The Nuclear Route: Sharp Asymptotics of ERM in Overparameterized Quadratic Networks

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, Florent Krzakala

EPFL SPOC lab, EPFL IDEPHICS lab, Lausanne

Question: Can we understand feature learning in 2-layer neural networks trained with GD?

We present a solvable, minimal yet expressive high-dimensional regression task that captures feature learning in 2-layer NN, with quadratic activation and quadratic target function. Even under simple assumptions on the data structure, our setting is rich enough to observe several non-trivial phenomena, such as:

Feature learning: We show that the 2-layer NN outperforms linear and kernel methods.
Induced regularisation: We show that ℓ_2 regularisation on the weights w of the network can be interpreted as a nuclear norm regularisation for the matrix ww^T, implying that the learned weights are effectively those of the narrowest neural network that can fit the data, favouring more compact features.
Spectral structure: We predict analytically the spectrum of the learned weight matrix, which aligns with the phenomenology observed in practice.
Interpolation peak: We predict analytically an interpolation peak in the test error.

1. High-dimensional ...

2-layer NN with quadratic activation

d: input dimension — scalar output
m: hidden layer dimension

$$f_w(x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \sigma_k\left(\frac{w^k \cdot x}{\sqrt{d}}\right), \quad x \in \mathbb{R}^d$$

$$\sigma_k(z) = z^2 - ||w^k||^2/d$$ ← zero-mean labels

Landscape properties studied in *Venturi '19, Gamarnik '20, Mannelli '20, ...*
$m = 1$ phase retrieval *Mondelli '18, ...*

Training dataset of n inputs/ outputs

$$x^\mu \sim N(0, \mathbb{I}_d), \quad y^\mu = \sum_{i,j=1}^{d} S^* x_i^\mu x_j^\mu + \sqrt{\Delta}\xi^\mu$$

Training loss

$$\mathscr{L}(w) = \sum_{\mu=1}^{n} ||f_w(x^\mu) - y^\mu||^2 + \lambda \sum_{k=1}^{m} ||w^k||^2$$

$$d \to \infty, \quad n = \alpha d^2, \quad m \geq d, \quad \text{rk}(S^*) = \kappa^* d$$

Challenging extensive-width regime.
Requires $O_d(d^2)$ samples *Maillard '25.*
$O_d(1)$-width regime widely studied: committee/single/multi-index models *Seung '90, Aubin '18, Damian '22, ...*

2. Main Result: analytic prediction of test error of loss minima

Main theorem

Call $\hat{w} \in \arg\min \mathscr{L}(w)$ any global minimum of the loss or output of gradient based method[1].

$$\lim_{d\to+\infty} \mathbb{E}_{\text{training set}}[e_{\text{test}}(\hat{w})] = 2\alpha\delta^2 - \Delta/2$$

$$\begin{cases} 4\alpha\delta - \delta\epsilon^{-1} = \partial_1 J(\delta, \lambda\sqrt{\kappa}\epsilon) \\ Q^* + \frac{\Delta}{2} + 2\alpha\delta^2 - \delta^2\epsilon^{-1} = (1 - \lambda\sqrt{\kappa}\epsilon\partial_2) J(\delta, \lambda\sqrt{\kappa}\epsilon) \end{cases}$$

$$J(a,b) = \int_b^{+\infty} dx\, \mu_a^*(x)\,(x-b)^2, \quad \mu_a^* = \mu^* \boxplus_{[2]} \mu_{\text{semicricle,a}}$$

Assumptions

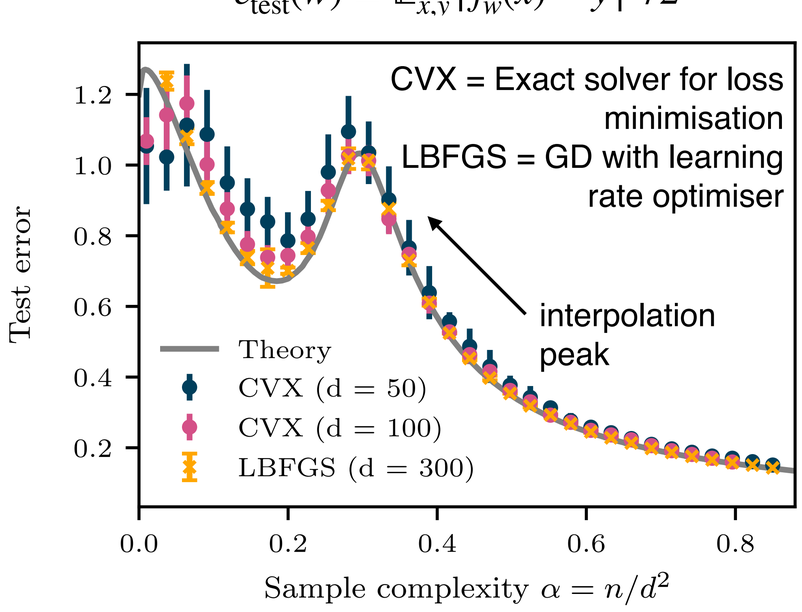
- S^* has spectral density μ^* with finite first and second moment as $d \to +\infty$
- fix $\kappa = m/d \geq 1$ and $\lambda > 0$

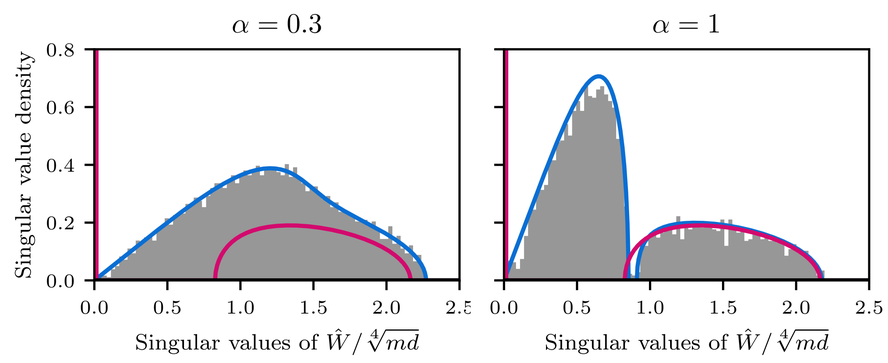
[1] *Arjevani et al. 2025* shows that $\mathscr{L}(w)$ has no spurious local minima, thus if GD converges to a minimum it must converge to a global minimum. We do not establish convergence to minima rigorously, but check it numerically.

[2] In $\mu_a^* = \mu^* \boxplus \mu_{\text{semicricle,a}}$, $\boxplus$ is the free convolution, i.e. μ_a^* is the spectral density of $S^* + aZ$ with Z Gaussian noise. μ_a^* is a close relative of the spectral density of the global minima of $\mathscr{L}(w)$.

Case study: $S^* \propto w_*^T w_*$, i.e. target function is also a two-layer quadratic NN, with $m^* = \kappa^* d$ hidden units ($\kappa^* = 0.2$ in plot

$$e_{\text{test}}(w) = \mathbb{E}_{x,y} |f_w(x) - y|^2/2$$

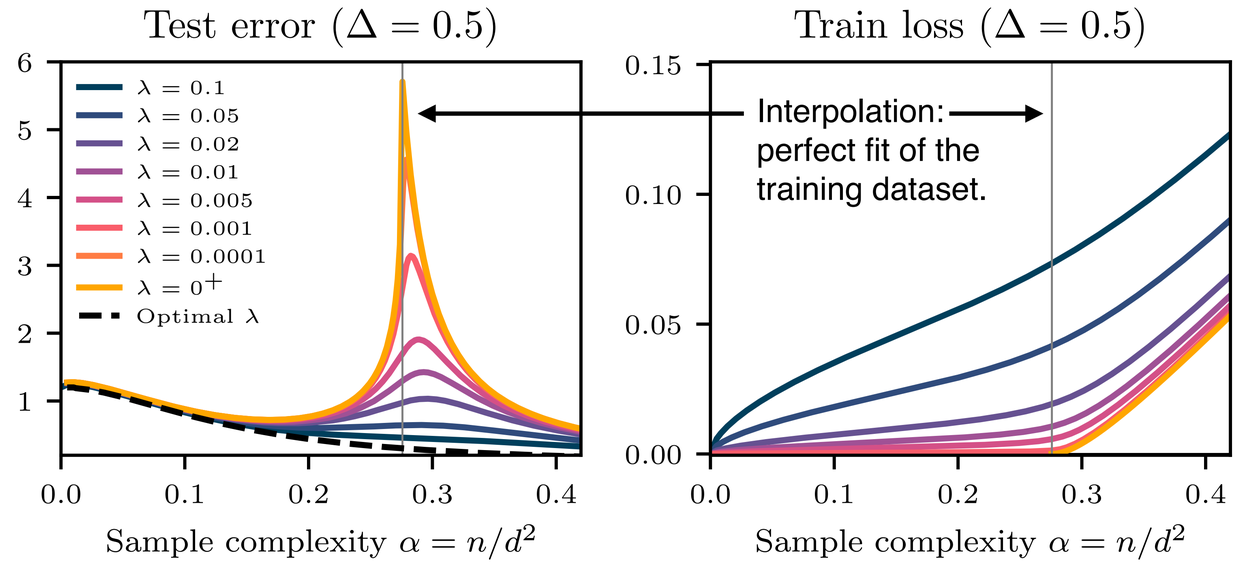


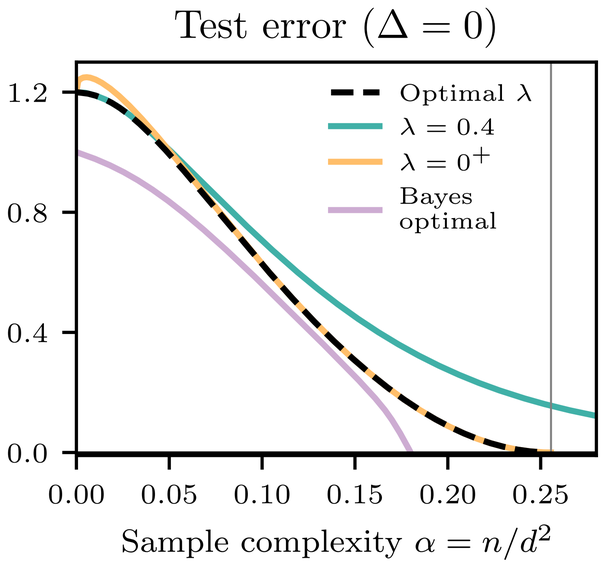


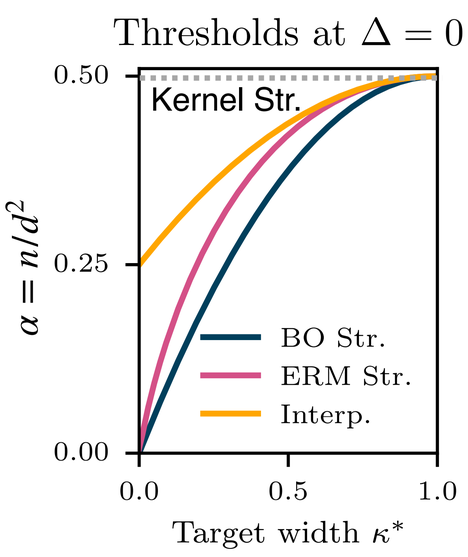
Blue: prediction of spectral density of trained weights
Red: spectral density of target function
Grey: histogram of of spectral density of trained weights (GD)
Phenomenology similar to applications:
Martin and Mahoney '18

3. Learning curves

$S^* \propto w_*^T w_*$, 2-layer NN with $m^* = \kappa^* d$ hidden units and $\kappa^* = 0.2$.







Str. = perfect recovery
Interp. = perfect fit of training data

BO = Bayes optimal
ERM = minimum of empirical loss
Kernel = quadratic kernel ridge

4. Sketch of proof: induced nuclear regularisation

1) Map problem in convex matrix compressed sensing (unique global minimum) for $S = \Sigma_k w_k w_k^T/\sqrt{md}$:
$\tilde{\mathscr{L}}(S) = \Sigma_\mu |\,\text{Tr}[(S - S^*)(xx^T - \mathbb{I})/\sqrt{d}\,|^2 + \text{Tr}(S)$
under the convex constraint that S is PSD
Induced reg. on S favours low rank/width
2) Use Gaussian universality on data
$(xx^T - \mathbb{I})/\sqrt{d} \to Z \sim \text{GOE(d)}$
3) Write appropriate AMP that, at convergence finds global minima of $\mathscr{L}$, and show that it converges
4) Use AMP's state evolution to characterise analytically the properties of the global minima

Matrix compressed sensing widely studied:
Fazel '08, Donoho '13, Amelunxen '14, Oymak '16
AMP standard proof technique for convex/RS optimisation:
Bayati '11, Donoho '16, Rangan '16, Vilucchio '25

arXiv:2505.17958

See preprint for more results and for discussion of many related works