GAUSSIAN UNIVERSALITY OF LINEAR CLASSIFIERS WITH RANDOM LABELS IN HIGH-DIMENSION

F. Gerace[1], F. Krzakala[2], B. Loureiro[2], L. Stephan[2], L. Zdeborová[2]

[1] International School of Advanced Studies *(SISSA), Trieste, Italy* [2] *École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

INTRODUCTION

While classical in many theoretical settings, *the assumption of i.i.d. Gaussian inputs* is often perceived as a strong limitation in the analysis of high-dimensional learning problems, out-of-touch with real-world practice. In this study, we redeem this line of work in the case of generalized linear classification with random labels. Our main contribution is a *rigorous proof* that data coming from a range of generative models in high-dimension have the same minimum training loss as Gaussian data with corresponding data covariance.

THE SETTING

We consider a dataset $\mathcal{D}=\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are the input vectors and $y_i \in \{-1,1\}$ the associated labels. On this dataset, we study the corresponding linear classification problem in the high-dimensional limit, e.g. $n,p\to\infty$ with $\alpha = \frac{n}{p} \sim O(1)$, and defined by the following empirical risk minimization:

$$\hat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}) = \inf_{\boldsymbol{\theta}\in S_p} \frac{1}{n}\sum_{i=1}^{n} \ell\left(\boldsymbol{\theta}^t \boldsymbol{x}_i, y_i\right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 \,,$$

where λ is the regularization strength and $\boldsymbol{\theta}$ is the vector of the learning model parameters, living in a compact subset S_p of $\mathbb{R}^p$. In particular, we mainly focus on the random label setting $y_i \sim \left(\frac{1}{2}\right)(\delta_{+1}+\delta_{-1})$ and we consider the following three types of input data models:

1. **The Gaussian Covariate (gc) model.** In this case, we independently sample the input vectors from a Gaussian distribution, e.g. $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$;

2. **The Gaussian Mixture (gm) model.** In this case, we independently sample the input vectors from a mixture of K different Gaussians, e.g. $\boldsymbol{x}_i \sim \sum_{c\in\mathcal{C}} \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, with $\mathcal{C} \equiv \{1,\dots,K\}$ indexing the K Gaussian clouds;

3. **The Neural Network Generative (nn) model.** In this case, we first sample a latent vector from a Gaussian Mixture distribution, e.g. $\boldsymbol{z}_i \sim \sum_{c\in\mathcal{C}} \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. We then generate the input vectors as:

$$\boldsymbol{x}_i = \Psi_{\text{nn}}(\boldsymbol{z}_i)$$

where Ψ_{nn} is the function parametrized by a neural network.

MAIN ANALYTICAL RESULT

Theorem 1. Assuming the following one-dimensional CLT to hold:

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\theta}\in S_p} \left|\mathbb{E}[\varphi(\boldsymbol{\theta}^t \boldsymbol{x})] - \mathbb{E}[\varphi(\boldsymbol{\theta}^t \boldsymbol{g})]\right| = 0,$$

with $\boldsymbol{g}_i \sim \sum_{c\in\mathcal{C}} \rho_c\, \mathcal{N}(\boldsymbol{\mu}_c^{\text{nn}}, \boldsymbol{\Sigma}_c^{\text{nn}})$, $\boldsymbol{\mu}_c^{\text{nn}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}[\Psi_{\text{nn}}(\boldsymbol{z})]$, $\boldsymbol{\Sigma}_c^{\text{nn}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}[(\Psi_{\text{nn}}(\boldsymbol{z})-\boldsymbol{\mu}_c^{\text{nn}})(\Psi_{\text{nn}}(\boldsymbol{z})-\boldsymbol{\mu}_c^{\text{nn}})^t]$ and $\boldsymbol{x}_i$ as in **3.**, for suitable regularity conditions on the loss and the labelling function η and for any bounded Lipshitz function $\Phi:\mathbb{R}\to\mathbb{R}$, we have:

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

with $y_i = \eta(\boldsymbol{\theta}_*^t \boldsymbol{x}_i, \epsilon_i)$, $\boldsymbol{\theta}_* \in S_p$ and ϵ_i i.i.d. noise. In particular:

$$\hat{\mathcal{R}}_n^*(\boldsymbol{X},\boldsymbol{y}(\boldsymbol{X})) \xrightarrow{\mathbb{P}} \varepsilon_{gm} \iff \hat{\mathcal{R}}_n^*(\boldsymbol{G},\boldsymbol{y}(\boldsymbol{G})) \xrightarrow{\mathbb{P}} \varepsilon_{gm},\ \forall \varepsilon_{gm}\in\mathbb{R}$$

Lemma 1. In the random label setting, if the loss is symmetric, e.g. $\ell(x,y)=\ell(-x,-y)$ for $x,y\in\mathbb{R}$, the limiting value ε_{gm} of the risk is independent from the means, that is:

$$\varepsilon_{gm}(\boldsymbol{\rho},\boldsymbol{M},\boldsymbol{\Sigma}^{\otimes}) = \varepsilon_{gm}(\boldsymbol{\rho},\boldsymbol{0},\boldsymbol{\Sigma}^{\otimes}),$$

with $\boldsymbol{\rho}\in[0,1]^K$ being the probability vector with entries ρ_c, $\boldsymbol{M}\in\mathbb{R}^{K\times p}$ the matrix of means and $\boldsymbol{\Sigma}^{\otimes}\in\mathbb{R}^{K\times p\times p}$ the concatenation of covariance matrices with rows $\boldsymbol{\Sigma}_c$.

Theorem 2. Given the assumptions in **Lemma 1**, and assuming the covariance matrices to be homogeneous, e.g. $\boldsymbol{\Sigma}_c=\boldsymbol{\Sigma}$ for all $c\in\mathcal{C}$, the asymptotic risk of a Gaussian Mixture is equivalent to that of a single Gaussian:

$$\varepsilon_{gm}(\boldsymbol{\rho},\boldsymbol{M},\boldsymbol{\Sigma}^{\otimes}) = \varepsilon_{gc}(\boldsymbol{0},\boldsymbol{\Sigma}).$$

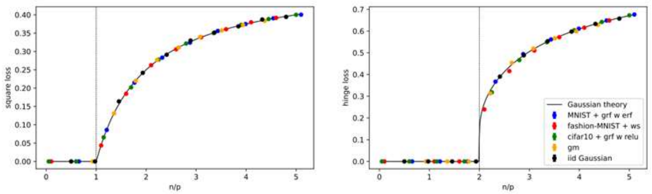
Theorem 3. Consider the same assumptions as in **Theorem 2**, if the minimizer is unique and the data matrix is full-rank, the asymptotic minimal loss for Gaussian data does not depend on the covariance for $\lambda=0$.

Theorem 4. In the specific case of ridge regression, when $\lambda\to 0^+$, we have:

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

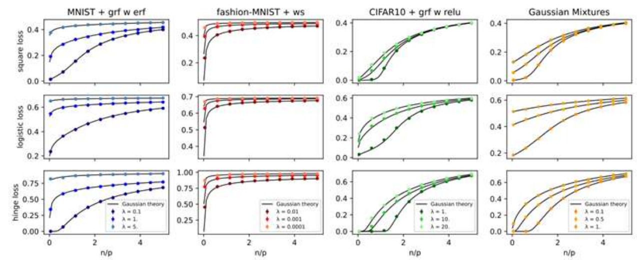
for any choice of $\boldsymbol{\rho}$, $\boldsymbol{M}$ and $\boldsymbol{\Sigma}^{\otimes}$.

GAUSSIAN UNIVERSALITY AT ZERO REGULARIZATION



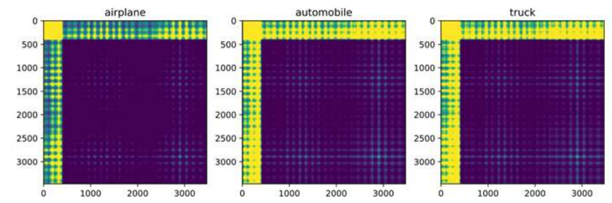
Training loss as a function of n/p at $\lambda=10^{-15}$. The black solid line represents the outcome of the theoretical prediction for $\boldsymbol{\Sigma}$ equal to the identity matrix $\boldsymbol{I}$. Colored dots refer to numerical simulations on MNIST pre-processed with Gaussian random features and error function non-linearity **(blue dots)**, fashion-MNIST pre-processed with wavelet scattering transform **(red dots)**, grayscale CIFAR10 pre-processed with Gaussian random features and relu non-linearity **(green dots)** and a mixture of two Gaussians with $\boldsymbol{\mu}_{1/2}=(\pm1,0,\dots0)$, $\boldsymbol{\Sigma}_{1/2}=\boldsymbol{I}$ and $\rho_{1,2}=1/2$ **(orange dots)**.

GAUSSIAN UNIVERSALITY AT FINITE REGULARIZATION



Training loss as a function of n/p at finite λ. The colored dots refer to numerical simulations on the same datasets of the previous plot. The black solid lines correspond to the theoretical predictions of the Gaussian Covariate model with $\boldsymbol{\Sigma}$ being the covariance matrix of the corresponding dataset.

HOMOGENITY ASSUMPTION



Input covariance matrices of grayscale CIFAR10 pre-processed with wavelet scattering transform. The covariances are conditioned on the true labels, e.g. airplane (rightmost), automobile (middle) and truck (rights most). Lighter colors refer to stronger correlations.