Entropy and mutual information in models of deep neural networks

Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, Lenka Zdeborová

Entropy and mutual information in models of deep neural networks

Marylou Gabrié*[1], Andre Manoel[2,3], Clément Luneau[4], Jean Barbier[4], Nicolas Macris[4], Florent Krzakala[1,5,6,7], Lenka Zdeborová[3,6]

[1]Laboratoire de Physique Statistique, École Normale Supérieure, Paris
[2]Parietal Team, INRIA Saclay Île-de-France
[3]Institut de Physique Théorique, Université Paris-Saclay, CNRS, CEA
[4]Laboratoire de Théorie des Communications, École Polytechnique Fédérale de Lausanne
[5]Sorbonne Universités
[6]Department of Mathematics, Duke University, Durham
[7]LightOn, Paris

May 25, 2018

Abstract

We examine a class of deep learning models with a tractable method to compute information-theoretic quantities. Our contributions are three-fold: (i) We show how entropies and mutual informations can be derived from heuristic statistical physics methods, under the assumption that weight matrices are independent and orthogonally-invariant. (ii) We extend particular cases in which this result is known to be rigorously exact by providing a proof for two-layers networks with Gaussian random weights, using the recently introduced adaptive interpolation method. (iii) We propose an experiment framework with generative models of synthetic datasets, on which we train deep neural networks with a weight constraint designed so that the assumption in (i) is verified during learning. We study the behavior of entropies and mutual informations throughout learning and conclude that, in the proposed setting, the relationship between compression and generalization remains elusive.

The successes of deep learning methods have spurred efforts towards quantitative modeling of the performance of deep neural networks. In particular, an information-theoretic approach linking generalization capabilities to compression has been receiving increasing interest. The intuition behind the study of mutual informations in latent variable models dates back to the information bottleneck (IB) theory of [1]. Although recently reformulated in the context of deep learning [2], verifying its relevance in practice requires the computation of mutual informations for high-dimensional variables, a notoriously hard problem. Thus, pioneering works in this direction focused either on small network models with discrete (continuous, eventually binned) activations [3], or on linear networks [4, 5].

In the present paper we follow a different direction, and build on recent results from statistical physics [6, 7] and information theory [8, 9] to propose, in Section 1, a formula to compute information-theoretic quantities for a class of deep neural network models. The models we approach, described in

*Corresponding author: marylou.gabrie@ens.fr

Section 2, are non-linear feed-forward neural networks trained on synthetic datasets with constrained weights. Such networks capture some of the key properties of the deep learning setting that are usually difficult to include in tractable frameworks: non-linearities, arbitrary large width and depth, and correlations in the input data. We demonstrate the proposed method in a series of numerical experiments in Section 3. First observations suggest a rather complex picture, where the role of compression in the generalization ability of deep neural networks is yet to be elucidated.

1 Multi-layer model and main theoretical results

A stochastic multi-layer model— We consider a model of multi-layer stochastic feed-forward neural network where each element x_i of the input layer $\boldsymbol{x} \in \mathbb{R}^{N_0}$ is distributed independently as $P_0(x_i)$, while hidden units $t_{\ell,i}$ at each successive layer $\boldsymbol{t}_\ell \in \mathbb{R}^{N_\ell}$ (vectors are column vectors) come from $P_\ell(t_{\ell,i}|\boldsymbol{W}_{\ell,i}^\intercal \boldsymbol{t}_{\ell-1})$, with $\boldsymbol{t}_0 \equiv \boldsymbol{x}$ and $\boldsymbol{W}_{\ell,i}$ denoting the i-th row of the matrix of weights $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$. In other words

$$t_{0,i} \equiv x_i \sim P_0(\cdot), \quad t_{1,i} \sim P_1(\cdot|\boldsymbol{W}_{1,i}^\intercal \boldsymbol{x}), \quad \dots \quad t_{L,i} \sim P_L(\cdot|\boldsymbol{W}_{L,i}^\intercal \boldsymbol{t}_{L-1}), \tag{1}$$

given a set of weight matrices $\{W_\ell\}_{\ell=1}^L$ and distributions $\{P_\ell\}_{\ell=1}^L$ which encode possible non-linearities and stochastic noise applied to the hidden layer variables, and P_0 that generates the visible variables. In particular, for a non-linearity $t_{\ell,i} = \varphi_\ell(h, \xi_{\ell,i})$, where $\xi_{\ell,i} \sim P_\xi(\cdot)$ is the stochastic noise (independent for each i), we have $P_\ell(t_{\ell,i}|\boldsymbol{W}_{\ell,i}^\intercal \boldsymbol{t}_{\ell-1}) = \int dP_\xi(\xi_{\ell,i})\, \delta\big(t_{\ell,i} - \varphi_\ell(\boldsymbol{W}_{\ell,i}^\intercal \boldsymbol{t}_{\ell-1}, \xi_{\ell,i})\big)$. Model (1) thus describes a Markov chain which we denote by $\boldsymbol{X} \to \boldsymbol{T}_1 \to \boldsymbol{T}_2 \to \cdots \to \boldsymbol{T}_L$, with $\boldsymbol{T}_\ell = \varphi_\ell(W_\ell \boldsymbol{T}_{\ell-1}, \boldsymbol{\xi}_\ell)$, $\boldsymbol{\xi}_\ell = \{\xi_{\ell,i}\}_{i=1}^{N_\ell}$, and the activation function φ_ℓ applied componentwise.

Replica formula— We shall work in the asymptotic high-dimensional statistics regime where all $\tilde{\alpha}_\ell \equiv N_\ell/N_0$ are of order one while $N_0 \to \infty$, and make the important assumption that all matrices W_ℓ are orthogonally-invariant random matrices independent from each other; in other words, each matrix $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ can be decomposed as a product of three matrices, $W_\ell = U_\ell S_\ell V_\ell$, where $U_\ell \in \mathrm{O}(N_\ell)$ and $V_\ell \in \mathrm{O}(N_{\ell-1})$ are independently sampled from the Haar measure, and S_ℓ is a diagonal matrix of singular values. The main technical tool we use is a formula for the entropies of the hidden variables, $H(\boldsymbol{T}_\ell) = -\mathbb{E}_{\boldsymbol{T}_\ell} \ln P_{\boldsymbol{T}_\ell}(\boldsymbol{t}_\ell)$, and the mutual information between adjacent layers $I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1}) = H(\boldsymbol{T}_\ell) + \mathbb{E}_{\boldsymbol{T}_\ell, \boldsymbol{T}_{\ell-1}} \ln P_{\boldsymbol{T}_\ell|\boldsymbol{T}_{\ell-1}}(\boldsymbol{t}_\ell|\boldsymbol{t}_{\ell-1})$, based on the heuristic replica method [10, 11, 6, 7, 8, 9]:

Claim 1 (Replica formula)**.** *Assume model (1) with L layers in the high-dimensional limit with componentwise activation functions and weight matrices generated from the ensemble described above, and denote by λ_{W_k} the eigenvalues of $W_k^\intercal W_k$. Then for any $\ell \in \{1, \dots, L\}$ the normalized entropy of $\boldsymbol{T}_\ell$ is given by the minimum among all stationary points of the replica potential:*

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\boldsymbol{T}_\ell) = \min \operatorname*{extr}_{\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}} \phi_\ell(\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}), \tag{2}$$

which depends on ℓ-dimensional vectors $\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}$, and is written in terms of mutual information

I and conditional entropies H of scalar variables as

$$\phi_\ell(\boldsymbol{A},\boldsymbol{V},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{V}}) = I\Big(t_0; t_0+\frac{\xi_0}{\sqrt{\tilde{A}_1}}\Big) - \frac{1}{2}\sum_{k=1}^{\ell}\tilde{\alpha}_{k-1}\big[\tilde{A}_k V_k + \alpha_k A_k \tilde{V}_k - F_{W_k}(A_k V_k)\big] + \sum_{k=1}^{\ell-1}\tilde{\alpha}_k\Big[H(t_k|\xi_k;\tilde{A}_{k+1},\tilde{V}_k,\tilde{\rho}_k) - \frac{1}{2}\log(2\pi e\tilde{A}_{k+1})\Big] + \tilde{\alpha}_\ell H(t_\ell|\xi_\ell;\tilde{V}_\ell,\tilde{\rho}_\ell), \tag{3}$$

where $\alpha_k = N_k/N_{k-1}$, $\tilde{\alpha}_k = N_k/N_0$, $\rho_k = \int dP_{k-1}(t)\, t^2$, $\tilde{\rho}_k = (\mathbb{E}_{\lambda_{W_k}}\lambda_{W_k})\rho_k/\alpha_k$, *and* $\xi_k \sim \mathcal{N}(0,1)$ *for* $k = 0,\dots,\ell$. *In the computation of the conditional entropies in (3), the scalar* t_k*-variables are generated from* $P(t_0) = P_0(t_0)$ *and*

$$P(t_k|\xi_k;A,V,\rho) = \mathbb{E}_{\tilde{\xi},\tilde{z}}\, P_k(t_k + \tilde{\xi}/\sqrt{A}|\sqrt{\rho - V}\xi_k + \sqrt{V}\tilde{z}), \quad k = 1,\dots,\ell-1, \tag{4}$$

$$P(t_\ell|\xi_\ell;V,\rho) = \mathbb{E}_{\tilde{z}}\, P_\ell(t_\ell|\sqrt{\rho - V}\xi_\ell + \sqrt{V}\tilde{z}), \tag{5}$$

where $\tilde{\xi}$ *and* $\tilde{z}$ *are independent* $\mathcal{N}(0,1)$ *random variables. Finally, the function* $F_{W_k}(x)$ *depends on the distribution of the eigenvalues* λ_{W_ℓ} *following*

$$F_{W_k}(x) = \min_{\theta\in\mathbb{R}}\ \big\{2\alpha_k\theta + (\alpha_k - 1)\ln(1-\theta) + \mathbb{E}_{\lambda_{W_k}}\ln[x\lambda_{W_k} + (1-\theta)(1-\alpha_k\theta)]\big\}. \tag{6}$$

The computation of the entropy in the large dimensional limit, a computationally difficult task, has thus been reduced to an extremization of a function of 4ℓ variables, that requires evaluating single or bidimensional integrals. This extremization can be done efficiently, as detailed in the Supplementary Material; a user-friendly Python package is provided [12], which performs the computation for different choices of prior P_0, activations φ_ℓ and spectra λ_{W_ℓ}. Finally, the mutual information between successive layers $I(\boldsymbol{T}_\ell;\boldsymbol{T}_{\ell-1})$ can be obtained from the entropy following the evaluation of an additional bidimensional integral, see Section 1.6.1 of the Supplementary Material.

Our approach in the derivation of (3) builds on recent progresses in statistical estimation and information theory for generalized linear models following the application of methods from statistical physics of disordered systems [10, 11] in communication [13], statistics [14] and machine learning problems [15, 16]. In particular, we use advanced mean field theory [17] and the heuristic replica method [10, 6], along with its recent extension to multi-layer estimation [7, 8, 9], in order to derive the above formula (3). This derivation is lengthy and thus given in the Supplementary Material.

Rigorous statement— We recall the assumptions under which the replica formula of Claim 1 is conjectured to be exact: *(i) weight matrices are drawn from an ensemble of random orthogonally-invariant matrices, (ii) matrices at different layers are statistically independent and (iii) layers have a large dimension and respective sizes of adjacent layers are such that weight matrices have aspect ratios* $\{\alpha_k,\tilde{\alpha}_k\}_{k=1}^{\ell}$ *of order one.* While we could not prove the replica prediction in full generality, we stress that it comes with multiple credentials: (i) for Gaussian prior P_0 and Gaussian distributions P_ℓ, it corresponds to the exact analytical solution when weight matrices are independent of each other (see Section 1.6.2 of the Supplementary Material). (ii) In the single-layer case with a Gaussian weight matrix, it reduces to formula (13) in the Supplementary Material, which has been recently rigorously proven for (almost) all activation functions φ [18]. (iii) In the case of Gaussian distributions P_ℓ, it has also been proven for a large ensemble of random matrices [19] and (iv) it is consistent with all the results of the AMP [20, 21, 22] and VAMP [23] algorithms, known to perform well for these

estimation problems. An equivalent formula was proposed by Reeves in [9] using different heuristic arguments.

In order to go beyond results for the single-layer problem and heuristic arguments, we prove Claim 1 for the more involved multi-layer case, assuming Gaussian i.i.d. matrices and two non-linear layers:

Theorem 1 (Two-layer Gaussian replica formula)**.** *Suppose* $(H1)$ *the input units distribution* P_0 *is separable and has bounded support;* $(H2)$ *the activations* φ_1 *and* φ_2 *corresponding to* $P_1(t_{1,i}|\boldsymbol{W}_{1,i}^\intercal \boldsymbol{x})$ *and* $P_2(t_{2,i}|\boldsymbol{W}_{2,i}^\intercal \boldsymbol{t}_1)$ *are bounded* $\mathcal{C}^2$ *with bounded first and second derivatives w.r.t their first argument; and* $(H3)$ *the weight matrices* W_1*,* W_2 *have Gaussian i.i.d. entries. Then for model* (1) *with two layers* $L = 2$ *the high-dimensional limit of the entropy verifies Claim 1.*

The theorem, that proves the conjecture presented in [7], is proven using the adaptive interpolation method of [24, 18] in a multi-layer setting, as first developed in [25]. The lengthy proof, presented in details in the Supplementary Material, is of independent interest and adds further credentials to the replica formula, as well as offers a clear direction to further developments. Note that, following the same approximation arguments as in [18] where the proof is given for the single-layer case, the hypothesis $(H1)$ can be relaxed to the existence of the second moment of the prior, $(H2)$ can be dropped and $(H3)$ extended to matrices with i.i.d. entries of zero mean, $O(1/N_0)$ variance and finite third moment.

2 Tractable models for deep learning

The multi-layer model presented above can be leveraged to simulate two prototypical settings of deep supervised learning on synthetic datasets amenable to the replica tractable computation of entropies and mutual informations.

The first scenario is the so-called *teacher-student* (see Figure 1, left). Here, we assume that the input $\boldsymbol{x}$ is distributed according to a *separable* prior distribution $P_X(\boldsymbol{x}) = \prod_i P_0(x_i)$, factorized in the components of $\boldsymbol{x}$, and the corresponding label $\boldsymbol{y}$ is given by applying a mapping $\boldsymbol{x} \to \boldsymbol{y}$, called *the teacher*. After generating a train and test set in this manner, we perform the training of a deep neural network, *the student*, on the synthetic dataset. In this case, the data themselves have a simple structure given by P_0.

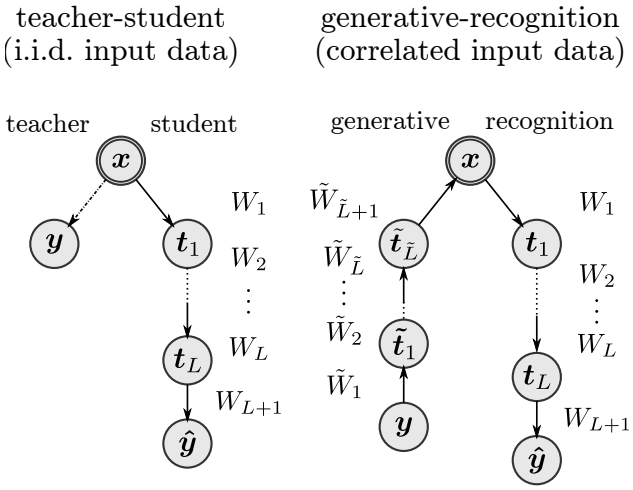

Figure 1: Two models of synthetic data

In constrast, the second scenario allows *generative models* (see Figure 1, right) that create more structure, and that are reminiscent of the *generative-recognition* pair of models of a Variational Autoencoder (VAE). A code vector $\boldsymbol{y}$ is sampled from a separable prior distribution $P_Y(\boldsymbol{y}) = \prod_i P_0(y_i)$ and a corresponding data point $\boldsymbol{x}$ is generated by a possibly stochastic neural network, *the generative model*. This setting allows to create input data $\boldsymbol{x}$ featuring correlations,

differently from the teacher-student scenario. The studied supervised learning task then consists in training a deep neural net, *the recognition model*, to recover the code $\boldsymbol{y}$ from $\boldsymbol{x}$.

In both cases, the chain going from $\boldsymbol{X}$ to any later layer is a Markov chain in the form of (1). In the first scenario, model (1) directly maps to the student network. In the second scenario however, model (1) actually maps to the feed-forward combination of the generative model followed by the recognition model. This shift is necessary to verify the assumption that the starting point (now given by $\boldsymbol{Y}$) has a separable distribution. In particular, it generates correlated input data $\boldsymbol{X}$ while still allowing for the computation of the entropy of any $\boldsymbol{T}_\ell$.

At the start of a neural network training, weight matrices initialized as i.i.d. Gaussian random matrices satisfy the necessary assumptions of the formula of Claim 1. In their singular value decomposition

$$W_\ell = U_\ell S_\ell V_\ell \tag{7}$$

the matrices $U_\ell \in \mathrm{O}(N_\ell)$ and $V_\ell \in \mathrm{O}(N_{\ell-1})$, are typical independent samples from the Haar measure across all layers. To make sure weight matrices remain close enough to independent during learning, we define a custom weight constraint which consists in keeping U_ℓ and V_ℓ fixed while only the matrix S_ℓ, constrained to be diagonal, is updated. The number of parameters is thus reduced from $N_\ell \times N_{\ell-1}$ to $\min(N_\ell, N_{\ell-1})$. We refer to layers following this weight constraint as USV-layers. For the replica formula of Claim 1 to be correct, the matrices S_ℓ from different layers should furthermore remain uncorrelated during the learning. In Section 3, we consider the training of linear networks for which information-theoretic quantities can be computed analytically, and confirm numerically that with USV-layers the replica predicted entropy is correct at all times. In the following, we assume that is also the case for non-linear networks.

In Section 3.2 of the Supplementary Material we train a neural network with USV-layers on a simple real-world dataset (MNIST), showing that these layers can learn to represent complex functions despite their restriction. We further note that such a product decomposition is reminiscent of a series of works on adaptative structured efficient linear layers (SELLs and ACDC) [26, 27] motivated this time by speed gains, where only diagonal matrices are learned (in these works the matrices U_ℓ and V_ℓ are chosen instead as permutations of Fourier or Hadamard matrices, so that the matrix multiplication can be replaced by fast transforms). In Section 3, we discuss learning experiments with USV-layers on synthetic datasets.

While we have defined model (1) as a stochastic model, traditional feed forward neural networks are deterministic. In the numerical experiments of Section 3, we train and test networks without injecting noise, and only assume a noise model in the computation of information-theoretic quantities. Indeed, for continuous variables the presence of noise is necessary for mutual informations to remain finite (see discussion of Appendix C in [5]). We assume at layer ℓ an additive white Gaussian noise of small amplitude just before passing through its activation function to obtain $H(\boldsymbol{T}_\ell)$ and $I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1})$, while keeping the mapping $\boldsymbol{X} \to \boldsymbol{T}_{\ell-1}$ deterministic. This choice attempts to stay as close as possible to the deterministic neural network, but remains inevitably somewhat arbitrary (see again discussion of Appendix C in [5]).

Other related works— The strategy of studying neural networks models, with random weight matrices and/or random data, using methods originated in statistical physics heuristics, such as the replica and the cavity methods [10] has a long history. Before the deep learning era, this approach led to pioneering results in learning for the Hopfield model [28] and for the random perceptron [29, 30, 15, 16].

Recently, the successes of deep learning along with the disqualifying complexity of studying real

world problems have sparked a revived interest in the direction of random weight matrices. Recent results –without exhaustivity– were obtained on the spectrum of the Gram matrix at each layer using random matrix theory [31, 32], on expressivity of deep neural networks [33], on the dynamics of propagation and learning [34, 35, 36, 37], on the high-dimensional non-convex landscape where the learning takes place [38], or on the universal random Gaussian neural nets of [39].

The information bottleneck theory [1] applied to neural networks consists in computing the mutual information between the data and the learned hidden representations on the one hand, and between labels and again hidden learned representations on the other hand [2, 3]. A successful training should maximize the information with respect to the labels and simultaneously minimize the information with respect to the input data, preventing overfitting and leading to a good generalization. While this intuition suggests new learning algorithms and regularizers [40, 41, 42, 43, 44, 45, 46], we can also hypothesize that this mechanism is already at play in a priori unrelated commonly used optimization methods, such as the simple stochastic gradient descent (SGD). It was first tested in practice by [3] on very small neural networks, to allow the entropy to be estimated by binning of the hidden neurons activities. Afterwards, the authors of [5] reproduced the results of [3] on small networks using the continuous entropy estimator of [44], but found that the overall behavior of mutual information during learning is greatly affected when changing the nature of non-linearities. Additionally, they investigate the training of larger linear networks on i.i.d. normally distributed inputs where entropies at each hidden layer can be computed analytically for an additive Gaussian noise. The strategy proposed in the present paper allows us to evaluate entropies and mutual informations in non-linear networks larger than in [5, 3].

3 Numerical experiments

Estimators and activation comparisons— Two non-parametric estimators have already been considered by [5] to compute entropies and/or mutual informations during learning. The kernel-density approach of Kolchinsky et. al. [44] consists in fitting a mixture of Gaussians (MoG) to samples of the variable of interest and subsequently compute an upper bound on the entropy of the MoG [47]. The method of Kraskov et al. [48] uses nearest neighbor distances between samples to directly build an estimate of the entropy. Both methods require the computation of the matrix of distances between samples. Recently, [45] proposed a new non-parametric estimator for mutual informations which involves the optimization of a neural network to tighten a bound. It is unfortunately computationally hard to test how these estimators behave in high dimension as even for a known distribution the computation of the entropy is intractable (#P-complete) in most cases. However the replica method proposed here is a valuable point of comparison for cases where it is rigorously exact.

In the first numerical experiment we place ourselves in the setting of Theorem 1: a 2-layer network with i.i.d weight matrices, where the formula of Claim 1 is thus rigorously exact in the limit of large networks, and we compare the replica results with the non-parametric estimators of [44] and [48]. Note that the requirement for smooth activations ($H2$) of Theorem 1 can be relaxed (see discussion below the Theorem). Additionally, non-smooth functions can be approximated arbitrarily closely by smooth functions with equal information-theoretic quantities, up to numerical precision.

We consider a neural network with layers of equal size $N = 1000$ that we denote: $\boldsymbol{X} \to \boldsymbol{T}_1 \to \boldsymbol{T}_2$. The input variable components are i.i.d. Gaussian with mean 0 and variance 1. The weight matrices entries are also i.i.d. Gaussian with mean 0. Their standard-deviation is rescaled by a factor $1/\sqrt{N}$ and then multiplied by a coefficient σ varying between 0.1 and 10, i.e. around the recommended value

for training initialization. To compute entropies, we consider noisy versions of the latent variables where an additive white Gaussian noise of very small variance ($\sigma^2_{\text{noise}} = 10^{-5}$) is added right before the activation function, $\boldsymbol{T}_1 = f(W_1\boldsymbol{X} + \boldsymbol{\epsilon}_1)$ and $\boldsymbol{T}_2 = f(W_2 f(W_1\boldsymbol{X}) + \boldsymbol{\epsilon}_2)$ with $\boldsymbol{\epsilon}_{1,2} \sim \mathcal{N}(0, \sigma^2_{\text{noise}} I_N)$, which is also done in the remaining experiments to guarantee the mutual informations to remain finite. The non-parametric estimators [44, 48] were evaluated using 1000 samples, as the cost of computing pairwise distances is significant in such high dimension and we checked that the entropy estimate is stable over independent draws of a sample of such a size (error bars smaller than marker size). On Figure 5, we compare the different estimates of $H(\boldsymbol{T}_1)$ and $H(\boldsymbol{T}_2)$ for different activation functions: linear, hardtanh or ReLU. The hardtanh activation is a piecewise linear approximation of the tanh, hardtanh$(x) = -1$ for $x < -1$, x for $-1 < x < 1$, and 1 for $x > 1$, for which the integrals in the replica formula can be evaluated faster than for the tanh.

In the linear and hardtanh case, the non-parametric methods are following the tendency of the replica estimate when σ is varied, but appear to systematically over-estimate the entropy. For linear networks with Gaussian inputs and additive Gaussian noise, every layer is also a multivariate Gaussian and therefore entropies can be directly computed in closed form (*exact* in the plot legend). When using the Kolchinsky estimate in the linear case we also check the consistency of two strategies, either fitting the MoG to the noisy sample or fitting the MoG to the deterministic part of the $\boldsymbol{T}_\ell$ and augment the resulting variance with σ^2_{noise}, as done in [44] (*Kolchinsky et al. parametric* in the plot legend). In the network with hardtanh non-linearities, we check that for small weight values, the entropies are the same as in a linear network with same weights (*linear approx* in the plot legend, computed using the exact analytical result for linear networks and therefore plotted in a similar color to *exact*). Lastly, in the case of the ReLU-ReLU network, we note that non-parametric methods are predicting an entropy increasing as the one of a linear network with identical weights, whereas the replica computation reflects its knowledge of the cut-off and accurately features a slope equal to half of the linear network entropy (*1/2 linear approx* in the plot legend). While non-parametric estimators are invaluable tools able to approximate entropies from the mere knowledge of samples,they inevitably introduce estimation errors. The replica method is taking the opposite view. While being restricted to a class of models, it can leverage its knowledge of the neural network structure to provide a reliable estimate. To our knowledge, there is no other entropy estimator able to incorporate such information about the underlying multi-layer model.

Beyond informing about estimators accuracy, this experiment also unveils a simple but possibly important distinction between activation functions. For the hardtanh activation, as the random weights magnitude increases, the entropies decrease after reaching a maximum, whereas they only increase for the unbounded activation functions we consider – even for the single-side saturating ReLU. This loss of information for bounded activations was also observed by [5], where entropies were computed by discretizing the output as a single neuron with bins of equal size. In this setting, as the tanh activation starts to saturate for large inputs, the extreme bins (at -1 and 1) concentrate more and more probability mass, which explains the information loss. Here we confirm that the phenomenon is also observed when computing the entropy of the hardtanh (without binning and with small noise injected before the non-linearity). We check via the replica formula that the same phenomenology arises for the mutual informations $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ (see Section3.1).

Learning experiments with linear networks— In the following, and in Section 3.3 of the the Supplementary Material, we discuss training experiments of different instances of the deep learning models defined in Section 2. We seek to study the simplest possible training strategies achieving good generalization. Hence for all experiments we use plain stochastic gradient descent (SGD) with

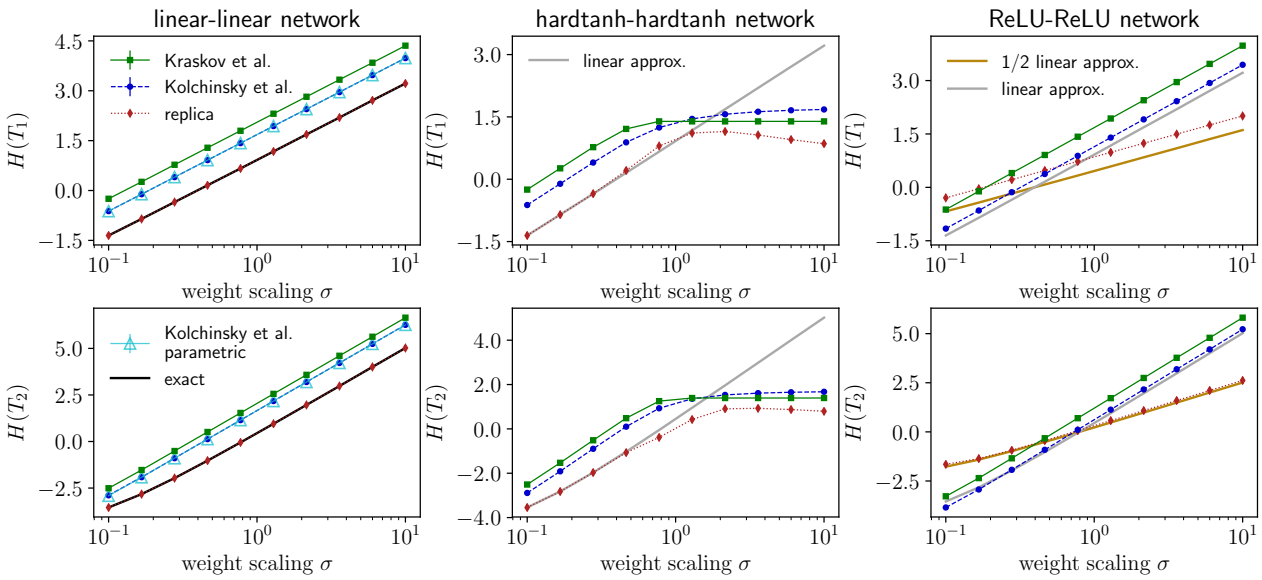

Figure 2: Entropy of latent variables in stochastic networks $\boldsymbol{X} \to \boldsymbol{T}_1 \to \boldsymbol{T}_2$, with equally sized layers $N = 1000$, inputs drawn from $\mathcal{N}(0, I_N)$, weights from $\mathcal{N}(0, \sigma^2 I_{N^2}/N)$, as a function of the weight scaling parameter σ. An additive white Gaussian noise $\mathcal{N}(0, 10^{-5} I_N)$ is added inside the non-linearity. Left column: linear network. Center column: hardtanh-hardtanh network. Right column: ReLU-ReLU network.

constant learning rates, without momentum and without any explicit form of regularization. The sizes of the training and testing sets are taken equal and scale typically as a few hundreds times the size of the input layer. Unless otherwise stated, plots correspond to single runs, yet we checked over a few repetitions that outcomes of independent runs lead to identical qualitative behaviors. The values of mutual informations $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ are computed by considering a noisy versions of the latent variables where an additive white Gaussian noise of very small variance ($\sigma^2_{\text{noise}} = 10^{-5}$) is added right before the activation function, as in the previous experiment. This noise is neither present at training time, where it could act as a regularizer, nor at testing time. Given the noise is only assumed at the last layer, the second to last layer is a deterministic mapping of the input variable; hence the replica formula yielding mutual informations between adjacent layers gives us directly $I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1}) = H(\boldsymbol{T}_\ell) - H(\boldsymbol{T}_\ell|\boldsymbol{T}_{\ell-1}) = H(\boldsymbol{T}_\ell) - H(\boldsymbol{T}_\ell|\boldsymbol{X}) = I(\boldsymbol{T}_\ell; \boldsymbol{X})$. We provide a second Python package [49] to implement in Keras learning experiments on synthetic datasets, using USV- layers and interfacing the first Python package [12] for replica computations.

To start with we consider the training of a linear network in the teacher-student scenario. The teacher has also to be linear to be learnable: we consider a simple single-layer network with additive white Gaussian noise, $\boldsymbol{Y} = \tilde{W}_{\text{teach}} \boldsymbol{X} + \boldsymbol{\epsilon}$, with input $\boldsymbol{x} \sim \mathcal{N}(0, I_N)$ of size N, teacher matrix $\tilde{W}_{\text{teach}}$ i.i.d. normally distributed as $\mathcal{N}(0, 1/N)$, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.01 I_N)$, and output of size $N_Y = 4$. We train a student network of three USV-layers, plus one fully connected unconstrained layer $\boldsymbol{X} \to \boldsymbol{T}_1 \to \boldsymbol{T}_2 \to \boldsymbol{T}_3 \to \hat{\boldsymbol{Y}}$ on the regression task, using plain SGD for the MSE loss $(\hat{\boldsymbol{Y}} - \boldsymbol{Y})^2$. We recall that in the USV-layers (7) only the diagonal matrix is updated during learning. On the left panel of Figure 3, we report the learning curve and the mutual informations between the hidden

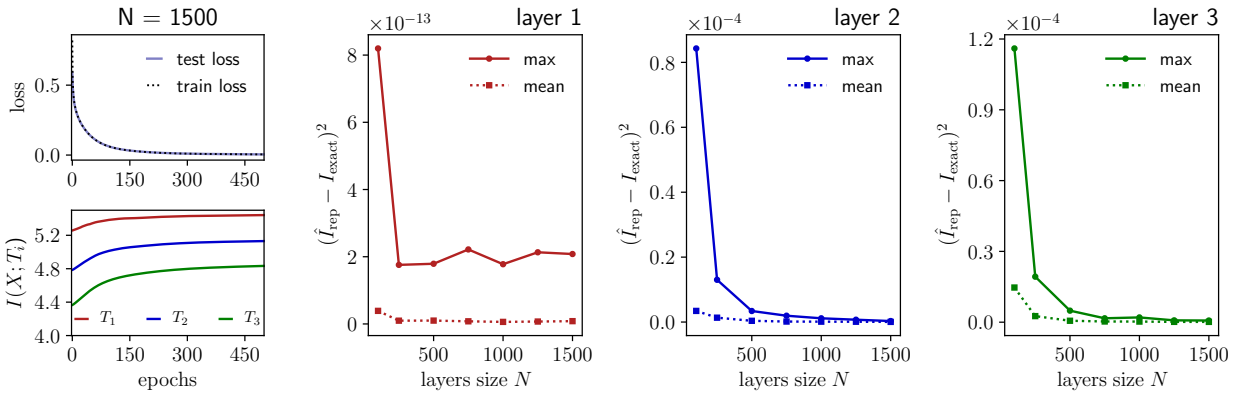

Figure 3: Training of a 4-layer linear student of varying size on a regression task generated by a linear teacher of output size $N_{\boldsymbol{Y}} = 4$. Upper-left: MSE loss on the training and testing sets during training by plain SGD for layers of size $N = 1500$. Best training loss is 0.004735, best testing loss is 0.004789. Lower-left: Corresponding mutual information evolution between hidden layers and input. Center-left, center-right, right: maximum and squared error of the replica estimation of the mutual information as a function of layers size N, over the course of 5 independent trainings for each value of N for the first, second and third hidden layer.

layers and the input in the case where all layers but outputs have size $N = 1500$. Again this linear setting is analytically tractable and does not require the replica formula, a similar situation was studied in [5]. In agreement with their observations, we find that the mutual informations $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ keep on increasing throughout the learning, without compromising the generalization ability of the student. Now, we also use this linear setting to demonstrate (i) that the replica formula remains correct throughout the learning of the USV-layers and (ii) that the replica method gets closer and closer to the exact result in the limit of large networks, as theoretically predicted (2). To this aim, we repeat the experiment for N varying between 100 and 1500, and report the maximum and the mean value of the squared error on the estimation of the $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ over all epochs of 5 independent training runs. We find that even if errors tend to increase with the number of layers, they remain objectively very small and decrease drastically as the size of the layers increases.

Learning experiments with deep non-linear networks— Finally, we apply the replica formula to estimate mutual informations during the training of non-linear networks on correlated input data.

We consider a simple single layer generative model $\boldsymbol{X} = \tilde{W}_{\rm gen}\boldsymbol{Y} + \boldsymbol{\epsilon}$ with normally distributed code $\boldsymbol{Y} \sim \mathcal{N}(0, I_{N_Y})$ of size $N_Y = 100$, data of size $N_X = 500$ generated with matrix $\tilde{W}_{\rm gen}$ i.i.d. normally distributed as $\mathcal{N}(0, 1/N_Y)$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.01 I_{N_X})$. We then train a recognition model to solve the binary classification problem of recovering the label $y = \text{sign}(Y_1)$, the sign of the first neuron in $\boldsymbol{Y}$, using plain SGD but this time to minimize the cross-entropy loss. Note that the rest of the initial code $(Y_2, ..Y_{N_{\boldsymbol{Y}}})$ acts as noise/nuisance with respect to the learning task. We compare two 5-layers recognition models with 4 USV- layers plus one unconstrained, of sizes 500-1000-500-250-100-2, and activations either linear-ReLU-linear-ReLU-softmax (top row of Figure 4) or linear-hardtanh-linear-hardtanh-softmax (bottom row). Because USV-layers only feature $O(N)$

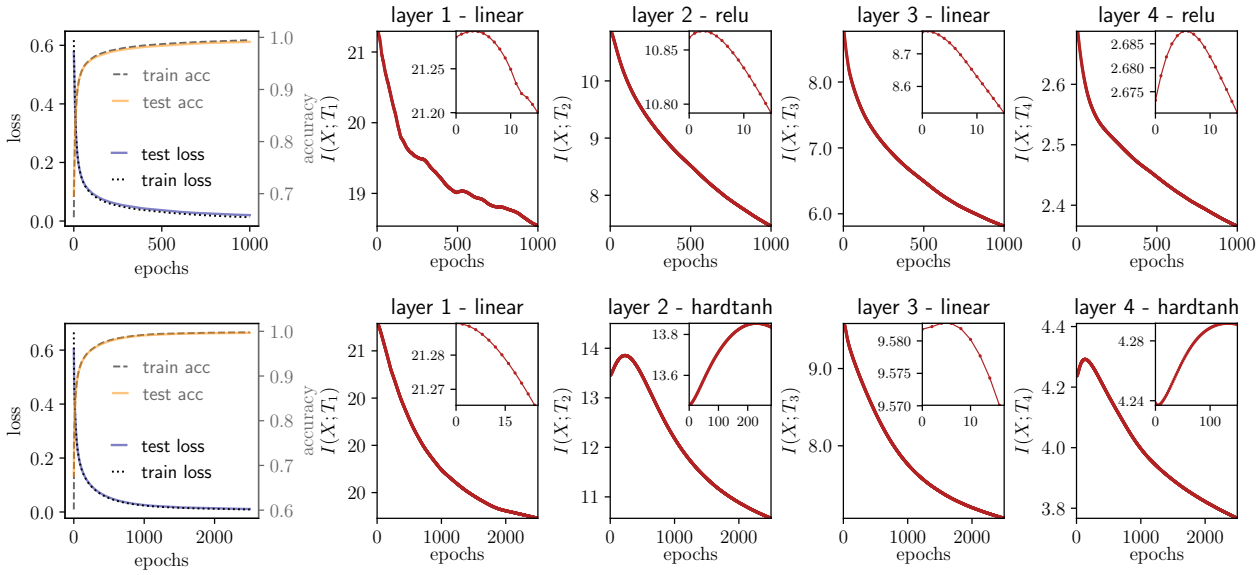

Figure 4: Training of two recognition models on a binary classification task with correlated input data and either ReLU (top) or hardtanh (bottom) non-linearities. Left: training and generalization cross-entropy loss (left axis) and accuracies (right axis) during learning. Best training-testing accuracies are 0.995 - 0.991 for ReLU version (top row) and 0.998 - 0.996 for hardtanh version (bottom row). Remaining colums: mutual information between the input and successive hidden layers. Insets zoom on the first epochs.

parameters instead of $O(N^2)$ we observe that they require more iterations to train in general. In the case of the ReLU network, adding interleaved linear layers was key to successful training with 2 non-linearities, which explains the somewhat unusual architecture proposed. For the recognition model using hardtanh, this was actually not an issue (see Supplementary Material for an experiment using only hardtanh activations), however, we consider a similar architecture for fair comparison. We discuss further the ability of learning of USV-layers in the Supplementary Material.

This experiment is reminiscent of the setting of [3], yet now tractable for networks of larger sizes. For both types of non-linearities we observe that the mutual information between the input and all hidden layers decrease during the learning, except for the very beginning of training where we can sometimes observe a short phase of increase (see zoom in insets). For the hardtanh layers this phase is longer and the initial increase of noticeable amplitude.

In this particular experiment, the claim of [3] that compression can occur during training even with non double-saturated activation seems corroborated (a phenomenon that was not observed by [5]). Yet we do not observe that the compression is more pronounced in deeper layers and its link to generalization remains elusive. For instance, we do not see a delay in the generalization w.r.t. training accuracy/loss in the recognition model with hardtanh despite of an initial phase without compression in two layers. Further learning experiments, including a second run of this last experiment, are presented in the Supplementary Material.

4 Conclusion and perspectives

We have presented a class of deep learning models together with a tractable method to compute entropy and mutual information between layers. This, we believe, offers a promising framework for further investigations, and to this aim we provide Python packages that facilitate both the computation of mutual informations and the training, for an arbitrary implementation of the model.

We observe in our high-dimensional experiments that compression does happen during learning, even when using ReLU activations. While we did not observe a clear link between generalization and compression in our setting, there are many directions to be further explored within the models presented in Section 2. Studying the entropic effect of regularizers is a natural step to formulate an entropic interpretation to generalization. Furthermore, while our experiments focused on the supervised learning, the replica formula derived for multi-layer models is general and can be applied in unsupervised contexts, for instance in the theory of VAEs. On the rigorous side, the greater perspective remains proving the replica formula in the general case of multi-layer models, and further confirm that the replica formula stays true after the learning of the USV-layers. Another question worth of future investigation is whether the replica method can be used to describe not only entropies and mutual informations for learned USV-layers, but also the optimal learning of the weights itself.

Acknowledgments

The authors would like to thank Léon Bottou, Antoine Maillard, Marc Mézard and Galen Reeves for insightful discussions. This work has been supported by the ERC under the European Union's FP7 Grant Agreement 307087-SPARCS and the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe, as well as by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL. Additional funding is acknowledged by MG from "Chaire de recherche sur les modèles et sciences des données", Fondation CFM pour la Recherche-ENS; by AM from Labex DigiCosme; and by CL from the Swiss National Science Foundation under grant 200021E-175541. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. *37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

[2] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, 2015.

[3] R. Shwartz-Ziv and N. Tishby. Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810*, 2017.

[4] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6(Jan):165–188, 2005.

[5] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the Information Bottleneck Theory of Deep Learning. In *International Conference on Learning Representations (ICLR)*, 2018.

[6] Y. Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series*, 95(1):012001, 2008.

[7] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová. Multi-layer generalized linear estimation. In *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[8] A. K. Fletcher and S. Rangan. Inference in Deep Networks in High Dimensions. *arXiv:1706.06549*, 2017.

[9] G. Reeves. Additivity of Information in Multilayer Networks via Additive Gaussian Noise Transforms. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[10] M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. World Scientific Publishing Company, 1987.

[11] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[12] dnner: Deep Neural Networks Entropy with Replicas, Python library. `https://github.com/sphinxteam/dnner`.

[13] A. M. Tulino, G. Caire, S. Verdú, and S. Shamai (Shitz). Support Recovery With Sparsely Sampled Free Random Matrices. *IEEE Transactions on Information Theory*, 59(7):4243–4271, 2013.

[14] D. Donoho and A. Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[15] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056, 1992.

[16] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[17] M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[18] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models. *arXiv:1708.03395*, 2017.

[19] J. Barbier, N. Macris, A. Maillard, and F. Krzakala. The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices. In *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[20] D. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[21] L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[22] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory (ISIT)*, 2011.

[23] S. Rangan, P. Schniter, and A. K. Fletcher. Vector approximate message passing. In *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[24] J. Barbier and N. Macris. The stochastic interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *arXiv:1705.02780*, 2017.

[25] J. Barbier, N. Macris, and L. Miolane. The Layered Structure of Tensor Estimation and its Mutual Information. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[26] M. Moczulski, M. Denil, J. Appleyard, and N. de Freitas. ACDC: A Structured Efficient Linear Layer. In *International Conference on Learning Representations (ICLR)*, 2016.

[27] Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang. Deep fried convnets. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[28] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

[29] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A*, 22(12):1983, 1989.

[30] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A*, 22(12):2181, 1989.

[31] C. Louart and R. Couillet. Harnessing neural networks: A random matrix approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[32] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[33] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the Expressive Power of Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, 2017.

[34] A. Saxe, J. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[35] S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, 2017.

[36] M. Advani and A. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667*, 2017.

[37] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina. Efficient supervised learning in networks with binary synapses. *Proceedings of the National Academy of Sciences*, 104:11079–11084, 2007.

[38] Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, 2014.

[39] R. Giryes, G. Sapiro, and A. M. Bronstein. Deep neural networks with random Gaussian weights: a universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016.

[40] M. Chalk, O. Marre, and G. Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, 2016.

[41] A. Achille and S. Soatto. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 2018.

[42] A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[43] A. Achille and S. Soatto. Emergence of Invariance and Disentangling in Deep Representations. In *ICML 2017 Workshop on Principled Approaches to Deep Learning*, 2017.

[44] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert. Nonlinear Information Bottleneck. *arXiv:1705.02436*, 2017.

[45] M.I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R.D. Hjelm. MINE: Mutual Information Neural Estimation. *arXiv:1801.04062*, 2017.

[46] S. Zhao, J. Song, and S. Ermon. InfoVAE: Information Maximizing Variational Autoencoders. *arXiv:1706.02262*, 2017.

[47] A. Kolchinsky and B. D. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

[48] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[49] lsd: Learning with Synthetic Data, Python library. `https://github.com/marylou-gabrie/learning-synthetic-data`.

Supplementary Material

Contents

1 Replica formula for the entropy

1.1 Background

The replica method [1, 2] was first developed in the context of disordered physical systems where the strength of interactions J are randomly distributed, $J \sim P_J(J)$. Given the distribution of microstates $\boldsymbol{x}$ at a fixed temperature β^{-1}, $P(\boldsymbol{x}|\beta, J) = \frac{1}{\mathcal{Z}(\beta,J)}\, e^{-\beta \mathcal{H}_J(\boldsymbol{x})}$, one is typically interested in the average free energy

$$\mathcal{F}(\beta) = -\lim_{N\to\infty} \frac{1}{\beta N} \mathbb{E}_J \log \mathcal{Z}(\beta, J), \tag{8}$$

from which typical macroscopic behavior is obtained. Computing (8) is hard in general, but can be done with the use of specific techniques. The replica method in particular employs the following mathematical identity

$$\mathbb{E}_J \log \mathcal{Z} = \lim_{n\to 0} \frac{\mathbb{E}_J \mathcal{Z}^n - 1}{n}. \tag{9}$$

Evaluating the average on the r.h.s. leads, under the *replica-symmetry* assumption, to an expression of the form $\mathbb{E}_J \mathcal{Z}^n = e^{-\beta N n \operatorname{extr}_{\boldsymbol{q}} \phi(\beta,\boldsymbol{q})}$, where $\phi(\beta, \boldsymbol{q})$ is known as the *replica-symmetric free energy*, and $\boldsymbol{q}$ are *order parameters* related to macroscopic quantities of the system. We then write $\mathcal{F}(\beta) = \operatorname{extr}_{\boldsymbol{q}} \phi(\beta, \boldsymbol{q})$, so that computing $\mathcal{F}$ depends on solving the saddle-point equations $\nabla_{\boldsymbol{q}} \phi\big|_{\boldsymbol{q}^*} = 0$.

Computing (8) is of interest in many problems outside of physics [3, 4]. Early applications of the replica method in machine learning include the evaluation of the optimal capacity and generalization error of the perceptron [5, 6, 7, 8, 9]. More recently it has also been used in the study of problems in telecommunications and signal processing, such as channel divison multiple access [10] and compressed sensing [11, 12, 13, 14]. For a review of these developments see [15].

These particular examples all share the following common probabilistic structure

$$\begin{cases} \boldsymbol{y} \sim P_{Y|Z}(\boldsymbol{y}|W\boldsymbol{x}), \\ \boldsymbol{x} \sim P_X(\boldsymbol{x}), \end{cases} \tag{10}$$

for fixed W and different choices of $P_{Y|Z}$ and P_X; in other words, they are all specific instances of *generalized linear models* (GLMs). Using Bayes theorem, one writes the posterior distribution of $\boldsymbol{x}$ as $P(\boldsymbol{x}|W, \boldsymbol{y}) = \frac{1}{P(W,\boldsymbol{y})}\, P_{Y|Z}(\boldsymbol{y}|W\boldsymbol{x})\, P_X(\boldsymbol{x})$; the replica method is then employed to evaluate the average log-marginal likelihood $\mathbb{E}_{W,\boldsymbol{y}} \log P(W, \boldsymbol{y})$, which gives us typical properties of the model. Note this quantity is nothing but the entropy of $\boldsymbol{y}$ given W, $H(\boldsymbol{y}|W)$.

The distribution P_J (or P_W in the notation above) is usually assumed to be i.i.d. on the elements of the matrix J. However, one can also use the same techniques to approach J belonging to arbitrary orthogonally-invariant ensembles. This approach was pioneered by [16, 17, 18, 19], and in the context of generalized linear models by [20, 21, 22, 23, 24, 25, 26].

Generalizing the analysis of (10) to multi-layer models has first been considered by [27] in the context of Gaussian i.i.d. matrices, and by [28, 29] for orthogonally-invariant ensembles. In particular, [29] has an expression for the replica free energy which should be in principle equivalent to the one we present, although its focus is in the derivation of this expression rather than applications or explicit computations.

Finally, it is worth mentioning that even though the replica method is usually considered to be non-rigorous, its results have been proven to be exact for different classes of models, including GLMs [30, 31, 32, 33, 34, 35], and are widely conjectured to be exact in general. In fact, in section 2 we show how to proove the formula in the particular case of two-layer with Gaussian matrices.

1.2 Entropy in single/multi-layer generalized linear models

1.2.1 Single-layer

For a single-layer generalized linear model

$$\begin{cases} \boldsymbol{x} \sim P_X(\boldsymbol{x}), \\ \boldsymbol{y} \sim P_{Y|Z}(\boldsymbol{y}|W\boldsymbol{x}). \end{cases} \tag{11}$$

with P_X and $P_{Y|Z}$ separable in the components of $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^M$, and $W \in \mathbb{R}^{M\times N}$ Gaussian i.i.d., $W_{\mu i} \sim \mathcal{N}(0, 1/N)$, define $\alpha = M/N$ and $\rho = \mathbb{E}_x x^2$. Then the entropy of $\boldsymbol{y}$ in the limit $N \to \infty$ is given by [15, 35]

$$\lim_{N\to\infty} N^{-1} H(\boldsymbol{y}|W) = \min \operatorname*{extr}_{A,V} \phi(A, V), \tag{12}$$

where

$$\phi(A, V) = -\frac{1}{2}AV + I(x; x + \frac{\xi_0}{\sqrt{A}}) + \alpha H(y|\xi_1; V, \rho), \tag{13}$$

with ξ_0, ξ_1 both normally distributed with zero mean and unit variance, and $P(y|\xi; V, \rho) = \int D\tilde{z}\, P_{Y|Z}(y|\sqrt{\rho - V}\xi + \sqrt{V}\tilde{z})$ (here $D\tilde{z}$ denotes integration over a standard Gaussian measure).

This can be adapted to orthogonally-invariant ensembles by using the techniques described in [22]. Let $W = USV^T$, where U is orthogonal, S diagonal and arbitrary and V is Haar distributed. We denote by $\pi_W(\lambda_W)$ the distribution of eigenvalues of W^TW, and the second moment of $\boldsymbol{z} = W\boldsymbol{x}$ by $\tilde{\rho} = \frac{\mathbb{E}\lambda_W}{\alpha}\rho$. The entropy is then written as $N^{-1}H(\boldsymbol{y}|W) = \min \operatorname{extr}_{A,V,\tilde{A},\tilde{V}} \phi(A, V, \tilde{A}, \tilde{V})$, where

$$\phi(A, V, \tilde{A}, \tilde{V}) = -\frac{1}{2}\big(\tilde{A}V + \alpha A\tilde{V} - F_W(AV)\big) + I(x; x + \frac{\xi_0}{\sqrt{\tilde{A}}}) + \alpha H(y|\xi_1; \tilde{V}, \tilde{\rho}), \tag{14}$$

and

$$F_W(x) = \min_\theta \big\{2\alpha\theta + (\alpha - 1)\log(1 - \theta) + \mathbb{E}_{\lambda_W} \log[x\lambda_W + (1 - \theta)(1 - \alpha\theta)]\big\}. \tag{15}$$

If the matrix is Gaussian i.i.d., $\pi_W(\lambda_W)$ is Marchenko-Pastur and $F_W(AV) = \alpha AV$. Extremizing over A gives $\tilde{V} = V$, so that (13) is recovered. In this precise case, it has been proven rigorously in [35].

1.2.2 Multi-layer

Consider the following multi-layer generalized linear model

$$\begin{cases} t_{0,i} \equiv x_i \sim P_0(x_i), \\ t_{1,i} \sim P_1(t_{1,i}|W_1\boldsymbol{x}), \\ t_{2,i} \sim P_2(t_{2,i}|W_2\boldsymbol{t}_1), \\ \vdots \\ t_{L,i} \equiv y_i \sim P_L(y|W_L\boldsymbol{t}_{L-1}), \end{cases} \tag{16}$$

where the $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ are fixed, and the i index runs from 0 to n_ℓ. Using Bayes' theorem we can write

$$P(\boldsymbol{t}_0|\boldsymbol{t}_L, \underline{W}) = \frac{1}{P(\boldsymbol{t}_L, \underline{W})} \int \prod_{\ell=1}^{L-1} d\boldsymbol{t}_\ell \prod_{\ell=1}^{L} P(\boldsymbol{t}_\ell|W_\ell\boldsymbol{t}_{\ell-1}) P(\boldsymbol{t}_0). \tag{17}$$

with $\underline{W} = \{W_\ell\}_{\ell=1,\dots,L}$. Performing posterior inference requires one to evaluate the marginal likelihood

$$P(\boldsymbol{t}_L, \underline{W}) = \int \prod_{\ell=0}^{L-1} d\boldsymbol{t}_\ell \prod_{\ell=1}^{L} P(\boldsymbol{t}_\ell | W_\ell \boldsymbol{t}_{\ell-1})\, P(\boldsymbol{t}_0), \tag{18}$$

which is in general hard to do. Our analysis employs the framework introduced in [27] to compute the entropy of $\boldsymbol{t}_L$ in the limit $n_0 \to \infty$ with $\tilde{\alpha}_\ell = n_\ell / n_0$ finite for $\ell = 1, \dots, L$

$$\lim_{n_0 \to \infty} n_0^{-1} H(\boldsymbol{t}_L | \underline{W}) = \min \operatorname*{extr}_{\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}} \phi(\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}), \tag{19}$$

with the replica potential ϕ given by

$$\begin{aligned} \phi(\boldsymbol{A}, \boldsymbol{V}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{V}}) = -\frac{1}{2} \sum_{\ell=1}^{L} \tilde{\alpha}_{\ell-1} \big[\tilde{A}_\ell V_\ell + \alpha_\ell A_\ell \tilde{V}_\ell - F_{W_\ell}(A_\ell V_\ell) \big] + I(t_0; t_0 + \frac{\xi_0}{\sqrt{\tilde{A}_1}}) + \\ + \sum_{\ell=1}^{L-1} \tilde{\alpha}_\ell \left[H(t_\ell | \xi_\ell; \tilde{A}_{\ell+1}, \tilde{V}_\ell, \tilde{\rho}_\ell) - \frac{1}{2} \log(2\pi e \tilde{A}_{\ell+1}) \right] + \tilde{\alpha}_L H(t_L | \xi_L; \tilde{V}_L, \tilde{\rho}_L). \end{aligned} \tag{20}$$

and the $\boldsymbol{\xi}$ normally distributed with zero mean and unit variance. The t_ℓ in the expression above are distributed as

$$P(t_\ell | \xi_\ell; A, V, \rho) = \int D\tilde{\xi} D\tilde{z}\, P_\ell(t_\ell + \sqrt{1/A}\tilde{\xi} | \sqrt{\rho - V}\xi_\ell + \sqrt{V}\tilde{z}), \tag{21}$$

$$P(t_L | \xi_L; V, \rho) = \int D\tilde{z}\, P_L(t_L | \sqrt{\rho - V}\xi_L + \sqrt{V}\tilde{z}). \tag{22}$$

where $\int Dz\,(\cdot) = \int dz\, \mathcal{N}(z; 0, 1)\,(\cdot)$ denotes the integration over the standard Gaussian measure.

1.3 A simple heuristic derivation of the multi-layer formula

Formula (20) can be derived using a simple argument. Consider the case $L = 2$, where the model reads

$$\begin{cases} \boldsymbol{t}_0 \sim P_0(\boldsymbol{t}_0), \\ \boldsymbol{t}_1 \sim P_1(\boldsymbol{t}_1 | W_1 \boldsymbol{t}_0), \\ \boldsymbol{t}_2 \sim P_2(\boldsymbol{t}_2 | W_2 \boldsymbol{t}_1), \end{cases} \tag{23}$$

with $\boldsymbol{t}_\ell \in \mathbb{R}^{n_\ell}$ and $W \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$. For the problem of estimating $\boldsymbol{t}_1$ given the knowledge of $\boldsymbol{t}_2$, we compute $\lim_{n_1 \to \infty} n_1^{-1} H(\boldsymbol{t}_2 | W_1)$ using the replica free energy (14)

$$\phi(A_2, V_2, \tilde{A}_2, \tilde{V}_2) = -\frac{1}{2} \big(\tilde{A}_2 V_2 + \alpha_2 A_2 \tilde{V}_2 - F_{W_2}(A_2 V_2) \big) + \tag{24}$$

$$+ I(t_1; t_1 + \frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}) + \alpha_2 H(t_2 | \xi_2; \tilde{V}_2, \tilde{\rho}_2). \tag{25}$$

Note that

$$\begin{aligned} I(t_1; t_1 + \frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}) &= H\big(t_1 + \frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}\big) - H\big(\frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}\big) \\ &= H\big(t_1 + \frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}\big) - \frac{1}{2} \log(2\pi e \tilde{A}_2). \end{aligned} \tag{26}$$

Moreover, $H(t_1 + \tilde{\xi}_1/\sqrt{\tilde{A}_2})$ can be obtained from the replica free energy of another problem: that of estimating $\boldsymbol{t}_0$ given the knowledge of (noisy) $\boldsymbol{t}_1$, which can again be written using (14)

$$\lim_{n_0\to\infty} n_0^{-1} H(t_1 + \frac{\tilde{\xi}_1}{\sqrt{\tilde{A}_2}}) = \min \underset{A_1,V_1,\tilde{A}_1,\tilde{V}_1}{\text{extr}} \phi_1(A_1, V_1, \tilde{A}_1, \tilde{V}_1), \tag{27}$$

with

$$\phi_1(A_1, V_1, \tilde{A}_1, \tilde{V}_1) = -\frac{1}{2}\big(\tilde{A}_1 V_1 + \alpha_1 A_1 \tilde{V}_1 - F_{W_1}(A_1 V_1)\big) + \tag{28}$$

$$+ I(t_0; t_0 + \frac{\xi_0}{\sqrt{\tilde{A}_1}}) + \alpha_1 H(t_1|\xi_1; \tilde{A}_1, \tilde{V}_1, \tilde{\rho}_1), \tag{29}$$

and the noise $\tilde{\xi}_1$ being integrated in the computation of $H(t_1|\xi_1)$, see (22). Replacing (26)-(29) in (25) gives our formula (20) for $L = 2$; further repeating this procedure allows one to write the equations for arbitrary L.

1.4 Formulation in terms of tractable integrals

While expression (20) is more easily written in terms of conditional entropies and mutual informations, evaluating it requires us to explicitely state it in terms of integrals, which we do below. We first consider the Gaussian i.i.d. In this case, the multi-layer formula was derived with the cavity and replica method by [27], and we shall use their results here. Assuming that $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ such that $W_{\ell,\mu i} \sim \mathcal{N}(0, 1/n_{\ell-1})$ and using the replica formalism, Claim 1 from the main text becomes, in this case

$$\lim_{n_0\to\infty} n_0^{-1} H(\boldsymbol{t}_L|\underline{W}) = \min \underset{\boldsymbol{A},\boldsymbol{V}}{\text{extr}}\, \phi(\boldsymbol{A}, \boldsymbol{V}), \tag{30}$$

with the replica potential ϕ evaluated from

$$\phi(\boldsymbol{A}, \boldsymbol{V}) = \frac{1}{2}\sum_{\ell=1}^{L} \tilde{\alpha}_{\ell-1} A_\ell(\rho_\ell - V_\ell) - \mathcal{K}(\boldsymbol{A}, \boldsymbol{V}, \boldsymbol{\rho}), \tag{31}$$

and

$$\mathcal{K}(\boldsymbol{A}, \boldsymbol{V}, \boldsymbol{\rho}) = K_0(A_1) + \sum_{\ell=1}^{L-1} \tilde{\alpha}_\ell K_\ell(A_{\ell+1}, V_\ell, \rho_\ell) + \tilde{\alpha_L} K_L(V_L, \rho_L). \tag{32}$$

The constants α_ℓ, $\tilde{\alpha}_\ell$ and ρ_ℓ are defined as following[1]: $\alpha_\ell = n_\ell/n_{\ell-1}$, $\tilde{\alpha}_\ell = n_\ell/n_0$, $\rho_\ell = \int dt\, P_{\ell-1}(t)\, t^2$. Moreover

$$K_\ell(A, V, \rho) = \mathbb{E}_{b,t,z,w|A,V,\rho} \log Z_\ell(A, b, V, w), \tag{33}$$

for $1 \le \ell \le L-1$, and

$$\begin{aligned} K_0(A) &= \mathbb{E}_{b,x|A} \log Z_0(A, b), \\ K_L(V, \rho) &= \mathbb{E}_{y,z,w|V,\rho} \log Z_L(y, V, w). \end{aligned} \tag{34}$$

[1]Note that due to the central limit theorem, ρ_ℓ can be evaluated from $\rho_{\ell-1}$ using $\rho_\ell = \int dtdz\, P_\ell(t|z)\mathcal{N}(z; 0, \rho_{\ell-1})\, t^2$.

where

$$
\begin{aligned}
&Z_0(A,B) = \int dx\, P_0(x) e^{-\frac{1}{2}Ax^2+Bx},\\
&Z_\ell(A,B,V,\omega) = \int dt dz\, P_\ell(t|z)\mathcal{N}(z;\omega,V) e^{-\frac{1}{2}At^2+Bt},\\
&Z_L(y,V,\omega) = \int dz\, P_L(y|z)\mathcal{N}(z;\omega,V).
\end{aligned} \tag{35}
$$

and the measures over which expectations are computed are

$$
\begin{aligned}
&p_0(b,x;A) = P_0(x)\mathcal{N}(b;Ax,A),\\
&p_\ell(b,t,z,w;A,V,\rho) = P_\ell(t|z)\mathcal{N}(b;At,A)\mathcal{N}(z;w,V)\mathcal{N}(w,0,m),\\
&p_L(y,z,w;V,\rho) = P_L(y|z)\mathcal{N}(z;w,V)\mathcal{N}(w;0,\rho-V).
\end{aligned} \tag{36}
$$

We typically pick the likelihoods P_ℓ so that Z_ℓ can be computed in closed-form, which allows for a number of activation functions – linear, probit, ReLU etc. However, our analysis is quite general and can be done for arbitrary likelihoods, as long as evaluating (33) and (34) is computationally feasible.

Finally, the replica potential above can be generalized to the orthogonally-invariant case using the framework of [22], which we have described in the previous subsection

$$
\begin{aligned}
\phi(\boldsymbol{A},\boldsymbol{V},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{V}}) = &-\frac{1}{2}\sum_{\ell=1}^{L}\tilde{\alpha}_{\ell-1}\left[\tilde{A}_\ell V_\ell + \alpha_\ell A_\ell \tilde{V}_\ell - F_{W_\ell}(A_\ell V_\ell)\right] + I(t_0; t_0 + \frac{\xi_0}{\sqrt{A_1}}) +\\
&+ \sum_{\ell=1}^{L-1}\tilde{\alpha}_\ell\left[H(t_\ell|\xi_\ell; A_{\ell+1}, V_\ell, \tilde{\rho}_\ell) - \frac{1}{2}\log(2\pi e A_\ell)\right] + \tilde{\alpha}_L H(t_L|\xi_L).
\end{aligned} \tag{37}
$$

If the matrix W_ℓ is Gaussian i.i.d., the distribution of eigenvalues of $W_\ell^T W_\ell$ is Marchenko-Pastur and one gets $F_{W_\ell}(A_\ell V_\ell) = \alpha_\ell A_\ell V_\ell$, $\tilde{A}_\ell = \alpha_\ell A_\ell$, $\tilde{V}_\ell = V_\ell$, so that (31) is recovered. Moreover, for $L = 1$, one obtains the replica free energy proposed by [22, 23, 24].

1.4.1 Recovering the formulation in terms of conditional entropies

One can rewrite the formulas above in a simpler way. By manipulating the measures (36) one obtains

$$
K_0(A,\rho) = -I(x;b) + \frac{1}{2}A\rho, \tag{38}
$$

for $x \sim P_0(x)$ and $b \sim \mathcal{N}(b;Ax,A)$. Introducing a standard normal variable ξ_0 and using the invariance of mutual informations, this can be written as

$$
K_0(A,\rho) = -I(x; x + \sqrt{1/A}\xi_0) + \frac{1}{2}A\rho. \tag{39}
$$

Similarly

$$
K_L(V,\rho) = -H(y|w;V), \tag{40}
$$

for $P(y|w;V) = \int dz P_L(y|z)\mathcal{N}(z;w,V)$ and $P(w;V,\rho) = \mathcal{N}(w;0,\rho-V)$. Introducing standard normal ξ_L

$$
K_L(V,\rho) = -H(y|\xi_L;V,\rho). \tag{41}
$$

where

$$P(y|\xi_L;V,\rho)=\int D\tilde{z}\,P_L(y|\sqrt{\rho-V}\xi_L+\sqrt{V}\tilde{z}), \tag{42}$$

and $\int D\tilde{z}\,(\cdot)=\int d\tilde{z}\,\mathcal{N}(z;0,1)\,(\cdot)$ denotes integration over the standard Gaussian measure.

Finally, for the K_ℓ

$$K_\ell(A,V,\rho)=-H(b|w;A,V,\rho)+\frac{1}{2}A\rho+\frac{1}{2}\log(2\pi eA), \tag{43}$$

for $P(b|w;A,V)=\int dtdz\,\mathcal{N}(b;At,A)P_\ell(t|z)\mathcal{N}(z;w,V)$ and $P(w;V,\rho)=\mathcal{N}(w;0,\rho-V)$. Introducing standard normal ξ_ℓ

$$K_\ell(A,V,\rho)=-H(t_\ell|\xi_\ell;A,V)+\frac{1}{2}A\rho+\frac{1}{2}\log(2\pi eA). \tag{44}$$

where

$$P(t_\ell|\xi_\ell;A,V,\rho)=\int D\tilde{\xi}D\tilde{z}\,P_\ell(t_\ell+\sqrt{1/A}\tilde{\xi}|\sqrt{\rho-V}\xi_\ell+\sqrt{V}\tilde{z}). \tag{45}$$

We can then rewrite (32) as

$$\begin{aligned}\mathcal{K}(\boldsymbol{A},\boldsymbol{V},\boldsymbol{\rho})=&\frac{1}{2}\sum_{\ell=1}^{L}\tilde{\alpha}_{\ell-1}A_\ell\rho_\ell-I(t_0;t_0+\frac{\xi_0}{\sqrt{A_1}})-\\&-\sum_{\ell=1}^{L-1}\tilde{\alpha}_\ell\left[H(t_\ell|\xi_\ell;A_{\ell+1},V_\ell,\rho_\ell)-\frac{1}{2}\log(2\pi eA_{\ell+1})\right]-\tilde{\alpha}_LH(t_L|\xi_L;V_L,\rho_L).\end{aligned} \tag{46}$$

Replacing in (31) yields

$$\begin{aligned}\phi(\boldsymbol{A},\boldsymbol{V})=&-\frac{1}{2}\sum_{\ell=1}^{L}\tilde{\alpha}_{\ell-1}A_\ell V_\ell+I(t_0;t_0+\frac{\xi_0}{\sqrt{A_1}})+\\&+\sum_{\ell=1}^{L-1}\tilde{\alpha}_\ell\left[H(t_\ell|\xi_\ell;A_{\ell+1},V_\ell,\rho_\ell)-\frac{1}{2}\log(2\pi eA_{\ell+1})\right]+\tilde{\alpha}_LH(t_L|\xi_L;V_L,\rho_L).\end{aligned} \tag{47}$$

1.5 Solving saddle-point equations

In order to deal with the extremization problem in

$$\lim_{n_0\to\infty}n_0^{-1}H(\boldsymbol{t}_L|\underline{W})=\min\operatorname*{extr}_{\boldsymbol{A},\boldsymbol{V},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{V}}}\phi(\boldsymbol{A},\boldsymbol{V},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{V}}), \tag{48}$$

one needs to solve the saddle-point equations $\nabla_{\{\boldsymbol{A},\boldsymbol{V},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{V}}\}}\phi=0$. In what follows we propose two different methods to do that: a fixed-point iteration, and the state evolution of the ML-VAMP algorithm [28].

1.5.1 Method 1: fixed-point iteration

We first introduce the following function, which is related to the derivatives of F_{W_ℓ}

$$\psi_\ell(\theta,\gamma) = 1 - \gamma\big[\mathcal{S}_\ell\big(-\gamma^{-1}(1-\theta)(1-\alpha_\ell\theta)\big)\big]^{-1}, \tag{49}$$

where $\mathcal{S}_\ell(z) = \mathbb{E}_{\lambda_\ell}\frac{1}{\lambda_\ell - z}$ is the Stieltjes transform of $W_\ell^T W_\ell$, see e.g. [36]. In our experiments we have evaluated $\mathcal{S}$ approximately by using the empirical distribution of eigenvalues.

The fixed point iteration consist in looping through layers L to 1, first computing the ϑ_ℓ which minimizes (15), and $\tilde{V}_\ell$

$$\begin{aligned}\vartheta_\ell^{(t)} &= \arg\min_\theta \Big[\theta - \psi_\ell(\theta, A_\ell^{(t)} V_\ell^{(t)})\Big]^2,\\ \tilde{V}_\ell^{(t)} &= \vartheta_\ell^{(t)}/A_\ell^{(t)},\end{aligned} \tag{50}$$

then $A_\ell^{(t+1)}$, which for layers $1 \le \ell \le L-1$ comes from

$$A_\ell^{(t+1)} = -\mathbb{E}_{b,t,z,w|\tilde{\rho}_\ell,\tilde{A}_{\ell+1},\tilde{V}_\ell}\partial_w^2 \log Z_\ell(\tilde{A}_{\ell+1}^{(t)}, b, \tilde{V}_\ell^{(t)}, w), \tag{51}$$

and for the L-th layer, from

$$A_L^{(t+1)} = -\mathbb{E}_{y,z,w|\tilde{\rho}_L,\tilde{V}_L}\partial_w^2 \log Z_L(y, \tilde{V}_L, w). \tag{52}$$

Finally, we recompute ϑ_ℓ using $A_\ell^{(t+1)}$, and $\tilde{A}_\ell$

$$\begin{aligned}\vartheta_\ell^{(t+\frac{1}{2})} &= \arg\min_\theta \Big[\theta - \psi_\ell(\theta, A_\ell^{(t+1)} V_\ell^{(t)})\Big]^2,\\ \tilde{A}_\ell^{(t)} &= \alpha_\ell \vartheta_\ell^{(t+\frac{1}{2})}/V_\ell^{(t)}.\end{aligned} \tag{53}$$

and move on to the next layer. After these quantities are computed for all layers, we compute all the V_ℓ; for $2 \le \ell \le L$

$$V_\ell^{(t+1)} = \mathbb{E}_{b,t,z,w|\tilde{\rho}_{\ell-1},\tilde{A}_\ell,\tilde{V}_{\ell-1}}\partial_b^2 \log Z_\ell(\tilde{A}_\ell^{(t)}, b, \tilde{V}_{\ell-1}^{(t)}, w), \tag{54}$$

and for the 1st layer

$$V_1^{(t+1)} = \mathbb{E}_{b,x|\tilde{A}_1}\partial_b \log Z_1(\tilde{A}_1^{(t)}, b). \tag{55}$$

This particular order has been chosen so that if W_ℓ is Gaussian i.i.d., $\theta_\ell^{(t)} = A_\ell^{(t)} V_\ell^{(t)}$ and one recovers the state evolution equations in [27].

The set of initial conditions is picked so as to cover the basin of attraction of typical fixed points. In our experiments we have chosen $(A_{i,\ell}^{(0)}, V_{i,\ell}^{(0)}) \in \{(\rho_\ell^{-1}, \rho_\ell), (\delta^{-1}, \delta)\}$, with $\delta = 10^{-10}$.

1.5.2 Method 2: ML-VAMP state evolution

While the fixed-point iteration above works well in most cases, it is not provably convergent. In particular, it relies on a solution for $\theta = \psi_\ell(\theta, A_\ell^{(t)} V_\ell^{(t)})$ being found, which might not happen throughout the iteration.

Algorithm 1 Compute entropy $H(\boldsymbol{y}_L|\underline{W})$

Require: $\{\boldsymbol{A}_i^{(1)}, \boldsymbol{V}_i^{(1)}\}_{i=1}^{n_{\text{init}}}$, ϵ, $t_{\max}$
  for $i = 1 \to n_{\text{init}}$ **do** *(loop through initial conditions)*
    $t \leftarrow 0$
    while $D < \epsilon$ or $t < t_{\max}$ **do** *(at each time step ...)*
      for $\ell = L \to 1$ **do** *(... loop through layers)*
        compute $\vartheta_\ell^{(t)}$, $\tilde{V}_\ell^{(t)}$ using (50)
        compute $A_\ell^{(t+1)}$ using (51) or (52)
        compute $\vartheta_\ell^{(t+\frac{1}{2})}$, $\tilde{A}_\ell^{(t)}$ using (53)
      end for
      compute $V_\ell^{(t+1)}\,\forall \ell$ using (54) or (55)
      $D \leftarrow \sum_\ell |V_\ell^{(t+1)} - V_\ell^{(t)}|$
      $t \leftarrow t + 1$
    end while
    $H_i \leftarrow \phi(\boldsymbol{A}^{(t)}, \boldsymbol{V}^{(t)}, \tilde{\boldsymbol{A}}^{(t)}, \tilde{\boldsymbol{V}}^{(t)})$
  end for
  return $\min_i H_i$

An alternative is to employ the state evolution (SE) of the ML-VAMP algorithm [28], which leads to the same fixed points as the scheme above under certain conditions. Let us first look at the single-layer case; the ML-VAMP SE equations read

$$A_x^+ = \frac{1}{V_x^+(A_x^-)} - A_x^-, \qquad A_z^+ = \frac{1}{V_z^+(A_x^+, 1/A_z^-)} - A_z^-, \tag{56}$$

$$A_x^- = \frac{1}{V_x^-(A_x^+, 1/A_z^-)} - A_x^+, \qquad A_z^- = \frac{1}{V_z^-(A_z^+)} - A_z^+, \tag{57}$$

where

$$V_x^+(A) = \mathbb{E}_{x,z}\partial_B^2 \log Z_0(A, Ax + \sqrt{A}z), \tag{58}$$

$$V_z^+(A, \sigma^2) = \sigma^2 \lim_{M\to\infty} \frac{1}{M} \operatorname{Tr}\left[\Phi(\Phi^T\Phi + A\sigma^2)^{-1}\Phi^T\right] = \alpha^{-1}\sigma^2\big(1 - A\sigma^2\,\mathcal{S}(-A\sigma^2)\big), \tag{59}$$

$$V_x^-(A, \sigma^2) = \sigma^2 \lim_{N\to\infty} \frac{1}{N} \operatorname{Tr}\left[(\Phi^T\Phi + A\sigma^2)^{-1}\right] = \sigma^2\,\mathcal{S}(-A\sigma^2), \tag{60}$$

$$V_z^-(A) = \frac{1}{A} + \frac{1}{A^2}\underbrace{\mathbb{E}_{y,w,z}\partial_w^2 \log Z_1(y, w, 1/A)}_{-\bar{g}(A)}. \tag{61}$$

Combining (57) and (61) yields

$$1/A_z^- = \frac{1}{\bar{g}(A_z^+)} - \frac{1}{A_z^+}. \tag{62}$$

At the fixed points

$$V_x \equiv V_x^+ = V_x^- = \frac{1}{A_x^+ + A_x^-}, \qquad V_z \equiv V_z^+ = V_z^- = \frac{1}{A_z^+ + A_z^-}. \tag{63}$$

as well as

$$V_z = \frac{1 - A_x^+ V_x}{\alpha A_z^-} = \frac{A_x^- V_x}{\alpha A_z^-} \Rightarrow \alpha \frac{A_z^-}{A_z^- + A_z^+} = \frac{A_x^-}{A_x^- + A_x^+} \tag{64}$$

One can show these conditions also hold for the above scheme, under the following mapping of variables:

$$V = V_x, \quad \tilde{A} = A_x^-, \quad \tilde{V} = 1/A_z^+, \quad A = A_z^- A_z^+ V_z = \bar{g}(A_z^+), \quad \theta = A_z^- V_z = \frac{1}{1 + \frac{A_z^+}{A_z^-}}. \tag{65}$$

These equations are easily generalizable to the multi-layer case; the equations for A_z^+ and A_x^- remain the same, while the equations for A_x^+ and A_z^- become

$$A_{x_\ell}^+ = \frac{1}{V_{x_\ell}^+(A_{x_\ell}^-, 1/A_{z_{\ell-1}}^+)} - A_{x_\ell}^-, \tag{66}$$

$$A_{z_{\ell-1}}^- = \frac{1}{V_{z_{\ell-1}}^-(A_{x_\ell}^-, 1/A_{z_{\ell-1}}^+)} - A_{z_{\ell-1}}^+, \tag{67}$$

where

$$V_{x_\ell}^+(A, V) = \mathbb{E}_{b,t,z,w}\, \partial_B^2 \log Z_\ell(A, b, V, w), \tag{68}$$

$$V_{z_{\ell-1}}^-(A, V) = \frac{1}{A} + \frac{1}{A^2} \mathbb{E}_{b,t,z,w}\, \partial_w^2 \log Z_\ell(A, b, V, w). \tag{69}$$

Note that the quantities in (58), (61), (68) and (69) were already being evaluated in the scheme described in the previous subsection.

1.6 Further considerations

1.6.1 Mutual information from entropy

While in our computations we focus on the entropy $H(\boldsymbol{T}_\ell)$, the mutual information $I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1})$ can be easily obtained from the chain rule relation

$$\begin{aligned} I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1}) &= H(\boldsymbol{T}_\ell) + \mathbb{E}_{\boldsymbol{T}_\ell, \boldsymbol{T}_{\ell-1}} \log P_{\boldsymbol{T}_\ell | \boldsymbol{T}_{\ell-1}}(\boldsymbol{t}_\ell | \boldsymbol{t}_{\ell-1}) \\ &= H(\boldsymbol{T}_\ell) + \int dz\, \mathcal{N}(z; 0, \tilde{\rho}_\ell) \int dh\, P_\ell(h|z) \log P_\ell(h|z), \end{aligned} \tag{70}$$

where in order to go from the first to the second line we have used the central limit theorem. In particular if the mapping $\boldsymbol{X} \to \boldsymbol{T}_{\ell-1}$ is deterministic, as typically enforced in the models we use in the experiments, then $I(\boldsymbol{T}_\ell; \boldsymbol{T}_{\ell-1}) = I(\boldsymbol{T}_\ell; \boldsymbol{X})$.

1.6.2 Equivalence in linear case

In the linear case, $\boldsymbol{Y} = W_L W_{L-1} \cdots W_1 \boldsymbol{X} + \mathcal{N}(0, \Delta)$, our formula reduces to [22, 25, 37]

$$\lim_{N \to \infty} N^{-1} I(\boldsymbol{Y}; \boldsymbol{X}) = \min \operatorname*{extr}_{A,V} \left\{ -\frac{1}{2} AV - \frac{1}{2} G(-V/\Delta) + I(x; x + \sqrt{1/A}\xi) \right\}, \tag{71}$$

where

$$G(x) = \operatorname*{extr}_{\Lambda} \{ -\mathbb{E}_\lambda \log |\lambda - \Lambda| + \Lambda x \} - (\log |x| + 1), \tag{72}$$

is also known as the integrated R-transform, with λ the eigenvalues of W^TW, $W \equiv W_L W_{L-1} \cdots W_1$. If P_0 is Gaussian, then $I(x; x + \sqrt{1/A}\xi) = \frac{1}{2}\log(1+A)$; extremizing over A and V then gives

$$A = 1/V - 1, \qquad V = \Delta \mathcal{S}(-\Delta), \tag{73}$$

where $\mathcal{S}(z)$ is the Stieltjes transform of W^TW. The mutual information can then be rewritten as

$$\lim_{N\to\infty} N^{-1} I(\boldsymbol{Y}; \boldsymbol{X}) = \frac{1}{2}\mathbb{E}_\lambda \log(\lambda + \Delta) - \frac{1}{2}\log \Delta. \tag{74}$$

This same result can be achieved analytically with much less effort, since in this case $P_{\boldsymbol{Y}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; 0, \Delta I_M + WW^T)$.

2 Proof of the replica formula by the adaptive interpolation method

2.1 Two-layer generalized linear estimation: Problem statement

One gives here a generic description of the observation model, that is a two-layer generalized linear model (GLM). Let $n_0, n_1, n_2 \in \mathbb{N}^*$ and define the triplet $\mathbf{n} = (n_0, n_1, n_2)$. Let P_0 be a probability distribution over $\mathbb{R}$ and let $(X_i^0)_{i=1}^{n_0} \overset{\text{i.i.d.}}{\sim} P_0$ be the components of a signal vector $\mathbf{X}^0$. One fixes two functions $\varphi_1 : \mathbb{R} \times \mathbb{R}^{k_1} \to \mathbb{R}$ and $\varphi_2 : \mathbb{R} \times \mathbb{R}^{k_2} \to \mathbb{R}$, k_1, $k_2 \in \mathbb{N}$. They act component-wise, i.e. if $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times k_i}$ then $\varphi_i(\mathbf{x}, \mathbf{A}) \in \mathbb{R}^m$ is a vector with entries $[\varphi_i(\mathbf{x}, \mathbf{A})]_\mu := \varphi_i(x_\mu, \mathbf{A}_\mu)$, $\mathbf{A}_\mu$ being the μ-th row of $\mathbf{A}$. For $i \in \{1, 2\}$, consider $(\mathbf{A}_{i,\mu})_{\mu=1}^{n_i} \overset{\text{i.i.d.}}{\sim} P_{A_i}$ where P_{A_i} is a probability distribution over $\mathbb{R}^{k_i}$. One acquires n_2 measurements through

$$Y_\mu = \varphi_2\Big(\frac{1}{\sqrt{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}, \mathbf{A}_1\Big)\Big]_\mu, \mathbf{A}_{2,\mu}\Big) + \sqrt{\Delta}Z_\mu\,, \qquad 1 \le \mu \le n_2\,. \tag{75}$$

Here $(Z_\mu)_{\mu=1}^{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is an additive Gaussian noise, $\Delta > 0$, and $\mathbf{W}_1 \in \mathbb{R}^{n_1 \times n_0}$, $\mathbf{W}_2 \in \mathbb{R}^{n_2 \times n_1}$ are measurement matrices whose entries are i.i.d. with respect to (w.r.t.) $\mathcal{N}(0,1)$. Equivalently,

$$Y_\mu \sim P_{\text{out},2}\Big(\,\cdot\,\Big|\frac{1}{\sqrt{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}, \mathbf{A}_1\Big)\Big]_\mu\Big) \tag{76}$$

where the transition density, w.r.t. Lebesgue's measure, is

$$P_{\text{out},2}(y|x) = \int dP_{A_2}(\mathbf{a}) \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}(y - \varphi_2(x,\mathbf{a}))^2}\,. \tag{77}$$

Our analysis uses both representations (75) and (76). The estimation problem is to recover $\mathbf{X}^0$ from the knowledge of $\mathbf{Y} = (Y_\mu)_{\mu=1}^{n_2}$, φ_1, φ_2, $\mathbf{W}_1$, $\mathbf{W}_2$, Δ, P_0.

In the language of statistical mechanics, the random variables $\mathbf{Y}$, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{X}^0$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{Z}$ are called *quenched* variables because once the measurements are acquired they have a "fixed realization". An expectation taken w.r.t. *all* quenched random variables appearing in an expression will simply be denoted by $\mathbb{E}$ *without* subscript. Subscripts are only used when the expectation carries over a subset of random variables appearing in an expression or when some confusion could arise.

After definition of the *Hamiltonian*

$$\mathcal{H}(\mathbf{x}, \mathbf{a}_1; \mathbf{Y}, \mathbf{W}_1, \mathbf{W}_2) := -\sum_{\mu=1}^{n_2} \ln P_{\text{out},2}\Big(Y_\mu\Big|\frac{1}{\sqrt{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}, \mathbf{a}_1\Big)\Big]_\mu\Big)\,, \tag{78}$$

the joint posterior distribution of $(\mathbf{x},\, \mathbf{a}_1)$ given the quenched variables $\mathbf{Y}$, $\mathbf{W}_1$, $\mathbf{W}_2$ reads (Bayes formula)

$$dP(\mathbf{x},\mathbf{a}_1|\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2) = \frac{1}{\mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)} dP_0(\mathbf{x}) dP_{A_1}(\mathbf{a}_1) e^{-\mathcal{H}(\mathbf{x},\mathbf{a}_1;\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)}\,; \tag{79}$$

$dP_0(\mathbf{x}) = \prod_{i=1}^{n_0} dP_0(x_i)$ being the prior over the signal and $dP_{A_1}(\mathbf{a}_1) := \prod_{i=1}^{n_1} dP_{A_1}(\mathbf{a}_{1,i})$. The *partition function* is defined as

$$\begin{aligned}&\mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)\\ &\quad := \int dP_0(\mathbf{x}) dP_{A_1}(\mathbf{a}_1) dP_{A_2}(\mathbf{a}_2) \prod_{\mu=1}^{n_2} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(Y_\mu - \varphi_2\left(\frac{1}{\sqrt{n_1}}\left[\mathbf{W}_2 \varphi_1\left(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}},\mathbf{a}_1\right)\right]_\mu, \mathbf{a}_{2,\mu}\right)\right)^2}.\end{aligned} \tag{80}$$

One introduces a standard statistical mechanics notation for the expectation w.r.t. the posterior (79), the so called *Gibbs bracket* $\langle - \rangle$ defined for any continuous bounded function g as

$$\langle g(\mathbf{x},\mathbf{a}_1)\rangle := \int dP(\mathbf{x},\mathbf{a}_1|\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2) g(\mathbf{x},\mathbf{a}_1) \tag{81}$$

One important quantity is the associated averaged *free entropy* (or minus the averaged *free energy*)

$$f_{\mathbf{n}} := \frac{1}{n_0} \mathbb{E} \ln \mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)\,. \tag{82}$$

It is perhaps useful to stress that $\mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)$ is nothing else than the density of $\mathbf{Y}$ conditioned on $\mathbf{W}_1, \mathbf{W}_2$; so we have the explicit representation (used later on)

$$\begin{aligned} f_{\mathbf{n}} &= \frac{1}{n_0} \mathbb{E}_{\mathbf{W}_1,\mathbf{W}_2} \int d\mathbf{Y}\, \mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2) \ln \mathcal{Z}(\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2) \\ &= \frac{1}{n_0} \mathbb{E}_{\mathbf{W}_1,\mathbf{W}_2} \Bigg[\int d\mathbf{Y} dP_0(\mathbf{X}^0) dP_{A_1}(\mathbf{A}_1) e^{-\mathcal{H}(\mathbf{X}^0,\mathbf{A}_1;\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)} \\ &\qquad\qquad \cdot \ln \int dP_0(\mathbf{x}) dP_{A_1}(\mathbf{a}_1) e^{-\mathcal{H}(\mathbf{x},\mathbf{a}_1;\mathbf{Y},\mathbf{W}_1,\mathbf{W}_2)} \Bigg], \end{aligned} \tag{83}$$

where $d\mathbf{Y} = \prod_{\mu=1}^{n_2} dY_\mu$.

This appendix presents the derivation, thanks to the adaptive interpolation method, of the thermodynamic limit $\lim_{\mathbf{n}\to\infty} f_{\mathbf{n}}$ in the "high-dimensional regime", namely when $n_0, n_1, n_2 \to +\infty$ such that $n_2/n_1 \to \alpha_2 > 0$, $n_1/n_0 \to \alpha_1 > 0$. In this high-dimensional regime, the "measurement rate" satisfies $n_2/n_0 \to \alpha := \alpha_1 \cdot \alpha_2$.

2.2 Important scalar inference channels

The thermodynamic limit of the free entropy will be expressed in terms of the free entropy of simple *scalar* inference channels. This "decoupling property" results from the mean-field approach in statistical physics, used through in the replica method to perform a formal calculation of the free entropy of the model [2, 4]. This section presents these three scalar denoising models.

The first channel is an additive Gaussian one. Let $r \geq 0$ play the role of a signal-to-noise ratio. Consider the inference problem consisting of retrieving $X_0 \sim P_0$ from the observation $Y_0 = \sqrt{r}\, X_0 + Z_0$, where $Z_0 \sim \mathcal{N}(0,1)$ independently of X_0. The associated posterior distribution is

$$dP(x|Y_0) = \frac{dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2}}{\int dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2}}\,. \tag{84}$$

The free entropy associated to this channel is just the expectation of the logarithm of the normalization factor

$$\psi_{P_0}(r) := \mathbb{E} \ln \int dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2}\,. \tag{85}$$

The second scalar channel appearing naturally in the problem is linked to $P_{\text{out},2}$ through the following inference model. Suppose that $V, U \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ where V is *known*, while the inference problem is to recover the unknown U from the observation

$$\widetilde{Y}_0 \sim P_{\text{out},2}\big(\,\cdot\,|\sqrt{q}\, V + \sqrt{\rho - q}\, U\big)\,, \tag{86}$$

where $\rho > 0$, $q \in [0, \rho]$. The free entropy for this model, again related to the normalization factor of the posterior $dP(u|\widetilde{Y}_0, V)$, is

$$\Psi_{P_{\text{out},2}}(q;\rho) := \mathbb{E} \ln \int \mathcal{D}u P_{\text{out},2}\big(\widetilde{Y}_0|\sqrt{q}\, V + \sqrt{\rho - q}\, u\big)\,, \tag{87}$$

where $\mathcal{D}u = du (2\pi)^{-1/2} e^{-w^2/2}$ is the standard Gaussian measure.

The third scalar channel to play a role is linked to the hidden layer $\mathbf{X}^1 := \varphi_1\Big(\mathbf{W}_1 \mathbf{X}^0/\sqrt{n_0}, \mathbf{A}_1\Big)$ of the two-layer GLM. Suppose that $V, U \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, where V is *known*. Consider the problem of recovering U from the observation $Y_0' = \sqrt{r}\varphi_1(\sqrt{q}\, V + \sqrt{\rho - q}\, U, \mathbf{A}_1) + Z'$ where $r \geq 0$, $\rho > 0$, $q \in [0, \rho]$, $Z' \sim \mathcal{N}(0,1)$ and $\mathbf{A}_1 \sim P_{A_1}$. Equivalently, $Y_0' \sim P^{(r)}_{\text{out},1}(\cdot|\sqrt{q}\, V + \sqrt{\rho - q}\, U)$ with

$$P^{(r)}_{\text{out},1}(y|x) := \int dP_{A_1}(\mathbf{a}) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y - \sqrt{r}\varphi_1(x,\mathbf{a}))^2}\,. \tag{88}$$

From this last description, it is easy to see that the free entropy for this model is given by a formula similar to (87). Introducing $\delta(\,\cdot - \varphi_1(x,\mathbf{a}))$, the Dirac measure centred on $\varphi_1(x,\mathbf{a})$, it reads

$$\begin{aligned}
\Psi_{P^{(r)}_{\text{out},1}}(q;\rho) &= \mathbb{E} \ln \int \mathcal{D}u P^{(r)}_{\text{out},1}\big(Y_0'|\sqrt{q}\, V + \sqrt{\rho - q}\, u\big) \\
&= \mathbb{E} \ln \int \mathcal{D}u dP_{A_1}(\mathbf{a}) dh \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Y_0' - \sqrt{r}h)^2} \delta\big(h - \varphi_1(\sqrt{q}\, V + \sqrt{\rho - q}\, u, \mathbf{a})\big) \\
&= -\frac{\ln(2\pi) + \mathbb{E}[(Y_0')^2]}{2} \\
&\quad + \mathbb{E} \ln \int \mathcal{D}u dP_{A_1}(\mathbf{a}) dh e^{\sqrt{r}hY_0' - \frac{rh^2}{2}} \delta\big(h - \varphi_1(\sqrt{q}\, V + \sqrt{\rho - q}\, u, \mathbf{a})\big)\,.
\end{aligned}$$

The second moment of Y_0' is simply $\mathbb{E}[(Y_0')^2] = r\mathbb{E}[\varphi_1^2(T, \mathbf{A}_1)] + 1$ with $T \sim \mathcal{N}(0,\rho)$, $\mathbf{A}_1 \sim P_{A_1}$. Hence

$$\Psi_{P^{(r)}_{\text{out},1}}(q;\rho) = -\frac{1 + \ln(2\pi) + r\mathbb{E}[\varphi_1^2(T, \mathbf{A}_1)]}{2} + \Psi_{\varphi_1}(q, r; \rho) \tag{89}$$

where

$$\Psi_{\varphi_1}(q, r; \rho) := \mathbb{E} \ln \int \mathcal{D}u dP_{A_1}(\mathbf{a}) dh e^{\sqrt{r}hY_0' - \frac{rh^2}{2}} \delta\big(h - \varphi_1(\sqrt{q}\, V + \sqrt{\rho - q}\, u, \mathbf{a})\big)\,. \tag{90}$$

2.3 Replica-symmetric formula and mutual information

Our goal is to prove Theorem 1 that gives a single-letter *replica-symmetric formula* for the asymptotic free entropy of model (75), (76). The result holds under the following hypotheses:

(H1) The prior distribution P_0 has a bounded support.
(H2) φ_1, φ_2 are bounded $\mathcal{C}^2$ functions with bounded first and second derivatives w.r.t. their first argument.
(H3) $\mathbf{W}_1$, $\mathbf{W}_2$ have entries i.i.d. with respect to $\mathcal{N}(0,1)$.

Let $\rho_0 := \mathbb{E}[(X^0)^2]$ where $X^0 \sim P_0$ and $\rho_1 := \mathbb{E}[\varphi_1^2(T, \mathbf{A}_1)]$ where $T \sim \mathcal{N}(0, \rho_0)$, $\mathbf{A}_1 \sim P_{A_1}$. The *replica-symmetric potential* (or just potential) is

$$f_{\rm RS}(q_0, r_0, q_1, r_1; \rho_0, \rho_1) := \psi_{P_0}(r_0) + \alpha_1 \Psi_{\varphi_1}(q_0, r_1; \rho_0) + \alpha \Psi_{P_{\rm out,2}}(q_1; \rho_1) - \frac{r_0 q_0}{2} - \alpha_1 \frac{r_1 q_1}{2} \,. \quad (91)$$

Theorem 1 (Replica-symmetric formula)**.** *Suppose that hypotheses (H1), (H2), (H3) hold. Then, the thermodynamic limit of the free entropy* (82) *for the two-layer generalized linear estimation model* (75), (76) *satisfies*

$$f_\infty := \lim_{\mathbf{n}\to\infty} f_{\mathbf{n}} = \sup_{q_1 \in [0,\rho_1]} \inf_{r_1 \geq 0} \sup_{q_0 \in [0,\rho_0]} \inf_{r_0 \geq 0} f_{\rm RS}(q_0, r_0, q_1, r_1; \rho_0, \rho_1) \,. \quad (92)$$

The limiting expression of the mutual information between the observations and the signal to recover follows immediately of Theorem 1.

Corollary 1 (Single-letter formula for the mutual information)**.** *The thermodynamic limit of the mutual information for model* (75), (76) *between the observations and the signal to recover verifies*

$$i_{\mathbf{n}} := \frac{1}{n_0} I(\mathbf{X}^0, \mathbf{A}_1, \mathbf{A}_2; \mathbf{Y} \,|\mathbf{W}_1, \mathbf{W}_2) \xrightarrow[\mathbf{n}\to\infty]{} i_\infty := -f_\infty - \frac{\alpha}{2}(1 + \ln(2\pi\Delta)) \,. \quad (93)$$

Proof. A simple calculation gives

$$\begin{aligned}
&\frac{1}{n_0} I(\mathbf{X}^0, \mathbf{A}_1, \mathbf{A}_2; \mathbf{Y}|\mathbf{W}_1, \mathbf{W}_2) \\
&\qquad = -\frac{1}{n_0}\mathbb{E} \ln P(\mathbf{Y}|\mathbf{W}_1, \mathbf{W}_2) + \frac{1}{n_0}\mathbb{E} \ln P(\mathbf{Y}|\mathbf{X}^0, \mathbf{A}_1, \mathbf{A}_2, \mathbf{W}_1, \mathbf{W}_2) \\
&\qquad = -f_{\mathbf{n}} - \frac{1}{2n_0\Delta}\mathbb{E}\left[\sum_{\mu=1}^{n_2} \Big(Y_\mu - \varphi_2\Big(\Big[\frac{\mathbf{W}_2\mathbf{X}^1}{\sqrt{n_1}}\Big]_\mu, \mathbf{A}_{2,\mu}\Big)\Big)^2\right] - \frac{n_2}{2n_0}\ln(2\pi\Delta) \\
&\qquad = -f_{\mathbf{n}} - \frac{n_2}{2n_0} - \frac{n_2}{2n_0}\ln(2\pi\Delta) \,.
\end{aligned}$$

□

2.4 Interpolating estimation problem

The proof of Theorem 1 follows the same steps than the proof of the replica formula for a one-layer GLM in [35].

One introduces an *interpolating estimation problem* that interpolates between the original problem (76) at $t = 0$, $t \in [0, 1]$ being the interpolation parameter, and two analytically tractable problems at $t = 1$.

Define $\rho_1(n_0) := \mathbb{E}\Big[\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i^1)^2\Big] = \mathbb{E}\Big[\varphi_1^2\big([\mathbf{W}_1\mathbf{X}^0/\sqrt{n_0}]_1, \mathbf{A}_{1,1}\big)\Big]$. In Appendix A.2 one shows

Proposition 1 (Convergence of $\rho_1(n_0)$ to ρ_1)**.** *Under the hypotheses (H1), (H2), (H3)*

$$\lim_{n_0\to+\infty}\rho_1(n_0)=\rho_1\,. \tag{94}$$

Let $q:[0,1]\to[0,\rho_1(n_0)]$ be a continuous *interpolation function* and $r\geq 0$. Define

$$S_{t,\mu}:=\sqrt{\frac{1-t}{n_1}}\,\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}},\mathbf{A}_1\Big)\Big]_\mu+\sqrt{\int_0^t q(v)dv}\,V_\mu+\sqrt{\int_0^t(\rho_1(n_0)-q(v))dv}\,U_\mu \tag{95}$$

where $V_\mu,U_\mu\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$. Assume $\mathbf{V}=(V_\mu)_{\mu=1}^{n_2}$ is *known* and two kinds of observations are given:

$$\begin{cases}Y_{t,\mu} & \sim P_{\text{out},2}(\,\cdot\,|\,S_{t,\mu})\,, & 1\leq\mu\leq n_2,\\ Y'_{t,i} & =\sqrt{r\,t}\,\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)+Z'_i\,, & 1\leq i\leq n_1,\end{cases} \tag{96}$$

where $(Z'_i)_{i=1}^{n_1}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$. $\mathbf{Y}_t=(Y_{t,\mu})_{\mu=1}^{n_2}$, $\mathbf{Y}'_t=(Y'_{t,i})_{i=1}^{n_1}$ are our "time-dependent" observations.

Define, with a slight abuse of notations, $s_{t,\mu}(\mathbf{x},\mathbf{a}_1,u_\mu)\equiv s_{t,\mu}$ as

$$s_{t,\mu}=\sqrt{\frac{1-t}{n_1}}\,\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}},\mathbf{a}_1\Big)\Big]_\mu+\sqrt{\int_0^t q(v)dv}\,V_\mu+\sqrt{\int_0^t(\rho_1(n_0)-q(v))dv}\,u_\mu\,. \tag{97}$$

One introduces the *interpolating Hamiltonian*

$$\begin{aligned}\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y}_t,\mathbf{Y}_t^{'},\mathbf{W}_1,\mathbf{W}_2,\mathbf{V}):=&-\sum_{\mu=1}^{n_2}\ln P_{\text{out},2}(Y_{t,\mu}|s_{t,\mu})\\ &+\frac{1}{2}\sum_{i=1}^{n_1}\Big[Y'_{t,i}-\sqrt{r\,t}\,\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\Big]^2.\end{aligned} \tag{98}$$

It depends on $\mathbf{W}_2$ and $\mathbf{V}$ through the terms $(s_{t,\mu})_{\mu=1}^{n_2}$, and on $\mathbf{W}_1$ through both $(s_{t,\mu})_{\mu=1}^{n_2}$ and the sum over $i\in\{1,\dots,n_1\}$. The corresponding Gibbs bracket $\langle-\rangle_t$, which is the expectation operator w.r.t. the t-dependent joint posterior distribution of $(\mathbf{x},\mathbf{a}_1,\mathbf{u})$ given $(\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})$ is defined for every continuous bounded function g on $\mathbb{R}^{n_0}\times\mathbb{R}^{n_2}$ as:

$$\langle g(\mathbf{x},\mathbf{a}_1,\mathbf{u})\rangle_t:=\frac{1}{\mathcal{Z}_t}\int dP_0(\mathbf{x})dP_{A_1}(\mathbf{a}_1)\mathcal{D}\mathbf{u}\,g(\mathbf{x},\mathbf{a}_1,\mathbf{u})\,e^{-\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})}\,. \tag{99}$$

In (99), $\mathcal{D}\mathbf{u}=(2\pi)^{-n_2/2}\prod_{\mu=1}^{n_2}du_\mu e^{-u_\mu^2/2}$ is the n_2-dimensional standard Gaussian distribution and $\mathcal{Z}_t\equiv\mathcal{Z}_t(\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})$ is the appropriate normalization, i.e.

$$\mathcal{Z}_t(\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V}):=\int dP_0(\mathbf{x})dP_{A_1}(\mathbf{a}_1)\mathcal{D}\mathbf{u}\,e^{-\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})}\,. \tag{100}$$

Finally, the *interpolating free entropy* is

$$f_{\mathbf{n}}(t):=\frac{1}{n_0}\mathbb{E}\ln\mathcal{Z}_t(\mathbf{Y}_t,\mathbf{Y}'_t,\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})\,. \tag{101}$$

2.5 Interpolating free entropy at t=0 and t=1

One verifies easily that

$$\left\{\begin{array}{rcl} f_{\mathbf{n}}(0) & = & f_{\mathbf{n}} - \frac{1}{2}\frac{n_1}{n_0}\,, \\ f_{\mathbf{n}}(1) & = & \tilde{f}_{n_0,n_1} + \frac{n_2}{n_0}\Psi_{P_{\mathrm{out},2}}\big(\int_0^1 q(t)dt; \rho_1(n_0)\big) + \frac{n_1}{n_0}\frac{\ln 2\pi}{2}\,. \end{array}\right. \tag{102}$$

In the expression of $f_{\mathbf{n}}(1)$, $\tilde{f}_{n_0,n_1}$ denotes the free entropy of the *one-layer* GLM

$$Y'_i = \sqrt{r}\,\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) + Z'_i\,, \qquad 1 \le i \le n_1, \tag{103}$$

with $(X_i^0)_{i=1}^{n_0} \overset{\text{i.i.d.}}{\sim} P_0$, $(\mathbf{A}_{1,i})_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{A_1}$ and $(Z'_i)_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Applying Theorem 1 of [35], then (89), the free entropy $\tilde{f}_{n_0,n_1}$ in the thermodynamic limit $n_0, n_1 \to +\infty$ such that $n_1/n_0 \to \alpha_1$ is

$$\begin{aligned} \lim_{n_0,n_1\to+\infty} \tilde{f}_{n_0,n_1} &= \sup_{q_0\in[0,\rho_0]} \inf_{r_0\ge 0}\ \psi_{P_0}(r_0) + \alpha_1\Psi_{P_{\mathrm{out},1}^{(r)}}(q_0;\rho_0) - \frac{r_0q_0}{2} \\ &= -\alpha_1\frac{1+\ln 2\pi + r\rho_1}{2} + \sup_{q_0\in[0,\rho_0]} \inf_{r_0\ge 0}\ \psi_{P_0}(r_0) + \alpha_1\Psi_{\varphi_1}(q_0,r;\rho_0) - \frac{r_0q_0}{2}\,. \end{aligned} \tag{104}$$

From (102), by making use of (104) and Lemma 1 below, one obtains in the thermodynamic limit:

$$\begin{aligned} f_{\mathbf{n}}(1) = -\alpha_1\frac{1+r\rho_1}{2} &+ \alpha\Psi_{P_{\mathrm{out},2}}\Big(\int_0^1 q(t)dt;\rho_1(n_0)\Big) \\ &+ \sup_{q_0\in[0,\rho_0]} \inf_{r_0\ge 0}\ \Big\{\psi_{P_0}(r_0) + \alpha_1\Psi_{\varphi_1}(q_0,r;\rho_0) - \frac{r_0q_0}{2}\Big\} + \mathcal{O}_{\mathbf{n}}(1)\,. \end{aligned} \tag{105}$$

Here $\mathcal{O}_{\mathbf{n}}(1)$ is a quantity that vanishes uniformly in the limit $n_0, n_1, n_2 \to +\infty$. Lemma 1 justifies the identity $\frac{n_2}{n_0}\Psi_{P_{\mathrm{out},2}}\Big(\int_0^1 q(t)dt;\rho_1(n_0)\Big) = \alpha\Psi_{P_{\mathrm{out},2}}\Big(\int_0^1 q(t)dt;\rho_1(n_0)\Big) + \mathcal{O}_{\mathbf{n}}(1)$.

Lemma 1 (Uniform upperbound on $\Psi_{P_{\mathrm{out},2}}$)**.** *Assuming φ_2 is bounded, one has for all $\rho \ge 0$ and $q \in [0,\rho]$*

$$|\Psi_{P_{\mathrm{out},2}}(q;\rho)| \le \frac{1+\ln(2\pi\Delta)}{2} + \frac{2\sup|\varphi_2|^2}{\Delta}\,.$$

Proof. The upperbound $P_{\mathrm{out},2}\big(y\big|x\big) \le \sqrt{2\pi\Delta}^{-1}$ directly implies

$$\Psi_{P_{\mathrm{out},2}}(q;\rho) \le -\frac{1}{2}\ln(2\pi\Delta)\,.$$

By Jensen's inequality, one also has the lowerbound

$$\begin{aligned} \Psi_{P_{\mathrm{out},2}}(q;\rho) &\ge \mathbb{E}\int \mathcal{D}u dP_{A_2}(\mathbf{a}) \ln\frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}(\widetilde{Y}_0 - \varphi_2(\sqrt{q}\,V + \sqrt{\rho-q}\,u,\mathbf{a}))^2} \\ &\ge -\frac{1}{2\Delta}\mathbb{E}\int \mathcal{D}u dP_{A_2}(\mathbf{a})\big(\varphi_2(\sqrt{q}\,V + \sqrt{\rho-q}\,U,\mathbf{a}) - \varphi_2(\sqrt{q}\,V + \sqrt{\rho-q}\,u,\mathbf{a})\big)^2 \\ &\quad - \frac{1+\ln 2\pi\Delta}{2}\,. \end{aligned}$$

Put together, these lower and upper bounds give the lemma. □

To conclude on that section, the interpolating model is such that:

- **at t=0**, it recovers the two-layer GLM;
- **at t=1**, it reveals one scalar inference channel associated to the term $\Psi_{P_{\text{out},2}}$ and a one-layer GLM whose formula for the free entropy $\tilde{f}_{n_0,n_1}$ in the thermodynamic limit is already known from [35].

2.6 Free entropy variation along the interpolation path

From the Fundamental Theorem of Analysis $f_{\mathbf{n}}(1) - f_{\mathbf{n}}(0) = \int_0^1 \frac{df_{\mathbf{n}}(t)}{dt} dt$ and (102), (105), it follows

$$f_{\mathbf{n}} = -\alpha_1 \frac{r\rho_1}{2} + \alpha \Psi_{P_{\text{out},2}}\Big(\int_0^1 q(t)dt; \rho_1(n_0)\Big) - \int_0^1 \frac{df_{\mathbf{n}}(t)}{dt} dt \\ + \sup_{q_0 \in [0,\rho_0]} \inf_{r_0 \geq 0} \Big\{ \psi_{P_0}(r_0) + \alpha_1 \Psi_{\varphi_1}(q_0, r; \rho_0) - \frac{r_0 q_0}{2} \Big\} + \mathcal{O}_{\mathbf{n}}(1)\,. \quad (106)$$

Most of the terms that form the potential (91) can already be identified in the expression (106). For the missing terms to appear, the t-derivative of the free entropy has to be computed first.

Define $u_y(x) := \ln P_{\text{out},2}(y|x)$. Let $u'_y(x)$ be the derivative (w.r.t. x). In Appendix B one shows

Proposition 2 (Free entropy variation)**.** *The derivative of the free entropy* (101) *verifies, for all* $t \in (0,1)$*,*

$$\frac{df_{\mathbf{n}}(t)}{dt} = -\frac{1}{2}\frac{n_1}{n_0}\mathbb{E}\Big\langle \Big(\frac{1}{n_1}\sum_{\mu=1}^{n_2} u'_{Y_{t,\mu}}(S_{t,\mu}) u'_{Y_{t,\mu}}(s_{t,\mu}) - r \Big) \big(\hat{Q} - q(t)\big) \Big\rangle_t \\ + \frac{n_1}{n_0}\left(\frac{rq(t)}{2} - \frac{r\rho_1(n_0)}{2} \right) + \mathcal{O}_{\mathbf{n}}(1)\,, \quad (107)$$

where $\mathcal{O}_{\mathbf{n}}(1)$ *is a quantity that goes to* 0 *in the limit* $n_0, n_1, n_2 \to +\infty$*,* ***uniformly*** *in* $t \in [0,1]$*, and the overlap is*

$$\hat{Q} := \frac{1}{n_1}\sum_{i=1}^{n_1} \varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) \varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)\,. \quad (108)$$

2.7 Overlap concentration

We already know from [35] that the overlap $Q = \frac{1}{n_0}\sum_{j=1}^{n_0} x_j \cdot X_j^0$ concentrates. This concentration plays a key role in the proof of the thermodynamic limit of the free entropy $\tilde{f}_{n_0,n_1}$, still in [35]. The next lemma states that the overlap $\hat{Q}$ concentrates around its mean.

As in the one-layer case, a "small" perturbation to the interpolating estimation problem is introduced by *adding* to the Hamiltonian (98) a term

$$\sum_{i=1}^{n_1} \frac{\epsilon}{2}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big) - \epsilon \varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big) \varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \sqrt{\epsilon}\varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big) \widehat{Z}_i \quad (109)$$

where $(\widehat{Z}_i)_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. It corresponds to having extra observations coming from a side-channel

$$\widehat{Y}_i = \sqrt{\epsilon}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) + \widehat{Z}_i \quad \text{for } i = 1, \dots, n_1 \,. \tag{110}$$

It thus preserves the Nishimori identity (see Proposition 8). The new Hamiltonian defines a new Gibbs bracket $\langle - \rangle_{t,\epsilon}$ and free entropy $f_{\mathbf{n},\epsilon}(t)$. All the set up of Sec. 2.4 and Proposition 2 trivially extend. This perturbation induces only a small change in the free entropy, namely of the order of ϵ:

Lemma 2 (Small free entropy variation under perturbation)**.** *Let* $\sup \varphi_1^2$ *be the supremum of* φ_1^2 *(well-defined under the hypothesis (H2)). For all* $\epsilon > 0$ *and all* $t \in [0,1]$*,*

$$|f_{\mathbf{n},\epsilon}(t) - f_{\mathbf{n}}(t)| \leq \underbrace{\frac{n_1}{n_0}}_{\to \alpha_1} \frac{\sup \varphi_1^2}{2} \cdot \epsilon \,. \tag{111}$$

Proof. A simple computation gives $\frac{\partial f_{\mathbf{n},\epsilon}(t)}{\partial \epsilon} = -\frac{n_1}{n_0}\mathbb{E}\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon}$ where

$$\mathcal{L}_\epsilon := \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{2}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big) - \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \frac{1}{2\sqrt{\epsilon}}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\widehat{Z}_i \,.$$

In Appendix C.2 one proves Lemma 10, i.e. $\mathbb{E}\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon} = -\frac{1}{2}\mathbb{E}\langle \hat{Q} \rangle_{t,\epsilon}$. The trivial bound $|\hat{Q}| \leq \sup \varphi_1^2$ ends the proof. □

Besides, this small perturbation forces the overlap to concentrate around its mean:

Lemma 3 (Overlap concentration)**.** *For any* $0 < a < 1$*,*

$$\lim_{n_0 \to \infty} \int_a^1 d\epsilon \int_0^1 dt\, \mathbb{E}\big\langle \big(\hat{Q} - \mathbb{E}\langle \hat{Q} \rangle_{t,\epsilon}\big)^2 \big\rangle_{t,\epsilon} = 0 \,. \tag{112}$$

The proof of Lemma 3 is mostly the same as the one streamlined in Section V of [38]. One only needs to make slight changes to fit the proof to our problem. For this reason, Appendix C.2 sketches the main steps of the *adapted* proof and refers to [38] for details.
Lemma 3 implies that there exists a sequence $(\epsilon(n_0))_{n_0 \geq 1} \in (0,1)^{\mathbb{N}^*}$ that converges to 0 such that

$$\lim_{n_0 \to +\infty} \int_0^1 dt\, \mathbb{E}\big\langle \big(\hat{Q} - \mathbb{E}\langle \hat{Q} \rangle_{t,\epsilon(n_0)}\big)^2 \big\rangle_{t,\epsilon(n_0)} = 0 \,. \tag{113}$$

As $(\epsilon(n_0))_{n_0 \geq 1}$ converges to 0, $f_{\mathbf{n},\epsilon(n_0)}(t)$ and $f_{\mathbf{n}}(t)$ have the same limit (provided it exists) thanks to Lemma 2. In the next section, to lighten the notations, the perturbation subscript $\epsilon(n_0)$ is abusively removed since it makes no difference for computing the limit of the free entropy.

2.8 Canceling the remainder

Note from (106) and (91) that the second term appearing in (107) is precisely the missing one that is required to obtain the expression of the potential on the r.h.s. of (106) (recall Proposition 1 and make the identifications $r \leftrightarrow r_1$, $\int_0^1 q(t)dt \leftrightarrow q_1$). Hence, to prove Theorem 1, we would like to "cancel" the Gibbs bracket in (107), which is the so called *remainder* (once integrated over t). This is made possible thanks to the adaptive interpolating parameter q. One has to choose $q(t) = \mathbb{E}\langle \hat{Q} \rangle_t$, which is approximately equal to $\hat{Q}$ because it concentrates (see Lemma 3). However, $\mathbb{E}\langle \hat{Q} \rangle_t$ depends on $\int_0^t q(v)dv$ (and on r too). The equation $q(t) = \mathbb{E}\langle \hat{Q} \rangle_t$ is therefore a first order differential equation over $t \mapsto \int_0^t q(v)dv$.

Proposition 3 (Existence of the optimal interpolation function)**.** *For all $r \geq 0$ the differential equation*

$$q(t) = \mathbb{E}\langle \hat{Q} \rangle_t \tag{114}$$

admits a unique solution $q_{n_0}^{(r)}(t)$ on $[0, \rho_1(n_0)]$ and the mapping

$$r \geq 0 \mapsto \int_0^1 q_{n_0}^{(r)}(v)dv \tag{115}$$

is continuous.

Proof. Under (H2) one verifies easily that $\mathbb{E}\langle \hat{Q} \rangle_t$ is a bounded $\mathcal{C}^1$ function of $(\int_0^t q(v)dv, r)$. The proposition then follows from an application of the parametric Cauchy-Lipschitz theorem. □

This optimal choice for the interpolating function allows to relate the free entropy to the potential.

Proposition 4. *Let $r \geq 0$. For $n_0 \in \mathbb{N}^*$, $q_{n_0}^{(r)}$ is the solution of (114). Then*

$$f_{\mathbf{n}} = \sup_{q_0 \in [0,\rho_0]} \inf_{r_0 \geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, \int_0^1 q_{n_0}^{(r)}(v)dv, r; \rho_0, \rho_1(n_0)\Big) + \mathcal{O}_{\mathbf{n}}(1)\,. \tag{116}$$

Proof. By Cauchy-Schwarz inequality

$$\begin{aligned}
&\Big| \int_0^1 dt\, \mathbb{E}\Big\langle \Big(\frac{1}{n_1}\sum_{\mu=1}^{n_2} u'_{Y_{t,\mu}}(S_{t,\mu}) u'_{Y_{t,\mu}}(s_{t,\mu}) - r\Big)\big(\hat{Q} - q_{n_0}^{(r)}(t)\big)\Big\rangle_t \Big| \\
&\leq \Big[\int_0^1 dt\, \mathbb{E}\Big\langle \Big(\frac{1}{n_1}\sum_{\mu=1}^{n_2} u'_{Y_{t,\mu}}(S_{t,\mu}) u'_{Y_{t,\mu}}(s_{t,\mu}) - r\Big)^2\Big\rangle_t\Big]^{1/2}\Big[\int_0^1 dt\, \mathbb{E}\big\langle \big(\hat{Q} - q_{n_0}^{(r)}(t)\big)^2\big\rangle_t\Big]^{1/2} = \mathcal{O}_{\mathbf{n}}(1)\,.
\end{aligned}$$

The last equality uses that the first factor is bounded (independently of t) under assumptions (H1), (H2) and (H3) (similar proof to the one in Appendix A.5 of [35]), and that the second factor goes to 0 when $n_0, n_1, n_2 \to +\infty$ with $\epsilon = \epsilon(n_0)$ thanks to (113), (114). Making use of the latter result and $n_1/n_0 \to \alpha_1$, $\rho_1(n_0) \to \rho_1$, the integral of (107) reads

$$\int_0^1 \frac{df_{\mathbf{n}}(t)}{dt}dt = \alpha_1 \frac{r}{2}\int_0^1 q_{n_0}^{(r)}(t)dt - \alpha_1\frac{r\rho_1}{2} + \mathcal{O}_{\mathbf{n}}(1)\,, \tag{117}$$

Replacing this identity in (106) gives the desired result. □

2.9 Lower and upper matching bounds

To end the proof of Theorem 1 one has to go through the following two steps:

(i) Prove that under assumptions (H1), (H2) and (H3)

$$\lim_{\mathbf{n}\to\infty} f_{\mathbf{n}} = \sup_{r_1\geq 0} \inf_{q_1\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}(q_0, r_0, q_1, r_1; \rho_0, \rho_1)\,.$$

(ii) Invert the order of the optimizations on r_1 and q_1.

To tackle (i), one proves that $\liminf_{n_0\to\infty} f_{\mathbf{n}}$ and $\limsup_{n_0\to\infty} f_{\mathbf{n}}$ are – respectively – lowerbounded and upperbounded by the same quantity $\sup_{r_1\geq 0} \inf_{q_1\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}(q_0, r_0, q_1, r_1; \rho_0, \rho_1)$.

Proposition 5 (Lower bound)**.** *The free entropy* (82) *verifies*

$$\liminf_{n_0\to\infty} f_{\mathbf{n}} \geq \sup_{r\geq 0} \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big)\,. \tag{118}$$

Proof. By Proposition 4 we have that for any $r \geq 0$

$$f_{\mathbf{n}} \geq \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1(n_0)\Big) + \mathcal{O}_{\mathbf{n}}(1)\,. \tag{119}$$

By a continuity argument

$$\begin{aligned}\lim_{n_0\to\infty} \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1(n_0)\Big)& \\ = \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big)\,.&\end{aligned} \tag{120}$$

This limit, combined with (119), gives

$$\liminf_{n_0\to\infty} f_{\mathbf{n}} \geq \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big)\,. \tag{121}$$

This is true for all $r \geq 0$, thus we obtain Proposition 5. □

Proposition 6 (Upper bound)**.** *The free entropy* (82) *verifies*

$$\limsup_{n_0\to\infty} f_{\mathbf{n}} \leq \sup_{r\geq 0} \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\mathrm{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big)\,. \tag{122}$$

Proof. Let $K_{n_0} = 2\alpha_2 \Psi'_{P_{\mathrm{out},2}}(\rho_1(n_0); \rho_1(n_0))$, $\Psi'_{P_{\mathrm{out},2}}$ being the derivative of $\Psi_{P_{\mathrm{out},2}}$ w.r.t. its first argument. The latter is continuous and bounded (see Appendix A.2.2. of [35]). Also, (115) is a continuous mapping. It follows that

$$\begin{array}{ccc}[0, K_{n_0}] & \to & [0, K_{n_0}] \\ r & \mapsto & 2\alpha_2 \Psi'_{P_{\mathrm{out},2}}\big(\int_0^1 q_{n_0}^{(r)}(t)dt; \rho_1(n_0)\big)\end{array} \tag{123}$$

is continuous too. Therefore, it admits a fixed point

$$r^*(n_0) = 2\alpha_2 \Psi'_{P_{\text{out},2}}\left(\int_0^1 q_{n_0}^{(r^*(n_0))}(t)dt; \rho_1(n_0)\right) .$$

We now remark that

$$\sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, \int_0^1 q_{n_0}^{(r^*(n_0))}(t)dt, r^*(n_0); \rho_0, \rho_1(n_0)\Big)$$
$$= \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r^*(n_0); \rho_0, \rho_1(n_0)\Big) . \tag{124}$$

Indeed, the function

$$g_{r^*(n_0)} : q \in [0, \rho_1(n_0)] \mapsto \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r^*(n_0); \rho_0, \rho_1(n_0)\Big)$$

is convex. To see it, first remember that

$$g_{r^*(n_0)}(q) = \alpha \Psi_{P_{\text{out},2}}\big(q; \rho_1(n_0)\big) - \alpha_1 \frac{r^*(n_0)}{2} q + C , \tag{125}$$

where $C := \sup_{q_0} \inf_{r_0} \psi_{P_0}(r_0) + \alpha_1 \Psi_{\varphi_1}(q_0, r^*(n_0); \rho_0) - r_0 q_0/2$ does not depend on q. The convexity of $g_{r^*(n_0)}$ then follows of the convexity of $\Psi_{P_{\text{out},2}}\big(\,\cdot\,; \rho_1(n_0)\big)$ (see Proposition 11 in Appendix A.2.2. of [35]). $g_{r^*(n_0)}$ derivative is easily obtained from (125):

$$g'_{r^*(n_0)}(q) = \alpha \Psi'_{P_{\text{out},2}}\big(q; \rho_1(n_0)\big) - \alpha_1 \frac{r^*(n_0)}{2} . \tag{126}$$

By definition of $r^*(n_0)$, $g'_{r^*(n_0)}\left(\int_0^1 q_{n_0}^{(r^*(n_0))}(t)dt\right) = 0$ and the minimum of $g_{r^*(n_0)}$ is necessarily achieved at $\int_0^1 q_{n_0}^{(r^*(n_0))}(t)dt$. Proposition 4, combined with (124), gives

$$\begin{aligned} f_{\mathbf{n}} &= \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r^*(n_0); \rho_0, \rho_1(n_0)\Big) + \mathcal{O}_{\mathbf{n}}(1) \\ &\leq \sup_{r\geq 0} \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1(n_0)\Big) + \mathcal{O}_{\mathbf{n}}(1) . \end{aligned} \tag{127}$$

Finally, by a continuity argument, we have

$$\lim_{n_0\to\infty} \sup_{r\geq 0} \inf_{q\in[0,\rho_1(n_0)]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1(n_0)\Big)$$
$$= \sup_{r\geq 0} \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big) ,$$

and taking the limit in the inequality (127) ends the proof:

$$\limsup_{n_0\to\infty} f_{\mathbf{n}} \leq \sup_{r\geq 0} \inf_{q\in[0,\rho_1]} \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} f_{\text{RS}}\Big(q_0, r_0, q, r; \rho_0, \rho_1\Big) .$$

□

It remains to prove (ii), i.e.

Proposition 7 (Switch the optimization order)**.** *Under (H2), for any positive real numbers ρ_0 and ρ_1 one has*

$$\sup_{r_1\geq 0}\inf_{q_1\in[0,\rho_1]}\sup_{q_0\in[0,\rho_0]}\inf_{r_0\geq 0} f_{\mathrm{RS}}(q_0,r_0,q_1,r_1;\rho_0,\rho_1) \\ = \sup_{q_1\in[0,\rho_1]}\inf_{r_1\geq 0}\sup_{q_0\in[0,\rho_0]}\inf_{r_0\geq 0} f_{\mathrm{RS}}(q_0,r_0,q_1,r_1;\rho_0,\rho_1)\,.$$

Proof. Let $f:[0,+\infty[\to\mathbb{R}$ and $g:[0,\rho_1]\to\mathbb{R}$ be the two functions

$$f(r_1) := \sup_{q_0\in[0,\rho_0]}\inf_{r_0\geq 0}\left\{\psi_{P_0}(r_0)+\alpha_1\Psi_{\varphi_1}(q_0,r_1;\rho_0)-\frac{r_0q_0}{2}\right\},\ g(q_1):=\alpha\Psi_{P_{\mathrm{out},2}}(q_1;\rho_1)\,,$$

such that $\psi(r_1,q_1):=\sup_{q_0\in[0,\rho_0]}\inf_{r_0\geq 0} f_{\mathrm{RS}}(q_0,r_0,q_1,r_1;\rho_0,\rho_1)=f(r_1)+g(q_1)-\frac{\alpha_1}{2}r_1q_1$.
In Appendix A.3 it is shown that, under (H2), f is convex, Lipschitz and non-decreasing on $[0,+\infty[$. Proposition 11 in Appendix A.2.2 of [35] states that, under (H2), g is convex, Lipschitz and non-decreasing on $[0,\rho_1]$. The desired result is then obtained by applying Corollary 5 in Appendix E of [35]:

$$\sup_{r_1\geq 0}\inf_{q_1\in[0,\rho_1]}\psi(r_1,q_1)=\sup_{q_1\in[0,\rho_1]}\inf_{r_1\geq 0}\psi(r_1,q_1)\,.$$

□

3 Numerical experiments

3.1 Activations comparison in terms of mutual informations

Here we assume the exact same setting as the one presented in the main text to compare activation functions on a two-layer random weights network. We compare here the mutual information estimated with the proposed replica formula instead of the entropy behaviors discussed in the main text. As it was the case for entropies, we can see that the saturation of the double-side saturated hardtanh leads to a loss of information for large weights, while the mutual informations are always increasing for linear and ReLU activations.

3.2 Learning ability of USV-layers

To ensure weight matrices remain close enough to being independent during learning we instroduce USV-layers, corresponding to a custom type of weight constraint. We recall that in such layers, weight matrices are decomposed in the manner of a singular value decomposition, $W_\ell = U_\ell S_\ell V_\ell$, with by U_ℓ and V_ℓ drawn from the corresponding Haar measures (i.e. uniformly among the orthogonal matrices of given size), and S_ℓ contrained to be diagonal, being the only matrix being learned. In the main text, we demonstrate on a linear network that the USV-layers ensure that the assumptions necessary to our replica formula are met with learned matrices in the case of linear networks. Nevertheless, a USV-layer of size $N\times N$ has only N trainable parameters, which implies that they are harder to train than usual fully connected layers. In practice, we notice that they tend to require more parameter updates and that interleaving linear USV-layers to increase the number of parameters between non-linearities can significantly improve the final result of training.

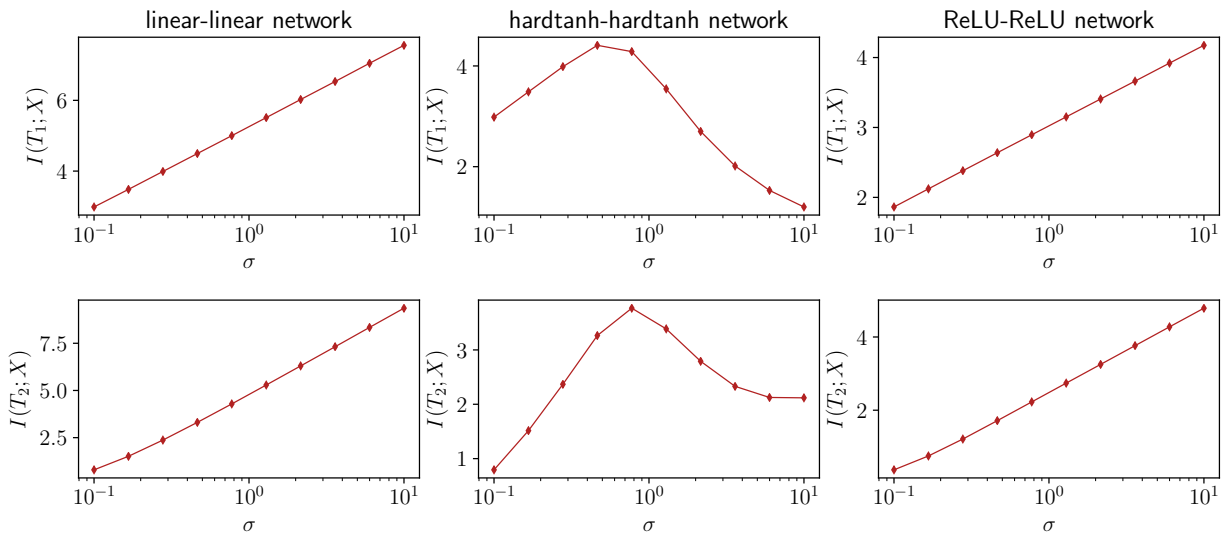

Figure 5: Replica mutual informations between latent and input variables in stochastic networks $\boldsymbol{X} \to \boldsymbol{T}_1 \to \boldsymbol{T}_2$, with equally sized layers $N = 1000$, inputs drawn from $\mathcal{N}(0, I_N)$, weights from $\mathcal{N}(0, \sigma^2 I_{N^2}/N)$, as a function of the weight scaling parameter σ. An additive white Gaussian noise $\mathcal{N}(0, 10^{-5} I_N)$ is added inside the non-linearity. Left column: linear network. Center column: hardtanh-hardtanh network. Right column: ReLU-ReLU network.

To convince ourselves that the training ability of USV-layers is still relevant to study learning dynamics on real data we conduct an experiment on the MNIST dataset. We study the classification problem of the classical MNIST data set (60 000 training images and 10 000 testing images) with a simple fully-connected network featuring one non-linear (ReLU) hidden layer of 500 neurones. On top of the ReLU-layer, we place a softmax output layer where the 500×10 parameters of the weight matrix are all being learned in all the versions of the experiments. Conversely, before the ReLU layer, we either (1) do not learn at all the 784×500 parameters which then define random projections, (2) learn all of them as a traditional fully connected network, (3) use a combination of 2 (3a), 3 (3b) or 6 (3c) consecutive USV-layers (without any intermediate non-linearity). The best train and test, losses and accuracies, for the different architectures are given in Table 1 and some learning curves are displayed on Figure 6. As expected we observe that USV-layers are acheiving better classification success than the random projections, yet worse than the unconstrained fully connected layer. Interestingly, stacking USV-layers to increase the number of trainable parameters allows to reach very good training accuracies, nevertheless, the testing accuracies do not benefit to the same extent from these additional parameters. On Figure 6, we can actually see that the version of the experiment with 6 USV-layers overfits the training set (green curves with testing losses growing towards the end of learning). Therefore, particularly in this case, adding regularizers might allow to improve the generalization performances of models with USV-layers.

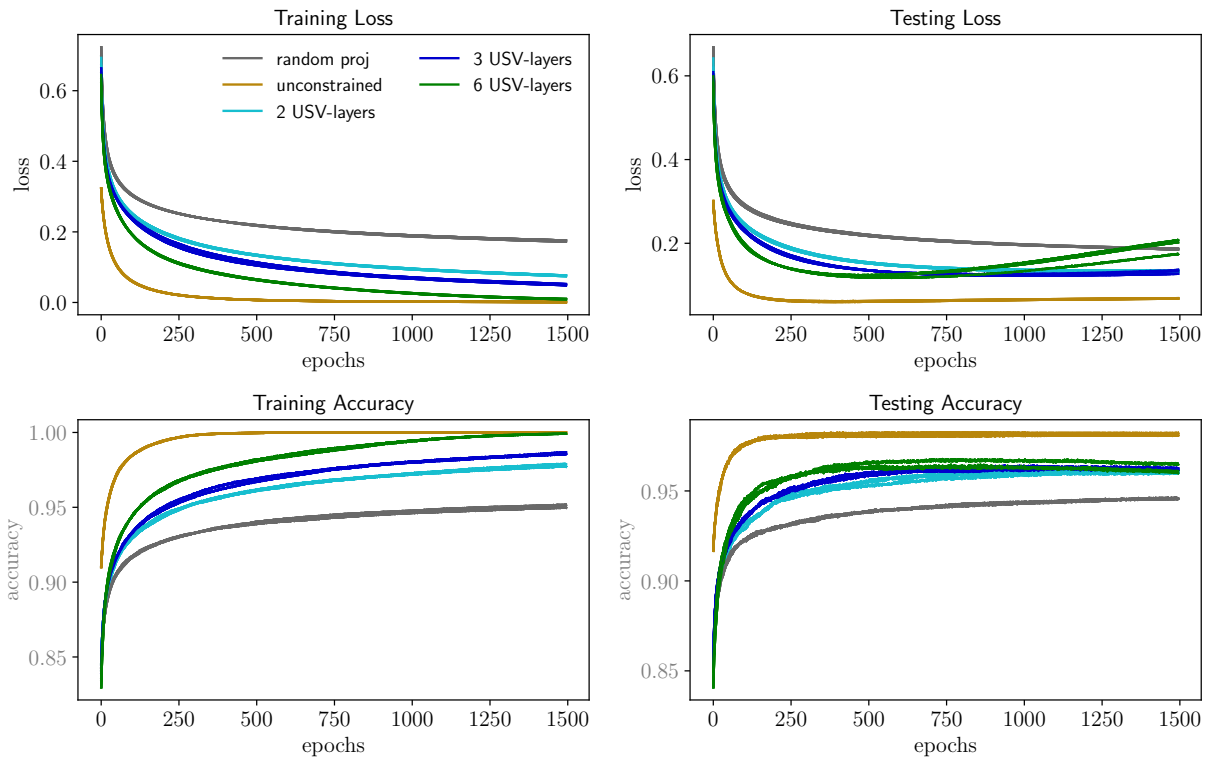

Figure 6: Training and testing curves for the training of a two-layer neural net on the classification of MNIST for different constraints on the first layer (further details are given in Section3.2). For each version of the experiment the outcomes of two independent runs are plotted with the same color, it is not always possible to distinguish the two runs as they overlap.

3.3 Additional learning experiments on synthetic data

Similarly to the experiments of the main text, we consider simple training schemes with constant learning rates, no momentum, and no explicit regularization.

We first include a second version of Figure 4 of the main text, corresponding to the exact same experiment with a different random seed and check that results are qualitatively identical.

We consider then a regression task created by a 2-layer teacher network of sizes 500-3-3, activations ReLU-linear, uncorrelated input data distribution $\mathcal{N}(0, I_{N_X})$ and additive white Gaussian noise at the output of variance 0.01. The matrices of the teacher network are i.i.d. normally distributed with a variance equal to the inverse of the layer input dimension. We train a student network with 2 ReLU layers of sizes 2500 and 1000, each featuring 5 stacked USV-layers of same size before the non linear activation, and with one final fully-connected linear layer. We use plain SGD with a constant learning rate of 0.01 and a batchsize of 50. In Figure 8 we plot the mutual informations with the input at the effective 10-hidden layers along the training. Except for the very first layer where we observe a slight initial increase, all mutual informations appear to only decrease during the learning, at least at this resolution (i.e. after the first epoch). We thus observe a compression

First layer type	#train params	Train loss	Test loss	Train acc	Test acc
Random (1)	0	0.1745	0.1860	95.05 (0.09)	94.61 (0.02)
Unconstrained (2)	784×500	0.0012	0.0605	100. (0.00)	98.18 (0.06)
2-USV (3a)	2×500	0.0758	0.1326	97.80 (0.07)	96.10 (0.03)
3-USV (3b)	3×500	0.0501	0.1238	98.62 (0.05)	96.35 (0.04)
6-USV (3c)	6×500	0.0092	0.1211	99.93 (0.01)	96.54 (0.17)

Table 1: Training results for MNIST classification of a fully connected 784-500-10 neural net with a ReLU non linearity. The different rows correspond to different specifications of trainable parameters in the first layer (1, 2, 3a, 3b, 3c) describe in the paragraph. We use plain SGD to minimize the cross-entropy loss. All experiments use the same learning rate 0.01 and batchsize of 100 samples. Results are averaged over 5 independent runs, and standard deviations are reported in parentheses.

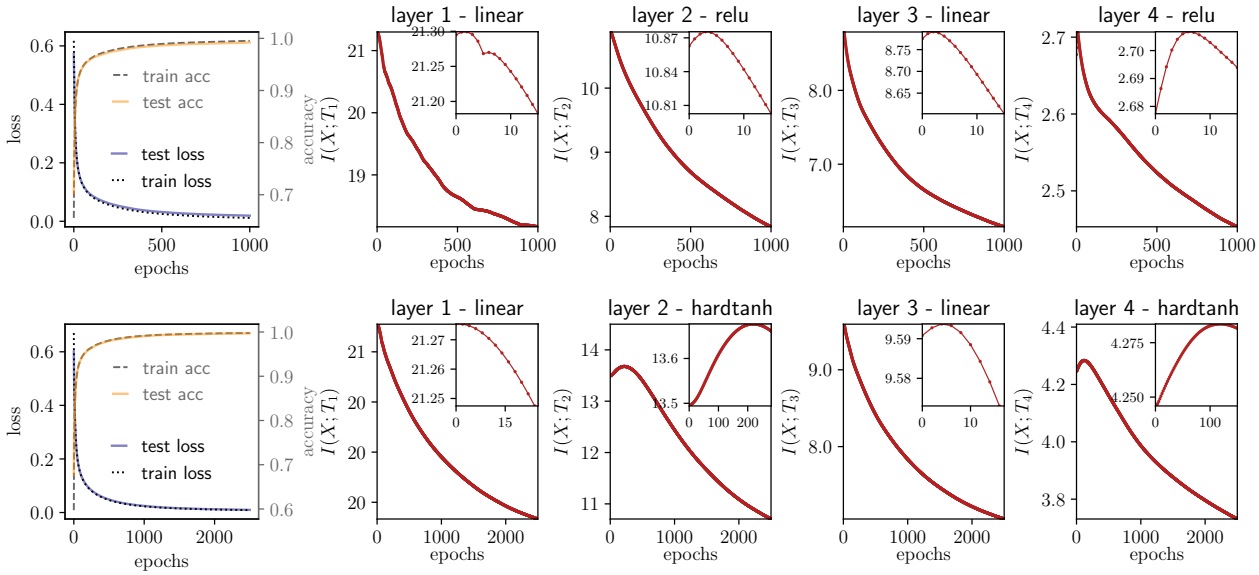

Figure 7: Independent run outcome for Figure 4 of the main text. Training of two recognition models on a binary classification task with correlated input data and either ReLU (top) or hardtanh (bottom) non-linearities. Left: training and generalization cross-entropy loss (left axis) and accuracies (right axis) during learning. Best training-testing accuracies are 0.995 - 0.992 for ReLU version (top row) and 0.998 - 0.997 for hardtanh version (bottom row). Remaining colums: mutual information between the input and successive hidden layers. Insets zoom on the first epochs.

even in the absence of double-saturated non-linearities. We further note that in this case we observe an accuentuation of the amount of compression with layer depth as observed by [39] (see second plot of first row of Figure 8), but which we did not observe in the binary classification experiment presented in the main text. On Figure 9, we reproduce the figure for a different seed.

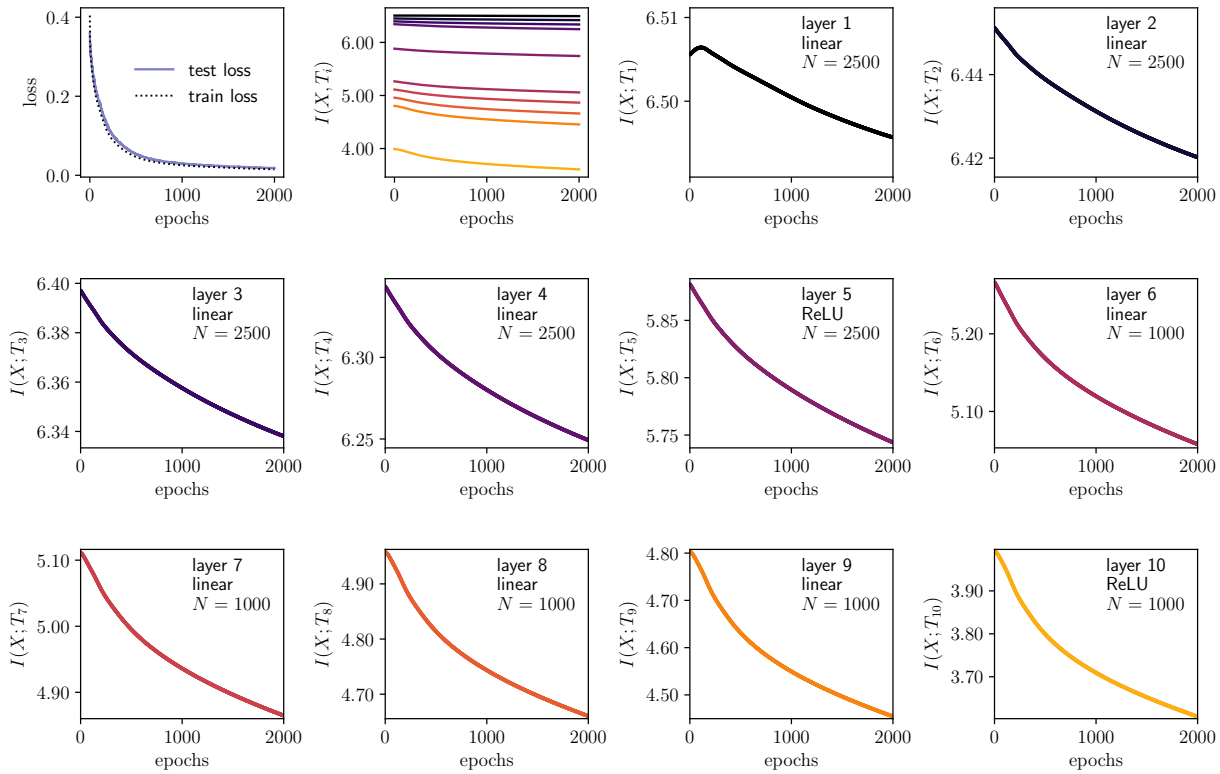

Figure 8: Example of regression with a 10 hidden-layer student network: 5 USV-layers - ReLU activation - 5 USV-layers - ReLu activation - 1 unconstrained final linear layer, on dataset generated by a non-linear teacher network: ReLu-linear. Top row, first plot: training and testing MSE loss along learning. Best train loss is 0.015, best test loss is 0.018. Top row, second plot: mutual informations curves of the 10 hidden layers showing the slight accentuation of compression in deeper layers. Remaining: mutual information from each layer displayed separately.

In a last experiment, we even show that merely changing the weight initialization can drastically change the behavior of mutual informations during training while resulting in identical training and testing final performances. We consider here a setting closely related to the classification on correlated data presented in the main text. The generative model is a a simple single layer generative model $\boldsymbol{X} = \tilde{W}_{\text{gen}}\boldsymbol{Y} + \epsilon$ with normally distributed code $\boldsymbol{Y} \sim \mathcal{N}(0, I_{N_Y})$ of size $N_Y = 100$, from which data of size $N_X = 500$ are generated with matrix $\tilde{W}_{\text{gen}}$ i.i.d. normally distributed as $\mathcal{N}(0, 1/\sqrt{N_Y})$ and noise $\epsilon \sim \mathcal{N}(0, 0.01 I_{N_X})$. The recognition model attempts to solve the binary classification problem of recovering the label $y = \text{sign}(Y_1)$, the sign of the first neuron in $\boldsymbol{Y}$. Again the training is done with plain SGD to minimize the cross- entropy loss and the rest of the initial code $(Y_2, ..Y_{N_Y})$ acts as noise/nuisance with respect to the learning task. On Figure 10 we compare 3 identical 5-layers recognition models with sizes 500-1000-500-250-100-2, and activations hardtanh-hardtanh-hardtanh-hartanh-softmax. For the model presented at the top row, initial weights were sampled according to $W_{\ell,ij} \sim \mathcal{N}(0, 4/N_{\ell-1})$, for the model of the middle row $\mathcal{N}(0, 1/N_{\ell-1})$ was used instead, and finally

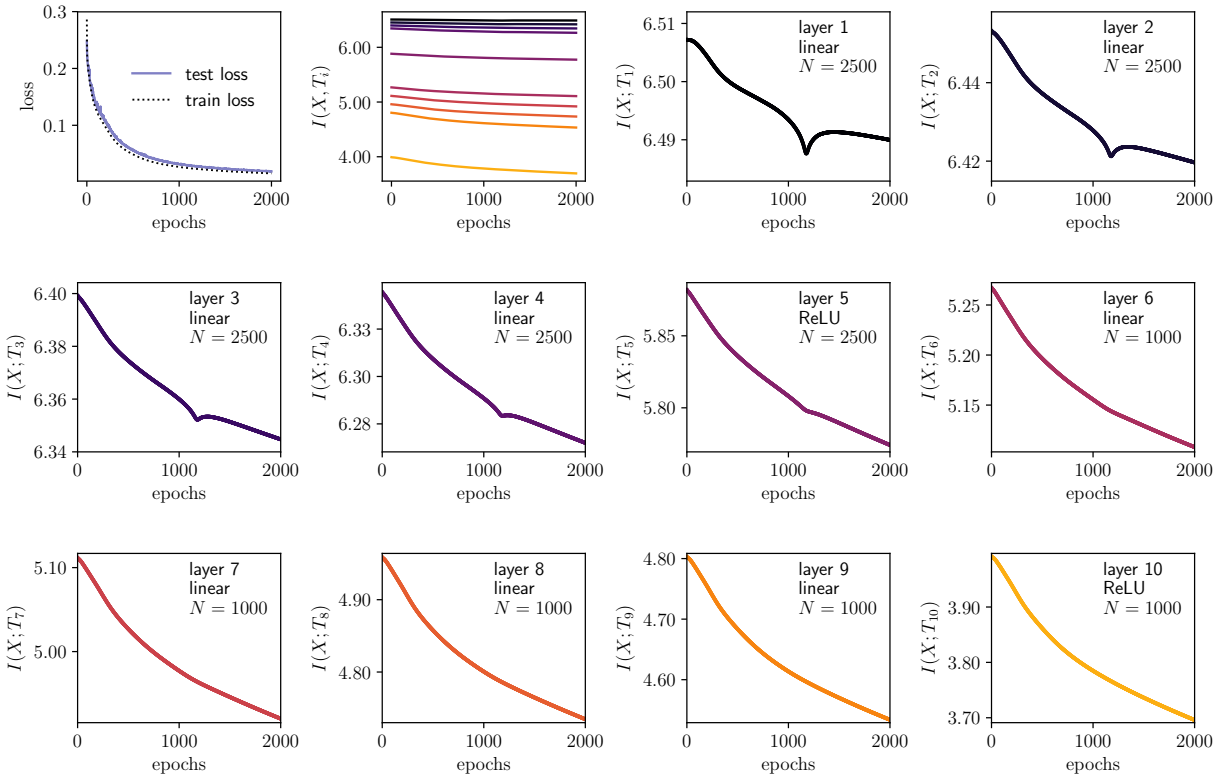

Figure 9: Independent run outcome for Figure 8 of the Supplementary Material. Example of regression with a 10 hidden-layer student network: 5 USV-layers - ReLU activation - 5 USV-layers - ReLu activation - 1 unconstrained final linear layer, on dataset generated by a non-linear teacher network: ReLu-linear. Top row, first plot: training and testing MSE loss along learning. Best train loss is 0.015, best test loss is 0.019. Top row, second plot: mutual informations curves of the 10 hidden layers showing the slight accentuation of compression in deeper layers. Remaining: mutual information from each layer displayed separately.

$\mathcal{N}(0, 1/4N_{\ell-1})$ for the bottom row. The first column shows that training is delayed for the weight initialized at smaller values, but eventually catches up and reaches accuracies superior to 0.97 both in training and testing. Meanwhile, mutual informations have different initial values for the different weight initializations and follow very different paths. They either decrease during the entire learning, or on the contrary are only increasing, or actually feature an hybrid path. We further note that it is to some extent surprising that the mutual information would increase at all in the first row if we expect the hardtanh saturation to instead induce compression. Figure 11 presents a second run of the same experiment with a different random seed. Findings are identical.

These observed differences and non-trivial observations raise numerous questions, and suggest that within the examined setting, a simple information theory of deep learning remains out-of-reach.

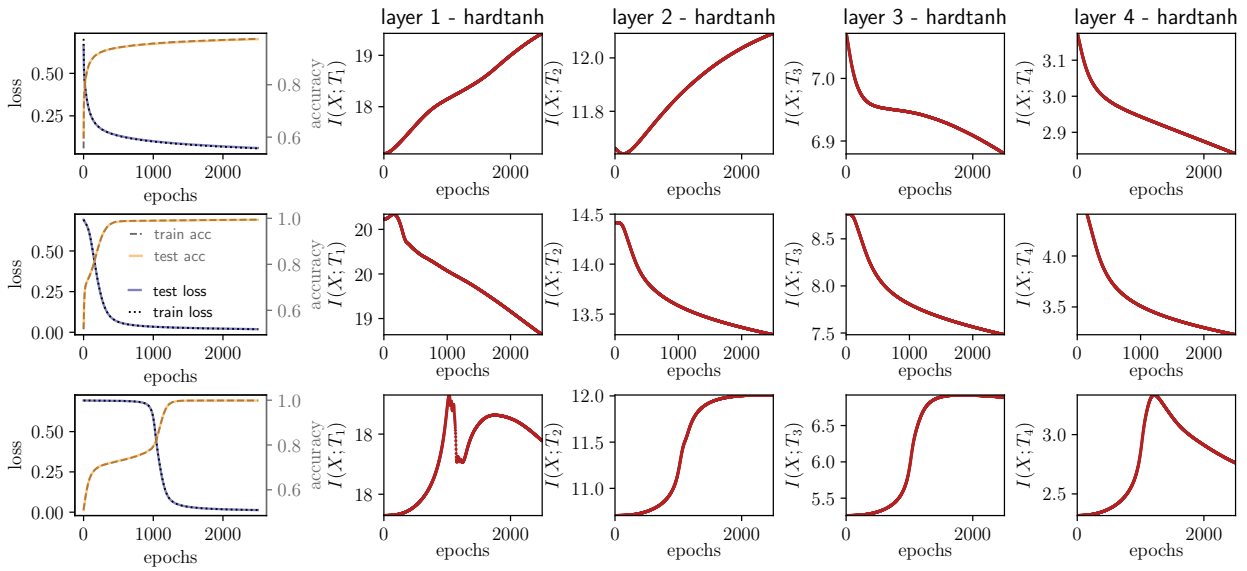

Figure 10: Learning and hidden-layers mutual information curves for a classification problem with correlated input data, using a 4-USV hardtanh layers and 1 unconstrained softmax layer, from 3 different initializations. Top: Initial weights at layer ℓ of variance $4/N_{\ell-1}$, best training accuracy 0.999, best test accuracy 0.994. Middle: Initial weights at layer ℓ of variance $1/N_{\ell-1}$, best train accuracy 0.994, best test accuracy 0.9937. Bottom: Initial weights at layer ℓ of variance $0.25/N_{\ell-1}$, best train accuracy 0.975, best test accuracy 0.974. The overall direction of evolution of the mutual information can be flipped by a change in weight initialization without changing drastically final performance in the classification task.

A Proofs of some technical propositions

A.1 The Nishimori identity

Proposition 8 (Nishimori identity)**.** *Let $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be a couple of random variables. Let $k \geq 1$ and let $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)}$ be k i.i.d. samples (given $\mathbf{Y}$) from the conditional distribution $P(\mathbf{X} = \cdot\,|\mathbf{Y})$, independently of every other random variables. Let us denote $\langle - \rangle$ the expectation operator w.r.t. $P(\mathbf{X} = \cdot\,|\mathbf{Y})$ and $\mathbb{E}$ the expectation w.r.t. $(\mathbf{X}, \mathbf{Y})$. Then, for all continuous bounded function g we have*

$$\mathbb{E}\langle g(\mathbf{Y}, \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)})\rangle = \mathbb{E}\langle g(\mathbf{Y}, \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k-1)}, \mathbf{X})\rangle\,. \tag{128}$$

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

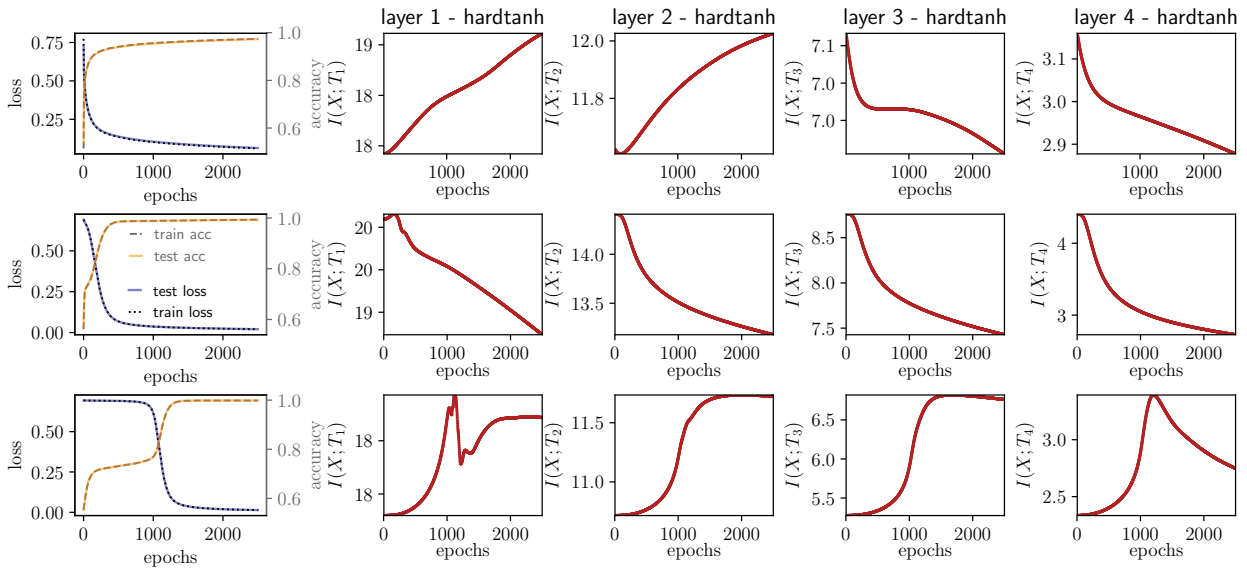

Figure 11: Independent run outcome for Figure 10 of the Supplementary Material. Learning and hidden-layers mutual information curves for a classification problem with correlated input data, using a 4-USV hardtanh layers and 1 unconstrained softmax layer, from 3 different initializations. Top: Initial weights at layer ℓ of variance $4/N_{\ell-1}$, best training accuracy 0.999, best test accuracy 0.998. Middle: Initial weights at layer ℓ of variance $1/N_{\ell-1}$, best train accuracy 0.9935, best test accuracy 0.9933. Bottom: Initial weights at layer ℓ of variance $0.25/N_{\ell-1}$, best train accuracy 0.974, best test accuracy 0.973. The overall direction of evolution of the mutual information can be flipped by a change in weight initialization without changing drastically final performance in the classification task.

A.2 Limit of the sequence $(\rho_1(n_0))_{n_0\geq 1}$

Here one proves Proposition 1, i.e. that the sequence $(\rho_1(n_0))_{n_0\geq 1}$ converges to $\rho_1 := \mathbb{E}[\varphi_1^2(T, \mathbf{A}_1)]$ where $T \sim \mathcal{N}(0, \rho_0)$, $\mathbf{A}_1 \sim P_{A_1}$ under the hypotheses (H1), (H2), (H3).

If $\rho_0 = 0$ then $\mathbf{X}^0 = 0$ almost surely (a.s.) and $\rho_1(n_0) = \mathbb{E}\varphi_1^2(0, \mathbf{A}_1) = \rho_1$ for every $n_0 \geq 1$, making the result trivial.

From now on, assume $\rho_0 > 0$. Given $\mathbf{X}^0$, one has $\left[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\right]_1 \sim \mathcal{N}\left(0, \frac{\|\mathbf{X}^0\|^2}{n_0}\right)$. Therefore

$$\rho_1(n_0) := \mathbb{E}\left[\varphi_1^2\left(\left[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\right]_1, \mathbf{A}_1\right)\right] = \mathbb{E}\int dt\, dP_{A_1}(\mathbf{a})\varphi_1^2(t,\mathbf{a})\frac{\exp -\frac{t^2}{2\frac{\|\mathbf{X}^0\|^2}{n_0}}}{\sqrt{2\pi\frac{\|\mathbf{X}^0\|^2}{n_0}}} = \mathbb{E}\left[h\left(\frac{\|\mathbf{X}^0\|^2}{n_0}\right)\right],$$

where $h : v \mapsto \int dt dP_{A_1}(\mathbf{a})\varphi_1^2(t,\mathbf{a})\frac{1}{\sqrt{2\pi v}}\exp(-t^2/2v)$ is a function on $]0, +\infty[$. It is easily shown to be continuous under (H2) thanks to the *dominated convergence theorem*.

By the Strong Law of Large Numbers, $\|\mathbf{X}^0\|^2/n_0$ converges a.s. to ρ_0. Combined with the continuity

of h, one has

$$\lim_{n_0\to+\infty} h\left(\frac{\|\mathbf{X}^0\|^2}{n_0}\right) \overset{\text{a.s.}}{=} h(\rho_0) = \rho_1\,.$$

Noticing that $|h\left(\|\mathbf{X}^0\|^2/n_0\right)| \leq \sup\varphi_1^2$, the *dominated convergence theorem* gives

$$\rho_1(n_0) = \mathbb{E}\left[h\left(\frac{\|\mathbf{X}^0\|^2}{n_0}\right)\right] \xrightarrow[n_0\to+\infty]{} \mathbb{E}\left[\lim_{n_0\to+\infty} h\left(\frac{\|\mathbf{X}^0\|^2}{n_0}\right)\right] = \rho_1\,.$$

A.3 Properties of the third scalar channel

Proposition 9. *Assume φ_1 is bounded (as it is the case under (H2)). Let $V, U \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\rho_0 \geq 0$, $q_0 \in [0,\rho_0]$. For any $r \geq 0$, $Y_0'^{(r)} = \sqrt{r}\varphi_1(\sqrt{q}\,V + \sqrt{\rho-q}\,U, \mathbf{A}_1) + Z'$ where $Z' \sim \mathcal{N}(0,1)$, $\mathbf{A}_1 \sim P_{A_1}$. The function*

$$\Psi_{\varphi_1}(q_0,\,\cdot\,;\rho_0) : r \mapsto \mathbb{E}\ln\int \mathcal{D}u P_{\text{out},1}^{(r)}\big(Y_0'^{(r)}\big|\sqrt{q}\,V + \sqrt{\rho-q}\,u\big).$$

is twice-differentiable, convex, non-decreasing and $\frac{\rho_1}{2}$-Lipschitz on $\mathbb{R}_+$. Then the function

$$f : r \mapsto \sup_{q_0\in[0,\rho_0]} \inf_{r_0\geq 0} \psi_{P_0}(r_0) + \alpha_1\Psi_\varphi(q_0,r;\rho_0) - \frac{r_0 q_0}{2}$$

is convex, non-decreasing and $\left(\alpha_1\frac{\rho_1}{2}\right)$-Lipschitz on $\mathbb{R}_+$.

Proof. For fixed ρ_0 and q_0, let $\Psi_{\varphi_1} \equiv \Psi_{\varphi_1}(q_0,\,\cdot\,;\rho_0)$. Note that

$$\Psi_{\varphi_1}(r) = \mathbb{E}\bigg[\int dy_0' \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_0' - \sqrt{r}\varphi_1(\sqrt{q_0}V + \sqrt{\rho_0-q_0}U, \mathbf{A}_1))^2} \cdot \ln\int \mathcal{D}u\, dP_{A_1}(\mathbf{a}) e^{\sqrt{r}y_0'\varphi_1(\sqrt{q_0}V+\sqrt{\rho_0-q_0}u,\mathbf{a}) - \frac{r}{2}\varphi_1^2(\sqrt{q_0}V+\sqrt{\rho_0-q_0}u,\mathbf{a})}\bigg].$$

With the properties imposed on φ_1, all the domination hypotheses to prove the twice-differentiability of ψ_{φ_1} are reunited. Denote $\langle - \rangle_r$ the expectation operator w.r.t. the joint posterior distribution

$$dP(u,\mathbf{a}|Y_0',V) = \frac{1}{\mathcal{Z}(Y_0',V)}\mathcal{D}u\, dP_{A_1}(\mathbf{a}) e^{\sqrt{r}y_0'\varphi_1(\sqrt{q_0}V+\sqrt{\rho_0-q_0}u,\mathbf{a}) - \frac{r}{2}\varphi_1^2(\sqrt{q_0}V+\sqrt{\rho_0-q_0}u,\mathbf{a})}\,,$$

where $\mathcal{Z}(Y_0',V)$ is a normalization factor. Using Gaussian integration by parts and the Nishimori property (Proposition 8), one verifies that for all $r \geq 0$

$$\begin{aligned}
\Psi_{\varphi_1}'(r) &= \frac{1}{2}\mathbb{E}\big[\langle\varphi_1(\sqrt{q_0}V + \sqrt{\rho_0-q_0}u, \mathbf{a})\rangle_r^2\big] \geq 0\,,\\
\Psi_{\varphi_1}''(r) &= \frac{1}{2}\mathbb{E}\big[\big(\langle\varphi_1^2(\sqrt{q_0}V + \sqrt{\rho_0-q_0}u, \mathbf{a})\rangle_r - \langle\varphi_1(\sqrt{q_0}V + \sqrt{\rho_0-q_0}u, \mathbf{a})\rangle_r^2\big)^2\big] \geq 0\,.
\end{aligned}$$

Hence Ψ_{φ_1} is non-decreasing and convex. The Lipschitzianity follows simply from

$$\big|\Psi_{\varphi_1}'(r)\big| \leq \frac{1}{2}\big|\mathbb{E}\langle\varphi_1^2(\sqrt{q_0}V + \sqrt{\rho_0-q_0}u, \mathbf{a})\rangle_r\big| = \frac{1}{2}\big|\mathbb{E}[\varphi_1^2(\sqrt{q_0}V + \sqrt{\rho_0-q_0}U, \mathbf{A}_1)]\big| = \frac{1}{2}\rho_1\,.$$

The Nishimori identity was used once again to obtain the penultimate equality. Finally, f properties are direct consequences of its definition as the "sup inf" of convex, non-decreasing, $\frac{\rho_1}{2}$-lipschitzian functions. □

B Proof of Proposition 2

Recall $u'_y(x)$ is the x-derivative of $u_y(x) := \ln P_{\text{out},2}(y|x)$. Denote $P'_{\text{out},2}(y|x)$ and $P''_{\text{out},2}(y|x)$ the first and second x-derivatives of $P_{\text{out},2}(y|x)$, respectively. First one shows that for all $t \in (0,1)$

$$\frac{df_{\mathbf{n}}(t)}{dt} = -\frac{1}{2}\frac{n_1}{n_0}\mathbb{E}\Big\langle\Big(\frac{1}{n_1}\sum_{\mu=1}^{n_2} u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu}) - r\Big)\big(\hat{Q} - q(t)\big)\Big\rangle_t + \frac{n_1}{n_0}\left(\frac{rq(t)}{2} - \frac{r\rho_1(n_0)}{2}\right) - \frac{A_{\mathbf{n}}(t)}{2}\,, \quad (129)$$

where the *overlap* is $\hat{Q} := \frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)$ and

$$A_{\mathbf{n}}(t) := \mathbb{E}\left[\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\text{out},2}(Y_{t,\mu}|S_{t,\mu})}{P_{\text{out},2}(Y_{t,\mu}|S_{t,\mu})}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\frac{\ln \mathcal{Z}_t}{n_0}\right]. \quad (130)$$

Once this is done, one proves that $A_{\mathbf{n}}(t)$ goes to 0 uniformly in $t \in [0,1]$ as $n_0, n_1, n_2 \to +\infty$ (while ${}^{n_1}\!/_{n_0} \to \alpha_1$, ${}^{n_2}\!/_{n_1} \to \alpha_2$), thus proving Proposition 2.

B.1 Proof of (129)

Recall definition (101). Once written as a function of the interpolating Hamiltonian (98), it becomes

$$f_{\mathbf{n}}(t) = \frac{1}{n_0}\mathbb{E}_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{V}}\left[\int d\mathbf{Y}d\mathbf{Y}'dP_0(\mathbf{X}^0)dP_{A_1}(\mathbf{A}_1)\mathcal{D}\mathbf{U}(2\pi)^{-\frac{n_1}{2}}e^{-\mathcal{H}_t(\mathbf{X}^0,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})} \cdot \ln\int dP_0(\mathbf{x})dP_{A_1}(\mathbf{a}_1)\mathcal{D}\mathbf{u}\, e^{-\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})}\right]. \quad (131)$$

Here, and from now on, one drops the dependence on t when writing $\mathbf{Y}$ and $\mathbf{Y}'$ as they are now dummy variables on which the integration is performed. We will need the Hamiltonian t-derivative $\mathcal{H}'_t$ given by

$$\mathcal{H}'_t(\mathbf{X}^0,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V}) = -\sum_{\mu=1}^{n_2}\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu}) - \frac{1}{2}\sqrt{\frac{r}{t}}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\left(Y'_i - \sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\right). \quad (132)$$

The derivative of the interpolating free entropy for $0 < t < 1$ thus reads

$$\frac{df_{\mathbf{n}}(t)}{dt} = \underbrace{-\frac{1}{n_0}\mathbb{E}\big[\mathcal{H}'_t(\mathbf{X}^*,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})\ln\mathcal{Z}_t\big]}_{T_1} \underbrace{-\frac{1}{n_0}\mathbb{E}\big\langle\mathcal{H}'_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})\big\rangle_t}_{T_2}, \quad (133)$$

where $\mathcal{Z}_t \equiv \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{W}_1, \mathbf{W}_2, \mathbf{V})$ is defined in (100).

In the remaining part of this subsection B.1, to lighten notations, one will omit the second argument of the function φ_1 except in a few occasions, i.e. one will write for $i = 1 \dots n_1$

$$\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big) \equiv \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\,,\ \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big) \equiv \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)\,.$$

It does not hurt the understanding of the derivation of (129) as the latter relies on integration by parts w.r.t. the Gaussian random variables $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{Z}'$.
Let first compute T_1. For $1 \le \mu \le n_2$ one has from (95)

$$\begin{aligned}
-\mathbb{E}\Big[\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Big] &= \frac{1}{2}\mathbb{E}\Big[\frac{1}{\sqrt{n_1(1-t)}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big)\Big]_\mu u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Big] \\
&\quad - \frac{1}{2}\mathbb{E}\Big[\Big(\frac{q(t)}{\sqrt{\int_0^t q(s)ds}}V_\mu + \frac{\rho_1(n_0)-q(t)}{\sqrt{\int_0^t(\rho_1(n_0)-q(s))ds}}U_\mu\Big)u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Big]\,. \quad (134)
\end{aligned}$$

By Gaussian integration by parts w.r.t $(\mathbf{W}_2)_{\mu i}$, $1 \le i \le n_1$, the first expectation becomes

$$\begin{aligned}
&\frac{1}{\sqrt{n_1(1-t)}}\mathbb{E}\Big[\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big)\Big]_\mu u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Big] \\
&= \frac{1}{\sqrt{n_1(1-t)}}\sum_{i=1}^{n_1}\mathbb{E}\Big[\int d\mathbf{Y}d\mathbf{Y}'(2\pi)^{-\frac{n_1}{2}}e^{-\mathcal{H}_t(\mathbf{X}^0,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})} \quad (135) \\
&\qquad\qquad \cdot(\mathbf{W}_2)_{\mu i}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Big] \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}\Big[\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\big(u''_{Y_\mu}(S_{t,\mu}) + u'_{Y_\mu}(S_{t,\mu})^2\big)\ln\mathcal{Z}_t\Big] \\
&\qquad + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}\Big\langle\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)u'_{Y_\mu}(S_{t,\mu})u'_{Y_\mu}(s_{t,\mu})\Big\rangle_t \\
&= \mathbb{E}\Big[\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\ln\mathcal{Z}_t\Big] \\
&\qquad + \mathbb{E}\Big\langle\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)u'_{Y_\mu}(S_{t,\mu})u'_{Y_\mu}(s_{t,\mu})\Big\rangle_t\,. \quad (136)
\end{aligned}$$

In the last equality we used the identity $u''_{Y_\mu}(x) + u'_{Y_\mu}(x)^2 = \frac{P''_{\mathrm{out},2}(Y_\mu|x)}{P_{\mathrm{out},2}(Y_\mu|x)}$.
Now one looks to the second expectation in the right hand side of (134). Using again Gaussian

integrations by parts, but this time w.r.t $V_\mu, U_\mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, one similarly obtains

$$\begin{aligned}
&\mathbb{E}\Bigg[\bigg(\frac{q(t)}{\sqrt{\int_0^t q(s)ds}}V_\mu + \frac{\rho_1(n_0)-q(t)}{\sqrt{\int_0^t(\rho_1(n_0)-q(s))ds}}U_\mu\bigg)u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Bigg] \\
&\qquad= \mathbb{E}\Bigg[\int d\mathbf{Y}d\mathbf{Y}'(2\pi)^{-\frac{n_1}{2}}e^{-\mathcal{H}_t(\mathbf{X}^0,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})} \\
&\qquad\qquad\cdot\bigg(\frac{q(t)}{\sqrt{\int_0^t q(s)ds}}V_\mu + \frac{\rho_1(n_0)-q(t)}{\sqrt{\int_0^t(\rho_1(n_0)-q(s))ds}}U_\mu\bigg)u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\Bigg] \\
&\qquad= \mathbb{E}\bigg[\rho_1(n_0)\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\ln\mathcal{Z}_t\bigg] + \mathbb{E}\big\langle q(t)u'_{Y_\mu}(S_{t,\mu})u'_{Y_\mu}(s_{t,\mu})\big\rangle_t\,. \qquad (137)
\end{aligned}$$

Combining equations (134), (135) and (137) together gives us

$$\begin{aligned}
-\mathbb{E}\bigg[\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu})\ln\mathcal{Z}_t\bigg] &= \frac{1}{2}\mathbb{E}\bigg[\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\Big(\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1^2\Big(\big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\big]_i\Big)-\rho_1(n_0)\Big)\ln\mathcal{Z}_t\bigg] \\
&\quad+\frac{1}{2}\mathbb{E}\bigg\langle\bigg\{\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)-q(t)\bigg\}u'_{Y_\mu}(S_{t,\mu})u'_{Y_\mu}(s_{t,\mu})\bigg\rangle_t\,. \qquad (138)
\end{aligned}$$

As seen from (132), (133) it remains to compute $\mathbb{E}\Big[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big(Y'_i-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big)\ln\mathcal{Z}_t\Big]$. Recalling that $Y'_i-\sqrt{rt}\varphi_1\Big(\big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\big]_i\Big)=Z'_i$ for $1\le i\le n_1$, and then integrating by parts w.r.t. $Z'_i\sim\mathcal{N}(0,1)$, it comes

$$\begin{aligned}
&\mathbb{E}\Big[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big(Y'_i-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big)\ln\mathcal{Z}_t\Big] \\
&\qquad= \mathbb{E}\Big[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)Z'_i\ln\mathcal{Z}_t\Big] \\
&\qquad= \mathbb{E}\bigg[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)Z'_i\ln\int dP_0(\mathbf{x})dP_{A_1}(\mathbf{a}_1)\mathcal{D}\mathbf{u}\,e^{-\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})}\bigg] \\
&\qquad= -\mathbb{E}\bigg[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big\langle\sqrt{rt}\Big(\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)-\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\Big)+Z'_i\Big\rangle_t\bigg] \\
&\qquad= -\sqrt{rt}\Big(\rho_1(n_0)-\mathbb{E}\Big\langle\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\Big\rangle_t\Big)\,. \qquad (139)
\end{aligned}$$

Thus, by taking the sum,

$$\begin{aligned}
&-\frac{1}{2}\sqrt{\frac{r}{t}}\mathbb{E}\bigg[\frac{1}{n_0}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big(Y'_i-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\Big)\ln\mathcal{Z}_t\bigg] \\
&\qquad\qquad= \frac{n_1}{n_0}\bigg(\frac{r\rho_1(n_0)}{2}-\frac{r}{2}\mathbb{E}\bigg\langle\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\bigg\rangle_t\bigg)\,. \qquad (140)
\end{aligned}$$

Therefore, for all $t \in (0,1)$,

$$T_1 = \frac{1}{2}\mathbb{E}\Big[\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big) - \rho_1(n_0)\Big)\Big\}\frac{1}{n_0}\ln\mathcal{Z}_t\Big]$$
$$+ \frac{1}{2}\frac{n_1}{n_0}\mathbb{E}\Big[\Big\langle\Big(\frac{1}{n_1}\sum_{\mu=1}^{n_2}u'_{Y_\mu}(S_{t,\mu})u'_{Y_\mu}(s_{t,\mu}) - r\Big)\Big(\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big) - q(t)\Big)\Big\rangle_t\Big]$$
$$+ \frac{n_1}{n_0}\Big(\frac{r\rho_1(n_0)}{2} - \frac{rq(t)}{2}\Big). \quad (141)$$

To obtain (129), it remains to show that T_2 is zero. According to the Nishimori identity (see Proposition 8), one has

$$T_2 = \frac{1}{n_0}\mathbb{E}\big\langle\mathcal{H}'_t(\mathbf{x},\mathbf{u},\mathbf{a}_1;\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2)\big\rangle_t = \frac{1}{n_0}\mathbb{E}\,\mathcal{H}'_t(\mathbf{X}^0,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2)\,. \quad (142)$$

From (132) one obtains

$$\mathbb{E}\,\mathcal{H}'_t(\mathbf{X}^0,\mathbf{A}_1,\mathbf{U};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})$$
$$= -\sum_{\mu=1}^{n_2}\mathbb{E}\Big[\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu})\Big] - \frac{1}{2}\sqrt{\frac{r}{t}}\sum_{i=1}^{n_1}\mathbb{E}\Big[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)Z'_i\Big]$$
$$= -\sum_{\mu=1}^{n_2}\mathbb{E}\Big[\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu})\Big]\,. \quad (143)$$

Performing the same integration by parts than the ones leading to (138), one gets

$$-\sum_{\mu=1}^{n_2}\mathbb{E}\Big[\frac{dS_{t,\mu}}{dt}u'_{Y_\mu}(S_{t,\mu})\Big]$$
$$= \frac{1}{2}\mathbb{E}\Big[\sum_{\mu}^{n_2}\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\Big(\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big] = 0\,. \quad (144)$$

The last equality follows from a computation in the next section, see (147). Combining (142), (143) and (144), one obtains $T_2 = 0$.

B.2 Proof that $A_{\mathbf{n}}(t)$ vanishes uniformly as $n_0 \to \infty$

To get Proposition 2, the last step is to prove that $A_{\mathbf{n}}(t)$ – see definition (130) – vanishes uniformly in $t \in [0,1]$ as $n_0 \to +\infty$, under conditions (H1)-(H2)-(H3). First we show that

$$f_{\mathbf{n}}(t)\mathbb{E}\Big[\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\Big(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big)\Big] = 0\,. \quad (145)$$

Once this is done, we use the fact that $\frac{1}{n_0}\ln\mathcal{Z}_t$ concentrates around $f_{\mathbf{n}}(t)$ to prove that $A_{\mathbf{n}}(t)$ vanishes uniformly.

Start by noticing the simple fact that $\int P''_{\mathrm{out},2}(y|s)dy = 0$ for all $s \in \mathbb{R}$. Consequently, for $\mu \in \{1, \dots, n_2\}$,

$$\mathbb{E}\Big[\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\,\Big|\,\mathbf{X}^0, \mathbf{A}_1, \mathbf{W}_1, \mathbf{S}_t\Big] = \int dY_\mu P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu}) = 0\,. \tag{146}$$

The "tower property" of the conditional expectation then gives

$$\begin{aligned}
&\mathbb{E}\Big[\Big\{\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\}\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big] \\
&= \mathbb{E}\Big[\Big\{\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\}\mathbb{E}\Big[\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\,\Big|\,\mathbf{X}^0, \mathbf{A}_1, \mathbf{W}_1, \mathbf{S}_t\Big]\Big] \\
&= 0\,.
\end{aligned} \tag{147}$$

This implies (145). Using successively (145) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|A_{\mathbf{n}}(t)| &= \Big|\mathbb{E}\Big[\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\} \\
&\qquad \cdot\Big(\frac{\ln \mathcal{Z}_t}{n_0} - f_{\mathbf{n}}(t)\Big)\Big]\Big| \\
&\le \mathbb{E}\Big[\Big(\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\}^2\Big]^{\frac{1}{2}} \\
&\qquad \cdot\mathbb{E}\Big[\Big(\frac{1}{n_0}\ln \mathcal{Z}_t - f_{\mathbf{n}}(t)\Big)^2\Big]^{\frac{1}{2}}.
\end{aligned} \tag{148}$$

Making once more use of the "tower property" of conditional expectation, one obtains

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\}^2\Big] \\
&\qquad = \mathbb{E}\Big[\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big\}^2 \\
&\qquad\qquad \cdot\mathbb{E}\Big[\Big(\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\,\Big|\,\mathbf{X}^0, \mathbf{A}_1\mathbf{W}_1, \mathbf{S}_t\Big]\Big]\,.
\end{aligned} \tag{149}$$

Conditionally on $\mathbf{S}_t$, the random variables $\Big(\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)_{1\le\mu\le n_2}$ are i.i.d. and centered. Therefore

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\,\Big|\,\mathbf{X}^0, \mathbf{A}_1, \mathbf{W}_1, \mathbf{S}_t\Big] &= \mathbb{E}\Big[\Big(\sum_{\mu=1}^{n_2}\frac{P''_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\,\Big|\,\mathbf{S}_t\Big] \\
&= n_2\mathbb{E}\Big[\Big(\frac{P''_{\mathrm{out},2}(Y_1|S_{t,1})}{P_{\mathrm{out},2}(Y_1|S_{t,1})}\Big)^2\,\Big|\,\mathbf{S}_t\Big]\,.
\end{aligned} \tag{150}$$

Under condition (H2), it is not difficult to show that there exists a constant $C > 0$ such that

$$\mathbb{E}\Big[\Big(\frac{P''_{\text{out},2}(Y_1|S_{t,1})}{P_{\text{out},2}(Y_1|S_{t,1})}\Big)^2 \,\Big|\, \mathbf{S}_t\Big] \leq C\,. \tag{151}$$

Combining now (151), (150) and (149) we obtain that

$$\begin{aligned}\mathbb{E}\Big[\Big(\frac{1}{\sqrt{n_1}}\sum_{\mu=1}^{n_2}\frac{P''_{\text{out},2}(Y_\mu|S_{t,\mu})}{P_{\text{out},2}(Y_\mu|S_{t,\mu})}\Big)^2\Big\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)-\rho_1(n_0)\Big)\Big\}^2\Big] \\ \leq \frac{n_2}{n_1}C\,\frac{1}{n_1}\mathbb{E}\Big[\Big\{\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)-\rho_1(n_0)\Big)\Big\}^2\Big]\,. \end{aligned}\tag{152}$$

It remains to prove the boundedness of $\frac{1}{n_1}\mathbb{E}\Big[\Big\{\sum_{i=1}^{n_1}\Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i\Big)-\rho_1(n_0)\Big)\Big\}^2\Big] = \frac{1}{n_1}\mathbb{V}\text{ar}(\overline{\mathbf{X}^1})$, where $\overline{\mathbf{X}^1} := \sum_{i=1}^{n_1} X_i^1$ is the sum of the random variables $X_i^1 := \varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)$, $1 \leq i \leq n_1$. To achieve that, one uses the identity

$$\frac{1}{n_1}\mathbb{V}\text{ar}(\overline{\mathbf{X}^1}) = \frac{1}{n_1}\mathbb{E}\big[\mathbb{V}\text{ar}(\overline{\mathbf{X}^1}|\mathbf{X}^0)\big] + \frac{1}{n_1}\mathbb{V}\text{ar}\big(\mathbb{E}[\overline{\mathbf{X}^1}|\mathbf{X}^0]\big) \tag{153}$$

and show that both terms in the right hand side are bounded.
First, the term $\mathbb{E}\big[\mathbb{V}\text{ar}(\overline{\mathbf{X}^1}|\mathbf{X}^0)\big]$. Conditionally on $\mathbf{X}^0$, the random variables $(X_i^1)_{1\leq i\leq n_1}$ are i.i.d. and

$$\mathbb{V}\text{ar}(\overline{\mathbf{X}^1}|\mathbf{X}^0) = \sum_{i=1}^{n_1}\mathbb{V}\text{ar}(X_i^1|\mathbf{X}^0) = n_1\mathbb{V}\text{ar}(X_1^1|\mathbf{X}^0)\,. \tag{154}$$

It follows that

$$\frac{1}{n_1}\mathbb{E}\big[\mathbb{V}\text{ar}(\overline{\mathbf{X}^1}|\mathbf{X}^0)\big] = \mathbb{E}\big[\mathbb{V}\text{ar}(X_1^1|\mathbf{X}^0)\big] \leq \mathbb{V}\text{ar}(X_1^1) \leq \mathbb{E}\Big[\varphi_1^4\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\Big]\,. \tag{155}$$

Under (H2), the expectation $\mathbb{E}\Big[\varphi_1^4\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\Big]$ is bounded because φ_1 is bounded.
Second, the term $\mathbb{V}\text{ar}\big(\mathbb{E}[\overline{\mathbf{X}^1}|\mathbf{X}^0]\big)$. We have

$$\mathbb{E}[\overline{\mathbf{X}^1}|\mathbf{X}^0] = n_1\cdot\mathbb{E}\Big[\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\Big|\mathbf{X}^0\Big] = n_1\cdot g(X_1^0,\dots,X_{n_0}^0) \tag{156}$$

where $g(x_1,\dots,x_{n_0}) = \mathbb{E}\Big[\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\Big]$. The partial derivatives of g satisfy for $1 \leq j \leq n_0$

$$\begin{aligned}\frac{\partial g}{\partial x_j}(x_1,\dots,x_{n_0}) &= \mathbb{E}\Big[2\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\varphi_1'\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\frac{(\mathbf{W}_1)_{1j}}{\sqrt{n_0}}\Big] \\ &= \frac{2x_j}{n_0}\mathbb{E}\Big[\varphi_1'^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)+\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\varphi_1''\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_1,\mathbf{A}_{1,1}\Big)\Big]\,, \end{aligned}\tag{157}$$

where (157) was obtained by integrating by parts w.r.t. $(\mathbf{W}_1)_{1j}$. Under the hypothesis (H1) the prior P_0 has bounded support $\mathcal{X} \subseteq [-S,S]$. Then, for every $\mathbf{x} \in \mathcal{X}^{n_0}$, we have

$$\frac{\partial g}{\partial x_j}(x_1,\dots,x_n) \leq \frac{2S}{n_0}\cdot(\sup|\varphi_1'|^2 + \sup|\varphi_1|\cdot\sup|\varphi_1''|) \leq \frac{C}{n_0}\,, \tag{158}$$

for some constant $C > 0$. Here the hypothesis (H2) was used to bound the expectation in (157). Thus, the function g satisfies the bounded difference property, i.e. $\forall j \in \{1, \dots, n_0\}$

$$\sup_{\mathbf{x} \in \mathcal{X}^{n_0}, x'_j \in \mathcal{X}} |g(x_1, \dots, x_j, \dots, x_{n_0}) - g(x_1, \dots, x'_j, \dots, x_n)| \leq \frac{C}{n_0} \sup_{x_j, x'_j \in \mathcal{X}} |x_j - x'_j| \leq \frac{2S \cdot C}{n_0} \,. \quad (159)$$

Applying Proposition 11 (see Appendix C.1) it comes

$$\mathbb{V}\mathrm{ar}\big(g(\mathbf{X}^0)\big) \leq \frac{1}{4} \sum_{j=1}^{n_0} \Big(\frac{2S \cdot C}{n_0}\Big)^2 = \frac{C'}{n_0} \,, \quad (160)$$

and

$$\frac{1}{n_1} \mathbb{V}\mathrm{ar}\big(\mathbb{E}[\overline{\mathbf{X}^1} | \mathbf{X}^0]\big) = n_1 \cdot \mathbb{V}\mathrm{ar}\big(g(\mathbf{X}^0)\big) \leq \frac{n_1}{n_0} C' \,. \quad (161)$$

It ends the proof of $\frac{1}{n_1}\mathbb{E}\Big[\Big(\sum_{i=1}^{n_1} \Big(\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1 \mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big) - \rho_1(n_0)\Big)\Big)^2\Big]$ boundedness in the limit $n_0 \to +\infty$. Combining (148), (152) and the latter, it comes

$$|A_{\mathbf{n}}(t)| \leq \; K \, \mathbb{E}\bigg[\bigg(\frac{1}{n_0} \ln \mathcal{Z}_t - f_{\mathbf{n}}(t)\bigg)^2\bigg]^{1/2} \quad (162)$$

for some constant $K > 0$ and n_0 large enough. The uniform convergence of $A_{\mathbf{n}}(t)$ then follows from (162) and Theorem 2 in Appendix C.1, that states $\mathbb{E}[(n_0^{-1} \ln \mathcal{Z}_t - f_{\mathbf{n}}(t))^2] \xrightarrow[n_0 \to +\infty]{} 0$ uniformly in $t \in [0, 1]$.

C Concentration of free entropy and overlaps

C.1 Concentration of the free entropy

In this section, one proves that the free entropy of the interpolation model studied in Sec. 2.4 concentrates around its expectation (uniformly in t), i.e. one proves Theorem 2 stated below. To lighten the notations, one uses $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)$ to denote a generic positive constant depending *only* on φ_1, φ_2, α_1, α_2, S. Remember that S is a bound on the signal absolute values. It is also understood that the dimensions n_0, n_1, n_2 are large enough and $n_1/n_0 \to \alpha_1$, $n_2/n_1 \to \alpha_2$.

Theorem 2. *Under assumptions (H1), (H2), (H3) one can find a positive constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)$ such that*

$$\mathbb{E}\left[\left(\frac{1}{n_0} \ln \mathcal{Z}_t - \mathbb{E}\left[\frac{1}{n_0} \ln \mathcal{Z}_t\right]\right)^2\right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0}. \quad (163)$$

One recalls some setups and notations for the reader's convenience. The interpolating Hamiltonian (97)-(98) is

$$-\sum_{\mu=1}^{n_2} \ln P_{\mathrm{out},2}(Y_\mu | s_{t,\mu}(\mathbf{x}, \mathbf{a}_1, u_\mu)) + \frac{1}{2} \sum_{i=1}^{n_1} \Big(Y'_i - \sqrt{r\,t}\, \varphi_1\Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)\Big)^2 \,, \quad (164)$$

where $s_{t,\mu}(\mathbf{x},\mathbf{a}_1,u_\mu)=\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}},\mathbf{a}_1\Big)\Big]_\mu+k_1(t)\,V_\mu+k_2(t)\,u_\mu$ with

$$k_1(t):=\sqrt{\int_0^t q(v)dv},\quad k_2(t):=\sqrt{\int_0^t(\rho_1(n_0)-q(v))dv}\,.$$

This Hamiltonian follows from the interpolating model

$$\begin{cases}Y_{t,\mu}=\varphi_2\Big(\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}},\mathbf{A}_1\Big)\Big]_\mu+k_1(t)\,V_\mu+k_2(t)\,U_\mu,\mathbf{A}_{2,\mu}\Big)+\sqrt{\Delta}Z_\mu\,, & 1\le\mu\le n_2,\\ Y'_{t,i}\;=\sqrt{r\,t}\,\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)+Z'_i\,, & 1\le i\le n_1\,,\end{cases}$$

where $(\mathbf{A}_{1,i})_{i=1}^{n_1}\overset{\text{i.i.d.}}{\sim}P_{A_1}$, $(\mathbf{A}_{2,\mu})_{\mu=1}^{n_2}\overset{\text{i.i.d.}}{\sim}P_{A_2}$, $(Z_\mu)_{\mu=1}^{n_2},(Z'_i)_{i=1}^{n_1}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$. Recall the definition $\mathbf{X}^1:=\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}},\mathbf{A}_1\Big)$. The channel $P_{\text{out},2}$ defined in (77) can be written as

$$\begin{aligned}P_{\text{out},2}(Y_{t,\mu}|s_{t,\mu}(\mathbf{x},\mathbf{a}_1,\mathbf{u}))&=\int dP_{A_2}(\mathbf{a}_{2,\mu})\frac{1}{\sqrt{2\pi\Delta}}e^{-\frac{1}{2\Delta}\left(Y_{t,\mu}-\varphi_2(s_{t,\mu}(\mathbf{x},\mathbf{a}_1,u_\mu),\mathbf{a}_{2,\mu})\right)^2}\\&=\int dP_{A_2}(\mathbf{a}_{2,\mu})\frac{1}{\sqrt{2\pi\Delta}}e^{-\frac{1}{2\Delta}(\Gamma_{t,\mu}(\mathbf{x},\mathbf{a}_1,\mathbf{a}_{2,\mu},u_\mu)+\sqrt{\Delta}Z_\mu)^2}\end{aligned}\tag{165}$$

with

$$\begin{aligned}\Gamma_{t,\mu}(\mathbf{x},\mathbf{a}_1,\mathbf{a}_{2,\mu},u_\mu):=\varphi_2\Big(&\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\mathbf{X}^1\Big]_\mu+k_1(t)\,V_\mu+k_2(t)\,U_\mu,\mathbf{A}_{2,\mu}\Big)\\&-\varphi_2\Big(\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}},\mathbf{a}_1\Big)\Big]_\mu+k_1(t)\,V_\mu+k_2(t)\,u_\mu,\mathbf{a}_{2,\mu}\Big)\,.\end{aligned}\tag{166}$$

From (164), (165), (166) the free entropy of the interpolating model reads

$$\frac{1}{n_0}\ln\mathcal{Z}_t=\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t-\frac{1}{2n_0}\sum_{\mu=1}^{n_2}Z_\mu^2-\frac{1}{2n_0}\sum_{i=1}^{n}Z_i'^2-\frac{n_2}{2n_0}\ln(2\pi\Delta)\tag{167}$$

where

$$\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t=\frac{1}{n_0}\ln\Big(\int dP_0(\mathbf{x})dP_{A_1}(\mathbf{a}_1)dP_{A_2}(\mathbf{a}_2)\mathcal{D}\mathbf{u}\,e^{-\hat{\mathcal{H}}_t(\mathbf{x},\mathbf{a}_1,\mathbf{a}_2,\mathbf{u})}\Big)\,,\tag{168}$$

and

$$\begin{aligned}\hat{\mathcal{H}}_t(\mathbf{x},\mathbf{a}_1,\mathbf{a}_2,\mathbf{u})=&\frac{1}{2\Delta}\sum_{\mu=1}^{n_2}\Big\{\Gamma_{t,\mu}(\mathbf{x},\mathbf{a}_1,\mathbf{a}_{2,\mu},u_\mu)^2+2\sqrt{\Delta}Z_\mu\Gamma_{t,\mu}(\mathbf{x},\mathbf{a}_1,\mathbf{a}_{2,\mu},u_\mu)\Big\}\\&+\frac{1}{2}\sum_{i=1}^{n_1}\Big[\sqrt{rt}X_i^1-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\Big]^2+2Z'_i\Big[\sqrt{rt}X_i^1-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\Big]\,.\end{aligned}\tag{169}$$

From (166), (168), (169), note that $\ln\hat{\mathcal{Z}}_t/n_0$ has been written as a function of $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$, $\mathbf{W}_1$, $\mathbf{A}_2$, $\mathbf{X}^1$. Our goal is to show that the free energy (167) concentrates around its expectation. It is enough to show that there exists a positive constant $C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$ such that $\mathbb{V}\text{ar}\left(\ln\hat{\mathcal{Z}}_t/n_0\right)\le\frac{C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)}{n_0}$. This concentration property together with (167) implies (163), i.e. Theorem 2.

First, the concentration w.r.t. all Gaussian variables $\mathbf{Z}, \mathbf{Z}', \mathbf{V}, \mathbf{U}, \mathbf{W}_2, \mathbf{W}_1$ is shown thanks to the classical Gaussian Poincaré inequality, then the concentration w.r.t. $\mathbf{A}_2$, $\mathbf{X}^1$ using classical bounded difference arguments. The order in which the concentrations are proved matters. These two variance bounds are recalled below. The reader can refer to [40] (Chapter 3) for detailed proofs of these statements.

Proposition 10 (Gaussian Poincaré inequality)**.** *Let $\mathbf{U} = (U_1, \dots, U_N)$ be a vector of N independent standard normal random variables. Let $g : \mathbb{R}^N \to \mathbb{R}$ be a continuously differentiable function. Then*

$$\mathbb{V}\mathrm{ar}(g(\mathbf{U})) \leq \mathbb{E}\left[\|\nabla g(\mathbf{U})\|^2\right] . \tag{170}$$

Proposition 11. *Let $\mathcal{U} \subset \mathbb{R}$. Let $g : \mathcal{U}^N \to \mathbb{R}$ a function that satisfies the bounded difference property, i.e., there exists some constants $c_1, \dots, c_N \geq 0$ such that*

$$\sup_{\substack{u_1, \dots u_N \in \mathcal{U}^N \\ u_i' \in \mathcal{U}}} |g(u_1, \dots, u_i, \dots, u_N) - g(u_1, \dots, u_i', \dots, u_N)| \leq c_i \,, \text{ for all } 1 \leq i \leq N \,.$$

Let $\boldsymbol{U} = (U_1, \dots, U_N)$ be a vector of N independent random variables that takes values in $\mathcal{U}$. Then

$$\mathbb{V}\mathrm{ar}(g(\mathbf{U})) \leq \frac{1}{4} \sum_{i=1}^{N} c_i^2 \,. \tag{171}$$

Finally, before starting the proof of Theorem 2, we point out that under the hypothesis (H2) all the suprema $\sup|\varphi_k|$, $\sup|\varphi_k'|$, $\sup|\varphi_k''|$ for $k \in \{1, 2\}$ are well-defined, and for all $i \in \{1, \dots, n_1\}$ $|\mathbf{X}_i^1| \leq \sup|\varphi_1|$ almost surely.

C.1.1 Concentration with respect to Gaussian random variables Z, Z′, V, U, W₂, W₁

In this subsection, as in B.1 and to lighten the notations, one will systemically omit the second argument in the functions φ_1, φ_2 and their first and second derivatives w.r.t. to their first argument. Here one proves $\ln \hat{\mathcal{Z}}_t / n_0$ is close to its expectation w.r.t. the Gaussian random variables $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$, $\mathbf{W}_1$, i.e.

Lemma 4. *Let $\mathbb{E}_{G'}$ denotes the expectation w.r.t. $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$, $\mathbf{W}_1$ only. There exists a positive constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)$ such that*

$$\mathbb{E}\left[\left(\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t - \mathbb{E}_{G'}\left[\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t\right]\right)^2\right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0} \,. \tag{172}$$

Lemma 4 follows, by Pythagorean theorem, from the Lemmas 5, 6, 7 proven below.

Lemma 5. *Let $\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}$ denotes the expectation w.r.t. $\mathbf{Z}, \mathbf{Z}'$ only. There exists a constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S) > 0$ such that*

$$\mathbb{E}\left[\left(\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t - \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\left[\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t\right]\right)^2\right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0} \,. \tag{173}$$

Proof. Here $g = {}^{\ln \hat{\mathcal{Z}}_t}/_{n_0}$ is seen as a function of $\mathbf{Z}$, $\mathbf{Z}'$ *only* and we work conditionally to all other random variables. The norm of the gradient of g reads

$$\|\nabla g\|^2 = \sum_{\mu=1}^{n_2} \left| \frac{\partial g}{\partial Z_\mu} \right|^2 + \sum_{i=1}^{n_1} \left| \frac{\partial g}{\partial Z'_i} \right|^2 . \tag{174}$$

Each of these partial derivatives are of the form $\partial g = -n_0^{-1} \langle \partial \hat{\mathcal{H}}_t \rangle_{\hat{\mathcal{H}}_t}$ where the Gibbs bracket $\langle - \rangle_{\hat{\mathcal{H}}_t}$ pertains to the effective Hamiltonian (169). One finds

$$\begin{aligned}
\left| \frac{\partial g}{\partial Z_\mu} \right| &= \frac{1}{n_0 \sqrt{\Delta}} \left| \langle \Gamma_{t,\mu} \rangle_{\hat{\mathcal{H}}_t} \right| \leq \frac{2}{n_0 \sqrt{\Delta}} \sup |\varphi_2| \, , \\
\left| \frac{\partial g}{\partial Z'_i} \right| &= \frac{1}{n_0} \left| \left\langle \sqrt{rt} X_i^1 - \sqrt{rt} \varphi_1 \Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}} \Big]_i \Big) \right\rangle_{\hat{\mathcal{H}}_t} \right| \leq \frac{2\sqrt{r}}{n_0} \sup |\varphi_1| \, ,
\end{aligned}$$

and, replacing in (174), one gets $\|\nabla g\|^2 \leq 4 n_0^{-1} \big(\frac{n_2}{n_0} \Delta^{-1} \sup |\varphi_2|^2 + r \frac{n_1}{n_0} \sup |\varphi_1|^2 \big)$. Applying Proposition 10, one obtains

$$\mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \left[\left(\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t - \mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \left[\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t \right] \right)^2 \right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0} \tag{175}$$

almost surely. Taking the expectation in (175) gives the lemma. □

Lemma 6. *Let $\mathbb{E}_G$ denote the expectation w.r.t. $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$ only. There exists a constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S) > 0$ such that*

$$\mathbb{E} \left[\left(\mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \left[\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t \right] - \mathbb{E}_G \left[\frac{1}{n_0} \ln \hat{\mathcal{Z}}_t \right] \right)^2 \right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0} \, . \tag{176}$$

Proof. Here $g = {}^{\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}[\ln \hat{\mathcal{Z}}_t]}/_{n_0}$ is seen as a function of $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$ and we work conditionally to all other random variables.

$$\begin{aligned}
\left| \frac{\partial g}{\partial V_\mu} \right| &= n_0^{-1} \Big| \mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \Big[\Big\langle (\Gamma_{t,\mu} + \sqrt{\Delta} Z_\mu) \Delta^{-1} \frac{\partial \Gamma_{t,\mu}}{\partial V_\mu} \Big\rangle_{\hat{\mathcal{H}}_t} \Big] \Big| \\
&\leq n_0^{-1} \mathbb{E}_{\mathbf{Z},\mathbf{Z}'} \Big[(2 \sup |\varphi_2| + \sqrt{\Delta} |Z_\mu|) \frac{\sqrt{\rho_1(n_0)}}{\Delta} 2 \sup |\varphi_2'| \Big] \\
&= n_0^{-1} \left(2 \sup |\varphi_2| + \sqrt{\frac{2\Delta}{\pi}} \right) \frac{\sqrt{\rho_1(n_0)}}{\Delta} 2 \sup |\varphi_2'|
\end{aligned}$$

The same inequality holds for $\big| \frac{\partial g}{\partial U_\mu} \big|$. To compute the derivative w.r.t. $(\mathbf{W}_2)_{\mu i}$, first remark that

$$\begin{aligned}
\frac{\partial \Gamma_{t,\mu}}{\partial (\mathbf{W}_2)_{\mu i}} = \sqrt{\frac{1-t}{n_1}} \Big\{ & X_i^1 \varphi_2' \Big(\sqrt{\frac{1-t}{n_1}} \Big[\mathbf{W}_2 \mathbf{X}^1 \Big]_\mu + k_1(t) \, V_\mu + k_2(t) \, U_\mu \Big) \\
& - \varphi_1 \Big(\Big[\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}} \Big]_i \Big) \varphi_2' \Big(\sqrt{\frac{1-t}{n_1}} \Big[\mathbf{W}_2 \varphi_1 \Big(\frac{\mathbf{W}_1 \mathbf{x}}{\sqrt{n_0}} \Big) \Big]_\mu + k_1(t) \, V_\mu + k_2(t) \, u_\mu \Big) \Big\}
\end{aligned}$$

Therefore

$$\begin{aligned}\left|\frac{\partial g}{\partial(\mathbf{W}_2)_{\mu i}}\right| &= n_0^{-1}\Big|\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[\Big\langle(\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{\mu i}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| \\ &\leq \frac{1}{n_0\sqrt{n_1}}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[(2\sup|\varphi_2|+\sqrt{\Delta}|Z_\mu|)\Delta^{-1}(2\sup|\varphi_1|\sup|\varphi_2'|)\Big] \\ &= \frac{1}{n_0\sqrt{n_1}}\left(2\sup|\varphi_2|+\sqrt{\frac{2\Delta}{\pi}}\right)\Delta^{-1}(2\sup|\varphi_1|\sup|\varphi_2'|)\end{aligned}$$

Putting these inequalities together one ends up with

$$\begin{aligned}\|\nabla g\|^2 &= \sum_{\mu=1}^{n_2}\left|\frac{\partial g}{\partial V_\mu}\right|^2+\sum_{\mu=1}^{n_2}\left|\frac{\partial g}{\partial U_\mu}\right|^2+\sum_{\mu=1}^{n_2}\sum_{i=1}^{n_1}\left|\frac{\partial g}{\partial(\mathbf{W}_2)_{\mu i}}\right|^2 \\ &\leq \frac{n_2}{n_0^2}\Big(2\underbrace{\rho_1(n_0)}_{\to\rho_1}+\sup|\varphi_1|^2\Big)\left(\frac{2\sup|\varphi_2'|}{\Delta}\right)^2\left(2\sup|\varphi_2|+\sqrt{\frac{2\Delta}{\pi}}\right)^2.\end{aligned}$$

Then the lemma follows once again of Proposition 10. □

Lemma 7. *Let $\mathbb{E}_{G'}$ denote the expectation w.r.t. $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$, $\mathbf{W}_1$ only. There exists a positive constant $C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$ such that*

$$\mathbb{E}\left[\left(\mathbb{E}_G\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right]-\mathbb{E}_{G'}\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right]\right)^2\right]\leq\frac{C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)}{n_0}. \tag{177}$$

Proof. Here $g=\mathbb{E}_G[\ln\hat{\mathcal{Z}}_t]/n_0$ is seen as a function of $\mathbf{W}_1$ *only* and we work conditionally to the other random variables. The partial derivatives of g w.r.t. $(\mathbf{W}_1)_{ij}$ reads

$$\begin{aligned}\frac{\partial g}{\partial(\mathbf{W}_1)_{ij}} &= -\frac{1}{n_0}\sum_{\mu=1}^{n_2}\mathbb{E}_G\Big[\Big\langle(\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_1)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] \\ &\quad+\frac{\sqrt{rt}}{n_0^{3/2}}\mathbb{E}_G\left[\left\langle\left(\sqrt{rt}X_i^1-\sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)+Z_i'\right)x_j\varphi_1'\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\right\rangle_{\hat{\mathcal{H}}_t}\right]\end{aligned}$$

In a similar fashion to what has be done previously, the absolute value of the second term in this partial derivative can be upperbounded by

$$\frac{\sqrt{r}}{n_0^{3/2}}\left(2\sqrt{r}\sup|\varphi_1|+\sqrt{\frac{2}{\pi}}\right)\cdot S\,\sup|\varphi_1'|\,.$$

The first term requires more work. First notice that

$$\begin{aligned}&\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_1)_{ij}} \\ &\quad=-x_j\sqrt{\frac{1-t}{n_0n_1}}(\mathbf{W}_2)_{\mu i}\varphi_1'\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\varphi_2'\Big(\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big)\Big]_\mu+k_1(t)\,V_\mu+k_2(t)\,u_\mu\Big).\end{aligned}$$

It follows that

$$\mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_1)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] = \Delta^{-1}\sqrt{\frac{1-t}{n_0 n_1}}\mathbb{E}_G\Big[(\mathbf{W}_2)_{\mu i}\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big],$$

where

$$\tilde{\Gamma}^{(ij)}_{t,\mu} = -x_j\varphi_1'\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big)\varphi_2'\Big(\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\varphi_1\Big(\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big)\Big]_\mu + k_1(t)\,V_\mu + k_2(t)\,u_\mu\Big).$$

Integrating by parts w.r.t. $(\mathbf{W}_2)_{\mu i}$ we get

$$\begin{aligned}
&\mathbb{E}_G\Big[(\mathbf{W}_2)_{\mu i}\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] \\
&\qquad = \mathbb{E}_G\Big[\Big\langle \frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{\mu i}}\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] + \mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\frac{\partial\tilde{\Gamma}^{(ij)}_{t,\mu}}{\partial(\mathbf{W}_2)_{\mu i}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] \\
&\qquad\qquad - \mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)^2\,\tilde{\Gamma}^{(ij)}_{t,\mu}\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big] \\
&\qquad\qquad\qquad + \mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big].
\end{aligned}$$

The first two expectations satisfy

$$\begin{aligned}
\Big|\mathbb{E}_G\Big[\Big\langle \frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{\mu i}}\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| &\le \frac{2\sup|\varphi_1|\sup|\varphi_2'|}{\sqrt{n_1}}\cdot S\sup|\varphi_1'|\sup|\varphi_2'|\,, \\
\Big|\mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\frac{\partial\tilde{\Gamma}^{(ij)}_{t,\mu}}{\partial(\mathbf{W}_2)_{\mu i}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| &\le \Big(2\sup|\varphi_2|+\sqrt{\frac{2\Delta}{\pi}}\Big)\cdot\frac{S\sup|\varphi_1'|\sup|\varphi_1|\sup|\varphi_2''|}{\sqrt{n_1}}\,,
\end{aligned}$$

while for the last two we have

$$\begin{aligned}
&\Big|\mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)^2\,\tilde{\Gamma}^{(ij)}_{t,\mu}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| \\
&\qquad \le \mathbb{E}\Big[(2\sup|\varphi_2|+\sqrt{\Delta}|Z_\mu|)^2\Big]\cdot S\sup|\varphi_1'|\sup|\varphi_2'|\cdot\frac{2\sup|\varphi_1|\sup|\varphi_2'|}{\sqrt{n_1}} \\
&\qquad = \Big(4\sup|\varphi_2|^2+\Delta+2\sup|\varphi_2|\sqrt{\frac{2\Delta}{\pi}}\Big)\cdot S\sup|\varphi_1'|\sup|\varphi_2'|\cdot\frac{2\sup|\varphi_1|\sup|\varphi_2'|}{\sqrt{n_1}}\,, \\
&\Big|\mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\tilde{\Gamma}^{(ij)}_{t,\mu}\Big\rangle_{\hat{\mathcal{H}}_t}\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_2)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| \\
&\qquad \le \Big(4\sup|\varphi_2|^2+\Delta+2\sup|\varphi_2|\sqrt{\frac{2\Delta}{\pi}}\Big)\cdot S\sup|\varphi_1'|\sup|\varphi_2'|\cdot\frac{2\sup|\varphi_1|\sup|\varphi_2'|}{\sqrt{n_1}}\,.
\end{aligned}$$

Putting all these inequalities together gives the existence of a positive constant $C_1(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$ such that

$$\Big|n_0^{-1}\sum_{\mu=1}^{n_2}\mathbb{E}_G\Big[\Big\langle (\Gamma_{t,\mu}+\sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial(\mathbf{W}_1)_{ij}}\Big\rangle_{\hat{\mathcal{H}}_t}\Big]\Big| \le \frac{1}{n_0^{3/2}}\cdot\frac{n_2}{n_1}\cdot C_1(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$$

Thus, it exists a positive constant $C_2(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$ satisfying $\left|\frac{\partial g}{\partial(\mathbf{W}_1)_{ij}}\right| \leq {C_2(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)}/{n_0^{3/2}}$ for any $(i,j)\in\{1,\dots,n_1\}\times\{1,\dots,n_0\}$, and

$$\|\nabla g\|^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\left|\frac{\partial g}{\partial(\mathbf{W}_1)_{ij}}\right|^2 \leq \frac{1}{n_0}\cdot\frac{n_1}{n_0}\cdot C_2^2(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)\,.$$

Applying Proposition 10 ends the proof. □

C.1.2 Bounded difference with respect to $\mathbf{A}_{2,\mu}$

Next one applies the variance bound of Lemma 11 to show $n_0^{-1}\mathbb{E}_{G'}[\ln\hat{\mathcal{Z}}_t]$ concentrates w.r.t. $\mathbf{A}_2$, while keeping $\mathbf{X}^1$ fixed for the moment.

Lemma 8. *Let $\mathbb{E}_{G',\mathbf{A}_2}$ denotes the expectation w.r.t. $\mathbf{Z}$, $\mathbf{Z}'$, $\mathbf{V}$, $\mathbf{U}$, $\mathbf{W}_2$, $\mathbf{W}_1$, $\mathbf{A}_2$ only. There exists a positive constant $C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)$ such that*

$$\mathbb{E}\left[\left(\mathbb{E}_{G'}\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right]-\mathbb{E}_{G',\mathbf{A}_2}\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right]\right)^2\right]\leq\frac{C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)}{n_0}\,. \tag{178}$$

Proof. Consider $g={\mathbb{E}_{G'}[\ln\hat{\mathcal{Z}}_t]}/{n_0}$ as a function of $\mathbf{A}_2$ only.
Let $\nu\in\{1,\dots,n_2\}$. One wants to estimate the variation $g(\mathbf{A}_2)-g(\mathbf{A}_2^{(\nu)})$ for two configurations $\mathbf{A}_2$ and $\mathbf{A}_2^{(\nu)}$ with $A_{2,\mu}^{(\nu)}=A_{2,\mu}$ for $\mu\neq\nu$. The notations $\hat{\mathcal{H}}_t^{(\nu)}$ and $\Gamma_{t,\mu}^{(\nu)}$ will denote the quantities $\hat{\mathcal{H}}_t$ and $\Gamma_{t,\mu}$ where $\mathbf{A}_2$ is replaced by $\mathbf{A}_2^{(\nu)}$, respectively. By an application of Jensen's inequality one finds

$$\frac{1}{n_0}\mathbb{E}_{G'}\langle\hat{\mathcal{H}}_t^{(\nu)}-\hat{\mathcal{H}}_t\rangle_{\hat{\mathcal{H}}_t^{(\nu)}}\leq g(\mathbf{A})-g(\mathbf{A}^{(\nu)})\leq\frac{1}{n_0}\mathbb{E}_{G'}\langle\hat{\mathcal{H}}_t^{(\nu)}-\hat{\mathcal{H}}_t\rangle_{\hat{\mathcal{H}}_t} \tag{179}$$

where the Gibbs brackets pertain to the effective Hamiltonians (169). From (169) we obtain

$$\begin{aligned}&\hat{\mathcal{H}}_t^{(\nu)}-\hat{\mathcal{H}}_t\\&\quad=\frac{1}{2\Delta}\sum_{\mu=1}^{n_2}\left(\Gamma_{t,\mu}^{(\nu)2}-\Gamma_{t,\mu}^2+2Z_\mu(\Gamma_{t,\mu}^{(\nu)}-\Gamma_{t,\mu})\right)=\frac{1}{2\Delta}\left(\Gamma_{t,\nu}^{(\nu)2}-\Gamma_{t,\nu}^2+2Z_\nu(\Gamma_{t,\nu}^{(\nu)}-\Gamma_{t,\nu})\right)\,.\end{aligned}$$

Notice that $\left|\Gamma_{t,\nu}^{(\nu)2}-\Gamma_{t,\nu}^2+2Z_\nu(\Gamma_{t,\nu}^{(\nu)}-\Gamma_{t,\nu})\right|\leq 8\sup|\varphi_2|^2+4|Z_\nu|\sup|\varphi_2|$. From (179) we conclude that g satisfies the bounded difference property:

$$|g(\mathbf{A}_2)-g(\mathbf{A}_2^{(\nu)})|\leq\frac{2\sup|\varphi_2|}{\Delta n_0}\left(2\sup|\varphi_2|+\sqrt{\frac{2}{\pi}}\right). \tag{180}$$

Lemma 8 follows then by an application of Proposition 11. □

C.1.3 Bounded difference with respect to X_i^1

One now proves the last lemma needed to get Theorem 2, i.e.

Lemma 9. *There exists a positive constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)$ such that*

$$\mathbb{E}\left[\left(\mathbb{E}_{G',\mathbf{A}_2}\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right] - \mathbb{E}\left[\frac{1}{n_0}\ln\hat{\mathcal{Z}}_t\right]\right)^2\right] \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0}\,. \tag{181}$$

Proof. One sees that $n_0^{-1}\mathbb{E}_{G',\mathbf{A}_2}[\ln\hat{\mathcal{Z}}_t]$ is a function of $\mathbf{X}^1$ only. Consider $g(\mathbf{x}^1) = n_0^{-1}\mathbb{E}[\ln\hat{\mathcal{Z}}_t|\mathbf{X}^1 = \mathbf{x}^1]$. Note that $n_0^{-1}\mathbb{E}_{G',\mathbf{A}_2}[\ln\hat{\mathcal{Z}}_t] = g(\mathbf{X}^1)$. We will show that g satisfies a bounded difference property, then an application of Proposition 11 will end the proof.

Let $i \in \{1, \ldots, n_1\}$ and $\mathbf{x}^1, \mathbf{x}^{(i)} \in [-\sup|\varphi_1|, \sup|\varphi_1|]^{n_1}$ two vectors such that $x_j^{(i)} = x_j^1$ for $j \neq i$. For $s \in [0,1]$ we define $\psi(s) = g(s\mathbf{x}^1 + (1-s)\mathbf{x}^{(i)})$. Hence $\psi(1) = g(\mathbf{x}^1)$ and $\psi(0) = g(\mathbf{x}^{(i)})$. If we can prove that

$$|\psi'(s)| \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0} \quad \forall s \in [0,1]\,, \tag{182}$$

then the bounded difference property follows, namely

$$\sup_{\mathbf{x}^1, \mathbf{x}^{(i)}} |g(\mathbf{x}^1) - g(\mathbf{x}^{(i)})| \leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_0}\,.$$

Let $\tilde{\mathbb{E}}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|\mathbf{X}^1 = s\mathbf{x}^1 + (1-s)\mathbf{x}^{(i)}]$. The derivative of ψ satisfies

$$\begin{aligned}
|\psi'(s)| &= \frac{|x_i^1 - x_i^{(i)}|}{n_0}\tilde{\mathbb{E}}\left[\left\langle \frac{\partial\hat{\mathcal{H}}_t}{\partial X_i^1}\right\rangle_{\hat{\mathcal{H}}_t}\right] \\
&\leq \frac{2\sup|\varphi_1|}{n_0}\sum_{\mu=1}^{n_2}\left|\tilde{\mathbb{E}}\left[\left\langle (\Gamma_{t,\mu} + \sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial X_i^1}\right\rangle_{\hat{\mathcal{H}}_t}\right]\right| \\
&\qquad + \frac{2\sup|\varphi_1|}{n_0}\left|\tilde{\mathbb{E}}\left[\sqrt{rt}\left\langle \sqrt{rt}X_i^1 - \sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big) + Z_i'\right\rangle_{\hat{\mathcal{H}}_t}\right]\right|
\end{aligned}$$

One has

$$\begin{aligned}
\left|\tilde{\mathbb{E}}\left[\sqrt{rt}\left\langle \sqrt{rt}X_i^1 - \sqrt{rt}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i\Big) + Z_i'\right\rangle_{\hat{\mathcal{H}}_t}\right]\right| &\leq \tilde{\mathbb{E}}\left[\sqrt{rt}(\sqrt{rt}\,2\sup|\varphi_1| + |Z_i'|)\right] \\
&\leq rt\,2\sup|\varphi_1| + \sqrt{\frac{2rt}{\pi}}\,,
\end{aligned}$$

while for every $\mu \in \{1, \ldots, n_2\}$ integration by parts w.r.t. $(\mathbf{W}_2)_{\mu i}$ gives

$$\begin{aligned}
&\left|\tilde{\mathbb{E}}\left[\left\langle (\Gamma_{t,\mu} + \sqrt{\Delta}Z_\mu)\Delta^{-1}\frac{\partial\Gamma_{t,\mu}}{\partial X_i^1}\right\rangle_{\hat{\mathcal{H}}_t}\right]\right| \\
&= \left|\tilde{\mathbb{E}}\left[\left\langle (\Gamma_{t,\mu} + \sqrt{\Delta}Z_\mu)\sqrt{\frac{1-t}{n_1}}(\mathbf{W}_2)_{\mu i}\varphi_2'\Big(\sqrt{\frac{1-t}{n_1}}\Big[\mathbf{W}_2\mathbf{X}^1\Big]_\mu + k_1(t)\,V_\mu + k_2(t)\,U_\mu, \mathbf{A}_{2,\mu}\Big)\right\rangle_{\hat{\mathcal{H}}_t}\right]\right| \\
&\leq \frac{C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)}{n_1}
\end{aligned}$$

for some positive constant $C(\varphi_1, \varphi_2, \alpha_1, \alpha_2, S)$. Hence the condition (182) is satisfied, ending the proof of the lemma. □

C.1.4 Proof of Theorem 2

From Lemmas 4, 8, 9 above, one obtains the bound

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right) &= \mathbb{E}\left[\left(\frac{\ln\hat{\mathcal{Z}}_t}{n_0}-\mathbb{E}_{G'}\left[\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}_{G'}\left[\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right]-\mathbb{E}_{G',\mathbf{A}_2}\left[\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right]\right)^2\right]\\ &\quad+\mathbb{E}\left[\left(\mathbb{E}_{G',\mathbf{A}_2}\left[\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right]-\mathbb{E}\left[\frac{\ln\hat{\mathcal{Z}}_t}{n_0}\right]\right)^2\right]\\ &\leq \frac{C(\varphi_1,\varphi_2,\alpha_1,\alpha_2,S)}{n_0}\,,\end{aligned}$$

where the equality of the first line follows simply of the Pythagorean theorem. As mentioned before, this implies Theorem 2 thanks to (167).

C.2 Concentration of the overlap

This section presents the main steps towards proving Lemma 3. The interested reader can find more details in Section V of [38] where the proof method has been streamlined.

One denotes by $\langle-\rangle_{t,\epsilon}$ the Gibbs measure associated to the perturbed Hamiltonian

$$\begin{aligned}\mathcal{H}_t(\mathbf{x},\mathbf{a}_1,\mathbf{u};\mathbf{Y},\mathbf{Y}',\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})+\Bigg\{\sum_{i=1}^{n_1}\frac{\epsilon}{2}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\\ -\epsilon\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)-\sqrt{\epsilon}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\widehat{Z}_i\Bigg\}\end{aligned}$$

i.e., the sum of (98) and (109). As already explained, the addition of the second term can be seen as having an extra Gaussian side-channel, described by (110). Hence the Nishimori identity (Proposition 8) is preserved. The corresponding average free entropy is denoted $f_{\mathbf{n},\epsilon}(t)$ and we call $F_{\mathbf{n},\epsilon}(t)$ the free entropy for a realization of the quenched variables, that is $F_{\mathbf{n},\epsilon}(t)=n_0^{-1}\ln\mathcal{Z}_t(\mathbf{Y},\mathbf{Y}',\widehat{\mathbf{Y}},\mathbf{W}_1,\mathbf{W}_2,\mathbf{V})$. Let

$$\begin{aligned}\mathcal{L}_\epsilon:=\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{2}\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)-\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\\ -\frac{1}{2\sqrt{\epsilon}}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\widehat{Z}_i\,.\end{aligned}$$

Up to the prefactor n_1^{-1} this quantity is the derivative of the perturbation term in (109). The fluctuations of the overlap $\hat{Q}=\frac{1}{n_1}\sum_{i=1}^{n_1}\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i,\mathbf{A}_{1,i}\Big)$ and those of $\mathcal{L}_\epsilon$ are related through the identity

$$\begin{aligned}\mathbb{E}\big\langle(\mathcal{L}_\epsilon-\mathbb{E}\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon})^2\big\rangle_{t,\epsilon}=\frac{1}{4}\mathbb{E}\langle(\hat{Q}-\mathbb{E}\langle\hat{Q}\rangle_{n,t,\epsilon})^2\big\rangle_{t,\epsilon}+\frac{1}{2}\mathbb{E}[\langle\hat{Q}^2\rangle_{t,\epsilon}-\langle\hat{Q}\rangle_{t,\epsilon}^2]\\ +\frac{1}{4n_1^2\epsilon}\sum_{i=1}^{n_1}\mathbb{E}\left[\left\langle\varphi_1^2\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i,\mathbf{a}_{1,i}\Big)\right\rangle_{t,\epsilon}\right].\quad(183)\end{aligned}$$

To see it, first note that $\mathcal{L}_\epsilon = \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{2}(x_i^1)^2 - X_i^1 x_i^1 - \frac{1}{2\sqrt{\epsilon}} x_i^1 \widehat{Z}_i$ where $x_i^1 := \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)$, $X_i^i := \varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)$. Then the full derivation in Appendix IX of [38] can be reproduced exactly by doing the identifications $X_i^1 \leftrightarrow S_i$, $x_i^1 \leftrightarrow X_i$, $n_1 \leftrightarrow n$. Indeed, the proof in Appendix IX of [38] only involves some algebra using the Nishimori identity (here $\mathbf{x}^1$ plays the role of a sample obtained from the conditional distribution $P(\mathbf{X}^1 = \cdot|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V})$) and integration by parts w.r.t. the Gaussian $\hat{Z}_i$ in the perturbation term. Besides, Appendix F.2 of [35] already remarked that the precise form of the first term $\mathcal{H}_t$ does not matter as long as it is a Hamiltonian whose Gibbs distribution satisfies Nishimori identity. To illustrate this, we are going to prove the following lemma, that is used to obtain (183) and is also useful to prove Lemma 2.

Lemma 10 (Formula for $\mathbb{E}\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon}$)**.** *For any $\epsilon > 0$,*

$$\mathbb{E}\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon} = -\frac{1}{2}\mathbb{E}\langle\hat{Q}\rangle_{t,\epsilon} \,. \tag{184}$$

Proof. From $\mathcal{L}_\epsilon$ definition we directly get

$$\mathbb{E}\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon} = \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{2}\mathbb{E}[\langle(x_i^1)^2\rangle_{t,\epsilon}] - \mathbb{E}[X_i^1\langle x_i^1\rangle_{t,\epsilon}] - \frac{1}{2\sqrt{\epsilon}}\mathbb{E}[\langle x_i^1\rangle_{t,\epsilon}\widehat{Z}_i] \,. \tag{185}$$

The expectation $\mathbb{E}[X_i^1\langle x_i^1\rangle_{t,\epsilon}]$ in the sum easily simplifies to

$$\begin{aligned}
&\mathbb{E}\left[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\left\langle\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{x}}{\sqrt{n_0}}\Big]_i, \mathbf{a}_{1,i}\Big)\right\rangle_{t,\epsilon}\right] \\
&\qquad = \mathbb{E}\left[X_i^1\mathbb{E}\left[\varphi_1\Big(\Big[\frac{\mathbf{W}_1\mathbf{X}^0}{\sqrt{n_0}}\Big]_i, \mathbf{A}_{1,i}\Big)\Big|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}\right]\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}\left[X_i^1\mathbb{E}\left[X_i^1\Big|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}\right]\Big|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}\right]\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}\left[X_i^1\Big|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}\right]\cdot\mathbb{E}\left[X_i^1\Big|\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V}\right]\right] \\
&\qquad = \mathbb{E}\left[\langle x_i^1\rangle_{t,\epsilon}^2\right] \,.
\end{aligned} \tag{186}$$

These lines of computation just correspond to the Nishimori identity. We detailed them here to make clear how Nishimory identity still applies to the random variables $\mathbf{x}^1$, $\mathbf{X}^1$.
The third expectation is dealt with an integration by parts w.r.t. $\hat{Z}_i$:

$$\mathbb{E}[\langle x_i^1\rangle_{t,\epsilon}\widehat{Z}_i] = \mathbb{E}\left[\frac{\partial\langle x_i^1\rangle_{t,\epsilon}}{\partial\widehat{Z}_i}\right] = \mathbb{E}\left[\sqrt{\epsilon}(\langle(x_i^1)^2\rangle_{t,\epsilon} - \langle x_i^1\rangle_{t,\epsilon}^2)\right] \,. \tag{187}$$

Combining (185), (186), (187) gives the desired result:

$$\mathbb{E}\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon} = \frac{1}{n_1}\sum_{i=1}^{n_1} -\frac{1}{2}\mathbb{E}[\langle x_i^1\rangle_{t,\epsilon}^2] = -\frac{1}{2}\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}[\langle x_i^1\rangle_{t,\epsilon}X_i^1] \;= -\frac{1}{2}\mathbb{E}\langle\hat{Q}\rangle_{t,\epsilon} \,.$$

□

Lemma 3 is then a direct consequence of the following:

Proposition 12 (Concentration of $\mathcal{L}_\epsilon$ on)**.** *Under assumptions (H1), (H2) and (H3) we have for any* $0 < a < 1$*,*

$$\lim_{n_0 \to +\infty} \int_a^1 d\epsilon \, \mathbb{E}\big\langle (\mathcal{L}_\epsilon - \mathbb{E}\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon})^2 \big\rangle_{t,\epsilon} = 0. \tag{188}$$

As for the one-layer case, the proof of this proposition is broken in two parts. Notice that

$$\mathbb{E}\big\langle (\mathcal{L}_\epsilon - \mathbb{E}\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon})^2 \big\rangle_{t,\epsilon} = \mathbb{E}\big\langle (\mathcal{L}_\epsilon - \langle \mathcal{L}_\epsilon \rangle_{t,\epsilon})^2 \big\rangle_{t,\epsilon} + \mathbb{E}\big[(\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon} - \mathbb{E}\langle \mathcal{L}_\epsilon \rangle_{t,\epsilon})^2 \big]. \tag{189}$$

Thus it suffices to prove the two following lemmas. The first lemma expresses concentration w.r.t. the posterior distribution (or "thermal fluctuations").

Lemma 11 (Concentration of $\mathcal{L}_\epsilon$ on $\langle \mathcal{L}_\epsilon \rangle$ $\mathbb{E}\langle \mathcal{L}_\epsilon \rangle$)**.** *Under assumptions (H1), (H2) and (H3), we have for any* $0 < a < 1$*,*

$$\lim_{n_0 \to +\infty} \int_a^1 d\epsilon \, \mathbb{E}\big[\big\langle (\mathcal{L}_\epsilon - \langle \mathcal{L}_\epsilon \rangle_{t,\epsilon})^2 \big\rangle_{t,\epsilon} \big] = 0 \, . \tag{190}$$

Proof. The result is a consequence of the convexity properties of the free energy and the Nishimori identity. The proof is similar to the one of Lemma 5.2 in Sec. V of [38]. Here we go quickly through those steps just to illustrate the minor changes.
Let $F_{\mathbf{n},t}(\epsilon) = n_0^{-1} \ln \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \widehat{\mathbf{Y}}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{V})$ and $f_{\mathbf{n},t}(\epsilon) = \mathbb{E} F_{\mathbf{n},t}(\epsilon)$. Note that in [38] the authors work with free energies instead of free entropies, i.e. $F_{\mathbf{n},t}$ is defined with a minus sign in front of the logarithm. Here we have:

$$\begin{aligned}
\frac{dF_{\mathbf{n},t}(\epsilon)}{d\epsilon} &= -\frac{n_1}{n_0} \langle \mathcal{L}_\epsilon \rangle \, , \\
\frac{1}{n_0} \frac{d^2 F_{\mathbf{n},t}(\epsilon)}{d\epsilon^2} &= \left(\frac{n_1}{n_0} \right)^2 (\langle \mathcal{L}_\epsilon^2 \rangle - \langle \mathcal{L}_\epsilon \rangle^2) - \frac{1}{4 n_0^2 \epsilon^{3/2}} \sum_{i=1}^{n_1} \langle x_i^1 \rangle \widehat{Z}_i \, , \\
\frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon} &= -\frac{n_1}{n_0} \mathbb{E} \langle \mathcal{L}_\epsilon \rangle = \frac{1}{2 n_0} \sum_{i=1}^{n_1} \mathbb{E}[\langle x_i^1 \rangle^2] \, , \\
\frac{1}{n_0} \frac{d^2 f_{\mathbf{n},t}(\epsilon)}{d\epsilon^2} &= \left(\frac{n_1}{n_0} \right)^2 \mathbb{E}[\langle \mathcal{L}_\epsilon^2 \rangle - \langle \mathcal{L}_\epsilon \rangle^2] - \frac{1}{4 n_0^2 \epsilon} \sum_{i=1}^{n_1} \mathbb{E}[\langle (x_i^1)^2 \rangle - \langle x_i^1 \rangle^2] \, ,
\end{aligned}$$

where we made use of (184), (187) and dropped the indices in $\langle - \rangle_{t,\epsilon}$ to lighten the notations. From the last equation we get

$$\mathbb{E}[\langle \mathcal{L}_\epsilon^2 \rangle - \langle \mathcal{L}_\epsilon \rangle^2] = \frac{n_0}{n_1^2} \frac{d^2 f_{\mathbf{n},t}(\epsilon)}{d\epsilon^2} + \frac{1}{4 n_1^2 \epsilon} \sum_{i=1}^{n_1} \mathbb{E}[\langle (x_i^1)^2 \rangle - \langle x_i^1 \rangle^2] \leq \frac{n_0}{n_1^2} \frac{d^2 f_{\mathbf{n},t}(\epsilon)}{d\epsilon^2} + \frac{\sup \varphi_1^2}{4 n_1 \epsilon}$$

where the last line follows from $\mathbb{E}[\langle (x_i^1)^2 \rangle] \leq \sup \varphi_1^2$. An integration over $\epsilon \in [a, 1]$ gives

$$\begin{aligned}
&\int_a^1 d\epsilon \mathbb{E}[\langle \mathcal{L}_\epsilon^2 \rangle - \langle \mathcal{L}_\epsilon \rangle^2] \\
&\qquad \leq \frac{n_0}{n_1^2} \frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon} \Big|_{\epsilon=1} - \frac{n_0}{n_1^2} \frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon} \Big|_{\epsilon=a} + \frac{\sup \varphi_1^2}{4 n_1} |\ln a| \leq \frac{n_0}{n_1^2} \frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon} \Big|_{\epsilon=1} + \frac{\sup \varphi_1^2}{4 n_1} |\ln a| \, ,
\end{aligned}$$

where to obtain the last inequality we used that the derivative $\frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon}$ is non-negative. Finally

$$\frac{n_0}{n_1^2}\frac{df_{\mathbf{n},t}(\epsilon)}{d\epsilon}\Big|_{\epsilon=1} = \frac{1}{2n_1^2}\sum_{i=1}^{n_1}\mathbb{E}[\langle x_i^1\rangle^2] \leq \frac{\sup\varphi_1^2}{2n_1}$$

leads to

$$\int_a^1 d\epsilon\mathbb{E}[\langle\mathcal{L}_\epsilon^2\rangle - \langle\mathcal{L}_\epsilon\rangle^2] \leq \frac{\sup\varphi_1^2}{2n_1}\left(1+\frac{|\ln a|}{2}\right) .$$

□

The second lemma expresses the concentration of the average overlap w.r.t. the realizations of quenched disorder variables.

Lemma 12 (Concentration of $\langle\mathcal{L}_\epsilon\rangle$ on $\mathbb{E}\langle\mathcal{L}_\epsilon\rangle$)**.** *Under assumptions (H1), (H2) and (H3), we have for any* $0 < a < 1$*,*

$$\lim_{n_0\to+\infty}\int_a^1 d\epsilon\,\mathbb{E}\big[(\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon} - \mathbb{E}[\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon}])^2\big] = 0 . \tag{191}$$

Proof. It is a consequence of the concentration of the free energy (see Theorem 2 in Appendix C.1). The proof is similar to the one of Lemma 5.3 in Sec. V of [38], the main change being in the definition of the functions $\tilde{F}(\epsilon)$, $\tilde{f}(\epsilon)$:

$$\tilde{F}(\epsilon) := F_{\mathbf{n},t}(\epsilon) - \frac{\sqrt{\epsilon}}{n_0}\sum_{i=1}^{n_1}(\sup|\varphi_1|)\cdot|\widehat{Z}_i| \ , \ \tilde{f}(\epsilon) := f_{\mathbf{n},t}(\epsilon) - \frac{\sqrt{\epsilon}}{n_0}\sum_{i=1}^{n_1}(\sup|\varphi_1|)\cdot\mathbb{E}|\widehat{Z}_i| .$$

The addition of the second term makes $\tilde{F}(\epsilon)$ convex, while $\tilde{f}(\epsilon)$ is convex too (note that $f_{\mathbf{n},t}(\epsilon)$ was already convex, as it can be shown by the same method than the one in Sec. V of [38]). The proof of Lemma 5.3 in [38] can then be reproduced and choosing $\delta = n_0^{-1/4}$ leads to the bound

$$\int_a^1 d\epsilon\,\mathbb{E}\big[(\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon} - \mathbb{E}[\langle\mathcal{L}_\epsilon\rangle_{t,\epsilon}])^2\big] \leq \frac{C}{n_0^{1/4}}\cdot(\sup|\varphi_1|)^2\cdot\left(1+\sqrt{\frac{2}{\pi\cdot a}}\right) ,$$

for some positive constant C and n_0 large enough. □

References

[1] D. Sherrington and S. Kirkpatrick. Solvable Model of a Spin-Glass. *Physical Review Letters*, 35(26):1792, 1975.

[2] M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond.* World Scientific Publishing Company, 1987.

[3] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction.* Clarendon Press, 2001.

[4] M. Mézard and A. Montanari. *Information, Physics, and Computation.* Oxford University Press, 2009.

[5] E. Gardner. The space of interactions in neural network models. *Journal of Physics A*, 21(1):257, 1988.

[6] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A*, 21(1):271, 1988.

[7] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A*, 22(12):2181, 1989.

[8] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056, 1992.

[9] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[10] T. Tanaka. A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910, 2002.

[11] S. Rangan, V. Goyal, and A. K. Fletcher. Asymptotic analysis of MAP estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[12] Y. Kabashima, T. Wadayama, and T. Tanaka. A typical reconstruction limit for compressed sensing based on Lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

[13] S. Ganguli and H. Sompolinsky. Statistical Mechanics of Compressed Sensing. *Physical Review Letters*, 104(18):188701, 2010.

[14] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[15] L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[16] E. Marinari, G. Parisi, and F. Ritort. Replica field theory for deterministic models. II. A non-random spin glass with glassy behaviour. *Journal of Physics A*, 27(23):7647, 1994.

[17] G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A*, 28(18):5267, 1995.

[18] M. Opper and O. Winther. Tractable Approximations for Probabilistic Models: The Adaptive Thouless-Anderson-Palmer Mean Field Approach. *Physical Review Letters*, 86(17):3695, 2001.

[19] R. Cherrier, D. S. Dean, and A. Lefèvre. Role of the interaction matrix in mean-field spin glass models. *Physical Review E*, 67(4):046112, 2003.

[20] K. Takeda, S. Uda, and Y. Kabashima. Analysis of CDMA systems that are characterized by eigenvalue spectrum. *Europhysics Letters*, 76(6):1193, 2006.

[21] R. R. Müller, D. Guo, and A. L. Moustakas. Vector Precoding for Wireless MIMO Systems and its Replica Analysis. *IEEE Journal on Selected Areas in Communications*, 26(3):530–540, 2008.

[22] Y. Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series*, 95(1):012001, 2008.

[23] T. Shinzato and Y. Kabashima. Perceptron capacity revisited: classification ability for correlated patterns. *Journal of Physics A*, 41(32):324013, 2008.

[24] T. Shinzato and Y. Kabashima. Learning from correlated patterns by simple perceptrons. *Journal of Physics A*, 42(1):015005, 2009.

[25] A. M. Tulino, G. Caire, S. Verdú, and S. Shamai (Shitz). Support Recovery With Sparsely Sampled Free Random Matrices. *IEEE Transactions on Information Theory*, 59(7):4243–4271, 2013.

[26] Y. Kabashima and M. Vehkaperä. Signal recovery using expectation consistent approximation for linear observations. In *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[27] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová. Multi-layer generalized linear estimation. In *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[28] A. K. Fletcher and S. Rangan. Inference in Deep Networks in High Dimensions. *arXiv:1706.06549*, 2017.

[29] G. Reeves. Additivity of Information in Multilayer Networks via Additive Gaussian Noise Transforms. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[30] M. Talagrand. *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models*. Springer Science & Business Media, 2003.

[31] D. Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[32] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová. Mutual Information for Symmetric Rank-one Matrix Estimation: A Proof of the Replica Formula. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[33] M. Lelarge and L. Miolane. Fundamental limits of symmetric low-rank matrix estimation. *arXiv:1611.03888*, 2016.

[34] G. Reeves and H. D. Pfister. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *IEEE International Symposium on Information Theory (ISIT)*, 2016.

[35] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models. *arXiv:1708.03395*, 2017.

[36] A. M. Tulino and S. Verdú. *Random Matrix Theory and Wireless Communications*. Now Publishers Inc, 2004.

[37] J. Barbier, N. Macris, A. Maillard, and F. Krzakala. The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices. In *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[38] J. Barbier and N. Macris. The stochastic interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *arXiv:1705.02780*, 2017.

[39] R. Shwartz-Ziv and N. Tishby. Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810*, 2017.

[40] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.