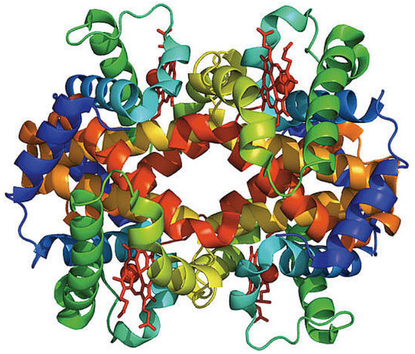
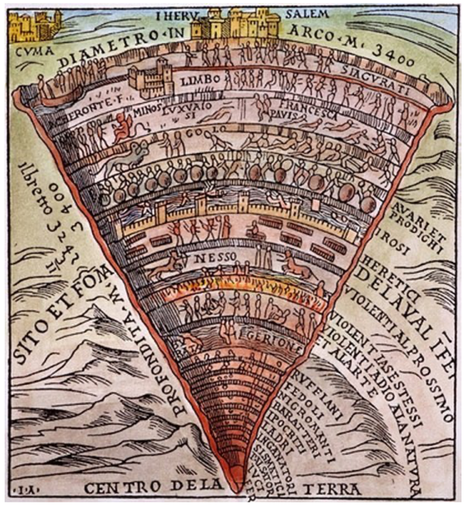

EPFL

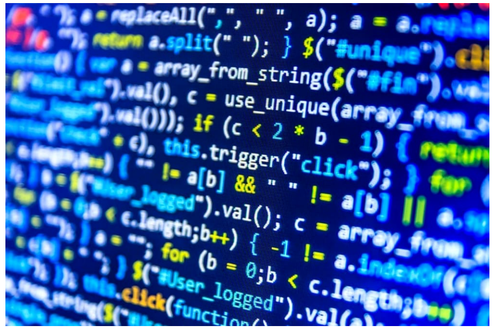

# Feature Learning in Quadratic Networks and Attention Layers

**Vittorio Erba**

Work in collaboration with:
Luca Biggio, Fabrizio Boncoraglio, Leonardo De Filippis, Julius Girardin, Florent Krzakala, Bruno Loureiro, Antoine Maillard, Emanuele Troiani, Yizhou Xu, Lenka Zdeborová





Understanding = **predicting** in **toy** models, micro to **macro**

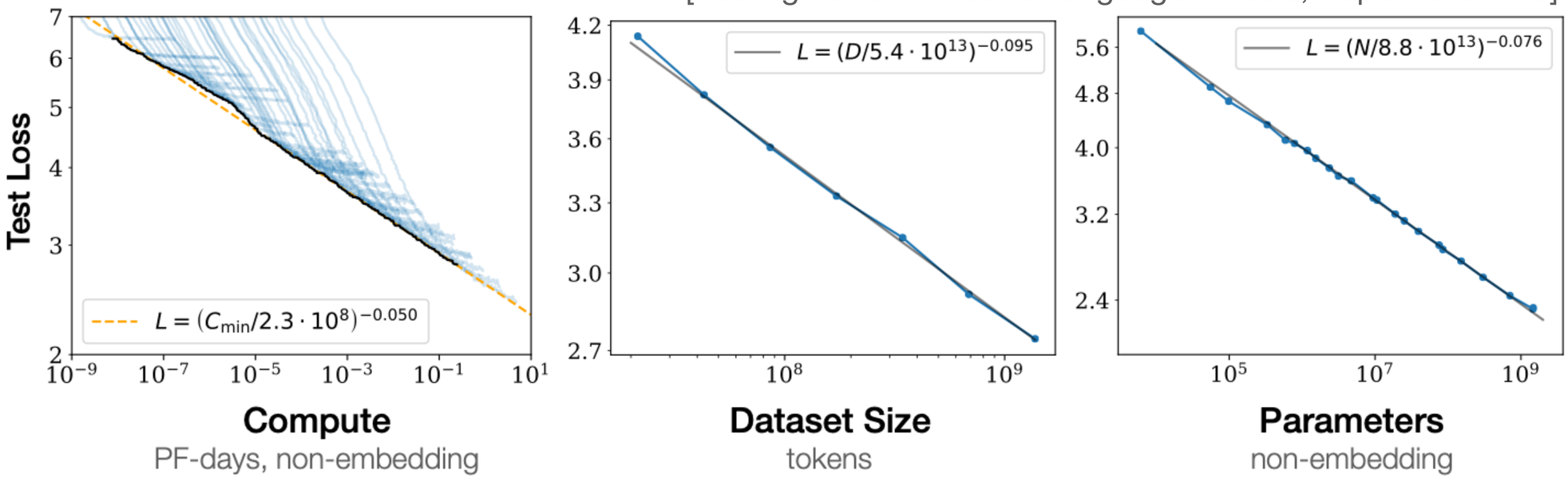
Quantification

Synthetic data
Synthetic tasks

High data dim
Low dim description

# Empirical observations (I): Scaling laws

[Scaling Laws for Neural Language Models, Kaplan et al. '20]




More compute (GPUs), larger datasets, more trainable parameters.

Scaling-laws are a key drive of current LLM industry (for better or worse)

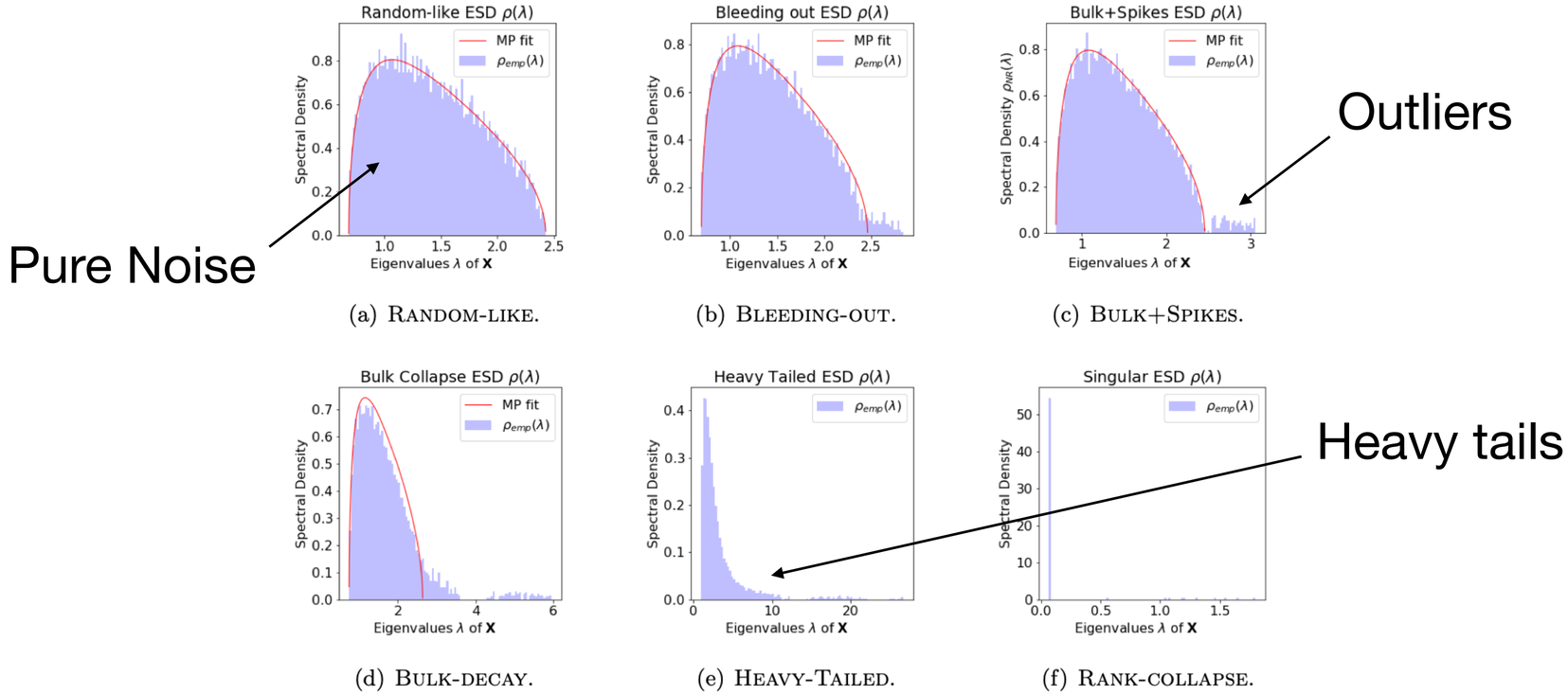
A large part* of the theory is for fixed/no kernel
Feature learning? Overparametrization/Width?

[Sharma '22, Cui et al. '23, Bahri et al. '24, Lin et al. '24, Bordelon et al. '24/'25, ...]

*But see the alternative approach at fixed data: [Li et al. '21, Pacelli et al. '23, ...]

# Empirical observations (II): Spectra

[Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning, Martin and Mahoney '21]




(a) RANDOM-LIKE. (b) BLEEDING-OUT. (c) BULK+SPIKES.

(d) BULK-DECAY. (e) HEAVY-TAILED. (f) RANK-COLLAPSE.

Understanding from first principles? Relation with test error?
Noise + signal of some kind?

[Martin et al '19/'21/..., Paquette et al. '24, Benigni at al. '25, ...]

# Today: theory of feature learning wide nets

**Questions**

- **Feature learning:** Why isn't over-parametrisation (learning with wide nets) hurtful?
- **Spectra:** Why does the spectrum of learned weights look like it does?
- **Scaling laws:** Can we understand scaling laws from simple models?

**Simplicity**: toy models with Gaussian data

**Insights**: gain some lore, not just a formula

**Exact asymptotics of quadratic fully-connected neural networks**

**Take home messages**

- $\ell_2$ regularisation $\Longrightarrow$ wide net learns narrow weights (spectral $\ell_1$)
- Spectrum: learned features + noise
- Modeling: zoology of spectral and scaling behaviours in minimal model
- **Bonus:** extension to attention layers and more (bilinear index models)

# Asymptotics of overparametrised neural nets

Supervised learning, regression

**High-dimensional data**

$$x \in \mathbb{R}^d, \quad y \in \mathbb{R}, \quad d \gg 1 \quad \text{training dataset} \quad \{x_\mu, y_\mu\}_{\mu=1}^n$$

**Two layer nets**

$$\hat{f}(x; w) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} a_i \, \sigma_i\left(\frac{w_i^\top x}{\sqrt{d}}\right) \qquad W = (w_1 \,|\, \cdots \,|\, w_p)$$

**Overparametrized**

Many hidden neurons: $p \geq d$

* can be relaxed heuristically to $p = O(d)$

Main technical problem $p = O(d)$!
- Matrix factorization
- Low degree: [Semerjian '23 + talk]
- Bayes: [Maillard et al. '24, Barbier et al. '25]

**Quadratic loss + $\ell_2$ regularization**

$$L(\{y_\mu, x_\mu\}_{\mu=1}^n, W) = \sum_{\mu=1}^{n} \left(y_\mu - \hat{f}(x_\mu, W)\right)^2 + \lambda \sum_{k=1}^{m} \|w_k\|_2^2$$

**ERM**

$$\hat{W} = \operatorname{argmin} L(\{y_\mu, x_\mu\}_{\mu=1}^n, W)$$

**Prediction**

$$\epsilon = \mathbb{E}_{\text{test}} \|y_{\text{test}} - \hat{f}(x_{\text{test}}; \hat{W})\|^2, \quad \text{spectrum of } \hat{W}$$

# Simplicity: simplest non-linear activation

$$\hat{f}(x;w) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} \sigma_i\left(\frac{w_i^\top x}{\sqrt{d}}\right)$$

Many hidden neurons: $p \geq d$

Linear activation $\sigma_i(z) = z$ Solvable, Kernel / Random features / NTK are linear in disguise

[where do I even start? For RF: ... Gerace et al. '20, Goldt et al. '22, Hu et al. '22, Montanari et al. '22, Dubova et al. '23, ...]


Polynomials: hierarchy of more and more non-linear functions

[Ge, Lee, Ma '17, ... , Mannelli, Vanden-Eijnden, Zdeborová '20, ... , Arjevani, Bruna, Kileel, Polak, Trager '25]

**Quadratic (centered) activation** $\sigma_i(z) = z^2 - \|w_i\|^2/d$ $(a_i = 1)$

Online dynamics: [Martin, Bach, Biroli '24]

$$\frac{1}{\sqrt{p}} \sum_{i=1}^{p} \sigma_i\left(\frac{w_i^\top x}{\sqrt{d}}\right) = \mathrm{Tr}\left[\frac{xx^\top - \mathbb{I}_d}{\sqrt{d}} \sum_{i=1}^{p} \frac{w_i w_i^\top}{\sqrt{pd}}\right] = \mathrm{Tr}\left[\frac{xx^\top - \mathbb{I}_d}{\sqrt{d}} S\right]$$

**Fit of a quadratic form:** $S \propto \Sigma_{i=1}^{p} w_i w_i^T$ $p \geq d$ ($S$ can represent any PSD matrix)

# Fitting quadratic forms: matrix compressed sensing

$$\hat{f}(x) = \mathrm{Tr}\left[(xx^\top - \mathbb{I}_d)/\sqrt{d} \cdot S\right] \qquad S \propto \Sigma_{i=1}^{p} w_i w_i^T = \text{any PSD matrix thanks to } p \geq d$$

**(i) Can fit only a *quadratic PSD form***. Choose a target

$$y_\mu = \mathrm{Tr}\left[(x_\mu x_\mu^\top - \mathbb{I}_d)/\sqrt{d} \cdot S^\star\right] + \sqrt{\Delta}\xi_\mu \qquad x_\mu \sim N(0,\mathbb{I}_d) \qquad \xi_\mu \sim N(0,1)$$

**(ii) Equivalent to PSD matrix compressed sensing problem** (convex)

**Loss:** $L(S) = \Sigma_{\mu=1}^{n}\left(\mathrm{Tr}\left[(xx^\top - \mathbb{I}_d)/\sqrt{d} \cdot (S - S^\star)\right]\right)^2 + \tilde{\lambda}\|S\|_*$

- $L(S)$ is MSE in $S$ + $\ell_2$ regularisation on $w$ becomes *nuclear* (spectral $\ell_1$) on $S \succeq 0$

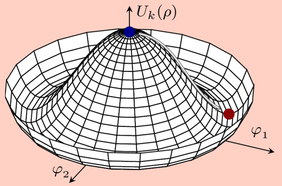
[Fazel et al '08; Donoho, Gavish et al '13; Gunasekar, et al '17, Kobayashi et al '24]

- Sensing matrix is non Gaussian $(xx^T - \mathbb{I}_d)/\sqrt{d}$

Sum of abs of eigenvalues

**(iii) Training is easy**. All local minima of $L(w)$ are global minima

[Venturi, Bandeira, Bruna '19, ...]



# Main theorem: asymptotics of quadratic networks

with Emanuele Troiani, Florent Krzakala, Lenka Zdeborová

$$y_\mu = \mathrm{Tr}\left[(x_\mu x_\mu^\top - \mathbb{I}_d)/\sqrt{d}\cdot S^\star\right] + \sqrt{\Delta}\xi_\mu \qquad \hat{f}(x) = \mathrm{Tr}\left[(xx^\top - \mathbb{I}_d)/\sqrt{d}\cdot S\right] \qquad L(S) = \Sigma_{\mu=1}^{n}\left(\mathrm{Tr}\left[(xx^\top - \mathbb{I}_d)/\sqrt{d}\cdot (S - S^\star)\right]\right)^2 + \tilde{\lambda}\|S\|_*$$

Data model — $\mu^\star$ — Learner — Loss

**Define:**

Samples $\alpha = n/d^2 = O(1)$

Width $\kappa = p/d \geq 1$

Noisy spectrum $\mu_\delta^\star = \mu^\star \boxplus \mu_{\mathrm{SC}}(\delta)$

$$Q^\star = \mathbb{E}_{x\sim\mu^\star}[x^2]$$

$$\tilde{\lambda} = \sqrt{\kappa}\lambda$$

**Find unique solution $(\bar{\delta}, \bar{\epsilon})$ of:**

$$\begin{cases} 4\alpha\delta - \dfrac{\delta}{\epsilon} = \partial_1 J(\delta, \tilde{\lambda}\epsilon) \\ Q^\star + \dfrac{\Delta}{2} + 2\alpha\delta^2 - \dfrac{\delta^2}{\epsilon} = (1 - \epsilon\tilde{\lambda}\partial_2) J(\delta, \tilde{\lambda}\epsilon) \end{cases}$$

$$J(a,b) = \int_b^{+\infty} dx\, \mu_a^\star(x)\,(x-b)^2$$

depending on $\alpha, \lambda, \Delta, \mu^\star$ but **not on $\kappa$ !**

**RESULT! Predict:**

Test error: $\lim_{d\to\infty} e_{\mathrm{test}}(\hat{S}) = 2\alpha\bar{\delta}^2 - \Delta/2$

Spectrum: $\lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \delta(x - \lambda_i(\hat{S})) = F_{\bar{\delta}}(\tilde{\lambda}\bar{\epsilon})\delta(x) + I(x>0)\,\mu_{\bar{\delta}}^\star(x + \tilde{\lambda}\bar{\epsilon})$

**Idea of proof = Morally a linear model**

Gaussian universality $(xx^\top - \mathbb{I}_d)/\sqrt{d} \approx$ GOE(d)

AMP/SE for non-separable priors — spectral $\ell_1$ regularisation

Trick to solve extensive multi-index model!

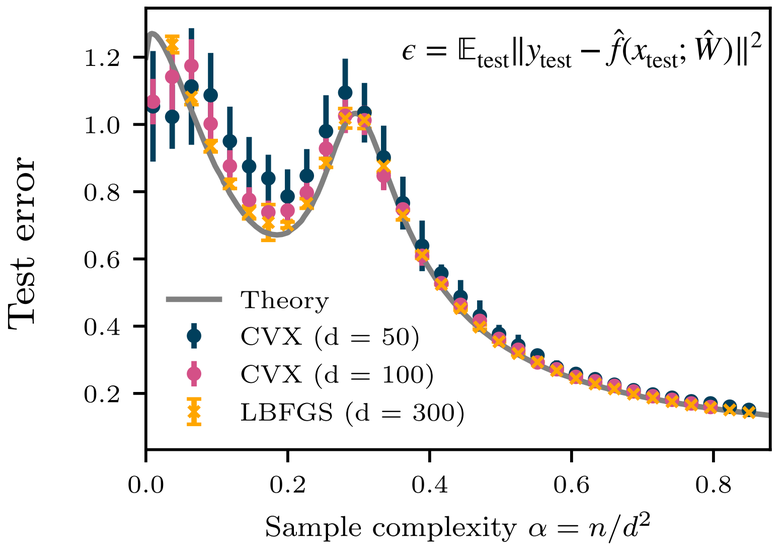
Before: only Bayes optimal estimator [Maillard et al. '24, Barbier et al. '25]

Universality: [Gerace et al. '20, Goldt et al. '22, Hu et al. '22, Montanari et al. '22, ...]

AMP: [Bayati et al. '11, Donoho et al. '16, ..., Berthier et al. '20, Gerbelot et al. '23, ....]

# Theory works! Comparison with solver and GD

**Setting:** $\mu^\star = MP(p^\star/d)$ Marchenko-Pastur distribution $\quad p^\star = 0.2\,d \quad \Delta = 0.5 \quad \lambda = 0.02$



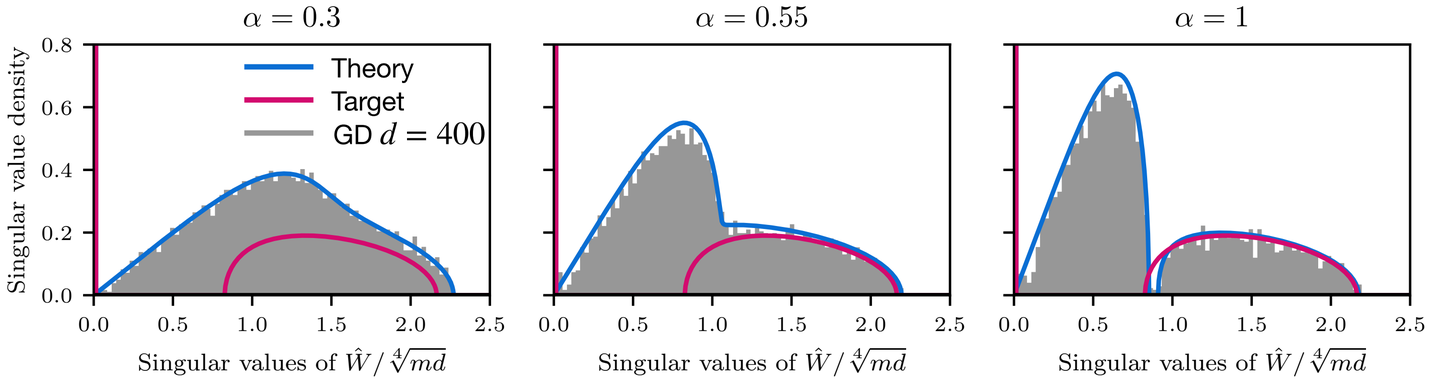

**CVX:** convex problem solver used on equivalent matrix compressed sensing problem (find $\hat{S}$)

**Theory matches with numerical global min**

**LBFGS:** GD + tricks to achieve faster convergence in convex landscapes on original network problem (find $\hat{w}$)

**Theory is predictive also for GD**

[Venturi, Bandeira, Bruna '19, ...]




**Theory is predictive for spectra found by GD**

$\ell_2 \implies \ell_1$ (spectral)
Matrix compressed sensing

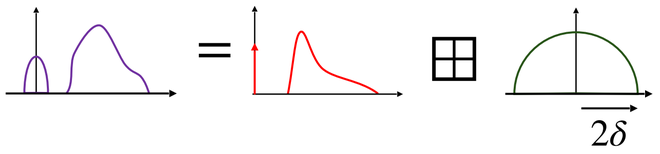
$$\mu(\hat{S}) = F_{\tilde{\delta}}(\tilde{\lambda}\bar{\epsilon})\delta(x) + I(x > 0)\,\mu_{\tilde{\delta}}(x + \tilde{\lambda}\bar{\epsilon})$$

Same minimum for $p \geq d$

**Take home 1:** regularisation induces sparse representations

**Take home 2a:** spectrum ~ noisy version of the target

$$\mu_\delta^\star = \mu^\star \boxplus \mu_{\mathrm{SC}}(\delta)$$



# Heavy-tailed target: spectra and scaling

with Leonardo De Filippis, Julius Girardin, Florent Krzakala, Bruno Loureiro, Emanuele Troiani, Yizhou Xu, Lenka Zdeborová

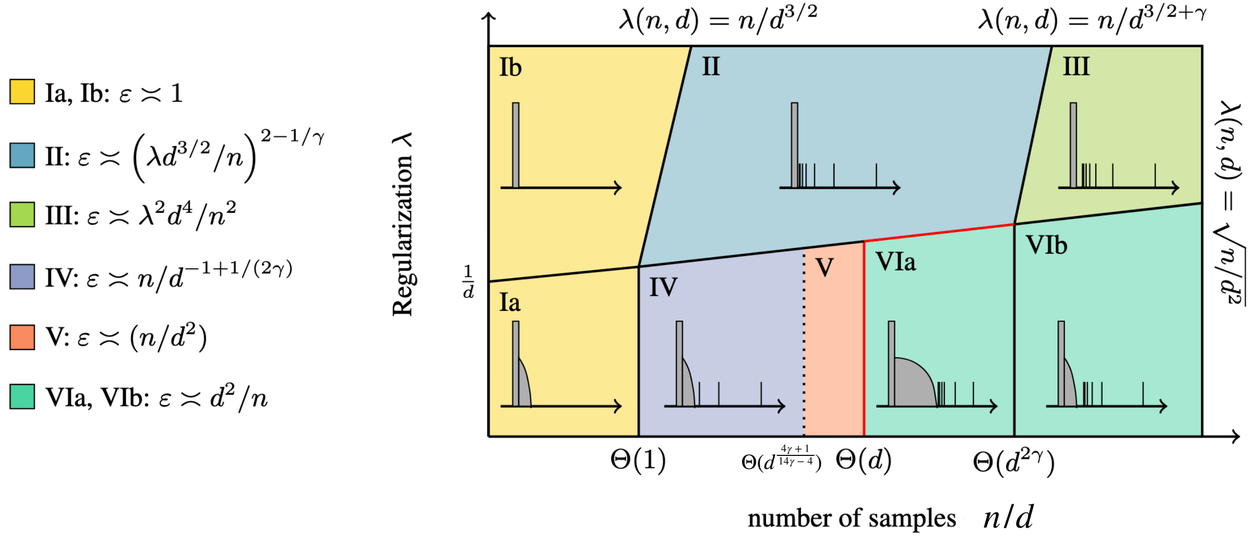
**Setting:** $\lambda_i(S^\star) = i^{-\gamma} \qquad \gamma > 1/2$

Compressed sensing: quasi-sparse target
Hierarchy of features contributing to label.

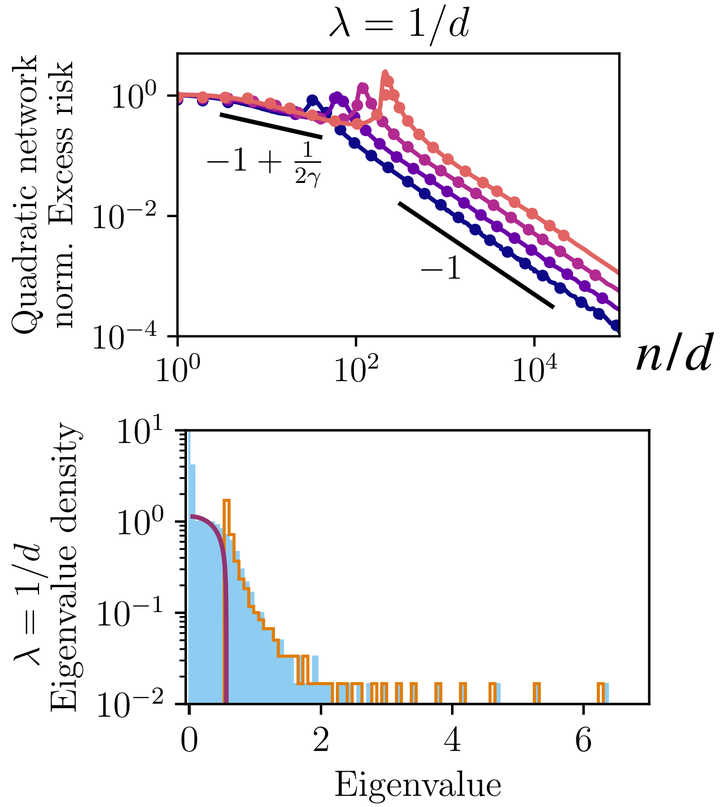
[Meyer '92; Donoho '98; Mallat '99, ..., Negahban et al. '11, Raskutti et al. '11, ...]

Dynamics in un-regularised / noiseless case: [Ben Arous, Erdogdu, Vural, Wu'25]

Full characterisation of test and spectrum as function of $d, n, \lambda, \Delta, \gamma$




Comparison: GD, theory, analytical slopes




Heuristic use of main theorem in **non-asymptotic setting**, where $n, d, \lambda$ are not obeying any scaling

Before: as $d \gg 1$
$\alpha = n/d^2 = O(1)$
$\lambda = O(1)$

Proven in ridge + random features: [L Defilippis, B Loureiro, T Misiakiewicz '24]

# Heavy-tailed target: interpretable test error

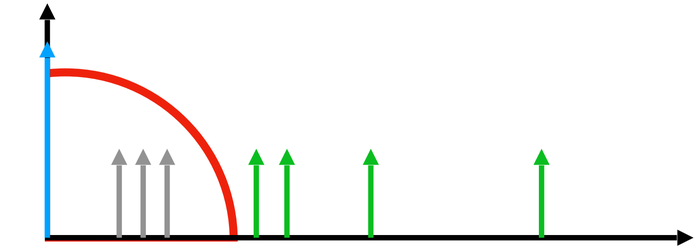
with Leonardo De Filippis, Julius Girardin, Florent Krzakala, Bruno Loureiro, Emanuele Troiani, Yizhou Xu, Lenka Zdeborová

**Setting:** $\lambda_i(S^\star) = i^{-\gamma} \qquad \gamma > 1/2$

Compressed sensing: quasi-sparse target
Hierarchy of features contributing to label.

[Meyer '92; Donoho '98; Mallat '99, ..., Negahban et al. '11, Raskutti et al. '11, ...]

Dynamics in un-regularised / noiseless case: [Ben Arous, Erdogdu, Vural, Wu'25]



BBP-style theory of spectrum

Test error =

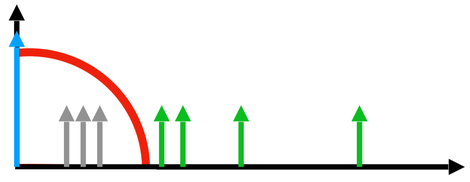
Regularisation (spike in zero)

Noise overfitting (bulk)

Features not learned (spikes in bulk)

Features learned (spikes out of bulk)

$$\mathrm{R} = \underbrace{\delta^2 \int_{\lambda\epsilon/\delta}^{2} \mu_{\mathrm{s.c.}}(\mathrm{d}x) \left(x - \frac{\lambda\epsilon}{\delta}\right)^2}_{\substack{\textit{overfitting} \\ \textit{(learned noise)}}} + \underbrace{\frac{1}{d} \sum_{i=K(\delta)}^{d} s_i^2}_{\substack{\textit{underfitting} \\ \textit{(not learned features)}}} + \underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta)} \left[ \left( \frac{\delta^2}{s_i} - \lambda\epsilon \right)^2 + \frac{\delta^2}{s_i} \left( s_i + \frac{\delta^2}{s_i} - \lambda\epsilon \right) \right]}_{\substack{\textit{approximation error} \\ \textit{for learned features}}}$$



Test error =

Regularisation (spike in zero)

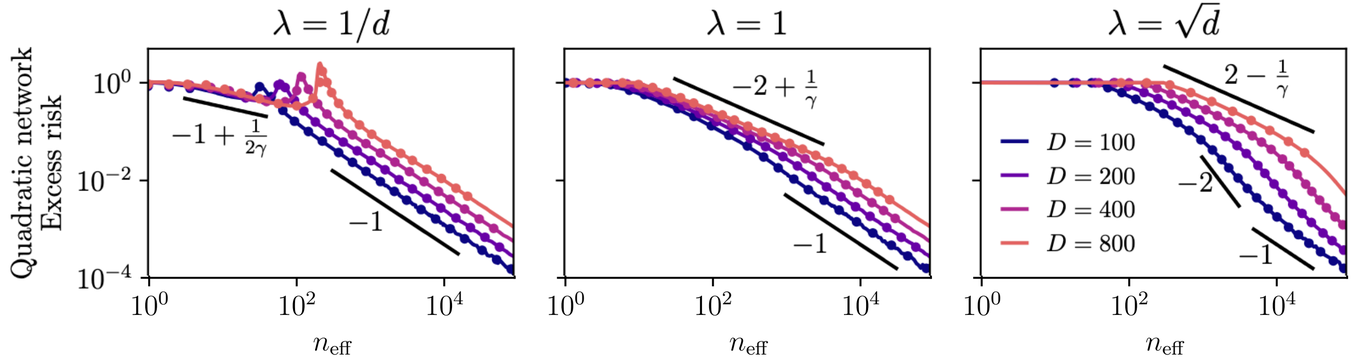
Noise overfitting (bulk)

Features not learned (spikes in bulk)

Features learned (spikes out of bulk)

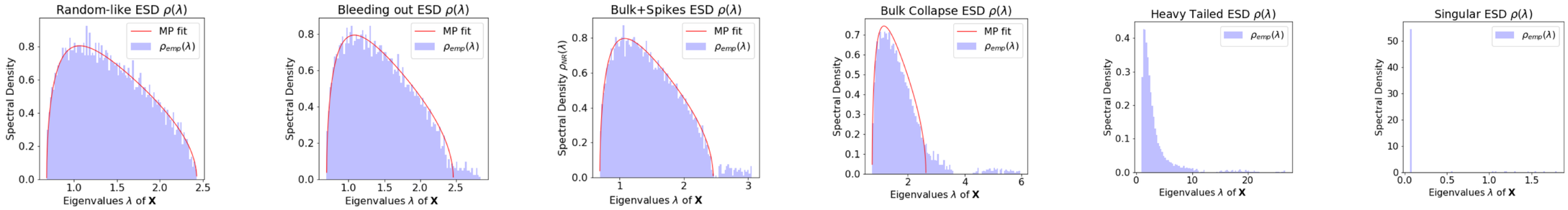
**Take home 2b:** spectrum $\Longrightarrow$ test error decomposition

**Take home 3:** hierarchical target $\Longrightarrow$ zoology of scaling laws



# Modeling: Full zoology of empirical spectra

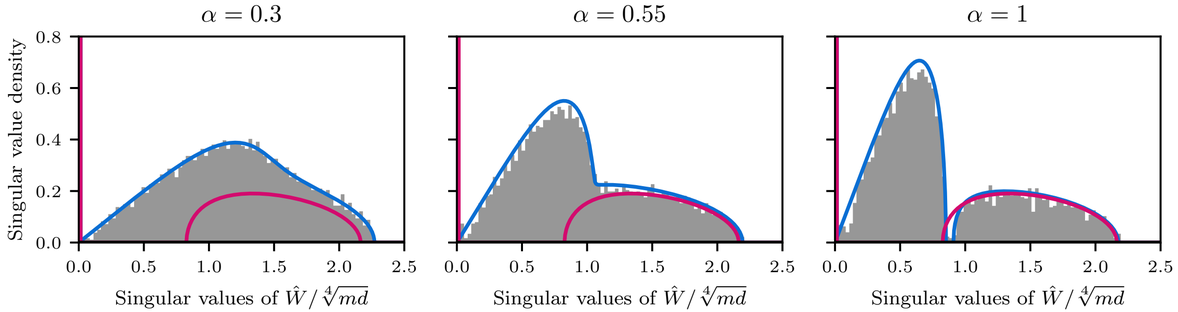
## Experiments

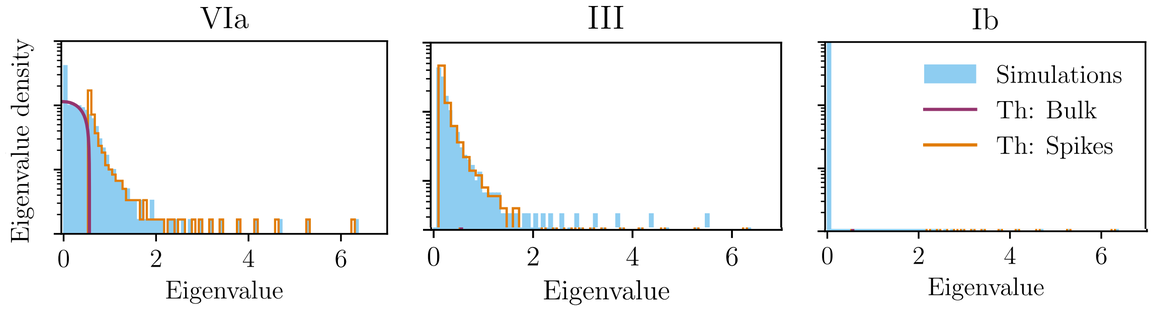



(a) Random-like. (b) Bleeding-out. (c) Bulk+Spikes. (d) Bulk-decay. (e) Heavy-Tailed. (f) Rank-collapse.

[Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning, Martin and Mahoney '21]

## Our theory




$\mu^\star = MP(p^\star/d)$

Marchenko-Pastur distribution


**Claim:** quadratic nets are expressive enough

$\lambda_i(S^\star) = i^{-\gamma}$

Heavy-tailed distribution

# Beyond quadratic networks: attention layers

with Fabrizio Boncoraglio, Yizhou Xu, Florent Krzakala, Emanuele Troiani, Lenka Zdeborová

**Sequential data** $X_{ai}$

Token $T$

Embedding $d$

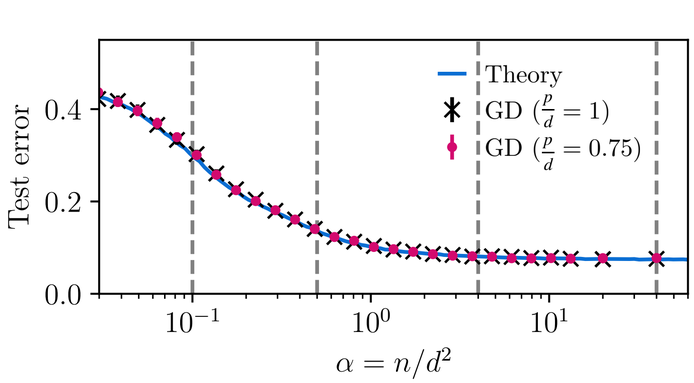
**Attention mechanism**: token-to-token correlation

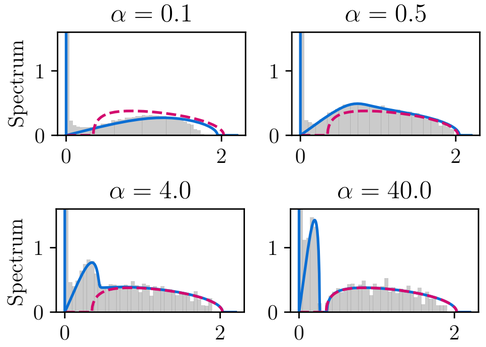
$$A(X, W) = \text{softmax}\left(X\, W_Q W_K^\top\, X^\top\right) \in \mathbb{R}^{T\times T}$$

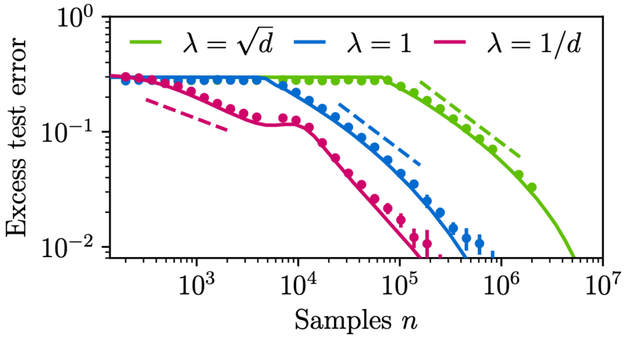
**Non-linearity + bilinear form of data!**

**Result:** same characterisation of ERM (and BO!) learning in single layer of attention for $T = O(1)$

Setting: regression
Student: attention
Target: noisy attention




**Take home (bonus):** asymptotic theory of **bilinear multi-index model**

$$y = g\left(x^\top S_1 x\,,\ \cdots\ , x^\top S_L x\right)$$

☑ $L = 1$ ☑ Multi-token $T = O(1)$ WORK IN PROGRESS $L > 1$

# Conclusion

**Take home 1:** regularisation induces sparse representations

**Take home 2:** spectrum = learned features + noise

**Take home 3:** zoology of spectral and scaling behaviours

**Take home (bonus):** asymptotic theory of bilinear multi-index model

# Next steps

1. Predictive power of scaling and spectra in real nets
2. Study multi-bilinear index models → Multi-matrix denoising
3. Modeling! Explore phenomena in feature learning setting / attention

# References

with Luca Biggio, Fabrizio Boncoraglio, Leonardo De Filippis, Julius Girardin, Florent Krzakala, Bruno Loureiro, Antoine Maillard, Emanuele Troiani, Yizhou Xu, Lenka Zdeborová


**Quadratic:**
[2025] The Nuclear Route: Sharp Asymptotics of ERM in Overparameterized Quadratic Networks
[2025] Scaling Laws and Spectra of Shallow Neural Networks in the Feature Learning Regime
**Attention:**
[2024] Bilinear sequence regression: A model for learning from long sequences of high-dimensional tokens
[2025] Bayes optimal learning of attention-indexed models
[2025] Single-Head Attention in High Dimensions: A Theory of Generalization, Weights Spectra, and Scaling Laws