HAL open science

Gaussian Universality of Perceptrons with Random Labels

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, Lenka Zdeborová

To cite this version:

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, Lenka Zdeborová. Gaussian Universality of Perceptrons with Random Labels. 2023. hal-04019749

HAL Id: hal-04019749

https://hal.science/hal-04019749v1

Preprint submitted on 8 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Gaussian Universality of Perceptrons with Random Labels

Federica Gerace[1,4], Florent Krzakala[2], Bruno Loureiro[2,3], Ludovic Stephan[2], and Lenka Zdeborová[4]
[1] *International School of Advanced Studies (SISSA). Trieste, Italy.*
[2] *EPFL, Information, Learning and Physics (IdePHICS) lab., Lausanne, Switzerland*
[3] *Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France and*
[4] *EPFL Statistical Physics of Computation (SPOC) lab.,Lausanne, Switzerland*
(Dated: March 3, 2023)

While classical in many theoretical settings — and in particular in statistical physics-inspired works— the assumption of Gaussian *i.i.d.* input data is often perceived as a strong limitation in the context of statistics and machine learning. In this study, we redeem this line of work in the case of generalized linear classification, a.k.a. the perceptron model, with random labels. We argue that there is a large universality class of high-dimensional input data for which we obtain the same minimum training loss as for Gaussian data with corresponding data covariance. In the limit of vanishing regularization, we further demonstrate that the training loss is independent of the data covariance. On the theoretical side, we prove this universality for an arbitrary mixture of homogeneous Gaussian clouds. Empirically, we show that the universality holds also for a broad range of real datasets.

arXiv:2205.13303v2 [stat.ML] 2 Mar 2023

I. INTRODUCTION

Statistical physics studies of artificial neural networks have a long history, including many works that continue to have an impact on the current investigations of deep neural networks. A large fraction of this continuing line of works has focused on Gaussian input data, see [1–3] for some of the earliest and most influential examples. However, the Gaussian data assumption is not limited to works from statistical physics of learning. Indeed, it is a widespread assumption in the high-dimensional statistics literature, where it is also known under the umbrella of *Gaussian design*, see for example [4–6]. Despite being both common and convenient for doing theory, *i.i.d.* Gaussian data might come across as a stringent limitation at first sight, out-of-touch with the real-world practice where data is structured. Indeed, an important branch of statistical learning theory is data-agnostic and avoids making too specific assumptions on the data distribution [7]. However, a number of recent observations (both heuristic and rigorous) suggest that the Gaussian assumption is not always that far-stretched for high-dimensional data (see for instance [8–12] and references therein). The goal of the present work is to redeem the Gaussian hypothesis for perhaps the simplest, yet deeply fundamental, problem of high-dimensional statistics: the perceptron problem, a.k.a. generalized linear classification, with random labels.

Models with random labels are ubiquitous in the theory of machine learning. The problem of how many randomly labelled Gaussian patterns a perceptron model can fit, known as the *storage capacity problem*, is at the root of the historical interest of the statistical physics community for machine learning problems. Indeed, works on this classical subject span more than four decades [1, 2, 13–19]. The interest for random labels is also not bound to the statistical physics of learning community. They appear in several contexts in statistical learning theory, such as in the definition of Rademacher complexities [7, 20], in Wendel/Cover's pioneering studies [21, 22] or in thought-provoking numerical experiments with deep learning [23, 24],

In this work, we ask: how would these theories for random labels change if using a realistic data set instead of a Gaussian one? We consider the training loss of generalized linear classifiers (perceptrons) trained on random labels, including ridge, hinge and logistic classification [25], but also kernel methods [26] and neural networks trained in the lazy regime [27] (the so-called neural tangent kernel [28]), as well as with engineered features such as the scattering transform [29]. We focus on the thermodynamic limit (known as the high-dimensional setting in statistics) where both n (the number of training samples) and p (the input dimension) go to infinity at a fixed rate $\alpha = n/p$.

Our main result is to argue that in the aforementioned setting with random labels many input data distributions actually have the same learning properties as Gaussian data, thus providing a rather surprising Gaussian universality result for this problem. In particular, the minimum training loss for a wide range of settings is the same as that of a corresponding Gaussian problem with matching data covariance. Furthermore, in the limit of vanishing regularization, we show that Gaussian universality is even stronger, as the minimum training loss is independent of the data covariance (and therefore the same as the one of *i.i.d.* Gaussian data). In other words: as far as random labels are concerned, it turns out that the theoretical results derived under the Gaussian data assumptions capture what is actually happening in practice. Certainly, the value of the interpolation (or capacity) threshold was known to be universal and occurs (for full-rank data) at $n = p$ for ridge regression, and for $n = 2p$ for linear classifiers (perceptrons) [22]; however, the fact that the loss itself is universal is a stronger statement that redeems an entire line of work using the Gaussian data assumption, and in particular a large part of those from statistical physics of learning.

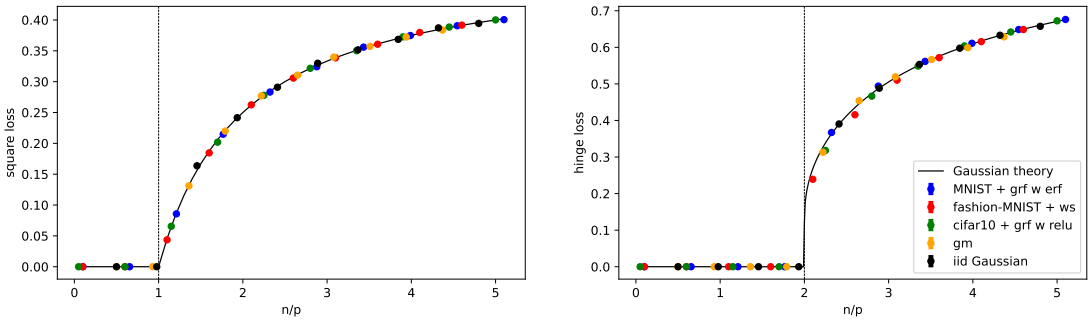

FIG. 1. Training loss as function of the number of samples n per input dimension p at regularization $\lambda = 10^{-15}$. In the left panel the square loss, and in the right panel the hinge loss. The black solid line represents the outcome of the replica calculation for *i.i.d* Gaussian inputs, namely when the covariance matrix Σ corresponds to the identity matrix. Dots refer to numerical simulations on different full-rank datasets. In particular, blue dots correspond to MNIST with Gaussian random features and error function non-linearity, red dots correspond to fashion-MNIST with wavelet scattering transform, green dots correspond to CIFAR10 in grayscale with Gaussian random features and ReLU non-linearity, yellow dots corresponds to a mixture of Gaussians, with means $\boldsymbol{\mu}_\pm = (\pm 1, 0, ..., 0)$, covariances $\Sigma_\pm$ both equal to the identity matrix and relative class proportions $\rho_\pm = {}^1\!/\!_2$. Finally, black dots correspond to *i.i.d.* Gaussian inputs.

Summary of main results

The main points of the present work can be summarized by Figures 1 and 2, which show the training loss of real-world data sets trained with random labels and various feature maps, compared with the (analytical) prediction derived for Gaussian data with matching covariance. The code used to run these experiments is publicly available in a GitHub repository. As illustrated in these plots, Gaussian universality seems to hold even for finite-dimensional data, and for actual real datasets. Notably, we observe that when using random labels the training losses plotted as a function of the ratio between the number of samples and the dimension $\alpha = n/p$ are indistinguishable from results obtained for Gaussian input data when using MNIST [30], fashion-MNIST [31], CIFAR10 [32] preprocessed through various standard feature maps. This conclusion seems robust and holds for different features of the raw data, such as random features [33] or the convolutional scattering transform [29, 34]. It also holds, as we prove, if we simply use a synthetic Gaussian mixture model, a classical model for complex multi-model data. The agreement between the real world and the asymptotic Gaussian theory is striking. While we may expect that such data could be approximated by a multimodal distribution such as a Gaussian mixture with enough modes, it should come as a rather puzzling fact that they lead to the same loss as a single Gaussian cloud. Our main contribution is to provide a rigorous theoretical foundation for these observations, that vindicates the classical line of works on Gaussian design, in particular the one stemming from statistical physics.

We list here our **main results**:

a) We provide a strong universality theorem for linear interpolators corresponding to ridgeless regression (with vanishing regularization) in high-dimensions and random labels, Theorem 5. Informally, we prove that a perceptron trained on randomly labelled Gaussian mixture data (a setting that encompasses complex multi-modal distributions) has the same minimum asymptotic loss as a perceptron trained on randomly labelled Gaussian data with isotropic covariance, that is $\mathcal{E}_\ell(\alpha) = {}^1\!/\!_2(1 - {}^1\!/\!_\alpha)_+$. This provides a theoretical explanation for the phenomena illustrated in Fig. 1 left.

b) Under an additional homogeneity assumption on the different modes of the data, Gaussian universality can be generalised to *any convex loss* (and we conjecture that it is valid for non-convex losses as well), Theorem 4. This provides a theoretical explanation of the phenomena illustrated in Fig. 1 (right).

c) At finite regularization and under the same homogeneity assumption, we show that the asymptotic training loss depends solely on the *data covariance matrix*, such that it is, again, the same loss as the one of a single Gaussian cloud with matching covariance, Theorem 3. This is illustrated in Fig. 2.

The proof technique used to establish these universality theorems has an interest on its own. It builds on recent progress in high-dimensional statistics and in mathematical insights drawn from the replica method in statistical physics. In particular, we provide an *explicit* matching of the expression (obtained from a rigorous proof of the replica prediction) for the asymptotic minimal loss [12, 35–37]. We further demonstrate the strong universality for ridge

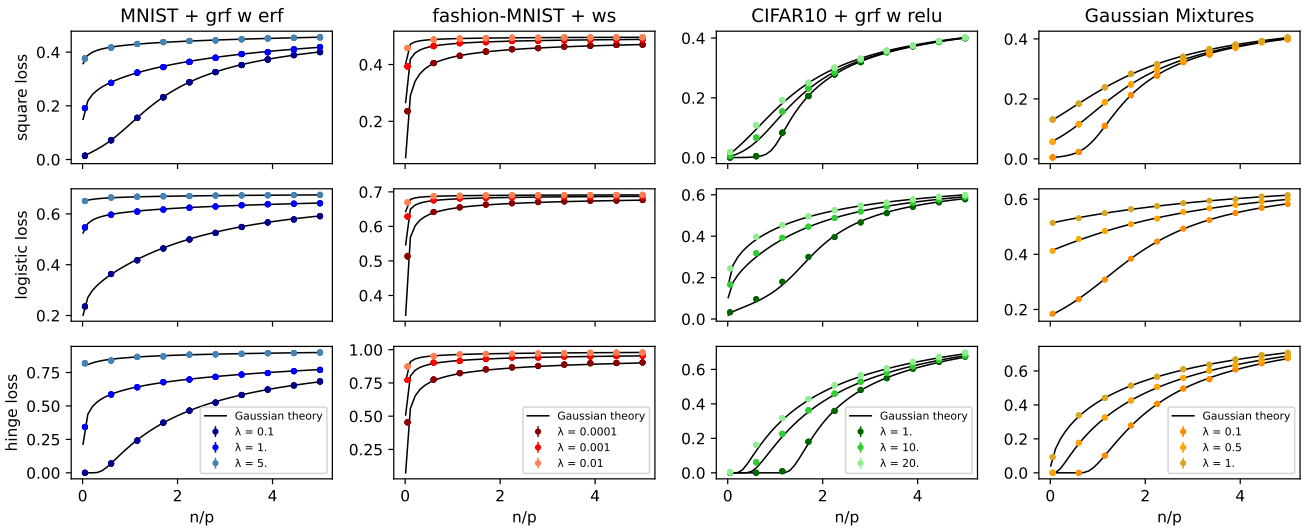

FIG. 2. This figure shows the training loss as a function of the number of samples n per dimension p at finite regularization λ. In the top panel the square loss, and in the bottom panel the hinge loss. The first column refers to MNIST with Gaussian random features and error function non-linearity, the second column corresponds to fashion-MNIST with wavelet scattering transform, the third column corresponds to CIFAR10 in grayscale with Gaussian random features and ReLU non-linearity, the fourth column corresponds to a mixture of Gaussians, with means $\boldsymbol{\mu}_\pm = (\pm 1, 0, ..., 0)$, covariances $\Sigma_\pm$ both equal to the identity matrix and relative class proportions $\rho_\pm = 1/2$. Black solid lines correspond to the outcome of the replica calculation, obtained by assigning to Σ the covariance matrix of each dataset plus the corresponding transformation. The coloured dots correspond to the simulations for different values of λ, as specified in the plot legend. Simulations are averaged over 10 samples & the error bars are not visible at the plot scale.

regression with vanishing regularization, again by showing explicitly that the exact solution [12, 38, 39] reduces to one of the homogeneous Gaussian cases. These results are obtained through methods that have been developed from statistical physics and mathematical physics-inspired techniques.

Further Related work

The perceptron — The question of how many samples can be perfectly fitted by a linear model is a classical one. For a ridge classifier, it amounts to asking whether a linear system of n equations with p unknowns is invertible so that for full-rank data the transition arises at $n = p$. For the $0/1$ loss or its convex surrogate such as the hinge loss, the question of linear separability was famously discussed by [22] who showed that for full-rank data the transition is given by $n = 2p$. In both cases, the transition is universal and does not depend on details of the data distribution (provided it is full rank, otherwise the rank replaces the dimension). For Gaussian data, such questions have witnessed a large amount of attention in the statistical physics community [1, 2, 13, 14, 40, 41] but also recently in theoretical computer science [15–19]. It is one of our goals to attract attention to these works, given the Gaussian universality we present shows that their relevance is not limited to idealistic Gaussian data.

Random Labels — Random labels are a fundamental and useful concept in machine learning. The pioneering work of [23], for instance, was instrumental in the modern critics of classical measures of model complexity, including the Rademacher complexity or the VC-dimension. These considerations have driven an entire line of research aiming to find substantial differences between learning with true and random labels, for instance in training time [42–44], in minima sharpness [45, 46] or in what neural networks can actually learn with random labels [24]. It has also been recently claimed [24] that pre-training on random labels under a given initial condition scaling can consistently speed up neural network training on both true and random labels, with respect to training from scratch.

Gaussian Universality — There has been much progress on a similar, though more restricted, Gaussian universality for random feature maps on Gaussian input data [33]. Following early insights by [47], the authors of [48, 49] showed that the empirical distribution of the Gram matrix of random features is asymptotically equivalent to a linear model with matched covariance. This was extended to generic convex losses by [50] using the heuristic replica method, and proven in [51]. A specific *Gaussian Equivalence Principle* [8] for learning with random features has been proven in a succession of works for convex penalties in [52, 53] and some non-convex ones in [54]. Early ideas on Gaussian universality have also appeared in the context of signal processing and compressed sensing in [55–59]. These theoretical

results, however, fall short when considering realistic datasets as we do in this work. Indeed, these previous works considered only uni-modal Gaussian data (observed through random feature maps), a situation far from realistic multi-modal, complex, real-world datasets. Instead, [9, 60, 61] argued that real datasets can be efficiently approximated in high dimensions by a finite *mixture of Gaussians*. These, of course, are multi-modal distributions that cannot be approximated by a single Gaussian. Gaussian mixtures will be the starting point of our theory.

Finally, we note that the observation that Gaussian data can fit or represent well some real data has been heuristically observed in many situations, but without theoretical justification and often limited to ridge regression, see e.g. [10, 12, 62–64].

II. SETTING, NOTATION, AND ASYMPTOTIC FORMULAS

The focus of the present work is the analysis of high-dimensional binary linear classification (aka perceptron) on a dataset $\mathcal{D} = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^n$. We shall consider a minimization problem of the form

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X}, \boldsymbol{y}) = \inf_{\boldsymbol{\theta}} \frac{1}{n} \sum_{\mu=1}^n \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_\mu, y_\mu) + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2, \tag{1}$$

where the $\boldsymbol{x}_\mu \in \mathbb{R}^p$ are input vectors, $y_\mu \in \{-1, +1\}$ are binary labels. We assume that the loss ℓ only depends on the inputs $\boldsymbol{x}_\mu$ through a one-dimensional projection $\boldsymbol{\theta}^\top \boldsymbol{x}_\mu$, and we work in the so-called *thermodynamic* or *proportional high-dimensional limit*, where n, p go to infinity with

$$\frac{n}{p} \to \alpha > 0.$$

In practice, practitioners seldom use the raw data $\mathbf{x}$ directly in their linear classifiers and usually perform a preprocessing step. For instance, instead of using the raw MNIST, a classical approach is to use a fixed feature map, and to observe the data as $\mathbf{x} = \sigma(F\mathbf{x})$, with F a random matrix. This is called the random feature map [33], and it has the advantage, among others, that the effective data $\mathbf{x}$ are full-rank. One may use more complicated such as the convolutional scattering transform [13, 29], or even pre-trained neural networks, a situation called transfer learning [65, 66]. We shall as well applied such transforms to our real data, in order to avoid theoretical pitfalls in direct space (in images some pixels are always zero for instance, so that the data may not even be full-rank). There is also one more advantage of using fixed features: it corresponds to deep learning (with actual multi-layer nets) in the so-called lazy regime [27, 28]. In this case, the feature matrix is a random matrix. Therefore, our results go beyond linear models and are also relevant to deep learning in the lazy regime. In our numerical experiments, we shall thus not only work with the original data (see appendix C, and in particular Fig.5), but also —and mainly— with random features maps and fixed features maps (as in Fig.1 and Fig.2)

For the labels, we shall focus in this work on the *random label* model, where the y_μ are independent of the inputs $\boldsymbol{x}_\mu$, and generated independently according to a Rademacher distribution:

$$y_\mu \sim \frac{1}{2} \left(\delta_{-1} + \delta_{+1} \right). \tag{2}$$

In our theoretical approach, we shall use mainly two data models:

• The simplest one is the **Gaussian covariate model** (GCM), where the inputs $\boldsymbol{x}_\mu \in \mathbb{R}^p$ are independently drawn from a Gaussian distribution:

$$\boldsymbol{x}_\mu \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3}$$

The Gaussian covariate model has been the subject of much attention recently [6, 12, 35, 38, 39, 49, 67–71]. In particular, the asymptotic statistics of the minimizer of eq. (1) for different models for the labels can be computed using the replica method, and rigorously proven as well. In particular, the random label limit relevant to our discussion can be obtained as a limit of the expressions derived using the replica method of statistical physics and mathematically proven in [12]. We shall use here the following expressions (see also Appendix A 3)

Theorem 1 (Asymptotics of the GCM for random labels, adapted from [12], informal)**.** *Consider the minimization problem in eq. (1), with the inputs $\boldsymbol{x}_\mu$ generated according to a Gaussian covariate model. Assume that the loss ℓ is strictly convex (or that ℓ is convex and $\lambda > 0$). Under mild regularity conditions on the $\boldsymbol{\mu}$, Σ, as well as the loss and*

regularizer, then we have the asymptotic training performance of the empirical risk minimizer eq. (A2) for the random label Gaussian mixture model satisfying the scalings (A4) in the proportional high-dimensional limit as $n \to \infty$:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X}, y(\boldsymbol{X})) \xrightarrow{\mathbb{P}} \mathcal{E}_\ell^{gcm}(\alpha, \lambda) := \frac{1}{2} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\ell(\mathrm{prox}_{V^\star \ell(\cdot,y)} \left(\sqrt{q^\star}\xi\right), y) \right] \tag{4}$$

where $\mathrm{prox}_{\tau f(\cdot)}$ *is the proximal operator associated with the loss:*

$$\mathrm{prox}_{\tau \ell(\cdot,y)}(x) := \arg\min_{z \in \mathcal{C}_1} \left[\frac{1}{2\tau}(z-x)^2 + \ell(z,y) \right] \tag{5}$$

and the parameters $(V^\star, q^\star)$ *are the (unique) fixed-point of the following self-consistent equations:*

$$\begin{cases} \hat{V} = \frac{\alpha}{2} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\partial_\omega f_\ell(y, \sqrt{q}\xi, V) \right] \\ \hat{q} = \frac{\alpha}{2} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[f_\ell(y, \sqrt{q}\xi, V)^2 \right] \end{cases}, \qquad \begin{cases} V &= \frac{1}{p} \operatorname{tr} \Sigma \left(\lambda \boldsymbol{I}_p + \hat{V}\Sigma \right)^{-1} \\ q &= \frac{1}{p} \hat{q} \operatorname{tr} \Sigma^2 \left(\lambda \boldsymbol{I}_p + \hat{V}\Sigma \right)^{-2} \end{cases} \tag{6}$$

where $f_g(y, \omega, V) := V^{-1} \left(\mathrm{prox}_{V\ell(\cdot,y)}(\omega) - \omega \right)$.

• A more generic model of data, which has the advantage of being multi-modal to be fit complex situations, is the **Gaussians Mixture Model** (GMM). In this case, the inputs $\boldsymbol{x}_\mu \in \mathbb{R}^p$ are independently generated as:

$$\boldsymbol{x}_\mu \sim \sum_{c \in \mathcal{C}} \rho_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{7}$$

where $\mathcal{C} := \{1, \cdots, K\}$ indexes the K Gaussian clouds and $\rho_c \in [0,1]$ is the density of points in each cloud and satisfies $\sum_{c \in \mathcal{C}} \rho_c = 1$. The analysis of Gaussian mixture models in the high-dimensional regime has been the subject of many works. The exact asymptotic expression for the minimum training loss has been derived for a range of particular cases in, between others, [72–77] and in full generality for arbitrary means and covariances in [37]. We shall thus use the random label limit of their expression in the binary classification case:

Theorem 2 (Asymptotics of the GMM for random labels, adapted from [37], informal)**.** *Consider the minimization problem in eq. (1), with the inputs* $\boldsymbol{x}_\mu$ *generated according to a Gaussian mixture as in (7). Assume that the loss* ℓ *is strictly convex (or that* ℓ *is convex and* $\lambda > 0$*). Under mild regularity conditions on the* $\boldsymbol{\mu}_c$*,* $\boldsymbol{\Sigma}_c$*, as well as the loss and regularizer, we have the training performance of the empirical risk minimizer eq. (A2) for the random label Gaussian mixture model satisfying the scalings (A4) are given by:*

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X}, y(\boldsymbol{X})) \xrightarrow{\mathbb{P}} \mathcal{E}_\ell^{gmm}(\alpha, \lambda, K) := \frac{1}{2} \sum_{c \in \mathcal{C}} \rho_c \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\ell(\mathrm{prox}_{V_c^\star \ell(\cdot,y)} \left(m_c^\star + \sqrt{q_c^\star}\xi \right), y) \right] \tag{8}$$

where ℓ *is the loss function used in the empirical risk minimization in eq. (A2),* $\mathrm{prox}_{\tau f(\cdot)}$ *is the proximal operator associated with the loss:*

$$\mathrm{prox}_{\tau \ell(\cdot,y)}(x) := \arg\min_{z \in \mathcal{C}_1} \left[\frac{1}{2\tau}(z-x)^2 + \ell(z,y) \right] \tag{9}$$

and $(m_c^\star, V_c^\star, q_c^\star)_{c \in \mathcal{C}}$ *are the* ***unique*** *fixed points of the following self-consistent equations:*

$$\begin{cases} \hat{V}_c = \frac{\alpha}{2} \rho_c \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\partial_\omega f_\ell(y, m_c + \sqrt{q_c}\xi, V_c) \right] \\ \hat{q}_c = \frac{\alpha}{2} p_c \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[f_\ell(y, m_c + \sqrt{q_c}\xi, V_c)^2 \right] \\ \hat{m}_c = \frac{\alpha}{2} p_c \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[f_\ell(y, m_c + \sqrt{q_c}\xi, V_c) \right] \end{cases}$$

$$\begin{cases} V_c &= \frac{1}{p} \operatorname{tr} \Sigma_c \left(\lambda \boldsymbol{I}_p + \sum_{c' \in \mathcal{C}} \hat{V}_{c'} \Sigma_{c'} \right)^{-1} \\ q_c &= \frac{1}{p} \sum_{c' \in \mathcal{C}} \left[\operatorname{tr} \left(\hat{q}_{c'} \Sigma_{c'} + \hat{m}_c \hat{m}_{c'} \boldsymbol{\mu}_{c'} \boldsymbol{\mu}_c^\top \right) \Sigma_c \left(\lambda \boldsymbol{I}_p + \sum_{c'' \in \mathcal{C}} \hat{V}_{c''} \Sigma_{c''} \right)^{-2} \right] \\ m_c &= \frac{1}{p} \sum_{c' \in \mathcal{C}} \hat{m}_c \hat{m}_{c'} \left[\operatorname{tr} \boldsymbol{\mu}_{c'} \boldsymbol{\mu}_c^\top \left(\lambda \boldsymbol{I}_p + \sum_{c'' \in \mathcal{C}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \right] \end{cases} \tag{10}$$

where $f_\ell(y, \omega, V) := V^{-1} \left(\mathrm{prox}_{V\ell(\cdot,y)}(\omega) - \omega \right)$.

III. THE MAIN THEORETICAL RESULTS: FROM MIXTURES TO A SINGLE GAUSSIAN

In this section, we present the main theoretical results of the present work and discuss their consequences: We show that with random labels, GMM models can be reduced to a single GCM model. This provides an explanation of the universality observed in Figs. 1 and 2.

We would like the reader to note that we state our results using theorems because indeed we were able to establish them mathematically rigorously. However, the proofs are deferred to the appendices and the reasoning and derivations presented in this section follow the level of rigour common in theoretical physics. We made this choice to ensure readability to both physics and mathematics-oriented audiences.

The starting point is the Gaussian Mixture Model (GMM). This is a very generic model of data and standard approximation results (e.g. the Stone-Weierstrass theorem) show in particular that one can approximate data density to arbitrary precision by Gaussian mixtures. While, in the worst case this would require a diverging number of Gaussian in the mixture, it can be shown that (as far as the generalized linear model is concerned) a mixture of a small number of Gaussians is actually able to approximate very complex data set in high-dimension [9, 60, 78]. More precisely, in the proportional high-dimensional regime, data generated by Generative Adversarial Networks (GAN), one of the state of the art techniques to generate realistic looking data, behave as Gaussian mixtures for such classifiers [61]. We shall thus use this model as our benchmark of "complex" data distribution.

If a mixture model is a good approximation of reality in high-dimension, the question remains: *why is it that we can fit real dataset with a single Gaussian.* Our main technical question will thus be: if we use random labels, what is the difference between a GMM and a single Gaussian model?

A. Mean invariance with random labels

We thus now move to the random label case and show how we can surprisingly use a simple Gaussian distribution instead of the GMM. We are going to use theorem 1 and 2. Note that the asymptotic value of the energy, or loss, only depends on the probability vector $\boldsymbol{\rho} \in [0,1]^K$ (with entries ρ_c corresponding to the respective sizes of the K clusters), the matrix of averages $\boldsymbol{M} \in \mathbb{R}^{K\times p}$ (with rows $\boldsymbol{\mu_c} \in \mathbb{R}^p$), and the concatenation of covariances $\boldsymbol{\Sigma}^{\otimes} \in \mathbb{R}^{K\times p\times p}$ (with rows $\boldsymbol{\Sigma}_c \in \mathbb{R}^{p\times p}$) and therefore we denote:

$$\mathcal{E}_\ell = \mathcal{E}_\ell^{\text{gmm}}(\boldsymbol{\rho}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\otimes}).$$

Similarly, for the Gaussian covariate model we define the limiting value

$$\mathcal{E}_\ell = \mathcal{E}_\ell^{\text{gcm}}(\boldsymbol{m}, \boldsymbol{\Sigma}).$$

where in both cases we omitted the explicit dependence on the parameters (α, λ). We are now in a position to state a lemma crucial to our first main universality result:

Lemma 1 (Single mean lemma for random labels)**.** *In the random label setting (2), assume that the loss ℓ is symmetric, in the sense that $\ell(x,y) = \ell(-x,-y)$ for $x, y \in \mathbb{R}$. Then, the limiting value $\mathcal{E}_\ell$ of the risk is independent from the means, i.e. for all choices of $\boldsymbol{\rho}$, $\boldsymbol{M}$ and $\boldsymbol{\Sigma}^{\otimes}$ we have*

$$\mathcal{E}_\ell^{gmm}(\boldsymbol{\rho}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\otimes}) = \mathcal{E}_\ell^{gmm}(\boldsymbol{\rho}, \boldsymbol{0}, \boldsymbol{\Sigma}^{\otimes})\,.$$

The symmetry condition on the loss is not really restrictive and is satisfied by virtually all losses used in binary classification (in particular margin-based losses of the form $\ell(x,y) = \phi(xy)$). Since a mixture of Gaussians with equal means and covariances is equivalent to a single Gaussian, we can now write the following theorem:

Theorem 3 (Gaussian universality for random labels)**.** *Consider the same assumptions as in Lemma 1, and assume further that the data is homogeneous, i.e.*

$$\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma} \quad \textit{for all} \quad c \in \mathcal{C}.$$

Then the asymptotic risk is equivalent to that of a single centered Gaussian:

$$\mathcal{E}_\ell^{gmm}(\boldsymbol{\rho}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\otimes}) = \mathcal{E}_\ell^{gcm}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

This is our first main universality theorem: a mixture of homogeneous Gaussians [79] can be replaced, when using random labels by a single Gaussian.

This surprising fact, alone, explains the empirical observation presented in Fig. 1 and Fig. 2, at least if we accept that the different modes are homogeneous (see discussion in Sec.IV).

Proof sketch — Both lemma 1 and Theorem 3 stem from the detailed analysis of the replica free energy for the GMM [37]. Indeed, to prove our claims, it suffices to show that the fixed points of the replica equations are the same. This is done in detail in appendix B, using the replica equation that we provide in appendix A. In a nutshell, we show that the expression of the GMM reduces to those of the GCM. □

B. Generic loss with vanishing regularisation

Additionally, we note that in Fig. 1 at vanishing regularization, we did not even require a matching covariance, and instead used a trivial one. This is because of the following consequence of Lemma 1:

Theorem 4 (Gaussian universality for vanishing regularization)**.** *Consider the same assumptions as in theorem 3, then if the minimizer of ℓ is unique and the data covariance full-rank, then the asymptotic minimal loss for Gaussian data does not depend on the covariance when the regularization is absent, $\lambda = 0$.*

Proof. Consider the empirical risk minimization problem in eq. (1) with data from the Gaussian covariate model eq. (3) with random labels. Without loss of generality, we can write $\boldsymbol{x}_\mu = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_\mu$, with $\boldsymbol{z}_\mu \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$. Then, making a change of variables $\boldsymbol{\theta}' = \boldsymbol{\Sigma}^{\mathbf{1/2}}\boldsymbol{\theta}$, we can write:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X}, \boldsymbol{y}) = \inf_{\boldsymbol{\theta}} \frac{1}{n}\sum_{\mu=1}^{n} \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_\mu, y_\mu) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 = \inf_{\boldsymbol{\theta}' \in \mathcal{S}_p'} \frac{1}{n}\sum_{\mu=1}^{n} \ell(\boldsymbol{\theta}'^\top \boldsymbol{z}_\mu, y_\mu) + \frac{\lambda}{2}||\Sigma^{-1/2}\boldsymbol{\theta}'||_2^2$$

where $\mathcal{S}_p' \subset \mathbb{R}^p$ is another compact set, and we have used the fact that y_μ are independent of $\boldsymbol{x}_\mu$. Since the minimizer of ℓ is unique, the result follows from taking $\lambda \to 0^+$. □

Note that in particular theorem 4 implies that for random labels, the GCM model with a covariance Σ is equivalent to a Gaussian i.i.d. model with a different regularization given by the norm $||\cdot||_{\Sigma^{-1}}$ induced by the inverse covariance matrix Σ^{-1}. Therefore, in the case in which ℓ has several minima, the $\lambda \to 0^+$ limit will give the performance of the solution with minimum $||\cdot||_{\Sigma^{-1}}$ norm.

Finally, we also note that this analysis also allows answering the important question: *what is being learned with random labels?*, discussed in particular in machine learning literature in [24]. For generalized linear models: the model is simply fitting the 2nd-order statistics (the total covariance $\boldsymbol{\Sigma}$).

C. Ridge regression with vanishing regularization

Even though it seems to be well obeyed in practice, one may wonder if we can in some cases get rid of the homogeneity condition. As we shall see, the answer is no: in general, a mixture of *inhomogeneous* Gaussian cannot be strictly replaced by a single one. It turns out, however, that there is one exception, and that the hypothesis can be lifted in one case, ridge regression with vanishing regularization with the squared loss $\ell(x, y) = \frac{1}{2}(x - y)^2$:

Theorem 5 (Strong universality for ridge loss)**.** *In the ridge regression case with vanishing regularization, i.e. when $\lambda \to 0^+$, we have*

$$\lim_{\lambda \to 0^+} \mathcal{E}_\ell^{gmm}(\boldsymbol{\rho}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\otimes}) = \frac{1}{2}\left(1 - \frac{1}{\alpha}\right)_+ ,$$

for any choice of $\boldsymbol{\rho}, \boldsymbol{M}$, or $\boldsymbol{\Sigma}^{\otimes}$.

In particular, it means that in the unregularized limit, any Gaussian mixture behaves in terms of its loss as a single cluster Gaussian model with identity covariance, whose asymptotic training loss is given by $\lim_{\lambda \to 0^+} \mathcal{E}_\ell^{\text{gcm}}(\alpha, \lambda) = {}^1\!/_2(1 - 1/\alpha)_+$.

Proof sketch — The proof of the strong universality, that follows from a rigorous analysis of the replica predictions, amounts to showing that the replica free energy for GMM reduces to the one of a single Gaussian. Interestingly, although the fixed points of the replica equations differ between the GMM and Gaussian case, they do give rise to the same free energy. Details can, again, be found in Appendix B 2. □

IV. NUMERICAL EXPERIMENTS

In this section, we describe more in detail the numerical experiments of Fig. 1 & Fig. 2. The coloured dots represent the outcome of the simulations on several full-rank datasets. In particular, the blue and the green dots refer to both MNIST and grayscale CIFAR-10 preprocessed with random Gaussian feature maps [33]. In this case, the input data points are constructed as $\boldsymbol{x}_\mu = \sigma\left(\boldsymbol{z}_\mu \boldsymbol{F}\right)$, with $\boldsymbol{z}_\mu \in \mathbb{R}^d$ being a sample from one of the two datasets, $\boldsymbol{F} \in \mathbb{R}^{d\times p}$ representing the matrix of random features, whose row elements are sampled according to a normal distribution, and σ being some point-wise non-linearity, namely erf for MNIST and relu for grayscale CIFAR-10. The red dots correspond instead to fashion-MNIST pre-processed with wavelet scattering transform, an ensemble of engineered features producing rotational and translational invariant representations of the input data points [29]. The orange dots correspond to simulations on the synthetic dataset built as a mixture of two Gaussians, with data covariance of the two clusters both equal to the identity matrix ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$, $\boldsymbol{\mu}_{1/2} = (\pm 1, 0, ..., 0)$ and $\rho_{1/2} = 1/2$. Further technical details are given in the appendix.

a. Experiments with finite regularization — Fig. 2 illustrates the Gaussian universality taking place at finite regularization. The coloured dots correspond to the outcome of the simulations for several values of the regularization strength. As we can see from this set of plots, the theoretical learning curve of a single Gaussian with matching covariance perfectly fits the behaviour of multi-modal and more realistic input data distributions. In fact, even though the experiment is performed for a realistic dataset and finite n and p, the asymptotic Gaussian theory gives a perfect fit of the data.

b. Experiments with vanishing regularization — Fig. 1 provides an illustration of the universality effect occurring at vanishing regularization. Here we use $\lambda \to 0$, and following theorem 4, we observe a collapse on a single curve given by the asymptotic theory for a single Gaussian with unit covariance. It is quite remarkable that our asymptotic theory, which is valid only in the infinite high-dimensional limit, is validated by such experiments with finite dimension, and finite sample size.

c. Homogeneity assumption — A remarkable point is that the homogeneity assumption (often called homoskedasticity in statistics) we use in theorem 3, which can be relaxed only for ridge regression, does not seem to be that important in practice, as we observed on our experiments on real data. One may thus wonder if the strong universality of Theorem 5 could be proved in full generality, and not only for the ridge loss. It turns out that the answer is no. Using Proposition 2, we can actually construct an artificial mixture of Gaussians, using *very different* covariances for each individual Gaussians, and observe small deviations from the strict universality. A mixture of *non-homogeneous* Gaussians is not strictly equivalent to a single one with random labels (except, as stated in Theorem 5, for the least squares that obey a strong universality). This is illustrated in Fig. 3 where we show the disagreement in the behaviour of the training loss between a single Gaussian and a mixture of two non-homogeneous Gaussians. This is a simple counter-example to the existence of a universal strong form of Gaussian universality, even for ridge regression (see discussions in e.g. [10, 52, 64, 80, 81]).

It may thus come as surprising that real datasets, which certainly will not obey such a strict homogeneity of the different modes, display such a spectacular agreement with the theory. We believe that this is due to two effects: first, the deviations we observed, even in our designed counter-example, are small, so they might not even be seen in practice. Secondly, and especially after observing the data through random or scattering features, it turns out that when we measure the empirical correlation matrix of the different modes, they look quite similar. In fact, it has even been suggested that neural networks are *precisely* learning representations that find such homogeneous Gaussian mixtures [82].

d. A remark on Rademacher complexity — A final comment is that the discussed universality indicates that, in high dimension, the Rademacher complexity can be effectively replaced by the one for Gaussian i.i.d. data. *Rademacher complexity* is a key quantity appearing in generalization bounds for binary classification problems [7, 20] that measures the ability of estimators in a hypothesis class $\mathcal{H}$ to fit i.i.d. random labels $y^\mu \sim \mathrm{Rad}(1/2)$:

$$\mathrm{Rad}_n\left(\mathcal{H}\right) = \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n} y_\mu h\left(\boldsymbol{x}_\mu\right)\right]. \tag{11}$$

It is explicitly dependent on the specific distribution of the input data points $\boldsymbol{x}_\mu$. As discussed in [17] there exists a direct mapping between the Rademacher complexity and the minimum 0/1 training loss - or ground state energy in the statistical physics parlance. Indeed, for a binary hypothesis class $\mathcal{H} = \{h : \mathbb{R}^p \to \{-1, +1\}\}$ the two are asymptotically related by the following equation:

$$\lim_{n\to\infty} \inf_{h\in\mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n} \mathbb{P}\left(h(\boldsymbol{x}_{\boldsymbol{\mu}}) \neq y_\mu\right) = \frac{\alpha}{2}\left[1 - \mathrm{Rad}_n\left(\mathcal{H}\right)\right]. \tag{12}$$

Moreover, [17] discussed how to explicitly compute the Rademacher complexity for Gaussian data using the replica method from statistical physics. This is actually a classical problem, studied by the pioneers of the application of the

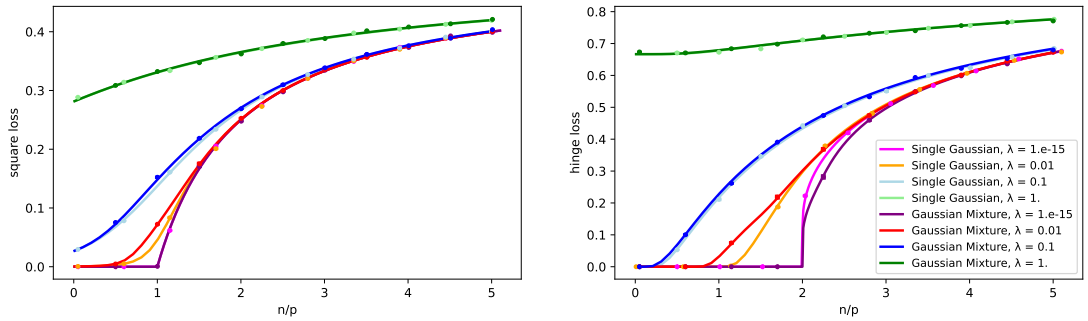

FIG. 3. Ridge/square loss (left) & hinge loss for a single Gaussian vs a mixture of inhomogeneous Gaussians at finite λ. Lines are the asymptotic exact results while dots are simulation ($p = 900$, dark lines for mixture, lighter ones for single Gaussian). When the homogeneity assumption is not obeyed, then a mixture of two Gaussians does not yield equal results to those of a single Gaussian with matching covariance. (Here, a mixture with zero mean and a block covariance with, resp. diagonal elements equal to 0.01, 0.98 and 0.01 for the first one, and 0.495 and 0.01, 0.495 for the second). Note however that the universality is restored in the Ridge case when $\lambda \to 0$, as stated in Theorem 5. It is also very well obeyed with large enough λ and deviations appear small in general.

replica method and spin glass theory to theoretical machine learning [1, 2, 40, 41]. Given the universality advocated in this present work, these Gaussian results thus seem to be of more relevance than previously thought, and in fact, allow us to compute a closed-form asymptotic expression for the Rademacher complexity for realistic data. This is a very interesting outcome of the Gaussian universality with random labels.

However, while we prove universality for convex losses, we so far only *conjecture it* for non-convex objectives, such as the ones appearing in the definition of the Rademacher complexity. The proof that a Gaussian mixture approximates well real datasets is still valid for non-convex losses. The identification of these mixtures to a single Gaussian is, however, using the replica formulas of [12, 37] which have been proven only for the case of convex losses. Our conjecture thus depends on proving a similar result for non-convex (as well as replica symmetry breaking) losses. This (and similar questions on multi-layer networks) is left for future work.

V. CONCLUSION

For the classical problem of fitting random labels with perceptrons aka generalized linear models in high dimensions, we showed that, far from being only a toy example, the Gaussian i.i.d. assumption is an excellent model of reality. The conclusion extends to deep-learning models in the lazy regimes as those are essentially random feature models. There are a number of potentially interesting extensions of this work, including non-convex losses and multi-layer neural networks, and beyond the random label cases, that should be investigated in the future.

These results, we believe, are of special interest given the number of theoretical studies with the Gaussian design and its variants, that are amenable to exact characterization, and that turn out to be less idealistic, and more realistic, than perhaps previously assumed. We believe, in particular, that these considerably strengthen the ensemble of results obtained within the statistical physics community, as well as in the statistical analysis of high-dimensional data. We anticipate that such redemption of the Gaussian assumption will lead to more work in this direction both those using Gaussian assumption and those aiming to extend out universality results.

ACKNOWLEDGEMENTS

We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe, as well as by the Swiss National Science Foundation grant SNFS OperaGOST,

200021_200390 and the *Choose France - CNRS AI Rising Talents* program.

[1] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

[2] Werner Krauth and Marc Mézard. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066, 1989.

[3] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[4] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[5] Emmanuel J Candès, Pragya Sur, et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[6] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[7] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[8] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. *Physical Review X*, 10:041044, 2019.

[9] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 8573–8582. JMLR.org, July 2020.

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

[11] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

[12] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] N Brunel, J-P Nadal, and G Toulouse. Information capacity of a perceptron. *Journal of Physics A: Mathematical and General*, 25(19):5017, 1992.

[14] Silvio Franz, Giorgio Parisi, Maxime Sevelev, Pierfrancesco Urbani, and Francesco Zamponi. Universality of the sat-unsat (jamming) threshold in non-convex continuous constraint satisfaction problems. *SciPost Physics*, 2(3):019, 2017.

[15] Jian Ding and Nike Sun. Capacity lower bound for the ising perceptron. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 816–827, 2019.

[16] Benjamin Aubin, Will Perkins, and Lenka Zdeborová. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 52(29):294003, 2019.

[17] Alia Abbara, Benjamin Aubin, Florent Krzakala, and Lenka Zdeborová. Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning. In *Mathematical and Scientific Machine Learning*, pages 27–54. PMLR, 2020.

[18] Andrea Montanari, Yiqiao Zhong, and Kangjie Zhou. Tractability from overparametrization: The example of the negative perceptron. *arXiv preprint arXiv:2110.15824*, 2021.

[19] Ahmed El Alaoui and Mark Sellke. Algorithmic pure states for the negative spherical perceptron. *arXiv preprint arXiv:2010.15811*, 2020.

[20] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[21] James G Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

[22] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[23] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[24] Hartmut Maennel, Ibrahim Alabdulmohsin, Ilya Tolstikhin, Robert JN Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? *arXiv preprint arXiv:2006.10455*, 2020.

[25] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[26] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[27] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[29] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[30] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[31] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[32] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[33] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

[34] Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, et al. Kymatio: Scattering transforms in python. *J. Mach. Learn. Res.*, 21(60):1–6, 2020.

[35] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[36] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[37] Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10144–10157. Curran Associates, Inc., 2021.

[38] Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[39] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[40] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[41] Bernard Derrida, RB Griffiths, and A Prugel-Bennett. Finite-size effects and bounds for perceptron models. *Journal of Physics A: Mathematical and General*, 24(20):4907, 1991.

[42] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[43] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018.

[44] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[45] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[46] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

[47] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[48] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

[49] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

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

[51] Oussama Dhifallah and Yue M. Lu. Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy*, 23(4):400, 2021.

[52] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 16–19 Aug 2022.

[53] Hong Hu and Yue M. Lu. Universality Laws for High-Dimensional Learning with Random Features. *arXiv:2009.07669 [cs, math]*, March 2021. arXiv: 2009.07669.

[54] Andrea Montanari and Basil Saeed. Universality of empirical risk minimization. *arXiv preprint arXiv:2202.08832*, 2022.

[55] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[56] Andrea Montanari and Phan-Minh Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342, 2017.

[57] Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE Transactions on Information Theory*, 57(4):2440–2450, 2011.

[58] Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[59] Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[60] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[61] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborova. Universality laws for gaussian mixtures in generalized linear models. *arXiv:2302.08933*, 2023.

[62] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

[63] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

[64] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks, 2022.

[65] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[66] Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 3(1):015030, 2022.

[67] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[68] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020.

[69] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

[70] Benjamin Aubin, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[71] Bruno Loureiro, Cédric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension, 2022.

[72] Ganesh Ramachandra Kini and Christos Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4020–4024. IEEE, 2021.

[73] Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. Phase transition in the hard-margin support vector machines. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 415–419. IEEE, 2019.

[74] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

[75] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.

[76] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Optimality of least-squares for classification in gaussian-mixture models. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2515–2520. IEEE, 2020.

[77] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[78] Mohamed El Amine Seddik, Cosme Louart, Romain Couillet, and Mohamed Tamaazousti. The unexpected deterministic and universal behavior of large softmax classifiers. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2021.

[79] Also called homoskedastic Gaussians, as opposed to heteroskedastic Gaussians.

[80] Umberto M Tomasini, Antonio Sclocchi, and Matthieu Wyart. Failure and success of the spectral bias prediction for kernel ridge regression: the case of low-dimensional data. *arXiv preprint arXiv:2202.03348*, 2022.

[81] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation, 2022.

[82] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[83] N Barkai and Haim Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Physical Review E*, 50(3):1766, 1994.

[84] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608. IEEE, 2016.

[85] Marc Lelarge and Léo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643. IEEE, 2019.

[86] Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4030–4034. IEEE, 2021.

[87] The general case is given by changing p for the rank of the design matrix.

[88] As discussed in Theorem 4, the fact that the loss is independent of Σ in this regime can be directly seen from the optimization.

[89] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database.

[90] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[91] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[92] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[93] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[94] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Appendix A: Exact asymptotic performances of GCM and GMM

In this appendix we summarize the exact asymptotic formulas for the performance of the generalized linear classifiers on random labels for the two structured data models studied in the the main body: the Gaussian covariate model (GCM) and the Gaussian mixture model (GMM).

1. Preliminaries: the setting

Before moving to the key formulas, let us recap the setting. We are interested in the performance of generalised linear classifiers:

$$\hat{y}(\boldsymbol{x}) = \text{sign}(\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}) \tag{A1}$$

where $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ is trained by minimising the following empirical risk on n independent training samples $(\boldsymbol{x}_\mu, y_\mu)_{\mu\in[n]} \in \mathbb{R}^p \times \{-1,+1\}$:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}) = \inf_{\boldsymbol{\theta}\in\mathcal{S}_p} \frac{1}{n}\sum_{\mu=1}^n \ell(\boldsymbol{\theta}^\top \boldsymbol{x}_\mu, y_\mu) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2, \tag{A2}$$

for a compact subset $\mathcal{S}_p \subset \mathbb{R}^p$ and convex loss function ℓ. In particular, we are interested in the case where the labels $y_\mu \in \{-1,+1\}$ are randomized (i.e. not correlated with the inputs $\boldsymbol{x}_\mu$),

$$y_\mu \sim \frac{1}{2}\left(\delta_{-1} + \delta_{+1}\right), \qquad \text{i.i.d.} \tag{A3}$$

and the inputs are generated independently from one of the following two structured models:

Gaussian covariate model (GCM): : $\boldsymbol{x}_\mu \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma})$,

Gaussian mixture model (GMM): : $\boldsymbol{x}_\mu \sim \sum_{c\in\mathcal{C}} \rho_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$,

where $\mathcal{C} = \{1,\cdots,K\}$ is the label set for the Gaussian clouds and $\rho_c \in [0,1]$ are the density of points in each class, satisfying $\sum_{c\in\mathcal{C}} \rho_c = 1$. Note that in this random label setting the GCM model is a special case of the GMM, where $K := |\mathcal{C}| = 1$ and $\boldsymbol{\mu}_1 = \boldsymbol{0}_p$.

In the following, we will be interested in describing the exact asymptotic limit of the following performance metrics in the proportional high-dimensional limit where $n, p \to \infty$ with the ratio $\alpha := n/p$ and the number of clusters K are fixed:

Training loss: : $\hat{\mathcal{E}}_\ell(\boldsymbol{X},\boldsymbol{y}) := \frac{1}{n}\sum_{\mu=1}^n \ell\left(\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}_\mu, y_\mu\right)$

0/1 training error: : $\hat{\mathcal{E}}_{0/1}(\boldsymbol{X},\boldsymbol{y}) := \frac{1}{n}\sum_{\mu=1}^n \mathbb{P}\left(\text{sign}(\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}_\mu) \neq y_\mu\right)$

where we have defined the design matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ and the label vector $\boldsymbol{y} \in \{-1,+1\}^n$. Note that for convenience we will focus the discussion in this appendix to these two measures. But all results could have been stated for $\hat{\mathcal{R}}_n^\star$ instead. In particular, the training loss $\hat{\mathcal{E}}_\ell$ differs from the empirical risk $\hat{\mathcal{R}}_n^\star$ by the regularisation term.

a. Note on scalings – Although the model above is well defined for any scaling, in the following we focus in the case in which the means are covariance satisfy:

$$||\boldsymbol{\mu}_c||_2^2 = O(1), \qquad \text{tr}\,\boldsymbol{\Sigma}_c = O(p). \tag{A4}$$

This scaling of the mean and variance is indeed the natural one (see e.g. [75, 83–86]) as well as the most interesting in high-dimensions. If the means have larger norm, then the problem becomes trivial (i.e. the Gaussians are trivially completely separable) while if the means are smaller it is impossible to separate them (i.e. they become trivially indistinguishable from a single Gaussian cloud).

b. Ridge and ordinary least-squares classification – Note that for the special case of the ridge classification in which $\ell(x,y) = 1/2(y-x)^2$, the empirical risk minimization problem defined in eq. (A2) admits a closed form solution:

$$\hat{\boldsymbol{\theta}} = \left(\lambda \boldsymbol{I}_p + \boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{X}\boldsymbol{y} \tag{A5}$$

and therefore the computation of the asymptotic training error or loss boils down to a Random Matrix Theory problem, with a solution equivalent to the one we will discuss shortly below. However, some qualitative features can be drawn just from this expression. First, note that for $\lambda > 0$, the ridge estimator above will always have a non-zero training loss because of the bias introduced by the regularization term $1/2\lambda||\boldsymbol{\theta}||_2^2$. This can only be achieved in the limit of vanishing regularization $\lambda \to 0^+$, in which case the ridge estimator simplifies to:

$$\hat{\boldsymbol{\theta}}_{\text{ols}} \coloneqq (\boldsymbol{X}^\top)^\dagger \boldsymbol{y} \tag{A6}$$

where $\boldsymbol{X}^\dagger \in \mathbb{R}^{n\times p}$ is the Moore-Penrose inverse of $\boldsymbol{X}$. In the simplest case in which $\boldsymbol{X}$ is a full-rank matrix (which ultimately depends on the covariances), it can be explicitly written as:

$$\boldsymbol{X}^\dagger \coloneqq \begin{cases} (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top & \text{if } \alpha < 1 \\ \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1} & \text{if } \alpha > 1 \end{cases} \tag{A7}$$

An important property of the estimator in eq. (A6) is that it corresponds to the least ℓ_2-norm interpolator when the system is underdetermined. Indeed, in the strict case when $\lambda = 0$ (i.e. least-squares regression) the ERM problem in eq. (A2) is equivalent to inverting a linear system:

$$\boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{\theta} \tag{A8}$$

i.e. to solve a system of n equations for p unknowns. Again, assuming the data is full-rank[87], for $\alpha = n/p < 1$ the system is *underdetermined*, meaning that there are infinitely many solutions that perfectly interpolate the data. Among all of them, $\hat{\boldsymbol{\theta}}_{\text{ols}}$ is the one that has lowest ℓ_2-norm. Instead, when $\alpha > 1$, the system is overdetermined, and no interpolating (zero-loss) solution exists.

2. Gaussian mixture model with general labels

Exact asymptotics of generalized linear classification with Gaussian Mixtures in the proportional regime have been derived under different settings in the literature [72–77]. Of particular interest to our work are the formulas proved in [37] under the most general setting of a multi-class learning problem with convex losses & penalties and generic means and covariances. In their work, the asymptotic performance of the minimiser in eq. (A2) was proven in the case where the labels are correlated to the mean. The formula we state in the text as theorem 2 is a straightforward adaptation of their result in the particular case of binary classification with K clusters and randomized labels.

a. Zero mean limit: Of particular interest for what follows is the zero-mean limit $\boldsymbol{\mu_c} = \boldsymbol{0}_p$ of the above equations, which is simply given by:

$$\begin{cases} \hat{m}_c = 0 \\ \hat{V}_c = \frac{\alpha}{2}\rho_c \sum\limits_{y\in\{-1,+1\}} \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\partial_\omega f_\ell(y, \sqrt{q_c}\xi, V_c)\right] \\ \hat{q}_c = \frac{\alpha}{2}p_c \sum\limits_{y\in\{-1,+1\}} \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[f_\ell(y, \sqrt{q_c}\xi, V_c)^2\right] \end{cases}$$
$$\begin{cases} m_c = 0 \\ V_c = \frac{1}{p}\operatorname{tr}\Sigma_c \left(\lambda \boldsymbol{I}_p + \sum\limits_{c'\in\mathcal{C}} \hat{V}_{c'}\Sigma_{c'}\right)^{-1} \\ q_c = \frac{1}{p}\sum\limits_{c'\in\mathcal{C}} \left[\hat{q}_{c'}\operatorname{tr}\Sigma_{c'}\Sigma_c \left(\lambda \boldsymbol{I}_p + \sum\limits_{c''\in\mathcal{C}} \hat{V}_{c''}\Sigma_{c''}\right)^{-2}\right] \end{cases} \tag{A9}$$

b. A particular case: ridge classification – The self-consistent equations above crucially depend on the loss function ℓ. A case particular case of interest in this work - and for which the expressions considerable simply - is the case of ridge regression where $\ell(x,y) = 1/2(x-y)^2$. In this case, the proximal can be explicitly writen as:

$$\operatorname{prox}_{\tau\ell(\cdot,y)}(x) = \frac{x+\tau y}{1+\tau} \quad \Leftrightarrow \quad f_\ell(y,\omega,V) = \frac{y-\omega}{1+V} \tag{A10}$$

and therefore the asymptotic training loss admits a closed-form expression:

$$\mathcal{E}_\ell^{\text{gmm}} = \sum_{c\in\mathcal{C}} \rho_c \frac{1+q_c^\star}{2(1+V_c^\star)^2} \tag{A11}$$

for $(V_c^\star, q_c^\star)_{c\in\mathcal{C}}$ solutions of the following simplified self-consistent equations:

$$\begin{cases} \hat{V}_c = \frac{\alpha\rho_c}{1+V_c} \\ \hat{q}_c = \alpha p_c \frac{1+q_c}{(1+V_c)^2} \end{cases}, \qquad \begin{cases} V_c &= \frac{1}{p}\operatorname{tr}\Sigma_c \left(\lambda \boldsymbol{I}_p + \sum_{c'\in\mathcal{C}} \hat{V}_{c'}\Sigma_{c'}\right)^{-1} \\ q_c &= \frac{1}{p}\sum_{c'\in\mathcal{C}} \left[\hat{q}_{c'} \operatorname{tr}\Sigma_{c'}\Sigma_c \left(\lambda \boldsymbol{I}_p + \sum_{c''\in\mathcal{C}} \hat{V}_{c''}\Sigma_{c''}\right)^{-2}\right] \end{cases} \tag{A12}$$

Note that in particular, at the fixed point, we can also express the training loss eq. (A11) as:

$$\mathcal{E}_\ell^{\text{gmm}} = \sum_{c\in\mathcal{C}} \frac{\hat{q}_c^*}{2\alpha}. \tag{A13}$$

3. Gaussian covariate model

The asymptotic training loss for the Gaussian covariate model for a fairly general teacher-student setting was first proven in [12]. Although the random label limit can be obtained from this work, as discussed in Sec. A 1 the random label Gaussian covariate model can also be seen as a particular case of the general Gaussian mixture model with $K = 1$ and $\boldsymbol{\mu}_1 = \boldsymbol{0}_p$. Therefore, its asymptotic performance is included in the discussion above. This lead to theorem 1 in the main text.

It is worth noting that, for the square loss the expressions simplify considerably. The training loss is given by:

$$\mathcal{E}_\ell^{\text{gmm}} = \frac{1+q^\star}{2(1+V^\star)^2} \tag{A14}$$

where $(V^\star, q^\star)$ are solutions of the following simplified self-consistent equations:

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{1+q}{(1+V)^2} \end{cases}, \qquad \begin{cases} V &= \frac{1}{p}\operatorname{tr}\Sigma \left(\lambda \boldsymbol{I}_p + \hat{V}\Sigma\right)^{-1} \\ q &= \frac{1}{p}\hat{q}\operatorname{tr}\Sigma^2 \left(\lambda \boldsymbol{I}_p + \hat{V}\Sigma\right)^{-2} \end{cases} \tag{A15}$$

Since the covariance Σ is positive-definite (and therefore invertible), in the overdetermined regime (for which the training loss is non-zero), the limit $\lambda \to 0^+$ can be easily taken, and the equations reduce to:

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{1+q}{(1+V)^2} \end{cases}, \qquad \begin{cases} V &= \frac{1}{\hat{V}} \\ q &= \frac{\hat{q}}{\hat{V}} \end{cases} \tag{A16}$$

which is completely independent of the covariance matrix Σ[88]. Moreover, it admits a closed-form solution given by:

$$V^\star = q^\star = \frac{1}{\alpha - 1}, \qquad \hat{V}^\star = \hat{q}^\star = \alpha - 1 \tag{A17}$$

Therefore, the full training loss is given by:

$$\lim_{\lambda\to 0^+} \mathcal{E}_\ell^{\text{gcm}}(\alpha,\lambda) = \begin{cases} 0 & \text{for } \alpha \le 1 \\ \frac{1}{2}\left(1 - \frac{1}{\alpha}\right) & \text{for } \alpha > 1 \end{cases} \tag{A18}$$

Appendix B: From Gaussian mixture to single Gaussian

1. Mixture of Gaussians with zero means

We first prove Lemma 1 in the main text. First, by Theorem 2, the asymptotic loss $\mathcal{E}_\ell^{\rm gmm}(\boldsymbol{\rho}, \boldsymbol{M}, \boldsymbol{\Sigma}^\otimes)$ (resp. $\mathcal{E}_\ell^{\rm gmm}(\boldsymbol{\rho}, \boldsymbol{0}, \boldsymbol{\Sigma}^\otimes)$) is a deterministic function of $(m_c^\star, q_c^\star, V_c^\star)_{c\in\mathcal{C}}$, which are the *unique* fixed points of (10) (resp (A9)). Since both saddle points equations differ only by setting $m_c = \hat{m}_c = 0$, Lemma 1 is a consequence of the following:

Lemma 2. *Let $(V_c^\star, q_c^\star)_{c\in\mathcal{C}}$ be the solutions of Eqs. (A9). Then, $(0, V_c^\star, q_c^\star)_{c\in\mathcal{C}}$ satisfy the general fixed point equations of (10).*

Proof. If we plug in $m_c = \hat{m}_c = 0$ for all $c \in \mathcal{C}$, the equations for $V_c, \hat{V}_c, q_c, \hat{q}_c$ become identical in (10) and (A9). It is also easy to check that $\hat{m}_c = 0$ for all c implies that $m_c = 0$; what remains is to show that the last equation holds, i.e.

$$\frac{\alpha}{2}\rho_c \sum_{y\in\{-1,+1\}} \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[f_\ell(y, \sqrt{q_c^\star}\xi, V_c^\star)\right] = 0. \tag{B1}$$

Define the function

$$g(\omega, V) = f_\ell(-1, \omega, V) + f_\ell(+1, \omega, V),$$

so that

$$\hat{m}_c^\star \propto \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[g(\sqrt{q_c^\star}\xi, V_c^\star)\right]$$

We shall show that g is odd in ω; since ξ is centered, the lemma will be proven. To do so, we shall show that

$$f_\ell(y, \omega, V) = -f_\ell(-y, -\omega, V),$$

for all $y \in \{-1, +1\}$, $\omega \in \mathbb{R}$, and $V \in \mathbb{R}$. By definition, we have

$$f_\ell(y, \omega, V) = V^{-1}\left(\mathrm{prox}_{V\ell(\cdot,y)}(\omega) - \omega\right),$$

and the linear term in ω is immediate. For the proximal operator, we use the symmetry of ℓ and write

$$\begin{aligned}
\mathrm{prox}_{V\ell(\cdot,y)}(\omega) &= \arg\min_{z\in\mathcal{C}_1}\left[\frac{1}{2\tau}(z-\omega)^2 + \ell(z,y)\right] \\
&= \arg\min_{z\in\mathcal{C}_1}\left[\frac{1}{2\tau}((-z)-(-\omega))^2 + \ell(-z,-y)\right] \\
&= -\,\mathrm{prox}_{V\ell(\cdot,-y)}(-\omega),
\end{aligned}$$

which concludes the proof. □

2. Strong universality of ordinary least-squares

We now have all elements we need to establish the universality of the ordinary least-squares estimator stated in Theorem 5 in the main. Our starting point is the ordinary least-squares problem for the Gaussian Mixture Model in the overdetermined regime $\alpha > 1$. In this cave, the training loss is given by eq. (A11) with $(V_c^\star, q_c^\star)_{c\in\mathcal{C}}$ unique solutions of the following equations:

$$\begin{cases} \hat{V}_c = \frac{\alpha\rho_c}{1+V_c} \\ \hat{q}_c = \alpha\rho_c\frac{1+q_c}{(1+V_c)^2} \end{cases}, \qquad \begin{cases} V_c &= \frac{1}{d}\,\mathrm{tr}\,\Sigma_c\left(\sum_{c'\in\mathcal{C}}\hat{V}_{c'}\Sigma_{c'}\right)^{-1} \\ q_c &= \frac{1}{d}\sum_{c'\in\mathcal{C}}\left[\hat{q}_{c'}\,\mathrm{tr}\,\Sigma_{c'}\Sigma_c\left(\sum_{c''\in\mathcal{C}}\hat{V}_{c''}\Sigma_{c''}\right)^{-2}\right] \end{cases} \tag{B2}$$

We shall now show how to reduce these equation to a simple analytical formula, equivalent to the one of a single Gaussian. Combining the equations for $\hat{V}_c$ and V_c, one sees that the fixed point must satisfy the following identity:

$$\sum_{c\in\mathcal{C}} \hat{V}_c^\star V_c^\star = 1 \tag{B3}$$

Similarly, multiplying the equation for q_c by $\hat{V}_c$, summing over $c \in \mathcal{C}$ and doing the same for the equation for $\hat{q}_c$ with V_c, we get a second identity satisfied by the fixed-point:

$$\sum_{c\in\mathcal{C}} \left(\hat{V}_c^\star q_c^\star - V_c^\star \hat{q}_c^\star\right) = 0 \tag{B4}$$

Note that, at this point these relations could have been derived for any loss functions. For the specific case of the square loss, further substituting the hat variables, these conditions are equivalent to:

$$\sum_{c\in\mathcal{C}} \rho_c \frac{V_c^\star}{1+V_c^\star} = \frac{1}{\alpha} \tag{B5}$$

$$\sum_{c\in\mathcal{C}} \rho_c \frac{V_c - q_c}{(1+V_c)^2} = 0 \tag{B6}$$

We thus find, combining the eq. (B2) for $\hat{q}_c$ with eq. (B6)

$$\sum_{c\in\mathcal{C}} \hat{q}_c^\star = \sum_{c\in\mathcal{C}} \alpha\rho_c \frac{1+V_c^\star}{(1+V_c^\star)^2} = \sum_{c\in\mathcal{C}} \alpha\rho_c \frac{1}{1+V_c^\star} \tag{B7}$$

Our goal is to evaluate the loss at the fixed point, which is given by eq. (A13):

$$\mathcal{E}_\ell^{\text{gmm}} = \sum_{c\in\mathcal{C}} \frac{\hat{q}_c^\star}{2\alpha} \tag{B8}$$

Combining this definition with eqs. (B6) and (B7), we find that

$$2\mathcal{E}_\ell^{\text{gmm}} + \frac{1}{\alpha} = \sum_{c\in\mathcal{C}} \rho_c \frac{1}{1+V_c^\star} + \sum_{c\in\mathcal{C}} \rho_c \frac{V_c^\star}{1+V_c^\star} = 1 \tag{B9}$$

so that finally, we reach the promised result:

$$\lim_{\lambda\to 0^+} \mathcal{E}_\ell^{\text{gmm}}(\alpha,\lambda,K) = \frac{1}{2}\left(1-\frac{1}{\alpha}\right)_+ = \lim_{\lambda\to 0^+} \mathcal{E}_\ell^{\text{gcm}}(\alpha,\lambda) \tag{B10}$$

as claimed in Theorem 5 in the main.

Appendix C: Numerical Simulations

In this section, we provide further details concerning the protocol we used to perform the numerical simulations, which corroborate the theoretical results exemplified in the main manuscript. All codes are publicly available on the GitHub repository associated to the current paper at https://github.com/IdePHICS/RandomLabelsUniversality. For concreteness and completeness, we illustrate these simulations with yet again a different case with respect to what was already presented in the main text in Fig. 4, where we use a smaller version of the well-known Imagenet benchmark [89]. It is made of 100.000 natural images, downsampled to 64×64 pixels each and grouped into 200 different classes.

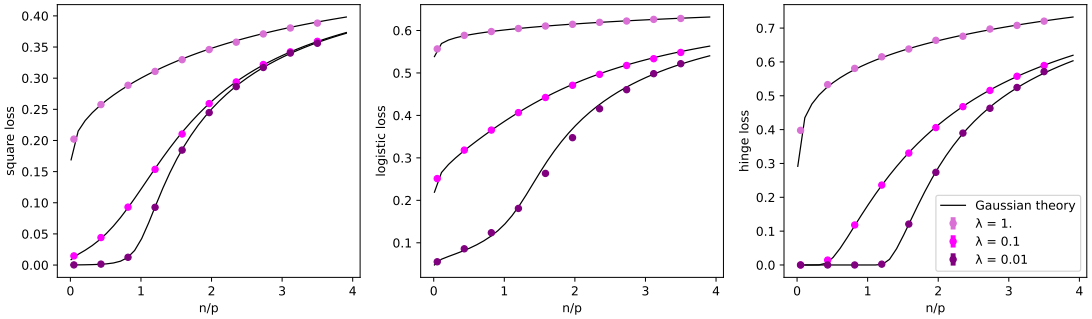

FIG. 4. Numerical simulations of universality: As in Fig. 2, this figure shows the training loss as a function of the number of samples n per dimension p at various values of λ for another data set we used here for completeness. Here we used a grayscale tiny-Imagenet pre-processed with Gaussian random features and tanh non-linearity. In the left panel the square loss, in the middle panel the logistic loss and in the right panel the hinge loss. The coloured dots refer to numerical simulations while the black solid lines correspond to the theoretical prediction of single Gaussian with corresponding input covariance matrices. The numerical simulations are averaged over 10 different realizations.

In all the numerical experiments on real datasets shown so far, we have both normalized and then pre-processed the datasets with either random features or wavelet-scattering transform. Fig. 5 compares instead the predictions of the Gaussian theory with respect to the numerical simulations on MNIST, fashion-MNIST, CIFAR10 and tiny ImageNet when no pre-processing is applied. As can be seen, despite the overall quite good agreement between theory and numerical experiments, we start observing some (very) small deviations from the Gaussian predictions. Indeed, as it is shown in sec. D, the covariance matrices associated to the different modes of the underlying real data distribution are, in this case, more heterogeneous than the ones observed when a pre-processing stage is applied. This is consistent with the homogeneous assumption in Theorem 3 and implying Gaussian Universality.

a. Dataset generation. As we have seen in the main manuscript, we basically deal with three different types of datasets. Two of them are synthetic datasets and correspond to i.i.d Gaussian input data-points and Gaussian Mixtures. The remaining one accounts for real datasets, such as MNIST [90], fashion-MNIST [31] and CIFAR10 [31] in grayscale, pre-processed with either random feature maps [33] or through wavelet scattering transform [29]. The procedure used to generate these kinds of datasets is exemplified in sec. IV. For the sake of clarity, we summarized it through the pseudo code in algorithm 1.

Algorithm 1 Generating dataset $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$

Input: Integer p, flag *dataset*, matrix $F \in \mathbb{R}^{d\times p}$ of random Gaussian features
If the *dataset type* is i.i.d. Gaussian:
  Sample each input data-point as $\boldsymbol{x}^\mu \sim \mathcal{N}(0, \boldsymbol{I})$, with $\boldsymbol{I} \in \mathbb{R}^{p\times p}$ the identity matrix;
Else if the *dataset type* is a Gaussian Mixture:
  Sample each input data-point as $\boldsymbol{x}^\mu \sim \sum_{k=1}^K \rho_k \, \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, with $\boldsymbol{\mu}_k$ being the centroid of the k-th cluster and Σ_k the corresponding covariance matrix;
Else if the *dataset type* is a real dataset pre-processed with random gaussian features:
  Load the real dataset samples $\boldsymbol{z}^\mu$ $\forall \mu = 1, ..., n$ with Pytorch dataloaders;
  Assign $\boldsymbol{x}^\mu \to \sigma(\boldsymbol{z}^\mu F)$;
Else if the *dataset type* is a real dataset pre-processed with wavelet scattering:
  Load the real dataset samples $\boldsymbol{z}^\mu$ $\forall \mu = 1, ..., n$;
  Apply wavelet scattering transform on $\boldsymbol{z}^\mu$;
Sample the labels according to the Rademacher distribution as $y^\mu \sim \frac{1}{2}(\delta_{+1} + \delta_{-1})$
Return: $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$

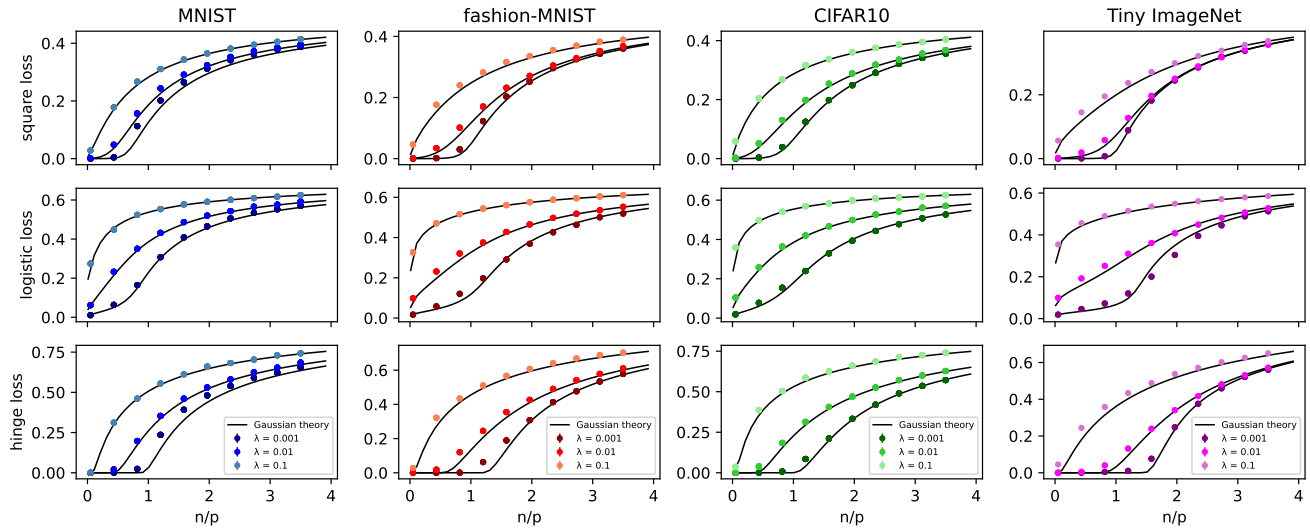

FIG. 5. This figure shows the training loss as a function of the number of samples n per dimension p at finite regularization λ. In the top panel the square loss, and in the bottom panel the hinge loss. The first column refers to MNIST, the second column corresponds to fashion-MNIST, the third column corresponds to CIFAR10 in grayscale, the fourth column corresponds to tiny ImageNet in grayscale. Black solid lines correspond to the outcome of the replica calculation, obtained by assigning to Σ the covariance matrix of each dataset. The coloured dots correspond to the simulations for different values of λ, as specified in the plot legend. Simulations are averaged over 10 samples & the error bars are not visible at the plot scale.

The real datasets are loaded through Pytorch dataloaders [91]. In particular, the dataloader of CIFAR10 includes a grayscale transformation of the dataset in order to reduce to one the number of input channels of the RGB colour encoding scheme. The wavelet scattering transform is instead implemented by means of the Kymatio Python library [34]. Note that, with the purpose of speeding up the realization of the learning curves and to reduce fluctuations, the pre-processed real datasets are generated once for all through algorithm 1 and then stored in a hdf5 file.

b. Learning phase. Given the dataset generated as in algorithm 1, the aim is to infer the estimator $\boldsymbol{\theta}$ minimizing the empirical risk as in eq. (1) of the main manuscript. In the present work we consider three distinct kinds of loss functions:

(i) **Square Loss.** In this specific case, the goal is to solve the following optimization problem:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}) = \inf_{\boldsymbol{\theta}\in\mathcal{S}_p} \frac{1}{2n}\sum_{\mu=1}^{n}(\boldsymbol{\theta}^\top \boldsymbol{x}_\mu - y_\mu)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2, \tag{C1}$$

The estimator can be here determined through the Moore-Penrose inverse as it follows, without relying on any learning algorithm:

$$\boldsymbol{\theta} = \begin{cases} \left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}, & \text{if } n > p \\ \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n\right)^{-1}\boldsymbol{y}, & \text{if } p > n \end{cases} \tag{C2}$$

(ii) **Logistic Loss.** In this specific case, the goal is to solve the following optimization problem:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}) = \inf_{\boldsymbol{\theta}\in\mathcal{S}_p} \frac{1}{n}\sum_{\mu=1}^{n}\log\left(1+\exp(-y_\mu\boldsymbol{\theta}^\top\boldsymbol{x}_\mu)\right) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2, \tag{C3}$$

Since the estimator of logistic regression can not be determined through an explicit closed formula, we here made use of the *lbgfs* solver with *penalty* set to ℓ_2. This optimizer corresponds to a Gradient Descent (GD)-like second order optimization method and it is implemented in the LogisticRegression class of the Scikit-Learn Python library [92]. The GD algorithm stops either if a maximum number of iterations has been reached or if the maximum component of the gradient goes below a certain threshold. We fixed this tolerance to $1e-5$ and the maximum number of iterations to $1e4$.

(iii) **Hinge Loss.** In this specific case, the goal is to solve the following optimization problem:

$$\widehat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}) = \inf_{\boldsymbol{\theta}\in\mathcal{S}_p} \frac{1}{n}\sum_{\mu=1}^{n} \max\left(0, 1 - y_\mu \boldsymbol{\theta}^\top \boldsymbol{x}_\mu)\right) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2, \tag{C4}$$

As for logistic regression, even in this case we can not rely on any explicit formula for the estimator, it has rather to be inferred by means of learning algorithm. In particular, for the simulations at finite regularization strength, we made use of the LinearSVC class provided by Scikit-Learn [92] and implementing the Support Vector Classification (SVC) with linear kernels and L_2 regularization if *penalty* is set to ℓ_2. In this case, we set the tolerance of convergence to $1e-5$ and the maximum number of iterations to $5e5$. Unfortunately, LinearSVC struggles to converge for vanishing regularization strengths. Therefore, we made use of CVXPY [93, 94] in order to perform the simulations at $\lambda = 1e-15$. CVXPY is an open-source Python-embedded modeling language for convex optimization problems. We set the *solver* option to None, in this way CVXPY chose automatically the most specialized solver for the optimization problem type. While being slower than LinearSVC, CVXPY guarantees convergence at vanishing regularization strengths.

At the end of the training process, we evaluate the training loss ℓ on the minimizer of the corresponding empirical risk minimization problem. To get the learning curves, we then repeat the whole process for a specified range of n/p and for a certain number of different realization of the learning problem, as exemplified in algorithm 2.

Algorithm 2 Learning curve

Input: range of n/p, flag *dataset type*, flag *which estimator*
For *seed* in a specified number of seeds **do**:
  For n/p in a specified range **do**:
    Choose the dataset according to *dataset type*;
    Compute the estimator according to the desired optimization problem as in (i)-(iii);
    Compute the training loss ℓ at fixed n/p;
  Update the mean train loss and its standard deviation with the new contribution from the current seed.
Return: Mean train loss and standard deviation as a function of n/p.

Appendix D: Empirical evidence of the homogenity assumption

As seen in the counter example illustrated in Fig. 3, in the case of very heterogeneous Gaussian Mixtures we can observe small deviations from universality both at zero and finite regularization. However, this disagreement between single Gaussian and Gaussian Mixtures does not appear in the experiments with real datasets of Fig. 1 and Fig. 2, despite their certainly multi-modal and mode-heterogeneity nature. First, we must acknowledge that deviations are, in general, observed to be small with respect to the homogeneous case, and that the data presented in Fig. 3 were carefully tuned so that the difference is visible.

Additionally, in this section, we also empirically demonstrate the similarity among the empirical correlation matrices of the various modes characterizing real dataset distributions. Fig. 6 shows the correlation matrix of all grayscale CIFAR-10 images depicting airplanes (leftmost), automobiles (middle) and trucks (rightmost) respectively. The point we wish to convey in this plot is that, despite the fact that there exists some modes of the CIFAR-10 empirical distribution which display a consistently different correlation structure (airplane mode) with respect to the other modes (automobile and truck mode), there exists some others which look like more similar among each other (automobile and truck mode).

As can be seen in Fig. 7 and Fig. 8, the structure similarity of the covariance matrices of the various mode is further enhanced when pre-processing grayscale CIFAR-10 with both Gaussian random feature maps and wavelet scattering transforms, at the point that even the less similar modes in the raw dataset conform to the others (see airplane mode).

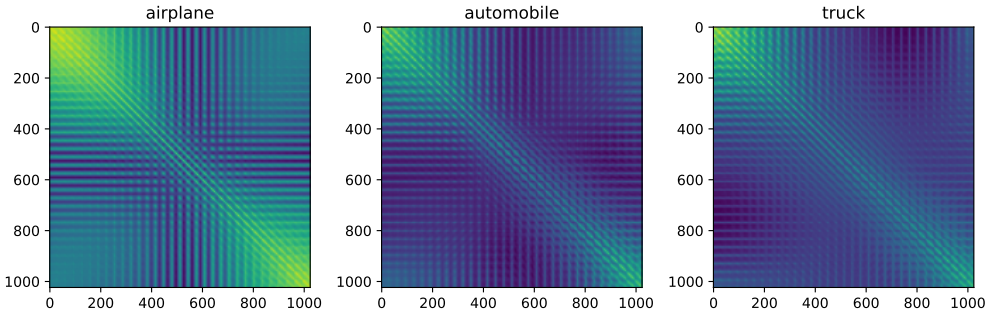

FIG. 6. Input data correlation matrix of grayscale CIFAR10, conditioned on the true labels, e.g. airplane (leftmost), automobile (middle), truck (rightmost). Lighter colors refer to stronger correlation.

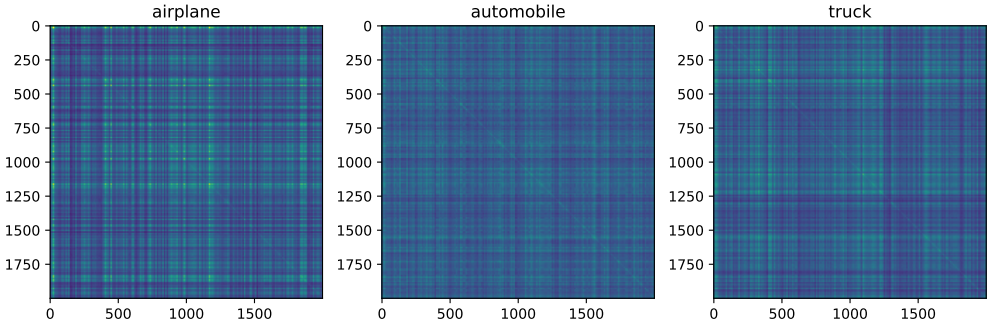

FIG. 7. Input data correlation matrix of grayscale CIFAR10 pre-processed with Gaussian random features and erf non-linearity. The correlation matrices are conditioned on the true labels, e.g. airplane (leftmost), automobile (middle), truck (rightmost). Lighter colors refer to stronger correlation.

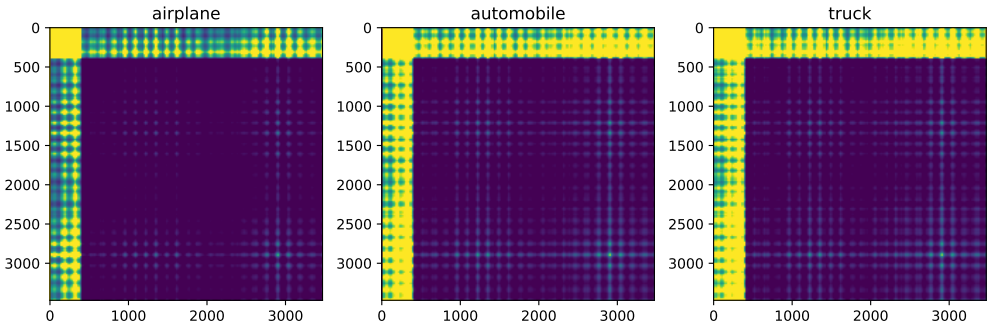

FIG. 8. Input data correlation matrix of grayscale CIFAR10 pre-processed with wavelet scattering transform. The correlation matrices are conditioned on the true labels, e.g. airplane (leftmost), automobile (middle), truck (rightmost). Lighter colors refer to stronger correlation.