The Breakdown of Gaussian Universality in Classification of High-Dimensional Linear Factor Mixtures

Xiaoyi MAI[⋆], Zhenyu LIAO[†]

[⋆]University of Toulouse - Jean Jaurès, Toulouse Mathematics Institute, France
[†]EIC Huazhong University of Science & Technology, China

ICLR, 2025

Supervised ML: building a classifier from a training set of $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with feature vectors $\mathbf{x}_i \in \mathbb{R}^p$ and class labels $y_i \in \{-1, 1\}$.

Supervised ML: building a classifier from a training set of $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with feature vectors $\mathbf{x}_i \in \mathbb{R}^p$ and class labels $y_i \in \{-1, 1\}$.

Empirical Risk Minimisation (ERM) : many supervised algorithms (e.g., SVM, LR, ANN with pretrained hidden layers) can be summarized by the following ERM

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i\right)}_{\text{Empirical loss}} + \underbrace{\lambda \|\boldsymbol{\beta}\|^2}_{\text{regularization term}} .$$

Supervised ML: building a classifier from a training set of $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with feature vectors $\mathbf{x}_i \in \mathbb{R}^p$ and class labels $y_i \in \{-1, 1\}$.

Empirical Risk Minimisation (ERM) : many supervised algorithms (e.g., SVM, LR, ANN with pretrained hidden layers) can be summarized by the following ERM

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i \right)}_{\text{Empirical loss}} + \underbrace{\lambda \|\boldsymbol{\beta}\|^2}_{\text{regularization term}} .$$

- Popular choices of ℓ: square loss for least squares method, logistic loss for LR, hinge loss for SVM, etc.

Modern ML: comparably numerous features and samples, i.e., $p \sim n \gg 1$.

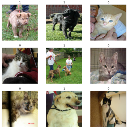


Figure: Image Classification

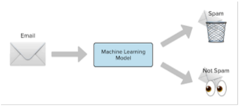


Figure: Spam Detection

Introduction: Performance Analysis in the Big Data Regime

Modern ML: comparably numerous features and samples, i.e., $p \sim n \gg 1$.

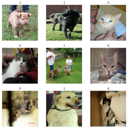


Figure: Image Classification

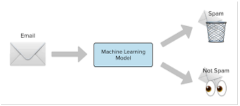


Figure: Spam Detection

Analysis of Modern ML:

- Complications of $n \sim p$:
  - Performance sensitive to the sample size n/p, and the hyperparameters ℓ, λ.
  - Random classifier $\hat{\boldsymbol{\beta}}$ depending on $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in a non-linear and implicit manner.

Introduction: Performance Analysis in the Big Data Regime

Modern ML: comparably numerous features and samples, i.e., $p \sim n \gg 1$.

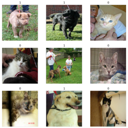


Figure: Image Classification

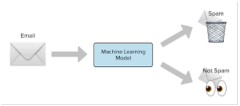


Figure: Spam Detection

Analysis of Modern ML:

- Complications of $n \sim p$:
  - Performance sensitive to the sample size n/p, and the hyperparameters ℓ, λ.
  - Random classifier $\hat{\boldsymbol{\beta}}$ depending on $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in a non-linear and implicit manner.
- Technical conveniences of $n, p \gg 1$:
  - Convergence of performance curve as a function of sample ratio n/p.
  - Gaussian universality (GU) induced by Central Limit Theorem (e.g., $\hat{\boldsymbol{\beta}} = \mathrm{avg}(y_i \mathbf{x}_i)$).

Implication of GU : universal performance depending on the first two moments.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^{\mathsf{T}}\mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^\top \mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^{\mathsf{T}}\mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

- Sharp performance analysis of ERM under linear factor mixture models (LFMM).
  - The development of leave-one-out approach to characterize performance beyond GU.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^\top \mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

- Sharp performance analysis of ERM under linear factor mixture models (LFMM).
  - The development of leave-one-out approach to characterize performance beyond GU.
- Condition of GU depending on LFMM. .
  - Explicit condition allowing insight into the impact of data structure.

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^{\mathsf{T}}\mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

- Sharp performance analysis of ERM under linear factor mixture models (LFMM).
  - The development of leave-one-out approach to characterize performance beyond GU.
- Condition of GU depending on LFMM. .
  - Explicit condition allowing insight into the impact of data structure.
- Implication of GU breakdown on the optimal choice of loss.

Introduction: Gaussian Universality of Empirical Risk Minimization

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^\mathsf{T}\mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

- Sharp performance analysis of ERM under linear factor mixture models (LFMM).
  - The development of leave-one-out approach to characterize performance beyond GU.
- Condition of GU depending on LFMM. .
  - Explicit condition allowing insight into the impact of data structure.
- Implication of GU breakdown on the optimal choice of loss.
  - Consequence: square loss no longer optimal as in GMM [TPT20; ML20].

Introduction: Gaussian Universality of Empirical Risk Minimization

Implication of GU : universal performance depending on the first two moments.

- Advantage: use of tools requiring data Gaussianity , e.g., CGMT, AMP, replica.
- Limitation: no insight into the learning of higher order data statistics.

Previous results on GU of ERM: proofs of GU under the assumptions of pointwise normality $\hat{\beta}^{\mathsf{T}}\mathbf{x}$ on the ERM solution $\hat{\beta}$ [MS22; Dan+24], not verifiable from the data distribution.

Our contributions [ML25]:

- Sharp performance analysis of ERM under linear factor mixture models (LFMM).
  - The development of leave-one-out approach to characterize performance beyond GU.
- Condition of GU depending on LFMM. .
  - Explicit condition allowing insight into the impact of data structure.
- Implication of GU breakdown on the optimal choice of loss.
  - Consequence: square loss no longer optimal as in GMM [TPT20; ML20].
  - Sharp results on the advantage of non-square losses in learning high-order data statistics.

Definition (Linear Factor Mixture Model (LFMM))

A data-label pair $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ with class label $y \in \{\pm 1\}$ is said to follow a linear factor mixture model if $\mathbf{x} \in \mathbb{R}^p$ is the linear combination of p **factors** $z_1, \ldots, z_p$

$$\mathbf{x} = \sum_{k=1}^{p} z_k \mathbf{v}_k = \sum_{k=1}^{p} (y s_k + e_k) \mathbf{v}_k, \tag{1}$$

for linearly independent deterministic $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ and standardized noises $e_1, \ldots, e_p \in \mathbb{R}$ independent of y with bounded fourth moments.

- q **informative factors** $z_1, \ldots, z_q$ with **deterministic signals** $s_k > 0$, $k \in \{1, \ldots, q\}$;

Definition (Linear Factor Mixture Model (LFMM))

A data-label pair $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ with class label $y \in \{\pm 1\}$ is said to follow a linear factor mixture model if $\mathbf{x} \in \mathbb{R}^p$ is the linear combination of p **factors** $z_1, \ldots, z_p$

$$\mathbf{x} = \sum_{k=1}^{p} z_k \mathbf{v}_k = \sum_{k=1}^{p} (y s_k + e_k) \mathbf{v}_k, \tag{1}$$

for linearly independent deterministic $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ and standardized noises $e_1, \ldots, e_p \in \mathbb{R}$ independent of y with bounded fourth moments.

- q **informative factors** $z_1, \ldots, z_q$ with **deterministic signals** $s_k > 0$, $k \in \{1, \ldots, q\}$;
- $p - q$ **noise factors** $z_{q+1}, \ldots, z_p$ with $s_k = 0$, $k \in \{q+1, \ldots, p\}$.

Definition (Linear Factor Mixture Model (LFMM))

A data-label pair $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ with class label $y \in \{\pm 1\}$ is said to follow a linear factor mixture model if $\mathbf{x} \in \mathbb{R}^p$ is the linear combination of p **factors** $z_1, \ldots, z_p$

$$\mathbf{x} = \sum_{k=1}^{p} z_k \mathbf{v}_k = \sum_{k=1}^{p} (y s_k + e_k)\mathbf{v}_k, \tag{1}$$

for linearly independent deterministic $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ and standardized noises $e_1, \ldots, e_p \in \mathbb{R}$ independent of y with bounded fourth moments.

- q **informative factors** $z_1, \ldots, z_q$ with **deterministic signals** $s_k > 0$, $k \in \{1, \ldots, q\}$;
- $p - q$ **noise factors** $z_{q+1}, \ldots, z_p$ with $s_k = 0$, $k \in \{q+1, \ldots, p\}$.
- $\mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ orthogonal to $\mathrm{Span}\{\mathbf{v}_{q+1}, \ldots, \mathbf{v}_p\}$

Definition (Linear Factor Mixture Model (LFMM))

A data-label pair $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ with class label $y \in \{\pm 1\}$ is said to follow a linear factor mixture model if $\mathbf{x} \in \mathbb{R}^p$ is the linear combination of p **factors** $z_1, \ldots, z_p$

$$\mathbf{x} = \sum_{k=1}^{p} z_k \mathbf{v}_k = \sum_{k=1}^{p} (y s_k + e_k) \mathbf{v}_k, \tag{1}$$

for linearly independent deterministic $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ and standardized noises $e_1, \ldots, e_p \in \mathbb{R}$ independent of y with bounded fourth moments.

- q **informative factors** $z_1, \ldots, z_q$ with **deterministic signals** $s_k > 0$, $k \in \{1, \ldots, q\}$;
- $p - q$ **noise factors** $z_{q+1}, \ldots, z_p$ with $s_k = 0$, $k \in \{q+1, \ldots, p\}$.
- $\mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ orthogonal to $\mathrm{Span}\{\mathbf{v}_{q+1}, \ldots, \mathbf{v}_p\}$
- class-conditional means and covariances of $\mathbf{x}$:

$$\mathbb{E}[\mathbf{x}|y] = y\boldsymbol{\mu}, \quad \mathrm{Cov}[\mathbf{x}|y] = \boldsymbol{\Sigma} \tag{2}$$

with $\boldsymbol{\mu} = \sum_{k=1}^{p} s_k \mathbf{v}_k$ and $\boldsymbol{\Sigma} = \sum_{k=1}^{p} \mathbf{v}_k \mathbf{v}_k^\mathsf{T}$.

Theorem (Asymptotic distribution of predicted scores under LFMM)

For ERM classifier $\hat{\beta}$ obtained on $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of size n drawn i.i.d. from an LFMM, we have that, for any bounded Lipschitz function $f \colon \mathbb{R} \to \mathbb{R}$,

$$\textit{Testing score:} \quad \boxed{\mathbb{E}\left[f(\hat{\boldsymbol{\beta}}^\mathsf{T} \boldsymbol{\nu})\right] - \mathbb{E}\left[f(\tilde{\boldsymbol{\beta}}^\mathsf{T} \boldsymbol{\nu})\right] \to 0}, \quad \forall \textit{ deterministic } \boldsymbol{\nu} \in \mathbb{R}^p$$

$$\textit{Training score:} \quad \boxed{\mathbb{E}[f(\hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i)] - \mathbb{E}[f(\operatorname{prox}_{\kappa, \ell(\cdot, y_i)}(\tilde{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i))] \to 0}, \quad \forall i \in \{1, \ldots, n\},$$

where

$$\tilde{\boldsymbol{\beta}} = (\lambda \mathbf{I}_p + \theta \boldsymbol{\Sigma})^{-1} \left(\eta \boldsymbol{\mu} + \sum_{k=1}^q \omega_k \mathbf{v}_k + \gamma \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u} \right), \tag{3}$$

for Gaussian vector $\mathbf{u} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p / n)$ independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and constants $\theta, \eta, \gamma, \omega_1, \ldots, \omega_q$ determined by a (known) self-consistent system of equations

$$[\theta, \eta, \gamma, \omega_1, \ldots, \omega_q] = G_{\ell, \lambda, n/p, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{v}_1, \ldots, \mathbf{v}_q, \mathcal{D}_{(z_1, \ldots, z_q)}} ([\theta, \eta, \gamma, \omega_1, \ldots, \omega_q]) .$$

Definition (Equivalent Gaussian mixture model (Equivalent GMM))

For an LFMM $\mathcal{D}_{(\mathbf{x},y)}$, we define its equivalent Gaussian mixture model $\mathcal{D}_{(\mathbf{g},y)}$ as the GMM with the same class-conditional means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$ as the LFMM in (2):

$$\mathbf{g} \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4}$$

We denote $\hat{\beta}^{\mathbf{g}}$ the ERM solution obtained on n i.i.d. $(\mathbf{g}_1, y_1), \ldots, (\mathbf{g}_n, y_n) \sim \mathcal{D}_{(\mathbf{g},y)}$.

Definition (Equivalent Gaussian mixture model (Equivalent GMM))

For an LFMM $\mathcal{D}_{(\mathbf{x},y)}$, we define its equivalent Gaussian mixture model $\mathcal{D}_{(\mathbf{g},y)}$ as the GMM with the same class-conditional means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$ as the LFMM in (2):

$$\mathbf{g} \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4}$$

We denote $\hat{\boldsymbol{\beta}}^{\mathbf{g}}$ the ERM solution obtained on n i.i.d. $(\mathbf{g}_1, y_1), \ldots, (\mathbf{g}_n, y_n) \sim \mathcal{D}_{(\mathbf{g},y)}$.

Definition (Gaussian universality under LFMM)

For an ERM solution $\hat{\boldsymbol{\beta}}$ on LFMM $\mathcal{D}_{(\mathbf{x},y)}$ and $\hat{\boldsymbol{\beta}}^{\mathbf{g}}$ on the equivalent GMM, we say Gaussian universality holds if

$$\Pr(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}} > 0) \simeq \Pr(y_i \mathbf{g}_i^\mathsf{T} \hat{\boldsymbol{\beta}}^{\mathbf{g}} > 0)$$
$$\Pr(y' \mathbf{x}'^\mathsf{T} \hat{\boldsymbol{\beta}} > 0) \simeq \Pr(y' \mathbf{g}'^\mathsf{T} \hat{\boldsymbol{\beta}}^{\mathbf{g}} > 0)$$

for $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$ independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and $(\mathbf{g}', y') \sim \mathcal{D}_{(\mathbf{g},y)}$ independent of $\{(\mathbf{g}_i, y_i)\}_{i=1}^n$

Corollary (Conditional Gaussian universality under LFMM)

Uner LFMM, the Gaussian universality of ERM holds if and only if the informative factors $z_1, \ldots, z_q$ of LFMM are class-conditional Gaussian.

Corollary (Conditional Gaussian universality under LFMM)

Uner LFMM, the Gaussian universality of ERM holds if and only if the informative factors $z_1, \ldots, z_q$ of LFMM are class-conditional Gaussian.

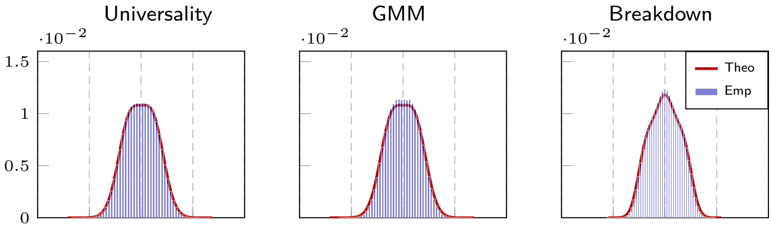


Figure: Theoretical and empirical distribution of predicted scores $\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}'$ for some fresh test data $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$ independent of $\hat{\boldsymbol{\beta}}$. The theoretical probability densities (**red**), and the empirical histograms (**blue**) are the values of $\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}'$ over 10^6 independent copies of $\mathbf{x}'$, for three different LFMMs with $n = 600$, $p = 200$, $\rho = 0.5$, $s = [\sqrt{2}; \mathbf{0}_{p-1}]$ (so that $q = 1$), and Haar distributed $\mathbf{V}$. **Left**: normal e_1 and uniformly distributed $e_2, \ldots, e_p$; **Middle**: normal $e_1, \ldots, e_p$; **Right**: uniformly distributed e_1, and normal $e_2, \ldots, e_p$.

Performance under Gaussian Universality Breakdown

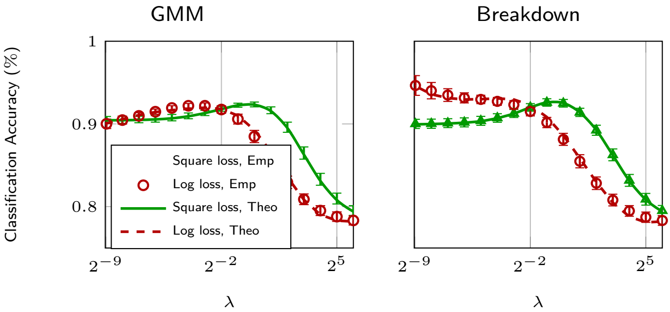


Figure: Empirical classification accuracy of $\hat{\mathbf{w}}_{\ell,\lambda}$ averaged over 100 trials with a width of ± 1 standard deviation, versus theoretical curve given by the square loss and the logistic loss on $n = 800$ training samples. **Left**: GMM under with $p = 200$, $\rho = 0.5$, $s = [1, 5; 0.5; \mathbf{0}_{p-2}]$ (so that $q = 2$), and $\mathbf{V} = \mathrm{diag}(2, \mathbf{1}_{p-1})\mathbf{H}$ with Haar distributed $\mathbf{H}$. **Right**: LFMM identical to the GMM in the left, but with Rademacher e_1.

Performance under Gaussian Universality Breakdown

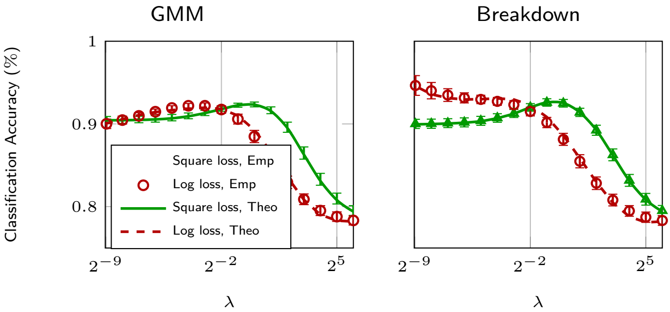
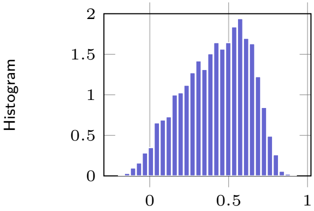


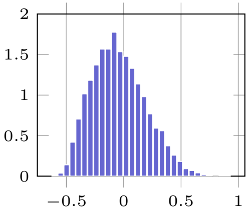
Figure: Empirical classification accuracy of $\hat{\mathbf{w}}_{\ell,\lambda}$ averaged over 100 trials with a width of ± 1 standard deviation, versus theoretical curve given by the square loss and the logistic loss on $n = 800$ training samples. **Left**: GMM under with $p = 200$, $\rho = 0.5$, $s = [1, 5; 0.5; \mathbf{0}_{p-2}]$ (so that $q = 2$), and $\mathbf{V} = \mathrm{diag}(2, \mathbf{1}_{p-1})\mathbf{H}$ with Haar distributed $\mathbf{H}$. **Right**: LFMM identical to the GMM in the left, but with Rademacher e_1.

Remark: under LFMM, the square loss is no longer optima as in GMM [TPT20; ML20].

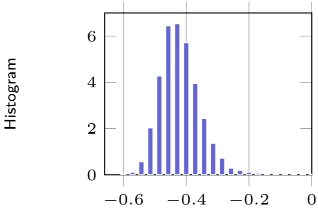
Case 1: Classes 3&7 of Fashion-MNIST data, for which *approximately Gaussian* informative factors (estimated by the principal components in PCA) can be observed.



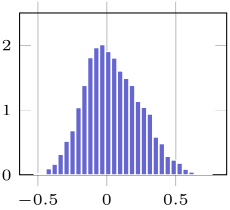
First principal component of Class 3



Second principal component of Class 3

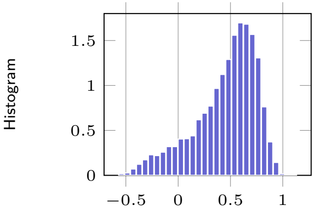


First principal component of Class 7

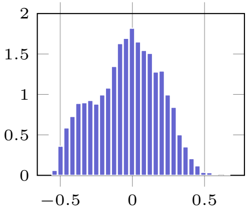


Second principal component of Class 7

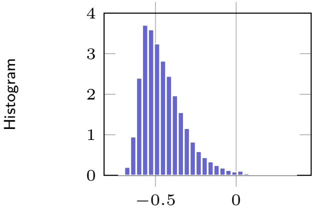
Case 2: Classes 4&5 of Fashion-MNIST data, as an example of *non-Gaussian* informative factors (estimated by the principal components in PCA).



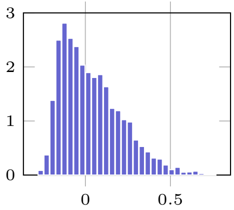
First principal componentr of Class 4



Second principal component of Class 4



First principal component of Class 5



Second principal component of Class 5

Experiments on Real Image Data

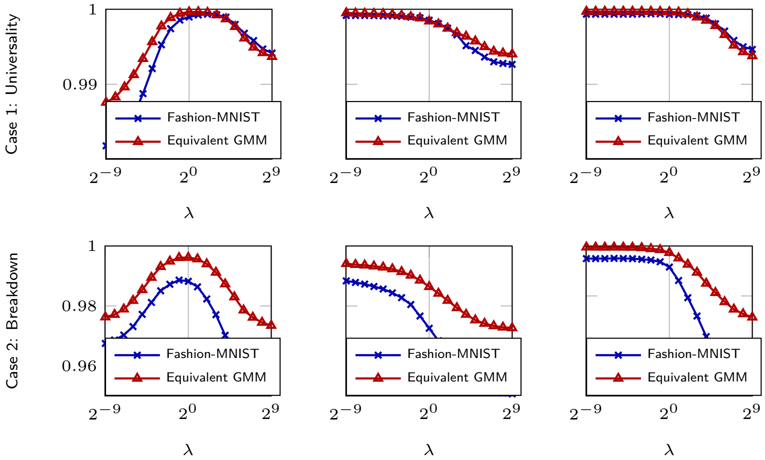


Figure: Classification accuracies as a function of the regularization penalty γ, for Fashion-MNIST data and Equivalent GMM of sample size $n = 512$, with square (**left**), logistic (**middle**), and square hinge (**right**) losses.

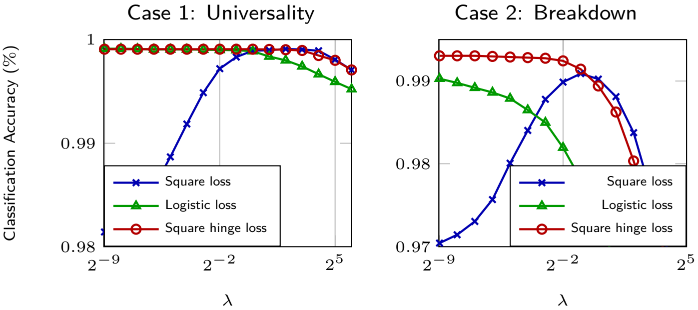


Figure: Classification accuracies as a function of the regularization penalty, for square, logistic, and square hinge loss, on Fashion-MNIST data of sample size $n = 512$. **Left**: Class 3 versus 7, as an example of (close-to) Gaussian information factors. **Right**: Class 4 versus 5, as an example of non-Gaussian information factors.

Take-away

Take-away

- Previous proofs of GU in ERM [MS22; Dan+24] required assumptions on the unknown statistical behaviour of $\hat{\beta}$, not verifiable from the data distribution.

Take-away

- Previous proofs of GU in ERM [MS22; Dan+24] required assumptions on the unknown statistical behaviour of $\hat{\beta}$, not verifiable from the data distribution.

- Our results demonstrate a breakdown of GU in ERM conditioned on the data structure: under LFMM, GU **only holds** in the case of class-condltiional Gaussian informative factors.

Take-away

- Previous proofs of GU in ERM [MS22; Dan+24] required assumptions on the unknown statistical behaviour of $\hat{\beta}$, not verifiable from the data distribution.
- Our results demonstrate a breakdown of GU in ERM conditioned on the data structure: under LFMM, GU **only holds** in the case of class-condltiional Gaussian informative factors.
- Our sharp performance analysis provides insight into the advantages of non-square losses in learning higher order data statistics under the GU breakdown.

Take-away

- Previous proofs of GU in ERM [MS22; Dan+24] required assumptions on the unknown statistical behaviour of $\hat{\beta}$, not verifiable from the data distribution.
- Our results demonstrate a breakdown of GU in ERM conditioned on the data structure: under LFMM, GU **only holds** in the case of class-condltiional Gaussian informative factors.
- Our sharp performance analysis provides insight into the advantages of non-square losses in learning higher order data statistics under the GU breakdown.
- Our results help predict GU on real-word data from the distributions of their principal components.

Take-away

- Previous proofs of GU in ERM [MS22; Dan+24] required assumptions on the unknown statistical behaviour of $\hat{\beta}$, not verifiable from the data distribution.

- Our results demonstrate a breakdown of GU in ERM conditioned on the data structure: under LFMM, GU **only holds** in the case of class-condltiional Gaussian informative factors.

- Our sharp performance analysis provides insight into the advantages of non-square losses in learning higher order data statistics under the GU breakdown.

- Our results help predict GU on real-word data from the distributions of their principal components.

Our Poster Session: Thu 24 Apr 10 a.m. SGT – 12:30 p.m. SGT.

References

[MS22] Andrea Montanari and Basil N Saeed. "Universality of empirical risk minimization". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4310–4312.

[Dan+24] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. "Universality laws for gaussian mixtures in generalized linear models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[ML25] Xiaoyi Mai and Zhenyu Liao. "The Breakdown of Gaussian Universality in Classification of High-dimensional Mixtures". In: *(accepted) 2025 International Conference on Learning Representations (ICLR)*. 2025.

[TPT20] 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)* 00 (2020), pp. 2515–2520.

[ML20] Xiaoyi Mai and Zhenyu Liao. *High Dimensional Classification via Regularized and Unregularized Empirical Risk Minimization: Precise Error and Optimal Loss*. Nov. 2020. arXiv: `1905.13742 [stat]`.

[Bea+13] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. "Optimal M-estimation in high-dimensional regression". In: *Proceedings of the National Academy of Sciences* 110.36 (2013), pp. 14563–14568.