How Width and Data Shape Generalization Scaling Laws in Quadratic Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Understanding how performance scales jointly with model size and data is a central problem in modern machine learning. Existing theoretical works on scaling laws typically describe generalization as a function of data or compute, often in fixed-feature or infinite-width regimes and for online SGD. Here, we instead study how generalization scales with the number of trainable parameters and the number of samples in a feature-learning model. We analyze ℓ_2-regularized empirical loss minimization for a quadratic two-layer network in a finite-sample setting with structured data. This setting allows for an explicit characterization of the generalization error as a function of the number of samples, model width, and regularization. Our results reveal a phase diagram with distinct scaling regimes as the number of parameters varies. In particular, the generalization error follows data-dependent power laws controlled by the spectral structure of the target. We further characterize the transitions between regimes, including the onset of interpolation, and their impact on generalization.

1 Introduction

Scaling laws have become a central tool for understanding modern machine learning systems, relating performance to the amount of data, model size, and compute (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022). On the theoretical side, a growing body of work has obtained precise scaling descriptions of generalization as a function of the number of samples or compute, see e.g. (Caponnetto & De Vito, 2007; Bordelon et al., 2020; Cui et al., 2021; Bahri et al., 2024; Paquette et al., 2024; Defilippis et al., 2026b;a; Cagnetta et al., 2026). A separate line of research studies how increasing the number of parameters affects optimization, initialization, or expressivity, without directly addressing data-dependent generalization (Yang et al., 2021; Bordelon et al., 2024b; Chizat & Netrapalli, 2024; Chaintron et al., 2026). Works that do analyze generalization as a function of model size typically focus on regimes with fixed representations, such as random feature models (Spigler et al., 2020; Maloney et al., 2022; Bahri et al., 2024; Atanasov et al., 2026; Bordelon et al., 2024a; Defilippis et al., 2024) where increasing the number of parameters does not alter the learned features, and the role of model size is therefore fundamentally limited.

Studies of scaling laws in model size in the feature learning regime were initiated only very recently (Ren et al., 2025; Ben Arous et al., 2026). In these works, the scaling laws are formulated for idealized dynamical limits, where performance is controlled by finite training time, or by the product of learning rate and number of iterations. A characterization of the scaling laws of the test error, train error, and their difference, the generalization gap, as a function of the number of samples and the model size in a feature learning setting remains open.

In this work, we fill this gap by characterizing how the generalization error of the empirical risk minimizer scales with the number of parameters and samples in a feature-learning setting. This allows us to disentangle their roles and characterize how data availability, width, regularization, and target structure jointly determine the performance of the trained predictor. To make this tractable, we consider a minimal yet non-trivial model: a quadratic two-layer neural network trained on high-dimensional data with a power-law structure. This setting captures a genuine feature-learning regime while remaining amenable to theoretical analysis. Our goal is not to model realistic architectures, but to isolate the mechanisms by which model size, data, and data structure interact to determine generalization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Model Setting. We consider the ℓ_2-regularized empirical risk minimization (ERM) problem

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) \quad \text{where} \quad \mathcal{L}(\boldsymbol{W}) = \sum_{\mu=1}^{n} \left(y_\mu - f_p(\boldsymbol{x}_\mu; \boldsymbol{W})\right)^2 + \lambda ||\boldsymbol{W}||_{\mathrm{F}}^2 \,, \tag{1}$$

for the class of quadratic two-layer neural networks with p hidden units

$$f_p \boldsymbol{x}; \boldsymbol{W}) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \sigma_j \left(\frac{\boldsymbol{w}_j^\top \boldsymbol{x}}{\sqrt{d}} \right) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left(\left(\frac{\boldsymbol{w}_j^\top \boldsymbol{x}}{\sqrt{d}} \right)^2 - \frac{||\boldsymbol{w}_j||_2^2}{d} \right) \tag{2}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1 | \dots | \boldsymbol{w}_p)^\top \in \mathbb{R}^{p \times d}$ are the first-layer trainable weights and we fixed the second layer weights to one. The activation σ_j is a centered quadratic function. In this work, we are interested in characterizing the properties of global minimizers of the objective in (1). For this purpose, the interplay between the network width p and the input dimension d plays a central role. Indeed, for the centered quadratic activation in (2), the network can be rewritten as a linear model with structured data and weight re-parametrization:

$$f_p(\boldsymbol{x}_\mu; \boldsymbol{W}) = \mathrm{Tr}[\boldsymbol{S}\, \boldsymbol{G}_\mu] \,, \qquad \boldsymbol{S} = \boldsymbol{W}^\top \boldsymbol{W} / \sqrt{pd} \in \mathbb{R}^{d \times d} \,, \qquad \boldsymbol{G}_\mu = (\boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top - \boldsymbol{I}_d)/\sqrt{d}. \tag{3}$$

As $\boldsymbol{W}$ varies over $\mathbb{R}^{p \times d}$, the matrix $\boldsymbol{S}$ ranges over the cone of positive semidefinite matrices with rank at most p. Therefore, minimizing over the network weights is equivalent to solving the constrained optimization problem

$$\hat{\boldsymbol{S}} = \underset{\boldsymbol{S} \succeq 0,\, \mathrm{rk}(\boldsymbol{S}) \le p}{\arg\min} \left\{ \sum_{\mu=1}^{n} \left(y_\mu - \mathrm{Tr}[\boldsymbol{S}\, \boldsymbol{G}_\mu]\right)^2 + d\tilde{\lambda}\, \mathrm{Tr}(\boldsymbol{S}) \right\}, \quad \text{with } \tilde{\lambda} := \lambda \sqrt{p/d}. \tag{4}$$

This formulation makes explicit that the width of the neural network acts as a rank constraint on the learned positive semidefinite matrix, while the weight decay in the original parametrization becomes a trace regularization in the matrix formulation, mapping the original problem to matrix compressed sensing with nuclear norm regularization (Fazel et al., 2008) and a non-convex rank constraint. Since the network is quadratic in the input, it can only learn target functions that lie in the span of quadratic features. We therefore consider labels generated by a centered quadratic model,

$$y_\mu = \mathrm{Tr}[\boldsymbol{S}_\star (\boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top - \boldsymbol{I}_d)/\sqrt{d}] + \sqrt{\Delta}\, \xi_\mu \,. \tag{5}$$

Here $\{\boldsymbol{x}_\mu\}_{\mu=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$ denotes the input samples, and $\{\xi_\mu\}_{\mu=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is independent Gaussian label noise, with variance parameter $\Delta \ge 0$. The quadratic form represents the structured component of the target, which is in the model class of the network and can therefore be learned from data. The noise term represents an unstructured component of the labels: it carries no predictive information about fresh inputs. We further assume that the target matrix $\boldsymbol{S}_\star$ is symmetric and has power-law decaying spectrum, i.e. that it has eigenvalues $\{\sqrt{d}\, i^{-\gamma}/\sqrt{\zeta(2\gamma)}\}_{i=1}^d$, where ζ is Riemann's zeta function and we impose $\gamma > 1/2$ to ensure square summability. The normalization is such that $\mathrm{Tr}(\boldsymbol{S}_\star^2)/d = 1 + o_d(1)$.

Our main goal in this paper is to study the excess test error

$$\mathcal{E}(\hat{\boldsymbol{W}}) = \frac{1}{2} \mathbb{E}_{\boldsymbol{x},y} \left[\left(y - f_p(\boldsymbol{x}; \hat{\boldsymbol{W}})\right)^2 \right] - \frac{\Delta}{2} \,, \tag{6}$$

of the minimizers $\hat{\boldsymbol{W}}$ of (1), where $(\boldsymbol{x}, y)$ is a new data pair with $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and y given by (5). We want to characterize its scaling behavior as a function of the number of trainable parameters $N_{\mathrm{param}} = dp$, the number of available training samples n and the regularization $\tilde{\lambda}$. For this purpose we define the following high-dimensional limit

$$d \gg 1, \qquad p = \Theta(d^\rho), \qquad n = \Theta(d^\alpha), \qquad \tilde{\lambda} = \Theta(d^\ell) \qquad \text{with} \qquad \alpha, \rho > 0 \,, \tag{7}$$

where d will be taken large enough, but still finite.

Summary of main results. Our results provide a predictive theory of how model size acts as a control parameter in a feature-learning model.

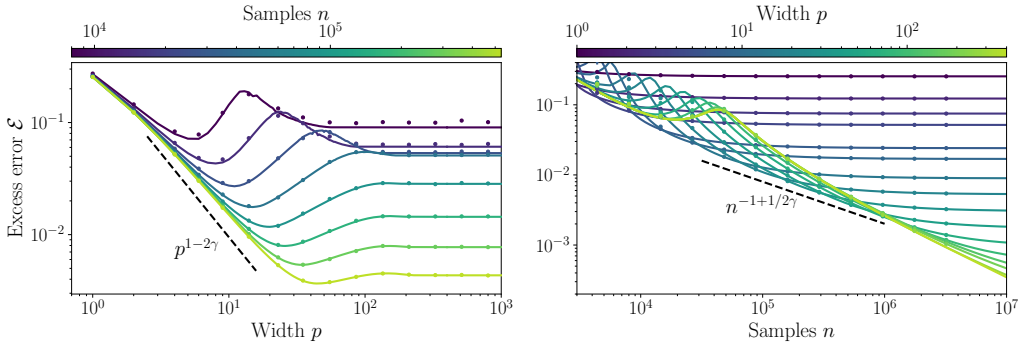


Figure 1: Excess test error vs width p (left) and number of samples n (right), for $d = 400$, $\gamma = 1.25$ and $\tilde{\lambda} = \sqrt{n/100d^2}$. Lines are theoretical predictions from Conjecture 1, dots are numerical simulations by LBFGS on (1), see Appendix D. As a function of the width, we highlight the low-p data-dependent decay of the error with exponent $1-2\gamma$ (dashed black). As a function of the number of samples, we highlight that networks at the optimal width, given in Eq. (16), (envelope of the curves) coincide with the Bayes optimal rate $-1+\frac{1}{2\gamma}$ from Defilippis et al. (2026b) (dashed black).

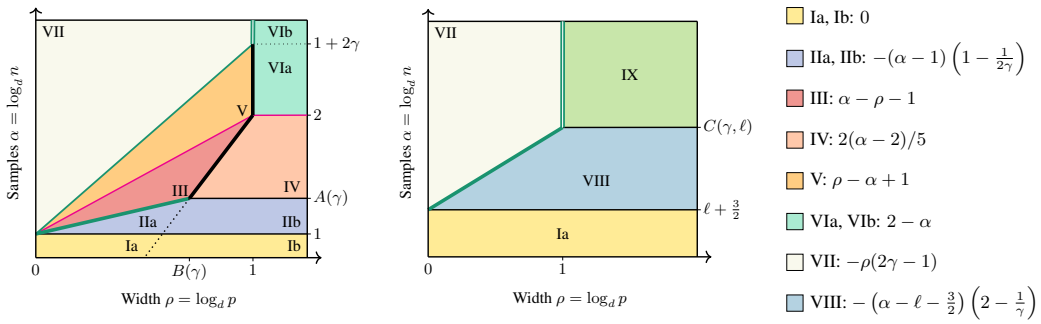
- We obtain explicit scaling laws for the excess test error as a function of $p, n, \tilde{\lambda}$; see Results 1 and 2. These laws show that width is not only a capacity parameter: it also acts as an implicit regularizer. At small width, the error is dominated by the unresolved tail of the power-law target. At larger width, finite-sample effects enter, and increasing the number of parameters can improve generalization, leave it essentially unchanged, or eventually worsen it by fitting noise-dominated directions.
- We characterize the optimal choices of width and regularization. Under mild assumptions, either optimally tuning the explicit regularization or selecting the optimal finite width achieves the Bayes-optimal excess-error rate. Thus the rank constraint induced by width can provide an implicit regularization mechanism that is, at the level of scaling exponents, as effective as an explicit optimal one.

Figure 1 summarizes these phenomena. The left panel shows the predicted and observed excess test error as a function of width, including the low-width decay controlled by the power-law exponent γ and the emergence of an optimal width. The right panel shows the corresponding sample-wise scaling: optimizing over the width traces the envelope of the curves and reaches the Bayes-optimal rate. The agreement between the state-evolution prediction and numerical optimization supports the use of Conjecture 1 as a quantitative description of the ERM.

Notations. We say that two functions $f(d) \sim g(d)$ if $\lim_{d\to+\infty} f(d)/g(d) = 1$, and that $f(d) = \Theta(g(d))$ if $cg(d) \leq f(d) \leq Cg(d)$ for some constants $c, C > 0$. Finally, we denote with $\overset{d}{\sim}$ convergence in distribution. GOE(d) denotes the ensemble of symmetric $d \times d$ matrices with Gaussian entries with mean zero and variance $1/d$ ($2/d$ on the diagonal).

2 Width-wise and sample-wise scaling laws

The main result is a characterization of the scaling of the excess test error (6) with the width p, sample size n, and effective regularization $\tilde{\lambda}$, as $d \to \infty$ with $\Delta = O(1)$. We express these laws through the exponents $\alpha, \rho > 0$ and ℓ in (7), fixing $\gamma > 1/2$ and $\Delta > 0$ (see Appendix C.2 for the noiseless case $\Delta = 0$). The resulting phase diagram is summarized in Figure 2; the derivation is sketched in Section 4 and the details, including asymptotic constants depending on γ and Δ, are given in Appendix C.1. We distinguish the under-regularized regime $\ell < \alpha/2 - 1$ and the over-regularized regime $\alpha/2 - 1 < \ell < \alpha - 3/2$. The case $\rho = 1$ is delicate, and our derivation of the scalings does not cover it: we recover it by taking the limits $\rho \to 1^{\pm}$.



(a) Under regularized $\ell < \alpha/2 - 1$ (b) Over regularized $\ell > \alpha/2 - 1$

Figure 2: Scaling laws as a function of the number of training samples $n = \Theta(d^\alpha)$, of the width of the network $p = \Theta(d^\rho)$ and of the decay of the target weights γ for the cases of low regularization (a) and high regularization (b). For each phase, we write the scaling exponent β of the test error $\mathcal{E} = \Theta(d^{\beta})$ on the right side. Major phases correspond to distinct scaling exponents β, whereas sub-phases sharing the same β are differentiated according to their spectral properties by a dotted line (see Appendix A). With green lines, we highlight the optimal width scaling (where two values of ρ are highlighted at same α, the optimal width can be either of both based on γ and Δ, see (16)). With thick lines, we highlight the effective target's rank ρ_{eff}, to the right of which the error scaling is the same as for full-width networks $p \geq d$. With a double line, we mark a transition across which the scaling exponent of the test error is discontinuous. With red line, we highlight the interpolation peak, below which the training set can be fitted almost exactly. We defined $A(\gamma) = (18\gamma - 5)/(14\gamma - 5)$, $B(\gamma) = (8\gamma - 2)/(14\gamma - 5)$ and $C(\gamma, \ell) = \ell + \gamma + 3/2$. Phases Ia, IIa, III, V, VII are newly derived scalings, while the other phases can be derived from the $p \geq d$ theory in Defilippis et al. (2026b).

Under-regularized regime. If $\ell < \alpha/2 - 1$ (low regularization), we obtain the following excess test error scaling behavior.

Result 1 (Scaling law for low regularization) *Consider any global minimum $\hat{\boldsymbol{W}}$ of (1) where the labels are generated from the model in (5) with $\Delta > 0$, and define $\hat{\boldsymbol{S}} = \hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}}/\sqrt{pd}$. Assume the target matrix $\boldsymbol{S}_\star \in \mathbb{R}^{d\times d}$ has eigenvalues $\{\sqrt{d}i^{-\gamma}/\sqrt{\zeta(2\gamma)}\}_{i=1}^d$, where ζ is Riemann's zeta function and $\gamma > 1/2$. Then for $d \gg 1$ and $\ell < \alpha/2 - 1$ we have $\mathcal{E} = \Theta(d^\beta)$, where*

$$\beta = \begin{cases} -(2\gamma-1)\rho & \text{if} \quad \rho < \min\{(\alpha-1)/(2\gamma), 1\} \\ \rho + 1 - \alpha & \text{if} \quad (\alpha-1)/(2\gamma) < \rho < \min(1, \alpha-1) \quad \text{and} \quad 1 < \alpha < 1 + 2\gamma \\ \alpha - 1 - \rho & \text{if} \quad \alpha - 1 < \rho < \rho_{\text{eff},1}(\alpha,\gamma) \quad \text{and} \quad 1 < \alpha < 2 \\ \phi_1(\alpha,\gamma) & \text{if} \quad \rho > \rho_{\text{eff},1}(\alpha,\gamma) \quad \text{and} \quad \alpha > 1 \\ 0 & \text{if} \quad \alpha < 1 \end{cases} \tag{8}$$

where, calling $A(\gamma) = (18\gamma - 5)/(14\gamma - 5) \in [9/7, 2]$ (recall $\gamma > 1/2$), we defined

$$\phi_1(\alpha,\gamma) = \begin{cases} -(\alpha-1)\left(1 - \frac{1}{2\gamma}\right) & \text{if} \quad \alpha < A(\gamma) \\ 2(\alpha-2)/5 & \text{if} \quad A(\gamma) < \alpha < 2 \\ 2 - \alpha & \text{if} \quad \alpha > 2 \end{cases}, \tag{9}$$

and

$$\rho_{\text{eff},1}(\alpha,\gamma) = \begin{cases} (\alpha-1)(4\gamma-1)/(2\gamma) & \text{if} \quad \alpha < A(\gamma) \\ (3\alpha-1)/5 & \text{if} \quad A(\gamma) < \alpha < 2 \\ 1 & \text{if} \quad \alpha > 2 \end{cases}. \tag{10}$$

In this regime, the regularization is low enough that its presence does not alter the rates. $\rho_{\text{eff},1}(\alpha,\gamma)$ can be interpreted as an effective rank of the target weights (as seen through the finite sample size n). For $\rho > \rho_{\text{eff},1}(\alpha,\gamma)$, the rates coincide with those of a full-rank student in Defilippis et al. (2026b).

Let us comment on the phases one by one. Two cases are present for all values of $\alpha > 1$.

- **Low width (Phase VII)** with data-dependent exponent $-\rho(2\gamma - 1)$. When the width scaling ρ is low enough the ERM weights have $p = \Theta(d^\rho)$ spikes in the spectrum, each correlating

with one of the $\Theta(d^\rho)$ dominant eigen-spaces of the target. The excess test error is composed of two terms, an approximation error due to the imperfect spike-to-eigen-space correlation, and a low-rank approximation error due to the $d - p$ un-recovered target eigen-spaces. The low-rank approximation error dominates and determines the asymptotic scaling of the excess test error.

- **Full width (Phases Ib, IIb, IV, VIa, VIb)** with ρ-independent exponents. For $\rho > \rho_{\text{eff},1}$, the ERM plateaus to a ρ-independent scaling. The value of the plateau depends on the number of samples. If $\alpha < 2$, the plateau showcases a non-trivial behavior first decreasing and then increasing as a function of α related to overfitting of the label noise (Phases Ib-IIb-IV). For $\alpha > 2$ instead, the plateau decreases as $-\alpha$ (Phases VIa, VIb), as the number of samples is large enough to avoid overfitting. We stress that for $\alpha < A(\gamma)$ and ρ small enough the width constraint is active (Phases Ia, IIa), but this has no effect at leading order for the error scaling.

For $\alpha > 1 + 2\gamma$ these are the only two phases. As α gets lowered, first getting below $1 + 2\gamma$ and then below 2, the other phases enter the phase diagram, providing a richer phenomenology between the low width and the full width phases.

- **Overfitting before interpolation (Phase V)** with γ-independent exponent $\rho - \alpha + 1$. The first additional behavior, present as soon as $\alpha < 1 + 2\gamma$, is a monotone increasing behavior of the excess test error in ρ. This is caused by the fact that the ERM weights have strictly less than p outlying eigen-spaces correlating with the target, while the rest of the non-zero eigenvalues are spurious, i.e. their eigenvectors do not correlate with the target and they are caused by label noise overfitting. As ρ increases, there are more and more of these spurious eigenvalues coalescing into a bulk, leading to a test error that increases with the width until the interpolation peak (the red line on the boundary between Phases V and III).
- **Decay after interpolation (Phase III)** with γ-independent exponent $\alpha - \rho - 1$. The second additional behavior, present as soon as $\alpha < 2$, is a further decay. As the width of the network increases, the bulk of spurious eigenvalues counter-intuitively shrinks, reducing the test error and providing a benefit of overparametrization.

Remarkably, in the extremely over-sampled scaling $\alpha > 2\gamma + 1$ where only the low width and full width are present, we observe a discontinuous jump in scaling between the two phases at $\rho = 1$: for $\rho \to 1^-$ (Phase VII) the error scales as $\Theta(d^{-2\gamma})$, while for $\rho > 1$ (Phase VIb) scales as $\Theta(d^{2-\alpha})$. For $\alpha < 2\gamma + 1$ instead, we observe no discontinuity at the boundary $\rho = 1$ between width-constrained models and non-width-constrained models (i.e., the boundary between Phases V and VIa, and inside Phases IV, IIb, Ib).

Over-regularized regime. If $\ell > \alpha/2 - 1$ (high regularization), we obtain the following excess test error scaling behavior.

Result 2 (Scaling law for large regularization) *In the setting of Result 1, when $d \gg 1$ and $\ell > \alpha/2 - 1$ we have $\mathcal{E} = \Theta(d^\beta)$, where*

$$\beta = \begin{cases} -(2\gamma - 1)\rho & \text{if} \quad \rho < \rho_{\text{eff},2}(\alpha, \gamma, \ell) \quad \text{and} \quad \alpha > \ell + 3/2 \\ \phi_2(\alpha, \gamma, \ell) & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha, \gamma, \ell) \quad \text{and} \quad \alpha > \ell + 3/2 \\ 0 & \text{if} \quad \alpha < \ell + 3/2 \end{cases} \tag{11}$$

where ϕ_2 is the following width-independent plateau

$$\phi_2(\alpha, \gamma, \ell) = \begin{cases} (\ell + 3/2 - \alpha)(2 - 1/\gamma) & \text{if} \quad \alpha < \ell + 3/2 + \gamma \\ 2(\ell + 2 - \alpha) & \text{if} \quad \alpha > \ell + 3/2 + \gamma \end{cases}, \tag{12}$$

and

$$\rho_{\text{eff},2}(\alpha, \gamma, \ell) = \begin{cases} (\alpha - \ell - 3/2)/\gamma & \text{if} \quad \alpha < \ell + 3/2 + \gamma \\ 1 & \text{if} \quad \alpha > \ell + 3/2 + \gamma \end{cases}. \tag{13}$$

In this over-regularized regime we observe a similar qualitative behavior as in the under-regularized regime, with two phases present for all $\alpha > \ell + 3/2$ at low width and full width with similar qualitative interpretation, separated by the threshold $\rho_{\text{eff},2}(\alpha, \gamma, \ell)$ (the boundary between Phases VII, and VIII, IX), which again can be interpreted as an effective target width. The higher regularization induces sparser weights spectra, and prevents the intermediate-ρ phases related to interpolation

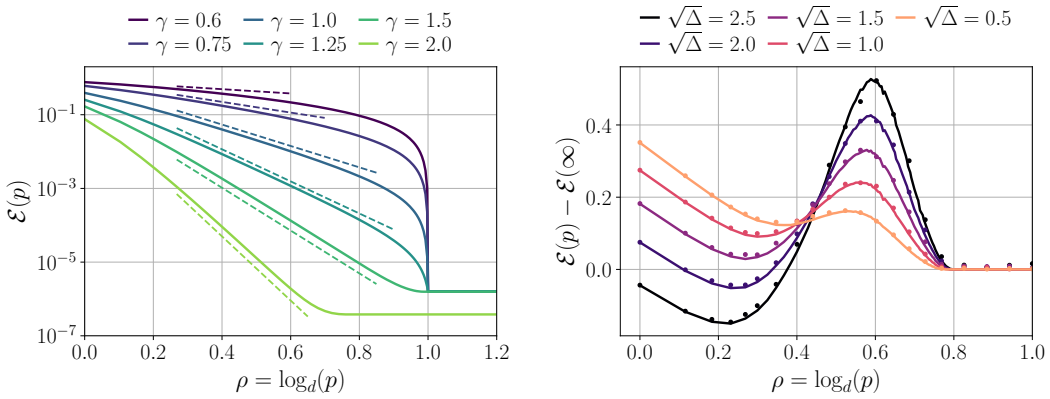


Figure 3: (Left) Test error scaling behavior in the over-regularized regime for different values of γ as a function of the width p, for fixed $d = 800$, $\sqrt{\Delta} = 0.5$, $\alpha = 4$ and $\ell = 1.2$. We observe the low-width data-dependent scaling with exponent $1 - 2\gamma$ (dashed lines, one per value of γ), as well as the large-width plateau. (Right) Dependence of the optimal width on (γ, Δ) in the low-regularization, low-samples regime. We fix $d = 400$, $\ell = -0.15$, $\alpha = 1.7$, $\gamma = 0.6$, and vary Δ. We observe that as the label noise decreases the optimal width jumps discontinuously from the position of the local minimizer to the start of the large-width plateau. In both plots, lines are theory (Conjecture 1), points numerical experiments (Appendix D).

observed at low number of samples in the under-regularized regime. At the boundary between width-constrained network and full width network $\rho = 1$ we again observe a discontinuity in the scaling of the test error for large number of samples $\alpha > \ell + 3/2 + \gamma$ (between Phases VII and IX).

Rank collapse regime (Phase Ia). If $\alpha < 1$ and $\rho < \frac{3\alpha - 1}{5}$ (extreme under-sampling) or $\ell > \alpha - 3/2$ (extreme over-regularization), the solution of the ERM (1) sets the weights of the network to zero, as if no samples were provided. The excess test error equals one at leading order, $\mathcal{E} \sim 1$. Note that if $\alpha < 1$ and $\rho > \frac{3\alpha - 1}{5}$ (Phase Ib), not all the weights are zero, but we still have $\mathcal{E} \sim 1$, where the overfitting only produces a second order correction (see Appendix C.1).

On different kinds of over-parametrization. The scaling diagram in Figure 2 allows to characterize different kinds of over-parametrization. In all under-regularized phases with $\alpha < \min(2, 1 + \rho)$, the training loss vanishes. In all phases with $\rho > \rho_{\text{eff},1/2}(\alpha, \gamma)$ (Phases Ib, IIb, IV, VIII), the network properties do not change when the width of the network is enlarged. In all phases with $\rho > 1$ (Phases VIa, VIb, IX, and some parts of Phases Ib, IIb, IV), the optimization problem (1) is benign: there are no spurious local minima.

Spectra and generalization. Appendix A provides a detailed characterization (including a schematic illustration in Figure 4) of the spectrum across each phase, where isolated spikes, the bulk, and spectral truncation correspond to learned features, overfitting noise, and regularization/finite-width constraints, respectively. This spectral perspective offers a heuristic explanation for the error scalings established in Results 1 and 2.

3 Rates for different regularization schemes

In this section, we derive scaling laws under optimal choices of regularization strength, network width, and post-training pruning. We treat ℓ_2 regularization, width constraints, and pruning as three distinct forms of regularization, and ask whether any one of them is inherently superior. We show that, across all three settings the resulting test error can achieve Bayes-optimal scaling rates (as characterized in (Defilippis et al., 2026b, Corollary 1)). This highlights a nontrivial interplay between these mechanisms: depending on the practical cost of cross-validating each form of regularization, one can recover optimal scaling behavior through multiple, interchangeable routes. For example, optimal post-training pruning requires a single training run on a convex objective, while the other regularization schemes may require multiple ones to perform cross-validation.

Optimal regularization under width constraint. From Result 1 and 2 we can see that the optimal effective regularization $\tilde{\lambda}$ is given by the scaling $\ell = \alpha/2 - 1$ for all $\rho > 0$ and $\alpha > 1$. Indeed, the over-regularized rates improve monotonically as the regularization decreases, while the under-regularized rates either do not depend on $\tilde{\lambda}$ (hence all $\ell < \alpha/2 - 1$ are optimal), or showcase a monotone increasing behavior with the width, which is sub-optimal (overfitting phase). We stress that the optimal regularization still depends on the width, as $\lambda_{\text{opt}} = \sqrt{d/p}\tilde{\lambda}_{\text{opt}} = \Theta(d^{(\alpha-\rho-1)/2})$. The associated test error then scales as

$$\log_d \mathcal{E}_{\text{opt reg}} \sim \begin{cases} -(2\gamma-1)\rho & \text{if} \quad \rho < \rho_{\text{eff},2}(\alpha,\gamma,\alpha/2-1) \quad \text{and} \quad \alpha > 1 \\ (1-\alpha)(1-1/(2\gamma)) & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha,\gamma,\alpha/2-1) \quad \text{and} \quad 1 < \alpha < 1+2\gamma \\ 2-\alpha & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha,\gamma,\alpha/2-1) \quad \text{and} \quad \alpha > 1+2\gamma \end{cases}, \tag{14}$$

For $\rho > \rho_{\text{eff},2}(\alpha,\gamma,\alpha/2-1)$, the rates obtained are Bayes optimal (the boundary between Phases VII and VIII), while for $\rho < \rho_{\text{eff},2}(\alpha,\gamma,\alpha/2-1)$ they are sub-optimal (Phase VII). As long as the network is not width-constrained, optimal regularization can achieve Bayes optimal rates.

Optimal width at fixed regularization. We are now interested in computing the optimal width, by which we mean the smallest width achieving the best test error scaling. In the over-regularized regime the test error scaling is always monotone decreasing and then plateauing in ρ, hence the optimal ρ is the smallest width in the plateauing phase, $\rho^{\text{over reg}}_{\text{opt width}} = \rho_{\text{eff},2}(\alpha,\gamma,\ell)$. The associated test error scaling is given by

$$\log_d \mathcal{E}^{\text{over reg}}_{\text{opt width}} \sim \begin{cases} (\ell+3/2-\alpha)(2-1/\gamma) & \text{if} \quad \alpha < \ell+3/2+\gamma \\ 2(\ell+2-\alpha) & \text{if} \quad \alpha > \ell+3/2+\gamma \end{cases} \tag{15}$$

which is Bayes optimal for $\ell = \alpha/2 - 1$, and is suboptimal for larger regularization.

In the under-regularized regime, the test error scaling may be non-monotone. In particular, one needs to compare the error at the end of the low-width decreasing phase (at width $\rho_{\text{eff},1}(\alpha,\gamma)$), where either the large width plateau starts or there is a local minimum of the test error scaling, with the value of the full width plateau. For $\alpha > 1 + 2\gamma$ the two coincide. For $A(\gamma) < \alpha < 1 + 2\gamma$, the local minimum at $(\alpha - 1)/(2\gamma)$ (the boundary between Phases VII and V) has smallest test error scaling. For $1 < \alpha < A(\gamma)$ the scaling of the test error at the local minimum and at the plateau are the same, so optimality is found by comparing the constants (i.e. comparing (1) the boundary between Phases VII and V with (2) Phases IIa, IIb). We find that there exists a function $G(\gamma, \Delta)$ such that

$$\rho^{\text{under reg}}_{\text{opt width}} = \begin{cases} \min\{(\alpha-1)/(2\gamma), 1\} & \text{if } \alpha > A(\gamma) \text{ or } (1 < \alpha < A(\gamma) \text{ and } G(\gamma,\Delta) > 0) \\ \rho_{\text{eff},1}(\alpha,\gamma) & \text{if } 1 < \alpha < A(\gamma) \text{ and } G(\gamma,\Delta) < 0 \end{cases}. \tag{16}$$

In particular, if $G(\gamma, \Delta) < 0$ the optimal width behaves discontinuously at $A(\gamma)$, jumping from a larger value at small sample scaling to a smaller value at larger sample scaling, revealing an interesting interplay between data structure γ and label noise Δ. To compute explicitly $G(\gamma, \Delta)$ and assess whether it can be negative for $d \gg 1$, we would need to control not only the asymptotic constant of the test error, but also that of the transition regime, which despite our best efforts, remains out of reach. We provide numerical evidence that this can be the case for $d = 400$ in Figure 3 right. The associated test error is given by

$$\log_d \mathcal{E}^{\text{under reg}}_{\text{opt width}} \sim \begin{cases} (1-\alpha)(1-1/(2\gamma)) & \text{if} \quad \alpha < 2\gamma+1 \\ 2-\alpha & \text{if} \quad \alpha > 2\gamma+1 \end{cases}, \tag{17}$$

which is Bayes optimal. Thus, as long as the network is not over-regularized, training at optimal width achieves Bayes optimal rates. In particular, optimally tuning the width regularizes the interpolation peak behavior, as we show in Figure 1 (right).

Optimal post-training pruning. As a third regularization mechanism, we consider after-training pruning. We first train a full-width network $p = d$, and write its learned matrix as $\hat{\boldsymbol{S}} = \boldsymbol{U}\text{diag}(s_1, \cdots, s_d)\boldsymbol{U}^\top$. We then apply spectral soft-thresholding $\hat{\boldsymbol{S}}_{\text{pruning}} = \boldsymbol{U}\text{diag}(\text{ReLU}(s_i - \tau_p))\boldsymbol{U}^\top$ for some threshold $\tau_p > 0$ such that there are p non-zero eigenvalues. For generic p, we give the error scaling in Appendix C.3, where it is shown that while soft thresholding and weight

decay share a similar regularization effect, their error scaling behaviors differ. We find that pruning to the optimal width scaling $\rho_{\text{opt pruning}}$ achieves Bayes optimal rates with

$$\rho_{\text{opt pruning}} = \frac{\alpha - 1}{2\gamma} \quad \text{if} \quad \ell < \alpha/2 - 1 \quad \text{and} \quad 1 < \alpha < 1 + 2\gamma. \tag{18}$$

We remark that outside the region $\ell < \alpha/2 - 1$ and $1 < \alpha < 1 + 2\gamma$ pruning does not help, because there is no overfitting. Moreover, $\rho_{\text{opt pruning}} = \rho^{\text{under reg}}_{\text{opt width}}$ in most of the parameter space, with the only possible exclusion of the region $G(\gamma, \Delta) < 0$.

4 Sketch of derivation

Results 1 and 2 are based on Conjecture 1, which we now discuss. It is derived heuristically based on the well-established replica method and approximate message passing in statistical physics (Mézard et al., 1987; Mézard & Montanari, 2009; Zdeborová & Krzakala, 2016), see details in Appendix B. While its rigorous proof remains an open challenge, numerically Conjecture 1 matches the error of numerical experiments of gradient descent at convergence with high accuracy, as can be seen in Figure 1. Moreover, in the asymptotic regime $n = \Theta(d^2)$, $p = \Theta(d)$ and $\tilde{\lambda} = \Theta(1)$, (20) coincides with the equations derived (also heuristically) in Martin et al. (2026) for the long-time behavior of gradient flow in the same model, providing an additional consistency check.

Conjecture 1 (Properties of the ERM) *In the setting of Results 1 and 2, for $n, p, d \gg 1$ we have*

$$\hat{\boldsymbol{S}} \overset{d}{\sim} \hat{\boldsymbol{S}}_{\delta,\tilde{\lambda}\epsilon} = (\boldsymbol{S}_\star + \delta \boldsymbol{Z} - \tilde{\lambda}\epsilon \boldsymbol{I}_d)^+_{(p)}\,, \quad \mathcal{E}(\hat{\boldsymbol{W}}) \sim \frac{2n}{d^2}\delta^2 - \frac{\Delta}{2}\,, \quad \frac{1}{n}\mathcal{L}(\hat{\boldsymbol{W}}) \sim \frac{\delta^2}{4\epsilon^2} + \frac{\tilde{\lambda}}{d}\,\mathrm{Tr}(\hat{\boldsymbol{S}}_{\delta,\tilde{\lambda}\epsilon})\,, \tag{19}$$

where $\boldsymbol{A}^+_{(p)}$ denotes the projection of the matrix $\boldsymbol{A}$ onto PSD, rank-p matrices[1]*, and $\boldsymbol{Z} \sim$ GOE(d). Here $(\delta, \epsilon) \in \mathbb{R}^2$ are functions of $(n, d, p, \lambda, \Delta, \gamma)$, found as a solution of the system*

$$\begin{cases} \dfrac{4n}{d^2}\,\delta - \dfrac{\delta}{\epsilon} = \partial_1 J_p(\delta, \tilde{\lambda}\epsilon), \\ 1 + \dfrac{\Delta}{2} + \dfrac{2n}{d^2}\,\delta^2 - \dfrac{\delta^2}{\epsilon} = \Big(1 - \tilde{\lambda}\epsilon\,\partial_2\Big)\, J_p(\delta, \tilde{\lambda}\epsilon), \end{cases} \quad \textit{with} \quad J_p(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d}\,\mathrm{Tr}\left(\hat{\boldsymbol{S}}^2_{\delta,\tilde{\lambda}\epsilon}\right)\,, \tag{20}$$

where ∂_1, ∂_2 indicate the partial derivative w.r.t. the first and second arguments of J_p respectively.

Conjecture 1 gives a closed state-evolution prediction for the test error, train error, and learned weights of the ERM. Given $n, p, d, \tilde{\lambda}$ and the spectrum of $\boldsymbol{S}_\star$, one solves (20) numerically and plugs the solution into (19); see Appendix D. The resulting picture is that the learned representation $\hat{\boldsymbol{S}}$ is a noisy, spectrally sparsified version of the target $\boldsymbol{S}_\star$: explicit regularization through $\tilde{\lambda}$ and the implicit rank constraint induced by the width p both bias the network towards the strongest target directions. We use these non-asymptotic equations as predictions across the scaling regimes (7), and validate them extensively by training the original quadratic network with gradient-based optimization. The agreement is excellent, already for moderate dimensions $d = 400$ and small width $p = 1$ (Figure 1).

From Conjecture 1, Results 1 and 2 can be derived by characterizing, for all possible scalings of $n, p, \tilde{\lambda}$ with d (i.e. for all possible values of α, ρ, ℓ), the solution of (20) at leading order in d, and hence deducing the leading order behavior of the quantities in (19).

To derive Conjecture 1, we first map the ERM problem to matrix compressed sensing with rank constraint (4), and extend the canonical computation (Erba et al., 2026) ($n = \Theta(d^2)$, $p \geq d$ and $\tilde{\lambda} = \Theta(1)$) to account for the rank constraint. This constraint makes the problem non-convex, and thus this extension should be thought of as the replica symmetric ansatz (Mézard et al., 1987; Mézard & Montanari, 2009) for the characterization of the global minimizer of (4). As a second step, we make the ansatz that the resulting characterization actually holds in a non-asymptotic sense, i.e. for generic scalings of $n, p, \tilde{\lambda}$. This is in line with previous work on non-width-constrained networks (Defilippis et al., 2026b), follows a long tradition in statistical physics, see Dietrich (2001) for early reports, and is corroborated by formal proof in the simpler setting of ridge regression (Defilippis et al., 2024). Proving Conjecture 1 would require controlling both the replica symmetric and the non-asymptotic ansatzes, a difficult technical challenge that goes beyond our aim in this paper: predicting simple and numerically observed sample-wise and width-wise scaling laws.

[1] $\boldsymbol{A}^+_{(p)}$ denotes the matrix $\boldsymbol{A}$, with the largest p eigenvalues ν_i mapped to $\max(\nu_i, 0)$, and the rest set to 0.

5 Related works

Our analysis leverages tools from the theory of Approximate Message Passing algorithms (AMP) (Bayati & Montanari, 2011; Donoho et al., 2013; Berthier et al., 2020; Gerbelot & Berthier, 2023; Vilucchio et al., 2025), which have been broadly applied to the study of learning problems in the high-dimensional limit (Donoho & Montanari, 2016; Rangan et al., 2016). In this work, we use the theory of AMP in a non-asymptotic regime as a heuristic extension of its standard asymptotic setting. This is in line with a common route in high-dimensional statistics, where heuristic predictions (Bordelon et al., 2020; Cui et al., 2021; Simon et al., 2023) often precede and guide later rigorous analyses (Cheng & Montanari, 2024; Misiakiewicz & Saeed, 2024; Defilippis et al., 2024). We view the rigorous justification of this non-asymptotic extension as a challenging mathematical problem in its own right, but distinct from the main goal of the present work.

Quadratic neural networks, their landscape, and sharp generalization properties were studied extensively in the full-rank case, $p \geq d$: Du & Lee (2018); Venturi et al. (2019); Soltanolkotabi et al. (2018); Gamarnik et al. (2019); Martin et al. (2024); Arjevani et al. (2026); Erba et al. (2026). In particular, their scaling laws were derived from the non-asymptotic AMP and for power-law targets in Defilippis et al. (2026b), still in the regime $p \geq d$, and extended to more general even activation functions in Defilippis et al. (2026a).

Ren et al. (2025); Ben Arous et al. (2026) study closely related hierarchical multi-index models and width-constrained quadratic networks from a dynamical perspective. While these works provide an important dynamical account of how width affects feature learning in terms of continuous-time population or online trajectories, they do not address the finite-dataset generalization problem studied in this paper, including train–test gaps, interpolation, and width-dependent implicit regularization.

The rank/width constraint regime, $1 \ll p \leq d$, was also studied in Martin et al. (2026), where the authors consider the long-time behavior of gradient descent $\boldsymbol{W}^{t+1} - \boldsymbol{W}^t = -\eta \nabla \mathcal{L}(\boldsymbol{W}^t)$ in the limit of $\eta \to 0$, $t \to \infty$ with initialization $\boldsymbol{w}_i^{t=0} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ for $1 \leq i \leq p$. The authors obtain an asymptotic characterization of the stationary states in the limit of large d and quadratic sample complexity $n \propto d^2$, which is consistent with our results in their asymptotic version.

Finally, quadratic neural networks are closely related to attention mechanisms, since both operate through bilinear projections of the inputs. Recent work (Boncoraglio et al., 2026) characterizes the learning curves of a single attention head, showing that the scaling laws for the quadratic model (Defilippis et al., 2026b) also arise for key and query width $p \geq d$. The finite-width analysis developed here extends similarly to attention heads with arbitrary width, as discussed in Appendix C.4.

6 Discussion and limitations

We studied the joint scaling of generalization with sample size, width, and regularization in a finite-width quadratic feature-learning model. Our analysis reveals a rich phase diagram, with distinct scaling regimes, plateaus, and non-monotone dependence on width. The main message is that width is not merely a capacity parameter; it acts as an implicit regularizer, selecting which directions are learned and which are suppressed. This allows us to characterize the optimal width and to compare explicit regularization by weight decay with implicit regularization by width constraint or pruning. In appropriate regimes, optimally tuning either regularization or width achieves optimal rates.

The main limitation of our approach is the price paid for mathematical tractability. We consider a deliberately stylized setting: a shallow quadratic network, Gaussian inputs, and a structured teacher. We therefore do not claim that the precise exponents transfer directly to realistic deep architectures. Rather, the value of the model is its solvability: it provides a controlled setting in which the roles of data availability, model size, target structure, and regularization can be understood explicitly.

Among the open problems, a natural direction is to turn the state-evolution prediction of Conjecture 1 into a rigorous theorem. This is not a mere technical gap. Making such finite-size state-evolution predictions fully rigorous requires significant progress in the control of AMP beyond standard asymptotic settings. In addition, the rank constraint induced by finite width makes the optimization problem non-convex, placing it beyond the existing convex AMP proof techniques. This therefore requires significant progress in the theory of AMP for non-convex problems. Another important direction is to analyze gradient descent and prove that it reaches the same solution as the ERM, as our numerical experiments very strongly suggest, even in this non-convex setting.

REFERENCES

Yossi Arjevani, Joan Bruna, Joe Kileel, Elzbieta Polak, and Matthew Trager. Geometry and optimization of shallow polynomial networks. *SIAM Journal on Applied Algebra and Geometry*, 10 (2):174–209, 2026.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4): 043404, 2026.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Jinho Baik, Gérard Ben Arous, and Sandrine Peche. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of probability*, 33(5):1643–1697, 2005.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

Mohsen Bayati, José Pereira, and Andrea Montanari. The lasso risk: asymptotic results and real world examples. *Advances in Neural Information Processing Systems*, 23, 2010.

Gerard Ben Arous, Murat Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws. *Advances in Neural Information Processing Systems*, 38:146725–146812, 2026.

Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, March 2020. ISSN 2049-8764, 2049-8772.

Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *ICML*, 2026. arxiv:2509.24914.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4345–4382, 2024a.

Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *12th International Conference on Learning Representations, ICLR 2024*, 2024b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. *arXiv preprint arXiv:2602.07488*, 2026.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Louis-Pierre Chaintron, Lénaïc Chizat, and Javier Maas. Resnets of all shapes and sizes: Convergence of training dynamics in the large-scale limit. *arXiv preprint arXiv:2603.18168*, 2026.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks. *Advances in Neural Information Processing Systems*, 37:62362–62383, 2024.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. Optimal scaling laws in learning hierarchical multi-index models. *arXiv preprint arXiv:2602.05846*, 2026a.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2026b. ICLR.

Rainer Dietrich. *Statistical Mechanics of Neural Networks: Enhancement by Weighting of Examples*. Citeseer, 2001.

David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

David L. Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, May 2013. ISSN 1091-6490.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1329–1338. PMLR, 10–15 Jul 2018.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks. *Advances in Neural Information Processing Systems*, 38:88862–88901, 2026.

Maryam Fazel, Emmanuel Candes, Ben Recht, and Pablo Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1043–1047. IEEE, 2008.

David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv preprint arXiv:1912.01599*, 2019.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

Jiaoyang Huang. Mesoscopic perturbations of large random matrices. *Random Matrices: Theory and Applications*, 7(02):1850004, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Antoine Maillard, Emanuele Troiani, Simon Martin, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, 2024.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3655–3663. PMLR, 2024.

Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026.

Leon Mirsky. A trace inequality of john von neumann. *Monatshefte für mathematik*, 79(4):303–306, 1975.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *arXiv preprint arXiv:2403.08938*, 2024.

Andrea Montanari, YC Eldar, and G Kutyniok. Graphical models concepts in compressed sensing. *Compressed Sensing*, pp. 394–438, 2012.

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 62(12):7464–7474, 2016.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in sgd learning of shallow neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R DeWeese. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *Transactions on Machine Learning Research*, 2023.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

Luca Venturi, Afonso S. Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

Matteo Vilucchio, Yatin Dandi, Matéo Pirio Rossignol, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.

Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 5757–5823. PMLR, 2025.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

A SCHEMATIC ILLUSTRATION OF THE SPECTRA

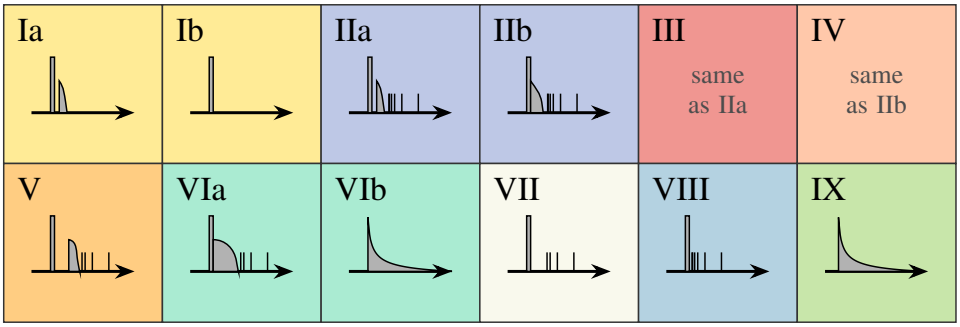


Figure 4: The spectra of $\hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}}$ in different phases.

Figure 4 illustrates the spectra of the learned weights across different phases, with detailed derivations provided in Appendix C.1.

In Phase Ib, most eigenvalues are zero, while a small non-zero fraction forms a separated bulk. Although this bulk reflects overfitting, it has only a secondary effect and does not affect the leading-order error. As regularization increases or the network width decreases, the system transitions into Phase Ia, and the bulk disappears accordingly.

In Phase IIb, the spectrum consists of several isolated spikes and a small bulk, where the spikes correspond to the learned features and the bulk reflects overfitting. Decreasing the network width drives the system into Phase IIa, where the bulk is truncated by the width constraint. Because the bulk's contribution is sub-leading in Phase IIb, Phases IIa and IIb achieve the same leading-order error. Phases III and IV exhibit qualitatively similar spectra to Phases IIa and IIb; in Phase IV, however, the bulk becomes the dominant (leading-order) term. Consequently, reducing the width transitions the system into Phase III, suppressing the bulk and thereby lowering the error.

In Phase VIa, the spectrum comprises several learned spikes and a bulk that characterizes overfitting. Increasing the sample size transitions the system into Phase VIb, where this overfitting bulk completely vanishes. Alternatively, reducing the network width truncates the bulk (Phase V), thereby mitigating overfitting and lowering the generalization error.

In Phase VIII, strong regularization completely suppresses the bulk, leaving only the spikes. This regime corresponds to underfitting, where fewer features are learned due to over regularization. Decreasing the network width transitions the system into Phase VII, where weaker spikes are truncated, leading to underfitting induced by limited representational capacity. Alternatively, increasing the sample size drives the system into Phase IX, whose spectrum is a shifted variant of Phase VIb; in this regime, the regularization shifts the feature spikes.

B Derivation of Conjecture 1: analytical test error characterization

We derive the characterization in Conjecture 1 by adapting the derivation of Erba et al. (2026) to power-law targets and width-constrained students. For the sake of the derivation, as discussed in the main text, we first assume that $n = \Theta(d^2)$, $p = \Theta(d)$, $\tilde{\lambda} = \Theta(1)$, that $\boldsymbol{S}_\star$ has a limiting spectral distribution for $d \gg 1$. We proceed under the *Replica Symmetric* (RS) assumption, a standard hypothesis in the statistical physics literature (Mézard et al., 1987; Mézard & Montanari, 2009). In the present setting, this assumption is equivalent to postulating that a rank-constrained extension of the Approximate Message Passing (AMP) algorithm introduced in Erba et al. (2026) tracks the global minimizer of the empirical risk minimization problem in the high-dimensional limit. This connection can be made rigorous when the loss is convex, as in the full-rank setting studied in Erba et al. (2026). In our rank-constrained formulation, however, the ERM problem is generally non-convex, and validating the replica-symmetric description requires additional checks, which we discuss below. In a second stage we will then assume that the derived equations can be relaxed to general scaling, and verify this by comparing with numerical experiments.

Gaussian universality and AMP. The derivation proceeds in two steps:

- **Gaussian universality:** one shows that for the asymptotic behavior of the global minimizer, the training data $\boldsymbol{G}_\mu = (\boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top - \boldsymbol{I}_d)/\sqrt{d}$ can be replaced with surrogate Gaussian data, i.e. $\tilde{\boldsymbol{G}}_\mu \sim \mathrm{GOE}(d)$. With this substitution, (4) becomes a Gaussian linear regression problem with matrix weights and data, plus a non-separable ℓ_1+PSD+rank-constraint regularization. (Maillard et al., 2024; Xu et al., 2025)
- **AMP and its analysis:** The asymptotic behavior of ERM in Gaussian linear models has been studied extensively (Bayati et al., 2010; Montanari et al., 2012; Rangan et al., 2016; Loureiro et al., 2021; Vilucchio et al., 2025), in particular through Approximate Message Passing (AMP). This framework allows to write an algorithm, the mentioned AMP, whose fixed points are stationary points of the original ERM loss, and that can be studied analytically for $d \gg 1$ through the state evolution framework. Crucially, AMP and state evolution can deal with non-separable regularization (Berthier et al., 2020; Gerbelot & Berthier, 2023), such as the one we need to treat here.

We do not repeat here the full derivation, as it is quite standard. We just mention where the role of the rank-constraint enters it, and what changes it induces. For an ERM problem of the form (4) with the data model (5), Erba et al. (2026) shows that the asymptotic behavior of the global minimum can be determined by solving the system of equations

$$\begin{cases} \widehat{\Sigma} = \dfrac{2n}{d^2}\dfrac{1}{\Sigma + \frac{1}{4}}, \\ \widehat{m} = \dfrac{2n}{d^2}\dfrac{1}{\Sigma + \frac{1}{4}}, \\ \widehat{q} = \dfrac{2n}{d^2}\dfrac{Q_\star - 2m + q + \frac{\Delta}{2}}{\left(\Sigma + \frac{1}{4}\right)^2}, \end{cases} \qquad \begin{cases} m = -2\,\partial_{\widehat{m}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}), \\ q = 4\,\partial_{\widehat{\Sigma}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}), \\ \Sigma = -4\,\partial_{\widehat{q}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}). \end{cases} \tag{21}$$

where

$$\Psi(\hat{\Sigma}, \hat{q}, \hat{m}) = \frac{1}{d}\mathbb{E}_{\boldsymbol{Z}} \min_{\boldsymbol{S}\succeq 0,\, \mathrm{rk}(\boldsymbol{S})\le p} \left[\tilde{\lambda}\,\mathrm{Tr}(\boldsymbol{S}) + \frac{\hat{\Sigma}}{4}\|\boldsymbol{S}\|_F^2 - \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{S}^\top(\hat{m}\boldsymbol{S}_\star + \sqrt{\hat{q}}\boldsymbol{Z})) \right], \tag{22}$$

and $\boldsymbol{Z} \sim \mathrm{GOE}(d)$. This requires the computation of the minimizer (which we will call also denoiser)

$$\boldsymbol{D}(\boldsymbol{Y}) = \operatorname*{argmin}_{\boldsymbol{S}\succeq 0,\, \mathrm{rk}(\boldsymbol{S})\le p} \left[\tilde{\lambda}\,\mathrm{Tr}(\boldsymbol{S}) + \frac{\hat{\Sigma}}{4}\|\boldsymbol{S}\|_F^2 - \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{S}^\top\boldsymbol{Y}) \right] \tag{23}$$

for a fixed symmetric matrix $\boldsymbol{Y} \in \mathbb{R}^{d\times d}$. Notice that m and q have a clear interpretation. For any fixed value of $\hat{q}, \hat{m}, \hat{\Sigma}, p$ and $\boldsymbol{P} = \boldsymbol{D}(\hat{m}\boldsymbol{S}_\star + \sqrt{\hat{q}}\boldsymbol{Z})$, we have that (Erba et al., 2026, Appendix

A.4.3)

$$\begin{aligned} m &= \frac{1}{d}\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Tr}(\boldsymbol{P}^\top \boldsymbol{S}_\star)\right] = -2\,\partial_{\widehat{m}}\Psi(\widehat{\Sigma},\widehat{q},\widehat{m}), \\ q &= \frac{1}{d}\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Tr}(\boldsymbol{P}^\top \boldsymbol{P})\right] = 4\,\partial_{\widehat{\Sigma}}\Psi(\widehat{\Sigma},\widehat{q},\widehat{m}), \end{aligned} \tag{24}$$

allowing to interpret m as an overlap between a ground truth matrix $\boldsymbol{S}_\star$ and a noisy version of it $\hat{m}\boldsymbol{S}_\star + \sqrt{\hat{q}}\boldsymbol{Z}$ denoised through $\boldsymbol{D}(\cdot)$, and q as its Frobenius norm squared. Consequently, we can also write the test error for any scalar multiple of such matrix as

$$\mathcal{E}(c\boldsymbol{P}) = \frac{1}{d}\|c\boldsymbol{P} - \boldsymbol{S}_\star\|_F^2 = Q_\star - 2cm + qc^2 = Q_\star + 4c\,\partial_{\widehat{m}}\Psi + 4c^2\,\partial_{\widehat{\Sigma}}\Psi \tag{25}$$

where $Q_\star = \mathrm{Tr}(\boldsymbol{S}_\star^2)/d$ and $c \in \mathbb{R}$

Computation of the denoiser and role of rank constraint. We now compute (23). We first remark that any $d \times d$ PSD matrix with rank at most $p < d$ can be decomposed spectrally as $\boldsymbol{S} = \boldsymbol{O}\boldsymbol{L}\boldsymbol{O}^\top$, where $\boldsymbol{O}$ is a $d \times d$ rotation matrix and $\boldsymbol{L} = \mathrm{diag}(L_1, \dots, L_p, 0, \dots, 0)$ with $L_1 \geq L_2 \geq \cdots \geq L_p \geq 0$. Thus

$$\boldsymbol{D}(\boldsymbol{Y}) = \underset{L_1 \geq \cdots \geq L_p \geq 0}{\mathrm{argmin}} \left[\tilde{\lambda}\sum_{i=1}^p L_i + \frac{\hat{\Sigma}}{4}\sum_{i=1}^p L_i^2 - \frac{1}{2}\underset{\boldsymbol{O}\text{ rotation}}{\mathrm{argmax}}\,\mathrm{Tr}(\boldsymbol{O}^\top \boldsymbol{L}\boldsymbol{O}\boldsymbol{Y})\right]. \tag{26}$$

By Von Neumann's inequality (Mirsky, 1975),

$$\underset{\boldsymbol{O}\text{ rotation}}{\mathrm{argmax}}\,\mathrm{Tr}(\boldsymbol{O}^\top \boldsymbol{L}\boldsymbol{O}\boldsymbol{Y}) = \sum_{i=1}^p L_i Y_i \tag{27}$$

where Y_i are the eigenvalues of $\boldsymbol{Y}$ sorted decreasingly, giving

$$\boldsymbol{D}(\boldsymbol{Y}) = \underset{L_1 \geq \cdots \geq L_p \geq 0}{\mathrm{argmin}} \left[\tilde{\lambda}\sum_{i=1}^p L_i + \frac{\hat{\Sigma}}{4}\sum_{i=1}^p L_i^2 - \frac{1}{2}\sum_{i=1}^p L_i Y_i\right]. \tag{28}$$

The minimization can be now solved independently for all $i = 1, \dots, p$ over the set $L_i \geq 0$, as the objective is a sum over functions depending on a single value of i, and checking a posteriori that the overall monotonicity constraint $L_1 \geq \cdots \geq L_p$ is satisfied. The fixed i solution gives

$$L_i = \frac{1}{\hat{\Sigma}}\mathrm{ReLU}(Y_i - 2\tilde{\lambda}) \tag{29}$$

from which we see that all the L_i are automatically sorted decreasing due to the sorting of the Y_i. Plugging back in we obtain

$$\Psi(\hat{\Sigma},\hat{q},\hat{m}) = -\frac{\hat{m}^2}{4\hat{\Sigma}}\mathbb{E}_{\boldsymbol{Z}\sim\mathrm{GOE}(d)}\frac{1}{d}\sum_{i=1}^p \mathrm{ReLU}(\nu_i - 2\tilde{\lambda}/\hat{m})^2 = -\frac{\hat{m}^2}{4\hat{\Sigma}}J_p\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right), \tag{30}$$

where we called ν_i the i-th eigenvalue sorted decreasingly of $\boldsymbol{S}_\star + \sqrt{\hat{q}}/\hat{m}\boldsymbol{Z}$, and finally used the notation of Conjecture 1. Algebraic manipulations of (21) under the change of variable $\delta = \sqrt{\hat{q}}/\hat{m}$ and $\epsilon = 2/\hat{m}$ leads to (20). We thus see that the rank constraint on the original problem induces an additional spectral cut in (30) (the sum going up to $p < d$, and not up to d as it would in the non-rank-constrained case).

This concludes the raw derivation of (20). The expression of the training loss, test error, and spectral distribution follow from the AMP/state evolution derivation (Erba et al., 2026), and are given by

$$\hat{\boldsymbol{S}} \overset{d}{\sim} (\boldsymbol{S}_\star + \delta\boldsymbol{Z} - \tilde{\lambda}\epsilon\boldsymbol{I}_d)^+_{(p)}, \quad \mathcal{E}(\hat{\boldsymbol{W}}) \sim \frac{2n}{d^2}\delta^2 - \frac{\Delta}{2}, \quad \frac{1}{n}\mathcal{L}(\hat{\boldsymbol{W}}) \sim \frac{\delta^2}{4\epsilon^2} + \frac{\tilde{\lambda}}{d}\mathrm{Tr}(\hat{\boldsymbol{S}}). \tag{31}$$

Moreover, combining (25) and (30) with $\boldsymbol{P} = (\hat{m}/\hat{\Sigma})\hat{\boldsymbol{S}}$ and $(\boldsymbol{S}_\star + \delta\boldsymbol{Z} - \tilde{\lambda}\epsilon\boldsymbol{I}_d)^+_{(p)}$, we get the alternative expression

$$\mathcal{E}(\hat{\boldsymbol{W}}) \sim Q_\star + (\delta\partial_1 + \tilde{\lambda}\epsilon\partial_2 - 1)J_p(\delta, \tilde{\lambda}\epsilon). \tag{32}$$

Validity. Now, as discussed in the main text, we take two leaps of faith. First, we assume that this derivation holds for different scaling regimes than the one in which these equations are usually derived or proven. Second, we assume that the non-convexity of the loss does not break the replica symmetric assumption The first point is tricky to control analytically, and we are not aware of any line of attack. The second point can be approached (in the strict asymptotic scaling for which the derivation formally holds) by checking the necessary (but not sufficient) replicon condition (see Vilucchio et al. (2025) for a related discussion)

$$\frac{1}{d^2}\left[2\sum_{i=1}^{d}(\partial\eta_i(\nu_i))^2+\sum_{i\neq i'}\left(\frac{\eta_i(\nu_i)-\eta_{i'}(\nu_{i'})}{\nu_i-\nu_{i'}}\right)^2\right]<\frac{2n}{d^2}\,, \tag{33}$$

where

$$\eta_i(x)=\theta(1\le i\le p)\mathrm{ReLU}(\nu_i-2\tilde{\lambda}/\hat{m})\,. \tag{34}$$

This condition is related to the linear stability of AMP at its fixed point. In order to state a sufficient condition for the replica symmetric assumption, we would have to work out the so-called replica symmetry breaking version of the analysis, as known in the physics of disordered systems (Mézard et al., 1987; Mézard & Montanari, 2009). In Appendix C.1, we discuss the replicon condition for each phase we consider in the main text, but we stress here that as soon as we enter the non-asymptotic scaling domain, there is no reason to believe that this condition would still be meaningful in the usual sense.

Pruning. The test error of the post-training pruning procedure sketched in the main text can be studied by setting $p=d$ (training at full width) and select a suitable threshold b_p such that $\hat{\boldsymbol{S}}_{\text{pruning}}=(\boldsymbol{S}_\star+\delta\boldsymbol{Z}-\tilde{\lambda}\epsilon\boldsymbol{I}_d)^+$ has rank p, where δ solve (20) for $p=d$. The corresponding generalization error is given by

$$\mathcal{E}(\hat{\boldsymbol{W}})\sim Q_\star+(\delta\partial_1+b_p\partial_2-1)J_d(\delta,b_p). \tag{35}$$

C Derivation of the scaling laws

Results 1, 2 are derived from Conjecture 1 by careful asymptotic analysis of (20). For notational simplicity we assume that $\boldsymbol{S}_\star$ has eigenvalues $\{\sqrt{d}\,i^{-\gamma}\}_{i=1}^d$ with $\gamma>1/2$, which can me mapped to the minimization objective in the main text (4) by rescaling $\Delta Q_\star\to\Delta$ and $\lambda\to\lambda/Q_\star$, where $Q_\star:=\zeta(2\gamma)$ alters only the asymptotic constants, not the scaling. The learned matrix $\hat{\boldsymbol{S}}$ in the appendix will be a factor $\sqrt{Q_\star}$ larger than in the main, and the generalization error and the loss will be $Q_\star$ larger.

The scaling laws obtained in this Appendix are a result of a simplifying assumption on the spectrum of the learned features $\hat{\boldsymbol{S}}$, which we do coherently with Defilippis et al. (2026b): calling $\mu_\delta(\cdot)$ the spectral density of $\boldsymbol{S}_\star+\delta\boldsymbol{Z}$, we approximate it as a semicircular bulk of radius 2δ of relative mass $d-K$ plus K outlying spikes. More explicitly, we model the eigenvalues of $\boldsymbol{S}_\star+\delta\boldsymbol{Z}$ as $\{\xi_i\}_{i=1,\dots d-K}\cup\{\nu_i\}_{i=1,\dots,K}$, where we assume that $\xi_i\overset{\text{i.i.d.}}{\sim}\mu_{\text{sc}}(x/\delta)/\delta$, with $\mu_{\text{sc}}(\cdot)$ the standard semi-circle law of GOE matrices

$$\mu_{\text{sc}}(x)=\frac{\sqrt{4-x^2}}{2\pi}\theta\big(|x|<2\big)\,, \tag{36}$$

and the top eigenvalues are the pushforward of $\{\sqrt{d}\,i^{-\gamma}\}_{i=1}^K$ through the BBP mapping (Baik et al., 2005; Huang, 2018) $f_\delta(x)=x+\delta^2/x$:

$$\nu_i=f_\delta\big(\sqrt{d}\,i^{-\gamma}\big)\,. \tag{37}$$

The number of spikes outside the semicircle is fixed by imposing that K is the largest integer such that

$$\sqrt{d}K^{-\gamma}\ge\delta \tag{38}$$

following again BBP theory. We expect this approximation to be good as long as $K\ll d$. In the following we will use the notation $\nu_{K+1}=\xi_1=2\delta$ to indicate the largest eigenvalue in the semicircle. Based on this simplification, we write the spectral density $\mu_\delta(\cdot)$ as

$$\mu_\delta(x)\approx\left(1-\frac{K}{d}\right)\mu_{\text{sc}}(x/\delta)/\delta+\frac{1}{d}\sum_{i=1}^{K}\delta\left(x-f_\delta\big(\sqrt{d}i^{-\gamma}\big)\right), \tag{39}$$

where $\delta(\cdot)$ is the Dirac delta function. The spectrum of $\hat{\boldsymbol{S}} = (\boldsymbol{S}_\star + \delta \boldsymbol{Z} - \tilde{\lambda}\epsilon \boldsymbol{I}_d)^+_{(p)}$ is thus given by

$$\mu(x) = \begin{cases} \mu_\delta(x + \tilde{\lambda}\epsilon), & x > x_0^+, \\ 0, & x \le x_0^+, \end{cases} \qquad x_0^+ := \max\{x_0, 0\}, \tag{40}$$

in addition to a Dirac delta in zero such that $\int_{\mathbb{R}} \mu(x) \mathrm{d}x = 1$. x_0 is a cutoff imposing that the rank of $\hat{\boldsymbol{S}}$ is at most p. In formulas this means x_0 is defined by imposing that $\hat{\boldsymbol{S}}$ has at least $\frac{d-p}{d}$ zero eigenvalues, implying

$$x_0 = \inf \left[t \,\middle|\, \int_t^\infty \mu_\delta(x + \tilde{\lambda}\epsilon) \, \mathrm{d}x \le \frac{p}{d} \right] \tag{41}$$

As long as $x_0 \le 0$, the rank constraint is not active as the PSD constraint already produces a matrix with low enough rank. With this form of the spectrum in mind, we can then study (20) in various regimes. In Appendix C.1 we study the scaling regimes for all noisy data models $\Delta > 0$, leaving the case $\Delta = 0$ to Appendix C.2. In Appendix C.3 we describe the scaling regimes of pruning.

Replicon. We discuss how the condition (33) can be specified for this target. Rewriting it in terms of density, it reads

$$\int dx \mu_\delta(x) \int dy \mu_\delta(y) \left(\frac{\eta(x) - \eta(y)}{x - y} \right)^2 < \frac{2n}{d^2}, \tag{42}$$

where

$$\eta(x) = \begin{cases} x - \tilde{\lambda}\epsilon, & x - \tilde{\lambda}\epsilon > x_0^+, \\ 0, & x - \tilde{\lambda}\epsilon \le x_0^+, \end{cases} \qquad x_0^+ := \max\{x_0, 0\}. \tag{43}$$

We then have three cases for the behaviour of this condition:

- When the rank-p cutoff has no effect, i.e. $x_0 < 0$, the spectrum is the same as the full rank case (Erba et al., 2026; Defilippis et al., 2026b), and the replicon condition is always satisfied by the convexity of the full rank optimization problem.
- When the cutoff is inside the bulk, i.e., $0 < \tilde{\lambda}\epsilon + x_0^+ \le \delta$, the left side of (33) diverges as $\int_{1/d}^L r^{-1} \, \mathrm{d}r \sim \log d$ because the mean spacing between eigenvalues is approximately d^{-1}. For this case, the replicon condition is satisfied only if $n \gg d^2 \log d$.
- When the cutoff is outside the bulk, i.e. $\nu_k < \tilde{\lambda}\epsilon + x_0^+ \le \nu_{k+1}$ for some $0 \le k \le K$, then the left side of (33) is of order

$$\frac{1}{d^2} \sum_{i \le k < j} \left(\frac{\eta(\nu_i) - \eta(\nu_j)}{\nu_i - \nu_j} \right)^2 \approx \frac{1}{d^2} \sum_{i \le k < j} \left(\frac{\sqrt{d} i^{-\gamma}}{\sqrt{d}\gamma(i-j) i^{-\gamma-1}} \right)^2 \approx \frac{k^2}{d^2 \gamma} \log d. \tag{44}$$

  Then replicon condition (33) thus holds as long as $k^2 \ll n / \log d$. This will be always true for the case $\gamma > 1/2$ we are analyzing, because by (Defilippis et al., 2026a, Theorem 3.2) the number of spikes is upper bounded by $K \le (n/d)^{1/2\gamma}$, and thus for $\gamma > 1/2$ the replicon condition is always satisfied once the cutoff is outside the bulk.

C.1 Derivation of the excess test error for width-constrained networks

In this section we derive the phase diagram. We will always assume $\rho > 0 \implies p \gg 1$.

We now proceed with the solution of the equations. As we derived in Appendix B, the excess test error is given by (32)

$$\mathcal{E}(\hat{\boldsymbol{W}}) \sim Q_\star + (\delta \partial_1 + \tilde{\lambda}\epsilon \partial_2 - 1) J_p(\delta, \tilde{\lambda}\epsilon) \,, \tag{45}$$

where $J_p(\delta, \tilde{\lambda}\epsilon)$ is the second spectral moment of $\hat{\boldsymbol{S}}$

$$J_p(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d} \operatorname{Tr} \left(\hat{\boldsymbol{S}}^2_{\delta, \tilde{\lambda}\epsilon} \right) . \tag{46}$$

The solution will consist in enumerating all the possible configurations for the spectrum of $\hat{\boldsymbol{S}}$ in terms of the magnitude of the spikes, the width of the bulk δ and the shift $\tilde{\lambda}\epsilon$, and computing the

leading behavior of J_p in each regime. These conditions will be translated into conditions on the network width, dimension, regularization and the number of samples, and one can verify that we spanned the whole space of possibilities when considering the union of all these regions.

We start by remarking that if $x_0^+ = 0$, that is when the ERM solution already has a rank less than p, we will obtain the same scaling results from (Defilippis et al., 2026b), since the limiting factor is not the width of the network but the regularization shift of the eigenvalues. We will thus focus on the cases $x_0^+ > 0$, in which the cutoff is active.

In the derivation, we will make use of the approximations valid for $s \ll 1$

$$\int_{2-s}^{2} \mu_{\rm sc}(x)\,\mathrm{d}x = \frac{2}{3\pi}s^{3/2} - \frac{1}{20\pi}s^{5/2} + \mathcal{O}(s^{7/2})\,, \tag{47}$$

as well as

$$\begin{aligned}\int_{2-s}^{2} (x-a)^2\,\mu_{\rm sc}(x)\,\mathrm{d}x = \frac{2}{3\pi}(2-a)^2 s^{3/2} \quad &+ \left[-\frac{4}{5\pi}(2-a) - \frac{(2-a)^2}{20\pi}\right] s^{5/2} \\ &+ \left[\frac{2}{7\pi} + \frac{(2-a)}{14\pi} - \frac{(2-a)^2}{448\pi}\right] s^{7/2} + \mathcal{O}(s^{9/2})\end{aligned} \tag{48}$$

We define the two natural parts into which J_p can be decomposed: one part for the semicircle contribution ($J_p^{(1)}$), and one for the spike contribution ($J_p^{(2)}$)

$$\begin{cases} J_p^{(1)}(\delta, \tilde{\lambda}\epsilon) = \displaystyle\int_{\tilde{\lambda}\epsilon + x_0^+}^{2\delta} (x - \tilde{\lambda}\epsilon)^2 \mu_{{\rm sc},\delta}(x)\mathrm{d}x = \delta^2 \int_{2-s}^{2} \left(x - \frac{\tilde{\lambda}\epsilon}{\delta}\right)^2 \mu_{\rm sc}(x)\mathrm{d}x \\ J_p^{(2)}(\delta, \tilde{\lambda}\epsilon) = \dfrac{1}{d} \displaystyle\sum_{i=1}^{\min(K,p)} \left[\sqrt{d} i^{-\gamma} + \frac{\delta^2}{\sqrt{d} i^{-\gamma}} - \tilde{\lambda}\epsilon\right]^2 \end{cases} \tag{49}$$

where $s := 2 - \frac{\tilde{\lambda}\epsilon + x_0^+}{\delta}$ and $1 \le K \le d$ is the largest integer such that $\sqrt{d}K^{-\gamma} \ge \delta$.

Non-rank-constrained phases. Before proceeding, we first recall the phase diagram from Defilippis et al. (2026b) (with λ replaced by $\tilde{\lambda}$ and the phases numbered according to our notation). Once the p is larger than the rank of the non-constrained ERM solution, which we compute for each phase below, the rank constraint is inactive and we obtain the same error rates. In general, the rank of the ERM can be computed by summing the amount of spikes outside of the bulk of the spectrum, plus the bulk contribution.

- **Phase Ib**: the excess error is given by

$$\mathcal{E} = Q^\star \quad \text{if} \quad \tilde{\lambda} \ll \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \ll d \quad \text{and} \quad p \ge d. \tag{50}$$

  The rank of the ERM solution is given by

$$d \int_{\tilde{\lambda}\epsilon}^{2\delta} \mu_{{\rm sc},\delta}(x)dx \approx \frac{2d}{3\pi}\left(2 - \frac{\tilde{\lambda}\epsilon}{\delta}\right)^{3/2} \approx \frac{2d}{3\pi}\left(\frac{15\pi n}{4d^2}\right)^{3/5} = \Theta(n^{3/5}d^{-1/5}), \tag{51}$$

  where we use $2 - \frac{\tilde{\lambda}\epsilon}{\delta} \approx \left(\frac{15\pi n}{4d^2}\right)^{2/5} \ll 1$ from Defilippis et al. (2026b) and (47). Therefore, we can obtain

$$\boxed{\mathcal{E} = Q^\star \quad \text{if} \quad \tilde{\lambda} \ll \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \ll d \quad \text{and} \quad p \gg n^{3/5}d^{-1/5}}. \tag{52}$$

- **Phase Ia**: the excess error is given by

$$\mathcal{E} = Q^\star \quad \text{if} \quad \tilde{\lambda} \gg \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \gg \tilde{\lambda}d^{3/2} \quad \text{and} \quad p \ge d. \tag{53}$$

In this phase the rank of the ERM solution is zero, the rank constraint is always inactive, and we can write

$$\boxed{\mathcal{E} = Q^\star \quad \text{if} \quad \tilde{\lambda} \gg \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \gg \tilde{\lambda} d^{3/2}}, \tag{54}$$

or equivalently

$$\boxed{\beta = 0 \quad \text{if} \quad \alpha < \ell + \frac{3}{2} \quad \text{and} \quad \ell > \frac{\alpha}{2} - 1}. \tag{55}$$

- **Phase VIII**: the excess error is given by

$$\mathcal{E} = \frac{2\gamma}{2\gamma - 1} \left(\tilde{\lambda} \frac{d^{3/2}}{4n} \right)^{\frac{2\gamma-1}{\gamma}} \quad \text{if} \quad \max\left(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{3/2+\gamma}} \right) \ll \tilde{\lambda} \ll \frac{n}{d^{3/2}} \quad \text{and} \quad p \geq d. \tag{56}$$

Moreover, there are $K = (4n/\tilde{\lambda} d^{3/2})^{1/\gamma}$ spikes, and thus if $p \gg K \implies \tilde{\lambda} \gg \frac{n}{d^{3/2} p^\gamma}$ the results remain unchanged. Therefore, we have

$$\boxed{\mathcal{E} = \frac{2\gamma}{2\gamma - 1} \left(\tilde{\lambda} \frac{d^{3/2}}{4n} \right)^{\frac{2\gamma-1}{\gamma}} \quad \text{if} \quad \max\left(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{3/2} p^\gamma}, \frac{n}{d^{3/2+\gamma}} \right) \ll \tilde{\lambda} \ll \frac{n}{d^{3/2}}}. \tag{57}$$

Equivalently, we can write

$$\beta = -\left(\alpha - \ell - \frac{3}{2} \right) \left(2 - \frac{1}{\gamma} \right) \quad \text{if} \quad \ell + \frac{3}{2} < \alpha < \ell + \frac{3}{2} + \gamma \min(\rho, 1). \tag{58}$$

This corresponds to Phase VIII in Figure 2(b).

- **Phase IX**: all spikes are recovered, thus the ERM solution has full rank, and the excess error is

$$\boxed{\mathcal{E} = \frac{\tilde{\lambda}^2 d^4}{16 n^2} \quad \text{if} \quad \tilde{\lambda} \gg \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \gg \tilde{\lambda} d^{3/2+\gamma} \quad \text{and} \quad p \geq d}, \tag{59}$$

or equivalently

$$\boxed{\beta = -2(\alpha - \ell - 2) \quad \text{if} \quad \ell > \frac{\alpha}{2} - 1 \quad \text{and} \quad \alpha > \ell + \frac{3}{2} + \gamma \quad \text{and} \quad \rho \geq 1}. \tag{60}$$

- **Phases IIb and IV**: the excess error is given by

$$\mathcal{E} = \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n} \right)^{1 - \frac{1}{2\gamma}} + \frac{\Delta}{7} \left(\frac{15\pi}{4} \right)^{2/5} \left(\frac{n}{d^2} \right)^{2/5} \quad \text{if} \quad d \ll n \ll d^2 \quad \text{and} \quad p \geq d \quad \text{and} \quad \tilde{\lambda} \ll \frac{\sqrt{n}}{d}. \tag{61}$$

The rank of the ERM solution is also given by (51), because the total number of spikes is subleading. Therefore, we can write

$$\boxed{\mathcal{E} = \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n} \right)^{1 - \frac{1}{2\gamma}} + \frac{\Delta}{7} \left(\frac{15\pi}{4} \right)^{2/5} \left(\frac{n}{d^2} \right)^{2/5} \quad \text{if} \quad d \ll n \ll d^2 \quad \text{and} \quad p \gg n^{3/5} d^{-1/5} \quad \text{and} \quad \tilde{\lambda} \ll \frac{\sqrt{n}}{d}}. \tag{62}$$

- **Phases VIa and VIb**: all spikes are recovered, thus the ERM solution has full rank, and the excess error is given by

$$\boxed{\mathcal{E} = \Theta\left(\frac{d^2}{n} \right) \quad \text{if} \quad \tilde{\lambda} \ll \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \gg d^2 \quad \text{and} \quad p \geq d}, \tag{63}$$

or equivalently

$$\boxed{\beta = 2 - \alpha \quad \text{if} \quad \ell < \frac{\alpha}{2} - 1 \quad \text{and} \quad \alpha > 2 \quad \text{and} \quad \rho \geq 1}. \tag{64}$$

Phase VII: rank constraint active, spikes only. We start by considering the case in which the final spectrum is composed by exactly p spikes. This is the case if the p-th target eigenvalue $\sqrt{d}p^{-\gamma}$ is pushed out of the bulk by the BBP mapping, $\sqrt{d}p^{-\gamma} \gg \delta$. Once this p eigenvalues are out, the rank-constraint will set to zero the remainder of the possibly pushed-out target eigenvalues, as well as the semicircular bulk. Additionally, the PSD+regularization constraint will set to zero all eigenvalues smaller than $\tilde{\lambda}\epsilon$, requiring for the p-th spike to be non zero that $\sqrt{d}p^{-\gamma} + \frac{\delta^2}{\sqrt{d}p^{-\gamma}} - \tilde{\lambda}\epsilon \gg 0$. Overall, this implies $\sqrt{d}p^{-\gamma} \gg \max(\delta, \tilde{\lambda}\epsilon)$. Notice that here we used $\gg$ where a priori only $\geq$ was needed in both conditions: this is a simplifying assumption that would alter the final phase diagram Figure 2 only in a zero measure set.

We can then write J_p as

$$J_p(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d}\sum_{i=1}^{p}\left[\sqrt{d}i^{-\gamma} + \frac{\delta^2}{\sqrt{d}i^{-\gamma}} - \tilde{\lambda}\epsilon\right]^2 \tag{65}$$

$$= \sum_{i=1}^{p}\left[i^{-2\gamma} + \frac{\delta^4}{d^2 i^{-2\gamma}} + \frac{1}{d}\lambda^2\epsilon^2 - 2\tilde{\lambda}\epsilon\frac{i^{-\gamma}}{\sqrt{d}} - 2\tilde{\lambda}\epsilon\frac{\delta^2}{d^{3/2}i^{-\gamma}} + 2\frac{\delta^2}{d}\right] \tag{66}$$

$$\approx Q^\star - \frac{p^{1-2\gamma}}{2\gamma-1} + \frac{\delta^4}{d^2}\frac{p^{1+2\gamma}}{1+2\gamma} + \frac{p}{d}(\lambda^2\epsilon^2 + 2\delta^2) - \frac{2\tilde{\lambda}\epsilon}{\sqrt{d}}\sum_{i=1}^{p}\left[i^{-\gamma} + \frac{\delta^2}{di^{-\gamma}}\right] \tag{67}$$

The derivatives with respect to $\tilde{\lambda}\epsilon$ and δ are

$$\begin{cases} \tilde{\lambda}\epsilon\partial_2 J_p(\delta, \tilde{\lambda}\epsilon) \approx 2\frac{p}{d}\lambda^2\epsilon^2 - \frac{2\tilde{\lambda}\epsilon}{\sqrt{d}}\sum_{i=1}^{p}\left[i^{-\gamma} + \frac{\delta^2}{di^{-\gamma}}\right] \\ \delta\partial_1 J_p(\delta, \tilde{\lambda}\epsilon) \approx 4\frac{\delta^4}{d^2}\frac{p^{1+2\gamma}}{1+2\gamma} + 4\frac{p}{d}\delta^2 - \frac{4\tilde{\lambda}\epsilon\delta^2}{d^{3/2}}\sum_{i=1}^{p} i^{\gamma} \end{cases} \tag{68}$$

such that we can write (20) as

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx 4\frac{\delta^4}{d^2}\frac{p^{1+2\gamma}}{1+2\gamma} + 4\frac{p}{d}\delta^2 - \frac{4\tilde{\lambda}\epsilon\delta^2}{d^{3/2}}\frac{p^{1+\gamma}}{1+\gamma} \ll \frac{p^{1-2\gamma}}{2\gamma-1} \\ Q^\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx Q^\star - \frac{p^{1-2\gamma}}{2\gamma-1} + \frac{\delta^4}{d^2}\frac{p^{1+2\gamma}}{1+2\gamma} + \frac{p}{d}(2\delta^2 - \lambda^2\epsilon^2) \approx Q^\star - \frac{p^{1-2\gamma}}{2\gamma-1} \end{cases} \tag{69}$$

$$\implies \mathcal{E}(p) \approx \frac{p^{1-2\gamma}}{2\gamma-1} \approx Q^\star - \sum_{i=1}^{p} i^{-2\gamma}, \tag{70}$$

where we used the conditions for having p spikes visible ($\sqrt{d}p^{-\gamma} \gg \max(\delta, \tilde{\lambda}\epsilon)$) to compare the terms and keep only the leading ones ($\propto p^{1-2\gamma}$). The first equation gives in this limit $\epsilon \approx \frac{d^2}{4n}$, while the second gives at zeroth order $\delta_0 \approx \frac{d}{2}\sqrt{\frac{\Delta}{n}}$, using that $p \gg 1$. The condition $\max(\delta_0, \tilde{\lambda}\epsilon) \ll \sqrt{d}p^{-\gamma}$ then translates to $n \gg \max\left(\Delta d p^{2\gamma}, \frac{1}{4}\lambda d^{3/2}p^{\gamma}\right)$. Thus the excess error is

$$\boxed{\mathcal{E} = \frac{1}{2\gamma-1}p^{1-2\gamma} \quad \text{if} \quad n \gg \max\left(dp^{2\gamma}, \lambda d^{3/2}p^{\gamma}\right).} \tag{71}$$

We can thus write this phase using the notation of the main text, also resolving the maximum, as

$$\boxed{\beta = \rho(1-2\gamma) \quad \text{if} \quad \begin{cases} \alpha > 1 + 2\gamma\rho \quad \text{and} \quad \ell < \frac{\alpha}{2} - 1 \\ \alpha > \ell + \frac{3}{2} + \gamma\rho \quad \text{and} \quad \ell > \frac{\alpha}{2} - 1 \end{cases}} \tag{72}$$

Phases IIa, III and V: rank constraint active, small bulk. We now consider the case in which the final spectrum is composed by $1 \ll K \ll p$ spikes, and by a slice of the semi-circular bulk. For this to be the case, we need to impose that:

- The p-th spike is not pushed out of the bulk by the BBP mapping, i.e. $\sqrt{d}p^{-\gamma} \ll \delta$, while the first spike is, i.e. $\sqrt{d} \gg \delta$. The fact that just K spikes are out imposes that $K^\gamma \approx \sqrt{d}/\delta$.
- We also need to impose that the rank cutoff is active, i.e. $x_0 > 0$ and that the overall cutoff at $x_0 + \tilde{\lambda}\epsilon$ is inside the bulk, i.e. $x_0 + \tilde{\lambda}\epsilon < 2\delta$. We will assume that $s := 2 - \frac{x_0+\tilde{\lambda}\epsilon}{\delta} \ll 1$.
- Finally, we need to impose that the remaining bulk is small compared with the non-width-constrained networks: $s \ll t := 2 - \frac{\tilde{\lambda}\epsilon}{\delta}$.

The spike contribution to J_p reads

$$J_p^{(2)}(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d}\sum_{i=1}^{K}\left[\sqrt{d}i^{-\gamma} + \frac{\delta^2}{\sqrt{d}i^{-\gamma}} - \tilde{\lambda}\epsilon\right]^2 \tag{73}$$

$$= \sum_{i=1}^{K}\left[i^{-2\gamma} + \frac{\delta^4}{d^2 i^{-2\gamma}} + \frac{1}{d}\tilde{\lambda}^2\epsilon^2 - 2\tilde{\lambda}\epsilon\frac{i^{-\gamma}}{\sqrt{d}} - 2\tilde{\lambda}\epsilon\frac{\delta^2}{d^{3/2}i^{-\gamma}} + 2\frac{\delta^2}{d}\right] \tag{74}$$

$$\approx Q^\star + \frac{K^{1-2\gamma}}{1-2\gamma} + \frac{K}{d}(\tilde{\lambda}^2\epsilon^2 + 2\delta^2) + \frac{\delta^4}{d^2}\frac{K^{1+2\gamma}}{1+2\gamma} - \frac{2\tilde{\lambda}\epsilon\delta^2}{d^{3/2}}\sum_{i=1}^{K} i^\gamma - \frac{2\tilde{\lambda}\epsilon}{\sqrt{d}}\sum_{i=1}^{K} i^{-\gamma} \tag{75}$$

For the bulk piece, using (48) and $s \ll t$ to get the leading order in s, we have

$$J_p^{(1)}(\delta, \tilde{\lambda}\epsilon) \approx \frac{2\delta^2}{3\pi}t^2 s^{3/2} - \frac{4\delta^2}{5\pi}t s^{5/2} + \frac{2\delta^2}{7\pi}s^{7/2} \approx \frac{2\delta^2}{3\pi}t^2 s^{3/2}\,, \tag{76}$$

Before computing the derivatives, we need to recall that $K^\gamma \approx \frac{\sqrt{d}}{\delta}$. We also need to consider how x_0 depends on p using (47) and (41):

$$p = K + d\int_{\tilde{\lambda}\epsilon + x_0}^{2\delta} \mu_{sc,\delta}(x)\mathrm{d}x = K + d\int_{2-s}^{2} \mu_{sc}(x)\mathrm{d}x \approx \frac{2d}{3\pi}s^{3/2}, \tag{77}$$

where for the last approximation we use $s \ll 1$ and $K \ll p$.

We use these relations to rewrite J_p and compute the derivatives

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx Q^\star + \frac{K^{1-2\gamma}}{1-2\gamma} + \frac{K}{d}(\tilde{\lambda}^2\epsilon^2 + 2\delta^2) + \frac{\delta^4}{d^2}\frac{K^{1+2\gamma}}{1+2\gamma} - \frac{2\tilde{\lambda}\epsilon\delta^2}{d^{3/2}}\sum_{i=1}^{K} i^\gamma - \frac{2\tilde{\lambda}\epsilon}{\sqrt{d}}\sum_{i=1}^{K} i^{-\gamma} + \frac{p}{d}t^2\delta^2 \tag{78}$$

$$\approx Q^\star + K^{1-2\gamma}\left[\frac{1}{1-2\gamma} + \frac{\tilde{\lambda}^2\epsilon^2}{\delta^2} + 2 + \frac{1}{1+2\gamma} - \left(2 - \frac{\tilde{\lambda}\epsilon}{\delta}\right)^2\right] - \frac{2\tilde{\lambda}\epsilon}{\delta}\frac{K^{1-2\gamma}}{1+\gamma} - \frac{2\tilde{\lambda}\epsilon}{\delta}\frac{K^{1-2\gamma}}{1-\gamma} + \frac{p}{d}t^2\delta^2 \tag{79}$$

$$\approx Q^\star + \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}\left[\frac{2}{1-4\gamma^2} - 2 - \frac{4\tilde{\lambda}\epsilon}{\delta}\frac{\gamma^2}{1-\gamma^2}\right] + \frac{p}{d}(2\delta - \tilde{\lambda}\epsilon)^2 \tag{80}$$

$$\implies \tilde{\lambda}\epsilon\partial_2 J_p(\delta, \tilde{\lambda}\epsilon) \approx -\frac{4\tilde{\lambda}\epsilon}{\delta}\frac{\gamma^2}{1-\gamma^2}\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} - \frac{2p}{d}\tilde{\lambda}\epsilon(2\delta - \tilde{\lambda}\epsilon) \tag{81}$$

$$\text{and} \quad \delta\partial_1 J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{2\gamma - 1}{\gamma}\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}\left[\frac{2}{1-4\gamma^2} - 2 - \frac{4\tilde{\lambda}\epsilon}{\delta}\frac{\gamma^2}{1-\gamma^2}\right] + \frac{4\tilde{\lambda}\epsilon}{\delta}\frac{\gamma^2}{1-\gamma^2}\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + \frac{4p}{d}\delta(2\delta - \tilde{\lambda}\epsilon) \tag{82}$$

$$= C\,(\gamma, t)\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + \frac{4p}{d}\delta^2 t \tag{83}$$

where $C(\gamma,t)=O(1)$ is a constant that depends on the value of t and γ. Now we can write the state evolution equations (20) as

$$\begin{cases} 4\frac{n}{d^2}\delta^2-\frac{\delta^2}{\epsilon}\approx C\left(\gamma,t\right)\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}+\frac{4p}{d}\delta^2 t \\ Q^\star+\frac{\Delta}{2}+2\frac{n}{d^2}\delta^2-\frac{\delta^2}{\epsilon}\approx Q^\star+\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}\left[\frac{2}{1-4\gamma^2}-2\right]+\frac{4p}{d}\delta^2 t-\frac{p}{d}\delta^2 t^2 \end{cases} \tag{84}$$

$$\implies \mathcal{E}(p)\approx\frac{p}{d}\delta^2 t^2+\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}\left[C\left(\gamma,t\right)-\frac{2}{1-4\gamma^2}+2\right] \tag{85}$$

We now need to solve the state evolution equations. Notice that $K\gg 1$ implies $\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}\ll 1$. There are two cases.

- The first case is when $\frac{p}{d}\delta^2 t,\frac{p}{d}\delta^2 t^2\ll 1$. Then (84) reduces to

$$\begin{cases} 4\frac{n}{d^2}\delta^2-\frac{\delta^2}{\epsilon}\approx 0 \\ Q^\star+\frac{\Delta}{2}+2\frac{n}{d^2}\delta^2-\frac{\delta^2}{\epsilon}\approx 0 \end{cases}, \tag{86}$$

  from which we obtain the leading order solution $\epsilon\approx\frac{d^2}{4n},\delta^2\approx\frac{\Delta d^2}{4n}$. This gives the excess risk

$$\mathcal{E}=\frac{\Delta pd}{n}\left(1-\frac{\tilde{\lambda}d}{4\sqrt{\Delta n}}\right)^2+\left[C\left(\gamma,t\right)-\frac{2}{1-4\gamma^2}+2\right]\left(\frac{\Delta d}{4n}\right)^{1-\frac{1}{2\gamma}}, \tag{87}$$

  where we use $t:=2-\frac{\tilde{\lambda}\epsilon}{\delta}=2-\frac{\tilde{\lambda}d}{\sqrt{\Delta n}}$. The conditions defining the boundaries of this phase are

$$\begin{cases} t>0 \\ s\ll t \\ 1\ll K\ll p \\ \frac{p}{d}\delta^2 t,\frac{p}{d}\delta^2 t^2\ll 1 \end{cases}\implies\begin{cases} \tilde{\lambda}<4\sqrt{\Delta}\frac{\sqrt{n}}{d} \\ p\ll d\left(2-\frac{\tilde{\lambda}d}{\sqrt{\Delta n}}\right)^{3/2} \\ 1\ll n\ll dp^{2\gamma} \\ \frac{n}{pd}\gg 2-\frac{\tilde{\lambda}d}{\sqrt{\Delta n}},\frac{pd}{n}\left(1-\frac{\tilde{\lambda}d}{4\sqrt{\Delta n}}\right)^2\ll 1 \end{cases} \tag{88}$$

  As a special case[2], we can write

$$\boxed{\mathcal{E}=\frac{\Delta pd}{n}\quad\text{if}\quad pd\ll n\ll p^{2\gamma}d\quad\text{and}\quad\tilde{\lambda}\ll\frac{\sqrt{n}}{d}}. \tag{89}$$

  Equivalently,

$$\boxed{\beta=\rho-\alpha+1\quad\text{if}\quad 1+2\gamma\rho>\alpha>1+\rho\quad\text{and}\quad\ell<\frac{\alpha}{2}-1} \tag{90}$$

  This corresponds to Phase V.

- The second case is when $\epsilon\gg\frac{d^2}{n}$ and $\frac{p}{d}\delta^2 t^2\ll\frac{p}{d}\delta^2 t=\Theta(1)$. Then (84) reduces to

$$\begin{cases} 4\frac{n}{d^2}\delta^2\approx\frac{4p}{d}\delta^2 t \\ Q^\star+\frac{\Delta}{2}+2\frac{n}{d^2}\delta^2\approx Q^\star+\frac{4p}{d}\delta^2 t \end{cases}. \tag{91}$$

  The leading order solution is thus $t\approx\frac{n}{pd}$ and $\delta^2\approx\frac{\Delta d^2}{4n}$ with $\epsilon=\frac{2\delta(1-t)}{\tilde{\lambda}}$. Taking it into (85), we obtain

$$\mathcal{E}=\frac{\Delta n}{4pd}+\left[C\left(\gamma,t\right)-\frac{2}{1-4\gamma^2}+2\right]\left(\frac{\Delta d}{4n}\right)^{1-\frac{1}{2\gamma}}. \tag{92}$$

[2]In fact (88) also includes some interesting boundary cases, which are dropped for simplicity because they are not the focus of our paper.

The conditions defining the boundaries of this phase are

$$\begin{cases} t > 0 \\ s \ll t \\ 1 \ll K \ll p \\ \frac{p}{d}\delta^2 t \ll \frac{p}{d}\delta^2 t^2 = \Theta(1) \\ \epsilon \gg \frac{d^2}{n} \end{cases} \implies \begin{cases} p \ll n^{3/5} d^{-1/5} \\ 1 \ll n \ll d p^{2\gamma} \\ n \ll pd \\ \tilde{\lambda} \ll \frac{\sqrt{n}}{d} \end{cases} \tag{93}$$

In conclusion, this phase can be written as

$$\boxed{\mathcal{E} = \frac{\Delta n}{4pd} + \frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)} \left(\frac{\Delta d}{4n}\right)^{1-\frac{1}{2\gamma}} \quad \text{if} \quad d \ll n \ll pd \quad \text{and} \quad p \ll n^{3/5} d^{-1/5} \quad \text{and} \quad \tilde{\lambda} \ll \frac{\sqrt{n}}{d}}, \tag{94}$$

where the constant $\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)}$ is obtained by $C(\gamma, 0) = \frac{8\gamma^2}{(\gamma+1)(2\gamma+1)}$ given that $t \ll 1$ in the phase. This corresponds to Phases III and IIa of Figure 2(a).
Combining (94) and (62), we can obtain the following two cases:

$$\boxed{\beta = \alpha - \rho - 1 \quad \text{if} \quad 1 + \rho > \alpha > \max\left(1 + \rho\frac{2\gamma}{4\gamma - 1}, \frac{5\rho + 1}{3}\right) \quad \text{and} \quad \ell < \frac{\alpha}{2} - 1}, \tag{95}$$

and

$$\boxed{\beta = -(\alpha - 1)\left(1 - \frac{1}{2\gamma}\right) \quad \text{if} \quad \min\left(\frac{18\gamma - 5}{14\gamma - 5}, 1 + \rho\frac{2\gamma}{4\gamma - 1}\right) > \alpha > 1 \quad \text{and} \quad \ell < \frac{\alpha}{2} - 1}. \tag{96}$$

They correspond to Phases III and II, respectively.

- The case in which $\frac{p}{d}\delta^2 t, \frac{p}{d}\delta^2 t^2 \gg 1$ is treated later when considering the interpolation peak.

Phase Ia: rank constraint active, no spikes. In this regime, the spectrum of $\hat{\boldsymbol{S}}$ is without spikes $K = 0$ ($\delta \gg \sqrt{d}$), thus we have

$$\mu_\delta(x) \approx \mu_{\rm sc}(x/\delta)/\delta\,, \tag{97}$$

The moment J_p becomes

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \int_{x_0^+}^{2\delta - \tilde{\lambda}\epsilon} x^2 \mu_{\rm s.c.}\big((x + \tilde{\lambda}\epsilon)/\delta\big)/\delta \,\mathrm{d}x = \delta^2 \int_{x_0^+/\delta + \tilde{\lambda}\epsilon/\delta}^{2} (x - \tilde{\lambda}\epsilon/\delta)^2 \mu_{\rm s.c.}(x)\,\mathrm{d}x\,. \tag{98}$$

Calling $t := 2 - \tilde{\lambda}\epsilon/\delta$ and $s := t - x_0^+/\delta$, we do the scaling ansatz $s \ll t \ll 1$ (i.e. the rank constraint dominates the spectral cut). This gives

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{2\delta^2}{3\pi} t^2 s^{3/2}\,. \tag{99}$$

Using the definition of the cutoff (41) we get

$$\frac{p}{d} = \int_{\tilde{\lambda}\epsilon + x_0^+}^{2\delta} \mu_\delta(x)\,\mathrm{d}x \;\approx\; \frac{2}{3\pi} s^{3/2}\,, \tag{100}$$

which gives

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{p}{d} t^2 \delta^2. \tag{101}$$

The derivatives of J_p include many terms, but the leading ones for $t \ll 1$ are given by

$$\partial_1 J_p(\delta, \tilde{\lambda}\epsilon) \approx 4\frac{p}{d} t\delta\,, \quad \text{and} \quad \tilde{\lambda}\epsilon \partial_2 J_p(\delta, \tilde{\lambda}\epsilon) \approx -4\frac{p}{d} t\delta^2\,, \tag{102}$$

where we further used that $\tilde{\lambda}\epsilon \approx 2\delta$ (due to $t \ll 1$). Plugging this into (20) we obtain

$$\begin{cases} \frac{4n}{d^2}\,\delta^2 - \frac{\delta^2}{\epsilon} \approx 4\frac{p}{d}t\delta^2, \\ Q^* + \frac{\Delta}{2} + \frac{2n}{d^2}\,\delta^2 - \frac{\delta^2}{\epsilon} \approx 4\frac{p}{d}t\delta^2, \end{cases} \tag{103}$$

where we use $t \ll 1$ again to state that the term J_p is much smaller than its derivatives. The first equation of (103) subtracted by the second yields the prediction for the test error

$$\mathcal{E} := 2\frac{n}{d^2}\delta^2 - \frac{\Delta}{2} \approx Q^*\,, \tag{104}$$

which is the same value of error as if $n = 0$. Solving for δ we get $\delta \approx \sqrt{(\Delta/2 + Q^*)d^2/n}$. To solve for t, we make another ansatz: $\epsilon \gg d^2/n$. Plugging this into the first equation of (103)

$$\frac{4n}{d^2}\delta^2 \approx 4\frac{p}{d}t\delta^2, \tag{105}$$

we can solve for t, obtaining $t \approx n/(pd)$.

The various assumptions that we made in the derivation lead to the following boundaries

$$\begin{cases} s \ll t \ll 1 \\ \epsilon \gg \frac{d^2}{n} \\ \delta \gg \sqrt{d} \end{cases} \implies \begin{cases} p \ll n^{3/5}d^{-1/5} \\ \tilde{\lambda} \ll \sqrt{\frac{n}{d^2}} \\ n \ll d. \end{cases} \tag{106}$$

This corresponds to Phase Ia in Figure 2(a).

Combining it with (52), we obtain

$$\boxed{\mathcal{E} = Q^\star \quad \text{if} \quad \tilde{\lambda} \ll \sqrt{\frac{n}{d^2}} \quad \text{and} \quad n \ll d}, \tag{107}$$

or equivalently

$$\boxed{\beta = 0 \quad \text{if} \quad \alpha < 1 \quad \text{and} \quad \ell > \frac{\alpha}{2} - 1}. \tag{108}$$

Interpolation peak. We start from (84) and (85) with the ansatz that $\frac{p}{d}\delta^2 t, \frac{p}{d}\delta^2 t^2 \gg 1$:

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx \frac{4p}{d}\delta^2 t \\ Q^\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx Q^\star + \frac{4p}{d}\delta^2 t - \frac{p}{d}\delta^2 t^2 \end{cases} \tag{109}$$

$$\implies \mathcal{E}(p) \approx \frac{p}{d}\delta^2 t^2, \tag{110}$$

where we use the fact that $\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} \ll 1$ as long as $K \gg 1$. (109) can be rewritten as

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = 8\frac{p}{d}\delta^2 - \frac{4p}{d}\tilde{\lambda}\epsilon\delta \\ Q_\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = Q_\star + \frac{4p}{d}\delta^2 - \frac{p}{d}\tilde{\lambda}^2\epsilon^2. \end{cases} \tag{111}$$

Assuming $n = 2pd$ and $\tilde{\lambda}\epsilon \ll \delta$ (i.e. $t \approx 2$), we can find the solution

$$\begin{cases} \delta \approx \left(\frac{\Delta^2 d}{4p\tilde{\lambda}}\right)^{1/3},\ \epsilon \approx \frac{\delta^2}{\Delta} \approx \left(\frac{d\sqrt{\Delta}}{4p\tilde{\lambda}}\right)^{2/3} \\ \mathcal{E} \approx \frac{4p}{d}\delta^2 \approx \tilde{\lambda}^{-2/3}\left(\frac{4p\Delta^4}{d}\right)^{1/3} \end{cases} \tag{112}$$

The conditions defining the boundaries of this phase are

$$\begin{cases} n = 2pd \\ s \ll 1 \\ \frac{p}{d}\delta^2 t, \frac{p}{d}\delta^2 t^2 \gg 1 \\ 1 \ll K \ll p \end{cases} \implies \begin{cases} n = 2pd \\ p \ll d \\ \frac{1}{p\sqrt{d}} \ll \tilde{\lambda} \ll \sqrt{\frac{p}{d}}. \end{cases} \tag{113}$$

In this phase the excess error then reads

$$\boxed{\mathcal{E} = 2\left(\frac{d\Delta^2}{16p}\right)^{2/3} \tilde{\lambda}^{-2/3} \quad \text{if} \quad n = 2pd \quad \text{and} \quad \frac{1}{p\sqrt{d}} \ll \tilde{\lambda} \ll \sqrt{\frac{p}{d}} \quad \text{and} \quad p \ll d}. \tag{114}$$

When $\tilde{\lambda} \ll \frac{1}{p\sqrt{d}}$, $K \gg 1$ is not satisfied and we have no spikes outside the bulk so the contribution from $J_p^{(1)}$ vanishes. Using the same assumptions on the bulk part as in (84), we get instead

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = 8\frac{p}{d}\delta^2 - \frac{4p}{d}\tilde{\lambda}\epsilon\delta \\ Q_\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = \frac{4p}{d}\delta^2 - \frac{p}{d}\tilde{\lambda}^2\epsilon^2. \end{cases} \tag{115}$$

from which we can easily see that the only difference is the constant (since the constant $\frac{\Delta}{2}$ is replaced with $Q^\star + \frac{\Delta}{2}$ in the equations), which gives

$$\boxed{\mathcal{E} = 2\left(\frac{d(\Delta + 2Q^\star)^2)}{16p}\right)^{2/3} \tilde{\lambda}^{-2/3} \quad \text{if} \quad n = 2pd \quad \text{and} \quad \tilde{\lambda} \ll \frac{1}{p\sqrt{d}} \quad \text{and} \quad p \ll d}. \tag{116}$$

Combining (114) and (116) we obtain the interpolation peak:

$$\boxed{\beta = -\frac{2}{3}\ell + \frac{\rho - 1}{3} \quad \text{for} \quad \ell < \frac{\alpha}{2} - 1 \quad \text{and} \quad \alpha = \rho + 1 \quad \text{and} \quad \rho < 1} \tag{117}$$

C.2 Derivation of the excess test error for width-constrained, noiseless networks

Non-rank-constrained phases. We first recall the results in Defilippis et al. (2026b).

- **Phase I**: Phases Ia and Ib in Appendix C.1 also hold for $\Delta = 0$, which gives:

$$\boxed{\mathcal{E} = Q^\star, \quad \text{if} \quad n \ll \max(d, \tilde{\lambda} d^{3/2})}, \tag{118}$$

  or equivalently

$$\boxed{\beta = 0 \quad \text{if} \quad \alpha < \min\left(1, \ell + \frac{3}{2}\right)}. \tag{119}$$

- **Phase V**: the excess error is given by

$$\mathcal{E} = \frac{2\gamma}{2\gamma - 1}\left(\tilde{\lambda}\frac{d^{3/2}}{4n}\right)^{\frac{2\gamma-1}{\gamma}} \quad \text{if} \quad \max\left(\frac{d^{\gamma-3/2}}{n^{\gamma-1}}, \frac{n}{d^{\gamma+\frac{3}{2}}}\right) \ll \tilde{\lambda} \ll \frac{d^{3/2}}{n} \quad \text{and} \quad p \geq d. \tag{120}$$

  Moreover, there are $K \approx (4n/\tilde{\lambda} d^{3/2})^{1/\gamma}$. The error is the same if $p \gg K \implies \tilde{\lambda} \gg \frac{n}{d^{3/2} p^\gamma}$, which gives

$$\boxed{\mathcal{E} = \frac{2\gamma}{2\gamma - 1}\left(\tilde{\lambda}\frac{d^{3/2}}{4n}\right)^{\frac{2\gamma-1}{\gamma}} \quad \text{if} \quad \max\left(\frac{d^{\gamma-3/2}}{n^{\gamma-1}}, \frac{n}{d^{\gamma+\frac{3}{2}}}, \frac{n}{d^{\frac{3}{2}} p^\gamma}\right) \ll \tilde{\lambda} \ll \frac{d^{3/2}}{n}}, \tag{121}$$

  or equivalently

$$\boxed{\beta = -\left(\alpha - \ell - \frac{3}{2}\right)\left(2 - \frac{1}{\gamma}\right) \quad \text{if} \quad \ell > \alpha + (1 - \alpha)\gamma - \frac{3}{2} \quad \text{and} \quad \ell + \frac{3}{2} < \alpha < \ell + \frac{3}{2} + \gamma \max(\rho, 1)}. \tag{122}$$

- **Phase III**: the ERM solution has full rank and the excess error is given by

$$\boxed{\mathcal{E} = \frac{\tilde{\lambda}^2 d^4}{16 n^2} \quad \text{if} \quad \tilde{\lambda} \ll \frac{n}{d^{\gamma+3/2}} \quad \text{and} \quad n \gg d^2 \quad \text{and} \quad p \geq d}, \tag{123}$$

or equivalently

$$\boxed{\beta = -2(\alpha - \ell - 2) \quad \text{if} \quad \alpha > \max(2, \ell + \gamma + \frac{3}{2}) \quad \text{and} \quad \rho \geq 1}. \tag{124}$$

- **Phase IIb**: the excess error is given by

$$\mathcal{E} = 2^{-2\gamma+1} \left(\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)} \right)^{2\gamma} \left(\frac{d}{n} \right)^{2\gamma-1} \quad \text{if} \quad \tilde{\lambda} \ll \frac{d^{\gamma-3/2}}{n^{\gamma-1}} \quad \text{and} \quad d \ll n \ll d^2 \quad \text{and} \quad p \geq d. \tag{125}$$

Because the rank of the ERM solution is still (51), we can write

$$\mathcal{E} = 2^{-2\gamma+1} \left(\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)} \right)^{2\gamma} \left(\frac{d}{n} \right)^{2\gamma-1} \quad \text{if} \quad p \gg n^{3/5} d^{-1/5}. \tag{126}$$

Phase IV: rank constraint active, spikes only The analysis of Phase VII in Appendix C.1 holds also for $\Delta = 0$, which gives

$$\begin{cases} \frac{4n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx 0, \\ Q^\star + \frac{2n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} \approx Q_\star - \frac{1}{2\gamma-1} p^{1-2\gamma}. \end{cases} \implies \begin{cases} \mathcal{E} := 2\frac{n}{d^2}\delta^2 \approx \frac{1}{2\gamma-1} p^{1-2\gamma} \\ \delta \approx \sqrt{\frac{p^{1-2\gamma} d^2}{2(2\gamma-1)n}}, \ \epsilon \approx \frac{d^2}{4n} \end{cases} \tag{127}$$

Since the leading order solution for δ is different, we need to compute the new boundaries from the conditions we assumed on the spectrum

$$\begin{cases} \sqrt{d} p^{-\gamma} \gg \tilde{\lambda}\epsilon + x_0^+ > 2\delta \\ \sqrt{d} p^{-\gamma} \gg \max(2\delta, \tilde{\lambda}\epsilon) \end{cases} \implies \begin{cases} n \gg pd \quad \text{if} \quad \tilde{\lambda} \ll \frac{p^{\frac{1}{2}-\gamma}\sqrt{n}}{d} \\ n \gg \tilde{\lambda} d^{3/2} p^\gamma \quad \text{if} \quad \tilde{\lambda} \gg \frac{p^{\frac{1}{2}-\gamma}\sqrt{n}}{d} \end{cases} \tag{128}$$

One can rewrite the regularization boundary to make it independent of p by using that at the boundary $n \approx pd \approx \tilde{\lambda} d^{3/2} p^\gamma \iff \alpha = 1 + \rho = \ell + \frac{3}{2} + \rho\gamma$, this gives the boundary $\ell_c := \alpha(1-\gamma) + \gamma - \frac{3}{2}$. We can thus write this phase as

$$\boxed{\mathcal{E} = \frac{1}{2\gamma-1} p^{1-2\gamma} \quad \text{if} \quad \tilde{\lambda} \ll \frac{n}{p^\gamma d^{3/2}} \quad \text{and} \quad n \gg pd}, \tag{129}$$

or equivalently

$$\boxed{\beta = -\rho(2\gamma-1) \quad \text{if} \quad \begin{cases} \alpha > 1 + \rho \quad \text{and} \quad \ell < \alpha(1-\gamma) + \gamma - \frac{3}{2} \\ \alpha > \ell + \frac{3}{2} + \rho\gamma \quad \text{and} \quad \ell > \alpha(1-\gamma) + \gamma - \frac{3}{2} \end{cases}}. \tag{130}$$

Phase IIa: rank constraint active, before interpolation. Phase IIa corresponds to (84) with $\Delta = 0$. We make an ansatz that $\frac{p}{d}\delta^2 t^2 \ll (\delta/\sqrt{d})^{2-1/\gamma}$, and then from (84) we approximately have

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = 4\frac{p}{d}\delta^2 t + (C_1(\gamma)(2 - 1/\gamma) + C_2(\gamma)) \left(\frac{\delta}{\sqrt{d}} \right)^{2-\frac{1}{\gamma}} \\ Q^\star + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = Q^\star + 4\frac{p}{d}\delta^2 t + (C_1(\gamma) + C_2(\gamma)) \left(\frac{\delta}{\sqrt{d}} \right)^{2-\frac{1}{\gamma}}, \end{cases} \tag{131}$$

for constants $C_1(\gamma) := -\frac{1}{2\gamma-1} + 6 - \frac{4}{1-\gamma} - \frac{4}{1+\gamma} + \frac{1}{1+2\gamma}$, $C_2(\gamma) := -4 + \frac{4}{1-\gamma} + \frac{4}{1+\gamma}$. We make another ansatz that $\frac{1}{\epsilon} \ll \frac{n}{d^2}$. The difference between the two equations in (131) gives

$$2\frac{n}{d^2}\delta^2 = \frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)}\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}. \tag{132}$$

Taking it back to (131), we obtain the leading order solution $t = \frac{cn}{pd}$ for some constant c depending on γ. We can solve δ from (132):

$$\delta = 2^{-\gamma}\left(\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)}\right)^{\gamma}\frac{d^{\gamma+1/2}}{n^\gamma}, \tag{133}$$

which gives

$$\mathcal{E} := 2\frac{n}{d^2}\delta^2 = 2^{-2\gamma+1}\left(\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)}\right)^{2\gamma}\left(\frac{d}{n}\right)^{2\gamma-1}. \tag{134}$$

ϵ is given by $\epsilon \approx \frac{2\delta}{\tilde{\lambda}}$. The conditions defining the boundaries of this phase are

$$\begin{cases} \frac{p}{d}\delta^2 t^2 \ll (\delta/\sqrt{d})^{2-1/\gamma} \\ s \ll t \ll 1 \\ \frac{1}{\epsilon} \ll \frac{n}{d^2} \\ \delta \ll \sqrt{d} \end{cases} \implies \begin{cases} p \ll n^{3/5}d^{-1/5} \\ d \ll n \ll pd \\ \tilde{\lambda} \ll \frac{d^{\gamma-3/2}}{n^{\gamma-1}} \end{cases} \tag{135}$$

Combining it with (126), the excess error reads

$$\boxed{\mathcal{E} = 2^{-2\gamma+1}\left(\frac{24\gamma^3}{(4\gamma^2-1)(\gamma+1)}\right)^{2\gamma}\left(\frac{d}{n}\right)^{2\gamma-1} \quad \text{if} \quad d \ll n \ll d^2 \quad \text{and} \quad \tilde{\lambda} \ll \frac{d^{\gamma-3/2}}{n^{\gamma-1}} \quad \text{and} \quad p \gg \frac{n}{d},} \tag{136}$$

or equivalently

$$\boxed{\beta = -(\alpha-1)(2\gamma-1) \quad \text{if} \quad 1 < \alpha < 2 \quad \text{and} \quad \ell < \alpha + (1-\alpha)\gamma - \frac{3}{2} \quad \text{and} \quad \rho > \alpha - 1} \tag{137}$$

in Phase II.

Summary The phase diagram is shown in Figure 5. For the under regularized case (Figure 5(a)), when fixing p, d and increasing n, the error decreases until $n = pd$, and then becomes a constant. When fixing n, d and increasing p, the error decreases as $p^{2\gamma-1}$ and becomes a constant after $p = \frac{n}{d}$. Thus large width and lower regularization is always preferred.

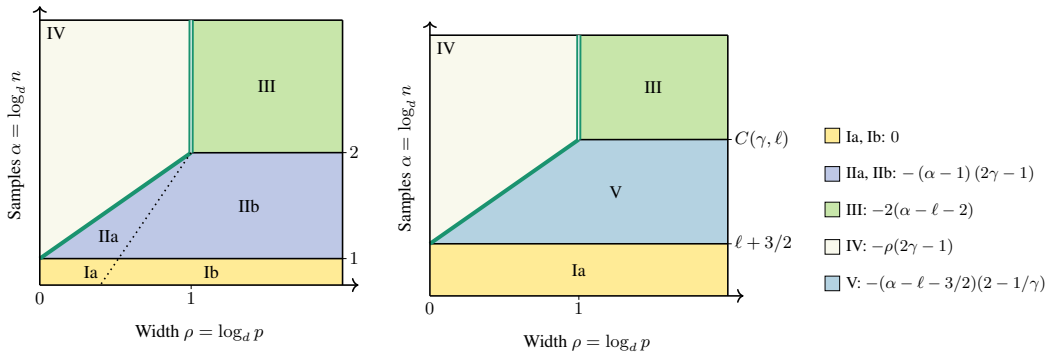
The spectra in the noiseless phases match those of the noisy phases shown in Figure 4 under the following correspondence: Ia $\leftrightarrow$ Ia, Ib $\leftrightarrow$ Ib, IIa $\leftrightarrow$ IIa, IIb $\leftrightarrow$ IIb, III $\leftrightarrow$ IX, IV $\leftrightarrow$ VII, and V $\leftrightarrow$ VIII.

C.3 Derivation of the excess test error for pruning

Now we consider pruning, i.e., first train a full width student and then apply soft-thresholding using $\mu \to \text{ReLU}(\mu - \mu_0)$ for some smallest $\mu_0 \geq 0$ such that at most p eigenvalues are non-zero. In this way the excess risk is given by $Q_\star + (\delta\partial_\delta + b\partial_b - 1)J(\delta, b)$ for some suitably chosen $b \geq \lambda\epsilon$. The phases in the following thus refer to the same phases as in Defilippis et al. (2026b).

For $n \ll d$ the excess test error is always $Q_\star$ and pruning has no effect. For $d \ll n \ll d^2$, the excess test error is given by the following

- When $\lambda \gg n/d^{3/2}$ the model is in phase I (all eigenvalues are zero). Thus pruning has no effect.
- When $\sqrt{n/d^2} \ll \lambda \ll n/d^{3/2}$, the model is in phase II, and there are $(4n/\lambda d^{3/2})^{1/\gamma}$ spikes. Therefore, when $p \gg (4n/\lambda d^{3/2})^{1/\gamma}$ nothing is pruned, and when $p \ll (4n/\lambda d^{3/2})^{1/\gamma}$, the error scales as $p^{1-2\gamma}$.



(a) Under regularized $\ell < \alpha + (1 - \alpha)\gamma - \frac{3}{2}$

(b) Over regularized $\ell > \alpha + (1-\alpha)\gamma - \frac{3}{2}$

Figure 5: Phase diagram of the noiseless estimation task. The dashed line and the blue lines represent the same meaning as in Figure 2. See Figure 7 for numerical verifications.

- When $\lambda \ll \sqrt{n/d^2}$, the model is in phases IV or V. There are $(n/d)^{1/2\gamma}$ spikes, a bulk, and other eigenvalues are zero. When $p/d \gg \int_{\lambda\epsilon}^{2\delta} \mu_{sc}(x/\delta)/\delta dx = \Theta((n/d^2)^{3/5})$, nothing is pruned. When $p \ll (n/d)^{1/2\gamma}$ the error scales as $p^{1-2\gamma}$. When $(n/d)^{1/2\gamma} \ll p \ll \Theta(d(n/d^2)^{3/5})$, the overfitting error depends on

$$J_1(\delta, b) := \int_b^{2\delta} \mu_{sc}(x/\delta)/\delta (x-b)^2 dx = \delta^2 \frac{16 t^{7/2}}{105\pi}, \tag{138}$$

where $t := 2 - \frac{b}{\delta}$ satisfies $\frac{p}{d} = \int_b^{2\delta} \mu_{sc}(x/\delta)/\delta dx = \frac{2}{3\pi} t^{3/2}$. Thus the total error scales as

$$\mathcal{E} = \frac{2^{5/3} \cdot 3^{4/3} \cdot \pi^{4/3}}{35} \left(\frac{p}{d}\right)^{7/3} \frac{\Delta d^2}{4n} + \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n}\right)^{1 - \frac{1}{2\gamma}}, \tag{139}$$

where we use the solution $\delta^2 = \frac{\Delta d^2}{4n}$ from Defilippis et al. (2026b). Consequently, the error is optimal (i.e. $(d/n)^{1-\frac{1}{2\gamma}}$) for $(n/d)^{1/2\gamma} \ll p \ll d^{10/7}(n/d)^{\frac{3}{14\gamma}}$. Note that we always have $(n/d)^{1/2\gamma} \ll d^{10/7}(n/d)^{\frac{3}{14\gamma}}$ for $n \ll d^{1+5\gamma}$.

For $n \gg d^2$, the excess test error is given by the following

- When $\lambda \gg \max(\sqrt{n/d^2}, n/d^{3/2+\gamma})$, the model is in phase II. Thus when $p \gg (4n/\lambda d^{3/2})^{1/\gamma}$ nothing is pruned, and when $p \ll (4n/\lambda d^{3/2})^{1/\gamma}$, the error scales as $p^{1-2\gamma}$.
- When $\lambda \ll n/d^{3/2+\gamma}$ and $n \gg d^{2\gamma}$, the model is in phases III or VIb. There are d spikes and thus the error scales as $p^{1-2\gamma}$.
- When $\lambda \ll \sqrt{n/d^2}$ and $d \ll n \ll d^{2\gamma}$, the model is in VIb. There are $(n/d)^{1/2\gamma}$ spikes and a bulk. When $p \ll (n/d)^{1/2\gamma}$ the error scales as $p^{1-2\gamma}$. Otherwise the error scales in the same way

$$\mathcal{E} = \frac{2^{5/3} \cdot 3^{4/3} \cdot \pi^{4/3}}{35} \left(\frac{p}{d}\right)^{7/3} \frac{\Delta d^2}{4n} + \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n}\right)^{1 - \frac{1}{2\gamma}}, \tag{140}$$

as in the previous case.

C.4 Generalization to the attention model

As shown by Boncoraglio et al. (2026), the scaling phase diagram obtained for full-rank quadratic networks in Defilippis et al. (2026b) remains very similar for the attention index model defined as

$$\hat{f}_{\text{sq}}(\boldsymbol{x}_{\text{in}}; W) = A_W(\boldsymbol{x}_{\text{in}})\boldsymbol{x}_{\text{in}}, \text{ or } \hat{f}_{\text{lb}}(\boldsymbol{x}_{\text{in}}; W) = A_W(\boldsymbol{x}_{\text{in}}) \tag{141}$$

for the seq2seq and seq2lab tasks, where $\boldsymbol{x}_{\rm in} \in \mathbb{R}^{T\times d}$ is a sequence of T tokens

$$A_W(\boldsymbol{x}_{\rm in}) = \sigma\left(\frac{\boldsymbol{x}_{\rm in} W W^\top \boldsymbol{x}_{\rm in}^\top - \mathbb{E}_{\rm tr}[\boldsymbol{x} W W^\top \boldsymbol{x}^\top]}{d\sqrt{p}}\right) \in \mathbb{R}^{T\times T} \tag{142}$$

with weights $W \in \mathbb{R}^{d\times p}$. $\sigma : \mathbb{R}^{T\times T} \to \mathbb{R}^{T\times T}$ is the activation (e.g. softmax) and $A_W(\boldsymbol{x})$ is the attention matrix with tied key and query matrices. $\mathbb{E}_{\rm tr}$ refers to the empirical average over the training set.

We consider a target function that lies within the expressivity class of the architecture in Eq. (142), and restrict the class of (possibly noisy) target functions to the ones of the form (respectively for the seq2seq and seq2lab tasks)

$$\boldsymbol{x}^\mu_{\rm out} = \sigma_*\left(R(\boldsymbol{x}_{\rm in})\right)\boldsymbol{x}_{\rm in} \quad \text{and} \quad y^\mu = \sigma_*\left(R(\boldsymbol{x}_{\rm in})\right)\,, \tag{143}$$

where $R(\boldsymbol{x}_{\rm in}) \in \mathbb{R}^{T\times T}$ is a centered pre-activation matrix, and σ_* is a possibly different activation function. We choose:

$$R(\boldsymbol{x}_{\rm in})_{ab} = \frac{\boldsymbol{x}_{{\rm in,a}}^\top S_\star \boldsymbol{x}_{\rm in,b} - \delta_{ab}\,{\rm Tr}(S_\star)}{\sqrt{d}} + \sqrt{\frac{\Delta}{2-\delta_{ab}}}\xi^\mu\,, \tag{144}$$

where $S_\star$ is the target function weight matrix with eigenvalues $\{\sqrt{d} i^{-\gamma}\}_{i=1}^d$.

We learn W by *empirical risk minimization* of the square loss with ℓ_2 regularization (weight decay), the latter being commonly used in large language models. We consider the following target functions for the seq2seq and seq2lab tasks: $\hat{W} = \arg\min_W[\mathcal{L}^{\rm data}(W) + \lambda\|W\|_F^2]$, where respectively for the two tasks

$$\begin{cases} \mathcal{L}^{\rm data}_{\rm sq}(W) := \frac{1}{d}\sum_{\mu=1}^n ||\boldsymbol{x}^\mu_{\rm out} - \hat{f}_{\rm sq}(\boldsymbol{x}^\mu_{\rm in}; W)||_F^2\,, \\ \mathcal{L}^{\rm data}_{\rm lb}(W) := \sum_{\mu=1}^n ||y^\mu - \hat{f}_{\rm lb}(\boldsymbol{x}^\mu_{\rm in}; W)||_F^2\,. \end{cases} \tag{145}$$

We measure the performance of the learned network through the test errors

$$\begin{aligned} e_{\rm test} &= \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_{\rm in},\boldsymbol{x}_{\rm out}}||\boldsymbol{x}_{\rm out} - \hat{f}(\boldsymbol{x}_{\rm in}; \hat{W})||_F^2, \\ e_{\rm test} &= \mathbb{E}_{\boldsymbol{x}_{\rm in},y}||y - \hat{f}(\boldsymbol{x}_{\rm in}; \hat{W})||_F^2\,, \end{aligned} \tag{146}$$

where $\mathbb{E}$ stands for an average over an appropriate test set.

Then we conjecture that the following excess test error satisfies the same phase diagram as in the main text

$$\mathcal{E} := e_{\rm test} - e_{\rm base},\ \text{where } e_{\rm base} := \mathcal{I}(m_0, m_0/Q_\star) \tag{147}$$

for the interpolation regime ($\tilde{\lambda} \ll \sqrt{n/d^2}$ and $n \ll dp$ such that the training loss is vanishing and the minimum is flat) and

$$e_{\rm base} := \inf_m \mathcal{I}(m, m/Q_\star) \tag{148}$$

otherwise. Here we define

$$\mathcal{I}(m,q) = \mathbb{E}_{z_0,z,\xi}||\tilde{\sigma}_*(z_0 + \sqrt{\Delta/2}\zeta) - \tilde{\sigma}(z)||^2, \tag{149}$$

where $(z_0, z) \sim \mathcal{N}\left(0, \begin{pmatrix} Q_0 & m \\ m & q \end{pmatrix}\right)$, $\zeta \sim \mathcal{N}(0,1)$. $\tilde{\sigma}_*$ is the effective teacher activation and $\tilde{\sigma}$ is the effective student activation, defined as $\tilde{\sigma}(A) := \sigma(\{\sqrt{1+\delta_{ab}}\,A_{ab}\}_{ab})$ for $A \in \mathbb{R}^{T\times T}$ and similarly $\tilde{\sigma}_*(A) := \sigma_*(\{\sqrt{1+\delta_{ab}}\,A_{ab}\}_{ab})$.

Finally $m_0 := \eta Q_\star$ is defined as the solution of the following equation

$$\eta = -\frac{1}{2}\left[\frac{\partial\mathcal{K}(m,q)}{\partial q}|_{\eta Q_\star, \eta^2 Q_\star}\right]^{-1}\frac{\partial\mathcal{K}(m,q)}{\partial m}|_{\eta Q_\star, \eta^2 Q_\star}, \tag{150}$$

where

$$\mathcal{K}(m,q) := \mathbb{E}_{z_0,z,\zeta} \inf_{h\in \arg\inf ||\tilde{\sigma}_*(z_0+\sqrt{\Delta/2}\zeta) - \tilde{\sigma}(\cdot)||^2} ||h - z||^2 \tag{151}$$

with $\zeta \sim \mathcal{N}(0,1)$ independent of (z_0, z).

D DETAILS OF NUMERICAL IMPLEMENTATIONS

Here we describe our numerical implementation. The code can be found in the repository `https://anonymous.4open.science/r/WidthQuadraticNetworks`.

Appendix D.1 shows how we solved the system (20) numerically to obtain the theoretical predictions for our figures, and are always plotted with continuous lines.

Appendix D.2 describes how we conducted the experiments, which we draw using dots with error bars indicating the standard deviation among at least two realizations for each value of the parameters. In all the plots that we show, the error bars are too small to be distinguished from the markers, we thus attribute the tiny deviations from the analytical theory either to failures of the training algorithm or to finite size effects that are expected to disappear by increasing d.

We include a working code to reproduce our results. The theory can be reproduced on a modern laptop. We ran our implementation on a MacBook Pro with Apple M4 Pro and 24/48GB of RAM. Some of the experiments require large memory, and they were conducted on a cluster with up to 1TB of RAM (56k CPU hours).

D.1 THEORETICAL LINES

To plot theoretical predictions, we solve (20) in the form (21), i.e.

$$\begin{cases} \widehat{\Sigma} = \dfrac{2n}{d^2}\,\dfrac{1}{\Sigma + \frac{1}{4}}, \\ \widehat{m} = \dfrac{2n}{d^2}\,\dfrac{1}{\Sigma + \frac{1}{4}}, \\ \widehat{q} = \dfrac{2n}{d^2}\,\dfrac{Q_\star - 2m + q + \frac{\Delta}{2}}{\left(\Sigma + \frac{1}{4}\right)^2}, \end{cases} \qquad \begin{cases} m = -2\,\partial_{\widehat{m}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}), \\ q = 4\,\partial_{\widehat{\Sigma}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}), \\ \Sigma = -4\,\partial_{\widehat{q}}\Psi(\widehat{\Sigma}, \widehat{q}, \widehat{m}). \end{cases} \tag{152}$$

with

$$\Psi(\hat{\Sigma}, \hat{q}, \hat{m}) = -\frac{\hat{m}^2}{4\hat{\Sigma}} J_p\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right). \tag{153}$$

We found that taking explicit derivatives of Ψ is numerically unstable for power-law targets, and for this reason we resorted to the solution of the equivalent set of equations

$$\begin{cases} m &= \frac{1}{d}\mathbb{E}_{\hat{\boldsymbol{S}}, \boldsymbol{S}_\star}\left[\mathrm{Tr}(\hat{\boldsymbol{S}}^\top \boldsymbol{S}_\star)\right] \\ q &= \frac{1}{d}\mathbb{E}_{\hat{\boldsymbol{S}}, \boldsymbol{S}_\star}\left[\mathrm{Tr}(\hat{\boldsymbol{S}}^\top \hat{\boldsymbol{S}})\right] \\ \Sigma &= \frac{2}{d}\mathbb{E}_{\hat{\boldsymbol{S}}}\left[\sum_{i=1}^p \frac{\Theta(\nu_i)}{\hat{\Sigma}^t} + \sum_{i<j} \frac{\tilde{\nu}_i - \tilde{\nu}_j}{\nu_i - \nu_j}\right] \end{cases} \tag{154}$$

where $\hat{\boldsymbol{S}} = (\boldsymbol{S}_\star + \delta \boldsymbol{Z} - \tilde{\lambda}\epsilon \boldsymbol{I}_d)^+_{(p)}$, ν_i is the i^{th} eigenvalue of the matrix $(\boldsymbol{S}_\star + \delta \boldsymbol{Z} - \tilde{\lambda}\epsilon \boldsymbol{I}_d)$ and $\tilde{\nu}_i$ is the i^{th} eigenvalue of $\hat{\boldsymbol{S}}$. These equations can be recast into the form (20) by defining $\delta = \sqrt{\hat{q}}/\hat{m}$ and $\epsilon = 2/\hat{m}$. The alternative expression for m, q, Σ can be found in (Erba et al., 2026, Appendix A.4.3), and is discussed in Appendix B. The expression for Σ is obtained computing the divergence (Erba et al., 2026, Eq. 54) explicitly in terms of eigenvalues.

In practice, we compute $\mathbb{E}_{\hat{\boldsymbol{S}}, \boldsymbol{S}_\star}$ by taking $\boldsymbol{S}_\star$ deterministic with power-law diagonal (as defined in the main text), by sampling $n_{\text{empirical}}$ GOE noise matrices $\boldsymbol{Z} \sim GOE(d)$, computing for each of the GOE matrices $(\hat{\boldsymbol{S}}, \nu_i, \tilde{\nu}_i)$, and taking the empirical means of the resulting (m, q, Σ) variables. In practice, we used $n_{\text{empirical}} = 16$ in our plots. We then iterate by fixed point scheme (154) until convergence. Crucially, we keep the same GOE matrices for all the iterations of the fixed point scheme in order to avoid unnecessarily long solution time.

D.2 EXPERIMENTS

To run experiments on this architecture, we use the solver LBFGS with the PyTorch implementation (Paszke et al., 2019) to minimize the objective (1). The specific hyperparameters used can be found

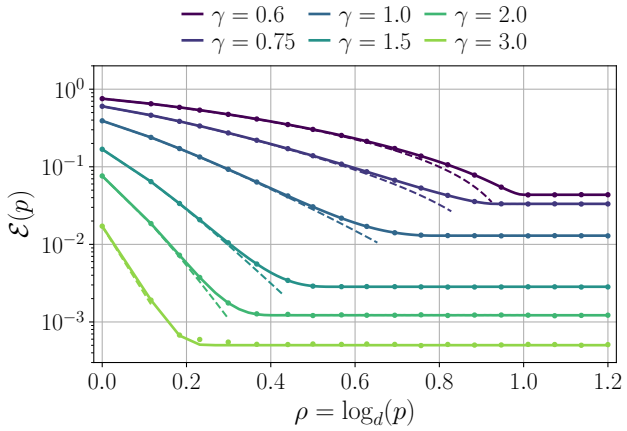


Figure 6: Test error scaling in the over-regularized regime for multiple data spectral exponents γ, with $d = 400$, $\alpha = 2.5$, $\sqrt{\Delta} = 0.5$ and $\ell = 0.45$. Full lines are non-asymptotic state evolution, dots are experiments (LBFGS) and dashed lines are the low rank estimation lower bound (155).

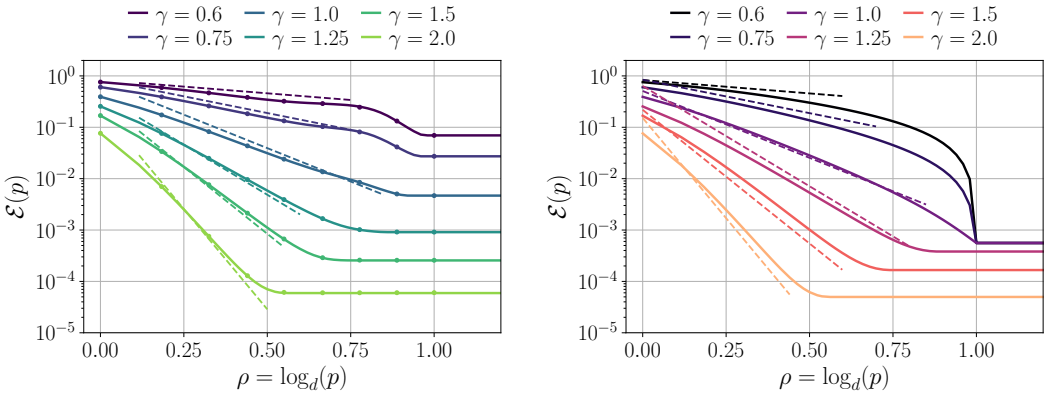


Figure 7: (Left) Test error scaling behavior in the under-regularized noiseless regime for different values of γ as a function of the width p, for fixed $d = 400$, $\alpha = 1.9$ and $\ell = -0.5$. (Right) Test error scaling behavior in the over-regularized noiseless regime for different values of γ as a function of the width p, for fixed $d = 400$, $\alpha = 2.5$ and $\ell = 0.1$.

in the code provided. As we remarked in the main text, the problem is non-convex, so we are not guaranteed to converge to the global minimum. Despite this, we observe a remarkably good agreement with our theory.

D.3 Additional figures

Here are some additional figures that show the match between our theory and experiments. In particular, in the regime with decay $\mathcal{E} = \Theta(d^{\rho(1-2\gamma)})$, where we expect to have enough samples to perfectly recover the dominant teacher eigenspaces, we compare our results to the lower bound provided by the Eckart-Young theorem (Eckart & Young, 1936) for our target structure (eigenvalues $\{\sqrt{d/\zeta(2\gamma)}i^{-\gamma}\}_{i=1}^{d}$)

$$\mathcal{E}(p) \geq \mathcal{E}_{EY}(p) = \frac{1}{\zeta(2\gamma)} \sum_{i=p+1}^{d} i^{-2\gamma} \tag{155}$$

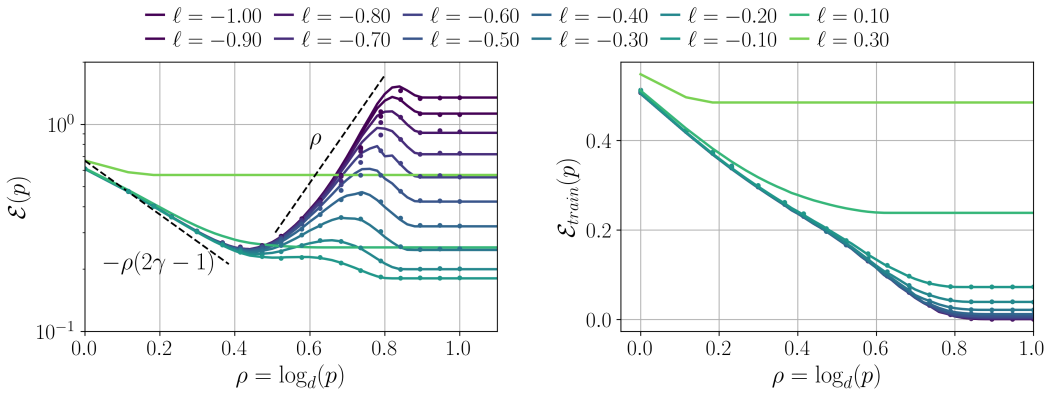


Figure 8: Test error (left) and training error (right) as functions of the network width for different regularization strengths ℓ. Parameters are $d = 400$, $\alpha = 1.8$, $\sqrt{\Delta} = 0.5$, and $\gamma = 0.75$. Solid lines correspond to the non-asymptotic state evolution predictions, while dots denote LBFGS experiments. The dashed lines show the corresponding scalings with ρ from Result 1. As the regularization is decreased, a double-descent phenomenon emerges. Increasing the width initially reduces both the training and test errors. Near the interpolation regime, the training error reaches a plateau while the test error exhibits a pronounced peak. Further increasing the width maintains a low training error and subsequently improves generalization performance.

E Declaration of LLM usage

We used LLMs in the following stages of the preparation of this work: editing (e.g., grammar, spelling, word choice), drafting sections of the paper, facilitating or running experiments, visualizing results for submission. Every LLM output was double-checked manually by the authors to ensure it performs as intended, in particular with regards to the numerical experiments.