential equation for covariance matrices of dimension $P \times P$. Since the submission of the first version of this paper, further progress has been made in this direction. Building on our representation, a complete characterization of the proportional limit $L/N \to \alpha > 0$ for deep linear networks is now available in the recent work Bassetti et al. (2024). This characterization has been made possible because our representation provides a fundamental dimensional reduction, which is not captured by the usual representation obtained by conditioning on the penultimate layer.

1.3 Kernel renormalization in the proportional limit

Bayesian networks in the infinite-width limit ($N \to \infty$ at finite P) are equivalent to Gaussian processes (NNGP), whose covariance is given by a kernel (Lee et al., 2018). A powerful view of Bayesian networks in the proportional limit (Li and Sompolinsky, 2021; Pacelli et al., 2023; Li and Sompolinsky, 2022) focuses on how the prior is modified layer by layer. For fully-connected architectures with a single output, the NNGP kernel at each layer is rescaled by a real number (a scalar order parameter in statistical mechanics). The order parameters satisfy a set of coupled equations that explicitly depend on the inputs and labels in the training set. This process is referred to as a scalar renormalization of the kernel. In fully-connected networks with multiple outputs, instead, the order parameters are $D \times D$ matrices (D being the size of the readout layer), and they modify the NNGP kernel via a Kronecker product. This process is called kernel shape renormalization (Li and Sompolinsky, 2021; Pacelli et al., 2023). Yet another transformation of the kernel, coined local kernel renormalization, was found for convolutional networks in Aiudi et al. (2025), where the NNGP kernel splits into several local components (Novak et al., 2019), labelled by two additional indices i, j that run across the patches in each element of the training set (e.g., the receptive fields of an image). In the proportional limit, each local component (i, j) of the kernel is modified by an order parameter Q_{ij}. An interesting type of kernel shape renormalization was found in globally gated deep linear networks too (Li and Sompolinsky, 2022), but the interpretation is different than in the convolutional case presented in Aiudi et al. (2025).

Common definitions

P	size of the training set
L	depth of the network
N_0	input dimension
θ	collection of all trainable weights, $\theta = \{W^{(\ell)}\}_{\ell=0}^{L}$
$\mathbf{x}^\mu$, X	input of the network in the training set, $X = [\mathbf{x}^\mu]_{\mu=1}^{P}$
$\mathbf{y}^\mu$, $\mathbf{y}_{1:P}$	label (response) of $\mathbf{x}^\mu$ in the training set, $\mathbf{y}_{1:P}^\top = (\mathbf{y}^{1\top}, \dots, \mathbf{y}^{P\top})$
$f_\theta(\mathbf{x}^\mu)$	output of the network from input $\mathbf{x}^\mu$

Fully-connected

$N_1, \cdots, N_L$	width of each hidden layer
$N_{L+1} = D$	output dimension
$\mathbf{S}^\mu$, $\mathbf{S}_{1:P}$	output of the network in $\mathbb{R}^D$, $\mathbf{S}^\mu = f_\theta(\mathbf{x}^\mu)$, $\mathbf{S}_{1:P}^\top = (\mathbf{S}^{1\top}, \dots, \mathbf{S}^{P\top})$

Convolutional

$C_0, \cdots, C_L$	number of channels in each hidden layer
M	dimension of the channel mask
S^μ, $\mathbf{S}_{1:P}$	output of the network in $\mathbb{R}$, $S^\mu = f_\theta(\mathbf{x}^\mu)$, $\mathbf{S}_{1:P}^\top = (S^1, \dots, S^P)$

Table 1: Table of notations

2. Problem setting

We consider a supervised learning problem with training set $\{\mathbf{x}^\mu, \mathbf{y}^\mu\}_{\mu=1}^P$, where each $\mathbf{x}^\mu \in \mathbb{R}^{N_0 \times C_0}$ and the corresponding labels (response) $\mathbf{y}^\mu \in \mathbb{R}^D$, with $C_0 = 1$ for the fully-connected architecture and, for simplicity, $D = 1$ for the convolutional architecture. The output will be denoted by $f_\theta(\mathbf{x}^\mu)$, where θ represents the collection of all the trainable weights of the network (see Table 1 for a summary of notations).

2.1 Fully-connected deep linear neural networks (FC-DLNs)

In fully-connected neural networks, the pre-activations of each layer $h_{i_\ell}^{(\ell)}$ ($i_\ell = 1, \dots, N_\ell$; $\ell = 1, \dots, L$) are given recursively as a function of the pre-activations of the previous layer $h_{i_{\ell-1}}^{(\ell-1)}$ ($i_{\ell-1} = 1, \dots, N_{\ell-1}$):

$$\begin{aligned} h_{i_1}^{(1)}(\mathbf{x}^\mu) &= \frac{1}{\sqrt{N_0}} \sum_{i_0=1}^{N_0} W_{i_1 i_0}^{(0)} x_{i_0}^\mu + b_{i_1}^{(0)} \,, \\ h_{i_\ell}^{(\ell)}(\mathbf{x}^\mu) &= \frac{1}{\sqrt{N_{\ell-1}}} \sum_{i_{\ell-1}=1}^{N_{\ell-1}} W_{i_\ell i_{\ell-1}}^{(\ell-1)} h_{i_{\ell-1}}^{(\ell-1)}(\mathbf{x}^\mu) + b_{i_\ell}^{(\ell-1)} \,, \end{aligned} \tag{1}$$

where $W^{(\ell-1)}$ and $b^{(\ell-1)}$ are respectively the weights and the biases of the ℓ-th layer, whereas the input layer has dimension N_0 (the input data dimension). In what follows we shall take $b_{i_\ell}^{(\ell-1)} = 0$ for every ℓ and every i_ℓ (without loss of generality, as one can include this contribution in the vectors of weights adding additional dimensions in input and hidden space).

We add one last readout layer with dimension $N_{L+1} = D$, and define the function implemented by the deep neural network as $f_\theta(\mathbf{x}^\mu) = (S_1^\mu, \dots, S_D^\mu)^\top$, where

$$S_i^\mu = \frac{1}{\sqrt{N_L}} \sum_{i_L=1}^{N_L} W_{i,i_L}^{(L)} h_{i_L}^{(L)}(\mathbf{x}^\mu) \qquad i = 1, \dots, D\,;\ \mu = 1, \dots, P, \tag{2}$$

$W_{i,i_L}^{(L)}$ being the weights of the last layer. The input training set is collected in a $N_0 \times P$ matrix X, that is

$$X = [\mathbf{x}^1, \dots, \mathbf{x}^P],$$

and the corresponding labels in a vector $\mathbf{y}_{1:P}^\top = (\mathbf{y}_1^\top, \dots, \mathbf{y}_P^\top)$. We also denote by

$$\mathbf{S}_{1:P} = (S_1^1, \dots, S_D^1, S_1^2, \dots, S_1^P, \dots, S_D^P)^\top$$

the P outputs ($\mathbf{S}^\mu = f_\theta(\mathbf{x}^\mu) : \mu = 1, \dots, P$) stacked in a vector.

2.2 Convolutional deep linear neural networks (C-DLNs)

For convolutional neural networks, the pre-activations at each layer $\ell = 1, \dots, L$ are labelled by two indices: (i) the channel index $a_\ell = 1, \dots, C_\ell$, where C_ℓ is the total number of channels in each layer; (ii) the spatial index $i = 1, \dots, N_0$, which runs over the input coordinates.

Similarly any input has two indices, that is $\mathbf{x}^\mu = [x^\mu_{a_0,i}]$, with $a_0 = 1, \ldots, C_0$ and $i-1, \ldots, N_0$. Pre-activations of the first layer are given by

$$h^{(1)}_{a_1,i}(\mathbf{x}^\mu) = \frac{1}{\sqrt{MC_0}} \sum_{a_0=1}^{C_0} \sum_{m=-\lfloor M/2 \rfloor}^{\lfloor M/2 \rfloor} W^{(0)}_{m,a_1 a_0} x^\mu_{a_0, i+m (\mathrm{mod} N_0)} \tag{3}$$

and for $\ell > 1$

$$h^{(\ell)}_{a_\ell,i}(\mathbf{x}^\mu) = \frac{1}{\sqrt{MC_{\ell-1}}} \sum_{a_{\ell-1}=1}^{C_{\ell-1}} \sum_{m=-\lfloor M/2 \rfloor}^{\lfloor M/2 \rfloor} W^{(\ell-1)}_{m,a_\ell a_{\ell-1}} h^{(\ell-1)}_{a_{\ell-1}, i+m (\mathrm{mod} N_0)}(\mathbf{x}^\mu)\,, \tag{4}$$

where M is the dimension of the channel mask. Here we restrict our analysis to $1d$ convolutions, periodic boundary conditions over the spatial index and unitary stride. This choice allows to keep the size of the spatial index constant through layers. Note that in the case of non-unitary stride, one should introduce a different spatial index i_ℓ for each layer, which runs from 1 to the integer part of the ratio between N_0 and the stride's ℓth power. The output of the network is given by

$$S^\mu = f_\theta(\mathbf{x}^\mu) = \frac{1}{\sqrt{C_L N_0}} \sum_{a_L=1}^{C_L} \sum_{i=1}^{N_0} W^{(L)}_{a_L i} h^{(L)}_{a_L,i}(\mathbf{x}^\mu)\,. \tag{5}$$

Again we denote by $\mathbf{S}_{1:P}$ the vector of outputs $(S^1, \ldots, S^P)$ and the data is now collected in a three-dimensional array $X = [x^\mu_{a_0,i}]_{\mu,a_0,i}$.

2.3 Bayesian setting

In a Bayesian neural network, a prior for the weights θ is specified, which translates in a prior for $f_\theta(\mathbf{x}^\mu)$. In other words, the prior (over the outputs) is the density of the random vector $\mathbf{S}_{1:P}$. This density evaluated for a given value $\mathbf{s}_{1:P}$ taken by $\mathbf{S}_{1:P}$, will be denoted by $p_{\mathrm{prior}}(\mathbf{s}_{1:P}|X)$. At this stage, one needs also a likelihood for the labels given the inputs and the outputs, denoted by $\mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}, \beta)$, where β is an additional parameter, which plays the role of the inverse temperature in statistical physics. The posterior density of $\mathbf{S}_{1:P}$ given $\mathbf{y}_{1:P}$ is (by Bayes theorem)

$$p_{\mathrm{post}}(\mathbf{s}_{1:P}|\mathbf{y}_{1:P}, X) \propto p_{\mathrm{prior}}(\mathbf{s}_{1:P}|X)\mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}).$$

In order to describe the posterior predictive, let $\mathbf{x}^0$ be a new input and set $\mathbf{S}^0 = f_\theta(\mathbf{x}^0)$. The posterior predictive density of $\mathbf{S}_0$ given $\mathbf{y}_{1:P}$ is obtained similarly by

$$p_{\mathrm{pred}}(\mathbf{s}_0|\mathbf{y}_{1:P}, \mathbf{x}^0, X) := Z^{-1}_{\mathbf{y}_{1:P},\beta} \int_{\mathbb{R}^{D\times P}} \mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}) p_{\mathrm{prior}}(\mathbf{s}_0, \mathbf{s}_{1:P}|\mathbf{x}^0, X) d\mathbf{s}_{1:P}$$

where $p_{\mathrm{prior}}(\mathbf{s}_0, \mathbf{s}_{1:P}|\mathbf{x}^0, X)$ is the prior for $(\mathbf{S}_0, \mathbf{S}_{1:P})$ and

$$Z_{\mathbf{y}_{1:P},\beta} = \int_{\mathbb{R}^{D(P+1)}} \mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}) p_{\mathrm{prior}}(\mathbf{s}_0, \mathbf{s}_{1:P}|\mathbf{x}^0, X) d\mathbf{s}_0 d\mathbf{s}_{1:P} \tag{6}$$

is the partition function (normalization constant). In the convolutional setting we have $D = 1$ and, consequently, we shall write s_0 in place of $\mathbf{s}_0$.

Note that, training the network with a quadratic loss function is equivalent to consider a Gaussian likelihood in the Bayesian setting, that is

$$\mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}, \beta) \propto e^{-\frac{\beta}{2}\|\mathbf{s}_{1:P}-\mathbf{y}_{1:P}\|^2}.$$

Note that in this case

$$\mathbf{y}^\mu = \mathbf{S}^\mu + \boldsymbol{\epsilon}^\mu = f_\theta(\mathbf{x}^\mu) + \boldsymbol{\epsilon}^\mu \quad \boldsymbol{\epsilon}^\mu \overset{iid}{\sim} \mathcal{N}(0, \beta^{-1}\mathbb{1}_D).$$

Here, we assume that the weights are independent normally distributed with zero means and variance $1/\lambda_\ell$ at layer ℓ. In summary:

- For the fully connected network (FC-DLN):

$$W_{ij}^{(\ell)} \overset{iid}{\sim} \mathcal{N}(0, \lambda_\ell^{-1}) \tag{7}$$

- For the convolutional network (C-DLN):

$$W_{m,ab}^{(\ell)} \overset{iid}{\sim} \mathcal{N}(0, \lambda_\ell^{-1})\,, \qquad W_{ai}^{(L)} \overset{iid}{\sim} \mathcal{N}(0, \lambda_L^{-1})\,. \tag{8}$$

3. Results

Our main results are collected in this section.

3.1 Exact non-asymptotic integral representation for the prior over the outputs

We start discussing the FC-DLN. To state the results, denote by $\mathcal{S}_D^+$ the set of symmetric positive definite matrices on $\mathbb{R}^D$ and recall that a random matrix Q taking values in $\mathcal{S}_D^+$ has Wishart distribution with $N > D$ degrees of freedom and scale matrix V if it has the following density (with respect to the Lebesgue measure on the cone of symmetric positive definite matrices)

$$\mathcal{W}_D(Q|V, N) = \frac{\det(Q)^{\frac{N-D-1}{2}} e^{-\frac{1}{2}\operatorname{tr}(V^{-1}Q)}}{\det(V)^{N/2} 2^{DN/2} (\pi)^{D(D-1)/4} \prod_{k=1}^D \Gamma\left(\frac{N-D+k}{2}\right)}.$$

Equivalently, Q has a Wishart distribution with N degrees of freedom and scale matrix V if its law corresponds to the law of $\sum_{i=1}^N \mathbf{Z}_i \mathbf{Z}_i^\top$ where $\mathbf{Z}_i$ are independent Gaussian vectors in $\mathbb{R}^D$ with zero mean and covariance matrix V (see Eaton, 2007 and the references therein). If $D = 1$, one has that $\mathcal{S}_D^+ = \mathbb{R}^+$ and the Wishart distribution reduces to a gamma distribution of parameters $\alpha = \frac{N}{2}$, $\beta = \frac{1}{2V}$, that is

$$\mathcal{W}_1(Q|V, N) = \frac{1}{(2V)^{N/2}\Gamma(N/2)} Q^{\frac{N}{2}-1} e^{-\frac{1}{2V}Q} \quad \text{for } Q > 0.$$

In our integral representations a fundamental role is played by the joint distribution of the vector of matrices

$$(Q_1, \dots, Q_L) := (\tilde{Q}_1/N_1, \dots, \tilde{Q}_L/N_L),$$

where the $\tilde{Q}_\ell$'s are $D \times D$ independent Wishart random matrices with N_ℓ degree of freedoms and identity scale matrix $\mathbb{1}_D$. This joint distribution of $(Q_1, \dots, Q_L)$ is

$$\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L) := \left(\prod_{\ell=1}^{L} \mathcal{W}_D\Big(Q_\ell \Big| \frac{1}{N_\ell}\mathbb{1}_D, N_\ell\Big) \right) dQ_1 \dots dQ_L \tag{9}$$

For the sake of notation, we write $\mathcal{Q}_{L,N}$ in place of the more correct $\mathcal{Q}_{L,N_1,\dots,N_L}$.

We now show that the prior distribution over $\mathbf{S}_{1:P}$ is a mixture of Gaussians where the covariance matrix is an explicit function of random Wishart matrices. In what follows $A \otimes B$ denotes the Kronecker product of matrices A and B. Moreover, we set $\lambda^* := \lambda_0 \dots \lambda_L$.

Proposition 1 (non-asymptotic integral representation for the prior of FC-DLNs with multiple outputs) *Let $\hat{p}_{\text{prior}}$ be the characteristic function of the outputs $\mathbf{S}_{1:P}$ of a FC-DLN, that is $\hat{p}_{\text{prior}}(\bar{\mathbf{s}}_{1:P}|X) = \mathbf{E}[\exp\{i\bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}\}]$. If $\min(N_\ell : \ell = 1, \dots, L) > D$, then*

$$\hat{p}_{\text{prior}}(\bar{\mathbf{s}}_{1:P}|X) = \int_{(\mathcal{S}_D^+)^L} e^{-\frac{1}{2}\bar{\mathbf{s}}_{1:P}^\top \mathcal{K}_{\text{FC}}(Q_1,\dots,Q_L)\bar{\mathbf{s}}_{1:P}} \mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)$$

where $\mathcal{Q}_{L,N}$ is defined by (9), $\mathcal{K}_{\text{FC}}(Q_1, \dots, Q_L) = (N_0\lambda^*)^{-1} X^\top X \otimes Q^{(L)}(Q_1, \dots, Q_L)$, and

$$Q^{(L)} = Q^{(L)}(Q_1, \dots, Q_L) := (U_1 \dots U_L)^\top U_1 \dots U_L,$$

U_ℓ being $D \times D$ matrices such that $U_\ell^\top U_\ell = Q_\ell$.

Remark 2 *In the previous representation one can choose any decomposition of Q_ℓ in $U_\ell^\top U_\ell$. For convenience in what follows we choose the Cholesky decomposition where $U_\ell^\top = U_\ell^\top(Q_\ell)$ is a lower triangular matrix with positive diagonal entries. This $U_\ell^\top$ is called Cholesky factor (or Cholesky square root). The Cholesky square root is one to one and continuous from $\mathcal{S}_D^+$ to its image. When Q_ℓ has a Wishart distribution the law of it Cholesky square root $U_\ell^\top$ is the so-called Bartlett distribution (see Kshirsagar, 1959).*

Remark 3 *The integral representation stated above is equivalent to say that the random matrix $S = [S_d^\mu]_{d=1,\dots,D;\mu=1,\dots,P}$ admits the explicit stochastic representation*

$$S = U_L^\top \cdots U_1^\top \frac{Z}{\sqrt{N_0\lambda^*}} X \tag{10}$$

where U_ℓ are $D \times D$ independent random matrices such that $Q_\ell = U_\ell^\top U_\ell$ has a Wishart distribution with N_ℓ degrees of freedom and scale matrix $\frac{1}{N_\ell}\mathbb{1}_D$ and Z is a $D \times N_0$ matrix of independent standard normal random variables. As mentioned in the introduction, all the N_ℓ's appear parametrically in the above representation, providing the anticipated dimensional reduction. For comparison, note that by (1) one can derive another explicit explicit expression for S, namely

$$S = \frac{W^{(L)}}{\sqrt{N_L}} \cdots \frac{W^{(0)}}{\sqrt{N_0}} X \tag{11}$$

where $W^{(\ell)} = [W^{(\ell)}_{i_{\ell+1},i_\ell}]$ are $N_{\ell+1} \times N_\ell$ matrices of normal random variables of zero mean and variance $1/\lambda_\ell$. This second representation is more direct (and easier to derive) and also shows that S is a mixture of Gaussians. Nevertheless, containing the $N_{\ell+1} \times N_\ell$ matrices $W^{(\ell)}$, it does not provide any dimensional reduction.

Remark 4 *As recalled above, for $D = 1$ the U_ℓ's are independent random variables distributed as the square root of a* $\mathrm{Gamma}(N_\ell/2, N_\ell/2)$ *random variable, that is $\tilde{Q}_\ell = N_\ell {U_\ell}^2$ is a χ^2 with N_ℓ degrees of freedom. The law of a product of independent gamma random variables is derived in Springer and Thompson (1970). Using this result, when $D = 1$, one gets that the rescaled random variable $\tilde{Q}^{(L)} = \prod_{i=1}^{L}(2\tilde{Q}_\ell/N_\ell)$ has density*

$$p_{\tilde{Q}^{(L)}}(q) = \frac{1}{\prod_{\ell=1}^{L}\Gamma(N_\ell/2)} G^{L0}_{0L}\left(q \,\Big|\, \begin{matrix} - \\ \frac{N_1}{2}-1, \cdots, \frac{N_L}{2}-1 \end{matrix}\right), \tag{12}$$

where $G^{m,n}_{p,q}$ is the Meijer G-function. In the case of Gaussian likelihood, one can employ (12) *to give an alternative derivation of the non-asymptotic expression of the partition function (Bayesian model evidence) given by Hanin and Zlokapa (2023). We provide more details on this derivation in Section 4.1. Note that the link between Meijer G-functions and the deep linear networks based on gamma functions was first established in Zavatone-Veth and Pehlevan (2021) in the limited setting of individual training patterns (i.e. for the prior over a single input vector).*

While in Hanin and Zlokapa (2023) the partition function is given in terms of Meijer G-functions, here it is expressed as an $L-$dimensional integral over a set of order parameters Q_ℓ. Using our alternative representation, it is possible to compare and find some equivalence with the results (in the proportional limit $P, N_\ell \to \infty$ at fixed $\alpha_\ell = P/N_\ell$) found in Li and Sompolinsky (2021) and Pacelli et al. (2023) (see the Discussion in Section 4.2).

In order to discuss a non-asymptotic representation formula for the convolutional architecture, it is convenient to introduce the translation operator T_m as

$$T_{m,ij} = \delta_{j,i+m(\mathrm{mod} N_0)}. \tag{13}$$

In this way, Eq. (4) takes the following form

$$h^{(\ell)}_{a_\ell,i}(\mathbf{x}^\mu) = \frac{1}{\sqrt{MC_{\ell-1}}} \sum_{a_{\ell-1}=1}^{C_{\ell-1}} \sum_{m=-\lfloor M/2 \rfloor}^{\lfloor M/2 \rfloor} W^{(\ell-1)}_{m,a_\ell a_{\ell-1}} \sum_{j=1}^{N_0} T_{m,ij} h^{(\ell-1)}_{a_{\ell-1},j}(\mathbf{x}^\mu)\,. \tag{14}$$

One can now define a map $\mathcal{T} : (\mathcal{S}^+_{N_0})^L \to \mathcal{S}^+_{N_0}$ that will play a role similar to the map $Q^{(L)} = Q^{(L)}(Q_1, \ldots, Q_L)$ introduced in Proposition 1. Given $Q_1, \ldots, Q_L$, one sets $\mathcal{T}(Q_1, \ldots, Q_L) = Q^*_1$ where $Q^*_L, \ldots, Q^*_1$ are defined by the following backward recursion:

$$\begin{aligned} Q^*_L &= \frac{1}{M} \sum_{m=-\lfloor M/2 \rfloor}^{\lfloor M/2 \rfloor} T^\top_m Q_L T_m\,, \\ Q^*_\ell &= \frac{1}{M} \sum_{m=-\lfloor M/2 \rfloor}^{\lfloor M/2 \rfloor} T^\top_m \left((U^*_{\ell+1})^\top Q_\ell U^*_{\ell+1}\right) T_m \quad \text{for } \ell = L-1, \ldots, 1 \end{aligned} \tag{15}$$

$U^*_{\ell+1}$ being a square root of $Q^*_{\ell+1}$, that is $(U^*_{\ell+1})^\top U^*_{\ell+1} = Q^*_{\ell+1}$. Note that $[T_m^\top Q T_m]_{i,j} = Q_{i+m(\mathrm{mod} N_0), j+m(\mathrm{mod} N_0)}$ for any matrix Q. Finally, we need the analogous of $\mathcal{Q}_{L,N}$ for the convolutional architecture. To this end, we set

$$\mathcal{Q}_{L,C}(dQ_1 \dots dQ_L) := \left(\prod_{\ell=1}^{L} \mathcal{W}_{N_0}\Big(Q_\ell \Big| \frac{1}{C_\ell} \mathbb{1}_{N_0}, C_\ell \Big) \right) dQ_1 \dots dQ_L. \tag{16}$$

Note that $\mathcal{Q}_{L,C}$ has exactly the same form of $\mathcal{Q}_{L,N}$ once $N_1, \dots, N_L$ are replaced by $C_1, \dots, C_L$.

Proposition 5 (non-asymptotic integral representation for the prior of C-DLNs with single output) *For a C-DLN with* $\min(C_\ell : \ell = 1, \dots, L) > N_0$*, the characteristic function* $\hat{p}_{\mathrm{prior}}(\bar{\mathbf{s}}_{1:P}|X) = \mathbf{E}[\exp\{i \bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}\}]$ *of the outputs* $\mathbf{S}_{1:P}$ *is given by*

$$\hat{p}_{\mathrm{prior}}(\bar{\mathbf{s}}_{1:P}|X) = \int_{(\mathcal{S}^+_{N_0})^L} e^{-\frac{1}{2}\bar{\mathbf{s}}_{1:P}^\top (\mathcal{K}_{\mathrm{C}}(Q_1,\dots,Q_L))\bar{\mathbf{s}}_{1:P}} \mathcal{Q}_{L,C}(dQ_1 \dots dQ_L)$$

where $\mathcal{Q}_{L,N}$ *is defined by* (16)*, and the covariance matrix* $\mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)$ *is defined as*

$$\mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L) := \Big[\sum_{r,s} \mathcal{T}(Q_1, \dots, Q_L)_{rs} \sum_{a_0} \frac{x^\mu_{a_0,r} x^\nu_{a_0,s}}{\lambda^* C_0} \Big]_{\mu=1,\dots,P;\nu=1,\dots,P}. \tag{17}$$

Note that the map $\mathcal{T}$ acts as on a four indices tensor, providing a partial trace wrt to indices r, s. This tensor is known in the literature as *local covariance matrix* from the seminal work Novak et al. (2019) on the Bayesian infinite-width limit in convolutional networks. The same partial trace structure has been also identified in the proportional limit for non-linear one-hidden-layer networks in Aiudi et al. (2025), where the authors coined the term *local kernel renormalization* to indicate the re-weighting of the infinite-width local NNGP kernel found in Novak et al. (2019).

3.2 Posterior distribution and predictor's statistics in the case of squared error loss function

In the case of quadratic loss function (Gaussian likelihood), it is possible to show that the predictive posterior of $\mathbf{S}_0 = f_\theta(\mathbf{x}^0)$ is a also mixture of Gaussian distributions, where the mixture is now both on the covariance matrix and on the vector of means and it is driven by an updated version of $\mathcal{Q}_{L,N}$. In the next result we assume that the matrix $\tilde{X} = [\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^P]$ is full rank.

Proposition 6 *Let* $\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)$ *be as in Eq.* (9) *and* $Q^{(L)} = Q^{(L)}(Q_1, \dots, Q_L)$ *be defined as in* Proposition 1*. Let* $p_{\mathrm{pred}}(\mathbf{s}_0|\mathbf{y}_{1:P}, \mathbf{x}_0, X)$ *be the posterior predictive of the FC-DLN and assume that* $\det(\tilde{X}^\top \tilde{X}) > 0$*. Then for every* $\mathbf{s}_0 \in \mathbb{R}^D$ *one has*

$$\begin{aligned} &p_{\mathrm{pred}}(\mathbf{s}_0|\mathbf{y}_{1:P}, \mathbf{x}_0, X) \\ &\quad = \int_{(\mathcal{S}^+_D)^L} \frac{e^{-\frac{1}{2}(\mathbf{s}_0 - \mathbf{m}_0)^\top (\Sigma_{00} - \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_{DP})^{-1}\Sigma_{01}^\top)^{-1}(\mathbf{s}_0 - \mathbf{m}_0)}}{(2\pi)^{\frac{D}{2}} \det(\Sigma_{00} - \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_{DP})^{-1}\Sigma_{01}^\top)^{\frac{1}{2}}} \mathcal{Q}_{L,N}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta) \end{aligned}$$

where:

$$\Sigma_{00} = \frac{\mathbf{x}_0^\top \mathbf{x}_0}{N_0 \lambda^*} \otimes Q^{(L)}, \quad \Sigma_{01} = \frac{\mathbf{x}_0^\top X}{N_0 \lambda^*} \otimes Q^{(L)}, \quad \Sigma_{11} = \frac{X^\top X}{N_0 \lambda^*} \otimes Q^{(L)}$$
$$\mathbf{m}_0 = \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P},$$

$$\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta) := \frac{e^{-\frac{1}{2}\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P})}\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)}{\int_{(\mathcal{S}_D^+)^L} e^{-\frac{1}{2}\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P})}\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)}$$

and $\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P}) := \mathbf{y}_{1:P}^\top(\Sigma_{11} + \beta^{-1}\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P} + \log(\det(\mathbb{1}_{DP} + \beta\Sigma_{11}))$.

An analogous result holds for C-DLN, where now the output is in dimension 1. Here we need to define $\tilde{X} = [x^\mu_{a_0,i}]_{\mu,a_0,i}$ where now $\mu = 0, \dots, P$ and, for every $a_0 = 1, \dots, N_0$, the matrix $\tilde{X}_{a_0} = [x^\mu_{a_0,i}]_{\mu,i}$.

Proposition 7 *Let* $\mathcal{Q}_{L,C}(dQ_1 \dots dQ_L)$ *be defined in Eq.* (16) *and* $p_{\text{pred}}(s_0|\mathbf{y}_{1:P}, \mathbf{x}_0, X)$ *be the posterior predictive of the C-DLN. Assume that* $\det(\sum_{a_0=1}^{N_0} \tilde{X}_{a_0}^\top \tilde{X}_{a_0}) > 0$*. Then, for any* $s_0 \in \mathbb{R}$ *one has*

$$p_{\text{pred}}(s_0|\mathbf{y}_{1:P}, \mathbf{x}_0, X)$$
$$= \int_{(\mathcal{S}_{N_0}^+)^L} \frac{e^{-\frac{1}{2}(\Sigma_{00} - \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_P)^{-1}\Sigma_{01}^\top)^{-1}(s_0 - m_0)^2}}{(2\pi)^{\frac{1}{2}} \det(\Sigma_{00} - \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_P)^{-1}\Sigma_{01}^\top)^{\frac{1}{2}}} \mathcal{Q}_{L,C}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta)$$

where:

$$\Sigma_{00} := \sum_{r,s} \mathcal{T}(Q_1, \dots, Q_1)_{rs} \sum_{a_0} \frac{x^0_{a_0,r} x^0_{a_0,s}}{\lambda^* C_0},$$
$$\Sigma_{01} := \Big[\sum_{r,s} \mathcal{T}(Q_1, \dots, Q_1)_{rs} \sum_{a_0} \frac{x^\mu_{a_0,r} x^0_{a_0,s}}{\lambda^* C_0}\Big]_{\mu=1,\dots,P}$$
$$\Sigma_{11} := \Big[\sum_{r,s} \mathcal{T}(Q_1, \dots, Q_1)_{rs} \sum_{a_0} \frac{x^\mu_{a_0,r} x^\nu_{a_0,s}}{\lambda^* C_0}\Big]_{\mu,\nu=1,\dots,P},$$
$$m_0 := \Sigma_{01}(\Sigma_{11} + \beta^{-1}\mathbb{1}_P)^{-1}\mathbf{y}_{1:P},$$
$$\mathcal{Q}_{L,C}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta) := \frac{e^{-\frac{1}{2}\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P})}\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)}{\int_{(\mathcal{S}_{N_0}^+)^L} e^{-\frac{1}{2}\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P})}\mathcal{Q}_{L,C}(dQ_1 \dots dQ_L)}$$

and $\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P}) := \mathbf{y}_{1:P}^\top(\Sigma_{11} + \beta^{-1}\mathbb{1}_P)^{-1}\mathbf{y}_{1:P} + \log(\det(\mathbb{1}_P + \beta\Sigma_{11}))$.

3.3 Quantitative description of the feature learning infinite-width regime

In this section, we give a quantitative description of the feature learning infinite-width regime, using large deviation theory. For concreteness, we specialize to the case of FC-DLNs with multiple outputs, but analogous statements hold for C-DLNs.

Following the literature (Chizat et al., 2019; Mei et al., 2018; Geiger et al., 2021, 2020; Yang and Hu, 2021), we recall that the easiest way to escape the lazy-training infinite-width limit is to consider the so-called *mean-field parametrization*, which corresponds to the following rescaling of the loss and output functions:

$$\mathcal{L}_N(\mathbf{y}_{1:P}|\mathbf{s}_{1:P},\beta) := \mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}/\sqrt{N}, N\beta). \tag{18}$$

We point out that the mean-field parameterization exhibits some pathological behavior in the Bayesian setting. In a sense, the scale of the prior is incorrect, since it looks like a delta function in the limit. Nevertheless, one can compute the posterior of the random covariance appearing in the mixture representation, and this posterior exhibits a well-defined and non-trivial limiting behavior. Comparing the large deviation asymptotics of the mean-field posterior covariance with those in the lazy-training infinite-width limit, one recognizes the presence of additional terms, which can be interpreted as an instance of feature learning.

As starting point, we briefly consider the lazy-training infinite-width limit. Here L and D are fixed and $N_1 = \ldots, N_L = N$ and $N \to +\infty$. It is easy to see that in this case $(\tilde{Q}_1/(\lambda_1 N_1), \ldots, \tilde{Q}_L/(\lambda_L N_L))$, converges (almost surely) to $(\mathbb{1}_D, \ldots, \mathbb{1}_D)$. This is a consequence of the law of large numbers, indeed $Q_\ell = \frac{\tilde{Q}_\ell}{N} = \frac{1}{N}\sum_{j=1}^N \mathbf{Z}_{j,\ell}\mathbf{Z}_{j,\ell}^\top$ with $\mathbf{Z}_{j,\ell}$ independent standard Gaussian's vectors. Since the map $(Q_1, \ldots, Q_L) \mapsto Q^{(L)}$ in Proposition 1 is continuous, also $Q^{(L)}$ converges almost surely to $\mathbb{1}_D$. In this way, one recovers the well-known Gaussian lazy-training infinite-width limit.

For L fixed, it is easy to derive a large deviation principle (LDP) for the random covariance matrix $Q^{(L)}$ appearing in the prior representation given in Proposition 1 or, equivalently, for $\mathcal{Q}_{L,N}$.

Roughly speaking, if $V_N \in \mathbb{R}^D$ satisfies an LDP with rate function I then $P\{V_N \in dx\} = P_N(dx) \simeq e^{-NI(x)}dx$. The formal definition (see, e.g., Dembo and Zeitouni, 2010) is as follows: a sequence of random elements $(V_N)_N$ taking values in a metric space $(\mathbb{V}, d_V)$ (or equivalently the sequence of their laws, say P_N) satisfies an LDP with rate function I if

$$\limsup_{N\to+\infty} \frac{1}{N}\log\Big(\mathbb{P}\{V_N \in C\}\Big) = \limsup_{N\to+\infty} \frac{1}{N}\log\Big(P_N(C)\Big) \le -\inf_{v\in C} I(v)$$

for any closed set $C \subset \mathbb{V}$ and

$$\liminf_{N\to+\infty} \frac{1}{N}\log\Big(\mathbb{P}\{V_N \in O\}\Big) = \liminf_{N\to+\infty} \frac{1}{N}\log\Big(P_N(O)\Big) \ge -\inf_{v\in O} I(v)$$

for any open set $O \subset \mathbb{V}$.

Proposition 8 *The sequence of measures $\mathcal{Q}_{L,N}$, satisfies an LDP on $(\mathcal{S}_D^+)^L$ as $N \to +\infty$ (L fixed) with rate function*

$$I(Q_1, \ldots, Q_L) := \frac{1}{2}\sum_{\ell=1}^L \Big(\operatorname{tr}(Q_\ell) - \log(\det(Q_\ell)\Big) - \frac{DL}{2}.$$

Moreover, the sequence of measures $\mathcal{Q}_{L,N}(dQ_1 \ldots dQ_L|\mathbf{y}_{1:P}, \beta)$ defined in Proposition 6*, satisfies an LDP with the same rate function, independently from β and $\mathbf{y}_{1:P}$.*

Remark 9 *As a corollary, using the well-known contraction principle (see, e.g., Theorem 4.2.1 in Dembo and Zeitouni, 2010), if* $(Q_1, \dots, Q_L)$ *has law* $\mathcal{Q}_{L,N}$*, then* $Q^{(L)}(Q_1, \dots, Q_L)$ *satisfies an LDP on* $\mathcal{S}_D^+$ *with rate function*

$$I^*(Q) = \inf\{I(Q_1, \dots, Q_L) : Q^{(L)}(Q_1, \dots, Q_L) = Q\}$$

After the re-scaling in Eq. (18), the posterior predictive distribution is the same mixture of Gaussians described in Proposition 6 with the posterior mixing measure $\mathcal{Q}_{L,N}(\cdot|\mathbf{y}_{1:P}, \beta)$ replaced by the rescaled mixing measure

$$\mathcal{Q}^\circ_{L,N}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta) := \frac{e^{-\frac{N}{2}\Phi^\circ_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P}) - \frac{1}{2}R(Q_1,\dots,Q_L)}\mathcal{Q}_{L,N}(Q_1, \dots, Q_L)}{\int_{(\mathcal{S}_D^+)^L} e^{-\frac{N}{2}\Phi^\circ_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P}) - \frac{1}{2}R(Q_1,\dots,Q_L)}\mathcal{Q}_{L,N}(Q_1, \dots, Q_L)},$$

where

$$\begin{aligned} \Phi^\circ_\beta(Q_1, \dots, Q_L, \mathbf{y}_{1:P}) &:= \mathbf{y}_{1:P}^\top(\Sigma_{11} + \beta^{-1}\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P} \\ R(Q_1, \dots, Q_L) &= \log(\det(\mathbb{1}_{DP} + \beta\Sigma_{11})), \end{aligned} \tag{19}$$

and $\mathbf{m}_0$ replaced by $\sqrt{N}\mathbf{m}_0$. Combining well-known Varadhan's lemma together with Proposition 8, one can prove that $\mathcal{Q}^\circ_{L,N}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P}, \beta)$ also satisfies an LDP.

Proposition 10 *The sequence of measures* $\mathcal{Q}^\circ_{L,N}(dQ_1, \dots, Q_L|\mathbf{y}_{1:P}, \beta)$ *as* $N \to +\infty$ *satisfies an LDP with rate function*

$$I^\circ(Q_1, \dots, Q_L) = \frac{1}{2}\sum_{\ell=1}^{L}\Big(\operatorname{tr}(Q_\ell) - \log(\det(Q_\ell))\Big) + \frac{1}{2}\Phi^\circ_\beta(Q_1, \dots, Q_L, \mathbf{y}_{1:P}) - \bar{I}_0.$$

where $\bar{I}_0 = \frac{1}{2}\inf_{Q_1,\dots,Q_L}\{\sum_{\ell=1}^{L}\Big(\operatorname{tr}(Q_\ell) - \log(\det(Q_\ell))\Big) + \Phi^\circ_\beta(Q_1, \dots, Q_L, \mathbf{y}_{1:P})\}$.

The rate function now explicitly depends on the training inputs and labels, showing that the measure over the Wishart ensembles this time concentrates around non-trivial data-dependent solutions, which quantitatively characterize feature learning in this limit (see also Discussion in Section 4.2 for a comparison with the proportional limit studied in Li and Sompolinsky, 2021; Pacelli et al., 2023; Hanin and Zlokapa, 2023). We observe that the result in Proposition 10 is complementary to the recent investigation in Chizat et al. (2022), where the feature learning infinite-width limit of DLNs under gradient descent dynamics was characterized.

Remark 11 Propositions 8 *and* 10 *hold for the C-DLN with* $\mathcal{Q}_{L,N}$ *and* $\mathcal{Q}^\circ_{L,N}$ *replaced by* $\mathcal{Q}_{L,C}$ *and* $\mathcal{Q}^\circ_{L,C}$*. In this case* $C \to +\infty$ *and* Φ°_β *has the same form given in* (19) *but* Σ_{11} *is defined as in Proposition 7.*

4. Discussion and conclusions

In this paper, we derived exact non-asymptotic integral representations of the prior distribution over the output and of the predictive posterior of deep linear Bayesian NNs, both for the case of fully-connected architectures with multiple outputs and of convolutional architectures. In this section, we provide a thorough discussion on how our results map to the ones already obtained in the past for special cases, as reported in Sec. 1.2, putting our work in perspective with the existing literature.

4.1 Connection with the formalism of Meijer G-functions for $D = 1$

In this section we provide a way to map our results for fully-connected NNs with scalar output ($D = 1$) to the formalism of Meijer G-functions (see, for example, Zavatone-Veth and Pehlevan (2021); Hanin and Zlokapa (2023)). The most direct way to obtain this mapping is to consider the Bayesian model evidence in the limit $\beta \to +\infty$ (zero temperature). With our notation (see Eq. (6)) and recalling that here $D = 1$

$$\begin{aligned}
Z_{\mathbf{y}_{1:P},\beta} &= \int_{\mathbb{R}^{P+1}} \mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}) p_{\text{prior}}(\mathbf{s}_0, \mathbf{s}_{1:P}|\mathbf{x}^0, X) d\mathbf{s}_0 d\mathbf{s}_{1:P} \\
&= \int_{\mathbb{R}^{P}} \mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P}) p_{\text{prior}}(\mathbf{s}_{1:P}|X) d\mathbf{s}_{1:P} \\
&= \int_{(\mathbb{R}^+)^L} e^{-\frac{1}{2}\Phi_\beta(Q_1 \dots Q_L, \mathbf{y}_{1:P})} \mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)
\end{aligned}$$

with

$$\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L) = \prod_{\ell=1}^{L} \frac{(N_\ell/2)^{N_\ell/2}}{\Gamma(N_\ell/2)} (Q_\ell)^{\frac{N_\ell}{2}-1} e^{-\frac{N_\ell Q_\ell}{2}} dQ_\ell.$$

Taking the limit $\beta \to \infty$ one obtains

$$Z_\infty := \lim_{\beta \to +\infty} \beta^{-P/2} Z_{\mathbf{y}_{1:P},\beta} = \int_{(\mathbb{R}^+)^L} e^{-\frac{\mathbf{y}_{1:P}^\top \tilde{\Sigma}_{11}^{-1} \mathbf{y}_{1:P}}{2Q_1 \dots, Q_L}} (Q_1 \dots, Q_L)^{-P/2} \mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)$$

where $\tilde{\Sigma}_{11} := \frac{X^\top X}{N_0 \lambda^*}$. The resulting Z_∞ is (up to a constant) the Bayesian evidence $Z_\infty(0)$ defined in equation [8] of Hanin and Zlokapa (2023). Recalling (12), the above integral becomes

$$Z_\infty = \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \int_{\mathbb{R}^+} e^{-\frac{\omega}{q}} q^{-P/2} G_{0L}^{L0}\left(q \;\middle|\; \begin{matrix} - \\ \frac{N_1}{2}-1, \cdots, \frac{N_L}{2}-1 \end{matrix} \right) dq\,,$$

where

$$\omega := \mathbf{y}_{1:P}^\top \tilde{\Sigma}_{11}^{-1} \mathbf{y}_{1:P} \frac{\prod_{\ell=1}^L N_\ell}{2^{L+1}} = \mathbf{y}_{1:P}^\top (XX^\top)^{-1} \mathbf{y}_{1:P} \frac{\prod_{\ell=0}^L \lambda_\ell N_\ell}{2^{L+1}}\,.$$

With the change of variables $u = q^{-1}$, this integral becomes

$$\begin{aligned}
Z_\infty &= \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \int_{\mathbb{R}^+} e^{-u\omega} u^{P/2-2} G_{0L}^{L0}\left(u^{-1} \;\middle|\; \begin{matrix} - \\ \frac{N_1}{2}-1, \cdots, \frac{N_L}{2}-1 \end{matrix} \right) du \\
&= \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \int_{\mathbb{R}^+} e^{-u\omega} u^{P/2-2} G_{L0}^{0L}\left(u \;\middle|\; \begin{matrix} 2-\frac{N_1}{2}, \cdots, 2-\frac{N_L}{2} \\ - \end{matrix} \right) du \\
&= \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \omega^{1-P/2} G_{L+1,0}^{0,L+1}\left(\omega^{-1} \;\middle|\; \begin{matrix} 2-\frac{P}{2}, 2-\frac{N_1}{2}, \cdots, 2-\frac{N_L}{2} \\ - \end{matrix} \right) \\
&= \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \omega^{1-P/2} G_{0,L+1}^{L+1,0}\left(\omega \;\middle|\; \begin{matrix} - \\ \frac{P}{2}-1, \frac{N_1}{2}-1, \cdots, \frac{N_L}{2}-1 \end{matrix} \right) \\
&= \frac{(\prod_{\ell=1}^L N_\ell/2)^{P/2}}{\prod_{\ell=1}^L \Gamma(N_\ell/2)} \omega^{-P/2} G_{0,L+1}^{L+1,0}\left(\omega \;\middle|\; \begin{matrix} - \\ \frac{P}{2}, \frac{N_1}{2}, \cdots, \frac{N_L}{2} \end{matrix} \right)\,,
\end{aligned}$$

where we used at each step known properties of the Meijer G-functions (see Supplementary Material of Hanin and Zlokapa (2023), Eq. (7, 8, 9)). In this way one recovers expression [14] of Theorem 1 in Hanin and Zlokapa (2023). Along these lines, one can re-obtain also the other results stated in Theorem 1 of Hanin and Zlokapa (2023) from our approach.

Moreover, Proposition 1 shows that $\mathbf{S}_{1:P}$ is a mixture of Gaussians with a Wishart matrix as covariance. It is known that the resulting distribution is the so-called multivariate Student's t-distribution (see, e.g., Representation B in Lin, 1972). One can show that the characteristic function of a multivariate Student's t can be expressed in term of Macdonald functions (Joarder and Alam, 1995). Since the Macdonald function is recovered as a special case of a Meijer G-function, this further clarifies (at least for $L = 1$) the link between our stochastic representation and the Fourier representation given in Zavatone-Veth and Pehlevan (2021).

4.2 From kernel renormalization to rigorous Bayesian statistics

In concluding, we discuss how our derivations provide a way to interpret the recent statistical physics literature on Bayesian deep learning theory in the proportional limit ($P, N_\ell \to \infty$, fixed $\alpha_\ell = P/N_\ell$) such as Li and Sompolinsky (2021, 2022); Pacelli et al. (2023); Aiudi et al. (2025).

Hanin and Zlokapa (2023) provides the first notable results that partly address this issue. However, the representation of the Bayesian model evidence (partition function) in terms of Meijer G-functions hides the possibility of a direct comparison with the previously-established physics results. On the contrary, our explicit non-asymptotic integral representation of the prior delivers a precise one-to-one mapping between the idea of kernel renormalization introduced by physicists (Li and Sompolinsky, 2021; Pacelli et al., 2023; Aiudi et al., 2025) and the Wishart ensembles that appear in our rigorous derivation.

Let us first consider the case of single output DLNs, $D = 1$. Using the explicit integral representation for the prior given in Proposition 1, and assuming Gaussian likelihood, we find the following non-asymptotic result for the partition function:

$$Z_{\mathbf{y}_{1:P},\beta} \propto \int_0^\infty \prod_{\ell=1}^{L} dQ_\ell \, e^{-\sum_{\ell=1}^{L}\left[\frac{N_\ell}{2}Q_\ell - \left(\frac{N_\ell}{2}-1\right)\log Q_\ell\right] - \frac{1}{2}\Phi_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P})}, \tag{20}$$

which is exact up to an overall unimportant constant. Note that here the first two terms in the exponential arise from the probability measure over the scalar variables Q_ℓ, while the third and fourth data-dependent terms are a consequence of the Gaussian integration over the output variables. This partition function can be compared with the result for non-linear networks found in the proportional limit in Pacelli et al. (2023). It turns out that if we restrict our focus to linear NNGP kernels in Pacelli et al. (2023), we find an effective action equivalent to the exponent in Eq. (20) at leading order in N_ℓ. Moreover, it is reasonable to expect that the third term in Eq. (20) scale linearly with P, under proper hypotheses on the training data (note that this requirement is similar to the one in Hanin and Zlokapa, 2023). This observation suggests that it is legitimate to employ large deviations techniques (or the more powerful saddle-point method) in the proportional limit, and it provides the interpretation mentioned at the beginning of this section, since one can prove that the saddle-point equations derived from Eq. (20) are exactly the same as those found in Li and

Sompolinsky (2021). Let us check this explicitly in the case $N_\ell = N$, $\forall \ell = 1, \ldots, L$. The leading order of the exponent (at large N and P) in Eq. (20) is given by:

$$-\frac{N}{2}\left\{\sum_{\ell=1}^{L}\left[Q_\ell - \log Q_\ell\right] + \frac{\alpha}{P}\mathbf{y}_{1:P}^\top(\Sigma_{11} + \beta^{-1}\mathbb{1}_P)^{-1}\mathbf{y}_{1:P} + \frac{\alpha}{P}\log(\det(\mathbb{1}_P + \beta\Sigma_{11}))\right\}, \quad (21)$$

where $\alpha = P/N$ and Σ_{11} is the same as defined in Proposition 6. Assuming that the third and fourth terms of the equation above converge to a finite value in the proportional limit (P, $N \to \infty$ with $\alpha = P/N$ fixed), we can employ the saddle-point method to estimate the integral in Eq. (20), which amounts to find the minima of the term in brackets in Eq. (21). After derivation wrt each Q_ℓ and by noting that the only solution of the resulting system of L equations has $Q_\ell = u_0$ $\forall \ell = 1, \ldots, L$, we recover Eq. (11) in Li and Sompolinsky (2021) if we take the limit $\beta \to \infty$.

By a line of reasoning similar to that exposed above for $D = 1$, we can also interpret the (heuristic) kernel shape renormalization pointed out in Li and Sompolinsky (2021) and Pacelli et al. (2023) for one-hidden-layer fully-connected architectures with multiple outputs, as well as the local kernel renormalization proposed to describe shallow CNNs in Aiudi et al. (2025). If we consider for instance the effective action derived in Aiudi et al. (2025), we again recognize that the first two terms arise from the probability measure of a Wishart ensemble, whereas the last two data-dependent terms are a consequence of the integration of the output variables. Even in this case, one can check a precise equivalence with the non-asymptotic partition function that one finds in the CNN case using our result for the prior in Proposition 6 for $L = 1$.

Finally, we point out that the feature learning infinite-width limit investigated in Proposition 10 shares qualitative similarities with the proportional limit studied in the physics literature. In both cases, the measure over the Wishart ensembles concentrates over data-dependent solutions of a rate function (or a saddle-point effective action). However, this happens for different reasons in the two settings: in the feature learning infinite-width limit, this effect is due to the additional data-dependent terms entering the rate function thanks to the mean-field scaling; in the proportional limit with standard scaling discussed in this section, data are instead entering the saddle point equations for the variables Q_ℓ in the action (21). Furthermore, the three terms of the rate function in Proposition 10 also appear in the effective action found in Pacelli et al. (2023), which is known to reproduce the saddle-point equations firstly identified in Li and Sompolinsky (2021) for the linear case.

Acknowledgments

P.R. is supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) "A Multiscale integrated approach to the study of the nervous system in health and disease" (DN. 1553 11.10.2022). F.B. is partially supported by the MUR - PRIN project "Discrete random structures for Bayesian learning and prediction" no. 2022CLTYP4.

Appendix A. Proofs

A.1 Preliminary facts

Matrix normal distribution. A random matrix Z of dimension $n_1 \times n_2$ has a centred matrix normal distribution with parameters (Σ_1, Σ_2) (with Σ_i's positive symmetric $n_i \times n_i$ matrices), if for any matrix S of dimension $n_2 \times n_1$

$$\mathbf{E}[e^{i\,\mathrm{tr}(SZ)}] = \exp\left\{-\frac{1}{2}\,\mathrm{tr}(S\Sigma_1 S^\top \Sigma_2)\right\}. \tag{22}$$

In symbols $Z \sim \mathcal{MN}(0, \Sigma_1, \Sigma_2)$.

We collect here some properties of multivariate normals, matrix normals and Wishart distributions to be used later. See Gupta and Nagar (2000) for details. In what follows $\mathrm{vec}(A)$ denotes the stacking of the *columns* of a matrix A to form a vector.

(P1) *Linear transformation of matrix normals.* Given two matrices H and K with compatible shape, if $Z \sim \mathcal{MN}(0, \Sigma_1, \Sigma_2)$ then

$$HZK \sim \mathcal{MN}(0, H\Sigma_1 H^\top, K^\top \Sigma_2 K).$$

(P2) *Equivalence with the multivariate normal.* $Z \sim \mathcal{MN}(0, \Sigma_1, \Sigma_2)$ if and only if $\mathrm{vec}(Z) \sim \mathcal{N}(0, \Sigma_2 \otimes \Sigma_1)$

(P3) *Laplace functional of a Wishart distribution.* Let S be a symmetric positive definite matrix. Then for any symmetric matrix C and any real number α such that all the eigenvalues of $(\mathbb{1}_D + \alpha SC)$ are strictly positive,

$$[\det(\mathbb{1}_D + \alpha SC)]^{-\frac{N}{2}} = \int_{\mathcal{S}_D^+} e^{-\frac{\alpha}{2}\,\mathrm{tr}(CQ)} \mathcal{W}_D(dQ|S, N). \tag{23}$$

Versions of this formula can be found in literature as *Ingham-Siegel integrals* (Ingham, 1933; Siegel, 1935, and Fyodorov, 2002 for an extension to Hermitian matrices); in this generality, the statement can be derived by Proposition 8.3 and its proof in Eaton (2007). In particular the formula above holds for $\alpha > 0$ and $S = \mathbb{1}_D$ and C symmetric and ≥ 0, showing that when $\mathbf{Z}_j \overset{iid}{\sim} \mathcal{N}(0, C)$, then

$$\mathbf{E}[e^{-\frac{\alpha}{2}\sum_{j=1}^{N}\|\mathbf{Z}_j\|^2}] = \int_{\mathcal{S}_D^+} e^{-\frac{\alpha}{2}\,\mathrm{tr}(CQ)} \mathcal{W}_D(dQ|\mathbb{1}_D, N). \tag{24}$$

A.2 Proofs for the FC-DLN

We set $h_{i_\ell,\mu}^{(\ell)} = h_{i_\ell}^{(\ell)}(\mathbf{x}^\mu)$ and

$$H^0 := X = [\mathbf{x}^1, \ldots, \mathbf{x}^P].$$

For $\ell \geq 2$ define the $N_{\ell-1} \times P$ matrix

$$H^{(\ell-1)} = [h_{i,\mu}^{(\ell-1)}]_{i,\mu}.$$

Notice that the weights of the network distributed as (7) can be arranged as rectangular matrices with law

$$W^{(\ell-1)} \sim \mathcal{MN}(0, \mathbb{1}_{N_\ell}, \lambda_{\ell-1}^{-1}\mathbb{1}_{N_{\ell-1}}).$$

With this notations, it is immediate to check that for any ℓ one has

$$H^{(\ell)} = \frac{1}{\sqrt{N_{\ell-1}}} W^{(\ell-1)} H^{(\ell-1)} . \tag{25}$$

(P4) Let $\mathbf{H}^{(\ell)}_{i_\ell} = (h^{(\ell-1)}_{i_\ell,1}, \dots, h^{(\ell-1)}_{i_\ell,P})^\top$ the i_ℓ-th row of $H^{(\ell)}$. Given $H^{(\ell-1)}$, the vectors $\mathbf{H}^{(\ell)}_{i_\ell}$ are independent with law $\mathcal{N}(0, (\lambda_{\ell-1}N_{\ell-1})^{-1} H^{(\ell-1)\top} H^{(\ell-1)})$. Indeed, $\mathbf{H}^{(\ell)}_{i_\ell} = N_{\ell-1}^{-\frac{1}{2}} W^{(\ell-1)}_{i_\ell,\cdot} H^{(\ell-1)}$ and the $W^{(\ell-1)}_{i_\ell,\cdot}$'s are independent and identically distributed. Recall that if $\mathbf{Z} \sim \mathcal{N}(0, C)$ and A is a matrix then $A\mathbf{Z} \sim \mathcal{N}(0, ACA^\top)$.

Proof of (11) **in Remark 3** Iterating (25) one gets

$$H^{(L+1)} = \frac{1}{\sqrt{N_L}} W^{(L)} H^{(L)} = \dots = \frac{1}{\sqrt{N_L}} W^{(L)} \cdots \frac{1}{\sqrt{N_0}} W^{(0)} X . \tag{26}$$

■

Proof of Proposition 1 and of (10) **in Remark 3** Starting from (25) with $\ell = L + 1$, conditioning on $H^{(L)}$, by (P1) and (22), it follows that for any $P \times N_{L+1} = P \times D$ matrix $\bar{S}$

$$\mathbf{E}[e^{i \operatorname{tr}(H^{(L+1)}\bar{S})}] = \mathbf{E}[e^{-\frac{1}{2N_L\lambda_L} \operatorname{tr}(\bar{S}^\top H^{(L)\top} H^{(L)} \bar{S})}] . \tag{27}$$

We divide the proof in few steps.

Step 1. For any $\ell \geq 1$ and any $P \times D$ matrix Θ

$$\operatorname{tr}(\Theta^\top H^{(\ell)\top} H^{(\ell)} \Theta) = \sum_{i_\ell=1}^{N_\ell} \|\mathbf{Y}^\ell_{i_\ell}\|^2 \tag{28}$$

with $\mathbf{Y}^\ell_{i_{\ell-1}} = H^{(\ell)}_{i_\ell,\cdot}\Theta$. To see this, note that

$$\begin{aligned}\operatorname{tr}(\Theta^\top H^{(\ell)\top} H^{(\ell)} \Theta) &= \sum_i \sum_{\mu,\nu} \Theta_{\mu,i} \Big(\sum_{i_\ell} h^{(\ell)}_{i_\ell,\mu} h^{(\ell)}_{i_\ell,\nu} \Big) \Theta_{\nu,i} = \sum_{i_\ell} \sum_i \sum_{\mu,\nu} \Theta_{\mu,i}\Theta_{\nu,i} h^{(\ell)}_{i_\ell,\mu} h^{(\ell)}_{i_\ell,\nu} \\ &= \sum_{i_\ell} \sum_i \Big(\sum_\mu h^{(\ell)}_{i_\ell,\mu} \Theta_{\mu,i} \Big)^2 = \sum_{i_\ell} \|\mathbf{Y}^\ell_{i_\ell}\|^2.\end{aligned}$$

Step 2. For any $\ell \geq 1$ and any $P \times D$ matrix Θ we claim that

$$\mathbf{E}[e^{-\frac{1}{2\lambda_\ell N_\ell} \operatorname{tr}(\Theta^\top H^{(\ell)\top} H^{(\ell)} \Theta)}] = \mathbf{E}[e^{-\frac{1}{2\lambda_\ell} \operatorname{tr}(K_{\ell-1} Q_\ell)}] \tag{29}$$

where Q_ℓ is a $D \times D$ random matrix with Wishart distribution of N_ℓ degree of freedom and scale matrix $\mathbb{1}_D / N_\ell$ and

$$K_{\ell-1} := \frac{1}{\lambda_{\ell-1} N_{\ell-1}} \Theta^\top H^{(\ell-1)\top} H^{(\ell-1)} \Theta.$$

Writing $Q_\ell = U_\ell^\top U_\ell$, with U_ℓ a $D \times D$ matrix, one has

$$\mathbf{E}[e^{-\frac{1}{2\lambda_\ell N_\ell}\operatorname{tr}(\Theta^\top H^{(\ell)\top} H^{(\ell)}\Theta)}] = \mathbf{E}[e^{-\frac{1}{2\lambda_\ell N_{\ell-1}\lambda_{\ell-1}}\operatorname{tr}(U_\ell \Theta^\top H^{(\ell-1)\top} H^{(\ell-1)}\Theta U_{\ell-1}^\top)}]. \tag{30}$$

To prove (29), note that by (P4) the $\mathbf{Y}_{i_\ell}^\ell$'s are independent and identically distributed conditionally on $H^{\ell-1}$, with

$$\mathbf{Y}_{i_\ell}^\ell | H^{(\ell-1)} \sim \mathcal{N}(0, K_{\ell-1}).$$

Hence using, (28) and (24)

$$\mathbf{E}[e^{-\frac{1}{2\lambda_\ell N_\ell}\operatorname{tr}(\Theta^\top H^{(\ell)\top} H^{(\ell)}\Theta)}] = \mathbf{E}[e^{-\frac{1}{2\lambda_\ell N_\ell}\sum_{i_\ell=1}^{N_\ell}\|\mathbf{Y}_{i_\ell}^\ell\|^2}] = \mathbf{E}[e^{-\frac{1}{2\lambda_\ell}\operatorname{tr}(K_{\ell-1}Q_\ell)}]. \tag{31}$$

Step 3. Starting from $\ell = L$ and $\Theta = \bar{S}$, iterating (30) one obtains

$$\mathbf{E}[e^{-\frac{1}{2\lambda_L N_L}\operatorname{tr}(\bar{S}^\top H^{(L)\top} H^{(L)}\bar{S})}] = \mathbf{E}[e^{-\frac{1}{2\lambda^* N_0}\operatorname{tr}(U_{1:L}\bar{S}^\top H^{(0)\top} H^{(0)}\bar{S} U_{1:L}^\top)}]$$

where $\lambda^* := \lambda_0 \dots \lambda_L$ and $U_{1:L} = U_1 \dots U_L$ for $Q_\ell = U_\ell^\top U_\ell$ independent $D \times D$ random matrix with Wishart distribution with N_ℓ degree of freedom and $\mathbb{1}_D/N_\ell$ as scale matrix. By (27) we conclude that

$$\begin{aligned}\mathbf{E}[e^{i\operatorname{tr}(H^{(L+1)}\bar{S})}] &= \mathbf{E}[e^{-\frac{1}{2\lambda_L N_L}\operatorname{tr}(\bar{S}^\top H^{(L)\top} H^{(L)}\bar{S})}] \\ &= \mathbf{E}[e^{-\frac{1}{2\lambda^* N_0}\operatorname{tr}(U_{1:L}\bar{S}^\top H^{(0)\top} H^{(0)}\bar{S} U_{1:L}^\top)}] \\ &= \mathbf{E}[e^{-\frac{1}{2\lambda^* N_0}\operatorname{tr}(\bar{S} U_{1:L}^\top U_{1:L}\bar{S}^\top H^{(0)\top} H^{(0)})}].\end{aligned}$$

Recalling that $H^{(0)\top}H^{(0)} := X^\top X$, using once again (22), this shows that, conditionally on $U_{1:L}^\top U_{1:L}$, $H^{(L+1)} \sim \mathcal{MN}(0, U_{1:L}^\top U_{1:L}, (\lambda^* N_0)^{-1} X^\top X)$. Note that this is (10) of Remark 3.

Since $\mathbf{S}_{1:P} = \operatorname{vec}[H^{(L+1)}]$, by (P2), one gets

$$\mathbf{S}_{1:P} | U_{1:L}^\top U_{1:L} \sim \mathcal{N}(0, (\lambda^* N_0)^{-1} X^\top X \otimes U_{1:L}^\top U_{1:L}).$$

and the thesis follows. ■

We now prove a slightly more general statement of the one given in Proposition 6. Let $\mathbf{s}^\top = ((\mathbf{s}^0)^\top, (\mathbf{s}^1)^\top, \dots, (\mathbf{s}^P)^\top)$ and $\tilde{X} = [\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^P]$. In what follows $\mathbb{1}_M$ is the identity matrix of dimension $M \times M$ and $0_{M\times N}$ is the zero matrix of dimension $M \times N$.

Proposition 12 *Let $\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L)$ be as in Eq. (9) and $Q^{(L)} = Q^{(L)}(Q_1, \dots, Q_L)$ be defined as in* Proposition 1. *Let $p_{\text{post}}(\mathbf{s}|\mathbf{y}_{1:P}, \mathbf{x}_0, X)$ be the posterior distribution of $(\mathbf{S}_0, \dots, \mathbf{S}_P)$ given $\mathbf{y}_{1:P}$ in a FC-DLN. If $\det(\tilde{X}^\top \tilde{X}) > 0$, then*

$$p_{\text{post}}(\mathbf{s}|\mathbf{y}_{1:P}, \tilde{X}) = \int_{(\mathcal{S}_D^+)^L} \frac{e^{-\frac{1}{2}(\mathbf{s}-\mathbf{m})^\top(\beta\Pi_0 + \Sigma^{-1})(\mathbf{s}-\mathbf{m})}}{(2\pi)^{\frac{D(P+1)}{2}} \det((\beta\Pi_0 + \Sigma^{-1})^{-1})^{\frac{1}{2}}} \mathcal{Q}_{L,N}(dQ_1 \dots dQ_L | \mathbf{y}_{1:P}, \tilde{X}, \beta)$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{01}^\top & \Sigma_{11} \end{pmatrix} := \frac{\tilde{X}^\top \tilde{X}}{N_0 \lambda^*} \otimes Q^{(L)}, \qquad \Pi_0 = \begin{pmatrix} 0_{D\times D} & 0_{P\times DP} \\ 0_{DP\times P} & \mathbb{1}_{DP} \end{pmatrix}$$

$$\mathbf{m} := \begin{pmatrix}\mathbf{m}_0\\ \mathbf{m}_1\end{pmatrix} = \begin{pmatrix}\Sigma_{01}(\Sigma_{11}+\beta^{-1}\mathbb{1}_{DP})^{-1}\\ \Sigma_{11}(\Sigma_{11}+\beta^{-1}\mathbb{1}_{DP})^{-1}\end{pmatrix}\mathbf{y}_{1:P},$$

$$\mathcal{Q}_{L,N}(dQ_1\dots dQ_L|\mathbf{y}_{1:P},\tilde{X},\beta) := \frac{e^{-\frac{1}{2}\Phi_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P})}\mathcal{Q}_{L,N}(dQ_1\dots dQ_L)}{\int_{(\mathcal{S}_D^+)^L} e^{-\frac{1}{2}\Phi_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P})}\mathcal{Q}_{L,N}(dQ_1\dots dQ_L)}$$

and $\Phi_\beta(Q_1,\dots,Q_L,\mathbf{y}_{1:P}) := \mathbf{y}_{1:P}^\top(\Sigma_{11}+\beta^{-1}\mathbb{1}_{DK})^{-1}\mathbf{y}_{1:P} + \log(\det(\mathbb{1}_{DP}+\beta\Sigma_{11}))$.

Proof Proposition 1 gives

$$\mathcal{L}(\mathbf{y}_{1:P}|\mathbf{s}_{1:P})p_{\text{prior}}(\mathbf{s}_0,\mathbf{s}_{1:P}|\mathbf{x}^0,X) = \int \frac{e^{-\frac{\beta}{2}(\mathbf{s}_{1:P}-\mathbf{y}_{1:P})^\top(\mathbf{s}_{1:P}-\mathbf{y}_{1:P})-\frac{1}{2}\mathbf{s}\Sigma^{-1}\mathbf{s}}}{\det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{D(P+1)}{2}}}\mathcal{Q}_{L,N}(dQ_1\dots dQ_L)$$

with $\Sigma := (N_0\lambda^*)^{-1}\tilde{X}^\top\tilde{X}\otimes Q^{(L)}$. Since $\tilde{X}^\top\tilde{X}$ is assumed to be strictly positive definite and hence invertible, using that $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$ (see, e.g., Proposition 1.28 Eaton, 2007) it follows that also Σ is invertible. Simple computations show that

$$\begin{aligned}\beta(\mathbf{s}_{1:P}-\mathbf{y}_{1:P})^\top(\mathbf{s}_{1:P}-\mathbf{y}_{1:P}) + \mathbf{s}\Sigma^{-1}\mathbf{s} &= (\mathbf{s}-\tilde{\mathbf{y}})^\top\Pi_0\beta(\mathbf{s}-\tilde{\mathbf{y}}) + \mathbf{s}\Sigma^{-1}\mathbf{s}\\ &= (\mathbf{s}-\mathbf{m})^\top(\Pi_0\beta+\Sigma^{-1})(\mathbf{s}-\mathbf{m}) + \varphi(\Sigma,\mathbf{y}_{1:P})\end{aligned}$$

where

$$\begin{aligned}&\tilde{\mathbf{y}} := (0_D^\top,\mathbf{y}_{1:P}^\top)^\top, \quad \mathbf{m} = (\Pi_0\beta+\Sigma^{-1})^{-1}\Pi_0\beta\tilde{\mathbf{y}} = \beta(\Pi_0\beta+\Sigma^{-1})^{-1}\tilde{\mathbf{y}}\,,\\ &\varphi(\Sigma,\mathbf{y}_{1:P}) = \beta\mathbf{y}_{1:P}^\top\mathbf{y}_{1:P} - (\mathbf{m})^\top(\Pi_0\beta+\Sigma^{-1})\mathbf{m}\,.\end{aligned}$$

Writing $\mathbf{m} := [\mathbf{m}_0,\mathbf{m}_1]$ one can check that $(\Sigma_{11}^{-1}+\beta\mathbb{1}_{DP})^{-1} = \Sigma_{11}(\mathbb{1}_{DP}+\beta\Sigma_{11})^{-1}$ and

$$\begin{aligned}&\mathbf{m}_1 = \beta(\Sigma_{11}^{-1}+\beta\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P} = \beta\Sigma_{11}(\mathbb{1}_{DP}+\beta\Sigma_{11})^{-1}\mathbf{y}_{1:P},\\ &\mathbf{m}_0 = \beta\Sigma_{01}\Sigma_{11}^{-1}(\Sigma_{11}^{-1}+\beta\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P} = \beta\Sigma_{01}(\mathbb{1}_{DP}+\beta\Sigma_{11})^{-1}\mathbf{y}_{1:P},\\ &\varphi(\Sigma,\mathbf{y}_{1:P}) = \beta\mathbf{y}_{1:P}^\top\mathbf{y}_{1:P} - \beta^2\mathbf{y}_{1:P}^\top(\Sigma_{11}^{-1}+\beta\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P} = \mathbf{y}_{1:P}^\top(\Sigma_{11}^{-1}+\mathbb{1}_{DP}/\beta)^{-1}\mathbf{y}_{1:P}.\end{aligned}$$

In conclusion

$$\frac{e^{-\frac{\beta}{2}(\mathbf{s}_{1:P}-\mathbf{y}_{1:P})^\top(\mathbf{s}_{1:P}-\mathbf{y}_{1:P})-\frac{1}{2}\mathbf{s}\Sigma^{-1}\mathbf{s}}}{|\Sigma|^{\frac{1}{2}}} = \frac{e^{-\frac{1}{2}\varphi(\Sigma,\mathbf{y}_{1:P})}}{\det(\Sigma)^{\frac{1}{2}}\det(\beta\Pi_0+\Sigma^{-1})^{\frac{1}{2}}}\frac{e^{-\frac{1}{2}(\mathbf{s}-\mathbf{m})^\top(\beta\Pi_0+\Sigma^{-1})(\mathbf{s}-\mathbf{m})}}{(\det(\beta\Pi_0+\Sigma^{-1})^{-1})^{\frac{1}{2}}}.$$

where $\det(\Sigma)^{\frac{1}{2}}\det(\beta\Pi_0+\Sigma^{-1})^{\frac{1}{2}} = \det(\mathbb{1}_{DP}+\beta\Sigma_{11})^{\frac{1}{2}}$.

■

Proof of Proposition 6 Applying the previous proposition, one gets that

$$p_{\text{pred}}(\mathbf{s}_0|\mathbf{y}_{1:P},\tilde{X}) = \int p_{\text{post}}(\mathbf{s}_0,\mathbf{s}_1,\dots,\mathbf{s}_P|\mathbf{y}_{1:P},\tilde{X})d\mathbf{s}_1\dots d\mathbf{s}_P$$

is again a mixture of gaussians with mean $\mathbf{m}_0 = \Sigma_{01}(\Sigma_{11} + \beta^{-1} I)^{-1}\mathbf{y}_{1:P}$ and (conditional) covariance

$$\Big[\Big(\beta\Pi_0 + \Sigma^{-1}\Big)^{-1}\Big]_{1:D,1:D}.$$

To conclude it remains to compute the previous matrix. We have already noted that, by $\tilde{X}^\top \tilde{X} > 0$, it follows that also Σ is invertible. Being $X^\top X$ and $\mathbf{x}_0^\top \mathbf{x}_0$ strictly positive, then the Schur complement of Σ_{00} in Σ, $\Sigma_{00}^* := \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{01}^\top$, is well-defined and strictly positive (Proposition 1.34 Eaton, 2007). In particular, using the Banachiewicz inversion formula, one can write the inverse of Σ as

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma_{00}^*)^{-1} & -(\Sigma_{00}^*)^{-1}\Sigma_{01}\Sigma_{11}^{-1} \\ -((\Sigma_{00}^*)^{-1}\Sigma_{01}\Sigma_{11}^{-1})^\top & \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{01}^\top(\Sigma_{00}^*)^{-1}\Sigma_{01}\Sigma_{11}^{-1} \end{pmatrix}.$$

See, e.g., Theorem 1.2 in Zhang (2005). Using again the Banachiewicz inversion formula on $\beta\Pi_0 + \Sigma^{-1}$ (which is again strictly positive), after some tedious computations, one gets that

$$\Big[\Big(\beta\Pi_0 + \Sigma^{-1}\Big)^{-1}\Big]_{1:D,1:D} = \Sigma_{00} - \Sigma_{01}\Big(\frac{1}{\beta}\mathbb{1}_{DP} + \Sigma_{11}\Big)^{-1}\Sigma_{01}^\top.$$

Observing that

$$\Sigma_{00} = \frac{\mathbf{x}_0^\top \mathbf{x}_0}{N_0\lambda^*} \otimes Q^{(L)}, \quad \Sigma_{01} = \frac{\mathbf{x}_0^\top X}{N_0\lambda^*} \otimes Q^{(L)}, \quad \Sigma_{11} = \frac{X^\top X}{N_0\lambda^*} \otimes Q^{(L)},$$

the proof of Proposition 6 is concluded. ■

Proof of Proposition 8 We give the proof in the case $L = 1$, the general case follows the same lines, or using Exercise 4.2.7 in Dembo and Zeitouni (2010) and the independence of the components. The result is a consequence of a very general version of the Cramér's theorem. See Theorem 6.1.3 and Corollary 6.1.6 in Dembo and Zeitouni (2010). Recall that if $Q_1 \sim \mathcal{W}_D(\cdot|\mathbb{1}_D, N)$ then $Q_\ell = \frac{1}{N}\sum_{j=1}^N \mathbf{Z}_j\mathbf{Z}_j^\top$ where $\mathbf{Z}_j\mathbf{Z}_j^\top \sim \mathcal{W}_D(\cdot|\mathbb{1}_D, 1)$ are independent and identically distributed taking values in $\mathcal{S}_D^+$. Now $\mathcal{S}_D^+$ is a convex open cone in the (topological) vector space $\mathcal{S}_D$ of the $D \times D$ symmetric matrices, endowed with the scalar product $(A, B) = \mathrm{tr}(AB)$. The space $\mathcal{S}_D$ is locally convex, Hausdorff and the closure of $\mathcal{S}_D^+$ in $\mathcal{S}_D$ is convex and separable and complete (with the topology induced by $\mathcal{S}_D$). Clearly $P\{\mathbf{Z}_j\mathbf{Z}_j^\top \in \overline{\mathcal{S}_D^+}\} = 1$ and any closed convex hull of a compact in $\overline{\mathcal{S}_D^+}$ is compact. Hence all the hypotheses of the General Cramér's theorem given in Chapter 6 of Dembo and Zeitouni (2010) are satisfied and hence Q_ℓ satisfies a (weak) LDP (when $N \to +\infty$) with rate function given by $I(Q) = \Lambda^*(Q) := \sup_{A\in\mathcal{S}_D}\{\mathrm{tr}(QA) - \Lambda(A)\}$ where

$$\Lambda(A) = \log\left(\int_{\mathcal{S}_D^+} e^{\mathrm{tr}(AQ)} P\{\mathbf{Z}_1\mathbf{Z}_1^\top \in dQ\}\right).$$

In point of fact, see Corollary 6.1.6 in Dembo and Zeitouni (2010), it holds also a strong LDP provided that 0 is in the interior of the domain of Λ, as we shall show. The well-known expression for the Laplace transform of a Wishart distribution with one degree of freedom (see (23)) gives $\Lambda(A) = -\frac{1}{2}\log(\det(\mathbb{1}_D - 2A))$ wherever the eigenvalues of $\mathbb{1}_D - 2A$ are positive and $+\infty$ otherwise. In computing $\Lambda^*(Q)$ we can restrict A to be such that the eigenvalues of $\mathbb{1}_D - 2A$ are positive, since otherwise $-\Lambda(A) = -\infty$. Now we check that

$$\Lambda^*(Q) = \sup_A\{\mathrm{tr}(QA) + \frac{1}{2}\log\det(\mathbb{1}_D - 2A)\} = \frac{1}{2}(\mathrm{tr}(Q) + \log(\det(Q^{-1})) - D).$$

Making the change of variables $B = \mathbb{1}_D - 2A$, we compute $\sup_B H(B,Q)$ for

$$H(B,Q) = \mathrm{tr}(Q(\mathbb{1}_D - B)/2) + \frac{1}{2}\log(\det(B))$$

where now the sup is over symmetric strictly positive matrices B. The gradient of $B \mapsto H(B,Q)$ is easily computed with the Jacobi's formula, $\partial_{B_{ij}}\Lambda(A) = -Q_{ij} + (B^{-1})_{ij}$ for $i<j$ and $\partial_{B_{ii}}\Lambda(A) = -Q_{ii}/2 + (B^{-1})_{ii}/2$ for $i=j$. Solving $\nabla H(B_*,Q) = 0$ we get $0 = -Q + B_*^{-1}$ and hence $B_* = Q^{-1}$. Since $B \mapsto -H(B,Q)$ is convex and $-H(B,Q) \to +\infty$ when B converges to the boundary of the convex cone $\mathcal{S}_D^+$ (since in this case $\log\det(B) \to -\infty$), the unique critical point B_* is also the unique minimum. Hence $\Lambda^*(Q) = H(B_*,Q) = \frac{1}{2}(\mathrm{tr}(Q) + \log(|Q^{-1}|) - D)$, as desired. The LDP for $\mathcal{Q}_{L,N}(dQ_1 \dots dQ_L|\mathbf{y}_{1:P},\beta)$ follows now using the next Lemma 13. Details are omitted since very similar to those given in the next proof of Proposition 10. ∎

In order to prove Proposition 10, we need the following variant of the Varadhan's lemma.

Lemma 13 *Let P_N be a sequence of probability measure on a convex closed subset $\mathcal{S} \subset \mathbb{R}^D$, satisfying an LDP with rate function I. Assume that*

- *Φ_0 is a bounded continuous function from $\mathcal{S} \to \mathbb{R}$;*
- *$\rho : \mathcal{S} \to [0,+\infty)$ such that $\sup_{O \cap B_R} |\rho(s)| < +\infty$ for any R, where $B_R = \{s \in \mathbb{R}^D : \|s\| < R\}$;*
- *the rate function I is such that $\inf_{s \in \mathcal{S} \cap B_R^c} I(s) \to +\infty$ as $R \to +\infty$ and $\inf_s I(s) < +\infty$.*

Then

$$P_N^\circ(ds) = \frac{e^{-(N\Phi_0(s)+\rho(s))}P_N(ds)}{\int_{\mathcal{S}} e^{-(N\Phi_0(s)+\rho(s))}P_N(ds)}$$

satisfies an LDP with rate function $I(s) + \Phi_0(s) - I_0$ where where $I_0 = \inf_s[I(s) + \Phi_0(s)]$.

Proof Varhadan Lemma gives that $\bar{P}_N(ds) = P_N(ds)/(\int_{\mathcal{S}} e^{-N\Phi_0(s)}P_N(ds))$ satifies an LDP with rate $I_{\Phi_0}(s) = I(s) + \Phi_0(s) - I_0$. Then,

$$\frac{1}{N}\log\Big(\int_{\mathcal{S}} e^{-(N\Phi_0(s)+\rho(s))}\bar{P}_N(ds)\Big) \leq \frac{1}{N}\log\Big(\int_{\mathcal{S}} e^{-N\Phi_0(s)}\bar{P}_N(ds)\Big) = 0$$

and

$$\begin{aligned}\liminf_N \frac{1}{N}\log\Big(\int_{\mathcal{S}} e^{-(N\Phi_0(s)+\rho(s))}\bar{P}_N(ds)\Big)\\ &\geq \liminf_N \frac{1}{N}\log\Big(\int_{B_R\cap\mathcal{S}} e^{-N\Phi_0(s)}\bar{P}_N(ds)e^{-\sup_{B_R\cap\mathcal{S}}|\rho(s)|}\Big)\\ &\geq -\inf_{s\in B_R\cap\mathcal{S}}[I(s)+\Phi_0(s)]-I_0.\end{aligned}$$

Since for $R\to+\infty$ one has $\inf_{s\in B_R\cap\mathcal{S}}[I(s)+\Phi_0(s)]-I_0\to 0$, one obtains that

$$\frac{1}{N}\log\Big(\int_{\mathcal{S}} e^{-(N\Phi_0(s)+\rho(s))}\bar{P}_N(ds)\Big)\to 0$$

which means that

$$\frac{1}{N}\log\Big(\frac{\int_{\mathcal{S}} e^{-(N\Phi_0(s)+\rho(s))}P_N(ds)}{\int_{\mathcal{S}} e^{-N\Phi_0(s)}P_N(ds)}\Big)\to 0.$$

If now C is a closed set, then

$$\begin{aligned}&\limsup_{N\to+\infty}\frac{1}{N}\log\Big(\int_C e^{-(N\Phi_0(s)+\rho(s))]}\bar{P}_N(ds)\Big)\leq \limsup_{N\to+\infty}\frac{1}{N}\log\Big(\int_C e^{-N\Phi_0(s)}\bar{P}_N(ds)\Big)\\ &\leq -\inf_{s\in C}[I(s)-\Phi_0(s)-I_0]\end{aligned}$$

On the other hand if O is open, then

$$\begin{aligned}&\liminf_{N\to+\infty}\frac{1}{N}\log\Big(\int_O e^{-(N\Phi_0(s)+\rho(s))]}\bar{P}_N(ds)\log\Big)\\ &\geq \liminf_{N\to+\infty}\frac{1}{N}\Big[\log\Big(\int_{O\cap B_R} e^{-N\Phi_0(s)}\bar{P}_N(ds)\log\Big)+\log\Big(e^{-\sup_{O\cap B_R}|\rho(s)|}\Big)\Big]\\ &\geq -\inf_{s\in O\cap B_R}[I(s)+\Phi_0(s)-I_0].\end{aligned}$$

Since $\inf_{s\in\mathcal{S}\cap B_R^c} I(s)\to+\infty$ as $R\to+\infty$, $\inf_s I(s)<+\infty$ and Φ_0 is bounded, then $\inf_{s\in\mathcal{S}\cap B_R^c}[I(s)+\Phi_0(s)]\to+\infty$ as $R\to+\infty$, which means that, for R big enough, $\inf_{s\in O\cap B_R}[I(s)+\Phi_0(s)-I_0]=\inf_{s\in O}[I(s)+\Phi_0(s)-I_0]$. Combining these facts the thesis follows. ■

Proof of Proposition 10 The proof follows by Lemma 13. To check the assumptions, note that

$$Q_1,\dots,Q_L\mapsto\Sigma(Q_1,\dots,Q_L)=\begin{pmatrix}\Sigma_{00} & \Sigma_{01}\\ \Sigma_{01}^\top & \Sigma_{11}\end{pmatrix}:=\frac{\tilde{X}^\top\tilde{X}}{N_0\lambda^*}\otimes Q^{(L)}$$

is a continuous function. Moreover, $\Sigma_{11}\mapsto\mathbf{y}_{1:P}^\top(\Sigma_{11}+\beta^{-1}\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P}$ is continuous and bounded (on $\mathcal{S}^+$) and positive, and hence the same holds for

$$Q_1,\dots,Q_L\mapsto\mathbf{y}_{1:P}^\top(\Sigma_{11}(Q_1,\dots,Q_L)+\beta^{-1}\mathbb{1}_{DP})^{-1}\mathbf{y}_{1:P}.$$

Let λ_i's be the eigenvalues of $\beta\Sigma_{11}$, then $\mathbb{1}_{DP} + \beta\Sigma_{11}$ has eigenvalues $1 + \lambda_i$ and

$$\log(\det(\mathbb{1}_{DP} + \beta\Sigma_{11})) = \sum_{i=1}^{DP} \log(1 + \lambda_i).$$

hence $\Sigma_{11} \mapsto \log(\det(\mathbb{1}_{DP}+\beta\Sigma_{11}))$ is bounded on bounded set of Σ_{11}. Using the continuity of $(Q_1, \dots, Q_L) \mapsto \Sigma_{11}(Q_1, \dots, Q_L)$ it follows that also $(Q_1, \dots, Q_L) \mapsto \log(\det(\mathbb{1}_{DP} + \beta\Sigma_{11}))$ is bounded when $\sum_{\ell=1}^{L} \|Q_\ell\|_2 \leq R$ for any R. It remains to check that

$$\lim_{R\to+\infty} \inf_{(Q_1,\dots,Q_L):\sum \|Q_\ell\|_2^2 > R^2} \sum_{\ell=1}^{L} \Big(\operatorname{tr}(Q_\ell) - \log(\det(Q_\ell)) \Big) = +\infty.$$

Letting $\lambda_{i\ell}$ the eigenvalues of Q_ℓ, the claim follows since $\operatorname{tr}(Q_\ell) - \log(\det(Q_\ell)) = \sum_{i=1}^{P} \Big(\lambda_{i\ell} + \log(\lambda_{i\ell})\Big)$. ■

A.3 Proofs for the C-DLN

Lemma 14 *Let $\mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)$ be defined as in (17). Then, if $Q_1, \dots, Q_L$ are strictly positive matrices, then $\mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)$ is positive. If in addition $\sum_{a_0} X_{a_0}^\top X_{a_0}$ is strictly positive, then $\mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)$ is strictly positive.*

Proof First of all note that if $Q_1, \dots, Q_L$ are strictly positive then $\mathcal{T}(Q_1, \dots, Q_L)$ is strictly positive. For $\bar{\mathbf{s}}_{1:P} \neq 0$ write

$$\begin{aligned}
\bar{\mathbf{s}}_{1:P}^\top \mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)\bar{\mathbf{s}}_{1:P} &= \sum_{\mu,\nu} \sum_{a_0} \sum_{r,s} \mathcal{T}(Q_1, \dots, Q_L)_{rs} \frac{x^\mu_{a_0,r} x^\nu_{a_0,s}}{\lambda^* C_0} \bar{s}_\mu \bar{s}_\nu \\
&= \sum_{a_0} \sum_{r,s} \mathcal{T}(Q_1, \dots, Q_L)_{rs} A^*_{r,a_0} A^*_{s,a_0}
\end{aligned}$$

with $A^*_{r,a_0} = \sum_\mu \bar{s}_\mu \frac{x^\mu_{a_0,r}}{\sqrt{\lambda^* C_0}}$. Using $\mathcal{T}(Q_1, \dots, Q_L) > 0$, one gets

$$\sum_{r,s} \mathcal{T}(Q_1, \dots, Q_L)_{rs} A^*_{r,a_0} A^*_{s,a_0} \geq 0$$

for every a_0, so that $\bar{\mathbf{s}}_{1:P}^\top \mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)\bar{\mathbf{s}}_{1:P} \geq 0$. If now $0 = \bar{\mathbf{s}}_{1:P}^\top \mathcal{K}_{\mathrm{C}}(Q_1, \dots, Q_L)\bar{\mathbf{s}}_{1:P}$, then $A^*_{r,a_0} = 0$ for every r and a_0. Hence,

$$0 = \sum_r (A^*_{r,a_0})^2 = \sum_{\mu,\nu} \bar{s}_\mu \bar{s}_\nu \sum_r \frac{x^\mu_{a_0,r} x^\nu_{a_0,r}}{\lambda^* C_0}$$

If $\sum_{a_0} X_{a_0}^\top X_{a_0} > 0$, this is possible only if $\bar{\mathbf{s}}_{1:P} = 0$. ■

Let $K = [K_{i,j,\mu,\nu}]_{i,j,\mu,\nu}$ a four index tensor with $i, j = 1, \dots, N_0$ and $\mu, \nu = 1, \dots, P$ corresponding to a covariance operator. One can identifies K with a symmetric and positive definite $(N_0 \times P) \times (N_0 \times P)$ matrix $K = [K_{i,j,\mu,\nu}]_{(i,\mu),(j,\nu)}$, where the multi-indices (i, μ)'s are properly ordered from 1 to $N_0 \times P$. With a slight abuse of language, we will use in the following $K^{(\ell-1)}$ to denote either the four indices tensor or its matrix representation, the case being clear from the context.

Lemma 15 *Let* $\mathbf{s}$ *be a vector in* $\mathbb{R}^P$*, then*

$$\det[\mathbb{1}_{N_0} \otimes \mathbb{1}_P + [\mathbb{1}_{N_0} \otimes (\mathbf{s}\mathbf{s}^\top)]K] = \det\Big(\mathbb{1}_{N_0} + \sum_{\mu,\nu} s_\mu s_\nu K_{\cdot,\mu\nu}\Big)\,, \tag{32}$$

where $K_{\cdot,\mu\nu} = [K_{i,j,\mu,\nu}]_{i,j=1}^{N_0}$.

Proof The eigenvalues of $[\mathbb{1}_{N_0} \otimes (\mathbf{s}\mathbf{s}^\top)]K$ are given by

$$[\mathbb{1}_{N_0} \otimes (\mathbf{s}\mathbf{s}^\top)]Kv = \lambda v$$

which reads, in components, as

$$\sum_{\mu,\nu,j} s_\mu K_{(i,\mu)(j,\nu)} v_{(j,\nu)} s_\rho = \lambda v_{(i,\rho)}\,.$$

This shows that if $\lambda \neq 0$ then $v_{(i,\mu)} = c_i s_\mu$ for some $\mathbf{c} \in \mathbb{R}^{N_0}$, such that

$$\sum_j \Bigg(\sum_{\mu,\nu} s_\mu K_{\cdot,\mu\nu} s_\nu\Bigg)_{ij} c_j = \lambda c_i\,,$$

that is $\lambda \in \mathrm{Sp}(\sum_{\mu,\nu} s_\mu K_{\cdot,\mu\nu} s_\nu)$, where Sp denotes the spectrum. The cardinality of this spectrum is N_0, all the other $N_0(P-1)$ eigenvalues must be zero. Hence,

$$\mathrm{Sp}(\mathbb{1}_{N_0} \otimes \mathbb{1}_P + [\mathbb{1}_{N_0} \otimes (\mathbf{s}\mathbf{s}^\top)]K) = \mathrm{Sp}\Big(\mathbb{1}_{N_0} + \sum_{\mu,\nu} s_\mu K_{\cdot,\mu\nu} s_\nu\Big) \cup \{1\}^{N_0(P-1)}\,,$$

and the thesis follows. ■

Proof of Proposition 5 From Eq. (5), we have

$$\mathbf{E}[e^{i\bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}}] = \mathbf{E}\left[\exp\left(-\frac{1}{2}\sum_{\mu,\nu} \bar{s}^\mu \bar{s}^\nu \sum_{a_L=1}^{C_L} \sum_{i=1}^{N_0} \frac{1}{\lambda_L C_L N_0} h^{(L)}_{a_L,i,\mu} h^{(L)}_{a_L,i,\nu}\right)\right]. \tag{33}$$

Let $h^{(\ell)}$ be the collection of all the variables $h^{(\ell)}_{a_\ell,i,\mu} = h^{(\ell)}_{a_\ell,i}(\mathbf{x}^\mu)$ where $a_\ell = 1, \dots, C_\ell$, $i = 1, \dots, N_0$, $\mu = 1, \dots, P$. By (4), it is clear that, conditionally on $h^{(\ell-1)}$, $h^{(\ell)}$ is a collection of jointly Gaussian random variables (a Gaussian field) with zero mean and covariance function given by

$$\mathrm{Cov}(h_{a,i,\mu}, h_{b,j,\nu}|h^{(\ell-1)}) = \delta_{a,b} K^{(\ell-1)}_{i,j,\mu,\nu}$$

where

$$K^{(\ell-1)}_{i,j,\mu,\nu} = \frac{1}{\lambda_{\ell-1}C_{\ell-1}M}\sum_{a_{\ell-1}=1}^{C_{\ell-1}}\sum_{m=-\lfloor M/2\rfloor}^{\lfloor M/2\rfloor}\left(\sum_{r=1}^{N_0}T_{m,ir}h^{(\ell-1)}_{a_{\ell-1},r,\mu}\right)\left(\sum_{s=1}^{N_0}T_{m,js}h^{(\ell-1)}_{a_{\ell-1},s,\nu}\right). \tag{34}$$

Considering $h^{(\ell)}$ as a vector of dimension $C_\ell \times P \times N_0$, the conditional distribution of $h^{(\ell)}$ given $h^{(\ell-1)}$ is

$$h^{(\ell)}|h^{(\ell-1)} \sim \mathcal{N}(0, \mathbb{1}_{C_\ell} \otimes K^{(\ell-1)})$$

(see the comment before Lemma 15 for the matrix representation of the tensor $K^{(\ell-1)}$). At this stage, by Gaussian integration, form (33) one gets

$$\begin{aligned}
\mathbf{E}[e^{i\bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}}] &= \mathbf{E}\Bigg[\det\left((\mathbb{1}_{C_L}\otimes K^{(L-1)})^{-1} + \frac{1}{\lambda_L C_L N_0}\mathbb{1}_{C_L}\otimes\mathbb{1}_{N_0}\otimes(\bar{\mathbf{s}}_{1:P}\bar{\mathbf{s}}_{1:P}^\top)\right)^{-\frac{1}{2}} \\
&\qquad \det(\mathbb{1}_{C_L}\otimes K^{(L-1)})^{-\frac{1}{2}}\Bigg] \\
&= \mathbf{E}\left[\det\left(\mathbb{1}_{N_0}\otimes\mathbb{1}_P + \frac{1}{\lambda_L C_L N_0}[\mathbb{1}_{N_0}\otimes(\bar{\mathbf{s}}_{1:P}\bar{\mathbf{s}}_{1:P}^\top)]K^{(L-1)}\right)^{-\frac{C_L}{2}}\right] \\
&= \mathbf{E}\left[\det\left(\mathbb{1}_{N_0} + \frac{1}{\lambda_L C_L N_0}\sum_{\mu,\nu}\bar{s}_\mu\bar{s}_\nu K^{(L-1)}_{\cdot,\mu\nu}\right)^{-\frac{C_L}{2}}\right].
\end{aligned} \tag{35}$$

The last step follows from Lemma 15. Using Eq. (23), we get

$$\begin{aligned}
&\mathbf{E}[e^{i\bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}}] \\
&= \mathbf{E}\left[\int_{\mathcal{S}^+_{N_0}}\exp\left\{-\frac{1}{2\lambda_L N_0}\sum_{i,j}Q_{L,ij}\sum_{\mu,\nu}\bar{s}_\mu\bar{s}_\nu K^{(L-1)}_{ij,\mu\nu}\right\}\mathcal{W}_{N_0}\left(dQ_L|\frac{1}{C_L}\mathbb{1}_{N_0}, C_L\right)\right] \\
&= \mathbf{E}\Bigg[\int_{\mathcal{S}^+_{N_0}}\exp\left\{-\frac{1}{2\lambda_L\lambda_{L-1}C_{L-1}N_0}\sum_{i,j}Q^*_{L,ij}\sum_{\mu,\nu}\bar{s}_\mu\bar{s}_\nu\sum_{a_{L-1}=1}^{C_{L-1}}h^{(L-1)}_{a_{L-1},i,\mu}h^{(L-1)}_{a_{L-1},j,\nu}\right\} \\
&\qquad \mathcal{W}_{N_0}\left(dQ_L|\frac{1}{C_L}\mathbb{1}_{N_0}, C_L\right)\Bigg]
\end{aligned} \tag{36}$$

where we used Eq. (15).

We could in principle proceed directly integrating out $h^{(L-1)}$ as we did for $h^{(L)}$, Eq. (35); however, in this way we would end up with a Wishart measure for Q_{L-1} dependent on Q_L. To keep factorized the different contributions, it is easier to change variables with the linear transformation

$$\tilde{h}^{(L-1)}_{a_{L-1},i,\mu} = \sum_j U^*_{L,ij}h^{(L-1)}_{a_{L-1},j,\mu}, \tag{37}$$

where $(U^*_L)^\top U^*_L = Q^*_L$. It is clear that the conditional distribution of $\tilde{h}^{(L-1)}$ given $h^{(L-2)}$ is again Gaussian, more precisely

$$\tilde{h}^{(L-1)}|h^{(L-2)} \sim \mathcal{N}(0, \mathbb{1}_{C_{L-1}}\otimes[U^*_L K^{(L-2)}(U^*_L)^\top]), \tag{38}$$

with the slight abuse of notation $U_L^* \equiv U_L^* \otimes \mathbb{1}_P$. Proceeding as before,

$$\begin{aligned}\mathbf{E}[e^{i\bar{\mathbf{s}}_{1:P}^\top \mathbf{S}_{1:P}}] &= \mathbf{E}\Bigg[\int_{(\mathcal{S}_{N_0}^+)^2} \exp\Bigg\{-\frac{1}{2\lambda_L\lambda_{L-1}N_0}\operatorname{tr}\Bigg(Q_{L-1}\sum_{\mu,\nu}\bar{s}_\mu\bar{s}_\nu[U_L^* K_{\cdot,\mu\nu}^{(L-2)}(U_L^*)^\top]\Bigg)\Bigg\} \\ &\qquad \prod_{\ell\in\{L,L-1\}}\mathcal{W}_{N_0}(dQ_\ell|\frac{1}{C_\ell}\mathbb{1}_{N_0},C_\ell)\Bigg] \\ &= \mathbf{E}\Bigg[\int_{(\mathcal{S}_{N_0}^+)^2}\exp\Bigg\{-\frac{1}{2\lambda_L\lambda_{L-1}\lambda_{L-2}C_{L-2}N_0}\sum_{i,j}Q_{L-1,ij}^*\sum_{\mu,\nu}\bar{s}_\mu\bar{s}_\nu \\ &\qquad \sum_{a_{L-2}=1}^{C_{L-2}} h_{a_{L-2},i,\mu}^{(L-2)} h_{a_{L-2},j,\nu}^{(L-2)}\Bigg\}\times\prod_{\ell\in\{L,L-1\}}\mathcal{W}_{N_0}\Big(dQ_\ell|\frac{1}{C_\ell}\mathbb{1}_{N_0},C_\ell\Big)\Bigg]\end{aligned} \tag{39}$$

where Q_{L-1}^* is again given by Eq. (15). Proposition 5 is proven once this procedure is straightforwardly iterated. ∎

Proof of Proposition 7 The proof of Proposition 7 is identical to the proof of Proposition 6 once one observes that the matrix Σ, which is $\mathcal{K}_{\rm C}(Q_1,\ldots,Q_L)$ for the enlarged dataset $\tilde{X}$, is strictly positive by Lemma 14. ∎

References

F. Aguirre-López, S. Franz, and M. Pastore. Random features and polynomial rules. *SciPost Phys.*, 18:039, 2025. doi: 10.21468/SciPostPhys.18.1.039. URL `https://scipost.org/10.21468/SciPostPhys.18.1.039`.

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`. eprint: `https://arxiv.org/abs/2307.11807`.

P. Baglioni, R. Pacelli, R. Aiudi, F. Di Renzo, A. Vezzani, R. Burioni, and P. Rotondo. Predictive power of a Bayesian effective action for fully connected one hidden layer neural networks in the proportional limit. *Phys. Rev. Lett.*, 133:027301, Jul 2024. doi: 10.1103/PhysRevLett.133.027301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.133.027301`. eprint: `https://arxiv.org/abs/2401.11004`.

J.-M. Bardet and D. Surgailis. Moment bounds and central limit theorems for Gaussian subordinated arrays. *Journal of Multivariate Analysis*, 114:457–473, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.002. URL `https://www.sciencedirect.com/science/article/pii/S0047259X1200200X`.

F. Bassetti, L. Ladelli, and P. Rotondo. Proportional infinite-width infinite-depth limit for deep linear neural networks, 2024. eprint: `https://arxiv.org/abs/2411.15267`.

B. Bordelon and C. Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32240–32256. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d027a5c93d484a4312cc486d399c62c1-Paper-Conference.pdf.

P. Breuer and P. Major. Central limit theorems for non-linear functionals of Gaussian fields. *Journal of Multivariate Analysis*, 13(3):425–441, 1983. ISSN 0047-259X. doi: 10.1016/0047-259X(83)90019-2. URL https://www.sciencedirect.com/science/article/pii/0047259X83900192.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

F. Camilli, D. Tieplova, and J. Barbier. Fundamental limits of overparametrized shallow neural networks for supervised learning, 2023. eprint: https://arxiv.org/abs/2307.05635.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. 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/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

L. Chizat, M. Colombo, X. Fernández-Real, and A. Figalli. Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, n/a(n/a), 2022. doi: 10.1002/cpa.22200. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22200.

S. Ciceri, L. Cassani, M. Osella, P. Rotondo, F. Valle, and M. Gherardi. Inversion dynamics of class manifolds in deep learning reveals tradeoffs underlying generalization. *Nature Machine Intelligence*, 6(1):40–47, 2024. doi: 10.1038/s42256-023-00772-9. URL https://doi.org/10.1038/s42256-023-00772-9. eprint: https://arxiv.org/abs/2303.05161.

H. Cui, F. Krzakala, and L. Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6468–6521. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cui23b.html.

A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. ISBN 978-3-642-03310-0. doi: 10.1007/978-3-642-03311-7. URL https://doi.org/10.1007/978-3-642-03311-7.

M. L. Eaton. *Multivariate statistics: A vector space approach*, volume 53 of *Institute of Mathematical Statistics Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH, 2007. ISBN 978-0-940600-69-0; 0-940600-69-2. doi: 10.1214/lnms/1196285102. URL https://doi.org/10.1214/lnms/1196285102.

A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: 10.1017/CBO9781139164542. URL https://doi.org/10.1017/CBO9781139164542.

S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati. Quantitative CLTs in deep neural networks. *Probability Theory and Related Fields*, Feb 2025. ISSN 1432-2064. doi: 10.1007/s00440-025-01360-1. URL https://doi.org/10.1007/s00440-025-01360-1. eprint: https://arxiv.org/abs/2307.06092.

K. Fischer, J. Lindner, D. Dahmen, Z. Ringel, M. Krämer, and M. Helias. Critical feature learning in deep neural networks. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13660–13690. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fischer24a.html.

Y. V. Fyodorov. Negative moments of characteristic polynomials of random matrices: Ingham–Siegel integral as an alternative to Hubbard–Stratonovich transformation. *Nuclear Physics B*, 621(3):643–674, 2002. ISSN 0550-3213. doi: 10.1016/S0550-3213(01)00508-9. URL https://www.sciencedirect.com/science/article/pii/S0550321301005089.

A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklfsi0cKm.

M. Geiger, S. Spigler, A. Jacot, and M. Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (11):113301, nov 2020. doi: 10.1088/1742-5468/abc4de. URL https://dx.doi.org/10.1088/1742-5468/abc4de.

M. Geiger, L. Petrini, and M. Wyart. Landscape and training regimes in deep learning. *Physics Reports*, 924:1–18, 2021. ISSN 0370-1573. doi: 10.1016/j.physrep.2021.04.001. URL https://www.sciencedirect.com/science/article/pii/S0370157321001290.

F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gerace20a.html.

F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborová. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109:034305, Mar 2024. doi: 10.1103/PhysRevE.109.034305. URL https://link.aps.org/doi/10.1103/PhysRevE.109.034305.

S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10: 041044, 12 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mezard, and L. Zdeborova. The Gaussian equivalence of generative models for learning with shallow neural networks. In J. Bruna, J. Hesthaven, and L. Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 16–19 Aug 2022. URL https://proceedings.mlr.press/v145/goldt22a.html.

A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2000. ISBN 1-58488-046-5. eBook.

B. Hanin. Random neural networks in the infinite width limit as Gaussian processes. *Ann. Appl. Probab.*, 33(6A):4798–4819, 2023. ISSN 1050-5164,2168-8737. doi: 10.1214/23-aap1933. URL https://doi.org/10.1214/23-aap1933.

B. Hanin. Random fully connected neural networks as perturbatively solvable hierarchies. *Journal of Machine Learning Research*, 25(267):1–58, 2024. URL http://jmlr.org/papers/v25/23-0643.html.

B. Hanin and A. Zlokapa. Bayesian interpolation with deep linear networks. *Proceedings of the National Academy of Sciences*, 120(23):e2301345120, 2023. doi: 10.1073/pnas.2301345120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2301345120.

J. N. Howard, R. Jefferson, A. Maiti, and Z. Ringel. Wilsonian renormalization of Neural Network Gaussian Processes, 2024. eprint: https://arxiv.org/abs/2405.06008.

J. N. Howard, M. S. Klinger, A. Maiti, and A. G. Stapleton. Bayesian RG flow in neural network field theories. *SciPost Phys. Core*, 8:027, 2025. doi: 10.21468/SciPostPhysCore.8.1.027. URL https://scipost.org/10.21468/SciPostPhysCore.8.1.027.

H. Hu and Y. M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698. eprint: https://arxiv.org/abs/2009.07669.

A. E. Ingham. An integral which occurs in statistics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(2):271–276, 1933. doi: 10.1017/S0305004100011075.

A. Ingrosso and S. Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119(40):e2201854119, 2022. doi: 10.1073/pnas.2201854119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2201854119.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

A. H. Joarder and A. U. Alam. The characteristic function of elliptical t-distribution using a conditional expectation approach. *J. Inform. Optim. Sci.*, 16(2):307–310, 1995. ISSN 0252-2667. doi: 10.1080/02522667.1995.10699228. URL https://doi.org/10.1080/02522667.1995.10699228.

A. M. Kshirsagar. Bartlett decomposition and Wishart distribution. *The Annals of Mathematical Statistics*, 30(1):239–241, 1959. ISSN 00034851. URL http://www.jstor.org/stable/2237140.

C. Lauditi, B. Bordelon, and C. Pehlevan. Adaptive kernel predictors from feature-learning infinite limits of neural networks, 2025. eprint: https://arxiv.org/abs/2502.07998.

J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. The large learning rate phase of deep learning, 2021. URL https://openreview.net/forum?id=TDDZxmr6851.

M. Li, M. Nica, and D. Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10795–10808. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/45fc4a0da7e7f6fbabaabe2d20a441d1-Paper-Conference.pdf.

Q. Li and H. Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. doi: 10.1103/PhysRevX.11.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031059.

Q. Li and H. Sompolinsky. Globally gated deep linear networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34789–34801. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e0f393e7980a24fd12fa6f15adfa25fb-Paper-Conference.pdf.

P. E. Lin. Some characterizations of the multivariate t distribution. *J. Multivariate Anal.*, 2:339–344, 1972. ISSN 0047-259X. doi: 10.1016/0047-259X(72)90021-8. URL https://doi.org/10.1016/0047-259X(72)90021-8.

S. Mei, A. Montanari, and P.-M. 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. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

G. Naveh and Z. Ringel. A self consistent theory of Gaussian processes captures feature learning effects in finite CNNs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21352–21364. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/b24d21019de5e59da180f1661904f49a-Paper.pdf.

R. 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.

I. Nourdin, G. Peccati, and M. Podolskij. Quantitative Breuer–Major theorems. *Stochastic Processes and their Applications*, 121(4):793–812, 2011. ISSN 0304-4149. doi: 10.1016/j.spa.2010.12.006. URL https://www.sciencedirect.com/science/article/pii/S0304414910002917.

R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

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*, 5:1497–1507, 2023. doi: 10.1038/s42256-023-00767-6. eprint: https://arxiv.org/abs/2209.04882.

L. Petrini, F. Cagnetta, U. M. Tomasini, A. Favero, and M. Wyart. How deep neural networks learn compositional data: The random hierarchy model. *CoRR*, abs/2307.02129, 2023a. URL https://doi.org/10.48550/arXiv.2307.02129.

L. Petrini, F. Cagnetta, E. Vanden-Eijnden, and M. Wyart. Learning sparse features can lead to overfitting in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114003, nov 2023b. doi: 10.1088/1742-5468/ad01b9. URL https://dx.doi.org/10.1088/1742-5468/ad01b9.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL https://doi.org/10.7551/mitpress/3206.001.0001.

G. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022. doi: 10.1002/cpa.22074. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22074.

N. Rubin, Z. Ringel, I. Seroussi, and M. Helias. A unified approach to feature learning in Bayesian neural networks. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. URL https://openreview.net/forum?id=ZmOSJ2MV2R.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6120.

A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019. doi: 10.1073/pnas.1820226116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1820226116.

A. Sclocchi and M. Wyart. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences*, 121(9):e2316301121, 2024. doi: 10.1073/pnas.2316301121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2316301121.

A. Sclocchi, M. Geiger, and M. Wyart. Dissecting the effects of SGD noise in distinct regimes of deep learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30381–30405. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sclocchi23a.html.

I. Seroussi, G. Naveh, and Z. Ringel. Separation of scales and a thermodynamic description of feature learning in some CNNs. *Nature Communications*, 14(1):908, 02 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://doi.org/10.1038/s41467-023-36361-y. eprint: https://arxiv.org/abs/2112.15383.

C. L. Siegel. Uber die analytische theorie der quadratischen formen. *Annals of Mathematics*, 36(3):527–606, 1935. ISSN 0003486X. URL http://www.jstor.org/stable/1968644.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020. doi: 10.1137/18M1192184. URL https://doi.org/10.1137/18M1192184.

M. D. Springer and W. E. Thompson. The distribution of products of beta, gamma and Gaussian random variables. *SIAM Journal on Applied Mathematics*, 18(4):721–737, 1970. ISSN 00361399. URL http://www.jstor.org/stable/2099424.

D. Trevisan. Wide deep neural networks with Gaussian weights are very close to Gaussian processes, 2023. eprint: https://arxiv.org/abs/2312.11737.

G. Yang and E. J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

J. Zavatone-Veth and C. Pehlevan. Exact marginal prior distributions of finite Bayesian neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3364–3375. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1baff70e2669e8376347efd3a874a341-Paper.pdf.

J. A. Zavatone-Veth, W. L. Tong, and C. Pehlevan. Contrasting random and learned features in deep Bayesian linear regression. *Phys. Rev. E*, 105:064118, Jun 2022. doi: 10.1103/PhysRevE.105.064118. URL https://link.aps.org/doi/10.1103/PhysRevE.105.064118.

F. Zhang, editor. *The Schur complement and its applications*, volume 4 of *Numerical Methods and Algorithms*. Springer-Verlag, New York, 2005. ISBN 0-387-24271-6. doi: 10.1007/b105056. URL https://doi.org/10.1007/b105056.