Examples and Counterexamples of Gaussian Universality in Large-dimensional Machine Learning

@ RMTA 2025

Zhenyu Liao

joint work Xiaoyi Mai (IMT, Toulouse, France).

School of Electronic Information and Communications
Huazhong University of Science and Technology

January 13, 2025

Outline

Motivation: Gaussian Universality

- RMT often assumes $\mathbf{x}$ are affine maps $\mathbf{Az} + \mathbf{b}$ of $\mathbf{z} \in \mathbb{R}^p$ with i.i.d. entries
- a lot of results on **large-dimensional universality**: the limiting behavior of statistics remain the same, regardless of the distribution of (the entries of $\mathbf{z}$), so long as the **first few-order moments** are matched
- for both eigenvalues and eigenvector distributions, in both global and local regimes, etc.

Concentrated random vectors [Led05]

For Lipschitz function $f \colon \mathbb{R}^p \mapsto \mathbb{R}$, there exists deterministic $m_f \in \mathbb{R}$

$$\Pr\left(\left| f(\mathbf{x}) - m_f \right| > t \right) \leq \exp(-g(t)), \quad \text{for some strictly increasing function } g(\cdot).$$

- example of concentrated random vectors: multivariate Gaussian, random vector having independent sub-gaussian (e.g., bounded) entries, and **their Lipschitz transformations** through $\phi \colon \mathbb{R}^p \to \mathbb{R}^d$
- for concentrated random vectors:
  - expected resolvent [Sed+20], and nonlinear kernel matrices [STC19] remains asymptotically the same (in spectral norm) as for GMM, connection to M-estimator of scatter [LC22] for elliptical distribution, etc.
  - and many more empirical observations in this vein

An example of Gaussian universal class in DNN

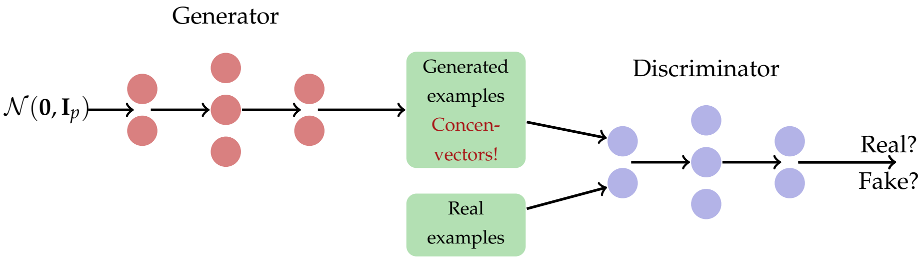


Figure: Illustration of a generative adversarial network (GAN) [Goo+14].

Figure: Images samples generated by BigGAN [BDS19]

Gaussian Universality in large-scale ML

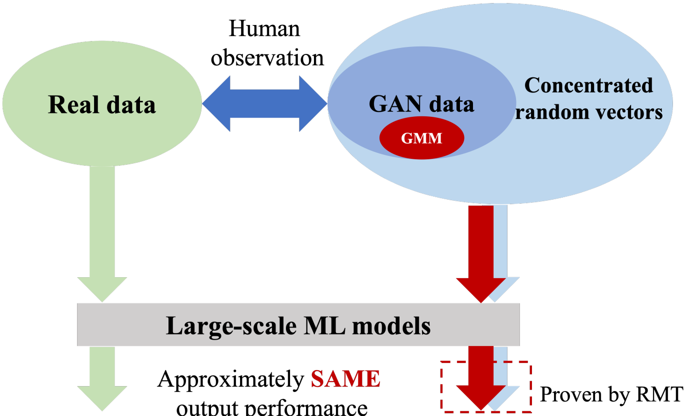


Figure: Concentrated random vectors and Gaussian Universality in large-scale ML

Empirical results on kernel eigenspectra: real data, GAN-generate data, and GMM

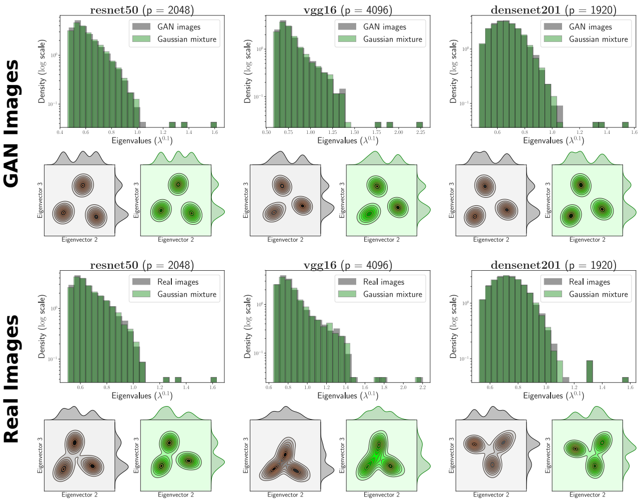


Figure: Figure from [Sed+20]

- eigenvalues and dominant eigenvectors of kernel matrix $\mathbf{K} = \{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$ for CNN features of real and GAN-generated images
- **left** to **right**: ResNet-50 [He+16], VGG-16 [SZ14], and DenseNet-201 [Hua+17]
- comparison between GAN-generated data **(top)** and real data **(bottom)**, empirically on the dataset (**gray**) and on independent GMM with same means and covariances (**green**).

Gaussian Universality and Non-universality in ML

- these universal results are nice, but at the same time discouraging: large-scale ML models, despite being nonlinear, **cannot** "learn" from input data distribution **beyond first few-order moments**
- many conjectures: due to (conditional) CLT? due to a specific choice of (tractable) ML and/or data model? due to the setting of "linear regime" $n \sim p$?

Empirical risk minimization (ERM)

For a set of n training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with feature vectors $\mathbf{x}_i \in \mathbb{R}^p$ and binary labels $y_i \in \{\pm 1\}$, a classifier is trained by minimizing the following ridge-regularized generic empirical risk:

$$\hat{\boldsymbol{\beta}}_{\ell,\lambda} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}, y_i) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2, \tag{1}$$

for some non-negative loss function $\ell \colon \mathbb{R} \times \{\pm 1\} \to \mathbb{R}_+$.

- logistic loss $\ell(\hat{y}, y) = -\ln(1/(1+e^{-y\hat{y}}))$ in logistic regression, square loss $\ell(\hat{y}, y) = (y - \hat{y})^2/2$ for least-squares classifier, and square hinge loss $\ell(\hat{y}, y) = \max\{0, 1 - y\hat{y}\}^2$; but NOT non-smooth in SVM
- **prediction**: fresh $\mathbf{x}'$ with negative scores $\boldsymbol{\beta}^\mathsf{T} \mathbf{x}'$ assigned to class of label $y = -1$, and positive to $y = 1$.

Technical challenges and our results

- **Technical challenge**: **NO** explicit solution, characterization only possible via a **system of (few) equations**
- Some technical approaches:
- convex Gaussian min-max theorem (CGMT) and approximate message passing (AMP): generally assumes Gaussianity
- Lindeberg method as in [HL22]: to establish universal results
- **Limitations**: either **only** holds in the Gaussian (mixture) setting, or can be extrapolated, but to prove **only universal** result
- **QUESTION**: under which condition the ERM solution $\hat{\beta}$ establishes universal **or** non-universal behavior?
- **OUR ANSWER**:
  - focus on not necessarily Gaussian linear factor mixture model (LFMM)
  - develop "**leave-one-out**" approach to **implicit ERM** solution $\hat{\beta}$, flexible enough to characterize non-Gaussian behavior (in LFMM)
  - **implication to ML**: two types of Gaussian universality, when they hold and when they collapse

Linear Factor Mixture Model (LFMM)

We say a data-label pair $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ with class label $y \in \{\pm 1\}$ and class priors $\Pr(y = -1) = \rho$, $\Pr(y = 1) = 1 - \rho$, follows a linear factor mixture model, if $\mathbf{x} \in \mathbb{R}^p$ is the linear combination of p **factors** $z_1, \ldots, z_p$

$$\mathbf{x} = \sum_{k=1}^{p} z_k \mathbf{v}_k = \sum_{k=1}^{p} (y s_k + e_k) \mathbf{v}_k, \quad \mathbf{x} = \mathbf{V}\mathbf{z}, \quad \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \in \mathbb{R}^{p \times p}, \quad \mathbf{z} = [z_1, \ldots, z_p]^\mathsf{T} \in \mathbb{R}^p, \tag{2}$$

for linearly independent deterministic $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ and zero-mean unit variance noises $e_1, \ldots, e_p \in \mathbb{R}$ of symmetric distribution and independent of y. We consider

- the factors $z_1, \ldots, z_p$ have bounded fourth moments; and
- the signal subspace $\operatorname{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ is orthogonal to the noise subspace $\operatorname{Span}\{\mathbf{v}_{q+1}, \ldots, \mathbf{v}_p\}$.

- 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 $s_k = 0, k \in \{q+1, \ldots, p\}$.
- class-conditional means and covariances of $\mathbf{x}$:

$$\boldsymbol{\mu} \equiv \mathbb{E}[\mathbf{x}|y = 1] = \sum_{k=1}^{q} s_k \mathbf{v}_k \in \mathbb{R}^p, \quad \mathbb{E}[\mathbf{x}|y = -1] = -\boldsymbol{\mu}, \quad \boldsymbol{\Sigma} \equiv \operatorname{Cov}[\mathbf{x}|y = \pm 1] = \mathbf{V}\mathbf{V}^\mathsf{T} \in \mathbb{R}^{p \times p}. \tag{3}$$

Connection to Gaussian Mixture Model

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 (3).:

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

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

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

- on **classifier** if $\hat{\boldsymbol{\beta}}$ has asymptotically the same predictive scores as $\hat{\boldsymbol{\beta}}^{\mathbf{g}}$ on any given test set, i.e., the two classifiers $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{\mathbf{g}}$ asymptotically "follows the same distribution distribution;"
- on **in-distribution performance** if the respective training and testing performances under $\mathcal{D}_{(\mathbf{x},y)}$ are asymptotically the same as under $\mathcal{D}_{(\mathbf{g},y)}$, that is $\Pr(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}} > 0) \simeq \Pr(y_i \mathbf{g}_i^\mathsf{T} \hat{\boldsymbol{\beta}}^{\mathbf{g}} > 0)$ and $\Pr(y' \mathbf{x}'^\mathsf{T} \hat{\boldsymbol{\beta}} > 0) \simeq \Pr(y' \mathbf{g}'^\mathsf{T} \hat{\boldsymbol{\beta}}^{\mathbf{g}} > 0)$, for $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$ a test sample independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and $(\mathbf{g}', y') \sim \mathcal{D}_{(\mathbf{g},y)}$ independent of $\{(\mathbf{g}_i, y_i)\}_{i=1}^n$.

Problem settings

Loss function

The loss function $\ell(\cdot, y) \colon \mathbb{R} \to \mathbb{R}_+$ in (1) is convex and continuously differentiable with $\ell(0, y) \neq 0$. Its second and third derivatives exist and are bounded, except on a finite set of points.

Large-dimensional asymptotics

As $n, p \to \infty$ with $n/p \to \bar{c} \in (0, \infty)$, $\|\boldsymbol{\mu}\|, \|\boldsymbol{\Sigma}\|, \|\boldsymbol{\Sigma}^{-1}\| = \Theta(1)$ and signals $s_1, \ldots, s_q = \Theta(1)$ with fixed q.

Main results: system of self-consistent equations

- for proximal operator $\operatorname{prox}_{\tau,f}(t) = \arg\min_{a\in\mathbb{R}} \left[f(a) + \frac{1}{2\tau}(a-t)^2 \right]$ for $\tau > 0$ and convex f, define

$$h_\kappa(t,y) = (\operatorname{prox}_{\kappa,\ell(\cdot,y)}(t) - t)/\kappa, \quad \text{for some } \kappa > 0. \tag{5}$$

- random variable $\boxed{r = ym + \sigma\tilde{e} + \sum_{k=1}^{q} \psi_k e_k}$, for **deterministic parameters** $m, \sigma, \psi_1, \ldots, \psi_q$, with label y and $e_1, \ldots, e_q$ the corresponding noise variables in the **informative factors** $z_1, \ldots, z_q$ of the LFMM, as well as $\tilde{e} \sim \mathcal{N}(0,1)$ independent of $y, z_1, \ldots, z_q$.
- distribution of r parameterized by $(m, \sigma^2, \psi_1, \ldots, \psi_q)$ **and** the distribution of $e_1, \ldots, e_q$
- **system of equations** on $q+3$ deterministic constants $\theta, \eta, \gamma, \omega_1, \ldots, \omega_q$ that fully characterize the asymptotic performance of ERM for LFMM:

$$\theta = -\mathbb{E}\left[\frac{\partial h_\kappa(r,y)}{\partial r}\right], \quad \eta = \mathbb{E}[y h_\kappa(r,y)], \quad \gamma = \sqrt{\mathbb{E}[h_\kappa^2(r,y)]}, \tag{6}$$

$$\omega_k = \mathbb{E}[h_\kappa(r,y)e_k] + \theta \cdot \mathbf{v}_k^\mathsf{T}\mathbf{Q}\boldsymbol{\xi}, \quad \psi_k = \mathbf{v}_k^\mathsf{T}\mathbf{Q}\boldsymbol{\xi}, \quad k \in \{1,\ldots,q\}, \tag{7}$$

$$\boldsymbol{\xi} = \eta\boldsymbol{\mu} + \sum_{k=1}^{q} \omega_k \mathbf{v}_k, \quad \mathbf{Q} = \left(\lambda\mathbf{I}_p + \theta\boldsymbol{\Sigma}\right)^{-1}, \quad \kappa = \frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{Q}, \quad m = \boldsymbol{\mu}^\mathsf{T}\mathbf{Q}\boldsymbol{\xi}, \quad \sigma^2 = \frac{\gamma^2}{n}\operatorname{tr}(\mathbf{Q}\boldsymbol{\Sigma})^2. \tag{8}$$

Main result: asymptotic distribution of predicted scores

Theorem (Asymptotic distribution of predicted scores)

For $\hat{\boldsymbol{\beta}}$ solution to the ERM problem in (1) *on a training set* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *of size n drawn i.i.d. from the LFMM, we have that, for any bounded Lipschitz function* $f : \mathbb{R} \to \mathbb{R}$,

$$\textit{Testing score:} \quad \boxed{\mathbb{E}\left[f(\hat{\boldsymbol{\beta}}^\mathsf{T} \boldsymbol{\nu})\right] - \mathbb{E}\left[f(\tilde{\boldsymbol{\beta}}^\mathsf{T} \boldsymbol{\nu})\right] \to 0} \tag{9}$$

$$\textit{Training score:} \quad \boxed{\mathbb{E}[f(\hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i)] - \mathbb{E}[f(\operatorname{prox}_{\kappa, \ell(\cdot, y_i)}(\tilde{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i))] \to 0} \quad i \in \{1, \ldots, n\}, \tag{10}$$

for any deterministic feature vector $\boldsymbol{\nu} \in \mathbb{R}^p$ *and*

$$\tilde{\boldsymbol{\beta}} = \left(\lambda \mathbf{I}_p + \theta \boldsymbol{\Sigma}\right)^{-1} \left(\eta \boldsymbol{\mu} + \sum_{k=1}^q \omega_k \mathbf{v}_k + \gamma \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u} \right), \tag{11}$$

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 the self-consistent system of equations.*

Asymptotic testing and training performances

Corollary (Asymptotic testing and training performances)

For any bounded Lipschitz function $f: \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\left[f(\tilde{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x})|y\right] - \mathbb{E}\left[f(r)|y\right] \to 0, \tag{12}$$

as $n, p \to \infty$, for $(\mathbf{x}, y) \sim \mathcal{D}_{(\mathbf{x},y)}$ independent of $\tilde{\boldsymbol{\beta}}$, r the "mixed" random variable, with $m, \sigma^2, \psi_1, \ldots, \psi_q$ determined by the system of equations. Consequently,

$$\Pr(y'\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}' > 0) - \Pr(yr > 0) \to 0, \tag{13}$$

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

$$\Pr(y_i\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x}_i > 0) - \Pr(y\,\mathrm{prox}_{\kappa,\ell(\cdot,y)}(r) > 0) \to 0, \quad i \in \{1, \ldots, n\}. \tag{14}$$

$$\tilde{\boldsymbol{\beta}} = (\lambda \mathbf{I}_p + \theta \boldsymbol{\Sigma})^{-1} \left(\eta \boldsymbol{\mu} + \sum_{k=1}^{q} \omega_k \mathbf{v}_k + \gamma \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u} \right), \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p / n).$$

- $\tilde{\boldsymbol{\beta}}$ is a "large-dimensional equivalent" to the (less accessible) ERM solution $\hat{\boldsymbol{\beta}}$, when training and testing performances are considered
- $\tilde{\boldsymbol{\beta}}$ is Gaussian, but having statistics (e.g., mean) **dependent on the deterministic parameters** $\omega_k = \mathbb{E}[h_\kappa(r, y) e_k] + \theta \cdot \mathbf{v}_k^\top \mathbf{Q} \boldsymbol{\xi}$ and "signals directions" $\mathbf{v}_k$, $k \in \{1, \ldots, q\}$, and thus of the distribution of the "mixed" random variable $\boxed{r = ym + \sigma \tilde{e} + \sum_{k=1}^{q} \psi_k e_k}$, through the fixed point proximal operator $h_\kappa(t, y) = (\mathrm{prox}_{\kappa, \ell(\cdot, y)}(t) - t)/\kappa$.
- if we understand the **interaction** between r and $h_\kappa(t, y) = (\mathrm{prox}_{\kappa, \ell(\cdot, y)}(t) - t)/\kappa$, we understand the universal versus non-universal behavior of the ERM $\hat{\boldsymbol{\beta}}$

Conditions for Gaussian universality under LFMM

Gaussian universality on in-distribution performance

The Gaussian universality of **in-distribution performance** (i.e., on the respective training/testing data) holds if and only if noises $e_1, \ldots, e_q$ of LFMM **informative factors are Gaussian**.

Gaussian universality on classifier

The Gaussian universality of **classifier** holds if and only if one of the following two conditions is met:

1. the informative factors $e_1, \ldots, e_q$ are Gaussian;
2. $\partial \ell(\hat{y}, y)/\partial \hat{y}$ is a linear function of $\hat{y}$, e.g., $\ell(\hat{y}, y) = (\hat{y} - y)^2/2$.

- **An important consequence**: any classifier $\hat{\boldsymbol{\beta}}$ trained using the square loss on generic LFMM $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}_{(\mathbf{x},y)}$ and $\hat{\boldsymbol{\beta}}^{\mathbf{g}}$ trained on equivalent GMM samples $\{(\mathbf{g}_i, y_i)\}_{i=1}^n \sim \mathcal{D}_{(\mathbf{g},y)}$ have asymptotically the same probability of correctly classifying a fresh LFMM test sample $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$.
- That is, ERM classifiers trained with square loss are **unable** to adapt to non-Gaussian informative factors of LFMM, contrarily to other (non-square) losses.
- **Remark**: proof of the classifier based on the fact that $h_\kappa(t, y)$ is linear if and only if $\partial \ell(\hat{y}, y)/\partial \hat{y}$ is linear.

Practical consequence of the theory: breakdown of Gaussian universality

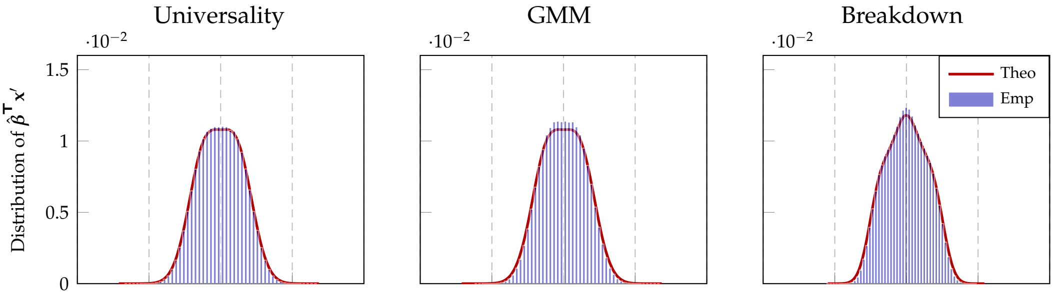


Figure: Theoretical and empirical distribution of predicted scores $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}'$ for some fresh test data $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$ independent of $\hat{\boldsymbol{\beta}}$. The theoretical probability densities (**red**), and the empirical 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 $e_2, \ldots, e_p$; normal **Middle**: $e_1, \ldots, e_p$; **Right**: uniformly distributed e_1, and normal $e_2, \ldots, e_p$.

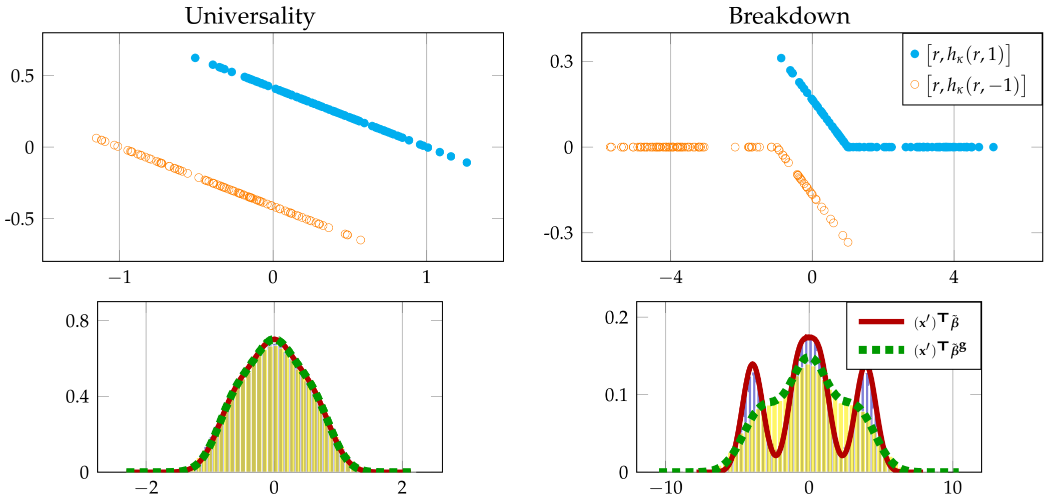


Figure: Empirical and theoretical results under an LFMM with $p = 200$, $\rho = 0.5$, $s = [\sqrt{2}; \mathbf{0}_{p-1}]$, Rademacher e_1, normal $e_2, \ldots, e_p$, and Haar distributed $\mathbf{V}$. **Top**: scatter plot of 200 independent $\left[r, h_\kappa(r, \pm 1)\right]$. **Bottom**: histograms of predicted scores on 10^6 fresh samples $(\mathbf{x}', y') \sim \mathcal{D}_{(\mathbf{x},y)}$ given by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^\mathbf{g}$, versus theoretical densities. **Left**: $n = 100$, square loss $\ell(\hat{y}, y) = (\hat{y} - y)^2/2$. **Right**: $n = 600$, square hinge loss $\ell(\hat{y}, y) = \max\{0, (1 - \hat{y}y)\}^2$.

Take-away

Take-away messages:

- a "leave-one-out" approach to assess the large-dimensional behavior of implicit ERM solution
- under LFMM, the **distribution** of noise random variables $e_1, \ldots, e_q$ in correspondence to the signals ($s_1, \ldots, s_q$ correlated to label y) determines the **universal** versus **non-universal** behaviors
- ERM solution/performance in general **non-universal** for structured data, unless for square loss

References:

- 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]
- Xiaoyi Mai and Zhenyu Liao. *The Breakdown of Gaussian Universality in Classification of High-dimensional Mixtures*. Oct. 2024. arXiv: 2410.05609 [stat]

References

[Led05] Michel Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. 2005.

[Sed+20] 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: *Proceedings of the 37th International Conference on Machine Learning*. Proceedings of Machine Learning Research. PMLR, 2020, pp. 8573–8582.

[STC19] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. "Kernel Random Matrices of Large Concentrated Data: the Example of GAN-Generated Images". In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). IEEE, 2019, pp. 7480–7484.

[LC22] Cosme Louart and Romain Couillet. "A Concentration of Measure and Random Matrix Approach to Large-Dimensional Robust Statistics". In: *The Annals of Applied Probability* 32.6 (Dec. 2022), pp. 4737–4762.

[Goo+14] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. NIPS'14. Curran Associates, Inc., 2014, pp. 2672–2680.

[BDS19] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *International Conference on Learning Representations*. ICLR'19. 2019.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp. 770–778.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *International Conference on Learning Representations*. ICLR'14. 2014.

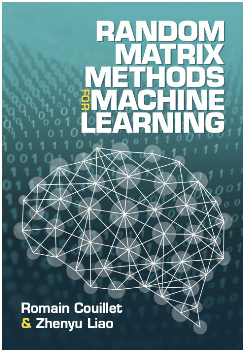
[Hua+17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 2261–2269.

[HL22] Hong Hu and Yue M. Lu. "Universality Laws for High-Dimensional Learning with Random Features". In: *IEEE Transactions on Information Theory* (2022), pp. 1–1.

RMT for machine learning: from theory to practice!

Random matrix theory (RMT) for machine learning:

- **change of intuition** from small to large dimensional learning paradigm!
- **better understanding** of existing methods: why they work if they do, and what the issue is if they do not
- **improved novel methods** with performance guarantee!



- book *"Random Matrix Methods for Machine Learning"*
- by Romain Couillet and **Zhenyu Liao**
- Cambridge University Press, 2022
- a pre-production version of the book and exercise solutions at `https://zhenyu-liao.github.io/book/`
- MATLAB and Python codes to reproduce all figures at `https://github.com/Zhenyu-LIAO/RMT4ML`

Thank you! Q & A?