Gaussian Universality of Linear Classifiers with Random Labels in High-Dimension

Federica Gerace¹, Florent Krzakala², Bruno Loureiro², Ludovic Stephan², and Lenka Zdeborová³

¹International School of Advanced Studies (SISSA). Trieste, Italy.
²Ecole Polytechnique Fédérale de Lausanne (EPFL). Information, Learning and Physics (IdePHICS) lab.
CH-1015 Lausanne, Switzerland.
³Ecole Polytechnique Fédérale de Lausanne (EPFL). Statistical Physics of Computation (SPOC) lab.
CH-1015 Lausanne, Switzerland.

Abstract

While classical in many theoretical settings, the assumption of Gaussian i.i.d. inputs is often perceived as a strong limitation in the analysis of high-dimensional learning. In this study, we redeem this line of work in the case of generalized linear classification with random labels. Our main contribution is a rigorous proof that data coming from a range of generative models in high-dimensions have the same minimum training loss as Gaussian data with corresponding data covariance. In particular, our theorem covers data created by an arbitrary mixture of homogeneous Gaussian clouds, as well as multi-modal generative neural networks. In the limit of vanishing regularization, we further demonstrate that the training loss is independent of the data covariance. Finally, we show that this universality property is observed in practice with real datasets and random labels.

1 Introduction

A line of work, particularly popular among theoreticians interested in high-dimensional statistics, focuses on the theory of machine learning for the simple case of Gaussian generated data (also known as Gaussian design), e.g. [1–4]. Despite being both common & convenient for doing theory, the Gaussian i.i.d. assumption on the data might seem at first a strong limitation, out-of-touch with real-world practice. Indeed, an important branch of statistical learning theory is data-agnostic, and avoids making too specific assumptions on the data distribution [5].

However, a number of recent observations (both heuristic and rigorous) suggests that the Gaussian assumption is not always that far-stretched for high-dimensional data (see for instance [6–10] and references therein). Our goal in the present work is to shed some light on the validity of the Gaussian hypothesis for perhaps the simplest, yet deeply fundamental, problem of high-dimensional statistics: generalized linear classification with random labels.

From the definition of Rademacher complexities in statistical learning theory [5, 11], passing by Wendel/Cover’s pioneering studies [12, 13] to thought-provoking numerical experiments with deep learning [14, 15], random labels are ubiquitous in the theory of machine learning. In fact, the capacity of generalized linear classifiers to fit random labels for the simple case of Gaussian generated data, is also a classical topic, e.g. [16–24].

In this work, we ask how would these theories change if using a realistic dataset? We consider the training loss of generalized linear classifiers 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 proportional high-dimensional setting when both \( n \) (the number of training samples) and \( p \) (the dimension) goes to infinity as \( \alpha = n/p \) is fixed.

Informally, our main result is to prove that in the aforementioned high-dimensional setting many data distributions actually have the same learning properties as Gaussian data, which turns out to be surprisingly universal for this problem. In particular, the minimum training loss for a wide range of settings turns out to be the same as 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 turns out to be 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 Gaussian data capture what is actually happening in practice.

This result is, we believe, counter-intuitive and rather interesting. Certainly, the value of the interpolation (or capacity) threshold was already known to be universal and occurs (for full-rank data) at \( n = p \) for ridge regression, and for \( n = 2p \)
Figure 1: Training loss as function of the number of samples $n$ per input dimension $p$ at $\lambda = 10^{-15}$. In the left panel the square loss, in the right panel the hinge loss. The black solid line represents the outcome of the replica calculation for iid Gaussian inputs, namely when the covariance matrix $\Sigma$ 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 $\mu_{\pm} = (\pm 1, 0, \ldots, 0)$, $\Sigma_{\pm}$ both equal to the identity matrix and $\rho_{\pm} = 1/2$, finally, black dots correspond to i.i.d. Gaussian inputs.

for linear classifiers [13]; however, the fact that the loss itself is universal is a stronger statement that redeems an entire line of work in statistics working under the Gaussian assumption.

Main Contributions

The main points of the present work can be summarized by figures 1 and 2, that show the training loss of real world data sets with various feature maps, trained with random labels, compared with the (analytical) prediction for Gaussian data with matching covariance. 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 loss plotted as a function of the ratio $\alpha = n/p$ for random labels using MNIST [30], fashion-MNIST [31], CIFAR10 [32], or a Gaussian mixture data, are indistinguishable from results obtained for a single Gaussian. The code used to run these experiments is publicly available in a GitHub repository. This conclusion seems robust and holds as well if we use more complicated features instead of the raw data, such as random features [33] or the convolutional scattering transform [29, 34]. 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. Our main contribution is to provide a rigorous theoretical foundation for these observations, that vindicates the classical theoretical line of works on Gaussian design. We list here our main contributions:

- We provide a strong universality theorem for ridge regression in high-dimensions with vanishing regularization and random labels. Informally, we prove that for random labels any data created by a multi-modal generative neural network have the same minimum asymptotic loss as Gaussian data generated with trivial covariance, that is $E_{\ell}(\alpha) = 1/2(1 - 1/\alpha)_+$. This provides a theoretical explanation for the phenomena illustrated in Fig. 1 left.

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

- 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. This is illustrated in Fig. 2.

The proof technique used to establish these universality theorems has an interest on his own, and builds on recent progress in high-dimensional statistics. Our technical contributions are threefold:
Figure 2: This figure shows the training loss as a function of the number of samples $n$ per dimension $p$ at finite $\lambda$. In the top panel the square loss, 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 $\mu_\pm = (\pm 1, 0, \ldots, 0)$, $\Sigma_\pm$ both equal to the identity matrix and $\rho_\pm = 1/2$. Black solid lines correspond to the outcome of the replica calculation, obtained assigning to $\Sigma$ the covariance matrix of each dataset plus the corresponding transformation. The coloured dots correspond to the simulations for different values of $\lambda$, as specified in the plot legend. Simulations are averaged over 10 samples & the error bars are not visible at the plot scale.

- First, we provide a proof generalizing a previous result by [7] showing that many real-world applications can be accurately approximated in high-dimensions by a non-trivial Gaussian mixture. Note that this contribution, that builds on very recent breakthrough [35–37], is valid beyond the random label assumption.

- We then move to the random labels case, and prove our second result: a mixture of many homogeneous Gaussian data has the same asymptotic learning properties as those of a single Gaussian distribution. This non-trivial map from a multi-modal to a single-mode distribution is achieved thanks to the explicit matching of the rigorous expression for the asymptotic minimal loss in both cases [10, 38–40].

- Finally, we further demonstrate the strong universality for ridge regression with vanishing regularization, again by showing explicitly that the exact solution [10, 41, 42] reduces to one of the homogeneous Gaussian cases.

**Related work**

**Universality** — There have been much progress on a similar, though more restricted, Gaussian universality for random feature maps on Gaussian input data [33]. Following early insights by [43], [44, 45] 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 [46] using the heuristic replica method, and proven in [47]. A general Gaussian Equivalence Principle [6] has been proven in a succession of works for convex penalties in [35, 36] and some non-convex ones in [37]. These tools fall short when considering realistic datasets. Indeed, these previous works considered just uni-modal Gaussian data (observed through random feature maps), a situation far from realistic multi-modal real world datasets.

Instead, [7, 48] argued that real datasets can be efficiently approximated in high-dimensions by a mixture of Gaussians, and provided deep results in particular for the softmax loss. One of our contributions here is to bridge this work with the aforementioned results of [36, 37]). Unifying these contribution allows us to generalize these results and to show that one can approximate realistic datasets in high-dimensions by multi-modal Gaussian mixtures.

We could not, however, find any reference discussing universality for random labels, which is the core of the present paper.

**Random Labels** — Random labels are a fundamental and useful concept in machine learning. The pioneering work of [14], 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 [49–51], in minima sharpness [52, 53] or in what neural networks can actually learn with random labels [15]. It has also been recently claimed [15] 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.

The perceptron — The question of how many samples can be perfectly fitted by a linear model is also 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 [13] 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, and recently in theoretical computer science [16–24, 54, 55]. It is one of our goals to attract the attention to these works, given the Gaussian universality we present shows that their relevance is not limited to idealistic Gaussian data.

2 Setting and notation

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

\[
\hat{R}_n^\star(X, y) = \inf_{\theta \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n \ell(\theta^\top x_i, y_i) + \frac{\lambda}{2} \|\theta\|^2,
\]

where the \( x_i \in \mathbb{R}^p \) are input vectors, \( y_\mu \in \{-1, +1\} \) are binary labels, and \( \mathcal{S}_p \) is a compact subset of \( \mathbb{R}^p \). We assume that the loss \( \ell \) only depends on the inputs \( x_i \) through a one-dimensional projection \( \theta^\top x_i \), and we work in the so-called proportional high-dimensional limit, where \( n, p \) go to infinity with

\[
\frac{n}{p} \to \alpha > 0.
\]

The problem defined in eq. (1) describes many settings of interest in machine learning, e.g. ridge or logistic regression, or (if the data are seen in a feature space) kernel methods, or neural networks in the lazy regime [27, 28]. Although some of our results hold in a more general setting, we focus in this work on the random label model, where the \( y_\mu \) are independent from the inputs \( x_\mu \), and generated independently according to a Rademacher distribution:

\[
y_\mu \sim \frac{1}{2} (\delta_{-1} + \delta_{+1})
\]

A number of data models will be of interest in the following. Here we define them:

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

\[
x_\mu \sim \mathcal{N}(0, \Sigma).
\]

The Gaussian covariate model has been the subject of much attention recently [4, 10, 38, 41, 42, 45, 56–60]. In particular, the asymptotic statistics of the minimizer of eq. (1) for different models for the labels is rigorously known in the literature. In particular, the random label limit relevant for our discussion can be obtained as a limit of the expressions proven in [10]. For completeness, we reproduce them in Appendix B.3.

• A more generic model of data, which has the advantage of being multi-modal, is the Gaussians Mixture Model (GMM). In this case, the inputs \( x_\mu \in \mathbb{R}^p \) are independently generated as:

\[
x_\mu \sim \sum_{c \in \mathcal{C}} \rho_c \mathcal{N}(\mu_c, \Sigma_c)
\]

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 exact asymptotic expression for the minimum training loss has been derived for a range of particular cases in, between others, [61–66] and in full generality for arbitrary means and covariances in [40]. For completeness, we reproduce in Appendix B.2 the random label limit of their expression in the binary classification case.
where \( \theta \) such as GANs [67] or normalizing flow [68]. This is a very generic setting, that can approximate any kind of data distribution and had immense empirical success, and we are going to use as the closest model to realistic data. In this case, we first sample a latent vector \( z_\mu \) from a Gaussian mixture distribution

\[
z_\mu \sim \sum_{c \in C} \rho_c N(\mu_c, \Sigma_c) \tag{5}
\]

and the data are generated by

\[
x_\mu = \Psi_{nn}(z_\mu), \tag{6}
\]

where \( \Psi_{nn} \) is a function parametrized by a neural network.

### 3 Main results

In this section we present the main theoretical contributions of the present work and discuss their consequences. Our first result is to show a Gaussian universality for the general empirical risk minimization setting introduced in eq. (1). The proof draws from recent results by [37], and generalizes the work of [7] that only considered universality of Gram matrices through the prism of random matrix theory. We also present a thorough, although informal, argument that the proof draws from recent results by [37], and generalizes the work of [7] that only considered universality of Gram matrices.

#### Universality of Gaussian mixtures

Consider the empirical risk minimization problem in eq. (1), and assume that the labels \( y_\mu \) are generated according to

\[
y_\mu = \eta(\theta^T x_\mu, \varepsilon_\mu),
\]

where \( \theta^c \in \mathbb{R}^p \) and \( \varepsilon_\mu \) is an i.i.d source of randomness. We also allow the function \( \eta \) to depend on the sample label cladd \( C_\mu \) if the \( x_\mu \) are generated according to a mixture model as in eq. (8). This encompasses diverse settings such as generalized linear models with noise, mixture classification, but also the random label setting (with \( \eta(x, \varepsilon) = \varepsilon \)). Let \( X \in \mathbb{R}^{p \times n} \) be a matrix with i.i.d columns, such that the \( x_\mu \) are distributed according to the neural network model of eq. (6). We define the equivalent Gaussian input matrix \( G \in \mathbb{R}^{p \times n} \) with independent columns \( g_1, \ldots, g_n \) such that

\[
g_\mu \sim \sum_{c \in C} \rho_c N(\mu_{c,n}, \Sigma_{c,n}), \tag{7}
\]

where

\[
\mu_{c,n} = \mathbb{E}_{z \sim N(\mu_c, \Sigma_c)}[\Psi_{mn}(z)] \quad \text{and} \quad \Sigma_{c,n} = \mathbb{E}_{z \sim N(\mu_c, \Sigma_c)}[(\Psi_{mn}(z) - \mu_c)(\Psi_{mn}(z) - \mu_c)^\top]
\]

The crucial assumption we make is a so-called one-dimensional CLT: for any Lipschitz function \( \varphi : \mathbb{R} \to \mathbb{R} \),

\[
\lim_{n,p \to \infty} \sup_{\theta \in \mathbb{R}_p} \left| \mathbb{E}[\varphi(\theta^T x)] - \mathbb{E}[\varphi(\theta^T g)] \right| = 0 \tag{8}
\]

Then, with suitable regularity conditions on the loss and the labeling functions, we show the following:

**Theorem 1** (Universality of GMM). For any bounded Lipschitz function \( \Phi : \mathbb{R} \to \mathbb{R} \), we have

\[
\lim_{n,p \to \infty} \left| \mathbb{E}\left[ \Phi\left( \hat{R}_n^*(X, y(X)) \right) \right] - \mathbb{E}\left[ \Phi\left( \hat{R}_n^*(G, y(G)) \right) \right] \right| = 0
\]

In particular, for any \( \varepsilon \in \mathbb{R} \),

\[
\hat{R}_n^*(X, y(X)) \xrightarrow{p} \varepsilon \quad \text{if and only if} \quad \hat{R}_n^*(G, y(G)) \xrightarrow{p} \varepsilon \tag{9}
\]

The full theorem, as well as its proof, is presented in Appendix A. In a nutshell, this theorem shows that the multimodal data generated by any generative neural network is equivalent to a finite mixture of Gaussian in high-dimensions: in other words, a finite mixture of Gaussians leads to the same loss as for data generated by (for instance) a GAN. Note that this theorem has an interest beyond the scope of this work, since it does not require random labels. The condition in eq. (8) might appear strong. However, such one-dimensional CLTs have been the subject of many recent works who proved them for many cases [35–37], including random features and two-layers neural tangent kernels.
A crucial remark is that Theorem 1 does not require that the data generated by GANs are Gaussians mixtures: rather, it is their one-dimensional projections along the directions in $S_p$ that should behave as such, which should be expected from a naive application of a the central limit theorem. Generic formal arguments can be given for the validity of this one-dimensional CLT, and we provide such arguments in the appendix. In particular we argue that a large class of distributions, including the generative models of eq. (6), do satisfy this condition, although for a possibly random choice of $S_p$. This condition can also be checked empirically in simulations [35, 69].

A second remark is that the interest of the theorem lies in the fact requires only a finite mixture to approximates the loss. Indeed, while we could use the standard approximation results (e.g. the Stone-Weierstrass theorem) to approximate the data density to arbitrary precision by Gaussian mixtures, this would require a diverging number of Gaussian in the mixture. The fact that loss is captured with finite $K$ is key to our approach.

**Deterministic limits and mean invariance** So far, we showed that real datasets can be approximated by Gaussian mixtures in high dimension. We now move to the random label case, and show how this allows for stronger statements, and in particular how we can surprisingly use a simple Gaussian distribution instead. The consequence (9) of Theorem 1 is of special importance. Indeed, [40] show how to characterise the asymptotic loss and provide the following theorem:

**Proposition 1 (Theorem 1 from [40]).** Consider the minimization problem in eq. (1), with the inputs $x_\mu$ generated according to a Gaussian mixture as in (5). Under mild regularity conditions on the $\mu_c$, $\Sigma_c$, as well as the loss and regularizer, we have

$$\hat{R}_n^\mu(X, y(X)) \overset{p}{\to} \mathcal{E}_\ell,$$

where $\mathcal{E}_\ell$ is a deterministic value given a the so-called replica equation.

We present the full theorem, as well as the deterministic (replica) equations to compute the limiting value $\mathcal{E}_\ell$, in Appendix B. This value only depends on the probability vector $\rho \in [0, 1]^K$ (with entries $\rho_c$, corresponding to the respective sizes of the $K$ clusters), the matrix of averages $M \in \mathbb{R}^{K \times p}$ (with rows $\mu_c \in \mathbb{R}^p$), and the concatenation of covariances $\Sigma^\otimes \in \mathbb{R}^{K \times p \times p}$ (with rows $\Sigma_c \in \mathbb{R}^{p \times p}$) and therefore we denote:

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

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

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

where in both cases we omitted the explicit dependence on the parameters $(\alpha, \lambda)$. We are now in a position to state a lemma crucial to our second main result:

**Lemma 1 (Single mode lemma).** In the random label setting (2), assume that the loss $\ell$ 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 $\rho$, $M$ and $\Sigma^\otimes$ we have

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

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 2 (Gaussian universality).** Consider the same assumptions as in lemma 1, and assume further that the data is homogeneous, i.e.

$$\Sigma_c = \Sigma \quad \text{for all} \quad c \in C.$$

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

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

This is our second main result: a mixture of homogeneous Gaussians can be replaced (with random labels) by a single Gaussian. This surprising fact explains the empirical observation presented in Fig. 1 and Fig. 2, at least if we accept that the different mode are homogeneous (see discussion in Sec.4). 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 3.** Consider the same assumptions as in theorem 2, then if the minimizer of $\ell$ is unique and the data covariance full-rank, then the asymptotic minimal loss for Gaussian data does not depends on the covariance when $\lambda = 0$. 

6
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 \( x_\mu = \Sigma^{1/2} x_\mu \), with \( z_\mu \sim N(0, I_p) \). Then, making a change of variables \( \theta' = \Sigma^{1/2} \theta \), we can write:
\[
\hat R_n(X, y) = \inf_{\theta \in S_\rho} \frac{1}{n} \sum_{\mu=1}^n \ell(\theta^T x_\mu, y_\mu) + \frac{\lambda}{2} \| \theta \|^2 = \inf_{\theta \in S_\rho} \frac{1}{n} \sum_{\mu=1}^n \ell(\theta^T z_\mu, y_\mu) + \frac{\lambda}{2} \| \Sigma^{-1/2} \theta' \|^2
\]
where \( S_\rho \subset \mathbb{R}^p \) is another compact set, and we have used the fact that \( y_\mu \) are independent of \( x_\mu \). Since the minimizer of \( \ell \) is unique, the result follows from taking \( \lambda \to 0^+ \).

Note that in particular Theorem 3 implies that for random labels, the GCM model with a covariance \( \Sigma \) 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 \( \Sigma^{-1} \). Therefore, in the case in which \( \ell \) has several minima, the \( \lambda \to 0^+ \) limit will give the performance of the solution with minimum \( \| \cdot \|_{\Sigma^{-1}} \) norm.

This analysis also allows to answer the question: what is being learned with random labels [15]? For generalized linear models: the model is simply fitting the 2nd order statistics (the total covariance \( \Sigma \)).

**Ridge regression with vanishing regularization** Even though it seems to be well obeyed in practice, one may wonder if we can in some case get rid of the homogeneity condition. As we shall see, the answer is not, and in general a mixture of inhomogeneous Gaussian cannot be strictly replaced a single one. It turns out, however, that there is one exception, and that the hypothesis can be lifted in one case, ordinary least square with the squared loss \( \ell(x, y) = \frac{1}{2}(x-y)^2 \):

**Theorem 4 (Strong universality of OLS).** In the ridge regression case, when \( \lambda \to 0^+ \), we have
\[
\lim_{\lambda \to 0^+} \mathcal{E}_\ell^{\text{gcm}}(\rho, M, \Sigma^\otimes) = \frac{1}{2} \left( 1 - \frac{1}{\alpha} \right)_{+},
\]
for any choice of \( \rho, M, \) or \( \Sigma^\otimes \).

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 \( \lim_{\lambda \to 0^+} \mathcal{E}_\ell^{\text{gcm}}(\alpha, \lambda) = \frac{1}{2}(1 - 1/\alpha)_{+} \). The proof of the strong universality is given in Appendix C.2.

## 4 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 pre-processed with random Gaussian feature maps [33]. In this case, the input data-points are constructed as \( x_\mu = \sigma \cdot (z_\mu, F) \), with \( z_\mu \in \mathbb{R}^d \) being a sample from one of the two datasets, \( F \in \mathbb{R}^{d \times p} \) representing the matrix of random features, whose row elements are sampled according to a normal distribution, and \( \sigma(\cdot) \) 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 build as a mixture of two Gaussians, with data covariance of the two clusters both equal to the identity matrix \( \Sigma_1 = \Sigma_2 = I \), \( \mu_{1/2} = (\pm 1, 0, \ldots, 0) \) and \( \rho_{1/2} = 1/2 \). Further technical details are given in appendix.

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 these 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.

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 3, 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.
Homogeneity assumption — A remarkable point is that the homogeneity assumption we use in Theorem 1 and Corollary 2, which can be relaxed only for ordinary least squared, does not seem to be that important in practice. One may thus wonder if the strong universality of Theorem 4 could be proved in full generality, and not only for OLS. It turns out that the answer is no. Using Proposition 1, 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 4, 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 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. [8, 35, 70–72]).

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 deviation we observed, even in our designed counter-example, are small, so that they might not even be seen in practice. Secondly, and especially after observing the data though random or scattering features, it turns out that when we measure the empirical correlation matrix of the different modes, they looks quite similar. In fact, it has even been advocated that neural networks are precisely learning representation that find such homogeneous Gaussian mixture [73].

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 [5, 11] that measures the ability of estimators in a hypothesis class $\mathcal{H}$ to fit i.i.d. random labels $y \sim \text{Rad}(1/2)$:

$$\text{Rad}_n (\mathcal{H}) = \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{\mu=1}^{n} y_{\mu} h(x_{\mu}) \right].$$

It is explicitly dependent on the specific distribution of the input data-points $x_{\mu}$. As discussed in [22] there exists a direct mapping between the Rademacher complexity and the minimum 0/1 training loss - or ground state energy in the statistical physics parlour. Indeed, for a binary hypothesis class $\mathcal{H} = \{ h : \mathbb{R}^p \rightarrow \{-1, +1\} \}$ the two are asymptotically related by the following equation:

$$\lim_{n \to \infty} \inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{\mu=1}^{n} \mathbb{P} ( h(x_{\mu}) \neq y_{\mu} ) = \frac{\alpha}{2} [ 1 - \text{Rad}_n (\mathcal{H}) ].$$

Moreover, [22] showed how to explicitly compute the Rademacher complexity for Gaussian data using the replica method from statistical physics. Given the universality advocated in this present work, this Gaussian results thus seem to be of more relevance than previously though, and in fact allows 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 label.

However, note that 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 [10, 40] 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.

5 Conclusion

For the classical problem of fitting random labels with generalized 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. There are a number of potential interesting extension of this work, including on non-convex loss and multi-layer neural networks that should be investigated in the future. Additionally, we proved, en passant, a result showing that finite Gaussian mixtures are also a good model in high-dimension—even with correlated labels—and this direction should be explored further. These results, we believe, are of special interest given the number of theoretical studies with the Gaussian design and their variants, that is amenable to exact characterization, and that turns out to be less idealistic, and more realistic, than perhaps previously assumed.

Acknowledgements

We acknowledge funding from the ERC under the European Union’s Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe.

References

[1] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.
[2] 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.
[3] 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.
[4] 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.
[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
[6] 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.
[7] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning parameters 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.
[8] 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.
[9] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.
[10] 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.
[11] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] Ahmed El Alaoui and Mark Sellke. Algorithmic pure states for the negative spherical perceptron. *arXiv preprint arXiv:2010.15811*, 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 Leonardiuzzi, 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] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joao 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

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

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

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

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

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

[46] 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.

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

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

[49] Devansh Arpit, Stanislaw Jastrzebski, 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.

[50] 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.

[51] 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.
[52] 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.

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

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

[55] 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.

[56] Denny Wu and Ji Xu. On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. In Advances in Neural Information Processing Systems, volume 33, 2020.

[57] 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.

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

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.

[68] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pages 1530–1538. PMLR, 2015.

[69] 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.

[70] Umberto M Tomasin, Antonio Scocchi, 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.
[71] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks, 2022.

[72] 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.

[73] 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.

[74] J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, December 1922.

[75] S. G. Bobkov. On concentration of distributions of random weighted sums. *The Annals of Probability*, 31(1):195–215, January 2003.

[76] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Math. Surv. Monogr*. American Mathematical Society (AMS), Providence, RI, 2001.

[77] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901.

[78] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[79] Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

[80] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data vs teacher-student paradigm. *arXiv preprint arXiv:1905.10843*, 2019.

[81] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[82] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, pages 1530–1538, Lille, France, July 2015. JMLR.org.

[83] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, pages 6241–6250, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[84] Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Joern-Henrik Jacobsen. Understanding and Mitigating Exploding Inverses in Invertible Neural Networks. pages 1792–1800. PMLR, March 2021.

[85] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

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

[87] 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, and Computing (Allerton)*, pages 601–608. IEEE, 2016.

[88] 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.

[89] 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.

[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 Mathematical arguments towards universality of Gaussian Mixtures

Setting and notations

We work in a setting slightly more general than the main text; the minimization problem we consider is of the form

\[ \hat{R}_n^*(X, y) = \inf_{\Theta \in S_p} \frac{1}{n} \sum_{\mu=1}^n \ell(\Theta^\top x_\mu, y_\mu) + r(\Theta), \]  

(12)

where the \( x_\mu \in \mathbb{R}^p \) are input vectors, \( y_\mu \in \mathbb{R} \) are one-dimensional labels, and \( S_p \) is a compact subset of \( \mathbb{R}^p \). We assume that the loss \( \ell \) only depends on the \( x_\mu \) through \( k \) one-dimensional projections \( \theta_1^\top x_\mu, \ldots, \theta_k^\top x_\mu \), and we work in the so-called proportional high-dimensional limit, where \( n, p \) go to infinity with

\[ \frac{n}{p} \to \alpha > 0, \]

while \( k \) stays fixed.

Throughout this section, \( \| \cdot \| \) will denote the spectral norm of a matrix, while \( \| \cdot \|_q \) for \( q > 0 \) will refer to the element-wise \( q \)-norms. For a subgaussian random variable \( Y \), its sub-gaussian norm \( \| Y \|_{\psi_2} \) is defined as

\[ \| Y \|_{\psi_2} = \inf\{ t > 0 \mid \mathbb{E}\left[ \exp\left( \frac{Y^2}{2t^2} \right) \right] \leq 2 \}. \]

A.1 State of the art

There have been many recent progress on Gaussian-type low-dimensional CLT and universality recently [35–37]. We shall leverage on these results to prove our first theorem.

In particular, the starting point for our mathematical proof will use the recent result of [37] which we shall now review. Consider the minimization problem (12), with \( (x_\mu, y_\mu) \) i.i.d random variables; the goal is to replace the \( x_\mu \) by their Gaussian equivalent model

\[ g_\mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma = \mathbb{E}[xx^\top]. \]

(13)

We make the following assumptions:

**Assumption 1** (Loss and regularization). The loss function \( \ell : \mathbb{R}^{k+1} \to \mathbb{R} \) is nonnegative and Lipschitz, and the regularization function \( r : \mathbb{R}^{p \times k} \to \mathbb{R} \) is locally Lipschitz, with constants independent from \( p \).

**Assumption 2** (Labels). The \( y_\mu \) are generated according to

\[ y_\mu = \eta(\Theta^* x_\mu, \epsilon_\mu), \]

(14)

where \( \eta : \mathbb{R}^{k^*+1} \to \mathbb{R} \) is a Lipschitz function, \( \Theta^* \in S_p^{k^*} \), and the \( \epsilon_\mu \) are i.i.d subgaussian random variables with

\[ \| \epsilon_\mu \|_{\psi_2} \leq M \]

for some constant \( M > 0 \).

**Assumption 3** (Concentration on the directions of \( S_p \)). We have

\[ \sup_{\Theta \in S_p, \| \Theta \|_2 \leq 1} \| \Theta^\top x \|_{\psi_2} \leq M \quad \text{and} \quad \sup_{\Theta \in S_p, \| \Theta \|_2 \leq 1} \| \Sigma \Theta \|_2 \leq M, \]

(15)

for some constant \( M > 0 \).

**Assumption 4** (One-dimensional CLT). For any bounded Lipschitz function \( \varphi : \mathbb{R} \to \mathbb{R} \),

\[ \lim_{p \to \infty} \sup_{\Theta \in S_p} \mathbb{E}\left[ \varphi(\Theta^\top x) - \varphi(\Theta^\top g) \right] = 0. \]

(16)

Building on those assumptions, [37] prove the following:
Theorem 5 (Theorem 1. in [37]). Suppose that Assumptions 1-4 hold. Then, for any bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$, we have
\[
\lim_{n,p \to \infty} \mathbb{E} \left[ \Phi \left( \hat{F}_n^*(X, y(X)) \right) \right] - \mathbb{E} \left[ \Phi \left( \hat{F}_n^*(G, y(G)) \right) \right] = 0
\]
In particular, for any $\rho \in \mathbb{R}$,
\[
\hat{F}_n^*(X, y(X)) \xrightarrow{p} \rho \quad \text{if and only if} \quad \hat{F}_n^*(G, y(G)) \xrightarrow{p} \rho
\]

A.2 Sketch of proof of Theorem 5, adapted from [37]

Free energy approximation Define the discretized free energy
\[
f_{\epsilon, \beta}(X) = \frac{1}{n \beta} \sum_{\Theta \in \mathbb{N}^n_\epsilon} \exp \left( -\beta \hat{F}_n(\Theta; X, y(X)) \right),
\]
where $\hat{F}_n(\Theta; X, y(X))$ is the quantity minimized in (12), and $\mathbb{N}_\epsilon$ is a minimal $\epsilon$-net of $S_p$. Using classical arguments from both the theory of $\epsilon$-nets and statistical physics, the authors show that
\[
\left| f_{\epsilon, \beta}(X) - \hat{F}_n(\Theta; X, y(X)) \right| \leq C_1(\epsilon) + \frac{C_2(\epsilon)}{\beta},
\]
and the same inequality holds for $G$. Since $C_1, C_2$ do not depend on $n, p$, it is possible to choose first $\epsilon$, then $\beta$ such that the RHS of (18) is as small as desired, and keep them fixed throughout the rest of the proof. We can therefore focus on studying the free energy approximation $f_{\epsilon, \beta}$ throughout the rest of the proof.

Interpolation path For any $0 \leq t \leq \pi/2$, define
\[
U_t = \cos(t)X + \sin(t)G
\]
Then $U_t$ is a smooth interpolation path with independent columns, ranging from $U_0 = X$ to $U_{\pi/2} = G$. We can write, for any differentiable function $\psi$,
\[
\left| \mathbb{E}[\psi(f_{\epsilon, \beta}(X))] - \mathbb{E}[\psi(f_{\epsilon, \beta}(G))] \right| \leq \int_0^{\pi/2} \left| \mathbb{E} \left[ \frac{d\psi(f_{\epsilon, \beta}(U_t))}{dt} \right] \right| dt,
\]
and by the dominated convergence theorem it suffices to show that the integrand converges to 0 for any $t$. The chain rule gives
\[
\frac{d\psi(f_{\epsilon, \beta}(U_t))}{dt} = \psi'(f_{\epsilon, \beta}(U_t)) \left( \sum_{\mu=1}^n \left( \frac{dU_t}{dt} \right)^\top \nabla u_{\mu, f_{\epsilon, \beta}(U_t)} \right),
\]
and the dependency in $\psi$ can be easily controlled. Since all columns of $U_t$ are i.i.d, we are left with showing
\[
\lim_{n,p \to \infty} n\mathbb{E}^{(1)} \left[ \left( \frac{dU_{t,1}}{dt} \right)^\top \nabla u_{1, f_{\epsilon, \beta}(U_t)} \right] = 0 \quad \text{a.s.,}
\]
where $\mathbb{E}^{(1)}$ denotes the expectation with respect to $(x_1, g_1, \epsilon_1)$.

Showing (19) Imagine for a moment that $x_1$ is Gaussian; then $u_{t,1}$ and $dU_{t,1}/dt$ are also jointly Gaussian, and we have
\[
\mathbb{E} \left[ \left( \frac{dU_{t,1}}{dt} \right)^\top u_{t,1} \right] = \mathbb{E} \left[ (-\sin(t)x_1 + \cos(t)g_1)^\top (\cos(t)x_1 + \sin(t)g_1) \right]
\]
\[
= 0,
\]
since $x_1$ and $g_1$ have the same covariance by definition. Therefore, they are independent, and we have
\[
\mathbb{E}^{(1)} \left[ \left( \frac{dU_{t,1}}{dt} \right)^\top \nabla u_{1, f_{\epsilon, \beta}(U_t)} \right] = \mathbb{E}^{(1)} \left[ \left( \frac{dU_{t,1}}{dt} \right)^\top \right] \mathbb{E}^{(1)} \left[ \nabla u_{1, f_{\epsilon, \beta}(U_t)} \right] = 0.
\]
On the other hand, it is possible to show that $x_1$ only appears in (19) through scalar products with $\Theta$ or $\Theta^*$. As a result, we can leverage Assumption 4 to replace $x_1$ by a Gaussian vector $w$ independent from $g_1$ as $p \to \infty$. Then, the reasoning above can be repeated with $w$ and $g_1$ to conclude the proof.
A.3 Proof of Theorem 1

In order to prove our theorem 1, we now aim to adapt the proof from [37] to the case where the distribution of $x$ can be a mixture of several other distributions, each with different mean and covariance. For a discrete set $C = \{1, \ldots, K\}$, we consider a family of distributions $(\nu_c)_{c \in C}$ on $\mathbb{R}^p$, with means and covariances

$$\mu_c = \mathbb{E}_{z \sim \nu_c}[z] \quad \text{and} \quad \Sigma_c = \mathbb{E}_{z \sim \nu_c}[zz^\top]$$

Given a type assignment $\sigma : [n] \to C$, each sample $x_\mu$ is then drawn independently from $\nu_{\sigma(\mu)}$. The equivalent Gaussian model is straightforward: we simply take $g_i \sim N(\mu_{\sigma(\mu)}, \Sigma_{\sigma(\mu)})$, independently from each other. An important special case of this setting is when $\sigma$ is itself random, independently from the $x_i$ and $g_i$: the law of $g_i$ is then a so-called Gaussian Mixture Model.

The assumptions of Theorem 5 are modified as follows:

(i) Assumption 1 is unchanged,

(ii) We relax (14) in Assumption 2 into

$$y_i = \eta_{\sigma(i)}(\Theta^* x_i, \varepsilon_i),$$

for a family $(\eta_c)_{c \in C}$ of Lipschitz functions. This allows in particular to incorporate classification problems in our setting, at no cost in the proof complexity.

(iii) We impose in Assumption 3 the stronger condition

$$\sup_{\theta \in \mathcal{S}_p, \|\theta\|_2 \leq 1} \|\theta^\top g\|_{\psi_2} \leq M,$$

which is a consequence of (15) when $g$ has zero-mean. This is in practice an additional condition on the means $\mu_c$; indeed, (20) is equivalent to

$$\sup_{\theta \in \mathcal{S}_p, \|\theta\|_2 \leq 1} \langle \mu_c, \theta \rangle \leq M' \quad \text{and} \quad \sup_{\theta \in \mathcal{S}_p, \|\theta\|_2 \leq 1} \|\Sigma_c \theta\|_2 \leq M',$$

for some different constant $M'$.

(iv) We suppose that Assumptions 3 and 4 hold for any possible distribution $\nu_c$ for $c \in C$ and its associated Gaussian equivalent model.

Our result is then an extension of Theorem 5 for mixtures:

**Theorem 1.** Suppose that the modified above assumptions hold. Then, for any bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\lim_{n, p \to \infty} \left| \mathbb{E} \left[ \Phi \left( \hat{R}_n^*(X, y(X)) \right) \right] - \mathbb{E} \left[ \Phi \left( \hat{R}_n^*(G, y(G)) \right) \right] \right| = 0$$

In particular, for any $\rho \in \mathbb{R}$,

$$\hat{R}_n^*(X, y(X)) \overset{p}{\to} \rho \quad \text{if and only if} \quad \hat{R}_n^*(G, y(G)) \overset{p}{\to} \rho$$

We now go through the proof of the previous section, highlighting the important changes.

**Free energy approximation** This section goes basically unchanged; the approximation between $\hat{R}_n^*(X, y(X))$ and $f_{\epsilon, \beta}(X)$ relies on Lipschitz arguments and concentration bounds on the $x_i$ and $g_i$, which are satisfied by our modification of Assumption 3.
Interpolation path  Recall that the important property of $U_t$ is that
\[ E \left[ \left( \frac{dU_t}{dt} \right)^\top U_t \right] = 0. \] (22)

To this end, we set
\[ u_{t,\mu} = \mu \sigma(\mu) + \cos(t)(x_{\mu} - \mu \sigma(\mu)) + \sin(t)(g_{\mu} - \mu \sigma(\mu)), \]
and it is easy to check that (22) is satisfied.

Another problem is that the columns of $U_t$ are not i.i.d anymore, so we have to control
\[ \frac{1}{n} \sum_{\mu=1}^{n} \left| E_{(\mu)} \left[ \left( \frac{du_{t,\mu}}{dt} \right)^\top \nabla u_{t,\mu} f_{e,\beta}(U_t) \right] \right|, \] (23)
where this time $E_{(\mu)}$ is the expectation w.r.t $(x_{\mu}, g_{\mu}, \varepsilon_{\mu})$. However, (23) is a weighted average over all values of $\sigma(\mu)$, and since $\mathcal{C}$ is finite is suffices to show (19) for any value of $\sigma(1)$.

Showing (19) This section again relies on concentration properties of the $x_i$ and $g_i$, as well as Assumption 4. The arguments thus translate directly from [37].

A.4 One-dimensional gaussian approximation

Although Theorem 5 is a powerful result, it still relies on very strong assumptions. In particular, given a distribution $\nu$ for the inputs $x_i$, characterizing the set of vectors $\theta$ such that Assumption 4 holds is in general a difficult task.

Rigorous results When the entries of $x$ are i.i.d subgaussian, a classical application of the Lindeberg method [74] shows that Assumptions 3 and 4 are satisfied with
\[ S_p = \{ \theta \in \mathbb{R}^p | \|\theta\|_\infty = o_p(1) \}. \]
More recently, this result (often used under the name “Gaussian Equivalence Theorem”) was extended to feature models [35] as well as to and to general feature models with approximate orthogonality constraints [35, 36], for the same choice of $S_p$. [37] also provides a central limit theorem result for the Neural Tangent Kernel [28], for a more convoluted parameter set $S_p$. While these papers provide a strong basis for the one-dimensional CLT, those rigorous results only concern (so far) a very restricted set of distributions.

Concentration of the norm Another, more informal line of work originating from [7], argues that most distributions found in the real world satisfy some form of central limit theorem. The starting point of this analysis is the following theorem, adapted from [75]:

Theorem 6 (Corollary 2.5 from [75]). Let $x \in \mathbb{R}^p$ be a random variable, with $E[xx^\top] = I_p$, and $\eta_p$ the smallest positive number such that
\[ P \left( \left| \frac{\|x\|_2^2}{\sqrt{p}} - 1 \right| \geq \eta_p \right) \leq \eta_p. \] (24)
Then for any $\delta > 0$, there exists a subset $S_p$ of the $p$-sphere $\mathbb{S}^{p-1}$ of measure at least $4p^3/8 e^{-c p^2}$, such that
\[ \sup_{\theta \in S_p} \sup_{t \in \mathbb{R}} |P(\theta^\top x \geq t) - \Phi(t)| \leq \delta + 4 \eta_p, \]
where $\Phi$ is the characteristic function of a standard Gaussian, and $c$ is a universal constant.

If both $\delta$ and $\eta_p$ are $o(1)$, Theorem 6 implies that Assumption 4 is satisfied for any compact subset $S_p^c \subseteq S_p$. This suggests that the norm concentration property of (24) is a convenient proxy for one-dimensional CLTs. However, the proof of this theorem uses isoperimetric inequalities, and is thus non-constructive; as a result, characterizing precisely the set $S_p$ remains an open and challenging mathematical problem.
Concentrated vectors In [7], the authors consider the concept of concentrated random variables, as defined in [76]:

**Definition 1.** Let $x \in \mathbb{R}^p$ be a random vector. $x$ is called (exponentially) concentrated if there exists two constants $C, c$ such that for any 1-Lipschitz function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, we have

$$\mathbb{P}(\|f(x) - \mathbb{E}[f(x)]\| \geq t) \leq Ce^{-ct^2}.$$ 

Since the norm function is 1-Lipschitz, it can be shown that any concentrated isotropic vector $x$ satisfies (24), with

$$\eta_p \propto \left(\frac{\log(p)}{p}\right)^{1/2}.$$ 

The converse is obviously not true; an exponential random vector still has $\eta_p \rightarrow 0$, but is not concentrated. However, even if it is stronger than (24), the concept of concentrated vectors has two important properties:

(i) a standard Gaussian vector $x \sim \mathcal{N}(0, I_p)$ satisfies Definition 1 with constants $C, c$ independent from $p$,

(ii) if $x \in \mathbb{R}^p$ is a concentrated vector with constants $C, c$ and $\Psi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is an $L$-Lipschitz function, then $\Psi(x)$ is also a concentrated vector, with constants only depending on $c, C$ and $L$.

Towards real-world datasets The real-world data considered in machine learning is often composed of very high-dimensional inputs, corresponding to $p \gg 1$ in our setting. However, it is generally accepted that this data actually lies on a low-dimensional manifold of dimension $d_0$; this is the idea behind many dimensionality reduction techniques, from PCA [77] to autoencoders [78]. Another, more recent line of work (see e.g. [79]) studies the estimation of the latent dimension $d_0$; results for the MNIST dataset ($p = 784$) yield $d_0 \approx 15$, while CIFAR-10 ($p = 3072$) has estimated intrinsic dimension $d_0 \approx 35$ [80].

Following this heuristic, the most widely used method to model realistic data is to learn a map $f : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^p$, usually through a deep neural network, and then generate the $x_i$ according to

$$x = f(z) \text{ with } z \sim \mathcal{N}(0, I_{d_0}) \quad (25)$$

Examples of functions $f$ include GANs [67], variational auto-encoders [81], or normalizing flows [82]. This ansatz has been studied theoretically, and the results compared with real-world datasets, in [6, 10]; the results indicate significant agreement between generated inputs and actual data.

Finally, we argue that for a large class of generative networks, the learned function $f$ is actually Lipschitz, with a bounded constant. This is even often a design choice; indeed, theoretical results such as [83] imply that a smaller Lipschitz constant improve the generalization capabilities of a network, or its numerical stability [84]. As a result, regularizations aimed at controlling the Lipschitz properties of a network are a common occurrence; see e.g. [85] for the spectral regularization of GANs. This indicates that concentrated vectors are indeed a good approximation for real-world data.
B  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 main body: the Gaussian covariate model (GCM) and the Gaussian mixture model (GMM).

B.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:

$$\hat{y}(x) = \text{sign}(\hat{\theta}^\top x)$$  \hspace{1cm} (26)

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

$$\hat{R}_n^\star(X, y) = \inf_{\theta \in S_p} \frac{1}{n} \sum_{\mu=1}^{n} \ell(\theta^\top x_\mu, y_\mu) + \frac{\lambda}{2} ||\theta||^2_2, \hspace{1cm} (27)$$

for a compact subset $S_p \subset \mathbb{R}^p$ and convex loss function $\ell$. 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 $x_\mu$),

$$y_\mu \sim \frac{1}{2} (\delta_{-1} + \delta_{+1}), \text{ i.i.d.} \hspace{1cm} (28)$$

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

**Gaussian covariate model (GCM):** $x_\mu \sim N(0_p, \Sigma)$, 

**Gaussian mixture model (GMM):** $x_\mu \sim \sum_{c \in C} \rho_c N(\mu_c, \Sigma_c)$,

where $C = \{1, \ldots, K\}$ is the label set for the Gaussian clouds and $\rho_c \in [0, 1]$ are the density of points in each class, satisfying $\sum_{c \in C} \rho_c = 1$. Note that in this random label setting the GCM model is a special case of the GMM, where $K := |C| = 1$ and $\mu_1 = 0_p$.

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 $\alpha := \frac{n}{p}$ and the number of clusters $K$ are fixed:

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

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

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

**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:

$$||\mu_c||^2_2 = O(1), \hspace{1cm} \text{tr} \Sigma_c = O(p). \hspace{1cm} (29)$$

Note that this is precisely the assumption 3 used in theorem 1 (see eq. (20)) and in theorem 7.

This scaling of the mean and variance is indeed the natural one (see e.g. [64, 86–89]) 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).
Ridge and ordinary least-squares classification – Note that for the special case of the ridge classification in which
\( \ell(x, y) = \frac{1}{2}(y - x)^2 \), the empirical risk minimization problem defined in eq. (27) admits a closed form solution:
\[
\hat{\theta} = (\lambda I_p + XX^\top)^{-1} X y
\]
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 \( \frac{1}{2}\lambda |\theta|^2 \). This can only be achieved in the limit of vanishing regularization \( \lambda \to 0^+ \), in which case the ridge estimator simplifies to:
\[
\hat{\theta}_{\text{ols}} := (X^\top)^+ y
\]
where \( X^\top \in \mathbb{R}^{n \times p} \) is the Moore-Penrose inverse of \( X \). In the simplest case in which \( X \) is a full-rank matrix (which ultimately depends on the covariances), it can be explicitly written as:
\[
X^\top := \begin{cases} 
(X^\top X)^{-1} X^\top & \text{if } \alpha < 1 \\
X^\top (XX^\top)^{-1} & \text{if } \alpha > 1
\end{cases}
\]
An important property of the estimator in eq. (31) is that it corresponds to the least \( \ell_2 \)-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. (27) is equivalent to inverting a linear system:
\[
y = X^\top \theta
\]
i.e. to solve a system of \( n \) equations for \( p \) unknowns. Again, assuming the data is full-rank\(^1\), 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, \( \hat{\theta}_{\text{ols}} \) is the one that has lowest \( \ell_2 \)-norm. Instead, when \( \alpha > 1 \), the system is overdetermined, and no interpolating (zero-loss) solution exists.

B.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 [61–66]. Of particular interest to our work are the formulas proved in [40] 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. (27) was proven in the case where \( \xi \sim N(0, 1) \). Theorem 7 (for random labels, informal) In the proportional high-dimensional limit where \( n, p \to \infty \) with fixed \( \alpha = n/p \) and \( K = |\mathcal{C}| \), the training performance of the empirical risk minimizer eq. (27) for the random label Gaussian mixture model satisfying the scalings (29) are given by:
\[
\hat{E}_\ell \to \hat{E}_\ell^{\text{gmm}}(\alpha, \lambda, K) := \frac{1}{2} \sum_{c \in \mathcal{C}} \sum_{y \in \{-1, +1\}} \mathbb{E}_{\xi \sim N(0, 1)} \left[ \ell\left( \text{prox}_{\ell_1}^\ell (\cdot, y) \left( \frac{m_c^* + \sqrt{q_c} \xi}{\sqrt{p_c}} \right), y \right) \right]
\]
\[
\hat{E}_{0/1} \to \hat{E}_{0/1}^{\text{gmm}}(\alpha, \lambda, K) := \frac{1}{2} \sum_{c \in \mathcal{C}} \sum_{y \in \{-1, +1\}} \mathbb{E}_{\xi \sim N(0, 1)} \left[ \mathbb{P}\left( \text{prox}_{\ell_1}^\ell (\cdot, y) \left( \frac{m_c^* + \sqrt{q_c} \xi}{\sqrt{p_c}} \right) \neq y \right) \right]
\]
where \( \ell \) is the loss function used in the empirical risk minimization in eq. (27), \( \text{prox}_{\ell_1}(\cdot) \) is the proximal operator associated with the loss:
\[
\text{prox}_{\ell_1}(\cdot, y) := \arg \min_{z \in \mathbb{R}} \left[ \frac{1}{2\tau} (z - x)^2 + \ell(z, y) \right]
\]
\(^{1}\)The general case is given by changing \( p \) for the rank of the design matrix.
and \((m^*_c, V^*_c, q^*_c)_{c \in \mathcal{E}}\) are the unique fixed points of the following self-consistent equations:

\[
\begin{align*}
V_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c'} \Sigma_{c''} \right)^{-1} \\
q_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{q}_{c'} \Sigma_{c'} - \mu_c \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \\
m_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{m}_{c'} \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1}
\end{align*}
\]

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

**Zero mean limit:** Of particular interest for what follows is the zero-mean limit \(\mu_c = 0_p\) of the above equations, which is simply given by:

\[
\begin{align*}
\hat{V}_c &= \frac{2}{p} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ \partial_\omega f_\ell(y, m_c + \sqrt{\gamma_c} \xi, V_c) \right] \\
\hat{q}_c &= \frac{2}{p} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ f_\ell(y, m_c + \sqrt{\gamma_c} \xi, V_c)^2 \right] \\
\hat{m}_c &= \frac{2}{p} \sum_{y \in \{-1,+1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ f_\ell(y, m_c + \sqrt{\gamma_c} \xi, V_c) \right] \\
V_c &= \frac{1}{p} \text{tr} \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \\
q_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{q}_{c'} \Sigma_{c'} - \mu_c \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \\
m_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{m}_{c'} \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1}
\end{align*}
\]

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

\[
\text{prox}_{\tau \ell(-y)}(x) = \frac{x + \tau y}{1 + \tau} \iff f_\ell(y, \omega, V) = \frac{y - \omega}{1 + V}
\]

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

\[
e^{\text{gmm}}_\ell = \sum_{c \in \mathcal{E}} \frac{1}{2} \frac{1 + q^*_c}{2(1 + V^*_c)^2}
\]

for \((V^*_c, q^*_c)_{c \in \mathcal{E}}\) solutions of the following simplified self-consistent equations:

\[
\begin{align*}
\hat{V}_c &= \frac{\alpha \rho_c}{1 + \rho_c V_c} \\
\hat{q}_c &= \frac{\alpha \rho_c}{1 + \rho_c V_c} \left( 1 + q^*_c \right) \\
\hat{m}_c &= \frac{1}{\alpha \rho_c} \frac{1 + q^*_c}{\left( 1 + \rho_c V_c \right)}
\end{align*}
\]

\[
\begin{align*}
V_c &= \frac{1}{p} \text{tr} \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \\
q_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{q}_{c'} \Sigma_{c'} - \mu_c \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1} \\
m_c &= \frac{1}{p} \sum_{c' \in \mathcal{E}} \hat{m}_{c'} \mu_c^\top \Sigma_c \left( \lambda I_p + \sum_{c'' \in \mathcal{E}} \hat{V}_{c''} \Sigma_{c''} \right)^{-1}
\end{align*}
\]

Note that in particular, at the fixed point, we can also express the training loss eq. (39) as:

\[
e^{\text{gmm}}_\ell = \sum_{c \in \mathcal{E}} \frac{\hat{q}^*_c}{\alpha}
\]
B.3 Gaussian covariate model

The asymptotic training loss for the Gaussian covariate model for a fairly general teacher–student setting was first proven in [10]. Although the random label limit can be obtained from this work, as discussed in Sec. B.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 $\mu_1 = 0$. Therefore, its asymptotic performance is included in the discussion above. Despite the redundancy, for clarity of the argument that will follow, we present it here as an independent case.

**Theorem 8 ([10] for random labels, informal).** In the proportional high-dimensional limit introduced above, the training loss and errors for the GCM model are given by:

$$
\hat{E}_f \rightarrow E^\text{gcm}_f(\alpha, \lambda) := \frac{1}{2} \sum_{y \in \{-1, 1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ \ell(\text{prox}_{V^*}(\cdot,y) (\sqrt{q^*} \xi), y) \right]
$$

$$
\hat{E}_{0/1} \rightarrow E^\text{gcm}_{0/1}(\alpha, \lambda) := \frac{1}{2} \sum_{y \in \{-1, 1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ \mathbb{P}(\text{prox}_{V^*}(\cdot,y) (\sqrt{q^*} \xi) \neq y) \right]
$$

where the proximal is given as before, and the parameters $(V^*, q^*)$ are the (unique) fixed-point of the following self-consistent equations:

$$
\begin{cases}
\hat{V} = \frac{q^*}{2} \sum_{y \in \{-1, 1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ \partial_x f_x(y, \sqrt{q^*} \xi, V) \right] \\
\hat{q} = \frac{q^*}{2} \sum_{y \in \{-1, 1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[ f_x(y, \sqrt{q^*} \xi, V)^2 \right]
\end{cases}, \\
\begin{cases}
V = \frac{1}{p} \text{tr } \Sigma \left( \lambda I_p + \hat{V} \Sigma \right)^{-1} \\
q = \frac{1}{p} \hat{q} \text{tr } \Sigma^2 \left( \lambda I_p + \hat{V} \Sigma \right)^{-2}
\end{cases}
$$

where, just as before $f_\beta(y, \omega, V) := V^{-1} \left( \text{prox}_{V^*}(\cdot,y)(\omega) - \omega \right)$.

**A particular case: ridge classification**  As before, for the square loss the expressions simplify considerably. The asymptotic training loss for the Gaussian covariate model for a fairly general teacher–student setting was first proven.

**Theorem 3.** The fact that the loss is independent of $\Sigma$ in this regime can be directly seen from the optimization.

$$
\mathbb{E}_f^\text{gcm}(\alpha, \lambda) = \begin{cases}
0 & \text{for } \alpha \leq 1 \\
\frac{1}{2} \left( 1 - \frac{1}{\alpha} \right) & \text{for } \alpha > 1
\end{cases}
$$

\[\]
C From Gaussian mixture to single Gaussian

C.1 Mixture of Gaussians with zero means

We first prove lemma 1 in the main text. In view of the asymptotic expressions of $\mathcal{E}_{\ell}^{gmm}$ in Theorem 7, and the unicity of the solution to Eqs (36), it suffices to show the following:

**Lemma 2.** Let $(V^*_c, q^*_c)_{c \in \mathcal{C}}$ be the solutions of Eqs. (37). Then, $(0, V^*_c, q^*_c)_{c \in \mathcal{C}}$ satisfy the general fixed point equations of (36).

**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 (36) and (37). It is also easy to check that $\hat{m}_c = 0$ for all $c$ implies that $\hat{m}_c = 0$; what remains is to show that the last equation holds, i.e.

$$\frac{\alpha}{2} \rho_c \sum_{\omega \in \{-1,+1\}} \mathbb{E}_{\xi \sim N(0,1)} \left[f_{\ell}(y, \sqrt{q^*_c} \xi, V^*_c)\right] = 0. \quad (49)$$

Define the function

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

so that

$$\hat{m}_c^* \propto \mathbb{E}_{\xi \sim N(0,1)} \left[g(\sqrt{q^*_c} \xi, V^*_c)\right].$$

We shall show that $g$ is odd in $\omega$; since $\xi$ 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

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

and the linear term in $\omega$ is immediate. For the proximal operator, we use the symmetry of $\ell$ and write

$$\text{prox}_{f_{\ell}(\cdot, y)}(\omega) = \arg \min_{z \in \mathbb{R}} \left[\frac{1}{2\tau} (z - \omega)^2 + \ell(z, y)\right]$$

$$= \arg \min_{z \in \mathbb{R}} \left[\frac{1}{2\tau} ((-z) - (-\omega))^2 + \ell(-z, -y)\right]$$

$$= -\text{prox}_{f_{\ell}(\cdot, -y)}(-\omega),$$

which concludes the proof. \qed

C.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 4 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 case, the training loss is given by eq. (39) with $(V^*_c, q^*_c)_{c \in \mathcal{C}}$ unique solutions of the following equations:

$$\begin{aligned}
\hat{V}_c &= \frac{\alpha \rho_c}{1 + \hat{q}_c} V^*_c, \\
\hat{q}_c &= \frac{\alpha \rho_c}{1 + \hat{q}_c} (1 + \hat{q}_c)^2.
\end{aligned} \quad (50)$$

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:

$$\mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[V^*_c \hat{V}_c^*\right] = 1. \quad (51)$$
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:

\[
\sum_{c \in \mathcal{C}} \left( \hat{V}_c^* q_c^* - V_c^* \hat{q}_c^* \right) = 0 \tag{52}
\]

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:

\[
\sum_{c \in \mathcal{C}} \rho_c \frac{V_c^*}{1 + V_c^*} = \frac{1}{\alpha} \tag{53}
\]

\[
\sum_{c \in \mathcal{C}} \rho_c \frac{V_c - q_c}{(1 + V_c)^2} = 0 \tag{54}
\]

We thus find, combining the eq. (50) for \( \hat{q}_c \) with eq. (54)

\[
\sum_{c \in \mathcal{C}} \hat{q}_c^* = \sum_{c \in \mathcal{C}} \alpha \rho_c \frac{1 + V_c^*}{(1 + V_c^*)^2} = \sum_{c \in \mathcal{C}} \alpha \rho_c \frac{1}{1 + V_c^*} \tag{55}
\]

Our goal is to evaluate the loss at the fixed point, which is given by eq. (41):

\[
\mathcal{E}_{gmm} = \sum_{c \in \mathcal{C}} \frac{\hat{q}_c^*}{2\alpha} \tag{56}
\]

Combining this definition with eqs. (54) and (55), we find that

\[
2\mathcal{E}_{gmm} + \frac{1}{\alpha} = \sum_{c \in \mathcal{C}} \rho_c \frac{1}{1 + V_c^*} + \sum_{c \in \mathcal{C}} \rho_c \frac{V_c^*}{1 + V_c^*} = 1 \tag{57}
\]

so that finally, we reach the promised result:

\[
\lim_{\lambda \to 0^+} \mathcal{E}_{gmm}(\alpha, \lambda, K) = \frac{1}{2} \left( 1 - \frac{1}{\alpha} \right) = \lim_{\lambda \to 0^+} \mathcal{E}_{gcm}(\alpha, \lambda) \tag{58}
\]

as claimed in Theorem 4 in the main.
D  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.

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. 4. For the sake of clarity, we summarized it through the pseudo code in algorithm 1.

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

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

The real datasets are loaded through Pytorch data loaders [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.

Learning phase. Given the dataset generated as in algorithm 1, the aim is to infer the estimator $\theta$ 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:

$$\hat{R}^n_2(X, y) = \inf_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{\mu=1}^n (\theta^\top x^\mu - y^\mu)^2 + \frac{\lambda}{2} ||\theta||_2^2,$$  \hspace{1cm} (59)

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

$$\theta = \begin{cases} (X^\top X + \lambda I_p)^{-1} X^\top y, & \text{if } n > p \\ X^\top (XX^\top + \lambda I_n)^{-1} y, & \text{if } p > n \end{cases}$$  \hspace{1cm} (60)

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

$$\hat{R}^n_2(X, y) = \inf_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{\mu=1}^n \log (1 + \exp(-y^\mu \theta^\top x^\mu)) + \frac{\lambda}{2} ||\theta||_2^2,$$  \hspace{1cm} (61)

Since the estimator of logistic regression can not be determined through an explicit closed formula, we here made use of the **lbfgs** solver with **penalty** set to $\ell_2$. 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 $1e4$. 

26
(iii) **Hinge Loss.** In this specific case, the goal is to solve the following optimization problem:

\[
\hat{R}_n^*(X, y) = \inf_{\theta \in S_{p}} \frac{1}{n} \sum_{\mu=1}^{n} \max \left( 0, 1 - y_\mu \theta^\top x_\mu \right) + \frac{\lambda}{2} ||\theta||_2^2,
\]

As for logistic regression, even in this case we cannot 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 \textit{penalty} is set to \(\ell_2\). 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 \textit{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 \(\ell\) 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 realizations of the learning problem, as exemplified in algorithm 2.

**Algorithm 2 Learning curve**

\begin{algorithm}
\textbf{Input:} range of \(n/p\), flag \textit{dataset type}, flag \textit{which estimator}
\For{seed in a specified number of seeds} do:
  \For{\(n/p\) in a specified range} do:
    Choose the dataset according to \textit{dataset type};
    Compute the estimator according to the desired optimization problem as in (i)-(iii);
    Compute the training loss \(\ell\) 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 \(n/p\).}
\end{algorithm}
E Empirical evidence of the homogeneity 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. 4 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).

Figure 4: 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.

As can be seen in Fig. 5 and Fig. 6, 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).

Figure 5: 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.
Figure 6: 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.