A Non-Universal Convex Non-Gaussian Min–Max Theorem and Its Implications for High-Dimensional ML

ICSA China 2026, Shenzhen

Zhenyu Liao

joint work with Xiaoyi Mai (IMT, France);
Chiheb Yaakoubi (CUHK-SZ), Cosme Louart (CUHK-SZ), and Malik Tiomoko (HUAWEI)

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

June 29, 2026

Outline

High-Dimensional ERM and the prediction score

Empirical risk minimization (ERM)

Given training samples $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$, consider ERM defined as

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(x_i^\top \theta, y_i) + \rho(\theta) \right\},$$

for smooth convex loss $L(x, y)$ and convex regularization $\rho(\cdot)$.

- **Object of interest**: prediction score on an independent test sample (x, y), in the high-dimensional regime as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$:

$$\mathsf{score}(x) = x^\top \hat{\theta}.$$

- Classification error, regression risk, calibration, and margins are all functions of the law of $\mathsf{score}(x) = x^\top \hat{\theta}$.

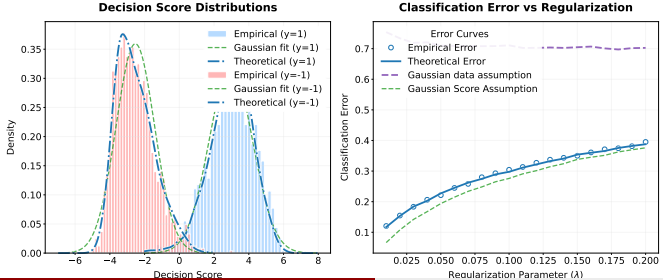
Three notions of Gaussian universality and a first counterexample

Gaussian universality in ERM

1. **Performance universality**: risk or classification error of $\hat{\theta}$ remains unchanged after Gaussian replacement.
2. **Score universality**: prediction score $x^\top \hat{\theta}$ is asymptotically Gaussian.
3. **Minimizer universality**: the ERM solution $\hat{\theta}$ itself is asymptotically Gaussian.

These notions related but not equivalent, e.g.,

- performance universality may hold even when score universality fails; and
- minimizer universality $\Longrightarrow$ score law characterization, but not necessarily score universality.



- Score universality breakdown on MNIST data
- Score distribution is data-dependent: Class 0 "less Gaussian" and Class 1 "more Gaussian"
- and (generalization) performance universality breakdown

Figure from [Yaa+26].

Existing toolkits and the bottleneck

- **Random matrix theory** (RMT) and high-dimensional statistics
- **Approximate message passing** (AMP) and state evolution
- **Gaussian convex min–max theorem** (CGMT): reduces high-dimensional problems to lower-dimensional
  - Connection: RMT or AMP fixed-point equations as reformulation from CGMT KKT condition

Major Bottleneck

For non-Gaussian data, the analysis often uses a universality "replacement":

$$x_i \overset{?}{\rightsquigarrow} g_i \sim \mathcal{N}(\mathbb{E}[x_i], \mathrm{Cov}[x_i]), \qquad x^\top \hat{\theta} \overset{?}{\rightsquigarrow} \text{Gaussian proxy.}$$

Goal: characterize **performance** and **score** universalities and breakdowns (in the general case)

Comparison to previous efforts

Table 1. **Comparison of related works and our method. Color code: Light Red = RMT, Light Green = CGMT, Light Blue = AMP, Light Purple = Lindeberg-type arguments (and variants). Notation:** ✓ indicates that the corresponding work covers the feature/setting; **R** denotes a proved statement; **A** denotes an assumption. Parenthesized **(R)** indicates that the statement holds only under a subset of the assumptions listed in that column. Note that **R** entries may rely on different baseline assumptions across columns and are therefore not always directly comparable. Papers are grouped within a single column when they share a closely related framework and essentially the same conclusions, although some modeling assumptions may differ.

		(KOR11)	(ELK13) (BEA13)	(STO13) (TIB18)	(DON16)	(PAN17) (HAS25)	(DOB18)	(ELK18)	(COU18) (MAI20)	(LOU21a) (LOU21b)	(MON22) (DAN25)	(HU22)	(PES23)	(ADO24)	(MAI25)	(ARI24) (ROS25)	(MAL25)	Our Method
Assumptions	**Statistics of x**																	
	$\mu_x \neq 0$ (non-zero mean)						✓		✓	✓	✓		✓	✓	✓	✓		✓
	$C_x \neq \sigma^2 I_p$ (correlated features)						✓		✓	✓	✓		✓		✓	✓	✓	✓
	Multiple statistics								✓	✓	✓		✓		✓	✓		✓
	Law type, default: $x \sim \mathcal{N}(\mu_x, C_x)$																	
	iid non-Gaussian entries	✓				✓	✓	✓			✓	✓					✓	✓
	Elliptical / linear factor							✓			✓	✓		✓	✓			✓
	Random features										✓	✓					✓	✓
	Heavy-tailed component			✓		✓								✓				
	Dependence $x \leftrightarrow y$																	
	$y = \theta^{\star\top} x + \varepsilon$	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓		✓	✓	✓
	y is class label of x						✓		✓	✓	✓				✓	✓		✓
	$y = f(\theta^{\star\top} x)$										✓	✓	✓				✓	✓
	Loss (default: ℓ_2)																	
	General convex on score		✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
	Regularization, default: None																	
	Ridge (ℓ_2)			✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Lasso (ℓ_1)	✓		✓		✓				✓			✓			✓		
	General convex			✓		✓				✓	✓	✓	✓	✓		✓		✓
Results	**Minimizer** universality		R		R			R										**A**
	Score universality		R								A						A	**(R)**
	Performance universality	R	R		R	R	R	R	R		R	(R)	(R)	(R)	(R)		R	**(R)**
	Non-universality **breakdown**											R	R		R			R
	Deterministic perf. **prediction**		R	R	R		R	R	R	R		R	R	R	R	R		**R**

Setup

Assumption 1 (Concentrated data and target)

For every 1-Lipschitz $f\colon (\mathbb{R}^p \times \mathcal{Y})^n \to \mathbb{R}$,

$$\mathbb{P}\left(|f(X,Y) - \mathbb{E}[f(X,Y)]| \geq t\right) \leq C\exp(-ct^2),$$

with constants $C, c > 0$ independent of n, p.

- possibly non-Gaussian
- includes Gaussian/sub-Gaussian vectors and their Lipschitz transformations; e.g., $y \in \{1,2\}, x = \phi_y(z)$ for $z \sim \mathcal{N}(0, I_p)$, with Lipschitz class-conditional maps ϕ_1, ϕ_2.

Assumption 2 (Convex and smooth loss and regularizer)

1. For each $y \in \mathcal{Y}$, the loss $L(\cdot, y)$ is convex, twice continuously differentiable, and has Lipschitz derivatives
2. The regularizer $\rho\colon \mathbb{R}^p \to \mathbb{R}$ is smooth and uniformly strongly convex.

- Covers logistic, squared, and smoothed hinge-type losses with smooth strongly convex regularization.
- Outside the main result: nonsmooth penalties such as Lasso and nonsmooth SVM hinge loss.

A first result: concentration of the minimizer

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(x_i^\top \theta, y_i) + \rho(\theta) \right\},$$

Theorem (Concentration of $\hat{\theta}$)

Under Assumptions 1 and 2, there exist constants $C, c > 0$ such that, for any 1-Lipschitz $f : \mathbb{R}^p \to \mathbb{R}$,

$$\mathbb{P}\left(\left| f(\hat{\theta}) - \mathbb{E} f(\hat{\theta}) \right| \geq t \right) \leq C \exp(-cnt^2),$$

in the high-dimensional $n \sim p$ regime, and for $\|\mathbb{E}[x]\| = \Theta(1)$, $\|\mathrm{Cov}[x]\| = \Theta(1)$ we have

$$\|\mathbb{E}[\theta]\| = \Theta(1) \quad \textit{and} \quad \left\|\mathrm{Cov}[\hat{\theta}]\right\| = \Theta(n^{-1}). \tag{1}$$

- This prepares the decomposition $\hat{\theta} = \mathbb{E}[\hat{\theta}] + \theta_0$, with $\mathbb{E}[\theta_0 \theta_0^\top] = O_{\|\cdot\|}(n^{-1})$.

A second result: quadratic universality for regularization

- **Concentration of** θ: Taylor expansion of smooth regularizer ρ for all θ close to $\hat{\theta}$:

$$\rho(\theta) \approx \rho(\mathbb{E}[\hat{\theta}]) + g_\rho(\mathbb{E}[\hat{\theta}])^\top(\theta - \mathbb{E}[\hat{\theta}]) + \frac{1}{2}(\theta - \mathbb{E}[\hat{\theta}])^\top H_\rho(\mathbb{E}[\hat{\theta}])(\theta - \mathbb{E}[\hat{\theta}]), \tag{2}$$

  for $g_\rho(\cdot)$ the **gradient** and $H_\rho(\cdot)$ the **Hessian** of ρ

Theorem (Quadratic universality of regularization)

Let $\hat{\theta}_q$ be the ERM solution with regularization ρ replaced by its quadratic surrogate ρ_q. Then for a fresh x, we have as $n, p \to \infty$

$$\left|\mathbb{E}[x^\top\hat{\theta}] - \mathbb{E}[x^\top\hat{\theta}_q]\right| \to 0, \qquad \left|\mathbb{E}[(x^\top\hat{\theta})^2] - \mathbb{E}[(x^\top\hat{\theta}_q)^2]\right| \to 0.$$

- simplifies to quadratic regularization (dependent on $\mathbb{E}[\hat{\theta}]$), yet tracking two score moments

Min–max formulation

- introduce $v = X^\top \theta$ for $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ and $L(v, Y) = \sum_{i=1}^n L(v_i, y_i)$
- then $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n L(x_i^\top \theta, y_i) + \rho(\theta) \right\} = \arg\min_{\theta \in \mathbb{R}^p, v \in \mathbb{R}^n} \left\{ \frac{1}{n} L(v, Y) + \rho(\theta) \right\}$, s.t. $v = X^\top \theta$
- introduce dual variable $w \in \mathbb{R}^n$ to get the Lagrangian

$$\mathcal{L}(\theta, v, w) = w^\top (X^\top \theta - v) + \frac{1}{n} L(v, Y) + \rho(\theta) \tag{3}$$

- **min–max formulation** (akin to CGMT):

CGMT-type formulation

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \max_{w \in \mathbb{R}^n} \left\{ \theta^\top X w - \frac{1}{n} \mathcal{L}^*(nw, Y) + \rho(\theta) \right\},$$

with $\mathcal{L}^*$ the Fenchel conjugate of $\mathcal{L}$, that is $\mathcal{L}^*(w) = \sup_{v \in \mathbb{R}^n} \{v^\top w - \mathcal{L}(v)\}$.

- For Gaussian X, classical CGMT replaces the random bilinear form $\theta^\top X w$ by a Gaussian auxiliary form.
- **Challenge** in general case: (1) law of $\theta^\top X w$ complex, and (2) in fact $\mathbb{E}[\theta^\top X w] \neq 0$.

A generalized non-Gaussian CGMT

A generalized CGMT conjecture

For concentrated column designs, the primary optimization is approximated (in the sense of first two moments of the solution) by an auxiliary **Gaussian** min–max problem:

$$\Phi(X) = \min_{\theta} \max_{w} \left\{ \theta^\top X w + \mu^\top X w + \psi(\theta, w, y) \right\} \quad \rightsquigarrow \quad \phi(g, h).$$

- allows for nonzero mean of the estimator
- key step that upgrades Gaussian-design CGMT into a non-Gaussian pipeline
- proven for Gaussian and sub-Gaussian data, expected to hold more generally

Fixed-point characterization

Limiting fixed-point equation

$$J(\mu,\alpha,\kappa;\beta,\nu)=\frac{\beta^2\kappa}{2}+\mathbb{E}\left[\mathrm{e}_L(\mu^\top x+\alpha z;\kappa)\right]+\rho(\mu)\ -\frac{\nu\alpha^2}{2}-\frac{\beta^2}{2n}\operatorname{tr}\left(\operatorname{Cov}[x](\nu\operatorname{Cov}[x]+H_\rho)^{-1}\right),\quad z\sim\mathcal{N}(0,1),$$

for $\mathrm{e}_L(u;\kappa)=\min_{v\in\mathbb{R}}\left\{\frac{1}{2\kappa}(u-v)^2+L(v,y)\right\}$ the Moreau envelope and

$$(\mu_*,\alpha_*)=\operatorname*{arg\,min}_{\mu\in\mathbb{R}^p,\alpha>0}\min_{\kappa>0}\max_{\beta,\nu>0}J(\mu,\alpha,\kappa;\beta,\nu).$$

Theorem (Limiting min–max characterization)

Under previously stated assumptions, $\left\|\mathbb{E}[\hat{\theta}]-\mu_*\right\|\to 0$ *and* $\operatorname{tr}(\operatorname{Cov}[x]\operatorname{Cov}[\hat{\theta}])-\alpha_*^2\to 0$ *as* $n,p\to\infty$.

- $\mu_*\in\mathbb{R}^p$: (deterministic) learned signal direction that "remembers" structured non-Gaussian information
- $\alpha_*\in\mathbb{R}$: scale of the high-dimensional fluctuation seen by a fresh sample

Everything needed for the law of score is "compressed" into (μ_*,α_*).

Characterization of the score distribution

Theorem (Asymptotic law of test score)

Under previously stated assumptions, we have, for any bounded measurable f,

$$\mathbb{E}[f(x^\top \hat{\theta}, y)] - \mathbb{E}[f(x^\top \mu_* + \alpha_* z, y)] \to 0.$$

- **Remark**: in passing, show that concentration data leads to minimizer universality: for all deterministic u of unit norm,

$$u^\top \hat{\theta} \quad \text{has the same asymptotic law as} \quad u^\top g, \qquad g \sim \mathcal{N}(\mathbb{E}[\hat{\theta}], \mathrm{Cov}[\hat{\theta}]).$$

The whole story in one line

$$\boxed{x^\top \hat{\theta} \;\simeq\; x^\top \mu_* + \alpha_* z, \qquad z \sim \mathcal{N}(0,1).}$$

Universality criterion

$$\boxed{x^\top \hat{\theta} \;\simeq\; x^\top \mu_* + \alpha_* z, \qquad z \sim \mathcal{N}(0,1).}$$

Score universality

$$x^\top \hat{\theta} \text{ Gaussian} \qquad \Longleftrightarrow \qquad x^\top \mu_* \text{ Gaussian.}$$

- $\alpha_* z$ is Gaussian
- score universality depends on $x^\top \mu_*$
- high-dimensional ERM may "Gaussianize" the many weak directions, yet still remember low-dimensional **non-Gaussian** structure

Example: squared loss may have non-Gaussian score but universal performance

Consider squared loss under a linear model

$$L(x, y) = \frac{1}{2}(x - y)^2, \qquad y = \theta_\star^\top x + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0,$$

and quadratic regularization.

Performance can be universal even when the score is not

The asymptotic generalization error depends on x **only** through first two moments and $\mathbb{E}[x]$ and $\mathrm{Cov}[x]$, although the score $x^\top \hat{\theta}$ may still be non-Gaussian.

- explains why squared loss can "hide" score-level non-universality
- performance universality $\not\Rightarrow$ score universality

Take-away

Take-away:

- Gaussian universality is **NOT** a single notion: performance, score, and minimizer universality can differ.
- For concentrated data, minimizer universality holds and general high-dimensional ERM score law satisfies

$$\boxed{x^\top \hat{\theta} \simeq x^\top \mu_* + \alpha_* z}$$

and learned structured projection $x^\top \mu_*$ is the "source" of non-Gaussianity

References:

- Xiaoyi Mai and Zhenyu Liao. "The Breakdown of Gaussian Universality in Classification of High-dimensional Linear Factor Mixtures". In: *The Thirteenth International Conference on Learning Representations*. 2025
- Chiheb Yaakoubi, Cosme Louart, Malik Tiomoko, and Zhenyu Liao. *Characterization of Gaussian Universality Breakdown in High-Dimensional Empirical Risk Minimization*. Apr. 2026. arXiv: 2604.03146 [stat]

Thank you! Q & A?