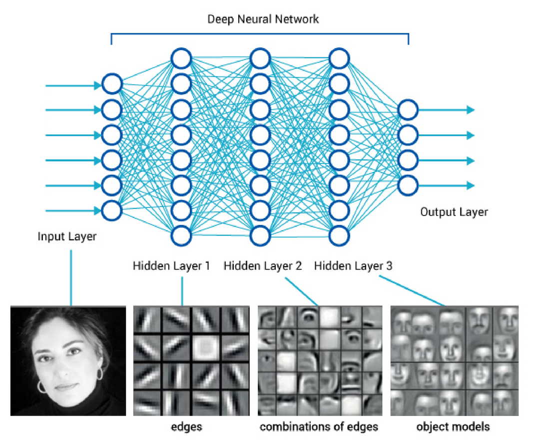
Learning Gaussian Mixtures with Generalised Linear Models : Precise Asymptotics in High-dimension

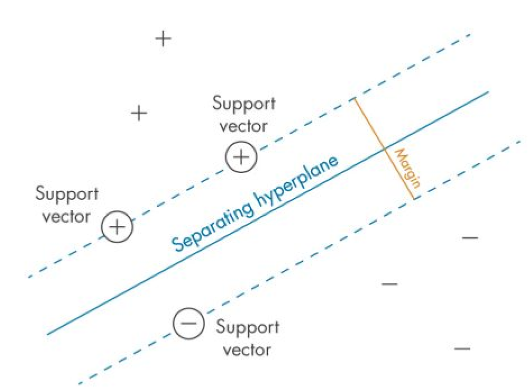
Bruno Loureiro[1], Gabriele Sicuro[2], **Cédric Gerbelot**[3], Alessandro Pacco[1], Florent Krzakala[1] and Lenka Zdeborova[4]
[1]EPFL IdePhics Lab., [2]King's College Math. Dept, [3]ENS Physics Dept., [4] EPFL SPOC Lab.

Motivation

Tractable models for deep learning



What we would like to understand



What we actually have theory for

- no quantitative theory for deep learning yet
- exact understanding of simple models (e.g., linear SVM)
- simple models sometimes capture deep learning phenomenology

Need for tractable, realistic surrogate models for deep learning

Ingredients for a surrogate model

- **learning architecture** (ridge regression, support vector machine, ...)
- **data/feature model** (i.i.d. Gaussian, non-diagonal covariance ...)
- training algorithm (not here, direct focus on estimators)

Examples

- Instances of ridge regression with i.i.d. coordinates captures the so-called **double descent** [BHMM19] phenomenon
- **Gaussian mixtures** are appropriate models for GAN data [SLTC20]
- **Convex Generalized Linear Models** (GLM) with correlated Gaussian designs [LGC+21] capture a wide range of single task regression problems, with structured data/feature maps (kernels, GAN data, ...)

Objective

Can we have a realistic benchmark for multiclass classification problems ?

Contributions

- **study classification of a high-dimensional K-Gaussian mixture with a convex Generalized Linear Model (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 $\{\mathcal{C}_k\}_{1\leqslant k\leqslant K}$:

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

- 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 $\mathbf{y}_k$ are **one-hot-encoded**, i.e. cluster $\mathcal{C}_k$ is denoted by the k-th basis vector in $\mathbb{R}^K$
- sample n pairs $(\mathbf{x}^\nu,\mathbf{y}^\nu)$ from Eq.(1)
- design matrix denoted $\mathbf{X}\in\mathbb{R}^{n\times d}$

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

The learning method : a convex GLM

Estimator obtained by minimising the empirical risk:

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

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

- $\mathbf{W}\in\mathbb{R}^{K\times d},\mathbf{b}\in\mathbb{R}^K$ are the weights and bias to be learned
- l is a convex loss function
- r is a convex regularisation function with strength $\lambda\in\mathbb{R}$

Examples : least-squares, logistic loss, ℓ_2 or ℓ_1 penalty, ...

Goal : asymptotic properties of W^*

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

In particular, we are interested in:

- the average training loss

$$\epsilon_\ell=\frac{1}{n}\sum_{\nu=1}^{n}\ell\left(\boldsymbol{y}^\nu,\frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}}+\boldsymbol{b}^\star\right), \qquad (4)$$

- the average training error ϵ_t and generalisation error ϵ_g:

$$\epsilon_t=\frac{1}{n}\sum_{\nu=1}^{n}\mathbb{I}\left[\boldsymbol{y}^\nu\neq\hat{\boldsymbol{y}}\left(\frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}}+\boldsymbol{b}^\star\right)\right], \qquad (5)$$

$$\epsilon_g=\mathbb{E}_{(\boldsymbol{x}^{\text{new}},\boldsymbol{y}^{\text{new}})}\left[\mathbb{I}\left[\boldsymbol{y}^{\text{new}}\neq\hat{\boldsymbol{y}}\left(\frac{\boldsymbol{W}^\star\boldsymbol{x}^{\text{new}}}{\sqrt{d}}+\boldsymbol{b}^\star\right)\right]\right],$$

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

Useful notation

Suppose that the matrix $\boldsymbol{G}=(G_{ki})_{ki}\in\mathbb{R}^{K\times d}$ is given, alongside the four-index tensor $\mathbf{A}=(A_{ki\,k'i'})_{ki\,k'i'}\in\mathbb{R}^{K\times d}\otimes\mathbb{R}^{K\times d}$. We will use the notation $\boldsymbol{G}\odot\mathbf{A}=\sum_{ki}G_{ki}A_{ki\,k'i'}\in\mathbb{R}^{K\times d}$. Similarly, given a four-index tensor $\mathbf{A}$, we will define $\sqrt{\mathbf{A}}$ as the tensor such that $\mathbf{A}=\sqrt{\mathbf{A}}\odot\sqrt{\mathbf{A}}$.

Main result : exact asymptotics [LSG+21]

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

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

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

where we have introduced the proximal for the loss:

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

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

$$\boldsymbol{G}=\mathbf{A}^{\frac{1}{2}}\odot\operatorname{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}}\odot\boldsymbol{B}),\qquad\mathbf{A}^{-1}\equiv\sum_k\hat{\boldsymbol{V}}_k\otimes\boldsymbol{\Sigma}_k,$$

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

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

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

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

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

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

Corollary : training and generalisation error

The training loss, the training error and the generalisation error are given by

$$\epsilon_\ell=\sum_{k=1}^{K}\rho_k\mathbb{E}_{\boldsymbol{\xi}}[\ell(\boldsymbol{e}_k,\boldsymbol{h}_k)], \qquad (6)$$

$$\epsilon_t=1-\sum_{k=1}^{K}\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{h}_k)\right], \qquad (7)$$

$$\epsilon_g=1-\sum_{k=1}^{K}\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{\omega}_k)\right]. \qquad (8)$$

Results on a synthetic dataset

- multiclass logistic regression with ridge penalty
- effect of sample complexity, number of clusters and regularisation strength
- recover and extend previous result on separability transition [MKL+20]

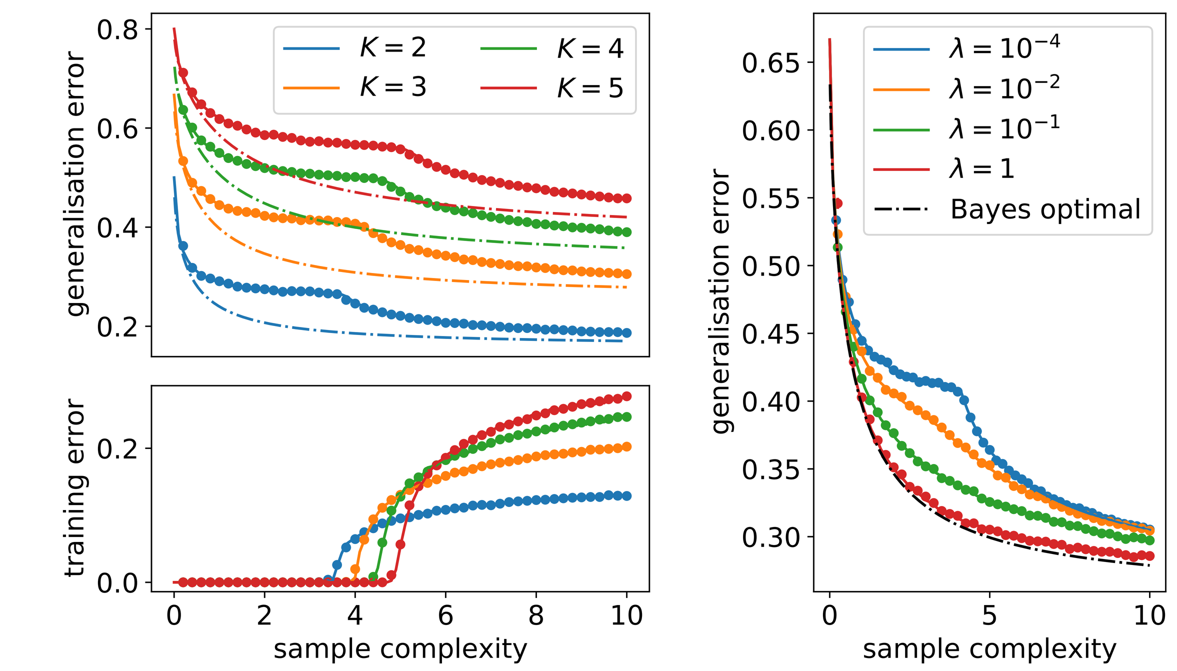


Figure: Gaussian means and $\Sigma_k\equiv\Sigma=1/2I_d$. (**Left**) Generalisation error ϵ_g (*top*) and training error ϵ_t (*bottom*) as function of α at $\lambda=10^{-4}$. Theoretical predictions (full lines) are compared with the results of numerical experiments (dots). Dash-dotted lines of the corresponding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

Results on a real dataset

- 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
- **the real learning curve is captured by the synthetic one**

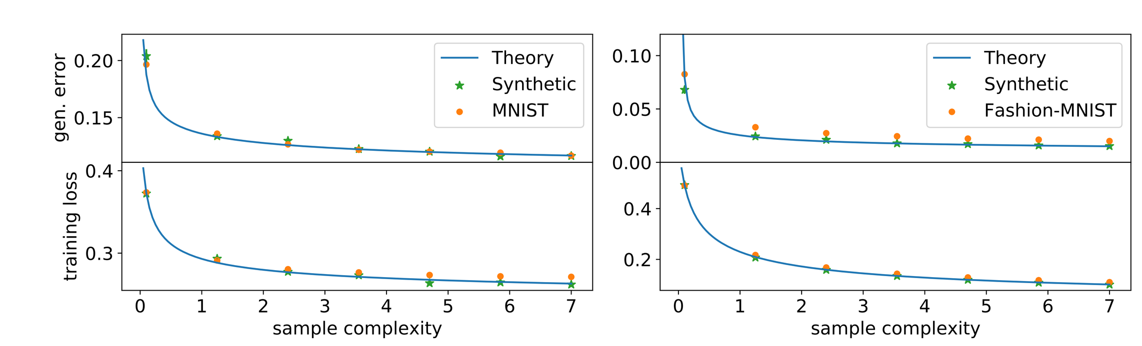


Figure: Generalisation error and training loss on MNIST with $\lambda=0.05$ (**left**) and on Fashion-MNIST with $\lambda=1$ (**right**)

References

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.

[LGC+21] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová.
Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model.
arXiv preprint arXiv:2102.08127, 2021.

[LSG+21] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová.
Learning gaussian mixtures with generalised linear models: Precise asymptotics in high-dimensions.
arXiv preprint arXiv:2106.03791, 2021.

[MKL+20] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova.
The role of regularization in classification of high-dimensional noisy Gaussian mixture.
In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6874–6883. PMLR, 13–18 Jul 2020.

[SLTC20] 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 Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8573–8582. PMLR, 13–18 Jul 2020.

Contact: cedric.gerbelot@ens.fr