Are Gaussian data all you need? Extents and limits of universality in high-dimensional generalized linear estimation

Luca Pesce, Florent Krzakala, Bruno Loureiro, Ludovic Stephan

Are Gaussian data all you need? Extents and limits of universality in high-dimensional generalized linear estimation

Luca Pesce[1], Florent Krzakala[1], Bruno Loureiro[2], and Ludovic Stephan[1]

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL). Information, Learning and Physics (IdePHICS) lab.
CH-1015 Lausanne, Switzerland.
[2]Département d'Informatique, École Normale Supérieure - PSL & CNRS, 45 rue d'Ulm,
F-75230 Paris cedex 05, France.

Abstract

In this manuscript we consider the problem of generalized linear estimation on Gaussian mixture data with labels given by a single-index model. Our first result is a sharp asymptotic expression for the test and training errors in the high-dimensional regime. Motivated by the recent stream of results on the Gaussian universality of the test and training errors in generalized linear estimation, we ask ourselves the question: "*when is a single Gaussian enough to characterize the error?*". Our formula allow us to give sharp answers to this question, both in the positive and negative directions. More precisely, we show that the sufficient conditions for Gaussian universality (or lack of thereof) crucially depend on the alignment between the target weights and the means and covariances of the mixture clusters, which we precisely quantify. In the particular case of least-squares interpolation, we prove a strong universality property of the training error, and show it follows a simple, closed-form expression. Finally, we apply our results to real datasets, clarifying some recent discussion in the literature about Gaussian universality of the errors in this context.

1 Introduction

It is commonsense in machine learning that structure in the data is an important ingredient for successful learning. Quantifying this statement, and in particular how structure in the features impact the training and generalization errors the most, is an important endeavor in the broad program of "seeing through" the modern machine learning black box. On the theoretical side, there has been some important recent progress in this direction in the context of generalized linear estimation. For instance, a recent line of work on linear regression trained on Gaussian data has shown that good generalization can arise even in the "overparametrized regime" where the training error is exactly zero [1–3]. This *benign overfitting* property crucially depends on the covariance structure, occurring when the signal components of the target align with a lower-dimensional of the data, leaving space for the noise to spread along the higher-dimensional orthogonal subspace [1]. Analogous conclusions hold, under similar conditions, to generalized linear tasks [4–6]. Indeed, this is only one example of many surprising insights learned from the study of generalized linear models on Gaussian data over the past few years [7–10]. Despite the seemingly constraining assumption on the distribution of the features, a recent line of work provides strong evidence for the *Gaussian universality* of the training and generalization errors in generalized estimation in different settings. These includes rigorous results for non-Gaussian designs [11, 12], random feature maps [13–16], neural tangent features [17], Gaussian mixtures with random labels [18], as well as extensive numerical evidence for other feature maps [15, 19] and even real datasets [19–21]. These works beg the question *"when are Gaussian features a good model for learning?"*. Our aim is to give precise answers to this question in the context of generalized linear estimation on a popular model for multi-modal data, known to be able to approximate any distribution: the Gaussian mixture model.

1.1 Main results

Our **main contributions** in this work are as follows:

• **Exact asymptotics of GLMs:** We provide the exact asymptotic limit for the training and test errors of a generalized linear model with convex loss in high dimensions, when the data is drawn from a Gaussian mixture model with a single-index target. These asymptotics are based on the so-called *replica method* from statistical physics, and follow the same line as [22], which considered instead the task of learning the mixture labels.

• **Universality of training and test errors:** We provide a set of sufficient conditions on the target weights θ_0 such that both the asymptotic training and test errors for a Gaussian mixture model are independent from the cluster means. In particular, these conditions are satisfied by a target whose direction is uniform on $\mathbb{S}^{d-1}$. In the case of ridge regression, we show an even stronger result: namely, the training loss is also independent from the cluster covariances, and reduces to that of a single Gaussian with identity covariance.

• **The importance of Homoscedasticity**: In the particular case of a homoscedastic Gaussian mixture (a mixture of Gaussians that share the same covariance matrix), we further demonstrate universality results that can actually be observed on real data after a random feature map (see e.g. Fig. 1 and 2). We also unveil the universal behavior of the linear separability transition, a phenomenon studied in detail for pure Gaussian data in [10] and that appears to be universal for a homoscedastic mixture.

• **Breaking universality:** In contrast to the results of the previous paragraph, we show that there are two ways to break Gaussian universality. First, strongly heteroscedasticity can break the universal behavior. Second, in the homoscedastic case we show that the correlation between the data and the task matters: even a small correlation between the target weights and the cluster means suffice to break universality, in the sense that the asymptotic errors of a model trained in a Gaussian mixture differs from the one of a model trained on Gaussian data. Rather than the structure of the data itself, what appears to matter is thus the correlation between this structure and the task to be learned.

1.2 Related works

Exact asymptotics: An appealing feature of Gaussian data is that the asymptotic performance of different models can be sharply characterized in the proportional high-dimensional limit where $n, d \to \infty$ at fixed *sample complexity* $\alpha := {}^{n}/_{d}$. This is particularly the case for ridge regression, because of the close connection to a random matrix theory problem; see e.g. [2, 3, 23]. Beyond the quadratic case, there exists many asymptotic rigorous studies, for instance [24] studied M-estimators, [25] the performance of the LASSO estimator, while [19] provided a general result for convex losses and penalties with arbitrary covariances.

In the case of Gaussian mixtures, most of the effort has been geared towards classification, i.e. recovering the cluster label instead of teacher-generated ones. For binary classification, examples include [26–28], the latter of which also shows an equivalence between classification and a single-index model. In the multi-class setting, [29] studied the performance of ridge regression classifiers; the most general result in this line is [22], which considers any convex (not necessarily separable) loss.

Gaussian universality: Remarkably, this model is also able to capture the errors of particular classes of non-Gaussian features. This *Gaussian universality property* (GEP) [30] was proven to hold for generalized linear estimation with random features [13–16, 31], neural tangent features [17] and kernel features [21, 32–35]. [36] showed that Gaussian universality is preserved on the random features model when the weights are trained at order one steps, but break if an extensive number of steps are taken. Beyond the realm of theorems, [19] provided numerical evidence of Gaussian universality for a broader class of realistic features from trained neural networks. On a close line, [18] studied the Gaussian universality for pure random binary labels, which was an important source of inspiration for the present work, while [19–21, 34, 37, 38] has numerically shown that for ridge regression in particular, the Gaussian formula captured the learning curves of some simple real datasets.

2 Setting & motivation

Let $(\boldsymbol{x}^\nu, y^\nu) \in \mathbb{R}^d \times \mathcal{Y}$ denote $\nu = 1, \cdots, n$ pairs of independently sampled training points. We shall be interested in studying the properties of generalized linear estimation $\hat{y}(\boldsymbol{x}) = \hat{f}(\boldsymbol{\theta}^\top \boldsymbol{x})$ with weights $\boldsymbol{\theta} \in \mathbb{R}^d$ learned from the training data by minimizing the following empirical risk:

$$\hat{\mathcal{R}}_n^\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^\nu, \frac{\boldsymbol{\theta}^\top \boldsymbol{x}}{\sqrt{d}}\right) + \lambda r(\boldsymbol{\theta}) \tag{1}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function and $r : \mathbb{R}^d \to \mathbb{R}$ is a convex penalty. For example, this includes the particular case of ridge regression where $\ell(y, \hat{y}) = (y - \hat{y})^2$ and $r(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_2^2$. The key quantities of interested in the

following will be the training and generalization error, defined as:

$$\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^{\nu}, \frac{\hat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}}\right) \tag{2}$$

$$\varepsilon_{\text{gen}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[g\left(y_{\text{new}}, \hat{f}\left(\frac{\hat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}_{\text{new}}}{\sqrt{d}}\right)\right)\right] \tag{3}$$

where $g : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a performance metric of the choice of the statistician, not necessarily equal to the loss function $\ell(y, \hat{y})$. For example, in the case of binary classification with $\mathcal{Y} = \{-1, +1\}$, we can take $\ell(y, \hat{y}) = \log(1 + e^{-y\hat{y}})$ to be the logistic loss, while taking $g(y, \hat{y}) = \mathbb{I}[y \neq \hat{y}]$ to be the classification error. Note that in eq. (3) the expectation is taken over a new data pair $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ which we assume is independently drawn from the same distribution of the training data.

In particular, we will be interested in characterizing these errors under the assumption that the labels have been generated by the following target distribution:

$$y^{\nu} \sim P_0\left(\cdot \,\middle|\, \frac{\boldsymbol{\theta}_0^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}}\right) \tag{4}$$

for some fixed vector $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and distribution P_0. A common choice for P_0 is $y^{\nu} = f_0\left(\frac{\boldsymbol{\theta}_0^{\top}\boldsymbol{x}^{\nu}}{\sqrt{d}} + \xi\right)$ for additive Gaussian noise $\xi \sim \mathcal{N}(0, \Delta)$. This setting is sometimes refereed to as a *teacher-student setting*. We will sometimes adopt this convenient terminology, refereeing to the target distribution P_0 as the *teacher* and $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ as the *teacher weights*. Similarly, we will sometimes refer to the model $\hat{y}$ as the *student* and $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ the *student weights*. As previously mentioned, we shall be considering both regression $\mathcal{Y} = \mathbb{R}$ and binary classification $\mathcal{Y} = \{-1, +1\}$.

This model has been the subject of a plethora of works in the high-dimensional statistics literature over the past few years, in particular under the *Gaussian design* assumption:

Model 2.1 (Gaussian covariate model)**.** In the Gaussian covariate model (GCM), we assume the inputs are independently drawn from a Gaussian distribution:

$$\boldsymbol{x}^{\nu} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}}{\sqrt{d}}, \Sigma\right), \qquad \nu = 1, \cdots, n. \tag{5}$$

We denote $\mathcal{G}_{\boldsymbol{\theta}_0, \Sigma, \boldsymbol{\mu}}$ the teacher-student problem under this data assumption.

A key motivation for this work is the common intuition that data from standard classification tasks such as MNIST are closer to multi-modal distributions than to a single-mode Gaussian. Therefore, in this manuscript we ask ourselves the question: *"when is Gaussian data all you need?"*. In particular, we focus our attention to a prototypical distribution to model multi-modal data (and a universal approximator of densities): the K cluster Gaussian mixture:

Model 2.2 (Gaussian mixture model)**.** In the Gaussian mixture model (GMM), we assume the inputs are independently drawn from a mixture of K Gaussians:

$$\boldsymbol{x}^{\nu} \sim \sum_{c \in \mathcal{C}} p_c \, \mathcal{N}\left(\frac{\boldsymbol{\mu}_c}{\sqrt{d}}, \Sigma_c\right), \qquad \nu = 1, \cdots, n. \tag{6}$$

where $\mathcal{C} \coloneqq \{1, \cdots, K\}$ is the set of possible clusters, and $p_c \in [0, 1]$ is the probability of belonging to cluster $c \in \mathcal{C}$, whose means and covariance are given by $(\boldsymbol{\mu}_c, \Sigma_c)$.

We note that despite Model 2.1 being a special case $K = 1$ of Model 2.2 when the labels y^{ν} do not depend on the input cluster, it will be instructive to treat the $K = 1$ and $K > 1$ case as two different models.

3 Main theoretical results

In this section, we introduce our main theoretical results concerning universality of high-dimensional generalized linear estimation of GMMs. Our result builds on a long line of works providing an exact asymptotic characterization of empirical risk minimizers (2) on the proportional high-dimensional limit for Model 2.1 [2, 19, 23, 24]. In particular, closer to our derivation are the rigorous results in [19, 22]. We prove that the training and generalization error concentrate in high-dimensions in a deterministic expression given by the solution of a set of self-consistent equations. Then we

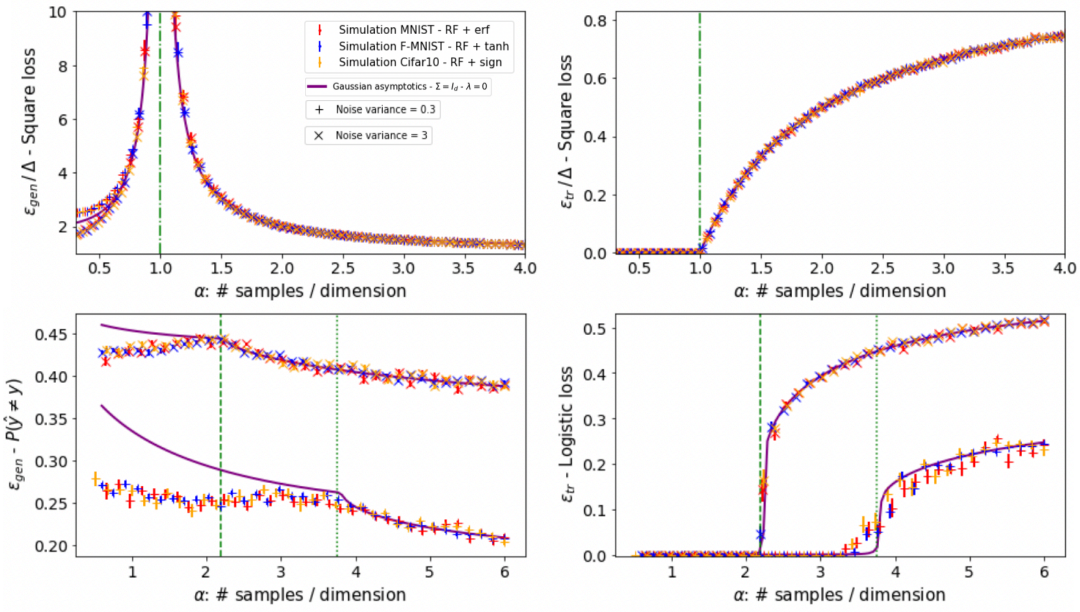

Figure 1: An illustration of Gaussian universality with vanishing regularization $\lambda = 0^+$ for a selection of datasets (MNIST, Fashion-MNIST, Cifar10) with a random teacher function, after a random feature map: Generalization (left) and training (right) errors as a function of the number of samples per dimension $\alpha = n/d$ for ridge regression (top pannels) and logistic classification (bottom pannels). The solid purple line is the exact asymptotics formula for Gaussian covariates with identity covariance, while the vertical green lines are the threshold values $\alpha^\star(\Delta)$ at which an unique minimizer of the loss starts to exist. Dots show numerical simulations for different real datasets with random features maps. All data follows the Gaussian predictions for the training loss, illustrating Theorems 3.4 and 3.6. In particular the separability, or interpolation, threshold is the one of the Gaussian model (see corollary 3.5). Each learning task was ran for two different value for the noise variance corrupting the labels - crosses are associated to $\Delta = 3$ while pluses to $\Delta = 0.3$. Error bars are built using standard deviation over 30 runs.

analyze Gaussian universality, provably characterizing a set of sufficient conditions the learning task must respect such that the test and training errors of Model 2.2 asymptotically agrees with the ones from Model 2.1.

Since we deal with sequences of random variables, we will need a rigorous definition of convergence. For two sequences of numbers $(a_n), (b_n)$, we write

$$a_n \simeq b_n \quad \text{iff} \quad \lim_{n\to\infty} a_n - b_n = 0. \tag{7}$$

Accordingly, for two sequences of random variables $(X_n), (Y_n)$, we define closeness in probability by

$$X_n \overset{P}{\simeq} Y_n \quad \text{iff} \quad X_n - Y_n \overset{P}{\to} 0, \tag{8}$$

where $\overset{P}{\to}$ denotes convergence in probability.

3.1 Exact asymptotics

Our first result is to give closed-form asymptotic characterization of the performance of the minimizer of 2 for the Gaussian Mixture model 2.2, generalizing the rigorous results of [22]:

Proposition 3.1. *(Exact asymptotics, informal statement) Consider the empirical risk minimization problem introduced in eq. (1) under Gaussian mixture data given by Model 2.2. For any pseudo-Lispchitz performance metric g, the training and generalization errors (2) converge in the high-dimensional limit of $n, d \to \infty$ with fixed ratio $\alpha = n/d$ to deterministic expressions which are entirely determined by the solution of a set of self-consistent replica saddle-point equations (33).*

$$\begin{aligned} \varepsilon_{tr}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \varepsilon_{tr}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \\ \varepsilon_{gen}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \varepsilon_{gen}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \end{aligned} \tag{9}$$

The key difference between Prop. 3.1 and Thm. 1 from [22] is the distribution of the labels. While the labels in [22] are given by the GMM cluster index, in Prop. 3.1 we consider labels generated by the target function (4). While we do not provide a formal proof of these results, note that the generic proof scheme from [22], that maps the solution to the study of a so-called approximate message passing algorithm [39], can be readily adapted to our setting. Indeed, the approximate message passing scheme in our scenario is the same, and the only difference in the proof is to include a teacher in the its asymptotic analysis, similarly as was done in [40].

Once we specified the GMM, and properly defined the ERM in eq. (1) we can evaluate the expressions in eqs. (2),(3) as a function of low-dimensional quantities which define the sufficient statistics of the asymptotic errors, and are also known as *order parameters*. We give the detail of the computation in Appendix A.

3.2 Uncorrelated teachers

An in-depth comparison of the asymptotic expressions for the errors of Models 2.1 & 2.2 reveals that their key difference lies on the way the leading target direction $\boldsymbol{\theta}_0$ correlate with the cluster means and covariances. A first step towards universality is therefore to characterize under which conditions the asymptotic errors are independent of the means. We make the following assumptions:

Assumption 1. The teacher $\boldsymbol{\theta}_0$ respects, $\forall (c, c') \in \mathcal{C} \times \mathcal{C}$:

$$\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0 \tag{10}$$

$$\lim_{n,d\to\infty} \frac{1}{d} \boldsymbol{\theta}_0^\top \Sigma_{c'} \left(\lambda + \sum_{c\in\mathcal{C}} \hat{V}_c^\star \Sigma_c \right)^{-1} \boldsymbol{\mu}_c \to 0 \tag{11}$$

where $\{\hat{V}_c^\star\}_{c=1}^K$ are the fixed points of the (replica) saddle point equations describing the centered GMM problem.

Assumption 2. The loss function, and the teacher distribution are both symmetric:

$$\ell(x, y) = \ell(-x, -y) \tag{12}$$

$$P_0\left(y|\tau\right) = P_0\left(-y|-\tau\right), \tag{13}$$

and the regularization is an ℓ^2 penalty $\lambda/2\|\cdot\|_2^2$.

Proposition 3.2. *(Mean Universality)*
Under Assumptions 1 and 2,the cluster means $\{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}$ are not relevant in high-dimensional ERM estimation:

$$\varepsilon_{gen}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \varepsilon_{gen}^{\text{GMM}}\left(\mathbf{0}, \{\Sigma_c\}_{c=1}^K\right) \tag{14}$$

$$\varepsilon_{tr}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \varepsilon_{tr}^{\text{GMM}}\left(\mathbf{0}, \{\Sigma_c\}_{c=1}^K\right) \tag{15}$$

Therefore, intuitively mean universality can be achieved when the label generation process is uncorrelated with the data structure, see Appendix B for a detailed discussion. The assumptions on the target and loss function are not restrictive, and are easily satisfied by odd target activation f_0 and margin-based losses of the form $l(y, z) = \ell(yz)$. Stronger universality can be shown by doing simplifying assumption on the data structure. Indeed, if the mixture is homogeneous (in the sense that the all the covariance are identical, a condition often called homoscedasticity in statistics) we have Gaussian universality:

Theorem 3.3. *(Gaussian Universality of homoscedastic GMMs) Under the assumption of 3.2, consider an homoscedastic GMM:*

$$\Sigma_c = \Sigma \qquad \forall c \in \mathcal{C}$$

Then for all α and Δ, the errors of the GMM are asymptotically equal to those of a GCM:

$$\varepsilon_{gen}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \Sigma\right) \simeq \varepsilon_{gen}^{\text{GCM}}(\boldsymbol{\theta}_0, \Sigma, \mathbf{0}) \tag{16}$$

$$\varepsilon_{tr}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \Sigma\right) \simeq \varepsilon_{tr}^{\text{GCM}}(\boldsymbol{\theta}_0, \Sigma, \mathbf{0}) \tag{17}$$

This results follow in a straightforward way from Theorem. 3.2. We have mapped under controllable assumptions a GMM problem to a simpler Gaussian one.

Additionally, if we consider vanishing regularization, we can prove that errors are independent from the shape of the covariance, and therefore we can take an isotropic mixture:

Theorem 3.4. *(Covariance universality for $\lambda = 0^+$) Under the assumption of Thm. 3.3, assume that it exists an unique minimizer of the empirical risk (1) with zero regularization. Then, the test & training errors a homoscedastic GMM (2.2) estimation problem asymptotically coincide with those of a centered, isotropic Gaussian model $\mathcal{G}_{(\boldsymbol{\theta}_0,\mathbf{I},\mathbf{0})}$ (2.1).*

The proof of this general result is given in Appendix B.

Since the training error for losses such as logistic or hinge characterize the separability transition at $\lambda = 0^+$, an interesting consequence of Thm. 3.4 is the following:

Corollary 3.5. *(Universality of linear separability for homoscedastic GMMs) The location of the separability / interpolation transition α_c above which the data stop to be linearly separable is the same for homoscedastic GMMs and the Gaussian model.*

This universality is particularly interesting in light of the detailed study of the separability transition for random teacher weights and *Gaussian data* in [10]. Similar universality phenomena was observed for the reconstruction transition in linear estimation in [41].

Surprisingly, if we further consider a square loss minimization, the estimation of *any* GMM (homoscedastic or not!) under mean universality condition can be mapped to those of a trivial Gaussian problem:

Theorem 3.6. *(Strong universality of the square loss for $\lambda = 0^+$) Consider underparametrized learning of GMMs with general means and covariance respecting the assumptions of Theorem. 3.2. Set in eq. (1): $l(y, \hat{y}) = (y - \hat{y})^2$ and consider a generic noisy linear teacher activation $P_0(\tau; \Delta) = \mathcal{N}(\tau, \Delta)$, with $\xi \sim \mathcal{N}(0, 1)$. Then, the training error for a GMM estimation problem is given by*

$$\varepsilon_{tr}^{\mathrm{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma_c\}_{c=1}^K\right) \simeq \frac{(\alpha - 1)\Delta}{\alpha} \tag{18}$$

Note that the strong universality statement does not hold for the test error (a counterexample is discussed in Appendix C using a strongly heteroscedastic case). However, as we see from Fig. 1, it seems surprisingly true that random teacher regression on real data follows the Gaussian asymptotic prediction in the underparametrized region. Moreover, we observe in the lower panel of Fig. 1 that even for non-quadratic losses we can draw a similar conclusion, and we investigate this further in next section.

Finally, we note that it the limit of infinite noise and binary labels, our results give back the ones observed for purely random Rachemacher labels proven in [18].

3.3 Correlated teachers

We now consider the general case where the target weights correlate with the structure in the data, relaxing the assumptions in Theorem. 3.2. Can we still say something? Our first result shows that the answer is yes. We focus on a simple controlled setting in which we can express the ERM performance in a closed form for *any* teacher vector:

Theorem 3.7. *(Exact asymptotics for isotropic covariance) Consider a ridge regression task with a 2-clusters GMM (2.2). Note that, without loss of generality we can take $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \boldsymbol{\mu}$. Assume isotropic covariances:*

$$\Sigma_+ = \Sigma_- = I_d, \tag{19}$$

and denote:

$$\rho = \lim_{d\to\infty} \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma \boldsymbol{\theta}_0, \qquad \gamma = \lim_{d\to\infty} \frac{1}{d}||\boldsymbol{\mu}||_2^2 \tag{20}$$

$$\pi = \lim_{d\to\infty} \frac{1}{d}\boldsymbol{\mu}^\top \boldsymbol{\theta}_0 \qquad P_0(\tau; \Delta) = \mathcal{N}(\tau, \Delta). \tag{21}$$

Defining:

$$A(\eta) = \frac{(\eta - 1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{(\Delta + (\eta - 1)^2\rho)\left(1 + \gamma\eta\right)^2} \quad \text{with} \tag{22}$$

$$\eta = 1 + \frac{1}{2}\left(\alpha - 1 + \lambda - \sqrt{4\lambda + (\alpha - 1 + \lambda)^2}\right)$$

The asymptotic errors admit a closed form expression in terms of η:

$$\varepsilon_{gen}^{\mathrm{GMM}} \simeq g(\alpha, \Delta, \rho, \eta)(1 - \pi^2 A(\eta)) \tag{23}$$

$$\varepsilon_{tr}^{\mathrm{GMM}} \simeq t(\alpha, \Delta, \rho, \eta)(1 - \pi^2 A(\eta)) \tag{24}$$

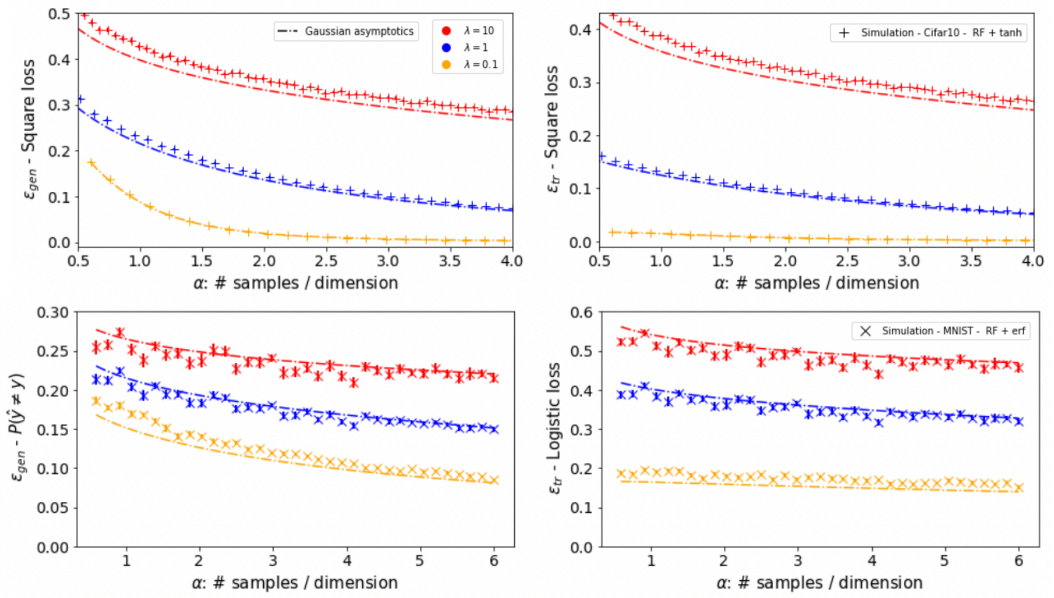

Figure 2: An illustration of Gaussian universality with finite regularization λ for a selection of datasets (MNIST, Fashion-MNIST, Cifar10) with a random teacher function, after a random feature map: Generalization (left) and training (right) errors as a function of the number of samples per dimension $\alpha = n/d$. In the upper panel we show ridge regression on Cifar10 preprocessed with RF and $\tanh$ activation, while the lower one is logistic regression on MNIST with RF and erf activation. The dashed curves are the exact asymtotics prediction of the Gaussian theory by matching covariance. The good agreement illustrates the property of theorem 3.3. The different colours represents different regularization strength $\lambda \in \{0.1, 1, 10\}$ respectively in yellow, blue, and red. Error bars are built over 30 runs.

where g and t are the asymptotic limit of the generalisation $\varepsilon_{gen}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, I, \mathbf{0})$ and training $\varepsilon_{tr}^{\mathrm{GCM}}(\boldsymbol{\theta}_0, I, \mathbf{0})$ error for a single Gaussian model with unit covariance and uncorrelated teacher:

$$g(\alpha, \Delta, \rho, \eta) = \frac{\alpha\left(\Delta + (\eta-1)^2\rho\right)}{(\alpha - \eta^2)} \tag{25}$$

$$t(\alpha, \Delta, \rho, \eta) = \frac{(\alpha-\eta)^2\left(\Delta + (\eta-1)^2\rho\right)}{\alpha\,(\alpha - \eta^2)} \tag{26}$$

The full closed form expressions and the extension to covariances of the form $\Sigma = \sigma I_d$ are derived in Appendix C. Note that the correction factor to the Gaussian performance scales with π^2, a measure of correlation between teacher vector and data structure. Hence, we would be tempted to state that targets that correlate with the data structure, i.e. $\pi \neq 0$, always break Gaussian universality. Although this is usually the case, in the limit of vanishing regularization we are in the position to present an interesting corollary of Theorem. 3.7:

Corollary 3.8. *(Restoration of universality for $\lambda = 0^+$) Consider the same setting of Theorem. 3.7, further consider underparametrized learning in the limit of vanishing regularization. Then, the test and training errors for a GMM estimation are equal to a Gaussian one for any teacher vector (that is eqs (23) and (24) at $\eta = 0$):*

$$\varepsilon_{gen}^{\mathrm{GMM}} = \Delta\frac{\alpha}{\alpha - 1} \tag{27}$$

$$\varepsilon_{tr}^{\mathrm{GMM}} = \Delta\frac{(\alpha-1)}{\alpha} \tag{28}$$

We thus restore Gaussian universality for correlated teachers in the underparametrized regime, in a similar fashion to what Theorem. 3.4 is stating for general convex losses and covariances under the mean universality condition. The universality property for correlated teachers is valid also for more general homoscedastic mixtures with identity covariance. For the sake of brevity we refer to Appendix C where we discuss in detail this interesting extension.

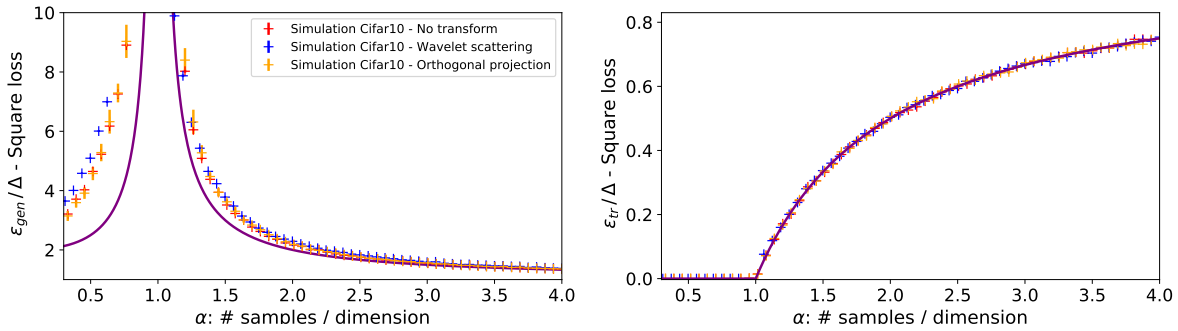

Figure 3: An illustration of Gaussian universality for Training, and lack thereof for Generalization at vanishing regularization. Instead of a random feature maps as in Fig. 1, we pre-processed grayscale Cifar10 with wavelet scattering (blue dots) or orthogonal hadamard projections (yellow dots), or used directly the raw data (red dots). Using again a random teacher function, the training error is found to agree perfectly with the universal Gaussian asymptotics predictions with identity covariance and $\lambda = 0^+$, as for theorem 3.6, but the generalization is found to show clear deviation. This highlight the roles of heteroscedasticity in the data. Error bars are built over 30 runs.

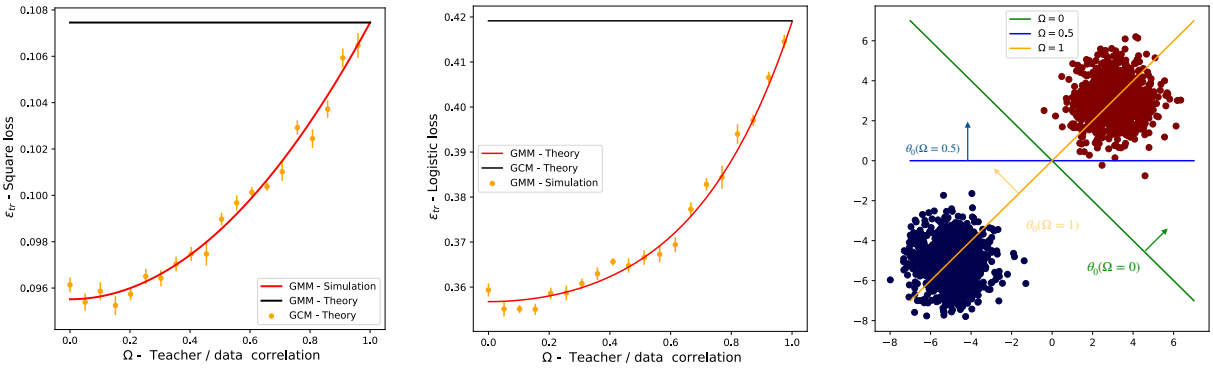

Figure 4: On the importance of the correlation of the task and the structure of the data: **Left & Center** : Training errors achieved in the ERM for estimation of 2-cluster GMM with $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_-$ and $\Sigma_+ = \Sigma_- = I_d$, plotted versus a correlation parameter Ω: we build a series of learning tasks with teacher weights dependent on Ω, namely we take $\boldsymbol{\theta}_0(\Omega) = \Omega\boldsymbol{\mu}_\perp + \sqrt{1-\Omega^2}\boldsymbol{\mu}_+$, with $\boldsymbol{\mu}_+^\top\boldsymbol{\mu}_\perp = 0$. We fix the dimension to be $d = 500$, and $(\alpha, \lambda) = (1.2, 0.7)$. The solid black line is the theoretical prediction coming from Gaussian asymptotics. The solid red line is the theoretical prediction for the GMM performance, while the orange dots are the numerical simulations which agree as expected with the theoretical prediction. In the left panel we have real labels and perform ridge regression, while on the central one we consider binary labels and perform logistic regression. The error bars are built using standard deviation over 30 runs. **Right**: Geometrical intuition plotted for $d = 2$. Three hyperplanes (lines in 2D) are displayed correspondent to $\Omega = \{0, 0.5, 1\}$, respectively in green, blue, and orange. As Ω increase the labels become more uncorrelated with the data structure.

4 Illustration on synthetic and real datasets

In this section we investigate the consequences of our main theoretical results. First, we consider the case of real data, illustrating the applicability of our universality theorems in cases in which the target is not correlated with the data structure. Based on these observations we move on studying simple synthetic settings described in Theorem 3.7 and for which we can derive analytical results and systematically probe Gaussian universality. All the code used in our experiments are available in a GitHub repository.

4.1 Random teacher universality

We analyze Gaussian universality of real data under the random teacher function assumptions of Prop. 3.2. We consider three standard datasets: MNIST, Fashion-MNIST, and grayscale CIFAR-10, as well as synthetic data.

Universality with random feature maps — To go beyond simple linear fit, we pre-process the data with a random feature map [42] as follows: take an image $\boldsymbol{\omega}_\nu \in \mathbb{R}^{d'}$ and map it to $\boldsymbol{x}_\nu = \sigma(F\boldsymbol{\omega}_\nu)$, where the elements $F \in \mathbb{R}^{d\times d'}$ are i.i.d drawn from $\mathcal{N}(0,1)$ and $\sigma(\cdot)$ is an activation function. As shown in [42], for $d \to \infty$ this converge to a kernel method. For each dataset we take a different non-linearity (erf,$\tanh$, and sign as described in Fig. 1). We then build new labels by plugging in eq. (4) $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, I_d)$, and we set for regression tasks $f_0(\tau) = \tau + \Delta$, while for classification tasks $f_0(\tau) = \mathrm{sign}(\tau + \Delta)$.

First, we analyze the vanishing regularization case in Fig. 1: random teacher regression on pre-processed real datasets respects the strong universality of the training loss as stated in Theorem. 3.6. Interestingly, we also observe a perfect match between the Gaussian asymptotics prediction and the simulations in Fig. 1 for the generalization error. This suggests that after the random features maps the data is sufficiently close to a homoscedastic mixture. Further, the lower panel of Fig. 1 shows that the strong universality property seems to hold beyond square loss minimization.

We analyze as well the finite regularization setting in Fig. 2: we compare the simulations on real data with the prediction of the exact Gaussian asymptotics: we match the covariance of each pre-processed real dataset and compute the performance of the ERM estimator thanks to the deterministic replica formula. 3.1. The predictions of the Gaussian theory matches the numerical simulations. As discussed for Fig. 1, it seems that the homoscedasticity assumption in Theorem. 3.3 can be sometimes relaxed.

Universality of the double descent phenomena — One finding in modern machine learning that goes against the classical statistical theory wisdom is the double descent behaviour of learning curves [13, 43–45], see upper panel of Fig. 1: the test error does not deteriorate as the number of parameters is increased with a characteristic divergence at $\alpha = 1$, known as the interpolation peak. As shown in Fig. 1 and Fig. 2, the generalization error and its characteristic "double descent" behavior for $\alpha > 1$ is universal for homoscedastic data, while the training error appears universal even for heteroscedastic data (see Fig. 3).

Universality of linear separability — Recently, [10] investigated the linear separability of Gaussian data with a random teacher and noise Δ, generalizing the classical results by [46] to a single-index target. [10] has shown the existence of a critical phase transition $\alpha_c(\Delta)$ that goes continuously from $\alpha_c(\infty) = 2$ (for infinite noise) to $\alpha_c > 2$ for finite Δ (with $\alpha_c \to \infty$ as $\Delta \to 0$). As discussed in corollary 3.5, this transition is universal for homoscedastic mixtures. The interest of this transition thus extends way beyond Gaussian data and in fact agrees with the real data experiment in Figs. 1 and 2. As for the double descent phenomena, we believe it is a very interesting consequence of our theorems that the theoretical works mentioned above are valid way beyond the simple Gaussian assumption.

Non-universal behavior for generalization — Finally, we implemented different transforms beyond random features for the pre-processing step: we considered the wavelet scattering transform [47], orthogonal random projections [48], and even no transform at all. Without the shuffling of the random projection, the fact that the data are complex and probably rather heteroscedastic than homoscedastic, we expected a weaker form of universality. We present the results of ridge regression with labels generated by a random teacher vector in Fig. 3. The strong universality statement in Theorem. 3.6, that did not required any assumption on the data, remains valid and we observed a perfect collapse of the training data. However, as expected, the generalization error shows clear deviations with respect to the Gaussian behavior. Presumably these transforms do not homogenize the data enough such that Gaussian universality for the test error to hold.

4.2 Correlated teacher

Previously, we showed that ridge regression with a random teacher model, even when the data is structured, can lead to universal behaviour. We now consider the dual task: take a simple homogeneous model for the data and study correlated target weights. Indeed, some recent works have studied examples of the lack of universality between Gaussian mixtures and Gaussian models [49, 50]; we want to follow this direction and use the setting described in Theorem. 3.7: we consider a 2-cluster GMM with opposite means of norm $\boldsymbol{\mu}$ and same covariance Σ. We build a series of learning tasks at fixed (α, λ) varying the overlap between the teacher vector and the cluster means as follows:

$$\boldsymbol{\theta}_0(\Omega) = \Omega\boldsymbol{\mu}_\perp + \sqrt{1-\Omega^2}\boldsymbol{\mu} \tag{29}$$

$$\text{such that} \qquad \boldsymbol{\mu}^\top\boldsymbol{\mu}_\perp = 0 \tag{30}$$

The results are presented in Fig. 4 for both ridge and logistic regression. They clearly show that a small correlation between the target weights and the mixture means breaks Gaussian Universality. A neat geometrical intuition of the result is given in Fig. 4: as we decrease Ω, we generate labels which are more and more correlated with the data distribution,

and consequently there is a correction factor to the Gaussian prediction as Theorem. 3.7 predicts. We refer to Appendix. C for a more detailed discussion on correlated teachers. We conclude that universality is broken in tasks where the labels are correlated with the structure. Note, however, that as proven in Corollary. 3.8: the discrepancy between the GMM prediction and the Gaussian one goes to zero *for any* correlation measure Ω as $\lambda \to 0^+$, where the universality is restored as suggested by Thm. 3.6.

5 Acknowledgments

We thank Yatin Dandi, Federica Gerace, Sebastian Goldt, Denny Wu & Lenka Zdeborová for useful discussions. We acknowledge funding from the Swiss National Science Foundation grant SNFS OperaGOST, 200021_200390 and the *Choose France - CNRS AI Rising Talents* program.

References

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

[2] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, April 2022.

[3] Denny Wu and Ji Xu. On the optimal weighted \ell_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc., 2020.

[4] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

[5] 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, 2021.

[6] Ohad Shamir. The implicit bias of benign overfitting. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 448–478. PMLR, 02–05 Jul 2022.

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

[8] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[9] 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.

[10] 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.

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

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

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

[14] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR, 13–18 Jul 2020.

[15] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *MSML*, 2021.

[16] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, pages 1–1, 2022.

[17] Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

[18] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of linear classifiers with random labels in high-dimension, 2022.

[19] 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. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151. Curran Associates, Inc., 2021.

[20] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15568–15578. Curran Associates, Inc., 2020.

[21] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. 2020.

[22] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalised linear models: Precise asymptotics in high-dimensions. 2021.

[23] Edgar Dobriban and Stefan Wager. High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[24] 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, August 2018.

[25] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing, 2020.

[26] 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, May 2019. ISSN: 2379-190X.

[27] Francesca Mignacco, Florent Krzakala, Yue M. Lu, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. 2020.

[28] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, June 2022.

[29] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: a high-dimensional asymptotic view. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 8907–8920, Red Hook, NY, USA, December 2020. Curran Associates Inc.

[30] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020.

[31] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning, 2023.

[32] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, pages 2757–2790, 2008.

[33] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[34] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10131–10143. Curran Associates, Inc., 2021.

[35] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error rates for kernel classification under source and capacity conditions, 2022.

[36] 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.

[37] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

[38] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.

[39] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[40] Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning curves for the multi-class teacher-student perceptron. *Machine Learning: Science and Technology*, 2022.

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

[42] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[43] Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

[44] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, jul 2019.

[45] S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, oct 2019.

[46] 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.

[47] Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, Joan Bruna, Vincent Lostanlen, Muawiz Chaudhary, Matthew J. Hirn, Edouard Oyallon, Sixin Zhang, Carmine Cella, and Michael Eickenberg. Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, 21(60):1–6, 2020.

[48] Krzysztof M Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[49] Umberto M. Tomasini, Antonio Sclocchi, and Matthieu Wyart. Failure and success of the spectral bias prediction for Laplace Kernel Ridge Regression: the case of low-dimensional data. pages 21548–21583. PMLR, June 2022.

[50] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Science*, 119(40):e2201854119, October 2022.

[51] Marc Mezard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Publishing Company, November 1987. Google-Books-ID: DwY8DQAAQBAJ.

[52] Michel Talagrand. *What Is a Quantum Field Theory?* Cambridge University Press, Cambridge, 2022.

[53] 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.

A Replica computation

A.1 Formal statement of the theorem

We first provide the full statement of Proposition 3.1. Consider a minimization problem of the form

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{\boldsymbol{d}}}{\arg\min} \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y^\nu, \boldsymbol{\theta}^\top \boldsymbol{x}^\nu\right) + r(\boldsymbol{\theta}), \tag{31}$$

where the data $(\boldsymbol{x}^\nu, y^\nu)$ is generated according to the following Gaussian mixture model:

$$\boldsymbol{x}^\nu \sim \sum_{c\in\mathcal{C}} p_c \,\mathcal{N}\left(\frac{\boldsymbol{\mu}_c}{\sqrt{d}}, \Sigma_c\right) \quad \text{and} \quad y^\nu \sim P_0(\cdot \,|\, \boldsymbol{\theta}_0^\top \boldsymbol{x}^\nu) \tag{32}$$

and assume the following:

1. The functions $\ell(y,\cdot)$ and r are continuous and coercive, and the function $\ell(y,\cdot)+r(\cdot)$ is strongly convex,
2. The covariance matrices Σ_c are positive definite, and their spectral norms are uniformly bounded,
3. The means $\boldsymbol{\mu_c}/\sqrt{d}$ and the teacher vector $\boldsymbol{\theta}_0$ are uniformly bounded,
4. the number of clusters $|\mathcal{C}|$ is finite,
5. the distribution P_0 is sub-gaussian with uniformly bounded norm.

Then, as $n, d \to \infty$ with $n/d \to \alpha > 0$, we have

$$\begin{aligned}
\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \sum_{c\in\mathcal{C}} p_c \,\mathbb{E}_{\omega_c^{(s)},\omega_c^{(t)},y}\left[\ell\left(y, \text{prox}_{V_c^\star \ell(y,\cdot)}(\omega_c^{(s)})\right)\right] &&=: \varepsilon_{\text{tr}}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}}) \\
\varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) &\overset{P}{\simeq} \sum_{c\in\mathcal{C}} p_c \,\mathbb{E}_{\omega_c^{(s)},\omega_c^{(t)},y}\left[\ell\left(y, \omega_c^{(s)}\right)\right] &&=: \varepsilon_{\text{tr}}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathcal{C}}, \{\Sigma_c\}_{c\in\mathcal{C}})
\end{aligned} \tag{33}$$

where

$$\begin{pmatrix}\omega_c^{(s)} \\ \omega_c^{(t)}\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\pi_c \\ h_c^\star\end{bmatrix}, \begin{bmatrix}\rho & m_c^\star \\ m_c^\star & q_c^\star\end{bmatrix}\right) \quad \text{and} \quad y \sim P_0\left(\cdot \,|\, \omega_c^{(t)}\right). \tag{34}$$

and the prox function is

$$\text{prox}_f(x) = \min_{z\in\mathbb{R}}\left[\frac{1}{2}(z-x)^2 + f(z)\right] \tag{35}$$

The overlaps used in the equation are defined as follows:

$$\rho_c = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}_0, \quad \pi_c = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c, \tag{36}$$

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

$$\begin{cases}
\hat{V}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) \partial_\omega f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\
\hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right] \\
\hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \partial_\omega \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\
\hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]
\end{cases} \tag{37}$$

$$\begin{cases}
V_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, I_d)}\left[\hat{\boldsymbol{\eta}}^\top \hat{q}_c^{-1/2} \Sigma_c^{1/2} \boldsymbol{\xi}_c\right] \\
q_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, I_d)}\left[\hat{\boldsymbol{\eta}}^\top \Sigma_c \hat{\boldsymbol{\eta}}\right] \\
m_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, I_d)}\left[\boldsymbol{\theta}_0^\top \Sigma_c \hat{\boldsymbol{\eta}}\right] \\
h_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, I_d)}\left[\boldsymbol{\mu}_c^\top \hat{\boldsymbol{\eta}}\right]
\end{cases}$$

and we have defined the following auxiliary functions:

$$\mathcal{Z}_0(y,\omega,V) = \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi V}} P_0(y|\lambda)\, e^{-\frac{(\lambda-\omega)^2}{2V}} \tag{38}$$

$$f_\ell(y,\omega,V) = -V^{-1}\partial_\omega \mathcal{M}_{V\ell(y,\cdot)}(\omega) = V^{-1}\left(\mathrm{prox}_{V\ell(y,\cdot)}(\omega) - \omega\right). \tag{39}$$

Finally, the auxiliary variable $\hat{\boldsymbol{\eta}}$ is a function of $\{\boldsymbol{\xi}_c, \hat{V}_c, \hat{m}_c, \hat{q}_c, \hat{h}_c\}$:

$$\hat{\boldsymbol{\eta}} = \mathrm{prox}_{r(\hat{\Sigma}^{-1/2}\cdot)}\left(\hat{\Sigma}^{-1/2}\left(\sum_{c\in\mathcal{C}} \hat{\boldsymbol{\mu}}_c + \sqrt{\hat{q}_c}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right)\right) \tag{40}$$

with

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{c}} = \hat{h}_c \boldsymbol{\mu}_{\boldsymbol{c}} + \hat{m}_c \Sigma_c \boldsymbol{\theta}_0 \quad \text{and} \quad \hat{\Sigma} = \sum_{c\in\mathcal{C}} \hat{V}_c \Sigma_c \tag{41}$$

We will not prove this theorem, since the proof is virtually equivalent to the one in [22, 40]. Instead, the next sections are dedicated to the derivation of these equations, using the so-called *replica method* from statistical physics [51]. Simpler particular cases of these equations can be found in Appendix A.9.

A.2 Gibbs measure and free energy

The starting point for our replica computation is to define the Gibbs measure over the weights $W \in \mathbb{R}^{K\times d}$:

$$\mu_\beta(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\nu=1}^n \ell\left(y^\nu, \boldsymbol{\theta}^\top \boldsymbol{x}^\nu\right) + \lambda r(\boldsymbol{\theta})\right]} = \frac{1}{\mathcal{Z}_\beta} e^{-\beta r(\boldsymbol{\theta})} \prod_{\nu=1}^n e^{-\beta\ell(y^\nu, \boldsymbol{\theta}^\top \boldsymbol{x}^\nu)} \tag{42}$$

where $\beta > 0$ is a parameter we eventually want to send to $\beta \to 0^+$, and $\mathcal{Z}_\beta \in \mathbb{R}$ is the partition function (the normalisation of the Gibbs measure). For convenience, we will define the following useful notation:

$$P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \equiv e^{-\beta r(\boldsymbol{\theta})}, \qquad P_\ell(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu) \equiv e^{-\beta\ell(y^\nu, \boldsymbol{\theta}^\top\boldsymbol{x}^\nu)} \tag{43}$$

which allow us to write the Gibbs measure as:

$$\mu_\beta(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_\beta} P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{\nu=1}^n P_\ell(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu) \tag{44}$$

When $\beta \to \infty$, the Gibbs measure μ_β will concentrate around the $\boldsymbol{\theta}$ that minimize the empirical risk (1). As a result, the free energy density defined as

$$-\beta f_\beta = \lim_{d\to\infty} \frac{1}{d} \mathbb{E} \log \mathcal{Z}_\beta \tag{45}$$

where the limit is taken with $n/d = \alpha$ fixed, and the expectation is over the distribution of the data, will concentrate around the minimum risk. In order to take the expectation explicitly, we use the replica trick:

$$\log \mathcal{Z}_\beta = \lim_{s\to 0^+} \frac{1}{s} \partial_s \mathcal{Z}_\beta^s|_{s=0} \tag{46}$$

Therefore, the replica computation boils down to the computation of the averaged replicated partition function:

$$\mathbb{E}\mathcal{Z}_\beta^s = \mathbb{E}\left[\int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}\; P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{\nu=1}^n P_\ell\left(y^\nu|\boldsymbol{\theta}^\top\boldsymbol{x}^\nu\right)\right]^s \tag{47}$$

$$= \prod_{\nu=1}^n \mathbb{E}_{(\boldsymbol{x}^\nu, y^\nu)} \int \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a\; P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a) P_\ell\left(y^\nu|\boldsymbol{\theta}^{a\top}\boldsymbol{x}^\nu\right) \tag{48}$$

A.3 Taking the average over the data

The main difference in this computation with respect to the usual binary Gaussian mixture is that the labels are generated by a teacher target function. In other words, the joint distribution of the data is given by:

$$P_{\boldsymbol{\theta}_0}(\boldsymbol{x}, y) = P_0(y|\tau) \sum_{c\in\mathcal{C}} \rho_c \, \mathcal{N}\left(\boldsymbol{x}; \frac{\boldsymbol{\mu}_c}{\sqrt{d}}, I_d\right) \tag{49}$$

where $P_0(y|\tau)$ is the probability induced by the teacher activation $f_0(\tau)$. The expression in (48) can then be written as:

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a) \left(\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{x} P_0(\boldsymbol{y}|\boldsymbol{\theta}_0^\top \boldsymbol{x}) P_{\boldsymbol{x}}(\boldsymbol{x}) \prod_{a=1}^{s} P_\ell\left(y^\nu | \boldsymbol{\theta}^{a\top} \boldsymbol{x}^\nu\right) \right)^n \tag{50}$$

We make a change of variables by introducing the *local fields* $\tau = \boldsymbol{\theta}_0^\top \boldsymbol{x}$ and $\lambda^a = \boldsymbol{\theta}^{a\top} \boldsymbol{x}$. The distribution of the local fields will be a Gaussian mixture itself, and we can compute the moments for each mode:

$$\begin{aligned}
\pi_c &\equiv \mathbb{E}_{\boldsymbol{x}^\nu}\left[\tau_c\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c \\
\rho_c &\equiv \mathrm{Var}_{\boldsymbol{x}^\nu}\left[\tau_c\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}_0 \\
h_c^a &\equiv \mathbb{E}_{\boldsymbol{x}^\nu}\left[\lambda_c^a\right] = \frac{1}{d}\boldsymbol{\theta}^{a\top} \boldsymbol{\mu}_c \\
q_c^{ab} &\equiv \mathrm{Cov}_{\boldsymbol{x}^\nu}\left[\lambda_c^a, \lambda_c^b\right] = \frac{1}{d}\boldsymbol{\theta}^{a\top} \Sigma_c \boldsymbol{\theta}^b \\
m_c^a &\equiv \mathrm{Cov}_{\boldsymbol{x}^\nu}\left[\tau_c, \lambda_c^a\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a
\end{aligned} \tag{51}$$

Equivalently, the local fields distribution is the following low-dimensional Gaussian mixture distribution:

$$P(\tau, \boldsymbol{\lambda}) = \sum_{c\in\mathcal{C}} p_c \, \mathcal{N}\left(\begin{bmatrix} \pi_c \\ \boldsymbol{h}_c \end{bmatrix}, \begin{bmatrix} \rho_c & \boldsymbol{m}_c^\top \\ \boldsymbol{m}_c & Q_c \end{bmatrix} \right), \tag{52}$$

where we have defined the vectors $\boldsymbol{m}, \boldsymbol{h} \in \mathbb{R}^s$ with entries m^a, h^a and the matrix $Q \in \mathbb{R}^{s\times s}$ with entries q^{ab}. Therefore, we can factorize the partition function by only integrating over the local field distribution:

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}^a) \left[\int \mathrm{d}y \int \mathrm{d}\tau \, P_0(y|\tau) \int \left(\prod_{a=1}^{s} \mathrm{d}\lambda^a P_\ell(y|\lambda^a) \right) P(\tau, \boldsymbol{\lambda} | \boldsymbol{m}, \boldsymbol{h}, q) \right]^n \tag{53}$$

A.4 Writing as a saddle-point problem

Note that (π_c, ρ_c) are fixed inputs in the problem. By Fourier transform arguments, we can write

$$\delta\left(m_c^a - \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right) = \int_{i\mathbb{R}} \frac{\mathrm{d}\hat{m}_c^a}{2\pi} \, e^{\hat{m}_c^a \left(d m_c^a - \boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right)}, \tag{54}$$

where the integral is on the imaginary line. By doing the same arguments for the q_c^{ab} and h_c^a, it ensues that

$$\mathbb{E}\mathcal{Z}_\beta^s = \int \prod_{c\in\mathcal{C}} \prod_{a=1}^{s} \frac{\mathrm{d}m_c^a \mathrm{d}\hat{m}_c^a}{2\pi} \frac{\mathrm{d}h_c^a \mathrm{d}\hat{h}_c^a}{2\pi} \prod_{1\le a\le b\le s} \frac{\mathrm{d}q_c^{ab} \mathrm{d}\hat{q}_c^{ab}}{2\pi} e^{d\Phi_{\mathrm{rs}}^{(s)}(\boldsymbol{m}, \hat{\boldsymbol{m}}, q, \hat{q})} \tag{55}$$

where we have defined the free energy potential:

$$\begin{aligned}
\Phi_{\mathrm{rs}}^{(s)}(\boldsymbol{m}, \hat{\boldsymbol{m}}, q, \hat{q}) &= \sum_{c\in\mathcal{C}} \left[\sum_{a=1}^{s} \hat{m}_c^a m_c^a + \sum_{a=1}^{s} \hat{h}_c^a h_c^a + \sum_{1\le a\le b\le s} \hat{q}_c^{ab} q_c^{ab} \right] - \alpha \Psi_g^{(s)}(\boldsymbol{m}, \boldsymbol{h}, q) - \Psi_\theta^{(s)}(\hat{m}, \hat{h}, \hat{q}) \\
\Psi_\ell^{(s)}(\boldsymbol{m}, \boldsymbol{h}, q) &\equiv \log \int \mathrm{d}y \int \mathrm{d}\tau \, P_0(y|\tau) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a P_\ell(y|\lambda^a) P(\tau, \boldsymbol{\lambda} | \boldsymbol{m}, \boldsymbol{h}, q) \\
\Psi_\theta^{(s)}(\hat{\boldsymbol{m}}, \hat{\boldsymbol{h}}, \hat{q}) &\equiv \frac{1}{d} \log \int \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \, P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{c\in\mathcal{C}} e^{\sum_{a=1}^{s} \left[\hat{h}_c^a \boldsymbol{\theta}^{a\top} \boldsymbol{\mu}_c + \hat{m}^a \boldsymbol{\theta}_0^\top \Sigma_c \boldsymbol{\theta}^a\right] + \sum_{1\le a\le b\le s} \hat{q}_c^{ab} \boldsymbol{\theta}^{a\top} \Sigma_c \boldsymbol{\theta}^b}
\end{aligned} \tag{56}$$

Therefore, as we take the $d \to \infty$ limit, we can apply the saddle-point method [51] to compute f_β

$$-\beta f_\beta = \underset{\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{h},\hat{\boldsymbol{h}},q,\hat{q}}{\operatorname{extr}} \lim_{s\to 0^+} \frac{\Phi^{(s)}(\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{h},\hat{\boldsymbol{h}},q,\hat{q})}{s} \tag{57}$$

Remark A.1. We have introduced the parameters $\hat{q}, \hat{m}, \hat{h}$ as pure imaginary numbers, but the optimization problem considers them as real numbers. This stems from the fact that the function Φ is holomorphic, so the integral is independent from the contour of integration. More details can be found in [52].

A.5 Replica symmetric ansatz

In order to make progress with the $s \to 0^+$ limit, we make the following replica symmetric ansatz:

$$\begin{aligned}
m_c^a &= m_c & \hat{m}_c^a &= \hat{m}_c, & a &= 1, \cdots, s \\
h_c^a &= h_c & \hat{h}_c^a &= \hat{h}_c, & a &= 1, \cdots, s \\
q_c^{aa} &= r_c, & \hat{q}_c^{aa} &= -\frac{1}{2}\hat{r}_c, & a &= 1, \cdots, s \\
q_c^{ab} &= q_c, & \hat{q}_c^{aa} &= \hat{q}_c, & 1 &\le a < b \le s
\end{aligned} \tag{58}$$

for all $c \in \mathcal{C}$. Since ℓ and r are convex, the function $\Phi^{(s)}$ has a unique saddle point, which must therefore coincide with the replica symmetric ansatz. We also define

$$V_c = r_c - q_c, \qquad \text{and} \qquad \hat{V}_c = \hat{r}_c - \hat{q}_c \tag{59}$$

By inserting this ansatz above, we can take the $s \to 0^+$ limit. This is done through a classical but computationally heavy method known as the *Hubbard-Stratonovich transform*; details can be found in [14], Appendix C. We simply reproduce the final results here:

$$-\beta f_\beta = \underset{\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}}{\operatorname{extr}} \Phi_{\text{rs}}(\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}) \tag{60}$$

The replicated free energy Φ_{rs} is given by the following formula:

$$\Phi_{\text{rs}}(\{m_c,\hat{m}_c,h_c,\hat{h}_c,q_c,\hat{q}_c,V_c,\hat{V}_c\}) = \sum_{c\in\mathcal{C}} \left[\frac{1}{2}\left(\hat{V}_c q_c - \hat{q}_c V_c\right) - \frac{1}{2}\hat{V}_c V_c + \hat{m}_c m_c + \hat{h}_c h_c\right] \tag{61}$$

$$- \alpha \Psi_\ell(\{m_c, h_c, q_c, V_c\}) - \Psi_{\boldsymbol{\theta}}(\{\hat{m}_c, \hat{h}_c, \hat{q}_c, \hat{V}_c\}) \tag{62}$$

where we have decomposed the contributions coming from the loss (Ψ_ℓ) and the regularization ($\Psi_{\boldsymbol{\theta}}$):

$$\Psi_\ell = \sum_{c\in\mathcal{C}} \mathbb{E}_{\xi_c \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)} \int \mathrm{d}y \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi + \pi_c, \rho - \frac{m_c^2}{q_c}\right) \log \mathcal{Z}_\ell(y, h_c + \sqrt{q_c}\,\xi_c, V_c) \tag{63}$$

$$\Psi_{\boldsymbol{\theta}} = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{\xi}_c \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, I_d)} \log \mathcal{Z}_{\boldsymbol{\theta}}\left(\sum_{c\in\mathcal{C}} \hat{V}_c \Sigma_c, \sum_{c\in\mathcal{C}} \hat{h}_c \boldsymbol{\mu}_c + \hat{m}_c \Sigma_c \boldsymbol{\theta}_0 + \sqrt{\hat{q}_c}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right), \tag{64}$$

and defined the following auxiliary free energies:

$$\mathcal{Z}_{\ell/0}(y,\omega,V) = \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi V}} P_{\ell/0}(y|\lambda)\, e^{-\frac{(\lambda-\omega)^2}{2V}} \tag{65}$$

$$\mathcal{Z}_{\boldsymbol{\theta}}(A,\boldsymbol{b}) = \int \mathrm{d}\boldsymbol{\theta} e^{-\beta r(\boldsymbol{\theta})} e^{-\frac{1}{2}\boldsymbol{\theta}^\top A \boldsymbol{\theta} + \boldsymbol{b}^\top \boldsymbol{\theta}} \tag{66}$$

A.6 Taking the zero temperature limit

In order to take the $\beta \to \infty$ limit, we make the following rescalings:

$$V_c \to \beta^{-1} V_c \qquad \hat{V}_c \to \beta \hat{V}_c \qquad \hat{q}_c \to \beta^2 \hat{q}_c \qquad \hat{m}_c \to \beta \hat{m}_c. \tag{67}$$

It is easy to check how this rescaling affects $\beta^{-1}\Phi^{(\mathrm{rs})}$, so we only need to consider Ψ_ℓ and $\Psi_{\boldsymbol{\theta}}$.

We start with the latter: letting

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{2}\,\boldsymbol{\theta}^\top A\boldsymbol{\theta} - \boldsymbol{b}^\top\boldsymbol{\theta} + r(\boldsymbol{\theta})$$

by the Laplace method for any $A, \boldsymbol{b}$,

$$\lim_{\beta\to\infty}\frac{1}{\beta}\log Z_{\boldsymbol{\theta}}(\beta A, \beta\boldsymbol{b}) = -\inf_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\eta}) \tag{68}$$

where

$$\boldsymbol{\eta} = \mathrm{prox}_{r(A^{1/2}\cdot)}(A^{1/2}\boldsymbol{b}). \tag{69}$$

As a result, every integral involved in $\Psi_{\boldsymbol{\theta}}$ (and, later, its partial derivatives) concentrates around its value at $\hat{\boldsymbol{\eta}}$ defined in (40). For the term Ψ_ℓ, the term $\mathcal{Z}_0$ is left unchanged, and by the same reasoning as above

$$\lim_{\beta\to\infty}\frac{1}{\beta}\log\mathcal{Z}_\ell(y,\omega,\beta V) = -V^{-1}\mathcal{M}_{V\ell(y,\cdot)}(\omega). \tag{70}$$

A.7 Saddle-point equations

The saddle-point equations are obtained by taking the derivatives of the free energy potential with respect to the overlap parameters. We obtain a set of self-consistent equations which we should solve in order to find a fixed point:

$$\begin{cases}
\hat{V}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right)\partial_\omega f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right]\\
\hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \partial_\omega\mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]\\
\hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right]
\end{cases} \tag{71}$$

$$\begin{cases}
V_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\boldsymbol{0},I_d)}\left[\hat{\boldsymbol{\eta}}^\top\hat{q}_c^{-1/2}\Sigma_c^{1/2}\boldsymbol{\xi}_c\right]\\
q_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\boldsymbol{0},I_d)}\left[\hat{\boldsymbol{\eta}}^\top\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
m_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\boldsymbol{0},I_d)}\left[\boldsymbol{\theta}_0^\top\Sigma_c\hat{\boldsymbol{\eta}}\right]\\
h_c &= \mathbb{E}_{\{\boldsymbol{\xi}_c\}\overset{\text{i.i.d}}{\sim}\mathcal{N}(\boldsymbol{0},I_d)}\left[\boldsymbol{\mu}_c^\top\hat{\boldsymbol{\eta}}\right]
\end{cases} \tag{72}$$

where all the relevant quantities have been defined in Appendix A.1.

A.8 Training and generalization errors

It now remains to compute the training and generalization errors from the free energy. Recalling that

$$\varepsilon_{\mathrm{tr}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{\nu=1}^{n}\ell\left(y^\nu, \hat{\boldsymbol{\theta}}^\top\boldsymbol{x}^\nu\right),$$

we can write using the definition of the free energy (45):

$$\lim_{n\to\infty}\varepsilon_{\mathrm{tr}} = \lim_{\beta\to\infty}\partial_\beta f_\beta - r(\hat{\boldsymbol{\theta}}).$$

Computing explicitly the derivative and averaging again over the data yields

$$\lim_{n\to\infty}\varepsilon_{\mathrm{tr}} = -\lim_{\beta\to\infty}\partial_\beta\Psi_\ell,$$

where Ψ_ℓ is the free energy contribution of the loss defined in (63). Writing explicitly this derivative,

$$-\partial_\beta\Psi_\ell = \sum_{c\in\mathcal{C}}\mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\int_{\mathbb{R}}\mathrm{d}y\frac{\mathcal{Z}_0\left(y,\omega_c^{(0)},\rho-\frac{m_c^2}{q_c}\right)}{\mathcal{Z}_\ell(y,\omega_c^{(\ell)},V_c)}\underbrace{\int_{\mathbb{R}}\frac{\mathrm{d}\lambda}{\sqrt{2\pi V_c}}e^{-(\lambda-\omega_c^{(\ell)})V_c^{-1}(\lambda-\omega_c^{(\ell)},V_c))-\beta\ell(y,\lambda)}\ell(y,\lambda)}_{\tilde{\mathcal{Z}}_\ell(y,\omega_c^{(\ell)},V_c)} \tag{73}$$

where

$$\omega_c^{(0)} = \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c \quad \text{and} \quad \omega_c^{(\ell)} = h_c + \sqrt{q_c}\xi_c.$$

We can now make the change of variables in (67), and use again Laplace's approximation: if

$$\eta_c = \text{prox}_{V\ell(y,\cdot)}(\omega_c^{(\ell)}) \quad \text{and} \quad \mathcal{L}(\lambda) = -(\lambda - \omega_c^{(\ell)})V_c^{-1}(\lambda - \omega_c^{(\ell)}) - \beta\ell(y,\lambda),$$

then

$$\mathcal{Z}_\ell(y, \omega_c^{(\ell)}, V_c) \sim e^{-\beta\mathcal{L}(\eta_c)} \quad \text{and} \quad \tilde{\mathcal{Z}}_\ell(y, \omega_c^{(\ell)}, V_c) \sim e^{-\beta\mathcal{L}(\eta_c)}\ell(y,\eta_c),$$

and all of the overlaps will concentrate around the solutions of (71), (72). Finally, the term containing $\mathcal{Z}_0$ is an expectation of y according to $P_0(y \mid \tau_c)$, where

$$\tau_c = \omega_c^{(0)} + \sqrt{\rho - \frac{m_c^2}{q_c}}\xi_c^{(0)}$$

and $\xi_c^{(0)}$ is a Gaussian variable independent from everything else. Putting all together, we can write

$$\lim_{n\to\infty} \varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) = \sum_{c\in\mathcal{C}} p_c \, \mathbb{E}_{\nu_c,\tau_c,y}\left[\ell\left(y, \text{prox}_{V_c^\star \ell(y,\cdot)}(\nu_c)\right)\right] \tag{74}$$

with

$$\begin{pmatrix}\nu_c \\ \tau_c\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\pi_c \\ h_c\end{bmatrix}, \begin{bmatrix}\rho_c & m_c^\star \\ m_c^\star & q_c^\star\end{bmatrix}\right) \quad \text{and} \quad y \sim P_0(\cdot \mid \tau_c). \tag{75}$$

The generalization error is much simpler to obtain: since $x_{\text{new}}, y_{\text{new}}$ are independent from the estimator $\hat{\boldsymbol{\theta}}$, we simply have

$$\lim_{n\to\infty} \varepsilon_{\text{gen}}(\hat{\boldsymbol{\theta}}) = \sum_{c\in\mathcal{C}} p_c \, \mathbb{E}_{\nu_c,\tau_c,y}\left[\ell\left(y,\nu_c\right)\right] \tag{76}$$

where ν_c, τ_c, y follow the same distribution as above.

A.9 Examples

Ridge penalty Consider a particular case of the general equations reported above: the case of a ridge penalty

$$r(\boldsymbol{\theta}) = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2.$$

We then have:

$$\text{prox}_r(\boldsymbol{x}) = (1+\lambda)^{-1}\boldsymbol{x} \tag{77}$$

which simplify the prior equations considerably. Indeed, we can now compute every expectation in (72), which yields the following fixed-point equations:

$$\begin{cases} \hat{V}_c &= -\alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right)\partial_\omega f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\ \hat{q}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)^2\right] \\ \hat{m}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \partial_\omega\mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \\ \hat{h}_c &= \alpha p_c \mathbb{E}_{\xi_c\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\; \mathcal{Z}_0\left(y, \pi_c + \frac{m_c}{\sqrt{q_c}}\xi_c, \rho - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi_c, V_c)\right] \end{cases} \tag{78}$$

$$\begin{cases} V_c &= \frac{1}{d}\,\text{tr}\left[\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right] \\ q_c &= \frac{1}{d}\,\text{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{q}_{c'}\Sigma_{c'} + \sum_{c',c''\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\hat{\boldsymbol{\mu}}_{c''}^\top\right)\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-2}\right] \\ m_c &= \frac{1}{d}\,\text{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\boldsymbol{\theta}_0^\top\right)\Sigma_c\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right] \\ h_c &= \frac{1}{d}\,\text{tr}\left[\left(\sum_{c'\in\mathcal{C}} \hat{\boldsymbol{\mu}}_{c'}\boldsymbol{\mu}_c^\top\right)\left(\lambda I_d + \hat{\Sigma}\right)^{-1}\right] \end{cases} \tag{79}$$

Gaussian covariate model The equations for the Gaussian covariate model can be found in [19]; they also correspond to taking $|\mathcal{C}| = 1$ in the ones above. We reproduce them here for completeness:

$$\begin{cases} \hat{V} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) \partial_\omega f_\ell(y, h + \sqrt{q}\xi, V) \right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V)^2 \right] \\ \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \partial_\omega \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V) \right] \\ \hat{h} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} \left[\int \mathrm{d}y \, \mathcal{Z}_0 \left(y, \pi + \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q} \right) f_\ell(y, h + \sqrt{q}\xi, V) \right] \end{cases} \tag{80}$$

$$\begin{cases} V &= \frac{1}{d} \operatorname{tr} \left[\Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ q &= \frac{1}{d} \operatorname{tr} \left[\left(\hat{q}\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \right) \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-2} \right] \\ m &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ h &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \end{cases} \tag{81}$$

where this time

$$\hat{\boldsymbol{\mu}} = \hat{h}\boldsymbol{\mu} + \hat{m}\Sigma\boldsymbol{\theta}_0 \qquad \text{and} \qquad \hat{\Sigma} = \hat{V}\Sigma. \tag{82}$$

The errors are given by

$$\lim_{n\to\infty} \varepsilon_{\text{tr}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\nu,\tau,y} \left[\ell \left(y, \operatorname{prox}_{V^\star \ell(y,\cdot)}(\nu) \right) \right] \tag{83}$$

$$\lim_{n\to\infty} \varepsilon_{\text{gen}}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\nu,\tau,y} \left[\ell \left(y, \nu \right) \right] \tag{84}$$

with

$$\begin{pmatrix} \nu \\ \tau \end{pmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \pi \\ h \end{bmatrix}, \begin{bmatrix} \rho & m^\star \\ m^\star & q^\star \end{bmatrix} \right) \quad \text{and} \quad y \sim P_0(\cdot \,|\, \tau). \tag{85}$$

Ridge regression We place ourselves in the ridge regression case, where

$$P_0(\tau|\Delta) = \mathcal{N}(\tau, \Delta) \qquad\qquad l(y, \hat{y}) = (y - \hat{y})^2. \tag{86}$$

In this case, the equations in (79) simplify even further, yielding

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\rho_c + \Delta + q_c - 2m_c + (h_c - \pi_c)^2) \\ \hat{m}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{h}_c &= \frac{\alpha p_c(\pi_c - h_c)}{1+V_c} \end{cases} \tag{87}$$

$$\begin{cases} V &= \frac{1}{d} \operatorname{tr} \left[\Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ q &= \frac{1}{d} \operatorname{tr} \left[\left(\hat{q}\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \right) \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-2} \right] \\ m &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top \Sigma \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \\ h &= \frac{1}{d} \operatorname{tr} \left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top \left(\lambda I_d + \hat{\Sigma} \right)^{-1} \right] \end{cases} \tag{88}$$

The training and generalization error also benefit from a very simple expression

$$\varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star}{\alpha} \qquad\qquad \varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star (1 + V_c^\star)^2}{\alpha}. \tag{89}$$

B Universality of Gaussian Mixture Models

In this section we prove the main results shown in in Sec. 3 regarding universality properties of Gaussian Mixture Models.

B.1 Mean Universality : proof of Proposition 3.2

Let us rewrite the assumptions here. Let $\{\hat{V}_c^\star\}_{c=1}^K$ be the fixed points of the (replica) saddle point equations describing the centered Gaussian mixture problem, i.e. the solutions of of eqs. (78), (79). We assume that the teacher vector and the data structure respect the following:

$$\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0 \qquad \forall c \in \mathcal{C} \tag{90}$$

$$\lim_{n,d\to\infty} \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_{c'} \left(\lambda + \sum_{c\in\mathcal{C}} \hat{V}_c^\star \Sigma_c\right)^{-1} \boldsymbol{\mu}_c \to 0 \qquad \forall (c,c') \in \mathcal{C}\times\mathcal{C} \tag{91}$$

and the loss and teacher are both symmetric:

$$\ell(x,y) = \ell(-x,-y) \tag{92}$$

$$P_0\left(y|\tau\right) = P_0\left(-y|-\tau\right). \tag{93}$$

In the saddle-point equations (78), (79), the mean vectors appear always coupled with the overlaps $\{h_c, \hat{h}_c\}_{c\in\mathcal{C}}$. Hence, to prove Prop. 3.2 it suffices to prove the following result:

Lemma B.1. *If $(a+b+c)$ hold, then $\{h_c = \hat{h}_c = 0\}_{c\in\mathcal{C}}$ is a fixed point for the problem.*

First consider the update equations for $\{h_c\}_{c\in\mathcal{C}}$:

$$h_c = \frac{1}{d}\sum_l \text{tr}\left[\left(\hat{h}_l \mu_l \mu_c^\top + \hat{m}_l \Sigma_l \boldsymbol{\theta}_0 \mu_c^\top\right)\left(\lambda I_d + \sum_{l'} \hat{V}_{l'}\Sigma_{l'}\right)^{-1}\right] \tag{94}$$

$$\underset{a+c}{\simeq} \quad \frac{1}{d}\sum_l \text{tr}\left[\left(\hat{h}_l \mu_l \mu_c^\top\right)\left(\lambda I_d + \sum_{l'} \hat{V}_{l'}\Sigma_{l'}\right)^{-1}\right] \tag{95}$$

By continuity of the saddle-point equations, if at the fixed point $\{\hat{h}_c^\star \to 0\}_{c\in\mathcal{C}}$ holds, we also easily have that $\{h_c^\star \to 0\}_{c\in\mathcal{C}}$. Now assume $\{h_c^\star = 0\}_{c\in\mathcal{C}}$ at the fixed point. We exploit symmetry argument inherent to the update functions to show that $\{\hat{h}_c^\star = 0\}_{c\in\mathcal{C}}$ under weak assumptions on the teacher and loss functions. We write the updates using assumption c):

$$\hat{h}_c = \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi + \pi_c, \rho_c - \frac{m_c^2}{q_c}\right) f_\ell(y, h_c + \sqrt{q_c}\xi, V)\right] \tag{96}$$

$$\underset{h_c\to 0}{\simeq} \quad \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi, \rho_c - \frac{m_c^2}{q_c}\right) f_\ell(y, \sqrt{q_c}\xi, V_c)\right] \tag{97}$$

Reminding the definition of the teacher measure term:

$$Z_0(y,\omega,V) = \mathbb{E}_{\lambda\sim\mathcal{N}(\omega,V)}\left[P_0(y|\lambda)\right] \qquad\qquad \lambda \equiv \frac{\boldsymbol{x}^\top\boldsymbol{\theta}_0}{\sqrt{d}} \tag{98}$$

we see that the symmetry conditions impose basic restrictions on the label generation:

$$f) \quad \mathbb{E}_{\lambda\sim\mathcal{N}(\omega,V)}\left[P_0(y|\lambda)\right] = \mathbb{E}_{\lambda\sim\mathcal{N}(-\omega,V)}\left[P_0(-y|\lambda)\right] \tag{99}$$

and hence $\mathcal{Z}_0$ is even in its second argument. Additionally, we have

$$f_\ell(y,\omega,V) = \frac{1}{V}(\text{prox}_{\,Vl(y,\cdot)}(\omega) - \omega) \qquad \text{prox}_{\,V\ell(y,\cdot)}(w) = \arg\min_z\left[\frac{1}{2V}(z-w)^2 + \ell(y,z)\right] \tag{100}$$

The symmetry condition on ℓ then implies that

$$\text{prox}_{\,V\ell(-y,\cdot)}(-\omega) = -\,\text{prox}_{\,V\ell(y,\cdot)}(\omega), \tag{101}$$

so f_ℓ is odd in its second argument. All that's left to notice is that (97) is an Gaussian integral of an odd function, hence it is equal to 0.

B.2 Gaussian Universality : proof of Proposition 3.3

We proved that the fixed point respects $\{h_c^\star, \hat{h}_c^\star = 0, 0\}_{c\in\mathcal{C}}$ under the hypothesis of Prop. 3.2. Assume now that the mixture is homogeneous:

$$\Sigma_c = \Sigma \qquad \forall c \in \mathcal{C} \tag{102}$$

Then we have

$$\varepsilon_{gen}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma\}_{c=1}^K\right) \simeq \varepsilon_{gen}^{\text{GMM}}\left(\mathbf{0}, \{\Sigma\}_{c=1}^K\right) \tag{103}$$

$$\varepsilon_{tr}^{\text{GMM}}\left(\{\boldsymbol{\mu}_c\}_{c=1}^K, \{\Sigma\}_{c=1}^K\right) \simeq \varepsilon_{tr}^{\text{GMM}}\left(\mathbf{0}, \{\Sigma\}_{c=1}^K\right) \tag{104}$$

But the right-hand side of those equations corresponds to a distribution of the form

$$\boldsymbol{x} \sim \sum_{c\in\mathcal{C}} \mathcal{N}(\mathbf{0}, \Sigma) = \mathcal{N}(\mathbf{0}, \Sigma)!$$

This proves Proposition 3.3

B.3 Covariance universality

Surprisingly in the limit of vanishing regularization, $\lambda \to 0^+$, the covariance is not relevant for the high dimensional learning problem. We show that when a unique minimizer of the loss exists, and we can take safely the limit $\lambda \to 0^+$ in the saddle point equations, the covariance Σ disappears completely from the overlap expression. Indeed assuming the minimizer $\hat{\boldsymbol{\theta}}$ is unique, we can safely simplify expression of the type:

$$\lim_{\lambda\to 0^+} \Sigma^n(\lambda I_d + \hat{V}\Sigma)^{n'} = \frac{1}{\hat{V}}\Sigma^{n-n'} \tag{105}$$

Then by plugging in the $\lambda \to 0^+$ simplification in eqs. (80), (81) we have complete independence from the covariance matrix Σ:

$$\begin{cases} \hat{V} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q_c}}\xi, \rho_c - \frac{m_c^2}{q_c}\right) \partial_\omega f_\ell(y, \sqrt{q_c}\xi, V_c)\right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m_c}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) f_\ell(y, \sqrt{q}\xi, V)^2\right] \\ \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\, \partial_\omega \mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) f_\ell(y, \sqrt{q}\xi, V)\right] \end{cases} \tag{106}$$

$$\begin{cases} V &= 1 \\ q &= \hat{q} + \rho\hat{m}^2 \\ m &= \hat{m}\rho \end{cases} \tag{107}$$

which concludes the proof.

B.4 Strong universality

We never made any specific assumption on the loss up to now, apart the very general symmetry condition in Assumption 2. In this section we show strong universality for square loss regression for *any GMM* estimation problem respecting the mean universality property (Prop. 3.2). We assume to consider a ridge regression problem in the underparametrized regime $\alpha > 1$:

$$P_0(\tau|\Delta) = \mathcal{N}(\tau, \Delta) \qquad\qquad \ell(y, \hat{y}) = (y - \hat{y})^2 \tag{108}$$

the replicas then correspond to equations (87), (88). In the $\lambda \to 0^+$ limit, the equation simplify greatly:

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\rho_c + \Delta + q_c - 2m_c) \\ \hat{m}_c &= \frac{\alpha p_c}{1+V_c} \end{cases} \tag{109}$$

$$\begin{cases} V_c &= \frac{1}{d}\operatorname{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}}\hat{V}_l\Sigma_l)^{-1}\right] \\ q_c &= \frac{1}{d}\sum_{l\in\mathcal{C}}^K \operatorname{tr}\left[\hat{q}_l\Sigma_l\Sigma_c(\sum_{l'\in\mathcal{C}}\hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \frac{1}{d}\sum_{(l,l)'\in\mathcal{C}\times\mathcal{C}}\operatorname{tr}\left[\hat{m}_l\hat{m}_{l'}\Sigma_l\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Sigma_{l'}\Sigma_c(\sum_{l''\in\mathcal{C}}\hat{V}_{l''}\Sigma_{l''})^{-2}\right] \\ m_c &= \frac{1}{d}\sum_{l=1}^K \operatorname{tr}\left[\hat{m}_l\Sigma_l\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\Sigma_c(\sum_{l'}\hat{V}_{l'}\Sigma_{l'})^{-1}\right] \end{cases} \tag{110}$$

Consider the equation for the overlaps $\{q_c\}_{c\in\mathcal{C}}$:

$$q_c = \frac{1}{d}\sum_{l=1}^{K} \operatorname{tr}\left[\hat{q}_l \Sigma_l \Sigma_c (\sum_{l'} \hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \frac{1}{d}\sum_{l,k=1}^{K} \operatorname{tr}\left[\hat{m}_l \hat{m}_k \Sigma_l \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^T \Sigma_k \Sigma_c (\sum_{l'} \hat{V}_{l'}\Sigma_{l'})^{-2}\right] \tag{111}$$

$$\equiv R_c + G_c \tag{112}$$

The error metrics can be decomposed as:

$$\varepsilon_{\text{tr}} = \sum_{c\in\mathcal{C}} \frac{p_c}{(1+V_c)^2}(R_c + \Delta) + \sum_{c\in\mathcal{C}} \frac{p_c}{(1+V_c)^2}(\rho_c + G_c - 2m_c) \tag{113}$$

$$\varepsilon_{\text{gen}} = \sum_{c\in\mathcal{C}} p_c(R_c + \Delta) + \sum_{c\in\mathcal{C}} p_c(\rho_c + G_c - 2m_c) \tag{114}$$

We focus on the second term and show that it is equal to zero. Looking at eqs. (110),(109), we note that:

$$\hat{m}_c = \hat{V}_c \qquad \forall c \in \mathcal{C} \tag{115}$$

By using eq. (115) we can simplify the equations for $\{\hat{m}_c, G_c\}$:

$$G_c = \rho_c, \qquad m_c = \rho_c \qquad \forall c \in \mathcal{C} \tag{116}$$

Plugging this relations in the saddle point equations we obtain:

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} = \hat{m}_c \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\Delta + R_c) \end{cases} \tag{117}$$

$$\begin{cases} V_c &= \frac{1}{d}\operatorname{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}} \hat{V}_l \Sigma_l)^{-1}\right] \\ q_c &= \frac{1}{d}\sum_{l=1}^{K} \operatorname{tr}\left[\hat{q}_l \Sigma_l \Sigma_c (\sum_{l'} \hat{V}_{l'}\Sigma_{l'})^{-2}\right] + \rho_c \equiv R_c + \rho_c \\ m_c &= \rho_c \end{cases} \tag{118}$$

The fixed point of the equations does not depend on $\{\rho_c\}_{\in\mathcal{C}}$ anymore and we are left with equations only for $\{\hat{V}_c, \hat{q}_c, V_c, R_c\}_{c\in\mathcal{C}}$:

$$\begin{cases} \hat{V}_c &= \frac{\alpha p_c}{1+V_c} \\ \hat{q}_c &= \frac{\alpha p_c}{(1+V_c)^2}(\Delta + R_c) \end{cases} \tag{119}$$

$$\begin{cases} V_c &= \frac{1}{d}\operatorname{tr}\left[\Sigma_c(\sum_{l\in\mathcal{C}} \hat{V}_l \Sigma_l)^{-1}\right] \\ R_c &= \frac{1}{d}\sum_{l=1}^{K} \operatorname{tr}\left[\hat{q}_l \Sigma_l \Sigma_c (\sum_{l'} \hat{V}_{l'}\Sigma_{l'})^{-2}\right] \end{cases} \tag{120}$$

The errors can be computed in a closed form with a series of algebraic manipulation. One can verify by algebraic manipulation the following relations:

$$\sum_c V_c \hat{V}_c = \sum_c \frac{1}{d}\operatorname{tr}\left[\hat{V}_c \Sigma_c (\sum_{l\in\mathcal{C}} \hat{V}_l \Sigma_l)^{-1}\right] = 1 \tag{121}$$

$$\sum_c R_c \hat{V}_c - V_c \hat{q}_c = \frac{1}{d}\sum_{l=1}^{K} \operatorname{tr}\left[\hat{q}_l \Sigma_l (\sum_c \hat{V}_c \Sigma_c)(\sum_{l'} \hat{V}_{l'}\Sigma_{l'})^{-2}\right] - \frac{1}{d}\operatorname{tr}\left[\sum_c \hat{q}_c \Sigma_c (\sum_{l\in\mathcal{C}} \hat{V}_l \Sigma_l)^{-1}\right] = 0 \tag{122}$$

Now we express everything in eq. (121) in terms of the non-hatted overlaps to obtain:

$$\star) \sum_c \frac{\alpha p_c}{1+V_c} V_c = 1 \tag{123}$$

$$\#) \sum_c \frac{\alpha p_c}{(1+V_c)^2} R_c = \sum_c \frac{\alpha p_c}{(1+V_c)^2} V_c \Delta \tag{124}$$

We remark that we can write the training and generalization error can be written as:

$$\varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c}{\alpha} \qquad\qquad \varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star (1+V_c^\star)^2}{\alpha} \tag{125}$$

So if we compute the quantity $\sum_c \hat{q}_c$ we conclude. By plugging in $(\#)$ into the expression above we obtain:

$$\sum_c \hat{q}_c = \sum_c \frac{\alpha p_c}{(1+V_c)^2}(R_c + \Delta) \tag{126}$$

$$\underset{(\#)}{=} \Delta \sum_c \frac{\alpha p_c}{(1+V_c)^2} + \Delta \sum_c \frac{\alpha p_c}{(1+V_c)^2} V_c \tag{127}$$

$$= \Delta \sum_c \frac{\alpha p_c}{1+V_c} \tag{128}$$

and finally using relation $(\star)$ we prove the theorem:

$$\varepsilon_{\text{tr}} = \sum_c \frac{\hat{q}_c^\star}{\alpha} \tag{129}$$

$$= \sum_c \frac{\hat{q}_c^\star}{\alpha} \pm \frac{\Delta}{\alpha} \tag{130}$$

$$\underset{(\star)}{=} \Delta(1 - \frac{1}{\alpha}) \tag{131}$$

On the other hand, the generalization error does not respect the strong universlaity statement as we analyze in the next section (See Fig. 5).

C Non-universality of Gaussian Mixture Models

In the previous section we enumerated a series of results unveiling universality of GMM. Now we want to study the dual task: when GMM model *are not* universal? We have two ways to break universality: a) allow strong heterogeneity in the data structure; b) consider labels which are strongly correlated with the data structure. The plan for this section is to review more in detail these two processes for universality breaking.

C.1 Strongly heterogeneous mixtures

We proved in Theorem. 3.6 a strong universality statement for the training loss of ridge regression. However in this section we want to clarify that the theorem is not valid beyond its assumption for general mixtures. We prove this by analyzing a counterexample: a strongly heterogeneous 2-clusters Gaussian Mixture:

$$\Sigma_+ = \text{diag}(0.1, \dots, 0.1, 1.9, \dots, 1.9) \qquad p_+ = 0.8 \tag{132}$$

$$\Sigma_- = \text{diag}(1.9, \dots, 1.9, 0.1, \dots, 0.1) \qquad p_- = 0.2 \tag{133}$$

We present in Fig. 5 the comparison of the heterogeneous GMM performance with the Gaussian theory: although the training errors coincide as predicted by Theorem. 3.6, the generalization errors are different. However, we remark that real data after preprocessing seem homogeneous enough to obtain a good agreement with the exact Gaussian asymtptotics as we see in Fig. 1.

C.2 Correlated teachers

In this section we investigate in deeper detail the controlled setting of Theorem. 3.7. We do not necessarily consider estimation problems which respects the assumption of mean universality property (Prop. 3.2), and in fact we precisely analyze the consequences if we relax these conditions. We perform ridge regression on a two-mixtures GMM with opposite means and same covariance matrix proportional to I_d. In this scenario the exact asymptotics for the performance of the ERM estimator admits a closed form expression. We remind here the main parameters for the theoretical analysis are:

$$\rho = \frac{1}{d}\boldsymbol{\theta}_0^\top \Sigma \boldsymbol{\theta}_0 = \frac{1}{d}\|\boldsymbol{\theta}_0\|_2^2 \qquad \gamma = \frac{1}{d}||\boldsymbol{\mu}||_2^2 \qquad \Delta = \mathbb{E}_\xi \xi^2 \qquad \hat{\boldsymbol{\mu}}_\pm = \hat{h}_\pm \boldsymbol{\mu}_\pm + \hat{m}_\pm \Sigma \boldsymbol{\theta}_0 \tag{134}$$

In this simple setting we can simplify the general mixture equations in eqs. (80),(81) without doing any assumption on the teacher. We first map the equation to a single Gaussian problem $\mathcal{G}_{\boldsymbol{\theta}, I_d, \boldsymbol{\mu}}$. We have to slightly modify the proof of

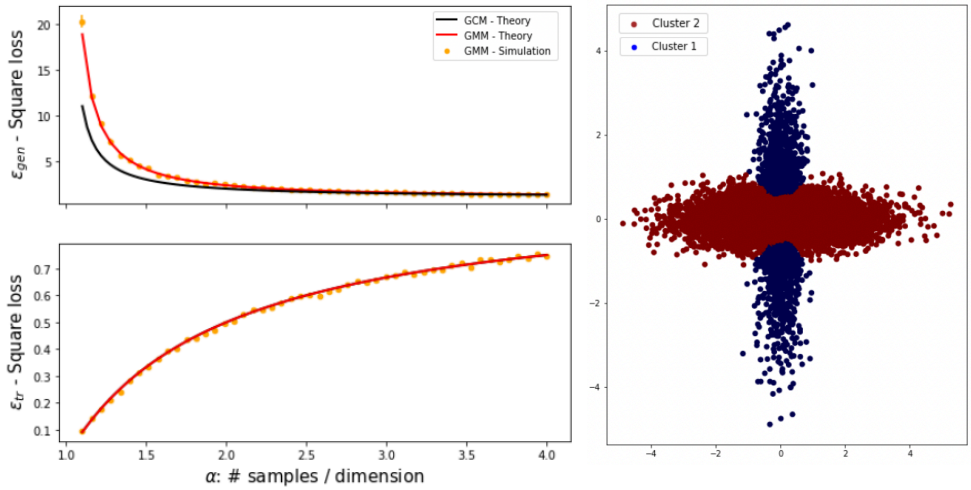

Figure 5: Training and generalization error for ridge regression on the GMM defined in eq. (132). **Left**: We compare the performance of the Gaussian asymtptotics (solid black line) with the GMM one (solid red line), while the orange dots represents simulations which agrees with the theoretical predictionas predicted by Theorem. 3.1. **Right**: Two dimensional toy plot of 10000 samples coming from the heterogeneous Gaussian mixture.

Prop. 3.3 presented in Sec. B.2, indeed we cannot blindly assume that the mean overlaps $\{h_+, h_-\}$ are zero but one can show that at the fixed point the overlaps respect:

$$\hat{h}_+^\star = -\hat{h}_-^\star, \qquad \hat{m}_+^\star = \hat{m}_-^\star \qquad \rightarrow \qquad \hat{\boldsymbol{\mu}}_+^\star = \hat{\boldsymbol{\mu}}_-^\star \tag{135}$$

$$\hat{V} = \hat{V}_+ + \hat{V}_- \qquad \hat{q} = \hat{q}_+ + \hat{q}_- \qquad \hat{m} = \hat{m}_+ + \hat{m}_- \qquad \hat{h} = \hat{h}_+ - \hat{h}_- \tag{136}$$

$$V = V_+ = V_- \qquad q = q_+ = q_- \qquad m = m_+ = m_- \qquad h = h_+ = -h_- \tag{137}$$

the fixed point of the replica equations for the mixture defined above are mapped to the one of a single Gaussian problem for the overlaps $\{V, m, q, h, \hat{V}, \hat{m}, \hat{q}, \hat{h}\}$. Now we can simplify them even further by plugging in the assumption on the covariance, all the traces simplify and we get:

$$\begin{cases} V &= \frac{1}{\lambda + \hat{V}} \\ q &= V^2\left(\hat{q} + \rho\hat{m}^2 + 2\pi\hat{m}\hat{h} + \gamma\hat{h}^2\right) \\ m &= V\left(\rho\hat{m} + \pi\hat{h}\right) \\ h &= V\left(\gamma\hat{h} + \pi\hat{m}\right) \end{cases} \tag{138}$$

These equations are actually solvable! Define

$$\eta := V\hat{V} = \frac{\alpha V}{1+V};$$

Now, we know that the generalization error satisfies:

$$V^2\hat{q} = \frac{V^2\alpha}{(1+V)^2}\varepsilon_{\text{gen}} = \frac{\eta^2}{\alpha}(\varepsilon_{\text{gen}}).$$

We can plug this into the equation for q to get

$$\varepsilon_{\text{gen}} = \rho + \underbrace{\left(\frac{\eta^2}{\alpha}(\varepsilon_{\text{gen}}) + \rho(V\hat{m})^2 + 2\pi(V\hat{m})(V\hat{h}) + \gamma(V\hat{h})^2\right)}_{q} - 2\underbrace{\left(\rho V\hat{m} + \pi V\hat{h}\right)}_{m} + (\pi - h)^2 \tag{139}$$

$$\varepsilon_{\text{tr}} = \frac{\varepsilon_{\text{gen}}}{(1+V)^2} \tag{140}$$

with the relations:

$$V = \frac{\eta}{\alpha - \eta}, \qquad V\hat{m} = \eta, \quad V\hat{h} = \eta(\pi - h)$$

and the expression of $\pi - h$ as a function of η:

$$\pi - h = \pi - V(\gamma\hat{h} + \pi\hat{m}) = \pi - \gamma\eta(\pi - h) - \pi\eta \iff \pi - h = \pi\frac{1-\eta}{1+\gamma\eta} \tag{141}$$

we simplify everything in terms only of η to get:

$$\varepsilon_{\text{gen}} = \frac{\alpha\left(\Delta + (\eta-1)^2\rho\right)}{(\alpha - \eta^2)}\left(1 - \pi^2\frac{(\eta-1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{\left(\Delta + (\eta-1)^2\rho\right)(1+\gamma\eta)^2}\right) \tag{142}$$

$$\varepsilon_{\text{tr}} = \frac{(\alpha - \eta)^2\left(\Delta + (\eta-1)^2\rho\right)}{\alpha\,(\alpha - \eta^2)}\left(1 - \pi^2\frac{(\eta-1)^2\left(\gamma\eta^2 + 2\eta - 1\right)}{\left(\Delta + (\eta-1)^2\rho\right)(1+\gamma\eta)^2}\right) \tag{143}$$

All that remains is to solve for η using the equations for V and $\hat{V}$, which yields

$$\eta = 1 + \frac{1}{2}\left(\alpha - 1 + \lambda - \sqrt{4\lambda + (\alpha - 1 + \lambda)^2}\right). \tag{144}$$

Vanishing regularization We can take $\lambda = 0$ inside (144) even when $\alpha < 1$, and we find

$$\eta(\lambda = 0) = \min(\alpha, 1) \tag{145}$$

Finally, this yields

$$\varepsilon_{gen} = \begin{cases} \frac{\alpha\Delta}{\alpha-1} & \alpha \geq 1 \\ \frac{\pi^2(\alpha-1)\left(\alpha^2\gamma + 2\alpha - 1\right)}{(\alpha\gamma+1)^2} - \frac{(\alpha-1)^2\rho + \Delta}{\alpha-1} & \text{else} \end{cases} \tag{146}$$

$$\varepsilon_{\text{train}} = \begin{cases} \frac{(\alpha-1)\Delta}{\alpha} & \alpha \geq 1 \\ 0 & \text{else} \end{cases} \tag{147}$$

We retrieve the results of Corollary. 3.8: even with correlated teachers we show that in the underparametrized regime we have Gaussian universality.

Extension of Corollary. 3.8: The Gaussian universality result for correlated teacher at vanishing regularization can be extended as well to general balanced mixtures with homoscedastic covariance, i.e. the case where

$$p_c = \frac{1}{|\mathcal{C}|}, \quad \Sigma_c = \Sigma, \quad \lambda = 0.$$

In order to do so, we consider the general saddle point equation for ridge regression in eq. (87),(88), and plug in the assumptions:

$$\begin{cases} \hat{V} &= \frac{\alpha}{|\mathcal{C}|(1+V)} \\ \hat{q}_c &= \frac{\alpha}{|\mathcal{C}|(1+V)^2}(\rho + \Delta + q_c - 2m + (h_c - \pi_c)^2) \\ \hat{m} &= \frac{\alpha}{|\mathcal{C}|(1+V)} \\ \hat{h}_c &= \frac{\alpha(\pi_c - h_c)}{|\mathcal{C}|(1+V)} \end{cases} \tag{148}$$

$$\begin{cases} V &= \frac{1}{d}\operatorname{tr}\left[\Sigma\hat{\Sigma}^{-1}\right] \\ q_c &= \frac{1}{d}\operatorname{tr}\left[\left(\hat{q}_c\Sigma + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right)\Sigma\hat{\Sigma}^{-2}\right] \\ m &= \frac{1}{d}\operatorname{tr}\left[\hat{\boldsymbol{\mu}}\boldsymbol{\theta}_0^\top\Sigma\hat{\Sigma}^{-1}\right] \\ h_c &= \frac{1}{d}\operatorname{tr}\left[\hat{\boldsymbol{\mu}}\boldsymbol{\mu}_c^\top\hat{\Sigma}^{-1}\right] \end{cases} \tag{149}$$

In particular, these assumptions imply that $V_c, \hat{V}_c, m_c, \hat{m}_c$ do not depend on the class labels c, and hence

$$\hat{\Sigma} = \hat{V}\Sigma.$$

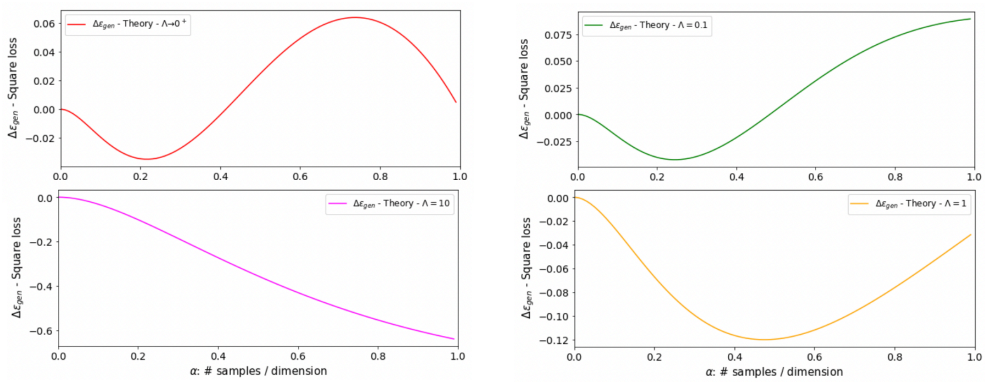

Figure 6: Theoretical prediction from eq. (158) for the difference between the generalization error computed from the Gaussian theory and GMM one plotted as a function of the number of samples used in the fit in Algorithm. C.3. We use different regularization strength $\Lambda \in \{1e-10, 0.1, 1, 10\}$, increasing clockwise in the figure.

Now, consider the assignment

$$\hat{h}_c = 0 \quad \text{and} \quad h_c = \pi_c; \tag{150}$$

is is easy to check that they are a fixed point of the above saddle-point equations, since then

$$h_c = \frac{\hat{m}\pi_c}{\hat{V}} = \pi_c$$

Plugging in this relation in the expression of the update for $\{\hat{q}_c\}_{c\in\mathcal{C}}$ we obtain:

$$\hat{q}_c^\star = \frac{\alpha p_c}{(1+V_c^\star)^2}(\rho_c + \Delta + q_c^\star - 2m_c^\star + (h_c^\star - \pi_c)^2) = \frac{\alpha p_c}{(1+V_c^\star)^2}(\rho_c + \Delta + q_c^\star - 2m_c^\star) \tag{151}$$

Hence by looking at the expression of generalization and training error for ridge regression in eq. (125) we conclude that they will not depend on the values of the mean overlaps $\{h_c^\star\}_{c\in\mathcal{C}}$ and we can prove Gaussian universality in the same way as we did in the proof of Theorem. 3.4.

C.3 Interpolating teachers

Different works observed over different datasets that if we find an interpolating teacher vector, in such a way that we can keep the real labels to study ERM performance, the theoretical predictions coming from Gaussian asymptotics would agree with simulations, [19] among others. This observation apparently contradicts the result of Theorem. 3.7 in which we show that correlated labels (as we expect real labels to be) break Gaussian universality. We try to motivate this in the same setting as Theorem. 3.7:

Theorem C.1. *(Real labels universality) Consider the same setting of Theorem. 3.7. Fix now the teacher vector to be the maximally correlated one $\boldsymbol{\theta} = \boldsymbol{\mu}$ and find an interpolating teacher $\tilde{\boldsymbol{\theta}}$ using $\tilde{n}$ samples coming from the GMM. If we compare the exact asymptotics of Gaussian and GMM for regularization Λ at fixed $\alpha = \frac{\tilde{n}}{d}$ we obtain a discrepancy:*

$$\Delta\varepsilon_{gen} = \frac{4(E-1)\alpha^2\left(E^2\gamma + 2e - 1\right)}{(\alpha+1)^2(E\gamma+1)^2} \tag{152}$$

with E(α,Λ) solution of:

$$E = 1 + \frac{1}{2}\left(\alpha - 1 + \Lambda - \sqrt{4\Lambda + (\alpha - 1 + \Lambda)^2}\right). \tag{153}$$

The result above solve the apparent contradiction: if we strongly overparametrize the fitting model, interpolating teachers can be uncorrelated with the data structure.

The proof of the results relies strongly on the fact that we can analyze analytically the performance in this scenario being it the same as the previous section. Reminding from eq. (145) that in the limit $\lambda \to 0$ we have in the overparametrized

Algorithm 1 Get theoretical learning curves keeping real labels

Data generation: Generate the data from the 2-clusters GMM described in C.1 and set:

$$\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \sqrt{\gamma}(1, \dots, 1)^\top, \qquad \Sigma_+ = \Sigma_- = I_d, \qquad \boldsymbol{\theta}_0 = \boldsymbol{\mu}_+ \tag{154}$$

Fit interpolating teacher Perform ridge regression for $\lambda \to 0$. Take $\alpha = \frac{n}{d} < 1$ so that the ERM estimator will interpolate the data.

Run exact asymptotics Compute the theoretical performance of the associated 2-GMM and equivalent GCM from the estimated $\hat{\theta}$ for the same α with a fixed regularization parameter Λ.

setting $\eta(\alpha, \lambda = 0) = \alpha$, we can write:

$$h_{erm} = \frac{\boldsymbol{\mu}^\top \hat{\boldsymbol{\theta}}}{d} = \gamma\alpha\frac{\gamma + 1}{1 + \gamma\alpha} \tag{155}$$

$$q_{erm} = \frac{\hat{\boldsymbol{\theta}}^\top \Sigma \hat{\boldsymbol{\theta}}}{d} = \alpha\left(-\frac{\Delta}{\alpha - 1} - \frac{(\alpha - 1)\pi^2}{\alpha\gamma + 1} + \rho\right) \tag{156}$$

We summarize the algorithmic procedure in in Algorithm. C.3. In order to compute the generalization error from eq. (142) we need the quantities $\{\rho, \pi\}$. Note that in Algorithm. C.3 we compute the theoretical curve for the GMM and the GCM from the estimated $\hat{\boldsymbol{\theta}}$. This translates into the fact that the estimated overlaps from numerical simulations in the previous step become the equivalent to:

$$\pi = h_{erm} \qquad\qquad \rho = q_{erm} \tag{157}$$

We fix the regularization for the theoretical prediction to be Λ, and for simplicity we use the same sample complexity parameter α. We are now in a position to compare Gaussian and GMM theoretical prediction, more precisely we analyze the difference of the generalization errors for the two data model. Recalling the expression for the generalization error in eq. (142), we just need to plug the estimated overalps in eq. (157) to obtain:

$$\Delta\varepsilon_{\text{gen}} = \frac{4\alpha^2(E - 1)\left(E^2\gamma + 2e - 1\right)}{(\alpha + 1)^2(E\gamma + 1)^2} \tag{158}$$

where $E(\alpha, \Lambda)$ is the solution of eq. (144), retrieving the result in Theorem. C.1. For the sake of clarity we stated the theorem in a simple setting, indeed once we estimated $\hat{\boldsymbol{\theta}}$ we could have decided to change the value of the sample complexity, as we are now interested in running theory curves. However, to simplify the equations we assumed to run the theoretical prediction for the same value of α as in the ERM fit. We present the results in Fig. 6 for $\gamma = 1$: although the true labels are maximally correlated with the data structure, we see that the Gaussian theoretical predictions with the interpolating teacher $\hat{\boldsymbol{\theta}}$ are very close to the GMM one. We remark that the upper-left plot in Fig. 1 is a nice characterization of Theorem. (3.8): as we reach the underparametrized regime we restore Gaussian universality for $\Lambda \to 0^+$ even for correlated teachers.

D Details on real dataset simulations

In this section we report the procedure to create the random regression task on real data, see Algorithm. D. The implementation of the different numerical simulations described in this work are available in a GitHub respository.

Algorithm 2 Random teacher regression on real data

Data processing: Load data and perform the preprocessing step with a transform matrix $F \in \mathbb{R}^{d\times d'}$, with d' dimension of the images in the chosen dataset. We choose for all the figures in this manuscript $d = 2000$.
Match covariance Compute $\hat{\Sigma} = \frac{1}{n}XX^\top$
Create new labels Forget the real labels associated with the dataset and create new label according to:

$$y_\nu = \begin{cases} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{x}_\nu}{\sqrt{d}} + \sqrt{\Delta}\xi & \text{Ridge regression} \\ \text{sign}\left(\frac{\boldsymbol{\theta}_0^\top \mathrm{x}_\nu}{\sqrt{\mathrm{d}}} + \sqrt{\Delta}\xi\right) & \text{Logistic regression} \end{cases} \tag{159}$$

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, I_d) \qquad \xi \sim \mathcal{N}(0,1) \tag{160}$$

Run learning curves Fix the regularization parameter λ.
for α in a given range **do**
Simulation Solve the ERM in eq. (1) using sklearn package *LogisticRegression* [53] or the Moore-Penrose pseudo-inverse for the ridge estimator:

$$\hat{\boldsymbol{\theta}} = \begin{cases} X^\top(XX^\top + \lambda I_d)^{-1}Y & n < d \\ (X^\top X + \lambda I_d)^{-1}X^\top y & n > d \end{cases} \tag{161}$$

Gaussian theory Solve saddle point equations in eqs. (80),(81) defining Gaussian asymptotics for the model. 2.1.
Compute erros Compute the training loss and generalization error using the metrics:

$$g(y,\hat{y}) = \begin{cases} (y-\hat{y})^2 & \text{Ridge regression} \\ \mathbb{P}(y \neq \hat{y}) & \text{Logistic regression} \end{cases} \tag{162}$$

end for