Universality laws for Gaussian mixtures in generalized linear models

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, Lenka Zdeborová

Universality laws for Gaussian mixtures in generalized linear models

Yatin Dandi[1], Ludovic Stephan[1], Florent Krzakala[1], Bruno Loureiro[2], and Lenka Zdeborová[3]

[1]École Polytechnique Fédérale de Lausanne (EPFL), IdePHICS Lab, CH-1015 Lausanne, Switzerland
[2]Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France
[3]École Polytechnique Fédérale de Lausanne (EPFL), SPOC Lab, CH-1015 Lausanne, Switzerland

February 20, 2023

Abstract

Let $(\boldsymbol{x}_i, y_i)_{i=1,\dots,n}$ denote independent samples from a general mixture distribution $\sum_{c\in\mathcal{C}} \rho_c P_c^{\boldsymbol{x}}$, and consider the hypothesis class of generalized linear models $\hat{y} = F(\boldsymbol{\Theta}^\top \boldsymbol{x})$. In this work, we investigate the asymptotic joint statistics of the family of generalized linear estimators $(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_M)$ obtained either from (a) minimizing an empirical risk $\hat{R}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})$ or (b) sampling from the associated Gibbs measure $\exp(-\beta n \hat{R}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}))$. Our main contribution is to characterize under which conditions the asymptotic joint statistics of this family depends (on a weak sense) only on the means and covariances of the class conditional features distribution $P_c^{\boldsymbol{x}}$. In particular, this allow us to prove the universality of different quantities of interest, such as the training and generalization errors, redeeming a recent line of work in high-dimensional statistics working under the Gaussian mixture hypothesis. Finally, we discuss the applications of our results to different machine learning tasks of interest, such as ensembling and uncertainty quantification.

1 Introduction

A recurrent topic in high-dimensional statistics is the investigation of the typical properties of signal processing and machine learning methods on synthetic, *i.i.d.* Gaussian data, a scenario often known under the umbrella of *Gaussian design* [Bartlett et al., 2020, Candès et al., 2020, Donoho and Tanner, 2009, Korada and Montanari, 2011, Monajemi et al., 2013]. A less restrictive assumption arises when considering that many machine learning tasks deal with data partitioned into a fixed number of classes. In these cases, the data distribution is naturally described by a *mixture model*, where each sample is generated *conditionally* on the class. In other words: data is generated by first sampling the class assignment and *then* generating the input conditioned on the class. Arguably the simplest example of such distributions is that of a *Gaussian mixture*, which shall be our focus in this work.

Gaussian mixtures are a popular model in high-dimensional statistics since, besides being an universal approximator, they often lead to mathematically tractable problems. Indeed, a recent line of work has analyzed the asymptotic performance of a large class of machine learning problems in the proportional high-dimensional limit under the Gaussian mixture data assumption, see e.g. Kini and Thrampoulidis [2021], Loureiro et al. [2021b], Mai and Liao [2019], Mignacco et al. [2020], Refinetti et al. [2021], Taheri et al.

[2020], Wang and Thrampoulidis [2021]. The key goal of the present work is to show that this assumption, and hence the conclusions derived therein, are far more general than previously anticipated.

We build on a recent line of works [Goldt et al., 2022, Hu and Lu, 2022, Montanari and Saeed, 2022] that have proven the asymptotic (single) Gaussian equivalence of generalized linear estimation on non-linear feature maps satisfying certain regularities conditions, a topic that has started with the work of El Karoui [2010] on kernel matrices. Furthermore, there is strong empirical evidence that Gaussian universality holds in a more general sense [Loureiro et al., 2021a]. A crucial limitation of the results in Hu and Lu [2022], Montanari and Saeed [2022], however, is the assumption of a target function depending on linear projections in the latent or feature space. Instead, we consider a rich class of mixture distributions, allowing arbitrary dependence between the class labels and the data.

Here, we extend this line of works and provide rigorous justification for universality in various settings such as empirical risk minimization (ERM), sampling, ensembling, etc. for *general mixture* distributions. Namely, we shall show that the statistics of generalized estimators obtained either from ERM or sampling on a mixture model asymptotically agrees (in a weak sense) with the statistics of estimators from the same class trained on a Gaussian mixture model with matching first and second order moments. In particular, this implies the universality of different quantities of interest, such as the training and generalization errors.

Our **main contributions** are as follows:

- We extend the Gaussian universality of empirical risk minimization theorems in Goldt et al. [2022], Hu and Lu [2022], Montanari and Saeed [2022] to generic mixture distribution and an equivalent mixture of Gaussians. In particular, we show that a Gaussian mixture observed through a random feature map is also a Gaussian mixture in the high-dimensional limit, a fact used for instance (without rigorous justification) in Loureiro et al. [2021b], Refinetti et al. [2021].
- A consequence of our results is that, with conditions on the matrix weights, data generated by conditional Generative Adversarial Networks (cGAN) behave as a Gaussian mixture when observed through the prism of generalized linear models (kernels, feature maps, etc...), as illustrated in Figs 1 and 2. This further generalizes the work of Seddik et al. [2020] that only considered the universality of Gram matrices for GAN generated data through the prism of random matrix theory.
- We consider setups involving multiple sets of parameters arising from simultaneous minimization of different objectives as well as sampling from Gibbs distributions defined by the empirical risk. This provides a unified framework for establishing the asymptotic universality of arbitrary functions of the set of minimizers or samples from different Gibbs distributions. For instance, it includes ensembling [Loureiro et al., 2022]) and uncertainty quantification [Clarté et al., 2022a,b] settings.
- We finally show how, in common setups, universality holds for a large class of functions, leading to the equivalence between the distributions of the minimizers themselves, and provide a theorem for their weak convergence.

Related work — Universality is an important topic in applied mathematics, as it motivates the scope of tractable mathematical models. It has been extensively studied in the context of random matrix theory [Tao and Vu, 2011, 2012], signal processing problems [Abbara et al., 2020, Abbasi et al., 2019, Donoho and Tanner, 2009, Dudeja et al., 2022, Korada and Montanari, 2011, Monajemi et al., 2013, Montanari and Nguyen, 2017, Panahi and Hassibi, 2017] and kernel methods [El Karoui, 2010, Lu and Yau, 2022, Misiakiewicz, 2022]. Closer to us is the recent stream of works that investigated the Gaussian universality of the asymptotic error of generalized linear models trained on non-linear features, starting from single-layer random feature maps [Dhifallah and Lu, 2020, Gerace et al., 2020, Hu and Lu, 2022, Montanari et al., 2019] and later extended to single-layer NTK [Montanari and Saeed, 2022] and deep random features [Schröder et al., 2023]. These

results, together with numerical observations that Gaussian universality holds for more general classes of features, led to the formulation of different *Gaussian equivalence* conjectures [Goldt et al., 2019, 2022, Loureiro et al., 2021a]. A complementary line of research has investigated cases in which the distribution of the features is multi-modal, suggesting a Gaussian mixture universality class instead [Louart et al., 2018, Seddik et al., 2020, 2021]. A bridge between these two lines of work has been recently investigated for generalized linear estimation with random labels in Gerace et al. [2022].

2 Setting and motivation

Consider a supervised learning problem where the training data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$, $i \in [n] \coloneqq \{1, \cdots, n\}$ is independently drawn from a mixture distribution:

$$\boldsymbol{x}_i \sim \sum_{c \in \mathcal{C}} \rho_c P_c^{\boldsymbol{x}}, \qquad \mathbb{P}(c_i = c) = \rho_c, \tag{1}$$

with c_i a categorical random variable denoting the cluster assignment for the i_{th} example $\mathbf{x}_i$. Let $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ denote the mean and covariance of $P_c^{\boldsymbol{x}}$, and $k = |\mathcal{C}|$. Further, assume that the labels y_i are generated from the following target function:

$$y_i(\boldsymbol{X}) = \eta(\boldsymbol{\Theta}_\star^\top \boldsymbol{x}_i, \varepsilon_i, c_i), \tag{2}$$

where $\boldsymbol{\Theta}_\star \in \mathbb{R}^{k \times p}$ and ε_i is an i.i.d source of randomness. It is important to stress that the class labels (2) are themselves not constrained to arise from a simple function of the inputs $\boldsymbol{x}_i$. For instance, the functional form in (2) includes the case where the labels are exclusively given by a function of the mixture index $y_i = \eta(c_i)$. This will allow us to handle complex targets, such as data generated using conditional Generative Adversarial Networks (cGANs).

In this manuscript, we will be interested in the hypothesis class defined by the following parametric predictor $\hat{y} = F(\boldsymbol{\Theta}^\top \boldsymbol{x})$, where $\boldsymbol{\Theta} \in \mathbb{R}^{k \times p}$ are the parameters and $F : \mathbb{R}^k \to \mathcal{Y}$ an activation function. For a given loss function $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ and regularization term $r : \mathbb{R}^{k \times p} \to \mathbb{R}_+$, define the (regularized) empirical risk over the training data:

$$\widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\Theta}^\top \boldsymbol{x}_i, y_i) + r(\boldsymbol{\Theta}), \tag{3}$$

where we have defined the feature matrix $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ by stacking the features $\boldsymbol{x}_i$ column-wise and the labels y_i in a vector $\boldsymbol{y} \in \mathcal{Y}^n$. In what follows, we will be interested in the following two tasks:
(i) Minimization: in a minimization task, the statistician's goal is to find a good predictor by minimizing the empirical risk (3), possibly over a constraint set $\mathcal{S}_p$:

$$\hat{\boldsymbol{\Theta}}_{\text{erm}}(\boldsymbol{X}, \boldsymbol{y}) \in \underset{\boldsymbol{\Theta} \in \mathcal{S}_p}{\arg\min}\, \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}), \tag{4}$$

This encompasses diverse settings such as generalized linear models with noise, mixture classification, but also the random label setting (with $\eta(x, \varepsilon) = \varepsilon$). In the following, we denote $\widehat{\mathcal{R}}_n^\star(\boldsymbol{X}, \boldsymbol{y}) \coloneqq \min_{\boldsymbol{\Theta}} \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})$
(ii) Sampling: here, instead of minimizing the empirical risk (3), the statistician's goal is to sample from a Gibbs distribution that weights different hypothesis according to their empirical error:

$$\boldsymbol{\Theta}_{\text{Bayes}}(\boldsymbol{X}, \boldsymbol{y}) \sim P_{\text{Bayes}}(\boldsymbol{\Theta}) \propto \exp\left(-\beta n \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})\right) \mathrm{d}\mu(\boldsymbol{\Theta}) \tag{5}$$

where μ is reference prior measure and $\beta > 0$ is a parameter known as the *inverse temperature*. Note that minimization can be seen as a particular example of sampling when $\beta \to \infty$, since in this limit the above measure peaks on the global minima of (4).

Applications of interest— So far, the setting defined above is quite generic, and the motivation to study this problem might not appear evident to the reader. Therefore, we briefly discuss a few scenarios of interest which are covered this model.

(i) *Conditional GANs (cGANs):* were introduced by Mirza and Osindero [2014] as a generative model to learn mixture distributions. Once trained in samples from the target distribution, they define a function Ψ that maps Gaussian mixtures (defining the latent space) to samples from the target mixture that preserve the label structure. In other words, conditioned on the label:

$$\forall c \in \mathcal{C}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \mapsto \boldsymbol{x}_c = \Psi(\boldsymbol{z}, c) \sim P_c^{\boldsymbol{x}} \tag{6}$$

The connection to model (1) is immediate. This scenario was extensively studied by Louart et al. [2018], Seddik et al. [2020, 2021], and is illustrated in Fig. 1. In Fig. 2 we report on a concrete experiment with a cGAN trained on the fashion-MNIST dataset.

(ii) *Multiple objectives:* Our framework also allows to characterize the joint statistics of estimators $(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_M)$ obtained from empirical risk minimization and/or sampling from different objective functions $\hat{R}_n^m$ defined on the same training data $(\boldsymbol{X}, \boldsymbol{y})$. This can be of interest in different scenarios. For instance, Clarté et al. [2022a,b] has characterized the correlation in the calibration of different uncertainty measures of interest, e.g. last-layer scores and Bayesian training of last-layer weights. This crucially depends on the correlation matrix $\hat{\boldsymbol{\Theta}}_{\text{erm}} \boldsymbol{\Theta}_{\text{Bayes}}^\top \in \mathbb{R}^{k \times k}$ which fits our framework.

(iii) *Ensemble of features:* Another example covered by the multi-objective framework above is that of ensembling. Let $(\boldsymbol{z}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$ denote some training data from a mixture model akin to (1). A popular ensembling scheme often employed in the context of deep learning [Lakshminarayanan et al., 2017] is to take a family of M feature maps $\boldsymbol{z}_i \mapsto \boldsymbol{x}_i^{(m)} = \varphi_m(\boldsymbol{z}_i)$ (e.g. neural network features trained from different random initialization) and train M independent learners:

$$\hat{\boldsymbol{\Theta}}_{\text{erm}}^{(m)} \in \underset{\boldsymbol{\Theta} \in \mathcal{S}_p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\Theta}^\top \boldsymbol{x}_i^{(m)}, y_i) + r(\boldsymbol{\Theta}) \tag{7}$$

Prediction on a new sample $\boldsymbol{z}$ is then made by assembling the independent learners, e.g. by taking their average $\hat{\boldsymbol{y}} = {}^1\!/_M \sum_{m=1}^M \hat{\boldsymbol{\Theta}}_{\text{erm}}^{(m)\top} \varphi_m(\boldsymbol{z})$. A closely related model was studied in d'Ascoli et al. [2020], Geiger et al. [2020], Loureiro et al. [2022].

Note that in all the applications above, having the labels depending on the features $\boldsymbol{X}$ would not be natural, since they are either generated from a latent space, as in (i), or chosen by the statistician, as in $(ii), (iii)$. Indeed, in these cases the most natural label model is given by the mixture index $y = c$ itself, which is a particular case of (2). This highlights the flexibility of the our target model with respect to prior work [Montanari and Saeed, 2022]. Instead, Hu and Lu [2022] assumes that the target is a function of a *latent variable*, which would correspond to a mismatched setting. The discussion here could be generalized also to this case, but would require an additional assumption, which we discuss in Appendix B.

Universality — Given these tasks, the goal of the statistician is to characterize different statistical properties of these predictors. These can be, for instance, point performance metrics such as the empirical and population risks, or uncertainty metrics such as the calibration of the predictor or moments of the posterior distribution

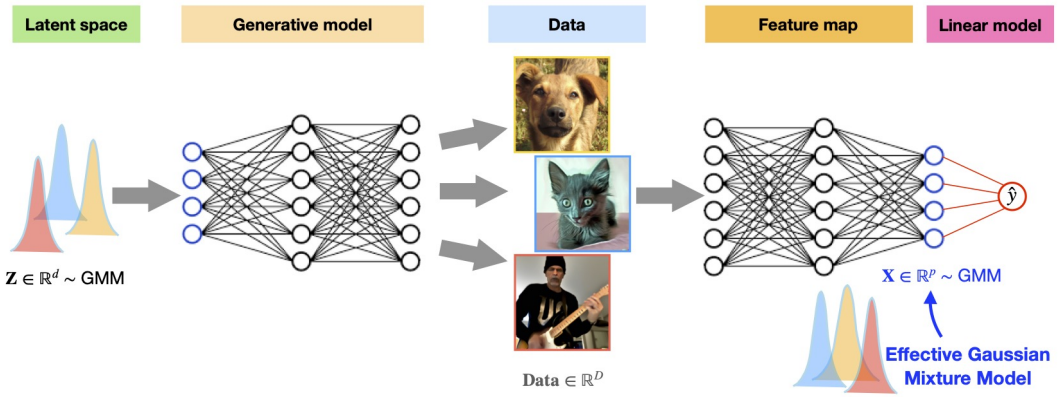

Figure 1: Illustration of Corollary 2: high-dimensional data generated by generative neural networks starting from a mixture of Gaussian in latent space ($\mathbf{z} \in \mathbb{R}^H$) are (with conditions on the weights matrices) equivalent, in high-dimension and for generalized linear models, to data sampled from a Gaussian mixture. A concrete example is shown in Fig. 2.

(5). These examples, as well as many different other quantities of interest, are functions of the joint statistics of the pre-activations $(\boldsymbol{\Theta}_\star^\top \boldsymbol{x}, \boldsymbol{\Theta}^\top \boldsymbol{x})$, for $\boldsymbol{x}$ either a test or training sample from (1). For instance, in a Gaussian mixture model, where $\boldsymbol{x} \sim \sum_{c\in\mathcal{C}} \rho_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, the sufficient statistics are simply given by the first two moments of these pre-activations. However, for a general mixture model (1), the sufficient statistics will generically depend on all moments of these pre-activations. Surprisingly, our key result in this work is to show that in the high-dimensional limit this is not the case. In other words, under some conditions which are made precise in Section 3, we show that expectations with respect to (1) can be exchanged by expectations over a Gaussian mixture with matching moments. This can be formalized as follows. Define an *equivalent Gaussian data set* $(\boldsymbol{g}_i, y_i)_{i=1}^n \in \mathbb{R}^p \times \mathcal{Y}$ with samples independently drawn from the *equivalent Gaussian mixture model*:

$$\boldsymbol{g}_i \sim \sum_{c\in\mathcal{C}} \rho_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \qquad y_i(\boldsymbol{G}) = \eta(\boldsymbol{\Theta}_\star^\top \boldsymbol{g}_i, \varepsilon_i, c_i). \tag{8}$$

We recall that $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ denotes the mean and covariance of $P_c^{\boldsymbol{x}}$ from (1). Consider a family of estimators $(\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_M)$ defined by minimization (3) and/or sampling (5) over the training data $(\boldsymbol{X}, \boldsymbol{y})$ from the mixture model (1). Let h be a statistical metric of interest. Then, in the proportional high-dimensional limit where $n, p \to \infty$ at fixed $k, k, M \in \mathbb{Z}_+$ sample complexity $\alpha = n/d > 0$, and where $\langle\cdot\rangle$ denote the expectation with respect to the Gibbs distribution (5), we define universality as:

$$\mathbb{E}_{\boldsymbol{X}} \left[\langle h(\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_M)\rangle_{\boldsymbol{X}}\right] \underset{n\to\infty}{\simeq} \mathbb{E}_{\boldsymbol{G}} \left[\langle h(\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_M)\rangle_{\boldsymbol{G}}\right] \tag{9}$$

The goal of the next section is to make this statement precise.

3 Main results

We now present the main theoretical contributions of the present work and discuss its consequences. We work under the following regularity and concentration assumptions:

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

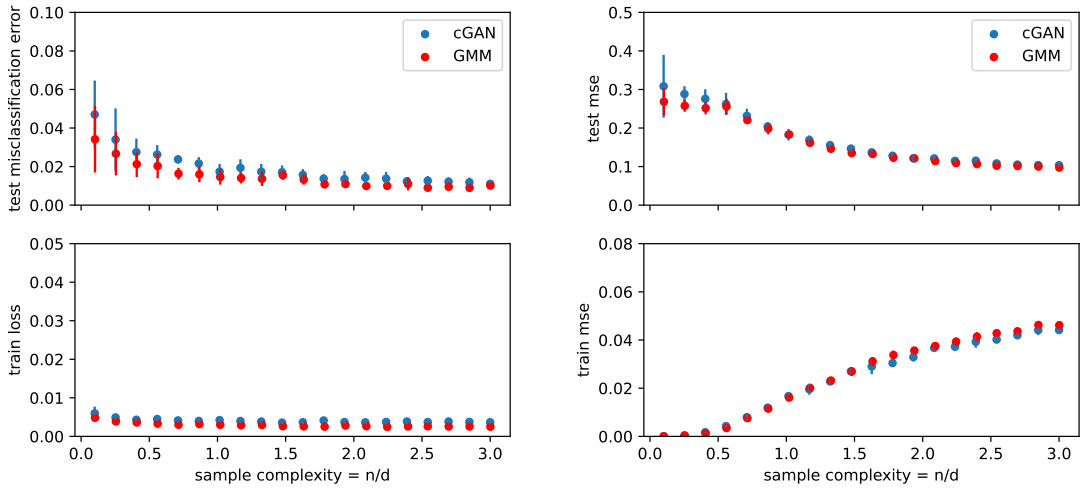

Figure 2: Illustration of the universality scenario described in Fig.1. Logistic (left) & ridge (right) regression test (up) and training (bottom) errors are shown versus the sample complexity $\alpha = n/d$ for an odd vs. even binary classification task on two data models: Blue dots data generated from a conditional GAN [Mirza and Osindero, 2014] trained on the fashion-MNIST dataset [Xiao et al., 2017] and pre-processed with a random features map $\boldsymbol{x} \mapsto \tanh(\mathrm{W}\boldsymbol{x})$ with Gaussian weights $W \in \mathbb{R}^{1176\times 784}$; Red dots are the 10- clusters Gaussian mixture model with means and covariances matching each fashion-MNIST cluster conditioned on labels (ℓ_2 regularization is $\lambda = 10^{-4}$). Details on the simulations are discussed in Appendix D.

Assumption 2 (Boundedness and concentration)**.** *The constraint set $\mathcal{S}_p$ is a compact subset of $\mathbb{R}^{k\times p}$. Further, there exists a constant $M > 0$ such that for any $c \geq 0$,*

$$\sup_{\boldsymbol{\theta}\in\mathcal{K}_p, \|\boldsymbol{\theta}\|_2\leq 1} \|\boldsymbol{\theta}^\top \boldsymbol{x}\|_{\psi_2} \leq M, \qquad \sup_{\boldsymbol{\theta}\in\mathcal{K}_p, \|\boldsymbol{\theta}\|_2\leq 1} \|\boldsymbol{\Sigma}_c^{1/2}\boldsymbol{\theta}\|_2 \leq M, \quad \text{and} \quad \|\boldsymbol{\mu}_c\|_2 \leq M \tag{10}$$

where $\|\cdot\|_{\psi_2}$ is the sub-gaussian norm, and $\mathcal{K}_p \subseteq \mathbb{R}^p$ is such that $\mathcal{S}_p \subseteq \mathcal{K}_p^k$.

Assumption 3 (Labels)**.** *The labeling function η is Lipschitz, the teacher vector $\boldsymbol{\Theta}$ belongs to $\mathcal{S}_p$, and the noise variables ε_i are i.i.d subgaussian with $\|\varepsilon_i\|_{\psi_2} \leq M$ for some constant $M > 0$.*

Those three assumptions are fairly technical, and it is possible that the universality properties proven in this article hold irrespective of these conditions. The crucial assumption in our theorems is that of a *conditional one-dimensional CLT*:

Assumption 4. *For any Lipschitz function $\varphi : \mathbb{R} \to \mathbb{R}$,*

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\theta}\in\mathcal{K}_p} \left| \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{x}) \,\middle|\, c_{\boldsymbol{x}} = c\right] - \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{g}) \,\middle|\, c_{\boldsymbol{g}} = c\right] \right| = 0, \quad \forall c \in \mathcal{C} \tag{11}$$

where $\boldsymbol{x}$ and $\boldsymbol{g}$ denote samples from the given mixture distribution and the equivalent gaussian mixture distribution in equations (1) and (8) respectively.

3.1 Universality of Mixture Models

We start by proving the universality of the free energy for a Gibbs distribution defined through the objective $\widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})$ for the data distribution defined in (2) and its equivalent Gaussian mixture distribution (8).

Theorem 1 (Universality of Free Energy)**.** *Let $\mu_p(\boldsymbol{\Theta})$ be a sequence of Borel probability measures with compact supports $\mathcal{S}_p$. Define the following free energy function:*

$$f_{\beta,n}(\boldsymbol{X}) = \int \exp\left(-\beta n \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X}))\right) d\mu_p(\boldsymbol{\Theta}) \tag{12}$$

Under Assumptions 1-4 on $\boldsymbol{X}$ and $\mathcal{S}_p$, for any bounded differentiable function Φ with bounded Lipschitz derivative, we have:

$$\lim_{n,p\to\infty} \left|\mathbb{E}\left[\Phi\left(f_{\beta,n}(\boldsymbol{X})\right)\right] - \mathbb{E}\left[\Phi\left(f_{\beta,n}(\boldsymbol{G})\right)\right]\right| = 0.$$

When μ_p corresponds to discrete measures supported on an ϵ-net in $\mathcal{S}_p$, using the reduction from Lemma 1 to Theorem 1 in Montanari and Saeed [2022], we obtain the following corollary:

Corollary 2. *[Universality of Training Error GMM] For any bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$:*

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

In particular, for any $\mathcal{E} \in \mathbb{R}$, and denoting $\xrightarrow{\mathbb{P}}$ the convergence in probability:

$$\widehat{\mathcal{R}}_n^\star(\boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X})) \xrightarrow{\mathbb{P}} \mathcal{E} \quad \textit{if and only if} \quad \widehat{\mathcal{R}}_n^\star(\boldsymbol{G}, \boldsymbol{y}(\boldsymbol{G})) \xrightarrow{\mathbb{P}} \mathcal{E}, \tag{13}$$

The full theorem, as well as its proof, is presented in Appendix A. In a nutshell, this theorem shows that the multi-modal data generated by any generative neural network is equivalent to a *finite* mixture of Gaussian in high-dimensions: in other words, a *finite* mixture of Gaussians leads to the same loss as for data generated by (for instance) a cGAN. Since the function $\ell : \mathbb{R}^{k+1} \to \mathbb{R}$ need not be convex, we can take

$$\ell(\boldsymbol{x}_{\text{out}}, y) = \ell'(\boldsymbol{\Psi}(\boldsymbol{x}_{\text{out}}), y),$$

where $\boldsymbol{\Psi}$ is an already pretrained neural network. In particular, if $\boldsymbol{x}$ is the output of all but the last layer of a neural net, we can view $\boldsymbol{\Psi}$ as the averaging procedure for a small committee machine.

Note that Corollary 2 depends crucially on assumption 4 (the one-dimensional CLT), which is by no means evident. We discuss the conditions on the weights matrix for which it can be proven in section 3.4. However, one can observe empirically that the validity of Corollary 2 goes well beyond what can be currently proven. A number of numerical illustrations of this property can be found in the work of Louart et al. [2018], Seddik et al. [2020, 2021], who already derived similar (albeit more limited) results using random matrix theory. Additionally, we observed that even with trained GANs, when we observed data through a random feature map [Rahimi and Recht, 2007], the Gaussian mixture universality is well obeyed. This scenario is illustrated in Fig.1, with a concrete example in Fig.2. Even though we did not prove the one-dimensional CLT for arbitrary learned matrices, and worked with finite moderate sizes, the realistic data generated by our cGAN behaves extremely closely with those of generated by the corresponding Gaussian mixture.

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

Proof sketch and remarks — While we provide a complete proof of these results in App. A, we believe it is useful to present a short intuitive presentation explaining why these results hold.

(i) A crucial first remark is that Theorem 1 and Corollary 2 do not require that the data generated by GANs are Gaussians mixtures: rather, it is their one-dimensional projections along the directions in $\mathcal{S}_p$ that should behave as such. The first, intuitive, explanation, is that indeed in high dimension, for a randomly chosen vector $\boldsymbol{\theta} \in \mathcal{S}_p$, it is natural to expect that $\boldsymbol{\theta}^\top \boldsymbol{x})$ behaves like a Gaussian mixture. Indeed, if we condition on a given label, then z is Gaussian, the random variable $\boldsymbol{x} = \Psi_{\text{nn}}(\mathbf{z})$, has a well defined mean and variance (at least if Ψ is a Lipschitz function), so that the central limit theorem shows that $\boldsymbol{\theta}^\top \boldsymbol{x})$ converges to a Gaussian variable.

(ii) We require, however, a slightly stronger condition in eq. (11): Indeed, it should be that, conditioned on a label, $\boldsymbol{\theta} \in \mathcal{S}_p$ is Gaussian *for all* $\boldsymbol{\theta} \in \mathcal{S}_p$, not a randomly chosen one (since we do certainly do not chose our weights randomly). This condition might appear strong. However, such one-dimensional CLTs have been the subject of many recent works which proved them for many cases [Goldt et al., 2022, Hu and Lu, 2022, Montanari and Saeed, 2022], including random features and two-layers neural tangent kernels. We extend the proof of the one-dimensional CLT to mixture models in section 3.4. We also provide further formal arguments in App. C. In particular, we argue that a large class of distributions, including deep generative models, do satisfy this condition that can also be checked empirically in simulations [Goldt et al., 2022, Seddik et al., 2021].

(iii) The one-dimensional CLT now implies that $\mathbb{E}\left[\widehat{\mathcal{R}}_n\left(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X})\right)\right] \simeq \mathbb{E}\left[\widehat{\mathcal{R}}_n\left(\boldsymbol{\Theta}; \boldsymbol{G}, \boldsymbol{y}(\boldsymbol{G})\right)\right]$ for any **fixed** choice of $\boldsymbol{\Theta}$, **independent from the data**. There is another, additional difficulty: when one performs empirical risk minimization, the minimizer $\hat{\boldsymbol{\Theta}}(\boldsymbol{X})$ **strongly depends** on the data $\boldsymbol{X}$, and the naive 1d-CLT simply does not apply! Solving the problem of the dependence of the estimator over the data is the main mathematical difficulty in proving Thm. 2. This is achieved in the Appendix by using a method due to Montanari and Saeed [2022], that leverage on the Guerra interpolation techniques used to prove the validity of the replica method Guerra [2003]. The idea is to define a t-dependent model that uses n dataset points at $t = 0$, and GMM ones at $t = 1$, and to show that the free energy (and all observables) remains constants at all "times" $t \in [0, 1]$). This establishes fully the universality advocated in our theorem. □

3.2 Convergence of expectations for Joint Minimization and Sampling

Our next result establishes a general relationship between the differentiability of the limit of expected training errors or free energies for empirical risk minimization or free energies for sampling and the universality of expectations of a given function of the parameters.

Setup Consider a sequence of M risks

$$\widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}) := \frac{1}{n} \sum_{i=1}^{n} \ell_m(\boldsymbol{\Theta}^\top \boldsymbol{x}_i^{(m)}, y_i^{(m)}) + r_m(\boldsymbol{\Theta}), \quad m \in [M] \tag{14}$$

with possibly different losses, regularizers and datasets. For simplicity, we assume that the objectives are defined on parameters having the same dimension $\boldsymbol{\Theta} \in \mathbb{R}^{p \times k}$. We aim to minimize M_1 of them:

$$\hat{\boldsymbol{\Theta}}^{(m)}(\boldsymbol{X}) \in \underset{\boldsymbol{\Theta} \in \mathcal{S}_p^{(m)}}{\arg\min}\ \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}), \quad m \in [M_1] \tag{15}$$

and the M_2 remaining parameters are independently sampled from a family of Gibbs distributions:

$$\boldsymbol{\Theta}^{(m)} \sim P_m(\boldsymbol{\Theta}) \propto \exp\left(-\beta_m \widehat{\mathcal{R}}_n^{(m)}\left(\boldsymbol{\Theta}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}\right)\right) d\mu_m(\boldsymbol{\Theta}), \quad m \in [M_1+1, M], \tag{16}$$

where $M = M_1 + M_2$. The joint distribution of the $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(1)}, \dots, \boldsymbol{x}_i^{(M)})$ is assumed to be a mixture of the form (1). However, we assume that the labels $\boldsymbol{y}_i^{(m)}$ only depend on the vectors $\boldsymbol{x}_i^{(m)}$:

$$y_i^{(m)}(\boldsymbol{X}^{(m)}) = \eta(\boldsymbol{\Theta}_\star^{(m)\top} \boldsymbol{x}_i^{(m)}, \varepsilon_i^{(m)}, c_i). \tag{17}$$

The equivalent Gaussian inputs $\boldsymbol{g}_i = (\boldsymbol{g}_i^{(1)}, \dots, \boldsymbol{g}_i^{(M)})$ and their associated labels $\boldsymbol{y}(\boldsymbol{G})$ are defined as in (8).

Statistical metric and free energy — Our goal is to study statistical metrics for some function $h : \mathbb{R}^{M\times k\times p} \to \mathbb{R}$ of the form $h(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)})$. For instance, the metric h could be the population risk (a.k.a. generalization error), or some overlap between $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_\star$. We define the following coupling free energy function:

$$f_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\log \int e^{-sn\, h(\boldsymbol{\Theta}^{(1)},\dots,\boldsymbol{\Theta}^{(M)})} dP^{(M_1+1):M}, \tag{18}$$

where $P^{(M_1+1):M}$ denotes the product measure of the P_m defined in (16). This gives rise to the following joint objective:

$$\widehat{\mathcal{R}}_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = \sum_{m=1}^{M_1} \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}^{(m)}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}) + f_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}). \tag{19}$$

In particular, when $s = 0$ we have $f_{n,0} = 0$ and the problem reduces to the joint minimization problem in (15). Our first result concerns the universality of the minimum of the above problem:

Proposition 3 (Universality for joint minimization and sampling)**.** *Under Assumptions 4 For any $s > 0$ and bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$, and denoting $\widehat{\mathcal{R}}^\star_{n,s}(\boldsymbol{X}, \boldsymbol{y}) \coloneqq \min \widehat{\mathcal{R}}_{n,s}(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})$:*

$$\lim_{n,p\to\infty} \left| \mathbb{E}\left[\Phi\left(\widehat{\mathcal{R}}^\star_{n,s}(\boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X}))\right)\right] - \mathbb{E}\left[\Phi\left(\widehat{\mathcal{R}}^\star_{n,s}(\boldsymbol{G}, \boldsymbol{y}(\boldsymbol{G}))\right)\right] \right| = 0$$

The proof is an easy reduction to Theorem 2, and can be found in Appendix A.5.

The next result concerns the value of h at the minimizers point $(\hat{\boldsymbol{\Theta}}^{(1)}, \dots, \hat{\boldsymbol{\Theta}}^{(M)})$. We make the following additional assumptions:

Assumption 5 (Differentiable Limit)**.** *There exists a neighborhood of 0 such that the function*

$$q_n(s) = \mathbb{E}\left[\widehat{\mathcal{R}}^\star_{n,s}(\boldsymbol{G}, \boldsymbol{y}(\boldsymbol{G}))\right] \tag{20}$$

converges pointwise to a function $q(s)$ that is differentiable at 0.

For a fixed realization of the dataset $\boldsymbol{X}$, we denote by $\left\langle h(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)})\right\rangle_{\boldsymbol{X}}$ the expected value of h when $(\hat{\boldsymbol{\Theta}}^{(1)}, \dots, \hat{\boldsymbol{\Theta}}^{(M_1)})$ are obtained through the minimization of (15) and $(\boldsymbol{\Theta}^{(M_1+1)}, \dots, \boldsymbol{\Theta}^{(M)})$ are sampled according to the Boltzmann distributions (16).

Assumption 6. *With high probability on $\boldsymbol{X}, \boldsymbol{G}$, the value $\left\langle h(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)})\right\rangle_{\boldsymbol{X}}$ (resp. the same for $\boldsymbol{G}$) is independent from the chosen minimizers in (15).*

In other words, the metric h respects the symmetries of the problem. Then the following holds:

Theorem 4. *Under assumptions 1-6, we have:*

$$\lim_{n,p\to\infty} \left| \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{X}}\right] - \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{G}}\right] \right| = 0, \tag{21}$$

Proof Sketch: Our proof relies on the observation that $q_n(s)$ is a concave function of s. We further have:

$$q_n'(0) = \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{G}}\right]. \tag{22}$$

Subsequently, we utilize a standard result from convex analysis relating the convergence of a sequence of convex or concave functions to the convergence of the corresponding derivatives. The above result shows that the expected value of $h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)$ for a multi-modal data satisfying the 1d CLT is equivalent to that of a mixture of Gaussians. It generalizes the universality of generalization error in Hu and Lu [2022], Montanari and Saeed [2022]to arbitrary function of parameters arising from an arbitrary number of Hamiltonians for mixture models. The full theorem is presented and proven in App. A. While we prove our result for convergence in expectations and assumption 5, the result can be generalized using standard techniques to convergence in probability and other related assumptions. We refer to the alternative assumptions in Theorem 3 of Montanari and Saeed [2022] for more details.

3.3 Universal Weak Convergence

Theorem 4 provides a general framework for proving the equivalence of arbitrary functions of parameters obtained by minimization/sampling on a given mixture dataset and the equivalent gaussian mixture distribution. However, it relies on the assumption of a differentiable limit of the free energy (assumption 5). If the assumption holds for a sequences of functions belonging to dense subsets of particular classes of functions, it allows us to prove convergence of minimizers themselves, in a weak sense. We illustrate this through a simple setup considered in Loureiro et al. [2021b], which precisely characterized the asymptotic distribution of the minimizers of empirical risk with GMM data in the strictly convex case. Consider the following setup:

$$\left(\hat{\boldsymbol{W}}^{\boldsymbol{X}}, \hat{\boldsymbol{b}}^{\boldsymbol{X}}\right) = \underset{\boldsymbol{W},\boldsymbol{b}}{\arg\min} \sum_{i=1}^{n} \ell\left(\frac{\boldsymbol{W}\boldsymbol{x}_i}{\sqrt{d}} + \boldsymbol{b}, \boldsymbol{y}_i\right) + \lambda r(\boldsymbol{W}), \tag{23}$$

where $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{C}|\times d}$, $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{C}|}$ and $\boldsymbol{y}_i \in \mathbb{R}^{|\mathcal{C}|}$ is the one-hot encoding of the class index c_i. For simplicity, we restrict ourselves to diagonal bounded covariances and bounded means.

Assumption 7. *All of the covariance matrices Σ_c are diagonal, with strictly positive eigenvalues $(\sigma_{c,i})_{i\in[d]}$, and there exists a constant $M > 0$ such that for any $c \in \mathcal{C}$*

$$\sigma_{c,i} \leq M \quad \textit{and} \quad \|\boldsymbol{\mu}_c\|_2 \leq M. \tag{24}$$

Secondly, since we aim at obtaining a result on the weak convergence of the estimators, we assume the same weak convergence for the means and covariances, and that the regularization only depends on the empirical measure of $\boldsymbol{W}$.

Assumption 8. *The empirical distribution of the $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ converges weakly as follows:*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{c\in\mathcal{C}}\delta(\mu_c - \sqrt{d}\mu_{c,i})\delta(\sigma_c - \sigma_{c,i}) \quad \xrightarrow[d\to\infty]{\mathcal{L}} \quad p(\boldsymbol{\sigma}, \boldsymbol{\mu}) \tag{25}$$

Assumption 9. *The regularizer $r(\cdot)$ is of the following form:*

$$r(\boldsymbol{W}) = \sum_{i=1}^{d} \psi_r(\boldsymbol{W}_i), \tag{26}$$

for some convex and twice differentiable function $\psi_r : \mathbb{R} \to \mathbb{R}$.

Under these conditions, the joint measure of the minimizers and of the data moments converges weakly to a fixed limit, independent of the data-distribution:

Theorem 5. *Assume that conditions 1-9 hold, and further that the function $\ell(\bullet, y) + r(\bullet)$ is convex and coercive. Then, for any bounded-Lipschitz function: $\Phi : \mathbb{R}^{3|\mathcal{C}|} \to \mathbb{R}$, we have:*

$$\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\Phi(\{(\hat{\boldsymbol{W}}^{\boldsymbol{X}})_{c,i}\}_{c\in\mathcal{C}}, \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}})\right] \xrightarrow{d\to+\infty} \mathbb{E}_{\tilde{p}}\left[\Phi(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma})\right], \tag{27}$$

where $\tilde{p}$ is a measure on $\mathbb{R}^{3|\mathcal{C}|}$, that is determined by the so-called replica equations*.*

Proof Sketch Let $\boldsymbol{\Theta}$ denote the combined set of parameters $\{\boldsymbol{W}, \boldsymbol{b}\}$. We first show that for $h(\boldsymbol{\Theta})$ having bounded second derivatives, the perturbation term $sh(\boldsymbol{\Theta})$ can be absorbed into the regularizer, while maintaining the assumptions in Loureiro et al. [2021b]. Next, we show that the sequence of solutions converge in a suitable sense, and are described through the limits of the replica equations from Loureiro et al. [2021b]. This allows us to prove that assumption 5 holds for functions that can be expressed as expectations w.r.t the joint empirical measure in 27. More details can be found in Appendix A.8.

3.4 One-dimensional CLT for Random Features

We finally show a conditional one-dimensional CLT for a random features map applied to a mixture of gaussians, in the vein of those shown in Goldt et al. [2022], Hu and Lu [2022], Montanari and Saeed [2022]. Concretely, we consider the following setup:

$$\boldsymbol{x}_i = \sigma(\boldsymbol{F}^\top \boldsymbol{z}_i), \quad \boldsymbol{z}_i \sim \sum_{c\in\mathcal{C}} \mathcal{N}(\boldsymbol{\mu}_c^{\boldsymbol{z}}, \boldsymbol{\Sigma}_c^{\boldsymbol{z}}), \tag{28}$$

where the feature matrix $\boldsymbol{F} \in \mathbb{R}^{d\times p}$ has i.i.d $\mathcal{N}(0, 1/d)$ entries. This setup is much more permissive than the ones in Hu and Lu [2022], Montanari and Saeed [2022], that restrict the samples $\boldsymbol{z}$ to standard normal vectors. However, we do require some technical assumptions:

Assumption 10. *The activation function σ is thrice differentiable, with $\|\sigma^{(i)}\| \leq M$ for some $M > 0$, and we have*

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\sigma(g)\right] = 0. \tag{29}$$

The cluster means and covariances of $\boldsymbol{z}$ satisfy for all $c \in \mathcal{C}$

$$\|\boldsymbol{\mu}_c^{\boldsymbol{z}}\| \leq M, \qquad \|\boldsymbol{\Sigma}_c^{\boldsymbol{z}}\|_{\text{op}} \leq M \tag{30}$$

for some constant $M > 0$.

We also place ourselves in the proportional regime, i.e. a regime where $p/d \in [\gamma^{-1}, \gamma]$ for some $\gamma > 0$. For simplicity, we will consider the case $k = 1$; and the constraint set $\mathcal{S}_p$ as follows:

$$\mathcal{S}_p = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \|\boldsymbol{\theta}\|_2 \leq R, \quad \|\boldsymbol{\theta}\|_\infty \leq Cp^{-\eta}\right\} \tag{31}$$

for a given $\eta > 0$. We show in the appendix the following theorem:

Theorem 6. *Under Assumption 10, and with high probability on the feature matrix $\boldsymbol{F}$, the data $\boldsymbol{X}$ satisfy the concentration assumption 2, as well as the one-dimensional CLT of Assumption 4. Consequently, the results of Theorems 1 and 4 apply to $\boldsymbol{X}$ and their Gaussian equivalent $\boldsymbol{G}$.*

Proof Sketch Our proof constructs a reduction to the one-dimensional CLT for random features in Goldt et al. [2022], Hu and Lu [2022], Montanari and Saeed [2022]. We first note that the one-dimensional CLT in Goldt et al. [2022], Hu and Lu [2022], Montanari and Saeed [2022] holds even when the activation functions differ across neurons. Subsequently, we proceed by defining the following neuron-wise activation functions:

$$\sigma_{i,c}(u) = \sigma(u + \boldsymbol{f}_i^\top \boldsymbol{\mu}_c). \tag{32}$$

However, the above activation functions depend on the means and the dimensions of the inputs. Our proof involves controlling the effect of this dependence. Additionally, we handle the effect of the covariance by observing that the features $\boldsymbol{x}$ can be equivalently generated by applying the modified weights $\boldsymbol{\Sigma}_c^{\boldsymbol{z}1/2}\boldsymbol{F}$ to isotropic Gaussian noise. Further details can be found in Appendix A.7.

While we prove the above result for random weights, we note, however that the non-asymptotic results in Goldt et al. [2022], Hu and Lu [2022] also hold for deterministic matrices satisfying approximate orthogonality conditions. Therefore, we expect the one-dimensional CLT to approximately hold for a much larger class of feature maps. Finally, we also note that the above extension of the one-dimensional CLT to mixture of gaussians also provides a proof for the asymptotic error for random features in Refinetti et al. [2021].

4 Conclusion

We demonstrate the universality of the Gaussian mixture assumption in high-dimension for various machine learning tasks such as empirical risk minimization, sampling and ensembling, in a variety of settings including random features or GAN generated data. We also show that universality holds for a large class of functions, and provide a weak convergence theorem. These results, we believe, vindicate the classical theoretical line of works on the Gaussian mixture design. We hope that our results will stimulate further research in this area.

Acknowledgements

We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe, the Swiss National Science Foundation grant SNFS OperaGOST, 200021_200390 and the *Choose France - CNRS AI Rising Talents* program.

References

Alia Abbara, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. On the universality of noiseless linear estimation with respect to the measurement matrix. *Journal of Physics A: Mathematical and Theoretical*, 53(16):164001, mar 2020. doi: 10.1088/1751-8121/ab59ef.

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.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. 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.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Joern-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1792–1800. PMLR, 13–15 Apr 2021.

S. G. Bobkov. On concentration of distributions of random weighted sums. *The Annals of Probability*, 31(1): 195–215, January 2003. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1046294309.

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.

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

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv: 1606.03657*, 2016. doi: 10.48550/ARXIV.1606.03657.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification. *arXiv: 2202.03295*, 2022a. doi: 10.48550/ARXIV.2202.03295.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. A study of uncertainty quantification in overparametrized high-dimensional models. *arXiv: 2210.12760*, 2022b. doi: 10.48550/ARXIV.2210.12760.

Oussama Dhifallah and Yue M. Lu. A precise performance analysis of learning with random features. *arXiv: 2008.11904*, 2020. doi: 10.48550/ARXIV.2008.11904.

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.

Rishabh Dudeja, Subhabrata Sen, and Yue M. Lu. Spectral universality of regularized linear regression with nearly deterministic sensing matrices. *arXiv: 2208.02753*, 2022. doi: 10.48550/ARXIV.2208.02753.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

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

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017. doi: 10.1038/s41598-017-11873-y.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, feb 2020. doi: 10.1088/1742-5468/ab633c.

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

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of linear classifiers with random labels in high-dimension. *arXiv: 2205.13303*, 2022. doi: 10.48550/ARXIV.2205.13303.

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

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

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

Francesco Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, pages 1–1, 2022. doi: 10.1109/TIT.2022.3217698.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv: 1312.6114*, 2013. doi: 10.48550/ARXIV.1312.6114.

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

Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE Transactions on Information Theory*, 57(4):2440–2450, 2011. doi: 10.1109/TIT.2011.2112231.

Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. ISSN 1547-5905. doi: 10.1002/aic.690370209.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 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.

Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Math. Surv. Monogr.* American Mathematical Society (AMS), Providence, RI, 2001. ISBN 9780821828649.

J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, December 1922. ISSN 1432-1823. doi: 10.1007/BF01494395.

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

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

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

Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. 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 14283–14314. PMLR, 17–23 Jul 2022.

Yue M. Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv: 2205.06308*, 2022. doi: 10.48550/ARXIV.2205.06308.

Xiaoyi Mai and Zhenyu Liao. High dimensional classification via empirical risk minimization: Improvements and optimality. *arXiv: 1905.13742*, 2019.

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

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv: 1411.1784*, 2014. doi: 10.48550/ARXIV.1411.1784.

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv: 2204.10425*, 2022. doi: 10.48550/ARXIV.2204.10425.

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

Hatef Monajemi, Sina Jafarpour, Matan Gavish, null null, David L. Donoho, Sivaram Ambikasaran, Sergio Bacallado, Dinesh Bharadia, Yuxin Chen, Young Choi, Mainak Chowdhury, Soham Chowdhury, Anil Damle, Will Fithian, Georges Goetz, Logan Grosenick, Sam Gross, Gage Hills, Michael Hornstein, Milinda Lakkam, Jason Lee, Jian Li, Linxi Liu, Carlos Sing-Long, Mike Marx, Akshay Mittal, Hatef Monajemi, Albert No, Reza Omrani, Leonid Pekelis, Junjie Qin, Kevin Raines, Ernest Ryu, Andrew Saxe, Dai Shi, Keith Siilats, David Strauss, Gary Tang, Chaojun Wang, Zoey Zhou, and Zhen Zhu. Deterministic matrices matching the compressed sensing phase transitions of gaussian random matrices. *Proceedings of the National Academy of Sciences*, 110(4):1181–1186, 2013. doi: 10.1073/pnas.1219540110.

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. doi: 10.1109/ISIT.2017.8006947.

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.

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

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.

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

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901. ISSN 1941-5982. doi: 10.1080/14786440109462720.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pages 8936–8947. PMLR, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. *arXiv: 2302.00401*, 2023. doi: 10.48550/ARXIV.2302.00401.

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

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

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, dec 2020. doi: 10.1088/1742-5468/abc61d.

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

Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127 – 204, 2011. doi: 10.1007/s11511-011-0061-3.

Terence Tao and Van Vu. Random matrices: universal properties of eigenvectors. *Random Matrices: Theory and Applications*, 01(01):1150001, 2012. doi: 10.1142/S2010326311500018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv: 1708.07747*, 2017. doi: 10.48550/ARXIV.1708.07747.

A Proofs of Main Results

A.1 Notation

We follow the setting defined in section 2. Throughout, we work in the so-called proportional high-dimensional limit, where n, p go to infinity with

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

while $\mathcal{C}$ stays fixed.

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

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

A.2 State of the art

There have been many recent progress on Gaussian-type low-dimensional CLT and universality recently Goldt et al. [2022], Hu and Lu [2022], Montanari and Saeed [2022]. We shall leverage on these results to prove our first theorem.

In particular, the starting point for our mathematical proof will use the recent result of Montanari and Saeed [2022] which we shall now review. Consider the minimization problem (4), with $(\boldsymbol{x}_\mu, y_\mu)$ i.i.d random variables; the goal is to replace the $\boldsymbol{x}_\mu$ by their Gaussian equivalent model

$$\boldsymbol{g}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{where} \qquad \boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}\right], \quad \boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]. \tag{33}$$

Montanari and Saeed [2022] make the following assumptions:

Assumption A1 (Loss and regularization)**.** *The loss function $\ell : \mathbb{R}^{k+1} \to \mathbb{R}$ is nonnegative and Lipschitz, and the regularization function $r : \mathbb{R}^{p\times k} \to \mathbb{R}$ is locally Lipschitz, with constants independent from p.*

Assumption A2 (Concentration on the directions of $\mathcal{S}_p$)**.** *We have*

$$\sup_{\boldsymbol{\theta}\in\mathcal{S}_p, \|\boldsymbol{\theta}\|_2\leq 1} \|\boldsymbol{\theta}^\top \boldsymbol{x}\|_{\psi_2} \leq M, \qquad \sup_{\boldsymbol{\theta}\in\mathcal{S}_p, \|\boldsymbol{\theta}\|_2\leq 1} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2 \leq M, \quad \textit{and} \quad \|\boldsymbol{\mu}\|_2 \leq M \tag{34}$$

for some constant $M > 0$.

Assumption A3 (One-dimensional CLT)**.** *For any bounded Lipschitz function $\varphi : \mathbb{R}^k \to \mathbb{R}$,*

$$\lim_{p\to\infty} \sup_{\boldsymbol{\theta}\in\mathcal{S}_p} \left| \mathbb{E}\left[\phi(\boldsymbol{\theta}^\top \boldsymbol{x})\right] - \mathbb{E}\left[\phi(\boldsymbol{\theta}^\top \boldsymbol{g})\right] \right| = 0. \tag{35}$$

Assumption A4 (Labels)**.** *The y_μ are generated according to*

$$y_i = \eta(\boldsymbol{\Theta}^* \boldsymbol{x}_i, \varepsilon_i, c_i), \tag{36}$$

where $\eta : \mathbb{R}^{k^*+1} \to \mathbb{R}$ is a Lipschitz function, $\boldsymbol{\Theta}^* \in \mathcal{S}_p^{k^*}$, and the ε_μ are i.i.d subgaussian random variables with

$$\|\varepsilon_i\|_{\psi_2} \leq M$$

for some constant $M > 0$.

Building on those assumptions, Montanari and Saeed [2022] prove the following:

Theorem 7 (Theorem 1. in Montanari and Saeed [2022]). *Suppose that Assumptions A1-A3 hold. Then, for any bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$, we have*

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

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

$$\widehat{\mathcal{R}}_n^\star(\boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X})) \xrightarrow{\mathbb{P}} \rho \quad \textit{if and only if} \quad \widehat{\mathcal{R}}_n^\star(\boldsymbol{G}, \boldsymbol{y}(\boldsymbol{G})) \xrightarrow{\mathbb{P}} \rho$$

Free energy approximation A crucial component of the proof in Montanari and Saeed [2022] is the approximation of the minimizer through a free energy function. Define the discretized free energy

$$f_{\epsilon,\beta}(\boldsymbol{X}) = \frac{1}{n\beta} \sum_{\boldsymbol{\Theta}\in\mathcal{N}_\epsilon^k} \exp\left(-\beta\, \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X}))\right), \tag{37}$$

where $\mathcal{N}_\epsilon$ is a minimal ϵ-net of $\mathcal{S}_p$.

Lemma 8 (Lemma 1 in Montanari and Saeed [2022]). *For any bounded differentiable function Φ with bounded Lipschitz derivative, and any $\epsilon > 0$ we have:*

$$\lim_{n,p\to\infty} \left|\mathbb{E}\left[\Phi\left(f_{\epsilon,\beta}(\boldsymbol{X})\right)\right] - \mathbb{E}\left[\Phi\left(f_{\epsilon,\beta}(\boldsymbol{G})\right)\right]\right| = 0.$$

Subsequently, using classical arguments from both the theory of ϵ-nets and statistical physics, the authors show that

$$\left| f_{\epsilon,\beta}(\boldsymbol{X}) - \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X}) \right| \le C_1(\epsilon) + \frac{C_2(\epsilon)}{\beta}, \tag{38}$$

and the same inequality holds for $\boldsymbol{G}$. Since C_1, C_2 do not depend on n, p, it is possible to choose first ϵ, then β so that the RHS of (38) is as small as desired.

Montanari and Saeed [2022] therefore used the universality of the free energy as an intermediate step wards proving the universality of the training error. We generalize this result to the free energy defined for a general class of Boltzmann distributions, allowing us to prove universality in applications related to sampling.

A.3 Sketch of proof of Lemma 8, adapted from Montanari and Saeed [2022]

Interpolation path For any $0 \le t \le \pi/2$, define

$$\boldsymbol{U}_t = \cos(t)\boldsymbol{X} + \sin(t)\boldsymbol{G}$$

Then $\boldsymbol{U}_t$ is a smooth interpolation path with independent columns, ranging from $\boldsymbol{U}_0 = \boldsymbol{X}$ to $\boldsymbol{U}_{\pi/2} = \boldsymbol{G}$. We can write, for any differentiable function ψ,

$$\left|\mathbb{E}\left[\psi(f_{\epsilon,\beta}(\boldsymbol{X}))\right] - \mathbb{E}\left[\psi(f_{\epsilon,\beta}(\boldsymbol{G}))\right]\right| \le \int_0^{\pi/2} \left| \mathbb{E}\left[\frac{d\psi(f_{\epsilon,\beta}(\boldsymbol{U}_t)))}{dt}\right]\right| dt,$$

and by the dominated convergence theorem it suffices to show that the integrand converges to 0 for any t. The chain rule gives:

$$\frac{d\psi(f_{\epsilon,\beta}(\boldsymbol{U}_t)))}{dt} = \psi'(f_{\epsilon,\beta}(\boldsymbol{U}_t))) \left(\sum_{\mu=1}^{n} \left(\frac{d\boldsymbol{u}_{t,\mu}}{dt} \right)^\top \nabla_{\boldsymbol{u}_{t,\mu}} f_{\epsilon,\beta}(\boldsymbol{U}_t) \right), \tag{39}$$

and the dependency in ψ can be easily controlled. Since all columns of $\boldsymbol{U}_t$ are i.i.d, we are left with showing

$$\lim_{n,p\to\infty} n\mathbb{E}_{(1)} \left[\left(\frac{d\boldsymbol{u}_{t,1}}{dt} \right)^\top \nabla_{\boldsymbol{u}_{t,1}} f_{\epsilon,\beta}(\boldsymbol{U}_t) \right] = 0 \quad \text{a.s.}, \tag{40}$$

where $\mathbb{E}_{(1)}$ denotes the expectation with respect to $(\boldsymbol{x}_1, \boldsymbol{g}_1, \varepsilon_1)$.

Showing (40) Imagine for a moment that $\boldsymbol{x}_1$ is Gaussian; then $\boldsymbol{u}_{t,1}$ and $d\boldsymbol{u}_{t,1}/dt$ are also jointly Gaussian, and we have

$$\begin{aligned} \mathbb{E}\left[\left(\frac{d\boldsymbol{u}_{t,1}}{dt} \right)^\top \boldsymbol{u}_{t,1} \right] &= \mathbb{E}\left[(-\sin(t)\boldsymbol{x}_1 + \cos(t)\boldsymbol{g}_1)^\top (\cos(t)\boldsymbol{x}_1 + \sin(t)\boldsymbol{g}_1) \right] \\ &= 0, \end{aligned}$$

since x_1 and g_1 have the same covariance by definition. Therefore, they are independent, and we have

$$\mathbb{E}_{(1)} \left[\left(\frac{d\boldsymbol{u}_{t,1}}{dt} \right)^\top \nabla_{\boldsymbol{u}_{t,1}} f_{\epsilon,\beta}(\boldsymbol{U}_t) \right] = \mathbb{E}_{(1)} \left[\left(\frac{d\boldsymbol{u}_{t,1}}{dt} \right) \right]^\top \mathbb{E}_{(1)} \left[\nabla_{\boldsymbol{u}_{t,1}} f_{\epsilon,\beta}(\boldsymbol{U}_t) \right] = 0.$$

On the other hand, it is possible to show that $\boldsymbol{x}_1$ only appears in (40) through scalar products with $\boldsymbol{\Theta}$ or $\boldsymbol{\Theta}^*$. As a result, we can leverage Assumption 4 to replace $\boldsymbol{x}_1$ by a Gaussian vector $\boldsymbol{w}$ independent from $\boldsymbol{g}_1$ as $p \to \infty$. Then, the reasoning above can be repeated with $\boldsymbol{w}$ and $\boldsymbol{g}_1$ to conclude the proof.

A.4 Proof of Theorem 2

In order to prove our theorem 2, we now aim to adapt the proof from Montanari and Saeed [2022] to the following case where the distribution of x can be a *mixture* of several other distributions, each with different mean and covariance. For a discrete set $\mathcal{C}$, we consider a family of distributions $(\nu_c)_{c\in\mathcal{C}}$ on $\mathbb{R}^p$, with means and covariances

$$\boldsymbol{\mu}_c = \mathbb{E}_{\mathbf{z}\sim\nu_c}[\mathbf{z}] \qquad \text{and} \qquad \boldsymbol{\Sigma}_c = \mathbb{E}_{\mathbf{z}\sim\nu_c}[\mathbf{z}\mathbf{z}^\top]$$

Given a type assignment $\sigma : [n] \to \mathcal{C}$, each sample x_i is then drawn independently from $\nu_{\sigma(i)}$. The equivalent Gaussian model is straightforward: we simply take

$$\boldsymbol{g}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\sigma(i)}, \boldsymbol{\Sigma}_{\sigma(i)}),$$

independently from each other. An important special case of this setting is when σ is itself random, independently from the $\boldsymbol{x}_i$ and $\boldsymbol{g}_i$: the law of $\boldsymbol{g}_i$ is then a so-called Gaussian Mixture Model. Note that in the Gaussian mixture setting, the existence of the labeling function σ implies that we coupled the labels for $\boldsymbol{X}$ and $\boldsymbol{G}$.

The main differences between our Assumptions 1-4 and Assumptions A1-A3 are the following:

(i) Assumption 1 is unchanged.

(ii) We relax (36) in Assumption 3 into

$$y_i = \eta_{\sigma(i)}(\boldsymbol{\Theta}^* \boldsymbol{x}_i, \varepsilon_i, c_i),$$

when η a Lipschitz function in its first two parameters. This allows in particular to incorporate classification problems in our setting, at no cost in the proof complexity.

(iii) We assume a more general setup where the constraint set $\mathcal{S}_p$ is not necessarily a product set. This slight generalization will be useful while proving a reduction to multiple objectives in Theorem 4.

(iv) We suppose that Assumptions 2 and 4 hold for any possible distribution ν_c for $c \in \mathcal{C}$ and its associated Gaussian equivalent model.

(v) We allow the reference measures to be any sequence of Borel measures with support on $\mathcal{S}_p$, instead of only the Dirac measure on the ϵ-net $\mathcal{N}_\epsilon$.

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

Free energy approximation This section goes basically unchanged; the approximation between $\widehat{\mathcal{R}}_n^*(\boldsymbol{X}, \boldsymbol{y}(\boldsymbol{X}))$ and $f_{\epsilon,\beta}(\boldsymbol{X})$ relies on Lipschitz arguments and concentration bounds on the $\boldsymbol{x}_i$ and $\boldsymbol{g}_i$, which are satisfied by our modification of Assumption 2.

General Reference Measures Our proof for the universality of free energy in Theorem 2 is a generalization of the proof of Lemma 1 in Montanari and Saeed [2022]. Compactness of the supports ensures that the corresponding free energy:

$$f_{\beta,n}(\boldsymbol{Z}) = \int \exp\left(-\beta n R_n(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{y})\right) d\mu(\boldsymbol{\Theta}), \tag{41}$$

is finite for any $\boldsymbol{Z}$.

The corresponding Boltzmann measures can then be defined by setting the Radon-Nikodym derivative/density to be:

$$\frac{d\tilde{\mu}}{d\mu} = \exp\left(-\beta n \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})\right). \tag{42}$$

The measure $\tilde{\mu}$ is then a Borel measure with support lying in $\mathcal{S}_p$. Therefore, through dominated convergence theorem, we can interchange differentiation and expectations w.r.t μ in the proof of Lemma 1 in Montanari and Saeed [2022]. For instance, equation (39) can be expressed as:

$$\mathbb{E}\left[\frac{\partial}{\partial t}\psi(f_{\beta,n}(\boldsymbol{U}_t))\right] = \mathbb{E}\left[\frac{\psi'(f_{\beta,n}(\boldsymbol{U}_t))}{n}\sum_{i=1}^{n}\frac{\int \widetilde{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{d}}_{t,i}(\boldsymbol{\Theta}) e^{-n\beta R_n(\boldsymbol{\Theta};\boldsymbol{Z},\boldsymbol{y})} d\mu_p}{\int e^{-n\beta R_n(\boldsymbol{\Theta};\boldsymbol{Z},\boldsymbol{y})} d\mu_p}\right]. \tag{43}$$

Similarly we substitute $\sum_{\boldsymbol{\Theta}}$ by $\int d\mu$ in the remaining arguments in the proof of Lemma 1 in Montanari and Saeed [2022].

Interpolation path Recall that the important property of $\boldsymbol{U}_t$ is that

$$\mathbb{E}\left[\left(\frac{d\boldsymbol{U}_t}{dt}\right)^\top \boldsymbol{U}_t\right] = 0. \tag{44}$$

To this end, we set

$$\boldsymbol{u}_{t,i} = \boldsymbol{\mu}_{\sigma(i)} + \cos(t)(\boldsymbol{x}_i - \boldsymbol{\mu}_{\sigma(i)}) + \sin(t)(\boldsymbol{g}_i - \boldsymbol{\mu}_{\sigma(i)}),$$

and it is easy to check that (44) is satisfied. Another problem is that the columns of $\boldsymbol{U}_t$ are not i.i.d anymore, so we have to control

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{(i)}\left[\left(\frac{d\boldsymbol{u}_{t,i}}{dt}\right)^\top \nabla_{\boldsymbol{u}_{t,i}} f_{\epsilon,\beta}(\boldsymbol{U}_t)\right]\right|, \tag{45}$$

where this time $\mathbb{E}_{(i)}$ is the expectation w.r.t $(\boldsymbol{x}_i, \boldsymbol{g}_i, \varepsilon_i)$. However, (45) is a weighted average over all values of $\sigma(\mu)$, and since $\mathcal{C}$ is finite is suffices to show (40) for any value of $\sigma(1)$.

Showing (40) This section again relies on concentration properties of the $\boldsymbol{x}_i$ and $\boldsymbol{g}_i$, as well as Assumption 4. The arguments thus translate directly from Montanari and Saeed [2022].

A.5 Proof of Proposition 3

We define the following free energy of the system:

$$f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\log \int e^{-n\sum_{m=1}^{M}\beta_m \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}^{(m)};\boldsymbol{X}^{(m)},\boldsymbol{y}^{(m)}) - sn\, h(\boldsymbol{\Theta}^{(1)},\dots,\boldsymbol{\Theta}^{(M)})} d\mu_\epsilon^{1:M_1} d\mu^{(M_1+1):M}, \tag{46}$$

where the μ_m are the reference measures for the Boltzmann distributions in (16), and μ_ϵ^m is the uniform measure supported on a minimal ϵ-net of $\mathcal{S}_p^{(i)}$:

$$\mu_\epsilon^m = \frac{1}{|\mathcal{N}_\epsilon^m|}\sum_{\boldsymbol{\Theta}\in\mathcal{N}_\epsilon^m}\delta_{\boldsymbol{\Theta}}, \tag{47}$$

where $\delta_{\boldsymbol{\Theta}}$ denotes the Dirac measure at $\boldsymbol{\Theta}$. We establish the following result:

Lemma 9 (Universality of the joint free energy)**.** *Under Assumptions 1-4, for any fixed $\epsilon > 0$ and any bounded differentiable function ψ with bounded Lipschitz derivative we have.*

$$\lim_{n\to\infty}\left|\mathbb{E}\left[\psi(f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}))\right] - \mathbb{E}\left[\psi(f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}))\right]\right| = 0.$$

Proof. We construct a reduction from the free energy of the form in equation 46 to the universality of free energy for a single objective in Theorem 1.

We construct an equivalent objective on the $pM \times kM$ dimensional space. Consider the following mapping:

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{\Theta}^{(2)} & \mathbf{0} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Theta}^{(M)} \end{pmatrix} \tag{48}$$

. Let $\mathcal{S}_{pM}$ be the set obtained by applying the mapping in equation 48 to $\mathcal{S}_p^{(1)}, \cdots, \mathcal{S}_p^{(M)}$. Let $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(M)})$ denote the combined input matrix with each row of dimension pM. We note that $\mathcal{S}_{pM}$ is a product of kM compact sets, each satisfying the assumption 2. Similarly, we have the combined output vector:

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}^{(1)} \\ \vdots \\ \boldsymbol{y}^{(M)} \end{pmatrix} \tag{49}$$

We define $\ell : \mathbb{R}^{k'M} \times \mathbb{R}^{kM} \to \mathbb{R}$ by:

$$\ell(\boldsymbol{u}, \mathbf{y}) = \sum_{m=1}^{M} \beta_m \ell_m(\boldsymbol{u}[(k-1)m : km], \mathbf{y}[(k-1)m : km]). \tag{50}$$

Similarly, we define the total regularization as:

$$\begin{aligned} r(\boldsymbol{\Theta}) = &\sum_{m=1}^{M} \beta_m r_m(\boldsymbol{\Theta}[(m-1)p : mp, (m-1) : k, mk]) \\ &+ sh(\boldsymbol{\Theta}[0:p, 0:k], \cdots, \boldsymbol{\Theta}[(M-1)p : Mp, (M-1)k : Mk]). \end{aligned} \tag{51}$$

Let $\widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})$ denote the following objective on the combined vector $\boldsymbol{\Theta}$:

$$\widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\Theta}^\top \boldsymbol{x}_i, \mathbf{y}_i) + r(\boldsymbol{\Theta}) \tag{52}$$

Then, using all the definitions above, we have

$$\widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{y}) = \sum_{m=1}^{M} \beta_m \, \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}^{(m)}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}) + s \, h(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}) \tag{53}$$

Therefore, we obtain that:

$$\begin{aligned} f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}) &= -\frac{1}{n} \log \int e^{-n \sum_{m=1}^{M} \beta_m \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}^{(m)}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}) - sn \, h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})} d\mu_{\epsilon}^{1:M_1} d\mu^{(M_1+1):M} \\ &= -\frac{1}{n} \log \int e^{-n \widehat{\mathcal{R}}_n(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{y})} d\mu_{\epsilon}^{1:M_1} d\mu^{(M_1+1):M}. \end{aligned} \tag{54}$$

We further note that the constraint sets on $\boldsymbol{\Theta}$, and the joint means, covariances on $\boldsymbol{X}$ satisfy the assumptions A2. Therefore, using Theorem 1, we obtain that:

$$\lim_{n \to \infty} |\mathbb{E}\left[\psi(f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}))\right] - \mathbb{E}\left[\psi(f_{n,s,\epsilon}(\boldsymbol{X}, \boldsymbol{y}))\right]| = 0$$

□

The proof of Proposition 3 then follows using the ϵ-net approximation in (38) for $\boldsymbol{\Theta}[1 : M_1]$.

A.6 Proof of Theorem 4

Our proof relies on the following result:

Lemma 10. *For any $n \in \mathbb{N}$, $q_n(s)$ is a concave function of s, that is differentiable at 0, and*

$$q_n'(0) = \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{G}}\right]. \tag{55}$$

Proof. Consider the joint free energy in Equation (18):

$$f_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\log \int e^{-sn\ h(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(M)})} dP^{(M_1+1):M}, \tag{56}$$

Let $\langle\cdot\rangle_{M_1+1:M}$ denote the expectation w.r.t the product measure

$$d\tilde{\mu}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = e^{-sn\ h(\boldsymbol{\Theta}^{(1)},\ldots,\boldsymbol{\Theta}^{(M)})} dP^{(M_1+1):M}.$$

We observe that:

$$\frac{df_{n,s}}{ds}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = \langle h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})\rangle_{M_1+1:M}. \tag{57}$$

Differentiating w.r.t s again, and using the dominated convergence theorem, we obtain:

$$-\frac{1}{n}\frac{d^2 f_{n,s}}{ds^2}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = \langle h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})^2\rangle_{M_1+1:M} - \langle h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})\rangle^2_{M_1+1:M}. \tag{58}$$

Since the R.H.S equals the variance of the variable $h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})$ w.r.t $\tilde{\mu}^{(M_1+1):M}$, we have:

$$\frac{d^2 f_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{Z}, s)}{ds^2} \leq 0. \tag{59}$$

Therefore, for fixed $\boldsymbol{\Theta}[1:M_1]$, we obtain that the function:

$$\widehat{\mathcal{R}}_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}) = \sum_{m=1}^{M_1} \widehat{\mathcal{R}}_n^{(m)}(\boldsymbol{\Theta}^{(m)}; \boldsymbol{X}^{(m)}, \boldsymbol{y}^{(m)}) + f_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y}). \tag{60}$$

is concave in s. Next, we recall that pointwise infimum of arbitrary collections of concave functions is concave [Rockafellar, 1970]. Therefore, we obtain that the function:

$$q_n(s) = \mathbb{E}\left[\min_{\boldsymbol{\Theta}[1:M_1]} \widehat{\mathcal{R}}_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{G}, \boldsymbol{y})\right], \tag{61}$$

is concave in s. Then, by Danskin's theorem, the subdifferential of q_n at zero is the set

$$\left\{\langle h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})^2\rangle_{M_1+1:M}, \boldsymbol{\Theta}[1:M_1] \in \arg\min \widehat{\mathcal{R}}_{n,0}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y})\right\}$$

But by Assumption 6, this set only has one element, and hence q_n is differentiable at 0. □

Next, we relate the convergence of the above functions to the expectation of $h(\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M)})$, through the following standard result from Convex Analysis:

Theorem 11. *(Theorem 25.7. in Rockafellar [1970]): Let C be an open convex set, and let f be a convex function which is finite and differentiable on C. Let $f_1, f_2, \ldots,$ be a sequence of convex functions finite and differentiable on C such that $\lim_{n\to\infty} f_n(x) = f(x)$ for every $x \in C$. Then*

$$\lim_{n\to\infty} \nabla f_n(x) = \nabla f(x), \quad \forall x \in C.$$

By Assumption 5,

$$\lim_{n\to\infty} q_{g,n}(s) = q(s). \tag{62}$$

Applying theorem 11 to the sequence $q_{g,n}(s)$ yields:

$$\lim_{n\to\infty} q_n'(0) = q'(0). \tag{63}$$

Now, consider the corresponding free energy for the data distribution $p_{\boldsymbol{x}}$:

$$q_{x,n}(s) = \mathbb{E}\left[\min_{\boldsymbol{\Theta}[1:M_1]} \widehat{\mathcal{R}}_{n,s}(\boldsymbol{\Theta}[1:M_1], \boldsymbol{X}, \boldsymbol{y})\right]. \tag{64}$$

We again have that $q_{x,n}(s)$ is a concave, differentiable function in s, with:

$$q_{x,n}'(0) = \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{X}}\right]. \tag{65}$$

Now, Proposition 3 and equation (62) imply that

$$\lim_{n\to\infty} q_{x,n}(s) = \lim_{n\to\infty} q_n(s) = q(s). \tag{66}$$

Therefore, Theorem 11 applied to the sequence of functions $q_{x,n}(s)$ implies that:

$$\lim_{n\to\infty} q_{x,n}'(0) = q'(0). \tag{67}$$

By equations 65 and Lemma 10, we then obtain:

$$\lim_{n,p\to\infty} \left|\mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{X}}\right] - \mathbb{E}\left[\left\langle h\left(\boldsymbol{\Theta}^{(1)}, \cdots, \boldsymbol{\Theta}^{(M)}\right)\right\rangle_{\boldsymbol{G}}\right]\right| = 0. \tag{68}$$

A.7 Proof of Theorem 6: One-dimensional CLT for Random Feature Models

Our proof relies on a reduction to Theorem 2 in Hu and Lu [2022] and the proof of corollary 2 in Montanari and Saeed [2022]. We first notice that it suffices to show the result when $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{z}}, \boldsymbol{\Sigma}^{\boldsymbol{z}})$ is a Gaussian variable; and upon rescaling of σ we shall assume that $\operatorname{tr}(\boldsymbol{\Sigma}^{\boldsymbol{z}}) = p$.

Let $\boldsymbol{V} = \boldsymbol{\Sigma}^{\boldsymbol{z}1/2}\boldsymbol{F}$. We express $\boldsymbol{z}$ as $\boldsymbol{z} = \boldsymbol{\mu}^{\boldsymbol{z}} + \boldsymbol{\Sigma}^{\boldsymbol{z}1/2}\boldsymbol{z}'$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Therefore, we have:

$$\boldsymbol{x} = \sigma(\boldsymbol{F}^\top \boldsymbol{z}) = \sigma(\boldsymbol{F}^\top \boldsymbol{\mu}^{\boldsymbol{z}} + \boldsymbol{V}^\top \boldsymbol{z}'). \tag{69}$$

We define the following events:

$$\mathcal{A}_1 = \left\{\sup_{i,j\in[d]} \left|\boldsymbol{v}_i^\top \boldsymbol{v}_j - \delta_{ij}\right| \le C_1 \left(\frac{\log d}{d}\right)^{1/2}\right\} \qquad \mathcal{A}_2 = \left\{\sum_{i\in[d]} \left|\|\boldsymbol{v}_i\|^2 - 1\right| \le C_2\right\}$$

$$\mathcal{A}_3 = \{\|\boldsymbol{F}\|_{\text{op}} \le C_3\} \qquad \mathcal{A}_4 = \{\|\boldsymbol{V}\|_{\text{op}} \le C_4\}$$

Since the $\boldsymbol{f}_i$ are independent and sub-gaussian, Lemma 22 in Montanari and Saeed [2022] implies that there exists constants C_1, C_2, C_3, C_4 such that $\mathcal{B} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$ is a high-probability event. Now, for $i \in [d]$, we define

$$\sigma_i(u) = \sigma(u + \boldsymbol{f}_i^\top \boldsymbol{\mu}^{\boldsymbol{z}}). \tag{70}$$

Now, as in Montanari and Saeed [2022], we argue that the proof of Theorem 2 in Hu and Lu [2022] still applies to our setting. Indeed:

- since $\boldsymbol{z}$ does not have identity covariance, we replace the conditions on $\boldsymbol{F}$ by the exact same ones on $\boldsymbol{V}$.
- the Stein's method they use proceeds term by term, so using a different σ_i in each term does not matter as long as they satisfy the boundedness assumptions above.
- since we match the means of $\boldsymbol{g}$ and those of $\boldsymbol{x}$, the requirement that σ be odd is unimportant in our setting.

In particular, for bounded Lipschitz test functions φ, the proof of Lemma 2 in Hu and Lu [2022] shows that for any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\left| \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{x})\right] - \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{g})\right] \right| \leq \frac{C\|\boldsymbol{\theta}\|_\infty \operatorname{polylog}(p)}{\nu^2} \tag{71}$$

where ν^2 is the variance of $\boldsymbol{\theta}^\top \boldsymbol{x}$:

$$\nu^2 = \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}. \tag{72}$$

We now place ourselves in the setting where $\boldsymbol{\theta} \in \mathcal{S}_p$ where $\mathcal{S}_p$ is defined in (31), and we consider two cases:

(i) if $\nu^2 > p^{-2\eta/3}$, then (71) reduces to

$$\left| \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{x})\right] - \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{g})\right] \right| \leq \frac{C \operatorname{polylog}(p)}{p^{\eta/3}}. \tag{73}$$

(ii) if instead $\nu^2 > p^{-2\eta/3}$, then by the Lipschitz property of φ we have

$$\left| \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \boldsymbol{x})\right] - \varphi(\boldsymbol{\mu}) \right| \leq C\sqrt{\nu^2} = \frac{C}{p^{\eta/3}}, \tag{74}$$

and the same holds for $\boldsymbol{g}$.

In both cases, the bounds goes to 0 uniformly over the whole constraint set $\mathcal{S}_p$, which shows that Assumption 4 holds.

We now move to checking Assumption 7; Lemma 8 in Hu and Lu [2022] (more precisely, eq. (159)) exactly shows that for $\boldsymbol{\theta} \in \mathcal{S}_p$, the random variable $\boldsymbol{\theta}^\top x - \boldsymbol{\theta}^\top \boldsymbol{\mu}$ is C-subgaussian for an absolute constant C. Hence, we only need to show that $\boldsymbol{\mu}$ is uniformly bounded. Recall that

$$\mu_i = \mathbb{E}\left[\sigma(\boldsymbol{f}_i^\top \boldsymbol{z})\right] = \mathbb{E}\left[\sigma\left(\boldsymbol{f}_i^\top \boldsymbol{\mu}^{\boldsymbol{z}} + (1+\tau_i)\tilde{z}\right)\right]$$

where $\tilde{z}$ is a standard normal variable and

$$\tau_i = \|\boldsymbol{v}_i\| - 1.$$

Then, we can write using the Lipschitz property of σ

$$\sigma(\boldsymbol{f}_i^\top \boldsymbol{z}) = \sigma(\tilde{z}) + (\boldsymbol{f}_i^\top \boldsymbol{\mu}^{\boldsymbol{z}} + \tau_i \tilde{z})\tilde{\sigma}(\boldsymbol{f}_i^\top \boldsymbol{z})$$

where $\tilde{\sigma}$ is a uniformly bounded function. By assumption, $\sigma(\tilde{z})$ has zero mean, and hence by the Cauchy-Schwarz inequality

$$\begin{aligned}
\|\boldsymbol{\mu}\|^2 &\leq C\left(\sum_{i\in[d]} (\boldsymbol{f}_i^\top \boldsymbol{\mu}^{\boldsymbol{z}})^2 + (\|\boldsymbol{v}_i\| - 1)^2\right) \\
&\leq C\left(\|\boldsymbol{F}\|_{\text{op}}^2 \|\boldsymbol{\mu}^{\boldsymbol{z}}\|^2 + \sum_{i\in[d]} \left|\|\boldsymbol{v}_i\|^2 - 1\right|\right) \\
&\leq C'
\end{aligned}$$

under the high-probability event $\mathcal{B}$. □

A.8 Proof of Theorem 5

Our proof utilizes the results in Loureiro et al. [2021b] that describe the asymptotic limits of the estimators obtained through empirical risk minimization on the mixture of gaussians dataset. We note that the assumptions A1-A5 of their Theorem 1 are satisfied by our setting.

Let $\boldsymbol{W}^\star$ denote the minimizer of the objective in equation 23, and let $\boldsymbol{Z}^\star = \boldsymbol{X}\boldsymbol{W}^\star$. Let $\boldsymbol{\xi}_{k\in[K]} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$, $\boldsymbol{\Xi}_k \in \mathbb{R}^{K\times d}$ be sets of K-dimensional vectors and dimensional matrices respectively, with i.i.d entries sampled from $\mathcal{N}(0,1)$.

Then, Theorem 1 in Loureiro et al. [2021b] proves that for any pseudo-lipschitz functions of finite order, $\phi_1 : \mathbb{R}^{K\times d} \to \mathbb{R}, \phi_2 : \mathbb{R}^{K\times n} \to \mathbb{R}$:

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\boldsymbol{G})\right], \qquad \phi_2(\boldsymbol{Z}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\xi}}\left[\phi_2(\boldsymbol{H})\right] \tag{75}$$

Here $\boldsymbol{G}$ and $\boldsymbol{H}$ are functions of certain finite dimensional parameters

$$\boldsymbol{u} := (\boldsymbol{Q}_k \in \mathbb{R}^{K\times K}, \boldsymbol{M}_k \in \mathbb{R}^K, \boldsymbol{V}_k \in \mathbb{R}^{K\times K}, \hat{\boldsymbol{Q}}_k \in \mathbb{R}^{K\times K}, \hat{\boldsymbol{m}}_k \in \mathbb{R}^K, \hat{\boldsymbol{V}}_k \in \mathbb{R}^{K\times K})_{k\in[K]}$$

and the random vectors $\boldsymbol{\xi}_{k\in[K]}, \boldsymbol{\Xi}_{k\in[K]}$. The matrix $\boldsymbol{H}$ is obtained by concatenating the following functions $\boldsymbol{h}_k$, $\rho_k n$ time for each k:

$$\boldsymbol{h}_k = \boldsymbol{V}_k^{1/2} \operatorname{Prox}_{\ell(\boldsymbol{e}_k, \boldsymbol{V}_k^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k) \in \mathbb{R}^K, \qquad \boldsymbol{\omega}_k \equiv \boldsymbol{M}_k + \boldsymbol{b} + \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k, \tag{76}$$

Similarly, the matrix $\boldsymbol{G} \in \mathbb{R}^{K\times d}$ is described by:

$$\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}} \odot \operatorname{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}} \odot \boldsymbol{B}), \quad \mathbf{A}^{-1} \equiv \sum_k \hat{\boldsymbol{V}}_k \otimes \boldsymbol{\Sigma}_k, \quad \boldsymbol{B} \equiv \sum_k \left(\boldsymbol{\mu}_k \hat{\boldsymbol{m}}_k^\top + \boldsymbol{\Xi}_k \odot \sqrt{\hat{\boldsymbol{Q}}_k \otimes \boldsymbol{\Sigma}_k}\right).$$

Further define the function: $\boldsymbol{f}_k \equiv \boldsymbol{V}_k^{-1}(\boldsymbol{h}_k - \boldsymbol{\omega}_k)$. The equivalent bias vector $\boldsymbol{b}^\star$ is defined through the linear equation:

$$\sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{V}_k \boldsymbol{f}_k\right] = \boldsymbol{0}, \tag{77}$$

and is therefore, unique, differentiable in $\boldsymbol{u}$. The parameters $\boldsymbol{u}$ satisfy the following equations:

$$\begin{cases} \boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\Sigma}_k\boldsymbol{G}^\top] \\ \boldsymbol{M}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\mu}_k] \\ \boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}\left[\left(\boldsymbol{G}\odot\left(\hat{\boldsymbol{Q}}_k\otimes\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}}\odot\left(\boldsymbol{I}_K\otimes\boldsymbol{\Sigma}_k\right)\right)\boldsymbol{\Xi}_k^\top\right] \end{cases} \quad \begin{cases} \hat{\boldsymbol{Q}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{f}_k^\top\right] \\ \hat{\boldsymbol{V}}_k = -\alpha\rho_k\boldsymbol{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{\xi}^\top\right] \\ \hat{\boldsymbol{m}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right]. \end{cases} \tag{78}$$

.

We observe that the system of equations (78) can be expressed as a multi-dimensional fixed point equation:

$$\boldsymbol{u} = F_n(\boldsymbol{u}). \tag{79}$$

We make the following assumption, which is slightly stronger than (A5) in Loureiro et al. [2021b]:

Assumption A5. *The fixed point equations $\boldsymbol{u} = F_n(\boldsymbol{u})$ have unique solutions $\forall n \in \mathbb{N}$. Let $\hat{\boldsymbol{u}}_n$ be the unique solution to $\boldsymbol{u} = F_n(\boldsymbol{u})$. We further assume the solutions are uniformly bounded, i.e:*

$$\|\hat{\boldsymbol{u}}_n\| \leq K \tag{80}$$

for some constant K, and the jacobian of the fixed point equations $I - \frac{dF_n}{d\boldsymbol{u}}$ is invertible. Furthermore, we assume that the same assumptions hold for the limiting equations $\boldsymbol{u} = F(\boldsymbol{u})$.

Remark: While we assume the above conditions, as noted in Loureiro et al. [2021b], the fixed point equations 78, correspond to the optimality conditions of a strictly convex-concave problem. This can be rigorously proven using the properties of Bregman envelopes, as in Celentano et al. [2020], Loureiro et al. [2021a]. The strict convexity-concavity then implies the uniqueness of the fixed points as well as the differentiability of the limits.

We now prove the following result:

Lemma 12. *Under assumption 8, the system of equations (78) converge uniformly to a limiting system of equations $\boldsymbol{u} = F(\boldsymbol{u})$.*

Proof. We first show that the coordinates of the equivalent minimizer $\boldsymbol{G}$ can be expressed as follows:

$$\boldsymbol{G}_i = g(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}, \xi_i, \boldsymbol{u}) \tag{81}$$

Where g is a differentiable function and ξ_i denote independent Gaussian random variables. Indeed, from the separability assumption on r, and the definition of the prox operator, we have

$$\begin{aligned} \operatorname{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}}\odot\boldsymbol{B}) &= \arg\min_{\mathbf{z}} r(\mathbf{A}^{1/2}\odot\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - (\mathbf{A}^{1/2}\odot\boldsymbol{B})\|^2 &(82)\\ &= \arg\min_{\mathbf{z}} \sum_{i=1}^{d}\psi_r((\mathbf{A}^{1/2}\odot\boldsymbol{z})_i) + \frac{1}{2}\sum_{i=1}^{d}(\boldsymbol{z}_i - (\mathbf{A}^{1/2}\odot\boldsymbol{B})_i)^2. &(83)\end{aligned}$$

$\boldsymbol{u}$ therefore only depends on the i_{th} entry of $(\mathbf{A}^{1/2}\odot\boldsymbol{B})$. Further, since all $\boldsymbol{\Sigma}_c$ are assumed diagonal, the entries of $(\mathbf{A}^{1/2}\odot\boldsymbol{z})_i$ and $(\mathbf{A}^{1/2}\odot\boldsymbol{B})_i$ only depend on $\boldsymbol{z}_i, \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}, \xi_i$, and the parameters $\boldsymbol{u}$. The differentiability of g then follows from the implicit function theorem applied to $\psi_r(\mathbf{A}^{1/2}\odot\bullet) + 1/2(\bullet - (\mathbf{A}^{1/2}\odot\boldsymbol{B})_i)^2$. The same holds for the matrix $\boldsymbol{H}$.

We next observe that each coordinate of F_n of the above system of equations 78 can be expressed as an expectation of a fixed continuous function w.r.t the joint empirical measure of $(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}})$. For instance, consider the $(i,j)_{th}$ entry of $\boldsymbol{Q}_k$. We have:

$$\boldsymbol{Q}_{k,ij} = F_{q,i,j,n} = \frac{1}{d}\sum_{\ell=1}^{d} \mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}_{i\ell}(\boldsymbol{\Sigma}_k)_{\ell\ell}\boldsymbol{G}_{\ell j}]. \tag{84}$$

Using equation (81), we have that $G_{i\ell}$ only depends on the ℓ_{th} coordinates of the means, covariances. Therefore, for fixed $\hat{\boldsymbol{u}}_n$, $\boldsymbol{Q}_{k,ij}$ is an expectation w.r.t the joint empirical measure of $(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}})$ of a continuous function. By Assumption 8, we have that

$$\lim_{n\to\infty} F_{q,i,j,n} = F_{q,i,j},$$

where $F_{q,i,j}$ denotes the expectation of $E_{\boldsymbol{\Xi}}[\boldsymbol{G}_{i\ell}(\boldsymbol{\Sigma}_k)_{\ell\ell}\boldsymbol{G}_{\ell j}]$ w.r.t the joint empirical measure.

To show that the convergence is uniform, we utilize assumptions 7 and A5. Since each term in F_n can be expressed as:

$$F_n^j(\boldsymbol{u}) = \frac{1}{d}\sum_{i=1}^{d} \Phi_j(\boldsymbol{u}, \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}), \tag{85}$$

for some function Φ_j with Lipschitz constant L_j. Therefore, for any $n \in \mathbb{N}$,$F_n^j(\boldsymbol{u})$ is L_j Lipschitz. This implies the uniform convergence of F_n to F. □

We now use the above result to prove convergence of the sequence of solutions $\hat{\boldsymbol{u}}_n$.

Lemma 13. *Under Assumptions 8 and A5:*

$$\lim_{n\to\infty} \hat{\boldsymbol{u}}_n = \hat{\boldsymbol{u}}, \tag{86}$$

where $\hat{\boldsymbol{u}}$ is the solution to the limiting equations $\boldsymbol{u} = F(\boldsymbol{u})$.

Proof. By assumption, $\hat{\boldsymbol{u}}_n$ are bounded. Therefore, by the Bolzano–Weierstrass theorem, there exists a convergent subsequence. Let $\hat{\boldsymbol{u}}_{n_j}$ be any such subsequence with corresponding limit $\tilde{\boldsymbol{u}}$. We have:

$$\|F(\tilde{\boldsymbol{u}}) - \tilde{\boldsymbol{u}}\| \le \|F(\tilde{\boldsymbol{u}}) - F(\hat{\boldsymbol{u}}_{n_j})\| + \|F(\hat{\boldsymbol{u}}_{n_j}) - F_{n_j}(\hat{\boldsymbol{u}}_{n_j})\|.$$

By uniform convergence of F_n to F (Lemma 12, we have that $\|F(\hat{\boldsymbol{u}}_{n_j}) - F_{n_j}(\hat{\boldsymbol{u}}_{n_j})\| \to 0$ while $\|F(\tilde{\boldsymbol{u}}) - F(\hat{\boldsymbol{u}}_{n_j})\| \to 0$ from the convergence of $\hat{\boldsymbol{u}}_{n_j}$ to $\tilde{\boldsymbol{u}}$ and the continuity of F. Therefore, we must have $F(\tilde{\boldsymbol{u}}) - \tilde{\boldsymbol{u}} = 0$ and thus $\tilde{\boldsymbol{u}} = \hat{\boldsymbol{u}}$. Therefore, any convergence subsequence of $\hat{\boldsymbol{u}}_n$ converges to $\hat{\boldsymbol{u}}$. Since $\hat{\boldsymbol{u}}_n$ are bounded, this implies that $\lim_{n\to\infty} \hat{\boldsymbol{u}}_n = \hat{\boldsymbol{u}}$. □

Now, let Φ be a twice differentiable test function with a Hessian having bounded spectral norm. We define:

$$h_d(\boldsymbol{W}) = \frac{1}{d}\sum_{i=1}^{d} \Phi(\{W_{c,i}\}_{c\in\mathcal{C}}, \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}). \tag{87}$$

Let $\hat{\boldsymbol{u}}(s)$ denote the solution to the limiting equation $\boldsymbol{u} = F_s(\boldsymbol{u})$ defined in Lemma 12 for regularization

$$r_s(\boldsymbol{W}) = \frac{\lambda}{n} r_d(\boldsymbol{W}) + s h_d(\boldsymbol{W})$$

By Assumption A5, $\boldsymbol{I} - \frac{dF_s}{d\boldsymbol{u}}$ is invertible in a neighbourhood of 0. Therefore, by implicit function theorem and uniqueness of the fixed points, $\hat{\boldsymbol{u}}(s)$, is a continuously differentiable function of s.

Using equation (81), we obtain that G is a differentiable function of s. We now consider the perturbed training loss:

$$\widehat{\mathcal{R}}(\boldsymbol{W}, \boldsymbol{b}, s) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\frac{\boldsymbol{W}\boldsymbol{x}_i}{\sqrt{d}} + \boldsymbol{b}, \boldsymbol{y}_i\right) + r_s(\boldsymbol{W}). \tag{88}$$

By the boundedness of the Hessian of Φ, $\widehat{\mathcal{R}}(\boldsymbol{W}, \boldsymbol{b}, s)$ satisfies the assumptions of strict convexity, coercivity for small enough s. We first note from (75) and Theorem 2 in Loureiro et al. [2021b], the expected training error converges to the following limit:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{n}\sum_{i=1}^{n} \ell\left(\frac{\boldsymbol{W}\boldsymbol{x}_i}{\sqrt{d}} + \boldsymbol{b}, \boldsymbol{y}_i\right)\right] \xrightarrow[n,d\to+\infty]{} \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}[\ell(\boldsymbol{e}_k, \boldsymbol{h}_k)]. \tag{89}$$

Similar to equation (81), we have that $\boldsymbol{h}_k$ are continuous, differentiable functions of $\hat{\boldsymbol{u}}(s)$.

From Assumption 9 and the form of the perturbation 87, we further have that $\lambda r(\boldsymbol{W}) + sh(\boldsymbol{W})$ is a pseudo-Lipschitz function of $\boldsymbol{W}$ of finite order. Therefore, Equation (75) gives:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{d}\lambda r(\boldsymbol{W}^*) + sh(\boldsymbol{W}^*)\right] \xrightarrow[n,d\to+\infty]{} \mathbb{E}_{\boldsymbol{\Xi}}\left[\frac{1}{d}\lambda r_d(\boldsymbol{G}_n) + sh_d(\boldsymbol{G}_n)\right]. \tag{90}$$

Define:

$$q_n(s, \boldsymbol{u}) = \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{\Xi}}\left[\psi_r(g(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}, \boldsymbol{\Sigma}_i^k, \xi_i, \boldsymbol{u}_n))\right] \tag{91}$$

$$+ s\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{\Xi}}\left[\Phi(g(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}, \boldsymbol{\Sigma}_i^k, \xi_i, \boldsymbol{u}), \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}})\right]. \tag{92}$$

From assumption 9 and equation 81, we have:

$$\mathbb{E}_{\boldsymbol{\Xi}}\left[\frac{1}{d}\lambda r_d(\boldsymbol{G}_n) + sh_d(\boldsymbol{G}_n)\right] = q_n(s, \hat{\boldsymbol{u}}_n(s)). \tag{93}$$

Similar to the proof of Lemma 12, we obtain that $q_n(s, \boldsymbol{u})$ converges uniformly to $q(s, \boldsymbol{u})$ given by the corresponding expectation w.r.t the limiting empirical measure:

$$q(s, \boldsymbol{u}) = \mathbb{E}_{p(\boldsymbol{\sigma},\boldsymbol{\mu})}\left[\mathbb{E}_{\boldsymbol{\Xi}}\left[\Phi(g(\{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}}, \boldsymbol{\Sigma}_i^k, \xi_i, \boldsymbol{u}), \{\mu_{c,i}\}_{c\in\mathcal{C}}, \{\sigma_{c,i}\}_{c\in\mathcal{C}})\right]\right]. \tag{94}$$

Due to the uniform convergence, we further have:

$$\lim_{n\to\infty} q_n(s, \hat{\boldsymbol{u}}_n(s)) = q(s, \hat{\boldsymbol{u}}(s)). \tag{95}$$

We conclude that:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{n}\sum_{i=1}^{n} \ell\left(\frac{\boldsymbol{W}^*\boldsymbol{x}_i}{\sqrt{d}} + \boldsymbol{b}^*, \boldsymbol{y}_i\right) + \frac{1}{d}\lambda r(\boldsymbol{W}^*) + sh(\boldsymbol{W}^*)\right] \to \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}[\ell(\boldsymbol{e}_k, \boldsymbol{h}_k)] + q(s, \hat{\boldsymbol{u}}(s)). \tag{96}$$

Since the RHS is a differentiable function in s, Assumption 5 is satisfied for the perturbation $h(\boldsymbol{W})$. Due to the coercivity of $\ell(\boldsymbol{y}, \bullet \boldsymbol{X}) + r(\bullet)$, there exists a sequence of fixed compact subsets containing the minimizers $\boldsymbol{W}^*$ with high probability as $n \to \infty$ (see Lemma 8 in Loureiro et al. [2021a]). Furthermore, since the input distribution is given by a mixture of gaussians with bounded means, Assumption 8 is satisfied for any such sequence of constraint sets. Therefore, the validity of assumption 5 through Equation 5 allows the applicability of Theorem 4 for the statistic $h_d(\boldsymbol{W})$. Through standard approximation techniques or the Stone–Weierstrass theorem, the restriction of differentiability and bounded Hessian of Φ can be removed. This completes the proof of Theorem 5 for general bounded Lipschitz Φ.

B Assumptions on the target function

In this section, we discuss possible generalizations of the assumptions on the target function. In (2), we assume a target function depending on a small number of linear projections in the input space, along with the class labels. However, when the inputs are generated through feature maps $\boldsymbol{x} = \psi(\boldsymbol{z})$, one may instead consider target functions depending directly on the latent vectors $\boldsymbol{z}$. This was the setup considered in Hu and Lu [2022] for random feature maps. For mixture models considered in our work, one may assume:

$$y_i(\boldsymbol{X}) = \eta(\boldsymbol{\Theta}_\star^\top \boldsymbol{z}_i, \varepsilon_i, c_i). \tag{97}$$

We conjecture that our results can be generalized to the above setup through the use of the following stronger assumption:

Assumption A6. *For any Lipschitz function $\varphi : \mathbb{R} \to \mathbb{R}$,*

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\Theta}_1 \in \mathcal{S}_p^{\boldsymbol{x}}, \boldsymbol{\Theta}_2 \in \mathcal{S}_d^{\boldsymbol{z}}} \left| \mathbb{E}\left[\varphi(\boldsymbol{\Theta}_1^\top \boldsymbol{x}, \boldsymbol{\Theta}_2^\top \boldsymbol{z}) \,\middle|\, c_{\boldsymbol{x}} = c\right] - \mathbb{E}\left[\varphi(\boldsymbol{\Theta}_1^\top \boldsymbol{g}, \boldsymbol{\Theta}_2^\top \boldsymbol{z}) \,\middle|\, c_{\boldsymbol{g}} = c\right] \right| = 0, \quad \forall c \in \mathcal{C}. \tag{98}$$

Here $S_p^{\boldsymbol{x}}$ is the constraint set for the training parameters $\boldsymbol{\Theta}_1$ while $\mathcal{S}_d^{\boldsymbol{z}}$ denotes a suitable constraint set on $\mathbb{R}^d$ where d denotes the dimension of the latent vectors. Under the above assumption

Such an assumption has been discussed in Goldt et al. [2019] under the term "Hidden Manifold Model", and was proven in Hu and Lu [2022] for random feature maps acting on Gaussian noise.

C One-dimensional gaussian approximation

Although Theorem 7 is a powerful result, it still relies on very strong assumptions. In particular, given a distribution ν for the inputs $\boldsymbol{x}_i$, characterizing the set of vectors $\boldsymbol{\theta}$ such that Assumption 4 holds is in general a difficult task.

Rigorous results When the entries of $\boldsymbol{x}$ are i.i.d subgaussian, a classical application of the Lindeberg method [Lindeberg, 1922] shows that Assumptions 2 and 4 are satisfied with

$$\mathcal{S}_p = \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \|\boldsymbol{\theta}\|_\infty = o_p(1)\}.$$

More recently, this result (often used under the name "Gaussian Equivalence Theorem") was extended to general feature models with approximate orthogonality constraints [Goldt et al., 2022, Hu and Lu, 2022], for

the same choice of $\mathcal{S}_p$. Montanari and Saeed [2022] also provides a central limit theorem result for the Neural Tangent Kernel of [Jacot et al., 2018], for a more convoluted parameter set $\mathcal{S}_p$. While these papers provide a strong basis for the one-dimensional CLT, those rigorous results only concern (so far) a very restricted set of distributions.

Concentration of the norm Another, more informal line of work originating from Seddik et al. [2020], argues that most distributions found in the real world satisfy some form of the central limit theorem. The starting point of this analysis is the following theorem, adapted from Bobkov [2003]:

Theorem 14 (Corollary 2.5 from Bobkov [2003])**.** *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a random variable, with $\mathbb{E} * \boldsymbol{x}\boldsymbol{x}^\top = \boldsymbol{I}_p$, and η_p the smallest positive number such that*

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{x}\|_2}{\sqrt{p}} - 1\right| \geq \eta_p\right) \leq \eta_p. \tag{99}$$

Then for any $\delta > 0$, there exists a subset $\mathcal{S}_p$ of the p-sphere $\mathbb{S}^{p-1}$ of measure at least $4p^{3/8}e^{-cp\delta^4}$, such that

$$\sup_{\boldsymbol{\theta} \in \mathcal{S}_p} \sup_{t \in \mathbb{R}} \left|\mathbb{P}(\boldsymbol{\theta}^\top \boldsymbol{x} \geq t) - \Phi(t)\right| \leq \delta + 4\eta_p,$$

where Φ is the characteristic function of a standard Gaussian, and c is a universal constant.

If both δ and η_p are $o(1)$, Theorem 14 implies that Assumption 4 is satisfied for any compact subset $\mathcal{S}'_p \subseteq \mathcal{S}_p$. This suggests that the norm concentration property of (99) is a convenient proxy for one-dimensional CLTs. However, the proof of this theorem uses isoperimetric inequalities, and is thus non-constructive; as a result, characterizing precisely the set $\mathcal{S}_p$ remains an open and challenging mathematical problem.

Concentrated vectors In Seddik et al. [2020], the authors consider the concept of *concentrated* random variables, as defined in Ledoux [2001]:

Definition 15. *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a random vector. $\boldsymbol{x}$ is called (exponentially) concentrated if there exists two constants C, c such that for any 1-Lipschitz function $f : \mathbb{R}^p \to \mathbb{R}$, we have*

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

Since the norm function is 1-Lipschitz, it can be shown that any concentrated isotropic vector $\boldsymbol{x}$ satisfies (99), with

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

The converse is obviously not true; an exponential random vector still has $\eta_p \to 0$, but is not concentrated. However, even if it is stronger that (99), the concept of concentrated vectors has two important properties:

(i) a standard Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ satisfies Definition 15 with constants C, c independent from p,

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

Towards real-world datasets The real-world data considered in machine learning is often composed of very high-dimensional inputs, corresponding to $p \gg 1$ in our setting. However, it is generally accepted that this data actually lies on a low-dimensional manifold of dimension d_0: this is the idea behind many dimensionality reduction techniques, from PCA [Pearson, 1901] to autoencoders [Kramer, 1991]. Another, more recent line of work (see e.g. Facco et al. [2017]) studies the estimation of the latent dimension d_0; results for the MNIST dataset ($p = 784$) yield $d_0 \approx 15$, while CIFAR-10 ($p = 3072$) has estimated intrinsic dimension $d_0 \approx 35$ [Spigler et al., 2020].

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

$$\boldsymbol{x} = f(\mathbf{z}) \quad \text{with} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_0}) \tag{100}$$

Examples of functions f include GANs [Goodfellow et al., 2014], variational auto-encoders [Kingma and Welling, 2013], or normalizing flows [Rezende and Mohamed, 2015]. This ansatz has been studied theoretically, and the results compared with real-world datasets, in Goldt et al. [2019], Loureiro et al. [2021a]; the results indicate significant agreement between generated inputs and actual data.

Finally, we argue that for a large class of generative networks, the learned function f is actually Lipschitz, with a bounded constant. This is even often a design choice; indeed, theoretical results such as Bartlett et al. [2017] imply that a smaller Lipschitz constant improve the generalization capabilities of a network, or its numerical stability [Behrmann et al., 2021]. As a result, regularizations aimed at controlling the Lipschitz properties of a network are a common occurrence; see e.g. Miyato et al. [2018] for the spectral regularization of GANs. This indicates that concentrated vectors are indeed a good approximation for real-world data.

D Details on the numerical simulations

In this appendix we expand on how Fig. 2 was generated. This closely follows the pipeline illustrated in Fig. 1.

Step 1: Training of the cGAN — The first step consists on training a cGAN on the real data set. For Fig. 2, we have used a PyTorch [Paszke et al., 2019] implementation of the architecture in Chen et al. [2016] publicly available at the pytorch-generative-model-collections repository. The cGAN was trained on the fashion-MNIST dataset Xiao et al. [2017] following the default procedure in the repository: i.e. training for 50 epochs and batch size 64 using Adam with learning rate 0.0002 and $(\beta_1, \beta_2) = (0.5, 0.999)$ on the binary cross entropy (BCE) loss (equal for both generator and discriminator). The evolution of the training loss during training is given in Fig. 3, and samples from the generator during different epochs are shown in Fig. 4.

Step 2: Evaluating the class means and covariances — With the trained cGAN in hands, we can generate as many fresh samples as needed for our experiments. Moreover, a feature map can be easily added on the top of the cGAN architecture, as illustrated in Fig. 1. For Fig. 2, we have added a random feature map [Rahimi and Recht, 2007] $\boldsymbol{x} \mapsto \tanh(\mathrm{F}\boldsymbol{x})$ with projection matrix $\mathrm{F} \in \mathbb{R}^{1176\times 784}$ with entries $\mathrm{F}_{ij} \sim \mathcal{N}(0, 1/d)$. In order to compare the performance of a model trained on cGAN+RF samples vs. the equivalent Gaussian mixture model, we need to compute the class-wise means and covariances $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. For Fig. 2, this was done with a standard Monte Carlo scheme over 10^6 samples.

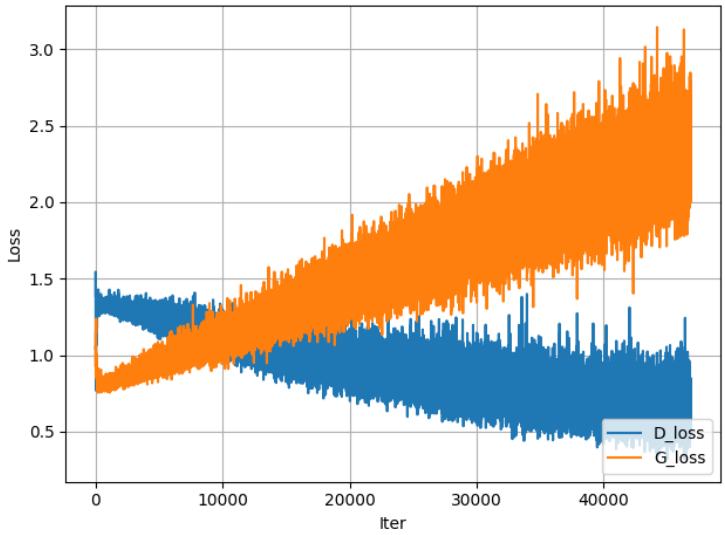

Figure 3: Evolution of the binary cross entropy training loss for the generator (orange) and discriminator (blue) during training.

Step 3: Learning curves — The last step consists of computing the curves for the test error of logistic and ridge regression trained on the cGAN+RF features. Each point in Fig. 2 corresponds to a fixed sample complexity $\alpha = {}^{n}/_{p}$. For each α, we generate fresh $n = \alpha \times p$ training features either from the cGAN+RF model (blue points) or from the equivalent Gaussian mixture model (red points). For the binary classification task, we split the samples over even vs. odds class labels. The SciPy [Virtanen et al., 2020] implementation of both ridge and logistic regression were used to train a classifier on the training data, from which both training error and test error were computed, using another batch of fresh samples for the latter. Finally, to reduce finite-size effects this procedure was repeated over 10 different seeds, with the average and standard deviation reported in Fig. 2.

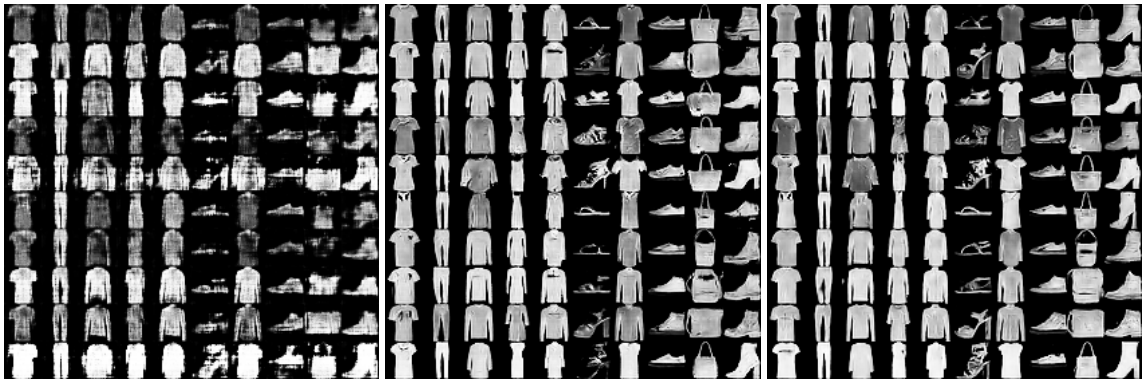

Figure 4: Fake fashion-MNIST Samples from the cGAN generator at the end of epoch 1 (left), 25 (middle) and 50 (right).