

Microscopic and collective signatures of feature learning in neural networks

A. Corti,^{1,2} R. Pacelli,³ P. Rotondo,⁴ and M. Gherardi^{1,2,*}

¹*Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*

²*I.N.F.N. sezione di Milano, Via Celoria 16, 20133 Milano, Italy*

³*I.N.F.N. sezione di Padova, Via Marzolo 8, 35131, Padova, Italy*

⁴*Dipartimento di Scienze Matematiche, Fisiche e Informatiche,
Università degli Studi di Parma, Parco Area delle Scienze, 7/A 43124 Parma, Italy*

Feature extraction — the ability to identify relevant properties of data — is a key factor underlying the success of deep learning. Yet, it has proved difficult to elucidate its nature within existing predictive theories, to the extent that there is no consensus on the very definition of feature learning. A promising hint in this direction comes from previous phenomenological observations of quasi-universal aspects in the training dynamics of neural networks, displayed by simple properties of feature geometry. We address this problem within a statistical-mechanics framework for Bayesian learning in one hidden layer neural networks with standard parameterization. Analytical computations in the proportional limit (when both the network width and the size of the training set are large) can quantify fingerprints of feature learning, both collective ones (related to manifold geometry) and microscopic ones (related to the weights). In particular, (i) the distance between different class manifolds in feature space is a nonmonotonic function of the temperature, which we interpret as the equilibrium counterpart of a phenomenon observed under gradient descent (GD) dynamics, and (ii) the microscopic learnable parameters in the network undergo a finite data-dependent displacement with respect to the infinite-width limit, and develop correlations. These results indicate that nontrivial feature learning is at play in a regime where the posterior predictive distribution is that of Gaussian process regression with a trivially rescaled prior.

Introduction—The empirical success of deep learning is fundamentally linked to the ability of neural networks (NNs) to extract meaningful features from raw data [1–3]. Nevertheless, the mechanistic definition of such process, referred to as *feature learning*, remains debated. Even for simple models, such as standard-scaled fully-connected (FC) NNs, there is no agreed-upon set of observables (properties) of their trainable parameters that definitively describe feature learning [4–8].

Recent progress on the theoretical analysis of NNs has established an equivalence between standard-scaled Bayesian FC NNs and kernel regressions, particularly in two limits. (i) In the infinite-width framework, where the width of the hidden layers N_1 is much larger than the number of training examples P , standard-scaled NNs are known to be mathematically equivalent to fixed Gaussian Processes (GPs), which only depend on the *prior* statistics of the weights [9–17]. In this case, training produces an infinitesimal displacement of both the network’s weights and the hidden-layer features, which is enough to fit the dataset examples. (ii) In the less overparameterized proportional regime, where $P/N_1 = \alpha > 0$, one lacks such formal equivalence, yet GPs still come to hand when analysing predictive performance. Specifically, the average generalisation error of the NN is found to be the same as that of a GP regression with a set of free hyperparameters that are fine-tuned to the task at hand. Differently from the infinite-width limit, this GP is not fixed at initialisation, and depends on the *posterior* statistics of the weights [18–27]. (This regime allows for the analysis of continual learning and transfer learning as well [28, 29].)

These descriptions in terms of GPs suggest that looking at the predictive distribution may not be enough to characterise feature learning in FC models. In fact, consistent experimental evidence points to the fact that finite, yet overparametrized, FC networks are eventually outperformed by a suitable GP [30] (at least in most computer vision tasks), prompting the identification of a different set of observables to describe feature learning in this context.

A key insight emerges from recent empirical studies of feature dynamics in FC NNs. Experiments show a robust behaviour that occurs during training: collective observables of the features are nonmonotonic in training time, with a quasi-universal inversion point that is consequential for good generalization properties [31, 32]. More precisely, for binary classification problems, the class manifolds (the two sets containing the representations in feature space of data with the same label [33–37]) initially become well separated, and then approach each other when the network is learning the most “challenging” data samples. These results, in addition to pre-existing numerical evidence [38, 39], indicate that the geometry of class manifolds in feature space is intimately linked to feature extraction.

In this manuscript, we investigate analytically both collective and microscopic equilibrium observables linked to hidden-layer features in Bayesian one-hidden-layer (1HL) FC NNs. In the proportional regime, we compute the posterior average squared distance $\langle D^2 \rangle$ between class manifolds in a binary classification problem, and analyse the second order statistics of the hidden

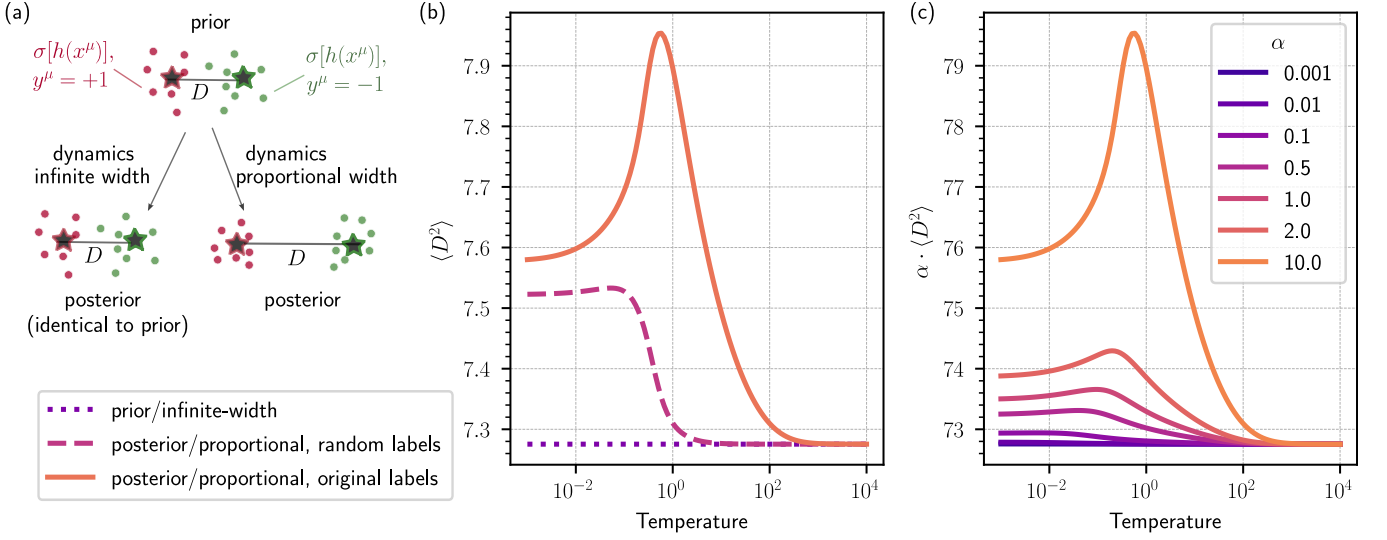


Figure 1. **(a) Separation of class manifolds in the lazy training infinite-width vs proportional regime.** In the infinite-width limit, the separation between class manifolds D^2 is unaffected by training. The weight displacement that occurs in this setting is not enough to produce changes in the collective observable D^2 . In the proportional regime, D^2 undergoes a finite shift, due to the microscopic weight displacements. **(b,c) The nonmonotonicity of the distance is a signature of feature learning at proportional width.** The average squared distance (y axis), Eq. (4), is a nonmonotonic function of the temperature T (x axis). The peak is not present for randomized labels (dashed line) or in the temperature-independent prior (dotted line). **(c) Moving away from the infinite-width limit (larger load α) makes the peak more prominent.** The distance (y axis) is rescaled by α so that the infinite-width limit is finite. All curves are computed using the first 1000 CIFAR10 samples, split into “airplane” and “automobile” classes, with activation function $\sigma = \text{erf}$.

layer weights. Our results can be summarised in three points. (i) The posterior distance remains unaltered in the infinite-width limit, while it departs from its prior value at proportional width (see sketch in panel (a) of Fig. 1). (ii) The observable $\langle D^2 \rangle$ is a non-monotonic function of the Gibbs temperature T . The nonmonotonicity disappears in the infinite-width limit, or when information about the features is removed from the data by randomising the labels. (iii) The hidden-layer weights become correlated in the proportional regime, undergoing a finite displacement, differently from the infinite-width limit, where the parameter statistics is known to be unaffected by training.

Setting of the learning problem— We consider a training set $\mathcal{D} = \{x^\mu, y^\mu\}_{\mu=1}^P$, with $x^\mu \in \mathbb{R}^{N_0}$ and $y^\mu \in \mathbb{R}$, and a 1HL FC network $f(x) = v \cdot \sigma[h(x)]/\sqrt{N_1}$, where the pre-activations read $h(x) = wx/\sqrt{N_0}$, and σ is a pointwise nonlinear activation function. The matrix $w \in \mathbb{R}^{N_1 \times N_0}$ and vector $v \in \mathbb{R}^{N_1}$ contain the microscopic degrees of freedom of the system. We refer to them collectively as $\theta = \{v, w\}$. Multiplying the network’s output by $1/\sqrt{N_1}$ corresponds to the standard scaling [4]. For the sake of simplicity, we restrict our analysis to zero-mean activation functions (e.g., erf, tanh). The statistics of v and w in the Bayesian setting are determined by their priors, which we take to be rescaled normals, and by a likelihood function, for which we use the mean-squared error loss. The properties of the Bayesian network are

then determined by the partition function $\mathcal{Z}(X, y) = \int d\mu(\theta) \exp(-\beta \|f_\theta(X) - y\|^2/2)$. The shorthand notation $f_\theta(X) = (f_\theta(x^\mu), x^\mu \in \mathcal{D})$ indicates the collection of the network’s outputs to the training samples, while the measure $d\mu(\theta)$ indicates integration over the prior on the weights $w \sim \mathcal{N}(\mathbf{0}, \mathbb{1}/\lambda_0)$, $v \sim \mathcal{N}(\mathbf{0}, \mathbb{1}/\lambda_1)$. As shown in [18, 19, 25] for linear activation functions and in [20] for general activations, the partition function can be evaluated analytically in the proportional-width limit, where $N_1, P \rightarrow \infty$ with $P/N_1 = \alpha$ fixed. The solution is obtained via a saddle-point evaluation of an integral, which corresponds to optimizing a scalar-dependent effective action $S(Q)$ through the condition $S'(\bar{Q}) = 0$.

Importantly, the optimization depends on the training dataset (besides the temperature $T = 1/\beta$), so that \bar{Q} contains information on both the input data and their corresponding labels. In the following, we will compute average observables over the posterior Gibbs distribution $P(\theta|X, y)$ associated to \mathcal{Z} , denoted with $\langle \star \rangle = \langle \star \rangle_{P(\theta|X, y)}$.

Collective displacement of the features— As shown in [24], within this setting it is possible to compute the posterior statistics of the features, which are collective variables of the first layer weights: $\sigma(h_j^\mu) = \sigma(\sum_{i_0} w_{ji_0} x_{i_0}^\mu / \sqrt{N_0})$. This is captured by the average

similarity matrix, which reads [24]:

$$\begin{aligned} \langle \sigma(h_i^\mu) \sigma(h_i^\nu) \rangle &= K_{\mu\nu} - \underbrace{\frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda, \delta=1}^P K_{\mu\lambda} K_{\nu\delta} (\tilde{K}_{(R)}^{-1})_{\lambda\delta}}_{\Delta_1^{\mu\nu}} + \\ &+ \underbrace{\frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda, \delta=1}^P K_{\mu\lambda} K_{\nu\delta} (\tilde{K}_{(R)}^{-1} y)_\lambda (\tilde{K}_{(R)}^{-1} y)_\delta}_{\Delta_2^{\mu\nu}}, \end{aligned} \quad (1)$$

The NNGP kernel $K_{\mu\nu}$ is defined as the averaged similarity matrix of the features over the prior distribution, $K_{\mu\nu} = \langle \sigma(h^\mu) \sigma(h^\nu) \rangle_{P(h|X)}$, and we have defined $\tilde{K}_{(R)} = \mathbb{1}/\beta + K_{(R)}$, where $K_{(R)} = \bar{Q}K/\lambda_1$ is the renormalized kernel. Both in the proportional and in the infinite-width limit, the posterior corrections $\Delta_1^{\mu\nu}$, $\Delta_2^{\mu\nu}$ are of order $\mathcal{O}(1/N_1)$, thus irrelevant whenever $N_1 \rightarrow \infty$. While this fact is evident in the infinite-width limit, where the sums involve a finite number of terms P , it requires further explanation when $P \rightarrow \infty$. Indeed, if $\beta \rightarrow \infty$, then $\tilde{K}_{(R)}^{-1}$ is equal to $\lambda_1 K^{-1}/\bar{Q}$. As a result, the first sum gives $\Delta_1^{\mu\nu} = K_{\mu\nu}/N_1$, and the second gives $\Delta_2^{\mu\nu} = (\lambda_1/\bar{Q}) y_\mu y_\nu / N_1$. On the other hand, if $\beta \rightarrow 0$, we have $\tilde{K}_{(R)}^{-1} \sim \beta \mathbb{1} \rightarrow 0$, which makes both the terms in the sum vanish. For all intermediate values of β , the system continuously interpolates between these two regimes and the two terms will remain of order $\mathcal{O}(1/N_1)$.

These considerations point to the fact that the second-order statistics of the features in the proportional-width limit is the same as in the infinite-width limit, where no feature learning happens [9, 40]. We now show that it is still possible to define observables that turn the subleading terms Δ_1 and Δ_2 into finite corrections in the proportional limit. While the previous results hold for a regression task with general training labels, in what follows we consider binary labels $y^\mu = \pm 1$, and training datasets where the data points are equally split into the two different classes of cardinality $P/2$ (the generalization to unbalanced datasets [41, 42] and other pairs of label values is shown in the Supplementary Material (SM)).

The squared distance D^2 , defined as the separation between the mean post-activations of the two different classes, has been used in [31] as a measure of class manifold separation between features:

$$D^2 = \left\| \frac{2}{P} \sum_{\mu=1}^{P/2} \sigma(\mathbf{h}_+^\mu) - \frac{2}{P} \sum_{\mu=1}^{P/2} \sigma(\mathbf{h}_-^\mu) \right\|^2, \quad (2)$$

where \mathbf{h}_\pm^μ is the pre-activation corresponding to the μ -th sample of the first and second class, respectively. Aver-

aging over the posterior, we obtain:

$$\begin{aligned} \langle D^2 \rangle &= \frac{4N_1}{P^2} \sum_{\mu, \nu=1}^P y_\mu y_\nu \langle \sigma(h_i^\mu) \sigma(h_i^\nu) \rangle = \\ &= \frac{4}{\alpha P} y^T K y - \frac{4}{\alpha P} y^T \Delta_1 y + \frac{4}{\alpha P} y^T \Delta_2 y, \end{aligned} \quad (3)$$

with Δ_1 and Δ_2 as in Eq. (1). It is important to emphasize that the y 's appearing explicitly in this formula do not originate from the posterior distribution, but rather from the relative signs in the definition (2). The labels stemming from the posterior average are instead embedded within Δ_1 and Δ_2 . In fact, averaging over the prior yields the first term alone, $\langle D^2 \rangle_{P(h|X)} = 4y^T K t / (\alpha P)$.

When $P \rightarrow \infty$, the scaling of the three terms in Eq. (3) is nontrivial, but it can be discussed in the zero-temperature limit $\beta \rightarrow \infty$, where the expressions simplify. In this case, the second term reduces to $4y^T K y / (\alpha P N_1)$, while the third simplifies to $\frac{4\lambda_1}{\bar{Q}}$. Thus, assuming that $y^T K y / P = \mathcal{O}(1)$, the second term vanishes, and the last term provides a finite, feature-dependent correction to the distance averaged over the prior. In general, the assumption is numerically satisfied in realistic regimes for computer vision tasks where the dataset is noisy enough (e.g., CIFAR10, see SM). In such cases, the large but finite-size regime, where the theory is applicable [26], creates a scenario in which the overall squared distance is primarily determined by the first and third terms. This numerical observation is an exact statement for iid random data: in this case, the components of K are independent random variables.

According to the central limit theorem, $\sum_\nu K_{\mu\nu} y_\nu \sim \sqrt{P}$ and the sum over two indices is $\sim P$. In the SM, we provide a numerical study of this scaling and more details about the typical situations where it does not hold. At finite temperature, disregarding the subleading (or effectively small) second term, the squared distance reads:

$$\langle D^2 \rangle \simeq \frac{4}{\alpha} \frac{1}{P} \sum_{\mu, \nu=1}^P y_\mu K_{\mu\nu} y_\nu + \frac{\lambda_1}{\bar{Q}} (\bar{f}_+ - \bar{f}_-)^2, \quad (4)$$

$$\bar{f}_\pm = \frac{2}{P} \sum_{\mu=1}^{P/2} \langle f_\pm^\mu \rangle, \quad \langle f_\pm^\mu \rangle = \sum_{\lambda, \sigma=1}^P (K_{(R)})_{\mu\lambda} (\tilde{K}_{(R)}^{-1})_{\lambda\sigma} y_\sigma. \quad (5)$$

where $\langle f_\pm^\mu \rangle$ are the posterior expected outputs of the NN in the proportional limit [20, 21]. The symbols \pm indicate that the index μ runs over the kernel evaluated on samples from the first and second class, respectively.

Note that the squared distance diverges in the infinite-width limit $\alpha \rightarrow 0$, since the post activations $\sigma(\mathbf{h})$ are N_1 -dimensional vectors, and their squared norms are $\mathcal{O}(N_1)$. We plot $\langle D^2 \rangle$ as a function of the temperature in Fig. 1(b) (and $\alpha \langle D^2 \rangle$ in Fig. 1(c), which is rescaled so that it converges in the infinite-width limit). The dependence of the squared distance on the temperature is

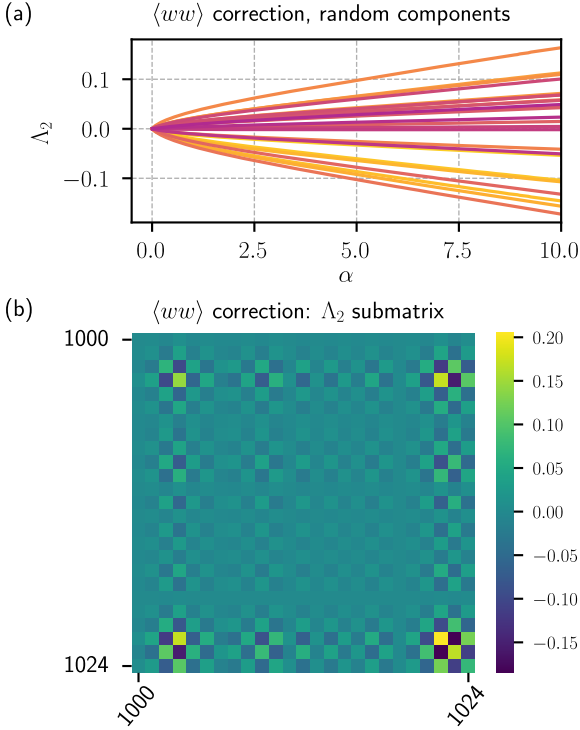


Figure 2. **Finite corrections to $\langle ww \rangle$ in the proportional regime.** (a) Values of 25 randomly selected components of the proportional correction Λ_2 , Eq. (8), as a function of α . As α increases, transitioning the system to the proportional regime, each component exhibits a linear trend, consistent with the scaling predicted by GD dynamics. (b) The heatmap shows a submatrix of Λ_2 at $\alpha = 2$. The magnitude of the matrix elements is not negligible compared to the value $\langle ww \rangle_{ij} = \delta_{ij}/\lambda_0$ expected at infinite width (here, $\lambda_0 = 1$). Simulations were performed with 1500 samples of CIFAR10 dataset, split into even and odd classes. We used a linear activation function $\sigma = \text{Id}$ and $T = 10^{-3}$.

nonmonotonic for $\alpha > 0$, with a characteristic peak at positive temperature. (An alternative way to normalize the distance is presented in the SM, with similar results as here.) The emergence of the term $(\bar{f}_+ - \bar{f}_-)^2 \lambda_1 / \bar{Q}$ in the proportional limit signals the network’s ability to learn the features by exploiting contributions from the posterior, which influences their separation. Thus, the strongly nonmonotonic behavior of $\langle D^2 \rangle$ is understood as a consequence of feature learning. This is confirmed by the fact that assigning random labels to the data nearly removes the peak (dashed line in Fig. 1(b)). Furthermore, reducing α towards the infinite-width limit reduces the relative contribution of the posterior with respect to the prior, while simultaneously eliminating the peak (Fig. 1(c)). This is in accordance with the fact that the posterior statistics becomes indistinguishable from that of the prior in the infinite-width limit.

Microscopic pairwise correlations of the weights— In the previous section, we isolated a *macroscopic* observable that carries information about feature learning at

proportional width. Here, we investigate *microscopic* features, showing that the statistics of the first-layer weights w depends on the data (which does not happen in the infinite-width limit). We compute the two-point function $\langle w_{1k} w_{1h} \rangle$ with respect to the posterior distribution. In the SM, we show that it is possible to use the effective action formalism to express the similarity matrix $\langle w_{1h} w_{1k} \rangle$ in terms of the order parameter \bar{Q} :

$$\langle w_{1h} w_{1k} \rangle = \frac{\delta_{hk}}{\lambda_0} + \frac{(\Lambda_1)_{hk}}{\lambda_0} + \frac{(\Lambda_2)_{hk}}{\lambda_0}, \quad (6)$$

$$(\Lambda_1)_{hk} = \frac{\alpha \bar{Q}}{2\lambda_1} \frac{1}{P} \sum_{\mu, \nu=1}^P (\tilde{K}_{(R)}^{-1})_{\mu\nu} \Delta_{hk}^{\mu\nu}, \quad (7)$$

$$(\Lambda_2)_{hk} = \frac{\alpha \bar{Q}}{2\lambda_1} \frac{1}{P} \sum_{\mu, \nu=1}^P (\tilde{K}_{(R)}^{-1})_{\mu} (\Delta_{hk}^{\mu\nu}) (\tilde{K}_{(R)}^{-1})_{\nu}. \quad (8)$$

$\Delta_{hk}^{\mu\nu}$ is a data-dependent kernel defined in the SM. In the case of a linear activation function ($\sigma = \text{Id}$), these results hold exactly for finite N_1 and P . This fact is in line with the use of the effective action formalism carried out in [25] for finite-width deep linear networks.

In the infinite-width limit, the terms Λ_1 and Λ_2 vanish, the components of w become uncorrelated, and averages over the prior and over the posterior coincide. In the proportional case, since the kernel $\Delta_{hk}^{\mu\nu}$ cannot be expressed in simple terms by using $K_{\mu\nu}$, the zero-temperature limit is not useful to investigate the magnitude of the proportional correction terms. In Fig. 2, we show results of numerical simulations performed with a linear activation function. In this case, the kernel is easily expressed through the data as $\Delta_{hk}^{\mu\nu} = -(x_h^\mu x_k^\nu + x_h^\nu x_k^\mu) / (\lambda_0 N_0)$. We note that, while the first correction Λ_1 is negligible, the second Λ_2 contributes to the statistics of the weights with a finite term that depends almost linearly on α . The linearity is consistent with a scaling argument that can be invoked also in the context of GD dynamics: in this case, the time derivative of the weights is proportional to $\sqrt{\alpha}$.

Discussion and conclusions— The squared distance D^2 is a collective observable of the features that signals non-trivial feature learning in Bayesian FC shallow networks. Other measures of feature segregation have been found to display interesting behavior empirically [7, 43, 44]; the Bayesian proportional-width framework employed here should be able to capture those as well.

Our computations show that the posterior average $\langle D^2 \rangle$ is a nonmonotonic function of the temperature T . The peak at the inversion point is more pronounced for larger values of α , and eventually disappears in the infinite-width limit $\alpha = 0$. In the Bayesian setting, the temperature plays the role of a regularizer, acting as early stopping in the optimization dynamics [45]. The behavior of the distance then aligns with the inversion dynamics observed in [31]. In that work, the nonmonotonic trend observed during training is interpreted as a trade-off between the segregation of the two class manifolds and the

fine tuning required for classifying the last hard samples. In this sense, feature learning manifests itself, during the optimization dynamics, through a well-defined transition between an easy and a hard phase of training. The results summarized in Fig. 1 can be seen as the equilibrium counterpart to this non-equilibrium phenomenon. A possible interpretation arises from the observation that increasing the temperature allows the posterior to sample regions of the loss landscape further away from the optima. The temperature at the inversion peak then corresponds to the typical loss values associated with the transition between easy and hard samples.

The (posterior) second-order statistics of the hidden-layer weights develop finite correlations in the proportional limit. This is remarkable. The Gaussian process of the output receives a trivial modification with respect to the infinite-width limit, because the scalar renormalization by \bar{Q} can be reabsorbed in the prior parameter λ_1 [24]. In contrast, the second-order statistics of w depends on the input patterns in a way that cannot be traced back to the infinite-width limit.

Finally, we would like to point out two distinctions between our work and other lines of research in this field. First, our results hold for standard-scaled neural networks. Another possibility would be to consider the mean-field scaling, where microscopic quantifiers of feature learning have been found already in the infinite-width limit [4, 5, 46–49]. Second, the proportional regime represents a genuinely overparameterized scenario. Therefore, our theory does not help to characterize feature learning closer to the interpolation threshold. This setting, which may also be relevant for modern deep learning applications, has been very recently investigated for Gaussian training inputs [50–53], leading to more complex feature learning mechanisms than the ones under scrutiny here [54–56].

Acknowledgements.— P.R. is supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) “A Multiscale integrated approach to the study of the nervous system in health and disease” (DN. 1553 11.10.2022). R.P. is funded by the MUR, project PRIN 2022HSKLK9.

* Correspondence email address: marco.gherardi@unimi.it

- [1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, *Nature* **323**, 533 (1986).
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM* **60**, 84–90 (2017).
- [3] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, Tech. Rep. 0 (University of Toronto, Toronto, Ontario, 2009).
- [4] G. Yang and E. J. Hu, Tensor programs iv: Feature learning in infinite-width neural networks, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 11727–11737.
- [5] I. Seroussi, G. Naveh, and Z. Ringel, Separation of scales and a thermodynamic description of feature learning in some cnns, *Nature Communications* **14**, 908 (2023).
- [6] A. X. Yang, M. Robeyns, E. Milsom, N. Schoots, and L. Aitchison, A theory of representation learning in deep neural networks gives a deep generalisation of kernel methods (2023), [arXiv:2108.13097 \[stat.ML\]](https://arxiv.org/abs/2108.13097).
- [7] C. Shi, L. Pan, and I. Dokmanić, Spring-block theory of feature learning in deep neural networks, *Phys. Rev. Lett.* **134**, 257301 (2025).
- [8] A. van Meegen and H. Sompolinsky, Coding schemes in neural networks learning classification tasks, *Nature Communications* **16**, 3354 (2025).
- [9] R. M. Neal, Priors for infinite networks, in *Bayesian Learning for Neural Networks* (Springer New York, New York, NY, 1996) pp. 29–53.
- [10] D. J. MacKay, Introduction to gaussian processes, NATO ASI Series F Computer and Systems Sciences **168**, 133 (1998).
- [11] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.*, Adaptive computation and machine learning (MIT Press, 2006) pp. I–XVIII, 1–248.
- [12] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, Deep neural networks as gaussian processes, in *International Conference on Learning Representations* (2018).
- [13] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-dickstein, Bayesian deep convolutional networks with many channels are gaussian processes, in *International Conference on Learning Representations* (2019).
- [14] G. Yang, Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, in *Neural Information Processing Systems* (2019).
- [15] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, *Nature communications* **12**, 1 (2021).
- [16] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
- [17] S. Favaro, B. Hanin, D. Marinucci, I. Nourdin, and G. Peccati, Quantitative clts in deep neural networks, *Probability Theory and Related Fields* **191**, 933 (2025).
- [18] Q. Li and H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization, *Phys. Rev. X* **11**, 031059 (2021).
- [19] B. Hanin and A. Zlokapa, Bayesian interpolation with deep linear networks, *Proceedings of the National Academy of Sciences* **120**, e2301345120 (2023), <https://www.pnas.org/doi/pdf/10.1073/pnas.2301345120>.
- [20] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo, A statistical mechanics framework for bayesian deep neural networks beyond the infinite-

- width limit, *Nature Machine Intelligence* **5**, 1497 (2023).
- [21] R. Pacelli, L. Giambagli, and P. Baglioni, Kernel shape renormalization in bayesian shallow networks: a gaussian process perspective, in *2024 IEEE Workshop on Complexity in Engineering (COMPENG)* (2024) pp. 1–6.
 - [22] H. Cui, F. Krzakala, and L. Zdeborová, Bayes-optimal learning of deep random networks of extensive-width, *Proceedings of the 40th International Conference on Machine Learning* (2023).
 - [23] F. Camilli, D. Tiepova, E. Bergamin, and J. Barbier, Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime, *arXiv preprint arXiv:2505.03577* (2025).
 - [24] R. Aiudi, R. Pacelli, P. Baglioni, A. Vezzani, R. Burioni, and P. Rotondo, Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks, *Nature Communications* **16**, 10.1038/s41467-024-55229-3 (2025).
 - [25] F. Bassetti, M. Gherardi, A. Ingrosso, M. Pastore, and P. Rotondo, Feature learning in finite-width bayesian deep linear networks with multiple outputs and convolutional layers, *Journal of Machine Learning Research* **26**, 1 (2025).
 - [26] P. Baglioni, R. Pacelli, R. Aiudi, F. Di Renzo, A. Vezzani, R. Burioni, and P. Rotondo, Predictive power of a bayesian effective action for fully connected one hidden layer neural networks in the proportional limit, *Phys. Rev. Lett.* **133**, 027301 (2024).
 - [27] P. Baglioni, L. Giambagli, A. Vezzani, R. Burioni, P. Rotondo, and R. Pacelli, Kernel shape renormalization explains output-output correlations in finite bayesian one-hidden-layer networks, *Phys. Rev. E* **111**, 065312 (2025).
 - [28] H. Shan, Q. Li, and H. Sompolinsky, Order parameters and phase transitions of continual learning in deep neural networks, *arXiv preprint arXiv:2407.10315* (2024).
 - [29] A. Ingrosso, R. Pacelli, P. Rotondo, and F. Gerace, Statistical mechanics of transfer learning in fully connected networks in the proportional limit, *Phys. Rev. Lett.* **134**, 177301 (2025).
 - [30] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, Finite versus infinite neural networks: an empirical study, in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020) pp. 15156–15172.
 - [31] S. Ciceri, L. Cassani, M. Osella, P. Rotondo, F. Valle, and M. Gherardi, Inversion dynamics of class manifolds in deep learning reveals tradeoffs underlying generalization, *Nature Machine Intelligence* **6**, 40 (2024).
 - [32] P. Pukowski and H. Lu, Investigating the impact of hard and easy samples on generalization reveals in-class data imbalance, in *AutoML 2024 Methods Track* (2024).
 - [33] S. Chung, D. D. Lee, and H. Sompolinsky, Classification and geometry of general perceptual manifolds, *Phys. Rev. X* **8**, 031003 (2018).
 - [34] P. Rotondo, M. Pastore, and M. Gherardi, Beyond the storage capacity: Data-driven satisfiability transition, *Phys. Rev. Lett.* **125**, 120601 (2020).
 - [35] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, *Phys. Rev. E* **102**, 032119 (2020).
 - [36] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, Separability and geometry of object manifolds in deep neural networks, *Nature Communications* **11**, 746 (2020).
 - [37] M. Pastore, Critical properties of the SAT/UNSAT transitions in the classification problem of structured data, *Journal of Statistical Mechanics: Theory and Experiment* **2021**, 113301 (2021).
 - [38] M. Farrell, S. Recanatesi, T. Moore, G. Lajoie, and E. Shea-Brown, Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion, *Nature Machine Intelligence* **4**, 564 (2022).
 - [39] K. Kamnitsas, D. Castro, L. L. Folgoc, I. Walker, R. Tanno, D. Rueckert, B. Glocker, A. Criminisi, and A. Nori, Semi-supervised learning via compact latent space clustering, in *Proceedings of the 35th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. 80, edited by J. Dy and A. Krause (PMLR, 2018) pp. 2459–2468.
 - [40] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, Vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018).
 - [41] E. Francazi, M. Baity-Jesi, and A. Lucchi, A theoretical analysis of the learning dynamics under class imbalance, in *Proceedings of the 40th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023) pp. 10285–10322.
 - [42] F. S. Pezzicoli, V. Ros, F. P. Landes, and M. Baity-Jesi, Anomaly detection with class imbalance: learning from exactly solvable models, in *The 28th International Conference on Artificial Intelligence and Statistics* (2025).
 - [43] N. Frosst, N. Papernot, and G. Hinton, Analyzing and improving representations with the soft nearest neighbor loss, in *Proceedings of the 36th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, 2019) pp. 2012–2020.
 - [44] H. He and W. J. Su, A law of data separation in deep learning, *Proceedings of the National Academy of Sciences* **120**, e2221704120 (2023), <https://www.pnas.org/doi/pdf/10.1073/pnas.2221704120>.
 - [45] A. Ali, J. Z. Kolter, and R. J. Tibshirani, A continuous-time view of early stopping for least squares regression, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, Vol. 89, edited by K. Chaudhuri and M. Sugiyama (PMLR, 2019) pp. 1370–1378.
 - [46] K. Fischer, J. Lindner, D. Dahmen, Z. Ringel, M. Krämer, and M. Helias, Critical feature learning in deep neural networks, in *Proceedings of the 41st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. 235, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (PMLR, 2024) pp. 13660–13690.
 - [47] N. Rubin, K. Fischer, J. Lindner, I. Seroussi, Z. Ringel, M. Krämer, and M. Helias, From kernels to features: A multi-scale adaptive theory of feature learning, in *Forty-second International Conference on Machine Learning* (2025).
 - [48] B. Bordelon and C. Pehlevan, Self-consistent dynamical

- field theory of kernel evolution in wide neural networks, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 32240–32256.
- [49] C. Lauditi, B. Bordelon, and C. Pehlevan, Adaptive kernel predictors from feature-learning infinite limits of neural networks, in *Forty-second International Conference on Machine Learning* (2025).
 - [50] F. Aguirre-López, S. Franz, and M. Pastore, Random features and polynomial rules, *SciPost Phys.* **18**, 039 (2025).
 - [51] J. Barbier, F. Camilli, M.-T. Nguyen, M. Pastore, and R. Skerk, Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation (2025), [arXiv:2501.18530 \[stat.ML\]](#).
 - [52] J. Barbier, F. Camilli, M.-T. Nguyen, M. Pastore, and R. Skerk, Statistical mechanics of extensive-width bayesian neural networks near interpolation (2025), [arXiv:2505.24849 \[stat.ML\]](#).
 - [53] V. Erba, E. Troiani, L. Zdeborová, and F. Krzakala, The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks (2025), [arXiv:2505.17958 \[stat.ML\]](#).
 - [54] A. Ingrosso and S. Goldt, Data-driven emergence of convolutional structure in neural networks, *Proceedings of the National Academy of Sciences* **119**, e2201854119 (2022), <https://www.pnas.org/doi/pdf/10.1073/pnas.2201854119>.
 - [55] Q. Li, B. Sorscher, and H. Sompolinsky, Representations and generalization in artificial and brain neural networks, *Proceedings of the National Academy of Sciences* **121**, e2311805121 (2024), <https://www.pnas.org/doi/pdf/10.1073/pnas.2311805121>.
 - [56] Z. Ringel, N. Rubin, E. Mor, M. Helias, and I. Seroussi, Applications of statistical field theory in deep learning, arXiv preprint [arXiv:2502.18553](#) (2025).

SUPPLEMENTARY MATERIAL

CONTENTS

References	5
Supplementary Material	8
Derivation of the average squared distance	8
Derivation of the second order statistics of the weights	11
Consistency check for linear activation function	14

DERIVATION OF THE AVERAGE SQUARED DISTANCE

In the main text, we mentioned that the squared distance can be defined to be finite also in the infinite-width regime and in the case of unbalanced datasets, where each class has a different number of data points. A more general definition can be considered:

$$D^2 = \frac{1}{N_1^\eta} \left\| \frac{1}{N_+^\delta} \sum_{\mu=1}^{N_+} \sigma(\mathbf{h}_+^\mu) - \frac{1}{N_-^\delta} \sum_{\mu=1}^{N_-} \sigma(\mathbf{h}_-^\mu) \right\|^2, \quad (9)$$

where $\eta \geq 0$, $\delta > 0$, and N_+ and N_- (with $N_+ + N_- = P$) denote the number of training data points belonging to the first and second class, respectively. Assuming that N_+ and N_- represent non-negligible fractions of the total number of points P — which is large in the proportional limit — we can define $\gamma = \left(\frac{N_+}{N_-}\right)^\delta$ and express the distance as

$$D^2 = \frac{1}{N_1^\eta N_+^{2\delta}} [\sigma(\mathbf{h}_+) \ \sigma(\mathbf{h}_-)] \cdot \begin{bmatrix} 1 & -\gamma \\ -\gamma & \gamma^2 \end{bmatrix} \cdot \begin{bmatrix} \sigma(\mathbf{h}_+) \\ \sigma(\mathbf{h}_-) \end{bmatrix}, \quad (10)$$

where the matrix and the vectors are to be understood in block form. By defining the vector z , with components

$$z_\mu = \begin{cases} 1 & \text{for } \mu = 1, \dots, N_+ \\ -\gamma & \text{for } \mu = N_+ + 1, \dots, P \end{cases},$$

it is possible to write the block matrix as a tensor product whose components are $z_\mu z_\nu$, which allows writing the squared distance as

$$D^2 = \frac{1}{N_1^\eta N_+^{2\delta}} \sum_{\mu\nu} z_\mu z_\nu \sum_i \sigma(h_i^\mu) \sigma(h_i^\nu). \quad (11)$$

Note that in the case of a balanced dataset with $N_+ = N_-$, we have $\gamma = 1$ and $z_\mu = \pm 1$, which are the values assumed by the labels y_μ , as in the case of the main text. Averaging over the posterior distribution, we have:

$$\langle D^2 \rangle = \frac{1}{N_1^\eta N_+^{2\delta}} \sum_{\mu\nu} z_\mu z_\nu \sum_i \langle \sigma(h_i^\mu) \sigma(h_i^\nu) \rangle. \quad (12)$$

The averaged similarity matrix has been computed in [24], which provides an explicit expression for the averaged similarity matrix:

$$\langle \sigma(h_i^\mu) \sigma(h_i^\nu) \rangle = K_{\mu\nu} - \frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda\delta} K_{\mu\lambda} K_{\nu\delta} \left[(\tilde{K}_{(R)}^{-1})_{\lambda\delta} - (\tilde{K}_{(R)}^{-1})_{\lambda\delta} (\tilde{K}_{(R)}^{-1})_{\delta\delta} \right]. \quad (13)$$

Since the averaged similarity matrix does not depend on the index i , we can trivially perform the internal sum to get:

$$\langle D^2 \rangle = \frac{1}{N_1^{\eta-1} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu z_\nu \langle \sigma(h_i^\mu) \sigma(h_i^\nu) \rangle, \quad (14)$$

which can be expressed as

$$\begin{aligned} \langle D^2 \rangle &= \frac{1}{N_1^{\eta-1} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu z_\nu \left[K_{\mu\nu} - \frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda\delta} K_{\mu\lambda} K_{\nu\delta} \left[(\tilde{K}_{(R)}^{-1})_{\lambda\delta} - (\tilde{K}_{(R)}^{-1} y)_\lambda (\tilde{K}_{(R)}^{-1} y)_\delta \right] \right] = \\ &= \tilde{D}_1^2 - \tilde{D}_2^2 + \tilde{D}_3^2. \end{aligned} \quad (15)$$

As mentioned in the main text, the terms that come from the posterior distribution (which involve y) are clearly distinguished from the ones which come from the relative minus sign of the norm (represented by z). We defined

$$\tilde{D}_1^2 = \frac{1}{N_1^{\eta-1} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu K_{\mu\nu} z_\nu, \quad (16)$$

$$\tilde{D}_2^2 = \frac{\bar{Q}}{\lambda} \frac{1}{N_1^{\eta} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu z_\nu \sum_{\lambda\delta} K_{\mu\lambda} K_{\nu\delta} (\tilde{K}_{(R)}^{-1})_{\lambda\delta}, \quad (17)$$

$$\begin{aligned} \tilde{D}_3^2 &= \frac{\bar{Q}}{\lambda_1} \frac{1}{N_1^{\eta} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu z_\nu \sum_{\lambda\delta} K_{\mu\lambda} K_{\nu\delta} \sum_{\sigma\rho} (\tilde{K}_{(R)}^{-1})_{\lambda\sigma} (\tilde{K}_{(R)}^{-1})_{\delta\rho} y_\sigma y_\rho = \\ &= \frac{\bar{Q}}{\lambda_1} \frac{1}{N_1^{\eta} N_+^{2\delta}} \sum_{\mu\nu}^P z_\mu z_\nu \left(\sum_{\lambda\sigma} K_{\mu\lambda} (\tilde{K}_{(R)}^{-1})_{\lambda\sigma} y_\sigma \right) \left(\sum_{\delta\rho} K_{\nu\delta} (\tilde{K}_{(R)}^{-1})_{\delta\rho} y_\rho \right) = \\ &= \frac{\lambda_1}{\bar{Q}} \frac{1}{N_1^{\eta} N_+^{2\delta}} \left(\sum_{\mu}^P z_\mu \langle f \rangle_\mu \right)^2, \end{aligned} \quad (18)$$

where $\langle f_\mu \rangle$ is defined as in the main text $\langle f_\mu \rangle = \sum_{\lambda,\sigma=1}^P (K_{(R)})_{\mu\lambda} (\tilde{K}_{(R)}^{-1})_{\lambda\sigma} y_\sigma$ and represents the expected output of the Gaussian Process associated with the Neural Network in the proportional limit. It can be observed that the expected output, although defined through a sum involving an increasing number of terms in the proportional limit, has finite size. In fact, if $\beta \rightarrow \infty$, then $\tilde{K}(R)^{-1}$ becomes proportional to K^{-1} . The result is thus a delta, and the output satisfies $\langle f_\mu \rangle = y_\mu$. On the other hand, if $\beta \rightarrow 0$, then $\tilde{K}(R)^{-1} \sim \beta \mathbb{1} \rightarrow 0$, which drives the outputs to zero. One can numerically check that all intermediate values of β smoothly interpolate between these two regimes, while keeping $\langle f_\mu \rangle$ of finite size. With this in place, we can proceed further by writing

$$\begin{aligned} \sum_{\mu}^P z_\mu \langle f_\mu \rangle &= \sum_{\mu}^{N_+} \langle f_+^\mu \rangle - \gamma \sum_{\mu}^{N_-} \langle f_-^\mu \rangle = \\ &= N_+ \left(\bar{f}_+ - \gamma^{\frac{\delta-1}{\delta}} \bar{f}_- \right). \end{aligned} \quad (19)$$

We defined $\bar{f}_\pm^\mu = 1/N_\pm \sum_{\mu} \langle f_\pm^\mu \rangle$. Since in the proportional limit all the quantities $N_1, P, N_+, N_- \rightarrow \infty$ at the same rate, it is useful to express everything as a function of P . From the relations $N_1 = P/\alpha$ and

$$P = \left(1 + \frac{1}{\sqrt{\delta/\gamma}} \right) N_+ = \tilde{\gamma} N_+, \quad (20)$$

we can explicitly write the three contributions to the distance:

$$\tilde{D}_1^2 = \frac{\alpha^{\eta-1} \tilde{\gamma}^{2\delta}}{P^{\eta+2\delta-1}} \sum_{\mu\nu}^P z_\mu K_{\mu\nu} z_\nu, \quad (21)$$

$$\tilde{D}_2^2 = \frac{\bar{Q}}{\lambda_1} \frac{\alpha^\eta \bar{\gamma}^{2\delta}}{P^{\eta+2\delta}} \sum_{\mu\nu} z_\mu z_\nu \sum_{\lambda\delta} K_{\mu\lambda} K_{\nu\delta} (\tilde{K}_{(R)}^{-1})_{\lambda\delta}, \quad (22)$$

$$\tilde{D}_3^2 = \frac{\lambda_1 \alpha^\eta \bar{\gamma}^{2\delta-2}}{\bar{Q}} \frac{1}{P^{\eta+2\delta-2}} \left(\bar{f}_+ - \gamma^{\frac{\delta-1}{\delta}} \bar{f}_- \right)^2. \quad (23)$$

The zero temperature limit can be used to justify the different scalings in P of the three quantities. When $T \rightarrow 0$, the internal sum of the second contribution returns $K_{\mu\nu}$, telling us that \tilde{D}_2^2 can be neglected. Under the assumption that $1/P \sum_{\mu\nu} z_\mu K_{\mu\nu} z_\nu \sim \mathcal{O}(1)$, the term \tilde{D}_1^2 is of the same order of \tilde{D}_3^2 , making it a feature dependent correction that reflects the fact that the posterior is learning through the labels. As mentioned in the main text, the assumption is true in the case of Gaussian data, where the kernel has independent entries, which make the sum $\sum_{\mu\nu} K_{\mu\nu} z_\mu z_\nu \sim \sqrt{P}$ and $1/P \sum_{\mu\nu} z_\mu K_{\mu\nu} z_\nu = \mathcal{O}(1)$ in virtue of the central limit theorem. In the case of fairly noisy data, it is reasonable to expect the term to remain finite also for large values of P . As shown in the graphs below ($\eta = 0, \delta = 1$), this is the case for CIFAR10 dataset.

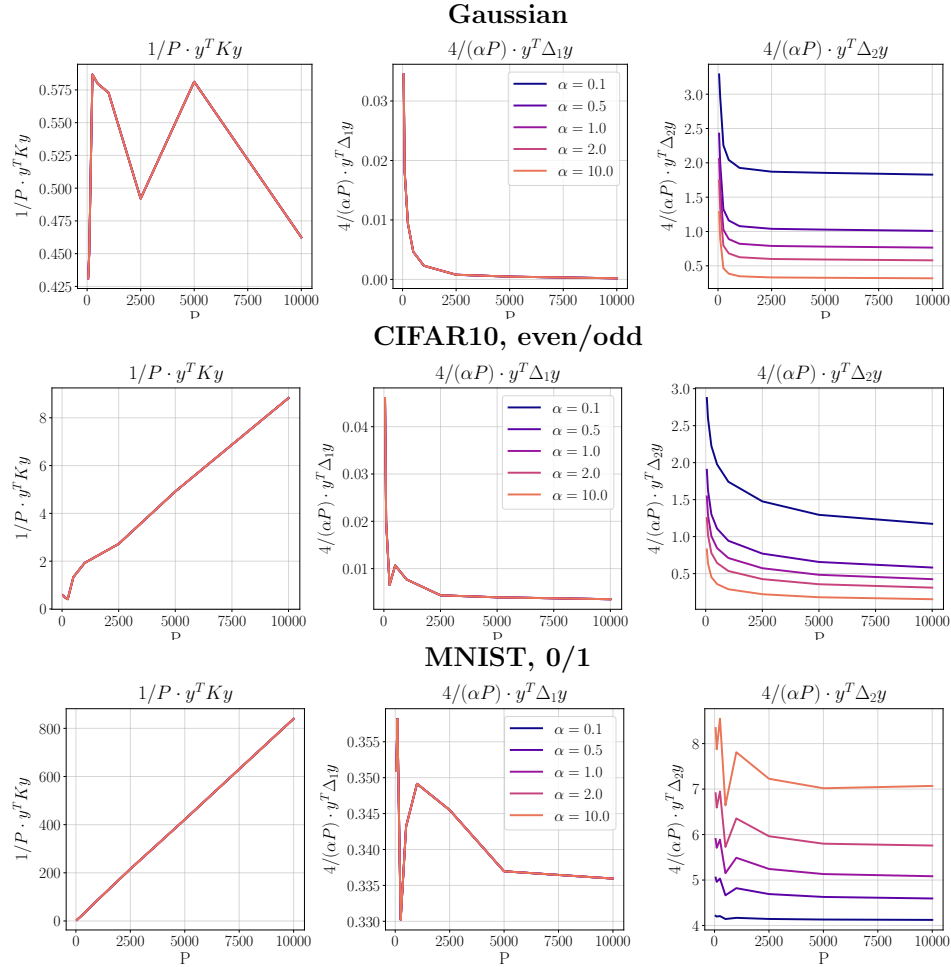


Figure 3. Gaussian data are associated with the theoretical scaling provided in the text. Strictly speaking, the scaling is not satisfied for CIFAR10 dataset, where $1/P z^T K z$ grows linearly. However, since the effective action formalism has been proven to work for large but finite P and N_1 , the central point is to identify a regime where the magnitude of \tilde{D}_1^2 is similar to the one of \tilde{D}_3^2 . In this case the first term grows slowly, allowing us to consider the two contributions effectively of the same order. This approach cannot be pursued in the case of MNIST dataset restricted to 0 and 1 classes. In this case, the kernel strongly correlates with the labels highlighting a linear scaling with a more pronounced growth rate. The graphs are plotted for the temperature value $T = 0.001$.

This type of argument holds for every choice of the scaling that meets the condition $\eta + 2\delta - 2 = 0$. One possible choice, which is the one presented in the main text, consists in selecting $\eta = 0, \delta = 1$, which returns (disregarding the

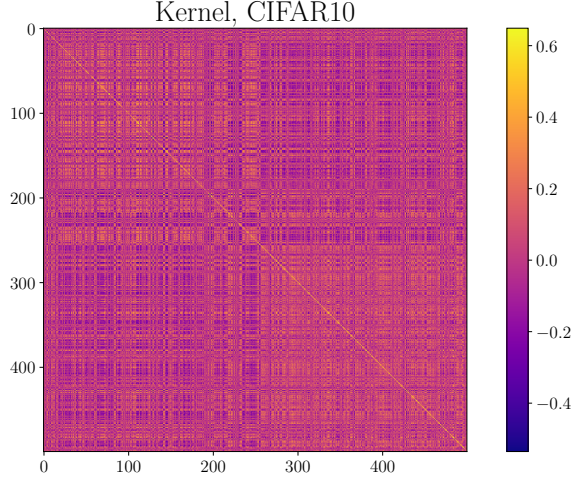


Figure 4. *
(a) CIFAR10, even vs. odd

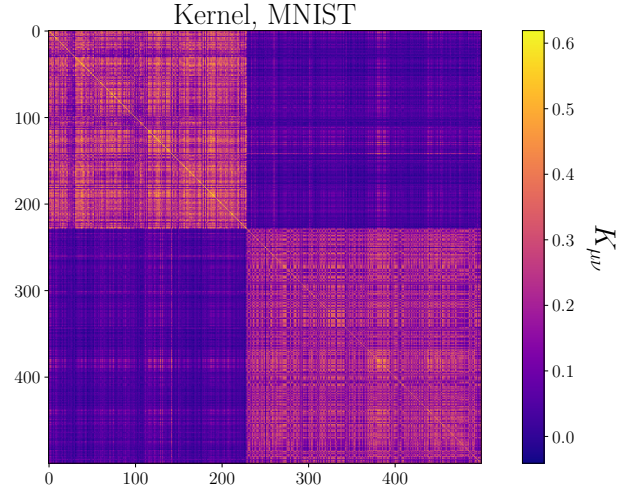


Figure 5. *
(b) MNIST, 0 vs. 1

Figure 6. For MNIST, the positive values of the kernel correlate with z , giving a linear scaling of the dominant contribution. For noisier datasets this effect is mitigated.

second contribution):

$$\langle D^2 \rangle = \frac{1}{\alpha} \left(\frac{\gamma + 1}{\gamma} \right)^2 \frac{1}{P} \sum_{\mu\nu}^P z_\mu K_{\mu\nu} z_\nu + \frac{\lambda_1}{Q} (\bar{f}_+ - \bar{f}_-)^2. \quad (24)$$

As argued in the main text, the divergence in the infinite-width limit $\alpha \rightarrow \infty$ is completely understood in terms of the original definition of the distance, which diverges as the norm of a vector with a growing number of components. As an alternative, the definition with $\eta = 1$ and $\delta = 1/2$ returns a well-defined infinite-width limit where the distance reads:

$$\langle D^2 \rangle = \frac{\gamma^2 + 1}{\gamma^2} \frac{1}{P} \sum_{\mu\nu} z_\mu K_{\mu\nu} z_\nu + \alpha \frac{\lambda_1}{Q} \frac{\gamma^2}{\gamma^2 + 1} \left(\bar{f}_+ - \frac{1}{\gamma} \bar{f}_- \right)^2 \quad (25)$$

Note that in both cases the relative magnitude between the prior term and the posterior correction is proportional to α . Because of this, in the infinite-width limit the prior term is dominant and the posterior correction can be disregarded, returning a distance that does not depend on the labels.

DERIVATION OF THE SECOND ORDER STATISTICS OF THE WEIGHTS

To compute the second order statistics of the internal weights, it is possible to introduce a partition function with a source term:

$$\mathcal{Z}(\{J\}) = \int dv dW e^{-\frac{\beta}{2} |f-y|^2} e^{-\frac{\lambda_0}{2} \|W\|^2} e^{-\frac{\lambda_1}{2} \|v\|^2} e^{-\frac{\lambda_0}{2} \left(\sum_{k=1}^{N_1} J_k W_{1k} \right)^2}. \quad (26)$$

With this definition, we have

$$\langle W_{1k} W_{1h} \rangle = -\frac{1}{\lambda_0} \frac{1}{\mathcal{Z}(0)} \frac{\partial^2 \mathcal{Z}(\{J\})}{\partial J_k \partial J_h} \Big|_{J=0}. \quad (27)$$

Note that, being interested in mean values, it is possible not to take care of the normalizations (unless they are J -dependent) while computing the partition function. Considering the inner integral, we separate the contribution

of the terms with $i \neq 1$ from that with $i = 1$ (omitting the products over the indices that are being integrated in an obvious way).

$$\int \prod_{i \neq 1}^{N_1} dW_{ij} \delta \left(h_i^\mu - \frac{1}{N_0} \sum_j W_{ij} x_j^\mu \right) e^{-\frac{\lambda_0}{2} \|W\|^2} \int dW_{1j} \delta \left(h_1^\mu - \frac{1}{N_0} \sum_j W_{1j} x_j^\mu \right) e^{-\frac{\lambda_0}{2} \|W\|^2} e^{-\frac{\lambda_0}{2} \left(\sum_{k=1}^{N_1} J_k W_{1k} \right)^2}. \quad (28)$$

After the integration, the previous expression assumes the form:

$$\left(\prod_{i \neq 1}^{N_1} \frac{1}{\sqrt{\text{Det}(C)}} e^{-\frac{1}{2} \sum_{\mu\nu} h_i^\mu C_{\mu\nu}^{-1} h_i^\nu} \right) \frac{1}{\sqrt{\text{Det}(D(J))}} e^{-\frac{1}{2} \sum_{\mu\nu} h_1^\mu D(J)_{\mu\nu}^{-1} h_1^\nu} \frac{1}{\sqrt{\text{Det}(\delta_{ij} + J_i J_j)}}, \quad (29)$$

where we defined $D(J)$ as:

$$\begin{aligned} D(J)_{\mu\nu} &= \sum_{ij} \frac{1}{N_0 \lambda_0} x_i^\mu x_j^\nu (\delta_{ij} + J_i J_j)^{-1} = \\ &= C_{ij} - \frac{1}{N_0 \lambda_0} \frac{1}{1 + J^2} \left(\sum_i J_i x_i^\mu \right) \left(\sum_i J_i x_i^\nu \right), \end{aligned} \quad (30)$$

The inverse is computed by employing the Sherman-Morrison formula. Note that we defined $J^2 = \sum_k J_k^2$ and that $\text{Det}(\delta_{ij} + J_i J_j) = 1 + J^2$. That said, the partition function is simplified by the introduction of an integral representation of a Dirac delta function $\prod_\mu \delta(s^\mu - f(v, h^\mu))$. Integrating the read-out weights, we obtain:

$$\begin{aligned} \mathcal{Z}(\{J\}) &= \int ds d\bar{s} e^{-\frac{\beta}{2} |s-y|^2} e^{is\bar{s}} \int \prod_{i \neq 1}^{N_1} dv_i dh_i^\mu e^{-\frac{\lambda_1}{2} \|v\|^2} e^{-\frac{1}{2} \sum_{\mu\nu} h_i^\mu C_{\mu\nu}^{-1} h_i^\nu} e^{-iv_i \frac{1}{\sqrt{N_1}} \sum_\mu \bar{s}^\mu \sigma(h_i^\mu)} \\ &\quad \int dv_1 dh_1^\mu e^{-\frac{\lambda_1}{2} v_1^2} \frac{1}{\sqrt{\text{Det}(D(J))}} e^{-\frac{1}{2} \sum_{\mu\nu} h_1^\mu D(J)_{\mu\nu}^{-1} h_1^\nu} e^{-iv_1 \frac{1}{\sqrt{N_1}} \sum_\mu \bar{s}^\mu \sigma(h_1^\mu)} \frac{1}{\sqrt{1 + J^2}} = \\ &= \int ds d\bar{s} e^{-\frac{\beta}{2} |s-y|^2} e^{is\bar{s}} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K \bar{s} \right]^{-\frac{N_1-1}{2}} \frac{1}{\sqrt{1 + J^2}} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K(J) \bar{s} \right]^{-\frac{1}{2}}, \end{aligned} \quad (31)$$

where the Kernel $K(J)$ is defined by

$$\begin{aligned} K(J)_{\mu\nu} &= \int dh \sigma(h^\mu) \sigma(h^\nu) \frac{e^{-\frac{1}{2} h D(J)^{-1} h}}{\sqrt{\text{Det}(D(J))}} = \\ &= \int dh d\bar{h} \sigma(h^\mu) \sigma(h^\nu) e^{-\frac{1}{2} \bar{h} D(J) \bar{h}} e^{ih\bar{h}}. \end{aligned} \quad (32)$$

A tedious but straightforward computation of the derivatives yields

$$\begin{aligned} \partial_h \partial_k \frac{1}{\sqrt{1 + J^2}} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K(J) \bar{s} \right]^{-\frac{1}{2}} \Big|_{J=0} &= \\ = -\delta_{hk} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K \bar{s} \right]^{-\frac{1}{2}} - \frac{1}{2} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K \bar{s} \right]^{-\frac{3}{2}} \frac{1}{\lambda_1 N_1} \bar{s} \left(\partial_h \partial_k K(J) \Big|_{J=0} \right) \bar{s}, \end{aligned} \quad (33)$$

where the second derivative of the Kernel assumes the form of

$$\left(\partial_h \partial_k K(J) \Big|_{J=0} \right)_{\mu\nu} = \int dh d\bar{h} \sigma(h^\mu) \sigma(h^\nu) \left(\bar{h} \frac{x_h x_k^T}{\lambda_0 N_0} \bar{h} \right) e^{-\frac{1}{2} \bar{h} C \bar{h}} e^{ih\bar{h}}. \quad (34)$$

The Fourier variable \bar{h} can be integrated out by means of a translation, leaving

$$\begin{aligned} \left(\partial_h \partial_k K(J) \Big|_{J=0} \right)_{\mu\nu} &= \frac{1}{\lambda_0 N_0} \left(\sum_{\lambda\sigma} x_h^\lambda x_k^\sigma C_{\lambda\sigma}^{-1} \right) \int dh e^{-\frac{1}{2} h C^{-1} h} \sigma(h^\mu) \sigma(h^\nu) - \\ &\quad - \frac{1}{\lambda_0 N_0} \sum_{\eta\delta\lambda\sigma} C_{\eta\lambda}^{-1} x_h^\lambda x_k^\sigma C_{\sigma\delta}^{-1} \int dh e^{-\frac{1}{2} h C^{-1} h} \sigma(h^\mu) \sigma(h^\nu) h^\eta h^\delta \end{aligned} \quad (35)$$

Note that the expression appears to be ill-defined since it depends on the single components of the inverse of the covariance matrix, which is non-existent as soon as $P > N_0$, which is typically the case of interest. It turns out that it is not the case and that the expression can be recast in such a way that the dependence on the single components of C^{-1} is removed. Note also that the presence of C^{-1} in the Gaussian measure associated with the variable h does not represent an issue since the distribution is well defined through its Fourier transform. To show the independence of the previous expression on C^{-1} , we first note that the integrals define two different kernels: while the first is the usual NNGP kernel, which we denote $K_{\mu\nu}$, the second depends on more indices and is defined as

$$K_{\mu\nu}^{\eta\delta} = \int dh e^{-\frac{1}{2}hC^{-1}h} \sigma(h^\mu) \sigma(h^\nu) h^\eta h^\delta. \quad (36)$$

To recast this expression, we express it in terms of the NNGP kernel. To do so, we introduce a source such that:

$$\begin{aligned} K_{\mu\nu}^{\eta\delta} &= \partial_\eta \partial_\delta (K(\{J\}))^{\mu\nu} \Big|_{J=0} = \\ &= \partial_\eta \partial_\delta \int dh e^{-\frac{1}{2}hC^{-1}h} \sigma(h^\mu) \sigma(h^\nu) e^{\sum_\alpha J_\alpha h^\alpha} \Big|_{J=0}. \end{aligned} \quad (37)$$

After completing the square and performing a translation, the kernel is brought in the form of:

$$(K(\{J\}))^{\mu\nu} = e^{\frac{1}{2}JCJ} \int dh e^{-\frac{1}{2}hC^{-1}h} \sigma(h^\mu - (CJ)^\mu) \sigma(h^\nu - (CJ)^\nu). \quad (38)$$

The computation of the derivatives returns:

$$K_{\mu\nu}^{\lambda\sigma} = C_{\lambda\sigma} K_{\mu\nu} + C_{\mu\sigma} C_{\mu\lambda} \int \mathcal{D}h \sigma''_\mu \sigma_\nu + C_{\nu\lambda} C_{\nu\sigma} \int \mathcal{D}h \sigma'_\nu \sigma_\mu + (C_{\mu\lambda} C_{\nu\sigma} + C_{\nu\lambda} C_{\mu\sigma}) \int \mathcal{D}h \sigma'_\mu \sigma'_\nu, \quad (39)$$

where σ'_μ and σ''_μ are the first and second derivatives of the activation function evaluated on h^μ . Inserting this expression in Eq. (35), the sums over the Greek indices remove the dependence on C^{-1} :

$$\begin{aligned} \left(\partial_h \partial_k K(J) \Big|_{J=0} \right)_{\mu\nu} &= - \int dh e^{-\frac{1}{2}hC^{-1}h} [\sigma''(h^\mu) \sigma(h^\nu) x_h^\mu x_k^\mu + \sigma(h^\mu) \sigma''(h^\nu) x_h^\nu x_k^\nu] - \\ &\quad - \frac{x_h^\mu x_k^\nu + x_h^\nu x_k^\mu}{\lambda_0 N_0} \int dh e^{-\frac{1}{2}hC^{-1}h} \sigma'(h^\mu) \sigma'(h^\nu), \end{aligned} \quad (40)$$

which is well defined also in case of $P > N_0$. Defining

$$(K'')^{\mu\nu}_{hk} = \int dh e^{-\frac{1}{2}hC^{-1}h} [\sigma''(h^\mu) \sigma(h^\nu) x_h^\mu x_k^\mu + \sigma(h^\mu) \sigma''(h^\nu) x_h^\nu x_k^\nu], \quad (41)$$

$$(K')^{\mu\nu} = \int dh e^{-\frac{1}{2}hC^{-1}h} \sigma'(h^\mu) \sigma'(h^\nu), \quad (42)$$

we denote for convenience $\left(\partial_h \partial_k K(J) \Big|_{J=0} \right)_{\mu\nu} = \Delta_{hk}^{\mu\nu}$, with:

$$\Delta_{hk}^{\mu\nu} = -(K'')^{\mu\nu}_{hk} - \frac{x_h^\mu x_k^\nu + x_h^\nu x_k^\mu}{\lambda_0 N_0} (K')^{\mu\nu}. \quad (43)$$

It is now possible to continue with the computation of the second order statistics. From the definition and the previous results, we have:

$$\begin{aligned} \langle W_{1h} W_{1k} \rangle &= \frac{\delta_{hk}}{\lambda_0 \mathcal{Z}(0)} \int ds d\bar{s} e^{-\frac{\beta}{2}|s-y|^2} e^{is\bar{s}} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K \bar{s} \right]^{-\frac{N_1}{2}} + \\ &\quad + \frac{1}{2\lambda_0 \mathcal{Z}(0)} \int ds d\bar{s} e^{-\frac{\beta}{2}|s-y|^2} e^{is\bar{s}} \left[1 + \frac{1}{\lambda_1 N_1} \bar{s} K \bar{s} \right]^{-\frac{N_1}{2}-1} \frac{1}{\lambda_1 N_1} \bar{s} \Delta_{hk} \bar{s}. \end{aligned} \quad (44)$$

From Eq. (31), by plugging $J = 0$ one can check that the integral in the first term is exactly $\mathcal{Z}(0)$, so that the first contribution to the expectation value is simply δ_{hk}/λ_0 . To make further progress with the second term, we use the integral definition of the Gamma function (the Gamma itself plays the role of a normalization constant and is not reported):

$$\begin{aligned} \left[1 + \frac{\bar{s}K\bar{s}}{\lambda_1 N_1}\right]^{-\frac{N_1}{2}-1} &= \int_{\tilde{Q}>0} d\tilde{Q} e^{-\tilde{Q}\left(1+\frac{\bar{s}K\bar{s}}{\lambda_1 N_1}\right)} \tilde{Q}^{\frac{N_1}{2}} = \\ &= \int_{Q>0} dQ e^{-\frac{N_1}{2}Q\left(1+\frac{\bar{s}K\bar{s}}{\lambda_1 N_1}\right)} Q^{\frac{N_1}{2}-1} Q \end{aligned} \quad (45)$$

The second line is obtained by rescaling the integration variable $\tilde{Q} = QN_1/2$. The integral in the second term reads:

$$\begin{aligned} &\int_{Q>0} dQ ds d\bar{s} e^{-\frac{N_1}{2}Q\left(1+\frac{\bar{s}K\bar{s}}{\lambda_1 N_1}\right)} Q^{\frac{N_1}{2}-1} Q e^{-\frac{\beta}{2}|s-y|^2} e^{is\bar{s}} \frac{\bar{s}\Delta_{hk}\bar{s}}{\lambda_1 N_1} = \\ &= \int_{Q>0} dQ e^{-\frac{N_1}{2}Q - \frac{N_1-2}{2}\log Q} \int d\bar{s} e^{-\frac{1}{2}(\bar{s}+i\tilde{K}_{(R)}^{-1}y)\tilde{K}_{(R)}(\bar{s}+i\tilde{K}_{(R)}^{-1}y)} e^{-\frac{1}{2}\log\beta} e^{-\frac{1}{2}y\tilde{K}_{(R)}^{-1}y} \frac{Q}{\lambda_1 N_1} \bar{s}\Delta_{hk}\bar{s} = \\ &= \int_{Q>0} dQ e^{-\frac{N_1}{2}Q - \frac{N_1-2}{2}\log Q} e^{-\frac{1}{2}\log\beta} e^{-\frac{1}{2}y\tilde{K}_{(R)}^{-1}y} \int d\bar{t} e^{-\frac{1}{2}\bar{t}\tilde{K}_{(R)}\bar{t}} \frac{Q}{\lambda_1 N_1} (\bar{t} - i\tilde{K}_{(R)}y) \Delta_{hk} (\bar{t} - i\tilde{K}_{(R)}y) = \\ &= \int_{Q>0} dQ e^{-\frac{N_1}{2}Q - \frac{N_1-2}{2}\log Q} e^{-\frac{1}{2}y\tilde{K}_{(R)}^{-1}y} e^{-\frac{1}{2}\log\text{Det}\beta\tilde{K}_{(R)}} \frac{Q}{\lambda_1 N_1} \sum_{\mu\nu} \left[(\tilde{K}_{(R)}^{-1})_{\mu\nu} \Delta_{hk}^{\mu\nu} - \sum_{\lambda\sigma} (\tilde{K}_{(R)}^{-1})_{\mu\lambda} y_\lambda (\Delta_{hk}^{\mu\nu}) (\tilde{K}_{(R)}^{-1})_{\nu\sigma} y_\sigma \right] = \\ &= \int_{Q>0} dQ e^{-\frac{N_1}{2}S(Q)} \frac{Q}{\lambda_1 N_1} \sum_{\mu\nu} \left[(\tilde{K}_{(R)}^{-1})_{\mu\nu} \Delta_{hk}^{\mu\nu} - \sum_{\lambda\sigma} (\tilde{K}_{(R)}^{-1})_{\mu\lambda} y_\lambda (\Delta_{hk}^{\mu\nu}) (\tilde{K}_{(R)}^{-1})_{\nu\sigma} y_\sigma \right], \end{aligned} \quad (46)$$

where the effective action $S(Q)$ is defined as

$$S(Q) = -Q + \frac{N_1-2}{N_1} \log Q - \frac{\alpha}{P} y^T \tilde{K}_{(R)}^{-1} y - \frac{\alpha}{P} \text{Tr} \log \beta \tilde{K}_{(R)}, \quad (47)$$

with

$$\tilde{K}_{(R)} = \frac{1}{\beta} + \frac{Q}{\lambda_1} K. \quad (48)$$

By introducing an integration variable Q in the same way as it was done in the previous lines, one can easily check that the partition function $\mathcal{Z}(0) = \int_{Q>0} \exp[-N_1 S(Q)/2]$. Because of that, in the proportional limit where N_1 is large, the integral is dominated by the saddle-point contribution (\bar{Q} such that $\partial_Q S(Q)|_{Q=\bar{Q}} = 0$), returning:

$$\langle W_{1h} W_{1k} \rangle = \frac{\delta_{hk}}{\lambda_0} + \frac{1}{\lambda_0} \frac{\bar{Q}}{2\lambda_1 N_1} \sum_{\mu,\nu=1}^P \left[(\tilde{K}_{(R)}^{-1})_{\mu\nu} \Delta_{hk}^{\mu\nu} - \sum_{\lambda\sigma} (\tilde{K}_{(R)}^{-1})_{\mu\lambda} y_\lambda (\Delta_{hk}^{\mu\nu}) (\tilde{K}_{(R)}^{-1})_{\nu\sigma} y_\sigma \right]. \quad (49)$$

As mentioned in the main text, the first term is obtained as the expectation value over the prior distribution of the weights W : $\langle W_{1h} W_{1k} \rangle_{P(W)} = \delta_{hk}/\lambda_0$ or over the posterior in the infinite-width limit. The latter statement can be easily checked by noting that when P is finite the term proportional to $1/N_1 \sum^P \rightarrow 0$. In the proportional limit, as argued in the main text, this term can be nontrivial and bring finite corrections to the two-point function.

CONSISTENCY CHECK FOR LINEAR ACTIVATION FUNCTION

In the case of a linear activation function $\sigma = Id$, we can compute

$$\langle \sigma(h_1^\mu) \sigma(h_1^\nu) \rangle = \langle h_1^\mu h_1^\nu \rangle = \frac{1}{N_0} \sum_{hk} x_h^\mu x_k^\nu \langle W_{1h} W_{1k} \rangle \quad (50)$$

and the result must match the one presented in Eq. (13), which is obtained by an independent computation. First of all, the expression of $\Delta_{hk}^{\mu\nu}$ simplifies as long as $\sigma'(h^\mu) = 1$ and $\sigma''(h^\mu) = 0$, returning

$$\Delta_{hk}^{\mu\nu} = -\frac{x_h^\mu x_k^\nu + x_h^\nu x_k^\mu}{\lambda_0 N_0}. \quad (51)$$

Furthermore, $K_{\mu\nu} = C_{\mu\nu}$. Plugging the previous equation into the expectation value:

$$\langle W_{1h} W_{1k} \rangle = \frac{\delta_{hk}}{\lambda_0} + \frac{1}{\lambda_0} \frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda\sigma} \left[-(\tilde{C}_{(R)}^{-1})_{\lambda\sigma} \frac{x_h^\lambda x_k^\sigma}{\lambda_0 N_0} + (\tilde{C}_{(R)}^{-1} y)_\lambda (\tilde{C}_{(R)}^{-1} y)_\sigma \frac{x_h^\lambda x_k^\sigma}{\lambda_0 N_0} \right], \quad (52)$$

where we used the notation $\tilde{C}_{(R)} = \mathbb{1}/\beta + \bar{Q}C/\lambda_1$. The computation of $\langle h_i^\mu h_i^\nu \rangle$ involves two further sums over the indices h, k . The terms involved return:

$$\sum_{hk} \frac{x_h^\mu x_k^\nu}{N_0} \frac{\delta_{hk}}{\lambda_0} = C_{\mu\nu}, \quad (53)$$

$$\frac{1}{\lambda_0} \sum_{hk} -\frac{x_h^\lambda x_k^\sigma}{\lambda_0 N_0} \frac{x_h^\mu x_k^\nu}{N_0} = -C_{\mu\lambda} C_{\nu\sigma}. \quad (54)$$

With this in mind, plugging the previous expressions in Eq. (50), we obtain:

$$\langle h_1^\mu h_1^\nu \rangle = C_{\mu\nu} - \frac{\bar{Q}}{\lambda_1 N_1} \sum_{\lambda\sigma}^P C_{\mu\lambda} C_{\nu\sigma} \left[(\tilde{C}_{(R)}^{-1})_{\lambda\sigma} - (\tilde{C}_{(R)}^{-1} y)_\lambda (\tilde{C}_{(R)}^{-1} y)_\sigma \right], \quad (55)$$

which is the equivalent of Eq. (13) in the case of linear activation function.