Learning Gaussian Mixtures with Generalised Linear Models: Precise Asymptotics in High-dimension

Bruno Loureiro[1], Gabriele Sicuro[2], **Cédric Gerbelot**[3], Alessandro Pacco[1], Florent Krzakala[1] and Lenka Zdeborova[4]
[1]EPFL IdePhics Lab., [2]King's College Math. Dept, [3]ENS Physics Dept., [4] EPFL SPOC Lab.

Motivation: high-dimensional asymptotics

Exact formulas for the performance of benchmark models in random design, high-dimensional setting

- quantitative theory for simple models (logistic regression, ...)
- difficult to extend to deep learning/elaborate feature maps
- simple models sometimes capture deep learning phenomenology

Need for tractable, realistic surrogate models for deep learning/complex feature maps

Ingredients for a surrogate model

- **learning architecture**
  ridge regression, support vector machine...
- **data/feature model**
  i.i.d. Gaussian with general covariance...
- training algorithm
  Not in this work, we directly focus on estimators.

Examples

- Instances of ridge regression with i.i.d. coordinates capture the so-called **double descent** phenomenon
- GAN data concentrates to **Gaussian mixtures**
- **Convex Generalized Linear Models** (GLM) with correlated Gaussian designs capture a wide range of single task regression problems, with structured data/feature maps

Objective

Can we have a realistic benchmark for multiclass classification problems?

Contributions

- **Study classification of a high-dimensional K-Gaussian mixture with a convex GLM**
- **Generic means and covariances for the clusters**
- **Exact asymptotic distribution of the estimator**
- **Study of both random design and real data problems**

The generative model: a K-Gaussian mixture

Consider the Gaussian mixture density with K cluster $\{\mathcal{C}_k\}_{1\leqslant k\leqslant K}$:

$$P(\mathbf{x},\mathbf{y})=\sum_{k=1}^{K} y_k\rho_k\mathcal{N}(\mathbf{x}\,|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{1}$$

- means $\boldsymbol{\mu}_k\in\mathbb{R}^d$, covariances $\boldsymbol{\Sigma}_k\in\mathbb{R}^{d\times d}$ positive definite.
- cluster membership $\rho_k\in[0,1]$ with $\sum_k\rho_k=1$.
- labels $\mathbf{y}\in\{\mathbf{e}_k\}_{k\in[K]}$ are **one-hot-encoded**: $\mathbf{x}\in\mathcal{C}_k \Leftrightarrow y_i=e_{ik}\equiv\delta_{ik}$.

Dataset obtained sampling n pairs $(\mathbf{x}^\nu,\mathbf{y}^\nu)_{\nu\in[n]}$ from Eq. (1).
We denote $\mathbf{X}=(x_i^\nu)_{\nu,i}\in\mathbb{R}^{n\times d}$.

The learning task

Learn K separating hyperplanes in $\mathbb{R}^d$: $\mathbf{W}^\star\in\mathbb{R}^{K\times d}$

The learning method: a convex GLM

Estimator obtained by minimising the empirical risk:

$$\mathcal{R}(\mathbf{W},\mathbf{b})\equiv\sum_{\nu=1}^{n}\ell\left(\mathbf{y}^\nu,\frac{\mathbf{W}\mathbf{x}^\nu}{\sqrt{d}}+\mathbf{b}\right)+\lambda r(\mathbf{W}), \tag{2}$$

$$(\mathbf{W}^\star,\mathbf{b}^\star)\equiv\underset{\mathbf{W}\in\mathbb{R}^{K\times d},\,\mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}}\ \mathcal{R}(\mathbf{W},\mathbf{b}), \tag{3}$$

- $\mathbf{W}\in\mathbb{R}^{K\times d}$, $\mathbf{b}\in\mathbb{R}^K$ are the weights and bias to be learned;
- ℓ convex loss and regularisation function (e.g., least-squares or logistic loss);
- r convex regularisation functions (e.g., ℓ_2 or ℓ_1 penalty).

Goal: asymptotic properties of $\mathbf{W}^\star$

High-dimensional limit: $n,d\to\infty$ **with fixed** $\alpha=n/d$
We characterise the asymptotic distribution of the estimator $(\mathbf{W}^\star,\mathbf{b}^\star)$.

Notation: If $\mathbf{G}=(G_{ki})_{ki}\in\mathbb{R}^{K\times d}$, $\mathbf{A}=(A_{ki\,k'i'})_{ki\,k'i'}\in\mathbb{R}^{K\times d}\otimes\mathbb{R}^{K\times d}$, then $\mathbf{G}\odot\mathbf{A}=\sum_{ki}G_{ki}A_{ki\,k'i'}\in\mathbb{R}^{K\times d}$. Moreover $\sqrt{\mathbf{A}}$ is the tensor such that $\mathbf{A}=\sqrt{\mathbf{A}}\odot\sqrt{\mathbf{A}}$.

Main result: exact asymptotics

- Let $\boldsymbol{\xi}_{k\in[K]}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_K)$ be collection of K-dimensional standard normal vectors independent of other quantities;
- let $\{\boldsymbol{\Xi}_k\}$ a set of K matrices, $\boldsymbol{\Xi}_k\in\mathbb{R}^{K\times d}$, with i.i.d. standard normal entries, independent of other quantities;
- let $\mathbf{Z}^\star=\frac{1}{\sqrt{d}}\mathbf{W}^\star\mathbf{X}\in\mathbb{R}^{K\times n}$.

Under mild feasibility and regularity assumptions, for any pseudo-Lispchitz functions $\phi_1:\mathbb{R}^{K\times d}\to\mathbb{R}$, $\phi_2:\mathbb{R}^{K\times n}\to\mathbb{R}$:

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

where we have introduced the proximal for the loss:

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

and $\boldsymbol{H}\in\mathbb{R}^{K\times n}$ is obtained by concatenating each $\mathbf{h}_k$, $\rho_k n$ times.
We have also introduced the matrix proximal $\mathbf{G}\in\mathbb{R}^{K\times d}$:

$$\mathbf{G}=\sqrt{\mathbf{A}}\odot\operatorname{Prox}_{r(\sqrt{\mathbf{A}}\odot\bullet)}(\sqrt{\mathbf{A}}\odot\mathbf{B}),\qquad \mathbf{A}^{-1}\equiv\sum_k\hat{\mathbf{V}}_k\otimes\boldsymbol{\Sigma}_k,$$

$$\mathbf{B}\equiv\sum_k\left(\boldsymbol{\mu}_k\hat{\mathbf{m}}_k^\top+\boldsymbol{\Xi}_k\odot\sqrt{\hat{\mathbf{Q}}_k\otimes\boldsymbol{\Sigma}_k}\right).$$

The collection of parameters $(\mathbf{Q}_k,\mathbf{m}_k,\mathbf{V}_k,\hat{\mathbf{Q}}_k,\hat{\mathbf{m}}_k,\hat{\mathbf{V}}_k)_{k\in[K]}$ is given by the fixed point of the following self-consistent equations:

$$\begin{cases}\mathbf{Q}_k=\frac{1}{d}\mathbb{E}_\Xi[\mathbf{G}\boldsymbol{\Sigma}_k\mathbf{G}^\top]\\ \mathbf{m}_k=\frac{1}{\sqrt{d}}\mathbb{E}_\Xi[\mathbf{G}\boldsymbol{\mu}_k]\\ \mathbf{V}_k=\frac{1}{d}\mathbb{E}_\Xi\left[\left(\mathbf{G}\odot(\hat{\mathbf{Q}}_k\otimes\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\odot(\mathbf{I}_K\otimes\boldsymbol{\Sigma}_k)\right)\boldsymbol{\Xi}_k^\top\right]\end{cases}$$

$$\begin{cases}\hat{\mathbf{Q}}_k=\alpha\rho_k\mathbb{E}_\xi\left[\mathbf{f}_k\mathbf{f}_k^\top\right]\\ \hat{\mathbf{V}}_k=-\alpha\rho_k\mathbf{Q}_k^{-\frac{1}{2}}\mathbb{E}_\xi\left[\mathbf{f}_k\boldsymbol{\xi}^\top\right]\\ \hat{\mathbf{m}}_k=\alpha\rho_k\mathbb{E}_\xi\left[\mathbf{f}_k\right]\end{cases}$$

Moreover

- $\mathbf{f}_k\equiv\mathbf{V}_k^{-1}(\mathbf{h}_k-\boldsymbol{\omega}_k)$;
- $\mathbf{b}^\star$ is such that $\sum_k\rho_k\mathbb{E}_\xi\left[\mathbf{V}_k\mathbf{f}_k\right]=\mathbf{0}$.

Important remarks

- Very generic statement.
- Proximal operators are easy to compute, summarize the effect of loss and penalty.
- Greatly simplifies with assumptions on covariances, separability of functions...
- In most cases reduces to low/one dimensional statement.

Sketch of proof

We use an approximate message passing iteration (AMP)

- AMP are iterations with exact asymptotics at each time step: the state evolution equations.
- Design an AMP sequence such that its fixed point matches the solution to Eq.(2)
- Find a converging trajectory (convexity is helpful).
- Use the fixed point of the state evolution equations.

Here a specific, block operating ("spatially coupled") AMP is used to handle the block covariance structure

Training and generalization error

- Average training loss

$$\epsilon_\ell=\frac{1}{n}\sum_{\nu=1}^{n}\ell\left(\mathbf{y}^\nu,\frac{\mathbf{W}^\star\mathbf{x}^\nu}{\sqrt{d}}+\mathbf{b}^\star\right)\xrightarrow[\alpha=n/d]{n\to+\infty}\sum_{k=1}^{K}\rho_k\mathbb{E}_\xi[\ell(\mathbf{e}_k,\mathbf{h}_k)].$$

- Average training error ϵ_t and generalisation error ϵ_g:

$$\epsilon_t=\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{I}\left[\mathbf{y}^\nu\neq\hat{\mathbf{y}}\left(\frac{\mathbf{W}^\star\mathbf{x}^\nu}{\sqrt{d}}+\mathbf{b}^\star\right)\right]\xrightarrow[\alpha=n/d]{n\to+\infty}1-\sum_{k=1}^{K}\rho_k\mathbb{E}_\xi[\hat{y}_k(\mathbf{h}_k)],$$

$$\epsilon_g=\mathbb{E}_*\left[\mathbb{I}\left[\mathbf{y}^*\neq\hat{\mathbf{y}}\left(\frac{\mathbf{W}^\star\mathbf{x}^*}{\sqrt{d}}+\mathbf{b}^\star\right)\right]\right]\xrightarrow[\alpha=n/d]{n\to+\infty}1-\sum_{k=1}^{K}\rho_k\mathbb{E}_\xi[\hat{y}_k(\mathbf{h}_k)],$$

where $(\mathbf{x}^*,\mathbf{y}^*)$ is a new sample from Eq. (1), and $\hat{y}_k(\mathbf{x})=\mathbb{I}(\max_\kappa x_\kappa=x_k)$.

Application: synthetic dataset

- Multiclass logistic regression with ridge penalty.
- Effect of sample complexity, number of clusters and regularisation strength is studied.
- Recover and extend previous results on separability transition.

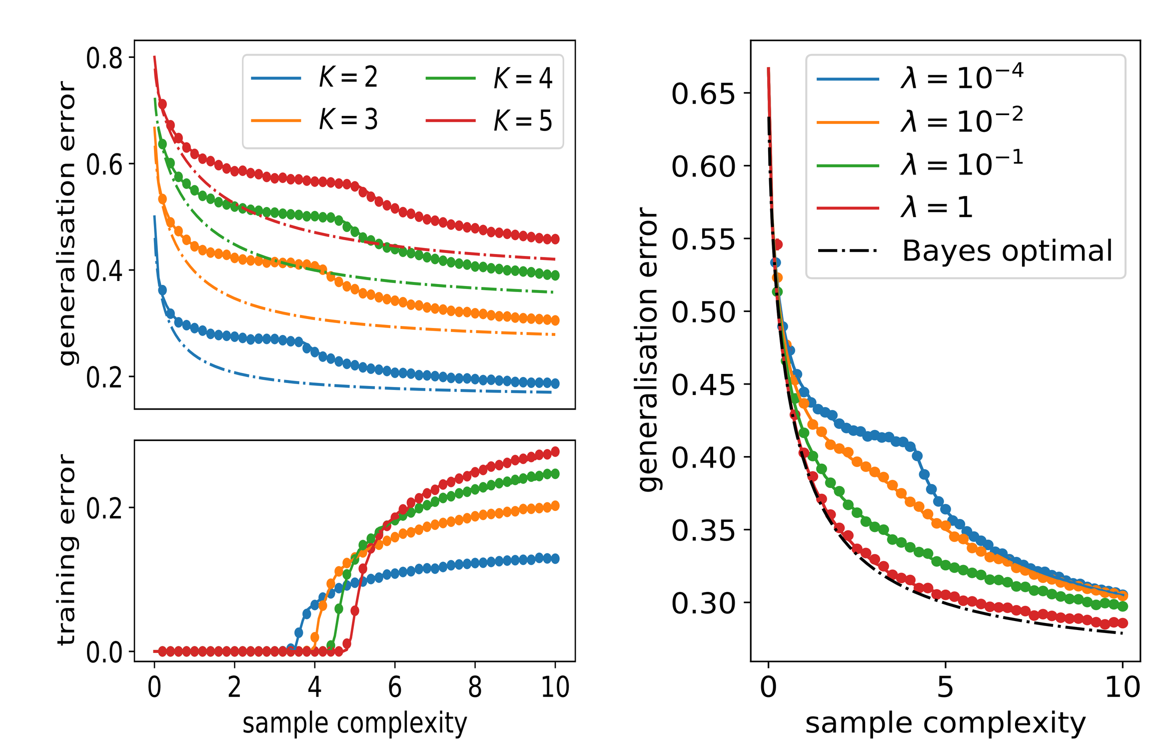

Figure: Gaussian means and $\boldsymbol{\Sigma}_k\equiv\boldsymbol{\Sigma}=1/2\mathbf{I}_d$. (**Left**) Generalisation error ϵ_g (*top*) and training error ϵ_t (*bottom*) as function of α at $\lambda=10^{-4}$. Theoretical predictions (full lines) are compared with the results of numerical experiments (dots). Dash-dotted lines of the corresponding color represent, for comparison, the Bayes-optimal error. (**Right**) Dependence of the generalisation error on the regularization λ for $K=3$ and $\Sigma=1/2\mathbf{I}_d$, $\rho_k=1/K$

Application: correlated sparse mixture

- model with strong and weak features
- sparse means $\boldsymbol{\mu}_k\in\mathbb{R}^d$ with sparsity $\rho\in[0,1]$.
- diagonal covariance $\Sigma_{ij}=\sigma_i\delta_{ij}$, with $\sigma_i\in\{\Delta_1,\Delta_2\}$.
- high/low σ_i aligned with non-zero components of means, i.e.

$$P(\boldsymbol{\mu},\boldsymbol{\sigma})=\prod_{i=1}^{d}\left\{\rho\mathcal{N}(\mu_i|0,1)\delta_{\sigma_i,\Delta_1}+(1-\rho)\delta_{\mu_i}\delta_{\sigma_i,\Delta_2}\right\}. \tag{4}$$

Binary classification on this model, with ℓ_1/ℓ_2 penalty

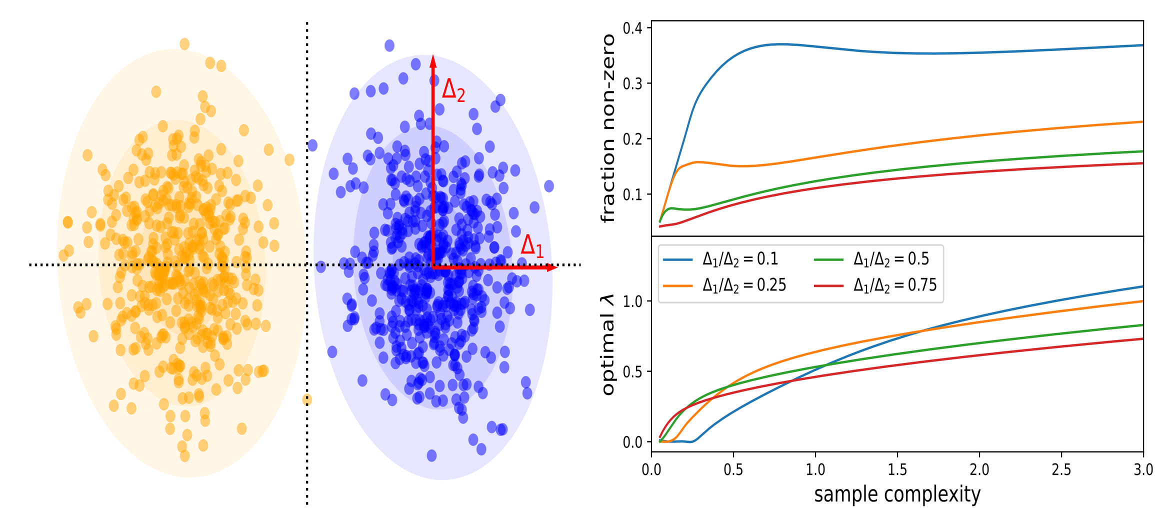

Figure: Two-dimensional projection of the Gaussian mixture introduced via Eq. (4) in which the sparse directions of the means are correlated with the weak/strong directions in the data. (**Right**) Fraction of non-zero elements of the lasso estimator (*top*) and optimal regularisation strength (*bottom*) as a function of $\alpha=n/d$, for varying Δ_1/Δ_2, at fixed sparsity $\rho=0.1$.

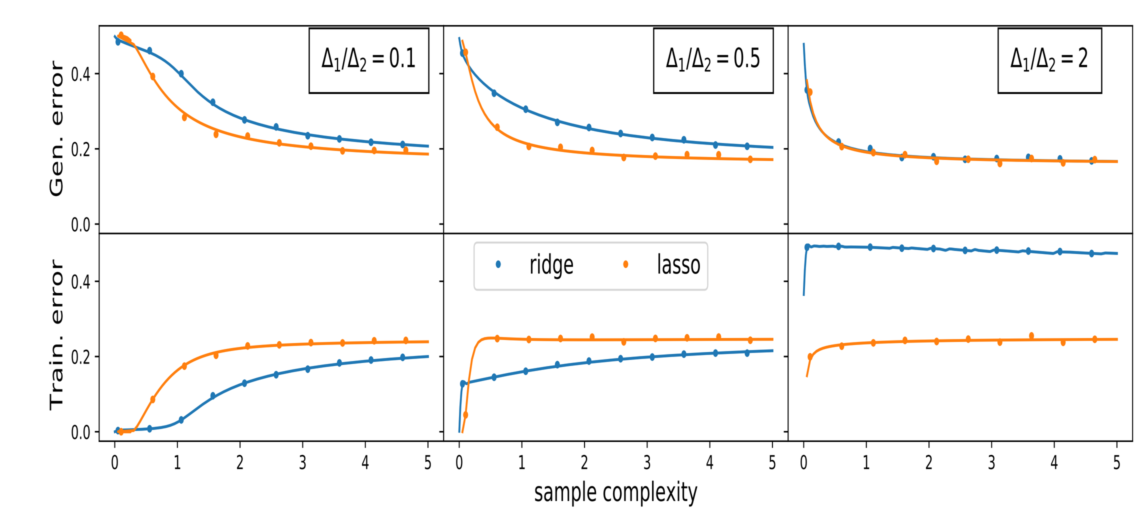

Figure: Performance of ridge (blue) and lasso (orange) estimators at optimal regularisation strength $\lambda^\star$ and for different values of Δ_1/Δ_2. Full lines denote the theoretical prediction, and dots denote finite instance simulations with $d=1000$. Above a certain sample complexity α, we can identify two regimes: a) a $\Delta_1/\Delta_2\lesssim1$ regime in which the ℓ_1 penalty improves significantly over ℓ_2; b) a $\Delta_1/\Delta_2\gtrsim1$ regime in which the performance is similar.

Application: real datasets

- Binary classification with the logistic loss on MNIST/Fashion-MNIST.
- Comparison between the estimator obtained with real data and a synthetic (Gaussian) approximation with matching covariances.
- **Real learning curve is captured by the synthetic one** feeded with real-data covariance matrices.

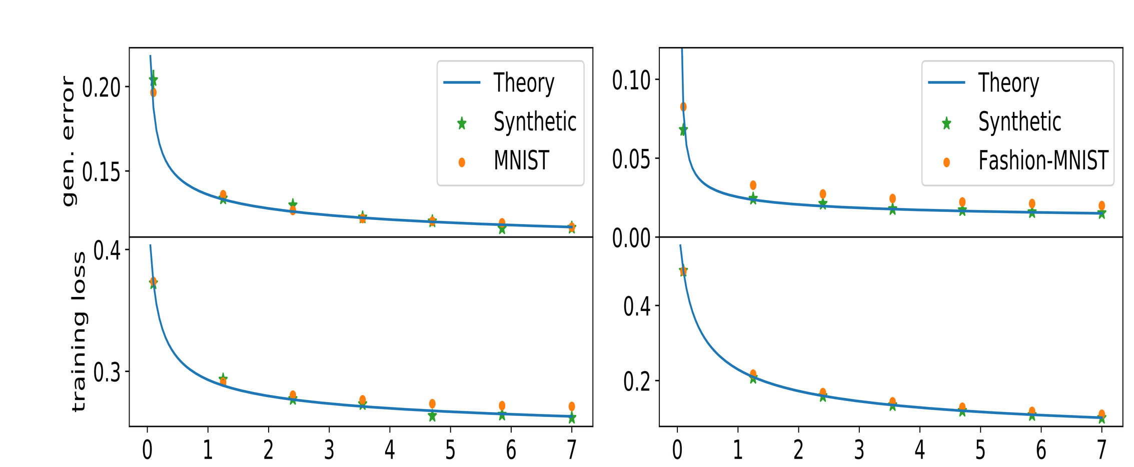

Figure: Generalisation error and training loss on MNIST with $\lambda=0.05$ (**left**) and on Fashion-MNIST with $\lambda=1$ (**right**)

Contact: cedric.gerbelot@ens.fr