Are Gaussian data all you need?

Extents and limits of universality in high dimensional GLMs

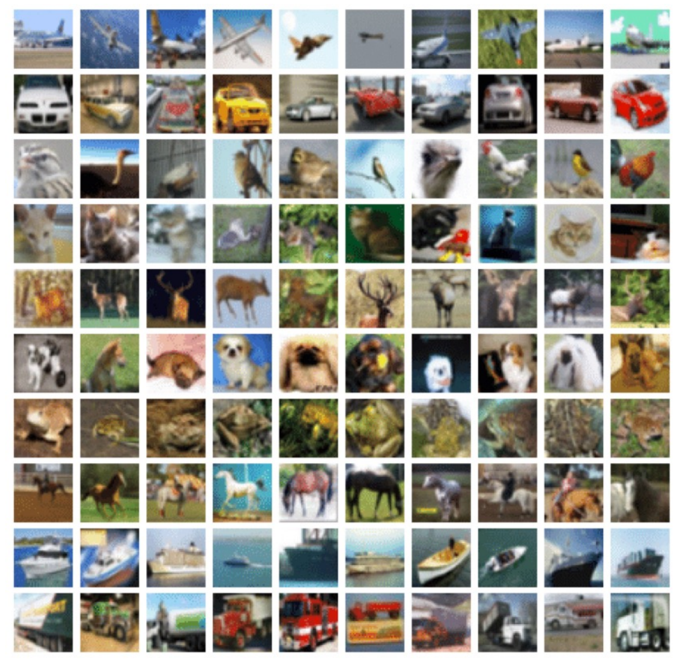
Luca Pesce[1], Florent Krzakala[1],
Bruno Loureiro[2], Ludovic Stephan[1]
[1] IdePHICS, EPFL; [2] ENS, Paris

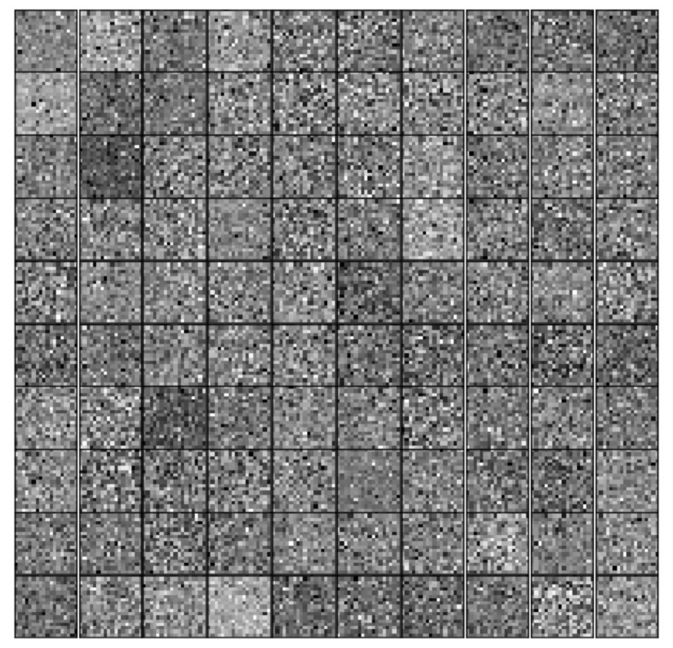
ICML International Conference On Machine Learning

Motivations

It is commonsense in machine learning that structure in the data is an important ingredient for successful learning. Quantifying this statement, and in particular how structure in the features impact the training and generalization errors the most, is an important endeavor in the broad program of "seeing through" the modern machine learning black box. Despite the seemingly constraining assumption on the distribution of the features, a recent line of work provides strong evidence for the Gaussian universality of the training and generalization errors in generalized estimation in different settings, e.g. closely related to this work [3,4,5]. These works beg the question **"when are Gaussian features a good model for learning?".**

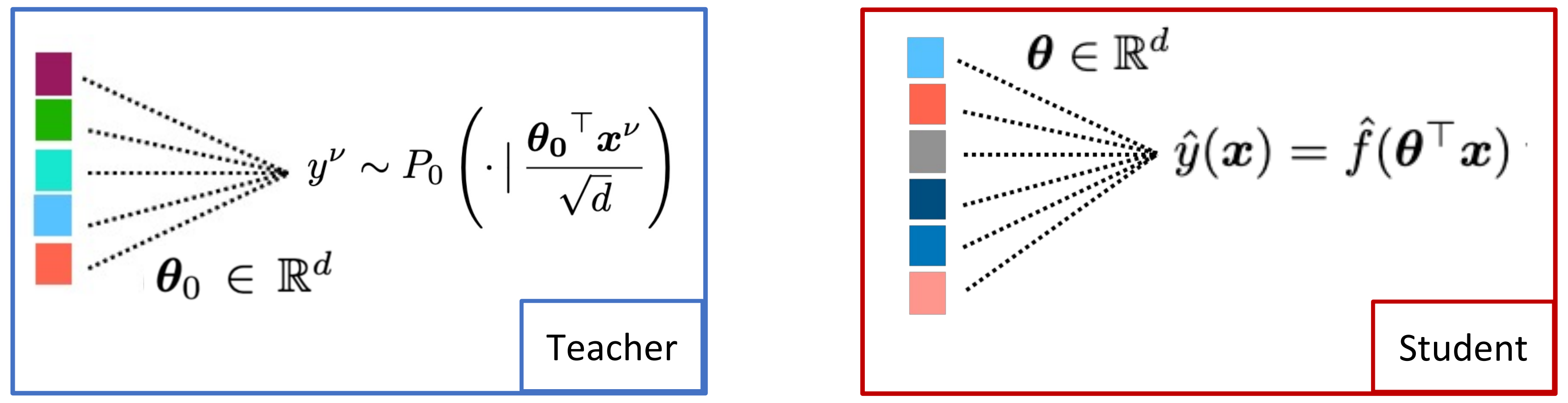
Real data are not Gaussian ... → ... Or are they?





Model & Setting

We are interested in studying the properties of generalized linear estimation, where the weights are found thanks to **Emprirical Risk Minimization (ERM)**.
In particular, we will be interested in characterizing the generalization and training errors under the so-called **teacher-student setting**: the label are generated from a target distribution depdendent on a fixed vector θ_0, called the teacher vector.



Teacher

$$y^\nu \sim P_0\left(\cdot \,\middle|\, \frac{\boldsymbol{\theta_0}^\top \boldsymbol{x}^\nu}{\sqrt{d}}\right)$$

$$\boldsymbol{\theta}_0 \in \mathbb{R}^d$$

Student

$$\boldsymbol{\theta} \in \mathbb{R}^d$$

$$\hat{y}(\boldsymbol{x}) = \hat{f}(\boldsymbol{\theta}^\top \boldsymbol{x})$$

ERM

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

Exact Asymtpotics of GMMs

Our aim is to analyze a popular model for multi-modal data, known to be able to approximate any distribution: the Gaussian Mixture Model (GMM). Our first result is to give closed-form **asymptotic characterization of the performance of the empirical risk minimizer** for the Gaussian Mixture model generalizing [5].

GMM

$$\boldsymbol{x}^\nu \sim \sum_{c\in\mathcal{C}} p_c \mathcal{N}\left(\frac{\boldsymbol{\mu}_c}{\sqrt{d}}, \Sigma_c\right)$$

$$p_c \in [0,1] \qquad \mathcal{C} := \{1, \cdots, K\}$$

→

Closed-form formula for the errors

$$\varepsilon_{tr}(\hat{\boldsymbol{\theta}}) \overset{P}{\simeq} \varepsilon_{tr}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathfrak{C}}, \{\Sigma_c\}_{c\in\mathfrak{C}})$$

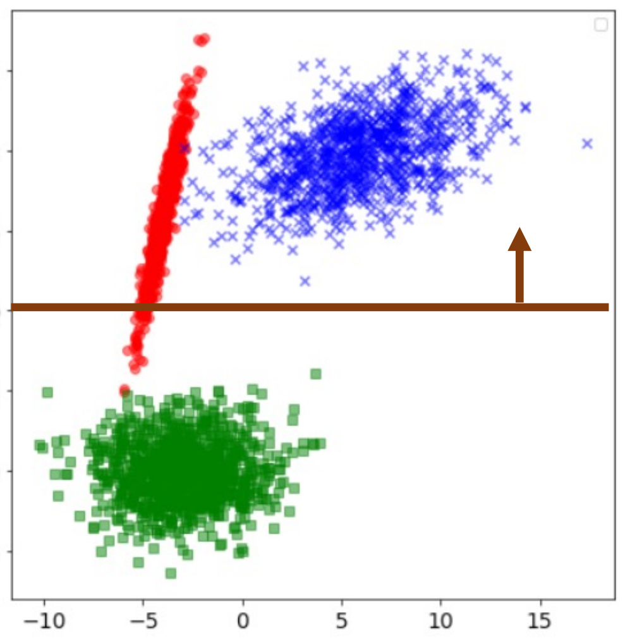
$$\varepsilon_{gen}(\hat{\boldsymbol{\theta}}) \overset{P}{\simeq} \varepsilon_{gen}(\boldsymbol{\theta}_0, \{\boldsymbol{\mu}_c\}_{c\in\mathfrak{C}}, \{\Sigma_c\}_{c\in\mathfrak{C}})$$

GMM Universality

The computation of the exact asymptotics for GMMs leads to the following key question: **when is the learning of GMMs equivalent to the Gaussian case?**

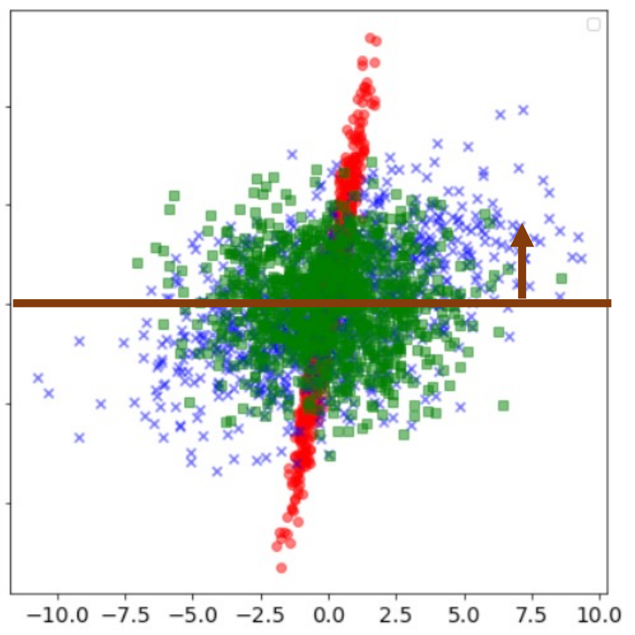
Mean Universality

The key difference between GMM and Gaussian exact asymptotics expression lies on the way the teacher vector correlates with the cluster means and covariances. A first step towards universality is therefore to **characterize under which conditions the asymptotic errors are independent of the means.**



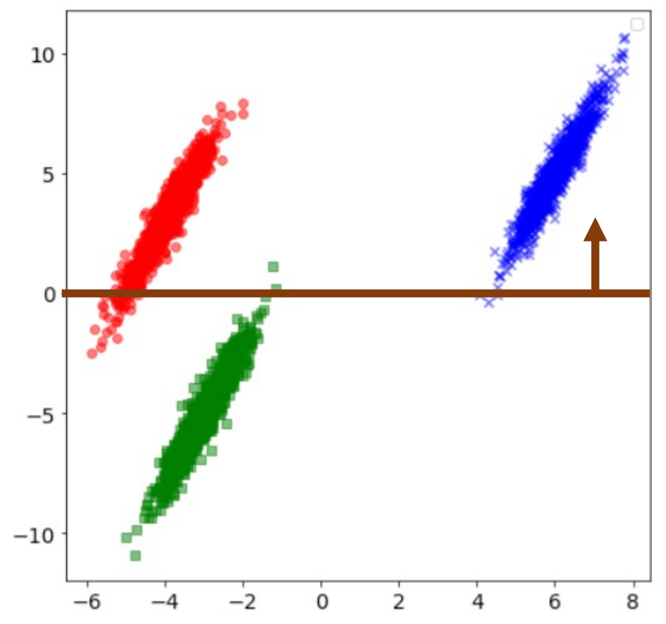
$$\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0$$

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



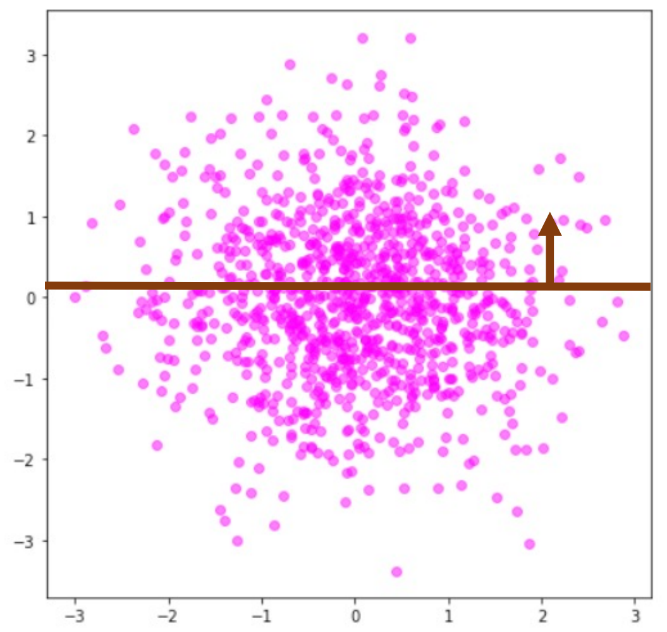
Covariance Universality

Stronger universality can be shown by doing simplifying assumptions on the data structure. Indeed, if the mixture is homogeneous (in the sense that the all the covariance are identical, a condition often called **homoscedasticity** in statistics) we have Gaussian universality. Moreover, if we consider the **unregularized case** and if it exists a **unique minimizer of the risk**, the form of the **covariance is irrelevant for the learning process.**



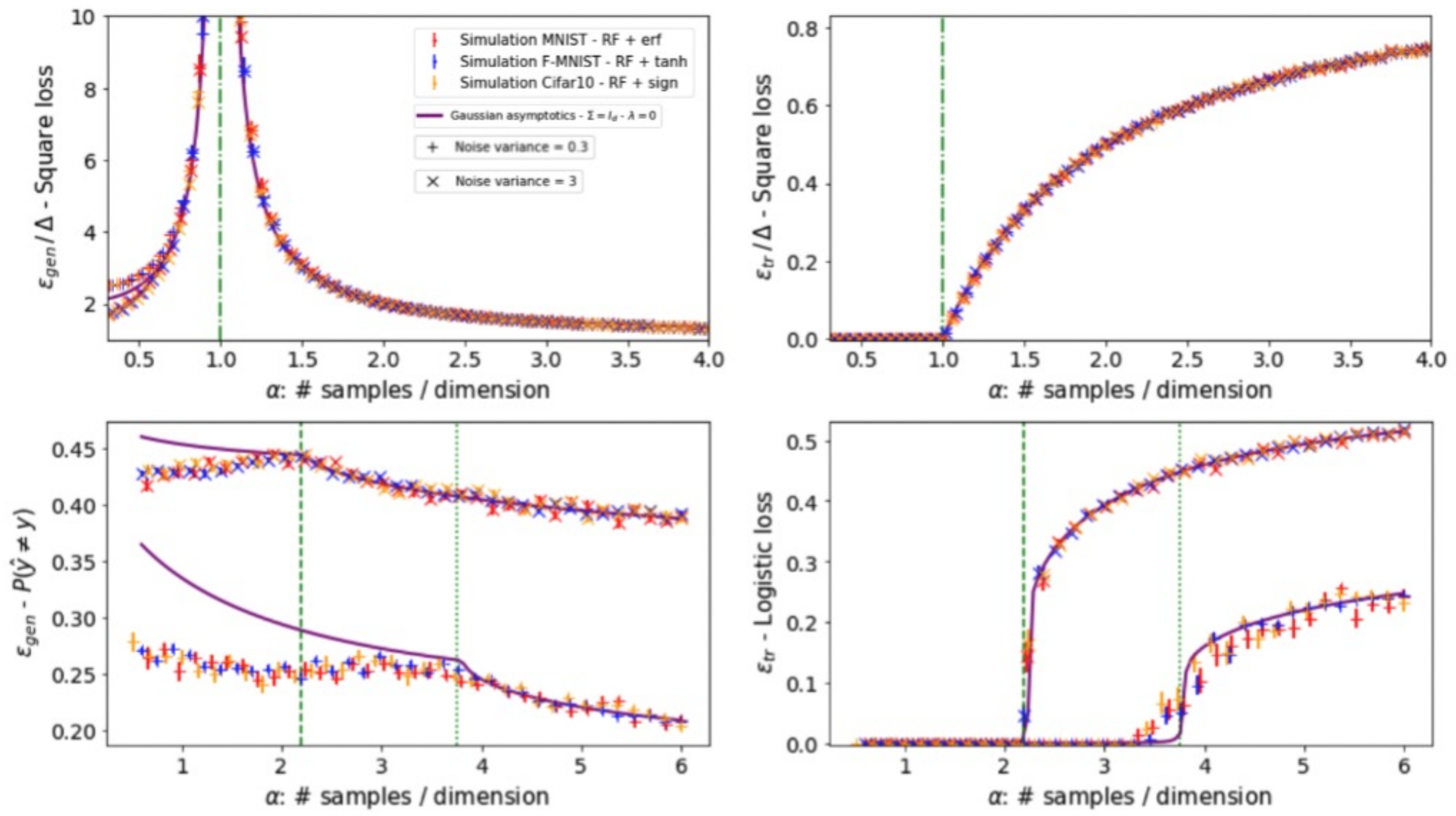
$$\Sigma_c = \Sigma \qquad \forall c \in \mathfrak{C}$$

$$\lambda = 0^+$$



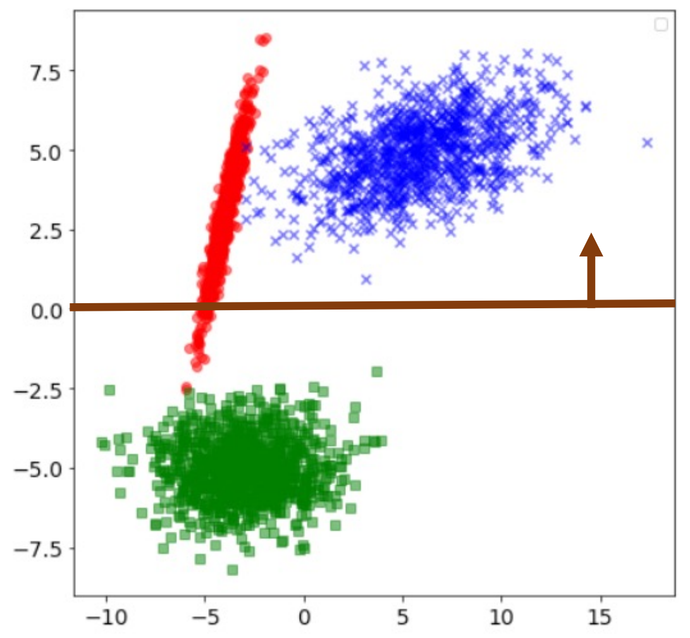
Real Data Universality

We analyze Gaussian universality of real data considering a random teacher function (to satisfy the mean universality conditions). We take three standard datasets: MNIST, Fashion-MNIST, and grayscale CIFAR-10, preprocessed with Random Feature (RF) map. The RF preprocessing step homogenize the data enough so that we observe Gaussian Universality for test and training errors in the interpolating regime, when an unique minimizer of the risk exists, in accordance with the Covariance Universality results.



Interpolators Strong Universality - I

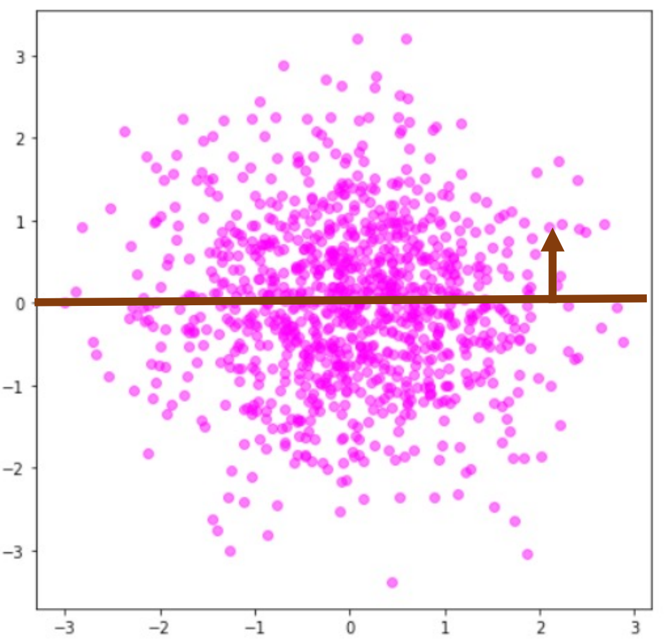
Surprisingly, if we further consider a **square loss** minimization, the estimation of **any GMM** (homoscedastic or not!) **under mean universality condition** can be mapped to those of a **trivial Gaussian** problem!



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

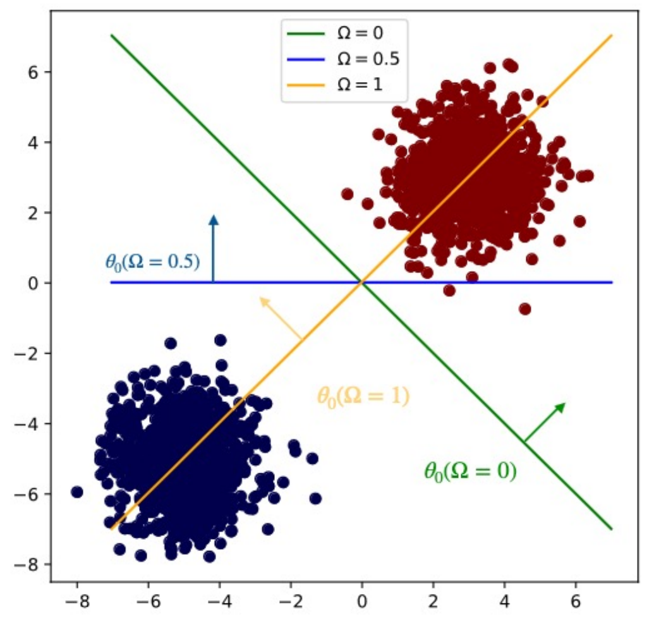
$$\lambda = 0^+$$

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$



Interpolators Strong Universality - II

We now consider the general case where the **target weights correlate with the structure** in the data, can we still say something? We focus on a simple controlled setting in which we can **express the ERM performance in a closed form for any teacher vector**. If we consider the **ridgeless** problem we again **restore Gaussian Universality for correlated teachers** (and simple mixture).

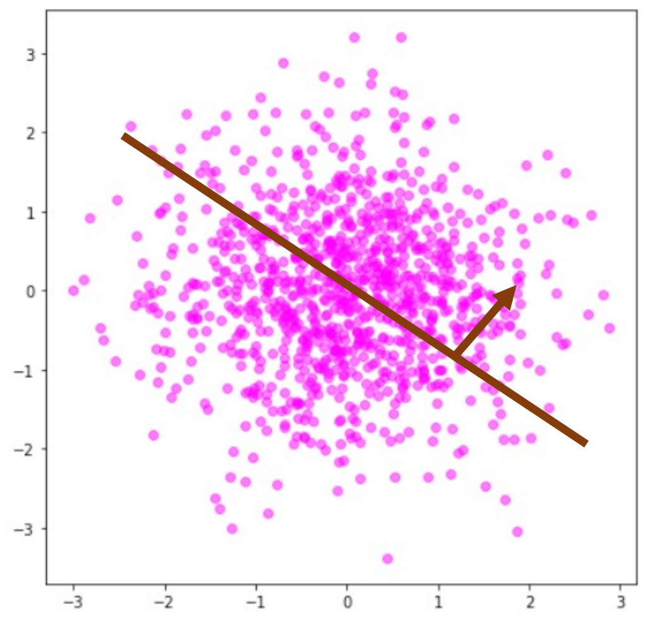


$$\cancel{\lim_{n,d\to\infty} \frac{\boldsymbol{\theta}_0^\top \boldsymbol{\mu}_c}{d} = 0}$$

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

$$\lambda = 0^+$$

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$



Conclusions

We provided a set of sufficient **conditions on the target function and the data structure** such that both the and test errors for a GMM are equivalent to the Gaussian case, characterizing the presence (or breaking) of Gaussian universality. We further demonstrated that the **Gaussian universality results can be observed on real data** after a random feature map to homogenize the data structure.

Rather than the structure of the data itself, what appears to matter is thus the correlation between this structure and the task to be learned.

Selected References

[1] **Are Gaussian data all you need? The extents and limits of high-dimensional generalized linear estimation**, L Pesce, F Krzakala, B Loureiro, L Stephan, arXiv: 2302.08923.
[2] Gaussian universality of perceptron with random labels, F Gerace, F Krzakala, B Loureiro, L Stephan, L Zdeborová, 2022.
[3] Montanari, A. and Saeed, B. N. Universality of empirical risk minimization, 2022.
[4] Hu, H. and Lu, Y. M. Universality laws for high-dimensional earning with random features, 2022.
[5] Learning gaussian mixtures with generalised linear models: Precise asymptotics in high-dimensions. B Loureiro, G Sicuro, C Gerbelot, A Pacco, F Krzakala, L Zdeborová, 2022.

Contact: luca.pesce@epfl.ch