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

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

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 test error minimization in 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 [14, 32, 34]. 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. [4, 11, 15, 16, 20, 22, 23, 46]. A separate line of research studies how increasing the number of parameters affects optimization, initialization, or expressivity, without directly addressing data-dependent generalization [13, 17, 19, 56]. Works that do analyze generalization as a function of model size typically focus on regimes with fixed representations, such as random feature models [3, 4, 12, 21, 37, 52] 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 [8, 49]. In these works, the scaling laws are formulated for idealized dynamical limits, where performance is controlled by training time, or by the product of learning rate and number of iterations, rather than by the generalization error of the trained empirical-risk minimizer at fixed dataset size. Consequently, 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.

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

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

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

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

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:

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

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

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

where $\tilde{\lambda} = \lambda \sqrt{p/d}$. 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 [29] 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

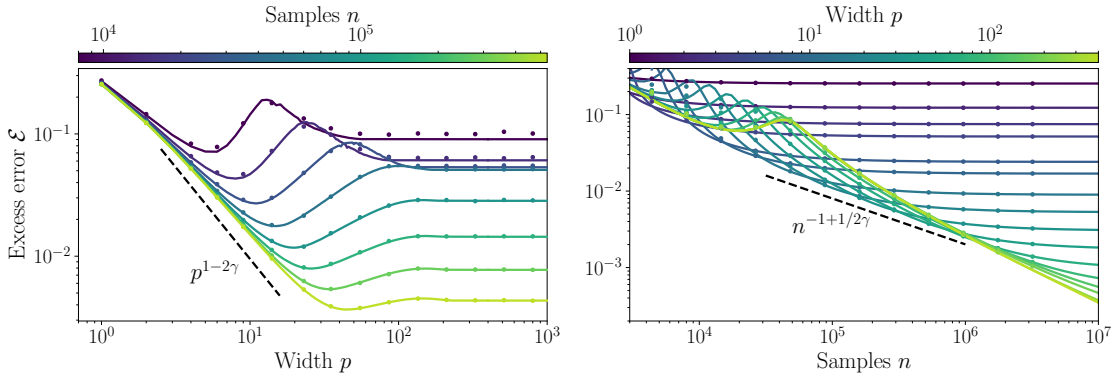


Figure 1: Excess test error vs width p (left) and number of samples n (right), for $d = 400$, $\gamma = 1.25$ and $\tilde{\lambda} = n^2/10d$. Lines are theoretical predictions from Result 1, dots are numerical simulations by LBFGS on (1), see Appendix F. As a function of the width, we highlight the 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. (18), (envelope of the curves) coincide with the Bayes optimal rate $-1 + 1/2\gamma$ from [23] (dashed black).

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

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

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 $N_{\text{param}} = dp$, the number of available training samples n and the regularization $\tilde{\lambda}$. We will be interested in general scaling regimes and 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 statistical control parameter in a feature-learning model.

- We derive a characterization of global minimizers of (1) as a function of the network width p, the number of samples n, the effective regularization $\tilde{\lambda} = \sqrt{p/d}\,\lambda$, and the spectral structure of the target, parametrized by γ; see Result 1. The equations predict the excess test error and reveal how

explicit regularization and the implicit rank constraint induced by finite width jointly shape the learned predictor.

- We obtain explicit scaling laws for the excess test error as a function of $p, n, \tilde{\lambda}$; see Results 2 and 3. 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 Result 1 as a quantitative description of the ERM.

2. Width-wise and sample-wise scaling laws

Our first result is the analytic characterization of the global minimizer $\hat{\boldsymbol{W}}$ of (1) through its equivalent characterization in terms of $\hat{\boldsymbol{S}} = \hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}}/\sqrt{pd}$ and its associated minimization problem (4).

Result 1 (Analytic properties of the ERM) *Consider any global minimum $\hat{\boldsymbol{W}}$ of (1) where the labels are generated from the model in (5) with $\Delta \geq 0$, 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 $n, p, d \gg 1$ we have*

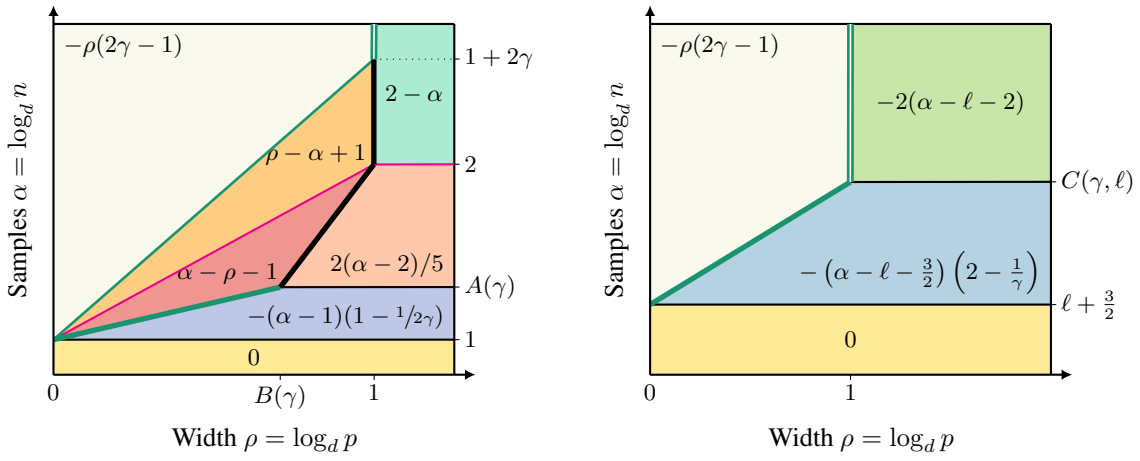
$$\hat{\boldsymbol{S}} \overset{d}{\sim} \hat{\boldsymbol{S}}_{\delta,\tilde{\lambda}\epsilon} = (\boldsymbol{S}_\star + \delta \boldsymbol{Z} - \tilde{\lambda}\epsilon \boldsymbol{I}_d)^+_{(p)}\,, \quad \mathcal{E}(\hat{\boldsymbol{W}}) \sim \frac{2n}{d^2}\delta^2 - \frac{\Delta}{2}\,, \quad \mathcal{L}(\hat{\boldsymbol{W}}) \sim \frac{\delta^2}{4\epsilon^2} + \frac{\tilde{\lambda}}{d}\operatorname{Tr}(\hat{\boldsymbol{S}}_{\delta,\tilde{\lambda}\epsilon})\,, \tag{8}$$

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*

$$\begin{cases} \dfrac{4n}{d^2}\,\delta - \dfrac{\delta}{\epsilon} = \partial_1 J_p(\delta, \tilde{\lambda}\epsilon), \\ 1 + \dfrac{\Delta}{2} + \dfrac{2n}{d^2}\,\delta^2 - \dfrac{\delta^2}{\epsilon} = \left(1 - \tilde{\lambda}\epsilon\,\partial_2\right) J_p(\delta, \tilde{\lambda}\epsilon), \end{cases} \quad \text{with} \quad J_p(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d}\operatorname{Tr}\left(\hat{\boldsymbol{S}}^2_{\delta,\tilde{\lambda}\epsilon}\right)\,, \tag{9}$$

where $\nu_i(\boldsymbol{A})$ denotes the i-th eigenvalue of the matrix $\boldsymbol{A}$ sorted decreasingly, and ∂_1, ∂_2 indicate the partial derivative with respect to the first and second arguments of J_p respectively.

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.



Under regularized $\ell < \alpha/2 - 1$ Over regularized $\alpha/2 - 1 < \ell < \alpha - 3/2$

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 (left) and large regularization (right). In each phase, we write the scaling exponent β of the test error $\mathcal{E} = \Theta(d^\beta)$. 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 (18)). With thick lines, we highlight the rank of the ERM, to the right of which the learned network is not width-constrained. With 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 exactly. We defined $A(\gamma) = (18\gamma - 5)/(14\gamma - 5)$, $B(\gamma) = (8\gamma - 2)/(14\gamma - 5)$ and $C(\gamma, \ell) = \ell + \gamma + 3/2$.

Result 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 (9) numerically and plugs the solution into (8); see Appendix F. 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).

The derivation in Appendix D follows the standard AMP/statistical-mechanics route and yields the *replica-symmetric* state-evolution prediction [1, 40, 41, 57]. The canonical computation is formulated in the regime $n = \Theta(d^2)$, $p = \Theta(d)$, $\tilde{\lambda} = \Theta(1)$, assuming a limiting spectral description for $\boldsymbol{S}_\star$. Extending the resulting equations to the power-law targets and to the regimes (7) should be viewed as a non-asymptotic extension, in the spirit of [23].

The main consequence of Result 1 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 $\alpha, \rho, \ell > 0$ in (7), fixing $\gamma > 1/2$ and $\Delta > 0$ (see Appendix E.2 for the noiseless case $\Delta = 0$). The resulting phase diagram is summarized in Figure 2; the mathematical description and its analysis is in Appendix B, while the derivation, including asymptotic constants depending on γ and Δ, is given in Appendix E.1.

References

[1] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

[2] 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.

[3] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[4] 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.

[5] 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.

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

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

[8] Gerard Ben Arous, Murat A Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] 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.

[10] 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.

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

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

[13] 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*, 2024.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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*, 2026. ICLR.

[24] 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.

[25] 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.

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

[27] 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.

[28] Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[29] 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*, pages 1043–1047. IEEE, 2008.

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

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

[32] 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.

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

[34] 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.

[35] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

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

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

[38] 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*, pages 3655–3663. PMLR, 2024.

[39] 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.

[40] M Mezard, G Parisi, and M Virasoro. Spin glasses, optimization, and biological applications, 1989.

[41] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[42] Marc Mézard, Giorgio Parisi, Miguel Angel Virasoro, and David J Thouless. Spin glass theory and beyond, 1988.

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

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5757–5823. PMLR, 30 Jun–04 Jul 2025.

[56] 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.

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

Appendix A. Further related works

Our analysis leverages tools from the theory of Approximate Message Passing algorithms (AMP) [6, 9, 25, 31, 54], which have been broadly applied to the study of learning problems in the high-dimensional limit [24, 48]. 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 [11, 20, 50] often precede and guide later rigorous analyses [18, 21, 44]. 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, $p \geq d$: [2, 26, 28, 30, 38, 51, 53]. In particular, their scaling laws were derived from the non-asymptotic AMP and for power-law targets in [23], still in the regime $p \geq d$, and extended to more general even activation functions in [22].

The works [8, 49] study closely related hierarchical multi-index models and width-constrained quadratic networks from a dynamical perspective. Their scaling laws describe idealized continuous-time population or online trajectories, rather than the discretized SGD dynamics or the finite-sample ERM considered here. While these works provide an important dynamical account of how width affects feature learning, 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 [39], 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 of the dynamics 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, we note that quadratic neural networks are closely related to attention mechanisms, since both operate through bilinear projections of the inputs, as discussed in Appendix E.4. This connection extends beyond structure to their high-dimensional analysis. Recent work [10] characterizes the learning curves of a single attention head, showing that the scaling laws obtained in [23] for the quadratic model also arise for key and query width $p \geq d$. We expect the finite-width analysis developed here to extend similarly to attention heads with arbitrary width, as discussed in Appendix E.4.

Appendix B. Detailed description of the phase diagram

We distinguish the under-regularized regime $\ell < \alpha/2 - 1$ with $\alpha > 1$, the over-regularized regime $\alpha/2 - 1 < \ell < \alpha - 3/2$ with $\alpha > \ell + 3/2$, and the rank-collapse regime $\alpha < \max(1, \ell + 3/2)$. The case $\rho = 1$ is delicate, and our derivation of the scalings does not cover it: we recover it by taking the limits $\rho \to 1^{\pm}$.

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

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

$$\beta = \begin{cases} -(2\gamma-1)\rho & \text{if} \quad \rho < \min\{(\alpha-1)/(2\gamma), 1\} \\ \rho + 1 - \alpha & \text{if} \quad (\alpha-1)/(2\gamma) < \rho < \min(1, \alpha-1) \quad \text{and} \quad \alpha < 1 + 2\gamma \\ \alpha - 1 - \rho & \text{if} \quad \min(1, \alpha-1) < \rho < \rho_{\text{eff},1}(\alpha,\gamma) \quad \text{and} \quad \alpha < 2 \\ \phi_1(\alpha,\gamma) & \text{if} \quad \rho > \rho_{\text{eff},1}(\alpha,\gamma) \end{cases} \tag{10}$$

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

$$\phi_1(\alpha,\gamma) = \begin{cases} -(\alpha-1)(1 - 1/2\gamma) & \text{if} \quad \alpha < A(\gamma) \\ 2(\alpha-2)/5 & \text{if} \quad A(\gamma) < \alpha < 2 \\ 2 - \alpha & \text{if} \quad \alpha > 2 \end{cases}, \tag{11}$$

and

$$\rho_{\text{eff},1}(\alpha,\gamma) = \begin{cases} (\alpha-1)(4\gamma-1)/(2\gamma) & \text{if} \quad \alpha < A(\gamma) \\ (3\alpha-1)/5 & \text{if} \quad A(\gamma) < \alpha < 2 \\ 1 & \text{if} \quad \alpha > 2 \end{cases}. \tag{12}$$

In this regime, the regularization is low enough that its presence does not alter the rates. The quantity $\rho_{\text{eff},1}(\alpha,\gamma)$ is the rank of the global minimizer, which 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 properties of the global minimum do not change, and the rates coincide with the ones in [23].

Let us comment the phases one by one (recalling that in this regime $\alpha > 1$ and $\ell < \alpha/2 - 1$); small arrows after the phase name indicate the monotonicity of the excess test error as a function of the network's width for fixed number of samples). We have two phases that are present for all values of sample scaling α.

- **Low width** ($\searrow$) 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** ($\rightarrow$) 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. For $\alpha > 2$ instead, the plateau decreases as $-\alpha$, as the number of samples is large enough to avoid overfitting.

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** ($\nearrow$) 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 then p outlying eigen-spaces correlating with the target, while the rest of the non-zero eigenvalues are spurious, i.e. their

eigenvector 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.

- **Decay after interpolation** ($\searrow$) 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 shrink, reducing the test error and providing a benefit of over-parametrization.

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 $\rho \to 1^-$ the error scales as $\Theta(d^{-2\gamma})$, while for $\rho > 1$ 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 model.

Over-regularized regime. If $\alpha > \ell + 3/2$ (enough samples) and $\alpha/2 - 1 < \ell < \alpha - 3/2$ (large regularization), we obtain the following excess test error scaling behavior.

Result 3 (Scaling law for large regularization) *In the setting of Result 1, when $\alpha > 1$ and $\alpha/2 - 1 < \ell < \alpha - 3/2$ we have $\mathcal{E} = \Theta(d^{\beta})$, where*

$$\beta = \begin{cases} -(2\gamma - 1)\rho & \text{if} \quad \rho < \rho_{\text{eff},2}(\alpha, \gamma, \ell) \\ \phi_2(\alpha, \gamma) & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha, \gamma, \ell) \end{cases} \tag{13}$$

where ϕ_2 is the following width-independent plateau

$$\phi_2(\alpha, \gamma) = \begin{cases} (\ell + 3/2 - \alpha)(2 - 1/\gamma) & \text{if} \quad \alpha < \ell + 3/2 + \gamma \\ 2(\ell + 2 - \alpha) & \text{if} \quad \alpha > \ell + 3/2 + \gamma \end{cases}, \tag{14}$$

and

$$\rho_{\text{eff},2}(\alpha, \gamma, \ell) = \begin{cases} (\alpha - \ell - 3/2)/\gamma & \text{if} \quad \alpha < \ell + 3/2 + \gamma \\ 1 & \text{if} \quad \alpha > \ell + 3/2 + \gamma \end{cases}. \tag{15}$$

In this over-regularized regime we observe a similar qualitative behavior as in the under-regularized regime, with two phases present for all $\alpha > 1$ at low width and full width with similar qualitative interpretation, separated by the threshold $\rho_{\text{eff},2}(\alpha, \gamma)$, which again is the rank of the ERM, and can be interpreted as an effective target width. The higher regularization induces sparser weights spectra, and prevents the intermediate-ρ phases related to interpolation 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 $\alpha > \ell + 3/2 + \gamma$.

Rank collapse regime. If $\alpha < 1$ (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, $\mathcal{E} \sim 1$.

On different kinds of over-parametrization. The scaling diagram in Figure 2 allows to characterize different kinds of over-parametrization. In all phases with $\alpha < \min(2, 1 + \rho)$, the network can fit perfectly the training set. In all phases with $\rho > \rho_{\text{eff},1/2}(\alpha, \gamma)$ (for both under- and over-regularization), the networks properties do not change when the width of the network is enlarged. In all phases with $\rho > 1$ (or $p \geq d$), the optimization problem (1) is benign: there are no spurious local minima.

Appendix C. 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 explicit regularization, and ask whether any one of them is inherently superior. We show that, across all three settings (under mild additional assumptions) the resulting test error can achieve Bayes-optimal scaling rates (as characterized in [23, Corollary 1]). This highlights a non-trivial 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 2 and 3 we can see that the optimal effective regularization $\tilde{\lambda}$ is given by the scaling $\ell = \alpha/2 - 1$ for all $\alpha, \rho > 0$. Indeed, the over-regularized rates improve monotonously 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

$$\log_d \mathcal{E}_{\text{opt reg}} \sim \begin{cases} -(2\gamma-1)\rho & \text{if} \quad \rho < \rho_{\text{eff},2}(\alpha,\gamma) \\ (1-\alpha)(1-1/(2\gamma)) & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha,\gamma) \quad \text{and} \quad \alpha < 1+2\gamma \\ 2-\alpha & \text{if} \quad \rho > \rho_{\text{eff},2}(\alpha,\gamma) \quad \text{and} \quad \alpha > 1+2\gamma \end{cases}, \tag{16}$$

For $\rho > \rho_{\text{eff},2}(\alpha,\gamma)$, the rates obtained are Bayes optimal, while for $\rho < \rho_{\text{eff},2}(\alpha,\gamma)$ they are sub-optimal. 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

$$\log_d \mathcal{E}^{\text{over reg}}_{\text{opt width}} \sim \begin{cases} (\ell+3/2-\alpha)(2-1/\gamma) & \text{if} \quad \alpha < \ell+3/2+\gamma \\ 2(\ell+2-\alpha) & \text{if} \quad \alpha > \ell+3/2+\gamma \end{cases} \tag{17}$$

which reduces to the Bayes optimal test error 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 $\alpha > 2$ the two coincide. For $\alpha > A(\gamma)$, the local minimum at $(\alpha-1)/(2\gamma)$ 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.

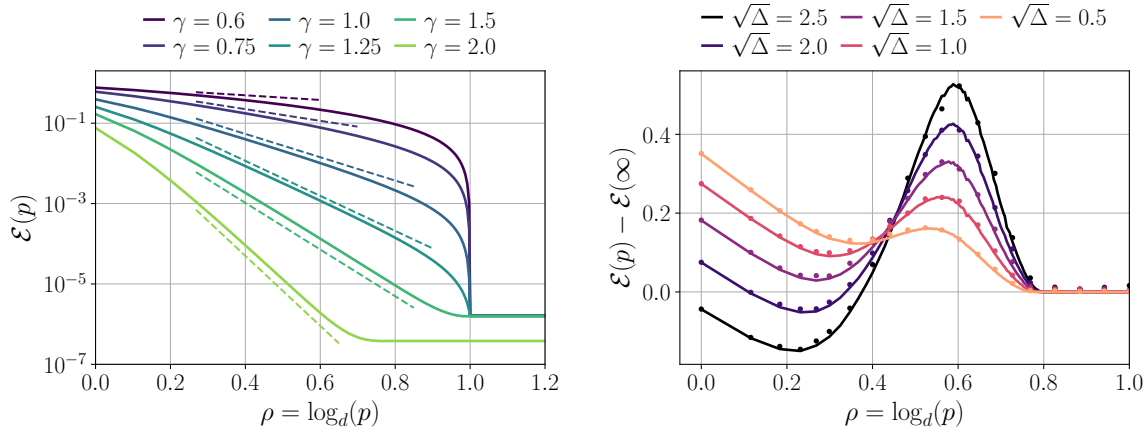


Figure 3: (Left) Test error scaling behavior in the over-regularized regime for different values of γ as a function of the width p, for fixed $d = 800$, $\sqrt{\Delta} = 0.5$, $\alpha = 4$ and $\ell = 1.2$. We observe the low-width data-dependent scaling with exponent $1 - 2\gamma$ (dashed lines, one per value of γ), as well as the large-width plateau. (Right) Dependency 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 (Result 1), points numerical experiments (Appendix F).

We find that there exists a function $G(\gamma, \Delta)$ such that

$$\rho^{\text{under reg}}_{\text{opt width}} = \begin{cases} \min\{(\alpha-1)/(2\gamma), 1\} & \text{if } \alpha > A(\gamma) \text{ or } (1 < \alpha < A(\gamma) \text{ and } G(\gamma,\Delta) > 0) \\ \rho_{\text{eff},1}(\alpha, \gamma) & \text{if } 1 < \alpha < A(\gamma) \text{ and } G(\gamma,\Delta) < 0 \end{cases} . \tag{18}$$

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 is to our best efforts 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

$$\log_d \mathcal{E}^{\text{under reg}}_{\text{opt width}} \sim \begin{cases} (1-\alpha)(1-1/(2\gamma)) & \text{if} \quad \alpha < 2\gamma + 1 \\ 2 - \alpha & \text{if} \quad \alpha > 2\gamma + 1 \end{cases} , \tag{19}$$

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 form of explicit regularization, we consider after-training pruning. This regularization consists in training a network at full width d, and later prune it

down to $p_{\text{pruning}} \ll d$ by substituting the weights with their best rank-p approximation (i.e., putting to zero all but the largest p_{pruning} eigenvalues). For generic final pruning width, we give the error scaling in Appendix E.3. We find, that pruning to the optimal width scaling $\rho_{\text{opt pruning}}$ achieves Bayes optimal rates with

$$\rho_{\text{opt pruning}} = \min\{(\alpha-1)/(2\gamma), 1\} \quad \text{if} \quad \ell < \alpha/2 - 1. \tag{20}$$

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 $G(\gamma, \Delta) < 0$.

Appendix D. Derivation of Result 1: analytical test error characterization

We derive the characterization in Result 1 by adapting the derivation of [28] to power-law targets and width-constrained students. For the sake of the derivation, as discussed in the main text, we assume that $n = \Theta(d^2)$, $p = \Theta(d)$, $\tilde{\lambda} = \Theta(1)$, that $\boldsymbol{S}_\star$ has a limiting spectral distribution for $d \gg 1$, and pretend that our minimization problem is convex. We will then check that this can be relaxed to general scaling, and that non-convexity is not an issue by comparing with numerical experiments. The key difference between that work and ours is the additional non-trivial width/rank constraint on the learned weights $\boldsymbol{W}$ (or $\boldsymbol{S}$ in the matrix representation (4)).

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. $\tilde{\boldsymbol{G}}_\mu \sim \text{GOE}(d)$. 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. [36, 55]
- **AMP and its analysis:** The asymptotic behavior of ERM in Gaussian linear models has been studied extensively [7, 35, 45, 48, 54], 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 [9, 31], 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 does it induce. For an ERM problem of the form (4) with the data model (5), [28] 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

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

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

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

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 [28, 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 [43],

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

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 Result 1. Algebraic manipulations of (21) under the change of variable $\delta = \sqrt{\hat{q}}/\hat{m}$ and $\epsilon = 2/\hat{m}$ leads to (9). 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 (9). The expression of the training loss, test error, and spectral distribution follow from the AMP/state evolution derivation [28], and are given by

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

where by $\mathrm{sp}[\cdot]$ we mean the empirical spectral distribution. 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 this equations can be formally proven. Second, we assume that the non-convexity of the loss does not alter the resulting equations. The first point is tricky to control analytically, and we are not aware of any line of attack. The second point could be approached (in the strict asymptotic scaling for which the derivation formally holds) in the same vein as [54], where the authors show that non-convex generalized linear models can be still studied through AMP and state evolution under the following sufficient condition, called replicon condition

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

where

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

This condition is related to the linear stability of AMP at its fixed point, but also to the so-called replica symmetry breaking phenomenon in the physics of disordered systems [41, 42]. In Appendix E.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.

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 projecting the resulting global minimizer to matrices with rank lower than $p_{\text{pruning}} < d$. Thus, we want to compute the generalization of $\hat{\boldsymbol{S}}_{\text{pruning}} = (\boldsymbol{S}_\star + \delta\boldsymbol{Z} - \tilde{\lambda}\epsilon\boldsymbol{I}_d)^+_{p_{\text{pruning}}}$, where δ, ϵ solve (9) for $p = d$. This can be computed using (25) and (30) with $\boldsymbol{P} = (\hat{m}/\hat{\Sigma})\hat{\boldsymbol{S}}_{\text{pruning}}$, giving

$$\mathcal{E}(\hat{\boldsymbol{W}}) \sim Q_\star + 4(\hat{m}/\hat{\Sigma})^{-1}\,\partial_{\hat{m}}\Psi + 4(\hat{m}/\hat{\Sigma})^{-2}\,\partial_{\hat{\Sigma}}\Psi = Q_\star + (\delta\partial_1 + \tilde{\lambda}\epsilon\partial_2 - 1)J_{p_{\text{pruning}}}(\delta, \tilde{\lambda}\epsilon) \tag{35}$$

This is equivalent to substituting $\tilde{\lambda}\epsilon$ at the solution of (9) with $\min(\tilde{\lambda}\epsilon, x_0)$ where x_0 is the value of the p-th eigenvalue of $\hat{S}$, sorted decreasingly.

Appendix E. Derivation of the scaling laws

Results 2, 3 are derived from Result 1 by careful asymptotic analysis of (9). 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 $Q_\star$ larger than in the main, as it will be the case for the generalization error and the loss. We further assume $1 \ll p \ll d$.

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 [23]: 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.

Said 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 $\xi_i \overset{\text{i.i.d.}}{\sim} \mu_{\rm sc}(x/\delta)/\delta$, with $\mu_{\rm sc}(\cdot)$ the standard semi-circle law of GOE matrices

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

and the top eigenvalues are the pushforward of $\{\sqrt{d}\, i^{-\gamma}\}_{i=d-K+1}^{d}$ though the BBP mapping [5, 33] $f_\delta(x) = x - \delta^2/x$:

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

The number of spikes outside the semicircle is fixed by imposing that K is the largest integer such that $\sqrt{d}K^{-\gamma} \geq \delta$ 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 $\nu_0 = \xi_{d-K} = 2\delta$ to indicate the largest eigenvalue in the semicircle. Based on this simplification, we write the spectral density $\mu_\delta(\cdot)$ as

$$\mu_\delta(x) \approx \left(1 - \frac{K}{d}\right)\mu_{\rm sc}(x/\delta)/\delta + \frac{K}{d}\sum_{i=1}^{K}\delta\left(x - f_\delta\big(\sqrt{d}i^{-\gamma}\big)\right), \tag{38}$$

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

$$\mu(x) = \begin{cases} \mu_\delta(x+\tilde{\lambda}\epsilon), & x > x_0^+, \\ 0, & x \leq x_0^+, \end{cases} \qquad x_0^+ := \max\{x_0, 0\}, \tag{39}$$

where x_0 is a cutoff imposing that the rank of $\hat{\boldsymbol{S}}$ is at most p. In formulas this means

$$x_0 = \inf\left\{t : \int_{t+\tilde{\lambda}\epsilon}^{\infty}\mu_\delta(x)\,{\rm d}x \leq \min\left(\frac{p}{d}, 1\right)\right\}. \tag{40}$$

With this form of the spectrum in mind, we can then study (9) in various regimes. In Appendix E.1 we study the scaling regimes for all noisy data models $\Delta > 0$, leaving the case $\Delta = 0$ to Appendix E.2. In Appendix E.3 we describe the scaling regimes of pruning. As a foreword, we discuss how the condition (33) can be specified for this target. Rewriting it in terms of density, it reads

$$\int dx \mu_\delta(x) \int dy \mu_\delta(y) \left(\frac{\eta(x) - \eta(y)}{x - y}\right)^2 < \frac{2n}{d^2}, \tag{41}$$

where

$$\eta(x) = \begin{cases} x - \tilde{\lambda}\epsilon, & x - \tilde{\lambda}\epsilon > x_0^+, \\ 0, & x - \tilde{\lambda}\epsilon \leq x_0^+, \end{cases} \qquad x_0^+ := \max\{x_0, 0\}. \tag{42}$$

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 [23, 28], and the replicon condition is always satisfied by the convexity of the optimization problem.
- When the cutoff is inside the bulk, i.e., $0 < x_0 \leq \delta$, 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. $\nu_k < x_0 \leq \nu_{k+1}$ for some $0 \leq k \leq K$, then the left side of (33) is of order

$$\frac{1}{d^2} \sum_{i \leq k < j} \left(\frac{\eta(\nu_i) - \eta(\nu_j)}{\nu_i - \nu_j} \right)^2 \approx \frac{1}{d^2} \sum_{i \leq k < j} \left(\frac{\sqrt{d} i^{-\gamma}}{\sqrt{d}\gamma(i-j)i^{-\gamma-1}} \right)^2 \approx \frac{k^2}{d^2\gamma} \log d. \tag{43}$$

  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 analysing, because by [22, Theorem 3.2] the number of spikes is upper bounded by $K \leq (n/d)^{1/2\gamma}$, and thus for $\gamma > 1/2$ the replicon condition is always satisfied once the cutoff is outside the bulk.

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

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

$$\mathcal{E}(\hat{\boldsymbol{W}}) \sim Q_\star + (\delta\partial_1 + \tilde{\lambda}\epsilon\partial_2 - 1) J_p(\delta, \tilde{\lambda}\epsilon)\,, \tag{44}$$

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

$$J_p(\delta, \tilde{\lambda}\epsilon) = \frac{1}{d} \operatorname{Tr}\left(\hat{\boldsymbol{S}}^2_{\delta, \tilde{\lambda}\epsilon} \right). \tag{45}$$

The solution will consist in enumerating all the possible configurations for the spectrum of $\hat{\boldsymbol{S}}$, and computing the leading behavior of J_p in each regime. A careful reader will notice that in some regimes we are assuming some scaling ansatz for our quantities. The consistency of these ansatz will be verified in each regime, and comparing the regimes as a whole one can verify we spanned all possible choices.

We start by remarking that if $x_0^+ = 0$, that is the rank-p cutoff has no effect, we will obtain the same scaling results from [23], obtained for $p \geq d$. We will thus focus on the cases $x_0^+ > 0$, in which the cutoff is active. This requires the condition

$$p > d \int_{\tilde{\lambda}\epsilon}^{\infty} \mu_\delta(x)\,\mathrm{d}x\,. \tag{46}$$

Additionally, this is also the condition under which the rates for $p < d$ are equivalent to the wide $p \geq d$ ones due to the rank of the ERM solution being deficient.

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

$$\int_{2-s}^{2} \mu_{\rm sc}(x)\,\mathrm{d}x \approx \frac{1}{\pi}\int_0^s \sqrt{x}\,\mathrm{d}x = \frac{2}{3\pi}s^{3/2}\,, \tag{47}$$

as well as

$$\int_{2-s}^{2} (x-a)^2\mu_{\rm sc}(x)\,\mathrm{d}x \approx \frac{2}{3\pi}(2-a)^2 s^{3/2} - \frac{4}{5\pi}(2-a)s^{5/2} + \frac{2}{7\pi}s^{7/2}\,, \tag{48}$$

Case I: rank collapse. In this regime, the spectrum of $\hat{\boldsymbol{S}}$ is without spikes $K = 0$, thus we have

$$\mu_\delta(x) \approx \mu_{\rm sc}(x/\delta)/\delta\,, \tag{49}$$

and the replicon condition is not satisfied. The moment J_p becomes

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \int_{x_0^+}^{2\delta-\tilde{\lambda}\epsilon} x^2 \mu_{\rm s.c.}\big((x+\tilde{\lambda}\epsilon)/\delta\big)/\delta\,\mathrm{d}x = \delta^2 \int_{x_0^+/\delta+\tilde{\lambda}\epsilon/\delta}^{2} (x-\tilde{\lambda}\epsilon/\delta)^2 \mu_{\rm s.c.}(x)\,\mathrm{d}x\,. \tag{50}$$

If $x_0^+ = 0$, the rank constraint because of the width has no effect, and from [23] we have $2\delta = \sqrt{\frac{d^2}{2n}(Q_\star + \Delta/2)}$ and $2\delta \approx \tilde{\lambda}\epsilon$, which result in a constant error. In order to have $x_0^+ > 0$ one needs

$$p \ll d \int_{\tilde{\lambda}\epsilon}^{2\delta} \mu_{\rm s.c.}(x/\delta)/\delta dx = \Theta(d(n/d^2)^{3/5})\,, \tag{51}$$

and this is the regime we will focus on. Calling $t_1 := 2 - \tilde{\lambda}\epsilon/\delta$ and $t_2 := t_1 - x_0^+/\delta$, we do the scaling ansatz $t_2 \ll t_1 \ll 1$ (i.e. the rank constraint dominates the spectral cut). This gives

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{2\delta^2}{3\pi} t_1^2 t_2^{3/2}\,. \tag{52}$$

Using the definition of the cutoff (40) we get

$$\frac{p}{d} = \int_{\tilde{\lambda}\epsilon + x_0^+}^{2\delta} \mu_\delta(x)\,\mathrm{d}x \approx \frac{2}{3\pi} t_2^{3/2}\,, \tag{53}$$

which gives

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{p}{d} t_1^2 \delta^2. \tag{54}$$

The derivatives of J_p include many terms, but the leading ones for $t_1 \ll 1$ are given by

$$\partial_1 J_p(\delta, \tilde{\lambda}\epsilon) \approx 4\frac{p}{d} t_1 \delta\,, \quad \text{and} \quad \tilde{\lambda}\epsilon \partial_2 J_p(\delta, \tilde{\lambda}\epsilon) \approx -4\frac{p}{d} t_1 \delta^2\,, \tag{55}$$

where we further used that $\tilde{\lambda}\epsilon \approx 2\delta$ (due to $t_1 \ll 1$). Plugging this into (9) we obtain

$$\begin{cases} \dfrac{4n}{d^2}\,\delta^2 - \dfrac{\delta^2}{\epsilon} \approx 4\dfrac{p}{d} t_1 \delta^2, \\ Q^* + \dfrac{\Delta}{2} + \dfrac{2n}{d^2}\,\delta^2 - \dfrac{\delta^2}{\epsilon} \approx 4\dfrac{p}{d} t_1 \delta^2, \end{cases} \tag{56}$$

where we use $t_1 \ll 1$ again to state that the term J_p is much smaller than its derivatives. The first equation of (56) subtracted by the second yields the prediction for the test error

$$\mathcal{E} := 2\frac{n}{d^2}\delta^2 - \frac{\Delta}{2} \approx Q^* \,, \tag{57}$$

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_1, we make another ansatz: $\epsilon \gg d^2/n$. Plugging this into the first equation of (56)

$$\frac{4n}{d^2}\delta^2 \approx 4\frac{p}{d}t_1\delta^2, \tag{58}$$

we can solve for t_1, obtaining $t_1 \approx n/(pd)$.

We now need to check the various assumptions that we made in the derivation, i.e. $t_2 \ll t_1 \ll 1$ and $\epsilon \gg d^2/n$. The condition $t_1 = n/pd \ll 1$ is satisfied for $n \ll d$ as $p \geq 1$. The condition $t_2 \ll t_1$ gives $p \ll n^{3/5}d^{-1/5}$ which is also satisfied (see (51)). From the explicit solution of δ we obtain $\delta^2 \approx d^2\Delta/4n$. Recall that $\tilde{\lambda}\epsilon \approx 2\delta$ because $t_1 \ll 1$. Then the anstaz $\epsilon \gg d^2/n$ requires

$$\tilde{\lambda} \ll \sqrt{n/d^2}. \tag{59}$$

If the regularization is too large, $\tilde{\lambda}\epsilon > 2\max(\delta, \sqrt{d})$, then all eigenvalues are zero. The rank contrain has no effect and from [23] one has the trivial error $\mathcal{E} = Q_\star$ under the condition $\lambda \gg \max(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{3/2}})$.

Case II: Over-regularization The second case is $\max(\delta, d^{-\gamma+1/2}) \ll \tilde{\lambda}\epsilon \ll \sqrt{d}$, where a part of the spikes and all the bulk are below $\tilde{\lambda}\epsilon$. Then the spectrum consists of K spikes. The replicon condition is always satisfied, and we have

$$J(\delta, \lambda\epsilon) = \frac{1}{d}\sum_{i=1}^{\min(p,K)}\left(\sqrt{d}i^{-\gamma} + \frac{\delta^2}{\sqrt{d}i^{-\gamma}} - \tilde{\lambda}\epsilon\right)^2. \tag{60}$$

For $p \gg K$, the rank constraint has no effect and thus the results are the same as [23], i.e., $\mathcal{E} = \Theta\left(\left(\tilde{\lambda}\frac{d^{3/2}}{4n}\right)^{\frac{2\gamma-1}{\gamma}}\right)$. For $1 \ll p \ll K$ we have

$$\begin{aligned} J(\delta, \tilde{\lambda}\epsilon) &\approx Q_\star - \frac{1}{2\gamma-1}p^{1-2\gamma} + \tilde{\lambda}^2\epsilon^2\frac{p}{d} + 2\delta^2\frac{p}{d} - \frac{2\tilde{\lambda}\epsilon}{\sqrt{d}}\left(\frac{p^{1-\gamma}}{1-\gamma} + \zeta(\gamma)\mathbf{1}_{\gamma>1}\right) \\ &\quad + \frac{\delta^4}{d^2}\frac{p^{2\gamma+1}}{2\gamma+1} - 2\frac{\tilde{\lambda}\epsilon\delta^2}{d\sqrt{d}}\frac{p^{\gamma+1}}{\gamma+1} \\ &\approx Q_\star - \frac{1}{2\gamma-1}p^{1-2\gamma}, \end{aligned} \tag{61}$$

under an ansatz that other terms are much smaller for $p \ll K$. Taking it into (9), we obtain $\delta \approx \sqrt{\frac{\Delta d^2}{4n}}$, $\epsilon \approx \frac{d^2}{4n}$ as the leading order solution and

$$\mathcal{E} := 2\frac{n}{d^2}\delta^2 - \frac{\Delta}{2} \approx \frac{1}{2\gamma-1}p^{1-2\gamma} \tag{62}$$

obtained by the first equation of (9) substracted by the second.

The ansatz $\max(\delta, d^{-\gamma+1/2}) \ll \lambda\epsilon \ll \sqrt{d}$ gives

$$\max\left(\frac{d^{\gamma-3/2}}{n^{\gamma-1}}, \frac{n}{d^{\gamma+\frac{3}{2}}}\right) \ll \lambda \ll \frac{d^{3/2}}{n}, \tag{63}$$

under which we can verify that $p^{1-2\gamma}$ is indeed the leading order in (61). Therefore, we have

$$\mathcal{E} = \Theta(\min(p, (n/\tilde{\lambda}d^{3/2})^{1/\gamma})^{1-2\gamma}) \tag{64}$$

under the condition $\max\left(\frac{d^{\gamma-3/2}}{n^{\gamma-1}}, \frac{n}{d^{\gamma+\frac{3}{2}}}\right) \ll \lambda \ll \frac{d^{3/2}}{n}$.

Case III: Intermediate over-regularization Case III is $\delta \ll \lambda\epsilon \ll d^{-\gamma+\frac{1}{2}}$. In this case the spectrum consists of d spikes. Thus the replicon condition is always satisfied and we have

$$\begin{aligned} J(\delta, \tilde{\lambda}\epsilon) &= \frac{1}{d}\sum_{i=1}^{p}(\sqrt{d}i^{-\gamma} + \frac{\delta^2}{\sqrt{d}i^{-\gamma}} - \tilde{\lambda}\epsilon)^2 \\ &\approx Q_\star - \frac{1}{2\gamma-1}p^{1-2\gamma} + \tilde{\lambda}\epsilon p^{\min(\gamma,1)-1} + \tilde{\lambda}^2\epsilon^2\frac{p}{d}. \end{aligned} \tag{65}$$

Then the first equation of (9) gives $\epsilon = \frac{1}{4\alpha}$ and the first equation of (9) subtracted by the second gives

$$\mathcal{E} \approx \frac{1}{2\gamma-1}p^{1-2\gamma} + \frac{\tilde{\lambda}^2 p d^3}{16n^2}. \tag{66}$$

The condition $\delta \ll \lambda\epsilon \ll d^{-\gamma+\frac{1}{2}}$ reduces to

$$\sqrt{\frac{n}{d^2}} \ll \lambda \ll \frac{n}{d^{\gamma+3/2}}, \tag{67}$$

which further requires $n \gg d^{2\gamma+1}$. Then we can verify that the $p^{1-2\gamma}$ term is always the leading term, i,e.,

$$\mathcal{E} \approx \frac{1}{2\gamma-1}p^{1-2\gamma} \tag{68}$$

under the condition $n \gg d^{2\gamma+1}$ and $\delta \ll \lambda\epsilon \ll d^{-\gamma+\frac{1}{2}}$.

Case IV: Large number of samples Case IV is $n \gg d^2$ and $\lambda\epsilon \ll \delta$. First, consider that the rank cutoff $x_0 > 2\delta$, the whole bulk is cut away and we have

$$J_p(\delta, \tilde{\lambda}\epsilon) \approx \frac{1}{d}\sum_{i=1}^{p}\left(\sqrt{d}i^{-\gamma} + \frac{\delta^2}{\sqrt{d}i^{-\gamma}} - \tilde{\lambda}\epsilon\right)^2 \approx Q_\star - \frac{1}{2\gamma-1}p^{2\gamma-1}, \tag{69}$$

which gives

$$\mathcal{E} = \frac{1}{2\gamma-1}p^{2\gamma-1} \tag{70}$$

similarly to Case II. The replicon condition is not satisfied if the cutoff is within the bulk: $(\Delta d/4n)^{-1/2\gamma} \ll p \ll d$, which further requires $d^2 \ll n \ll dp^{2\gamma}$. Then similarly to Case I the rank constrain gives

$$\frac{p}{d} = \int_{x_0}^{2\delta} \mu_{\text{s.c.}}(x/\delta)/\delta dx = \frac{2}{3\pi} t_2^{3/2}. \tag{71}$$

Then we obtain

$$\begin{aligned} J(\delta, \tilde{\lambda}\epsilon) &\approx \frac{1}{d} \sum_{i=1}^{(\delta/\sqrt{d})^{-1/\gamma}} (\sqrt{d} i^{-\gamma} + \frac{\delta^2}{\sqrt{d} i^{-\gamma}} - \lambda\epsilon)^2 + \int_{x_0}^{2\delta} \mu_{\text{s.c.}}(x/\delta)/\delta x^2 dx \\ &\approx Q_\star + C(\gamma) \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + 4\frac{p}{d}\delta^2, \end{aligned} \tag{72}$$

where $C(\gamma) := \frac{4-4\gamma^2}{1-4\gamma^2}$ is a constant. The SE (9) reduces to

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = (2 - 1/\gamma) C(\gamma) \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + 8\frac{p}{d}\delta^2 \\ Q_\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = C(\gamma) \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + 4\frac{p}{d}\delta^2, \end{cases} \tag{73}$$

As an ansatz, we assume that the right side is small. Then the leading order solution of (73) is $\delta^2 \approx \frac{d^2\Delta}{4n}$ and $\epsilon \approx \frac{d^2}{4n}$. The first equation (73) substracted by the second gives

$$\mathcal{E} \approx \frac{4(1+\gamma)(1-\gamma)^2}{4\gamma^2 - 1} \left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{p}{d}\frac{d^2\Delta}{n} \approx \frac{\Delta p d}{n}, \tag{74}$$

where the second term is leading because $\frac{pd}{n} \gg (d/n)^{1-1/2\gamma}$ as long as $p \gg (\Delta d/4n)^{-1/2\gamma}$.

The ansatz $\lambda\epsilon \ll \delta$ requires $\lambda \ll \sqrt{\frac{n}{d^2}}$. The ansatz $\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} \ll \frac{p}{d}\delta^2 \ll 1$ is always satisfied for $n \gg d^2$ and $p \gg (\Delta d/4n)^{-1/2\gamma}$.

In conclusion, in Case IV we have

$$\mathcal{E} \approx \begin{cases} \frac{1}{2\gamma-1} p^{1-2\gamma}, & p \ll (\Delta d/4n)^{-1/2\gamma} \\ \frac{pd\Delta}{n}, & p \gg (\Delta d/4n)^{-1/2\gamma} \end{cases} \tag{75}$$

under the condition $n \gg d^2$ and $\lambda \ll \sqrt{\frac{n}{d^2}}$

Case V: Overfitting Case V is $d \ll n \ll d^2$ and $0 < 2 - \frac{\lambda\epsilon}{\delta} \ll 1$, $d^{-\gamma+\frac{1}{2}} \ll \delta \ll \sqrt{d}$. Then if there is no rank constrain, from [23] there are $(\Delta d/4n)^{-1/2\gamma}$ spikes, a bulk, and $n^{3/5}d^{-1/5}$ zero eigenvalues.

Thus when $p \ll (\Delta d/4n)^{-1/2\gamma}$, the rank-$p$ cutoff is within the spikes and the replicon condition is satisfied. Similarly to Case IV We have

$$\mathcal{E} \approx \frac{1}{2\gamma - 1} p^{2\gamma - 1}. \tag{76}$$

When $p \gg n^{3/5}d^{-1/5}$ the rank-p cutoff has no effect and the replicon condition is also satisfied. The result is the same as [22], i.e.,

$$\mathcal{E} \approx \frac{24\gamma^3}{4\gamma^3+4\gamma^2-\gamma-1}\left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{\Delta}{7}\left(\frac{15\pi}{4}\right)^{2/5}\left(\frac{n}{d^2}\right)^{2/5}. \tag{77}$$

Otherwise the cutoff is within the bulk and the replicon condition is not satisfied. Then similarly to Case IV we have

$$J(\delta, \tilde{\lambda}\epsilon) \approx Q_\star + C(\gamma)\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + \frac{p}{d}\delta^2 t_1^2. \tag{78}$$

under the ansatz $t_1 \ll 1$. The SE (9) reduces to

$$\begin{cases} 4\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = \frac{4r}{d}\delta^2 t_1 + (2-1/\gamma)C(\gamma)\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + 2\frac{p}{d}\delta^2 t_1^2 \\ Q_\star + \frac{\Delta}{2} + 2\frac{n}{d^2}\delta^2 - \frac{\delta^2}{\epsilon} = \frac{4r}{d}\delta^2 t_1 + C(\gamma)\left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} + \frac{p}{d}\delta^2 t_1^2, \end{cases} \tag{79}$$

which solves $t_1 = \frac{n}{pd}$, $\delta \approx \sqrt{\frac{\Delta d^2}{4n}}$ (similarly to case I) and

$$\mathcal{E} := 2\frac{n}{d^2}\delta^2 - \frac{\Delta}{2} \approx \frac{24\gamma^3}{4\gamma^3+4\gamma^2-\gamma-1}\left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{\Delta n}{4pd}. \tag{80}$$

Then $t_1 \ll 1$ requires $p \gg \frac{n}{d}$. When $(\Delta d/4n)^{-1/2\gamma} \ll p \ll \frac{n}{d}$, we instead should use the ansatz $\tilde{\lambda}\epsilon \ll \delta$, which reduces to case IV, i.e.

$$\mathcal{E} \approx \frac{pd\Delta}{n}. \tag{81}$$

In this case we have $\epsilon \approx \frac{1}{4\alpha}$, and thus $\tilde{\lambda}\epsilon \ll \delta$ is always satisfied given $\tilde{\lambda} \ll \sqrt{\frac{n}{d^2}}$.

In conclusion, in case V we have

$$\mathcal{E} = \begin{cases} \frac{1}{2\gamma-1}p^{1-2\gamma}, \ p \ll (\Delta d/4n)^{-1/2\gamma} \\ \frac{pd\Delta}{n}, \ (\Delta d/4n)^{-1/2\gamma} \ll p \ll \frac{n}{d} \\ \Theta((\tilde{\lambda}p/d)^{-2/3}), \ p = \Theta(n/d) \\ \frac{24\gamma^3}{4\gamma^3+4\gamma^2-\gamma-1}\left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{\Delta n}{4pd}, \ \frac{n}{d} \ll p \ll n^{3/5}d^{-1/5} \\ \frac{24\gamma^3}{4\gamma^3+4\gamma^2-\gamma-1}\left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{\Delta}{7}\left(\frac{15\pi}{4}\right)^{2/5}\left(\frac{n}{d^2}\right)^{2/5}, \ p \gg n^{3/5}d^{-1/5} \end{cases} \tag{82}$$

under the condition $d \ll n \ll d^2$ and $\tilde{\lambda} \ll \sqrt{\frac{n}{d^2}}$.

Summary The excess test error is

$$\mathcal{E} = \begin{cases} \Theta\left((n/d)^{-1+1/(2\gamma)} + \rho(n,d,p)\right) & \text{if} \quad d \ll n \ll pd \quad \text{and} \quad \lambda \ll \sqrt{\frac{n}{pd}} \\ \Theta\left(\left(\lambda\sqrt{p}d/n\right)^{2-1/\gamma}\right) & \text{if} \quad \max\left(\sqrt{\frac{n}{pd}}, \frac{n}{p^{\gamma+1/2}d}\right) \ll \lambda \ll \frac{n}{\sqrt{pd}} \\ \Theta\left(p^{1-2\gamma}\right) & \text{if} \quad \lambda \ll \frac{n}{p^{\gamma+1/2}d} \quad \text{and} \quad n \gg dp^{2\gamma} \end{cases}, \tag{83}$$

where the overfitting rate is given by

$$\rho(n,d,p) := \begin{cases} \Theta((n/d^2)^{2/5}),\ n \ll (p^5 d)^{1/3} \\ \Theta(n/pd),\ n \gg (p^5 d)^{1/3}. \end{cases} \tag{84}$$

If we choose $p = \Theta(d^\rho), n = \Theta(d^\alpha)$ and $\lambda = \Theta(d^{\ell - \frac{1-\rho}{2}})$, we obtain $\mathcal{E} = \Theta(d^\beta)$ with

$$\beta = \begin{cases} (-1+1/(2\gamma))(\alpha-1) + (\alpha - 1 - \rho) & \text{if} \quad 1 < \alpha < \rho + 1 \quad \text{and} \quad \ell < \frac{\alpha - 3/2}{2} \quad \text{and} \quad \alpha > \frac{1}{3}(5\rho+1) \\ (-1+1/(2\gamma))(\alpha-1) + \frac{2}{5}(\alpha - 2) & \text{if} \quad 1 < \alpha < \rho + 1 \quad \text{and} \quad \ell < \frac{\alpha - 3/2}{2} \quad \text{and} \quad \alpha < \frac{1}{3}(5\rho+1) \\ (2 - 1/\gamma)(\ell + \frac{3}{2} - \alpha) & \text{if} \quad \max(\frac{\alpha-3/2}{2}, \alpha - \gamma\rho - \frac{3}{2}) < \ell < \alpha - \frac{3}{2} \\ \rho(1 - 2\gamma) & \text{if} \quad \ell < \alpha - \gamma\rho - \frac{3}{2} \quad \text{and} \quad \alpha > 2\gamma\rho + 1 \end{cases}. \tag{85}$$

Recall that for the derivation above we assume $\rho < 1$. Combining it with the results for $\rho = 1$ [23] we obtain the phase diagram in the main text.

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

Most of the analysis is the same for the noiseless setting. The main difference is that the overfitting term disappears. The excess test error is given by

$$\mathcal{E} = \begin{cases} \Theta\left((n/d)^{1-2\gamma}\right) & \text{if} \quad d \ll n \ll pd \quad \text{and} \quad \lambda \ll \frac{d^{\gamma-1}}{n^{\gamma-1}\sqrt{p}} \\ \Theta\left(\left(\lambda\sqrt{p}d/n\right)^{2-1/\gamma}\right) & \text{if} \quad \max\left(\frac{d^{\gamma-1}}{n^{\gamma-1}\sqrt{p}}, \frac{n}{p^{\gamma+1/2}d}\right) \ll \lambda \ll \frac{n}{\sqrt{p}d} \\ \Theta\left(p^{1-2\gamma}\right) & \text{if} \quad \lambda \ll \frac{n}{p^{\gamma+1/2}d} \quad \text{and} \quad n \gg pd \end{cases}, \tag{86}$$

If we choose $p = \Theta(d^\rho), n = \Theta(d^\alpha)$ and $\lambda = \Theta(d^{\ell - \frac{1-\rho}{2}})$, we obtain $\mathcal{E} = \Theta(d^\beta)$ with

$$\beta = \begin{cases} (\alpha - 1)(1 - 2\gamma) & \text{if} \quad 1 < \alpha < \rho + 1 \quad \text{and} \quad \ell < \alpha + (1-\alpha)\gamma - \frac{3}{2} \\ -\left(\alpha - \ell - \frac{3}{2}\right)\left(2 - \frac{1}{\gamma}\right) & \text{if} \quad \max\left\{\alpha + (1-\alpha)\gamma - \frac{3}{2}, \alpha - \rho\gamma - 3/2\right\} < \ell < \alpha - 3/2 \\ -\rho(2\gamma - 1) & \text{if} \quad \ell < \alpha - \rho\gamma - 3/2 \quad \text{and} \quad \alpha > \rho + 1 \end{cases}, \tag{87}$$

for $\rho < 1$. The phase diagram is plotted in Figure 4, where we also combine the results for $\rho = 1$ in [23]. At low regularization, 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 is always preferred.

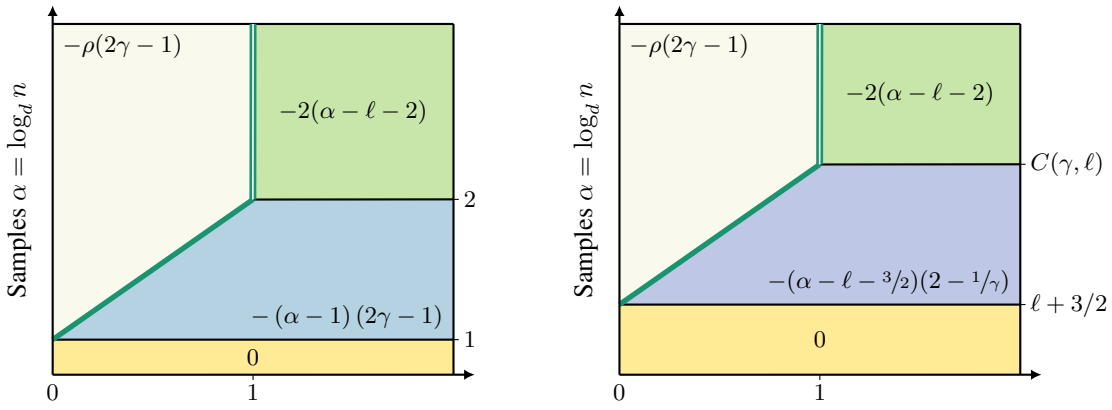
E.3. Derivation of the excess test error for pruning

Now we consider pruning, i.e., first train a full width student and only keep the leading $p \ll d$ eigenvalues. For technical simplicity, we cut the eigenvalues 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

$$Q_\star + (\delta\partial_\delta + b\partial_b - 1)J(\delta, b), \tag{88}$$

for some suitably chosen $b \geq \lambda\epsilon$, where δ remains the same as in [23]. The phases in the following thus refer to the same thing as in [23].

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



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

Figure 4: Phase diagram of the noiseless estimation task. See Fig 6 for numerical verification.

- 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 $p^{1-2\gamma}$.
- 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

$$J_1(\delta, b) := \int_b^{2\delta} \mu_{sc}(x/\delta)/\delta (x-b)^2 dx = \delta^2 \frac{16 t_1^{7/2}}{105\pi}, \tag{89}$$

where $t_1 := 2 - \frac{b}{\delta}$ satisfies $\frac{p}{d} = \int_b^{2\delta} \mu_{sc}(x/\delta)/\delta dx = \frac{2}{3\pi} t_1^{3/2}$. Thus the total error scales as

$$\mathcal{E} = \frac{2^{5/3} \cdot 3^{4/3} \cdot \pi^{4/3}}{35} \left(\frac{p}{d}\right)^{7/3} \frac{\Delta d^2}{4n} + \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}}, \tag{90}$$

where we use the solution $\delta^2 = \frac{\Delta d^2}{4n}$ from [23]. 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

$$\mathcal{E} = \frac{2^{5/3} \cdot 3^{4/3} \cdot \pi^{4/3}}{35} \left(\frac{p}{d}\right)^{7/3} \frac{\Delta d^2}{4n} + \frac{24\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta}{4n}\right)^{1-\frac{1}{2\gamma}}, \tag{91}$$

as in the previous case.

E.4. Generalization to the attention model

As shown by [10], the scaling phase diagram obtained for full-rank quadratic networks in [23] remains very similar for the attention index model defined as

$$\hat{f}_{\rm sq}(\boldsymbol{x}_{\rm in}; W) = A_W(\boldsymbol{x}_{\rm in})\boldsymbol{x}_{\rm in}, \text{ or } \hat{f}_{\rm lb}(\boldsymbol{x}_{\rm in}; W) = A_W(\boldsymbol{x}_{\rm in}) \tag{92}$$

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

$$A_W(\boldsymbol{x}_{\rm in}) = \sigma\left(\frac{\boldsymbol{x}_{\rm in} W W^T \boldsymbol{x}_{\rm in}^T - \mathbb{E}_{\rm tr}[\boldsymbol{x} W W^T \boldsymbol{x}^T]}{d\sqrt{p}}\right) \in \mathbb{R}^{T\times T} \tag{93}$$

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. (93), and restrict the class of (possibly noisy) target functions to the ones of the form (respectively for the seq2seq and seq2lab tasks)

$$\boldsymbol{x}_{\rm out}^\mu = \sigma_*\left(R(\boldsymbol{x}_{\rm in})\right)\boldsymbol{x}_{\rm in} \quad \text{and} \quad y^\mu = \sigma_*\left(R(\boldsymbol{x}_{\rm in})\right), \tag{94}$$

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:

$$R(\boldsymbol{x}_{\rm in})_{ab} = \frac{\boldsymbol{x}_{{\rm in,a}}^T S_0 \boldsymbol{x}_{\rm in,b} - \delta_{ab}\,{\rm Tr}(S_0)}{\sqrt{d}} + \sqrt{\frac{\Delta}{2-\delta_{ab}}}\xi^\mu, \tag{95}$$

where S_0 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 (or equivalently, weight decay) that is commonly used in practice in large language models, i.e. $\hat{W} = \arg\min_W[\mathcal{L}^{\rm data}(W) + \lambda\|W\|_F^2]$ where respectively for the two tasks

$$\begin{cases} \mathcal{L}_{\rm sq}^{\rm data}(W) := \frac{1}{d}\sum_{\mu=1}^n ||\boldsymbol{x}_{\rm out}^\mu - \hat{f}_{\rm sq}(\boldsymbol{x}_{\rm in}^\mu; W)||_F^2, \\ \mathcal{L}_{\rm lb}^{\rm data}(W) := \sum_{\mu=1}^n ||y^\mu - \hat{f}_{\rm lb}(\boldsymbol{x}_{\rm in}^\mu; W)||_F^2. \end{cases} \tag{96}$$

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

$$\begin{aligned} e_{\rm test} &= \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_{\rm in},\boldsymbol{x}_{\rm out}}||\boldsymbol{x}_{\rm out} - \hat{f}(\boldsymbol{x}_{\rm in}; \hat{W})||_F^2, \\ e_{\rm test} &= \mathbb{E}_{\boldsymbol{x}_{\rm in},y}||y - \hat{f}(\boldsymbol{x}_{\rm in}; \hat{W})||_F^2, \end{aligned} \tag{97}$$

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

$$\mathcal{E} := e_{\rm test} - e_{\rm base}, \tag{98}$$

where

$$e_{\rm base} := \inf_m \mathcal{I}(m_0, m_0/Q_\star) \tag{99}$$

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

$$e_{\rm base} := \inf_m \mathcal{I}(m, m/Q_\star) \tag{100}$$

otherwise. Here we define

$$\mathcal{I}(m,q) = \mathbb{E}_{z_0,z,\xi} ||\tilde{\sigma}_*(z_0 + \sqrt{\Delta/2}\zeta) - \tilde{\sigma}(z)||^2, \tag{101}$$

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 effetive 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

$$\eta = -\frac{1}{2}\left[\frac{\partial \mathcal{K}(m,q)}{\partial q}|_{\eta Q_\star, \eta^2 Q_\star}\right]^{-1} \frac{\partial \mathcal{K}(m,q)}{\partial m}|_{\eta Q_\star, \eta^2 Q_\star}, \tag{102}$$

where

$$\mathcal{K}(m,q) := \mathbb{E}_{z_0,z,\zeta} \inf_{h \in \arg\inf ||\tilde{\sigma}_*(z_0+\sqrt{\Delta/2}\zeta) - \tilde{\sigma}(\cdot)||^2} ||h - z||^2 \tag{103}$$

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

Appendix F. Details of numerical implementations

Here we describe our numerical implementation.

Appendix F.1 shows how we solved the system (9) numerically to obtain the theoretical predictions for our figures, and are always plotted with continuous lines.

Appendix F.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 to 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).

F.1. Theoretical lines

To plot theoretical predictions, we solve (9) in the form (21), i.e.

$$\begin{cases} \widehat{\Sigma} = \dfrac{2n}{d^2}\dfrac{1}{\Sigma+\frac{1}{4}}, \\ \widehat{m} = \dfrac{2n}{d^2}\dfrac{1}{\Sigma+\frac{1}{4}}, \\ \widehat{q} = \dfrac{2n}{d^2}\dfrac{Q_\star - 2m + q + \frac{\Delta}{2}}{\left(\Sigma+\frac{1}{4}\right)^2}, \end{cases} \qquad \begin{cases} m = -2\,\partial_{\widehat{m}}\Psi(\widehat{\Sigma},\widehat{q},\widehat{m}), \\ q = 4\,\partial_{\widehat{\Sigma}}\Psi(\widehat{\Sigma},\widehat{q},\widehat{m}), \\ \Sigma = -4\,\partial_{\widehat{q}}\Psi(\widehat{\Sigma},\widehat{q},\widehat{m}). \end{cases} \tag{104}$$

with

$$\Psi(\hat{\Sigma},\hat{q},\hat{m}) = -\frac{\hat{m}^2}{4\hat{\Sigma}} J_p\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right). \tag{105}$$

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

$$\begin{cases} m &= \frac{1}{d}\mathbb{E}_{\hat{\boldsymbol{S}},\boldsymbol{S}_\star}\left[\mathrm{Tr}(\hat{\boldsymbol{S}}^\top \boldsymbol{S}_\star)\right] \\ q &= \frac{1}{d}\mathbb{E}_{\hat{\boldsymbol{S}},\boldsymbol{S}_\star}\left[\mathrm{Tr}(\hat{\boldsymbol{S}}^\top \hat{\boldsymbol{S}})\right] \\ \Sigma &= \frac{2}{d}\mathbb{E}_{\hat{\boldsymbol{S}}}\left[\sum_{i=1}^{p}\frac{\Theta(\nu_i)}{\hat{\Sigma}^t} + \sum_{i<j}\frac{\tilde{\nu}_i-\tilde{\nu}_j}{\nu_i-\nu_j}\right] \end{cases} \tag{106}$$

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 (9) by defining $\delta = \sqrt{\hat{q}}/\hat{m}$ and $\epsilon = 2/\hat{m}$. The alternative expression for m, q, Σ can be found in [28, Appendix A.4.3], and is discussed in Appendix D. The expression for Σ is obtained computing the divergence [28, 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 (106) 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.

F.2. Experiments

To run experiments on this architecture, we use the solver LBFGS with the PyTorch implementation [47] to minimize the objective (1). The specific hyperparameters used can be found in the code provided. As we remarked in the main text, the problem is in non-convex, so we are not guaranteed to converge to the global minimum. Despite this, we observe a remarkably good agreement with our theory.

F.3. Additional figures

Here 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 [27] for our target structure (eigenvalues $\{\sqrt{d/\zeta(2\gamma)}i^{-2\gamma}\}_{i=1}^{d}$)

$$\mathcal{E}(p) \geq \mathcal{E}_{EY}(p) = \frac{1}{\zeta(2\gamma)} \sum_{i=p+1}^{d} i^{-2\gamma} \tag{107}$$

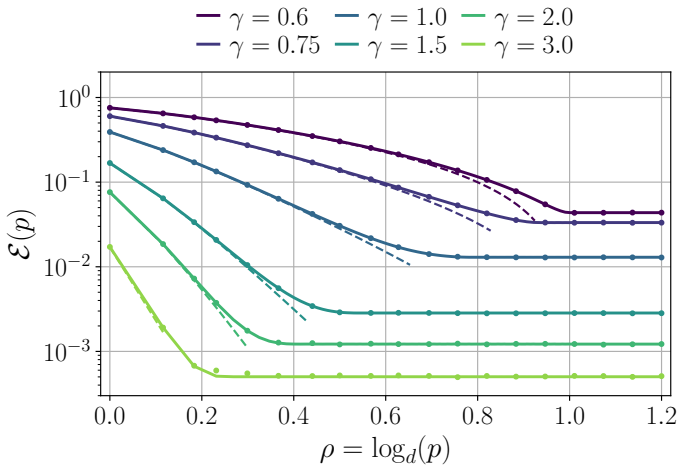


Figure 5: Test error scaling in the over-regularized regime for multiple data exponents γ, with $d = 400$, $\alpha = 2.5$, $\sqrt{\Delta} = 0.5$ and $\ell = 0.45$. Full lines are non-asymptotic state evolution, dots are experiments (LBFGS) and dashed lines are the low rank estimation lower bound (107).

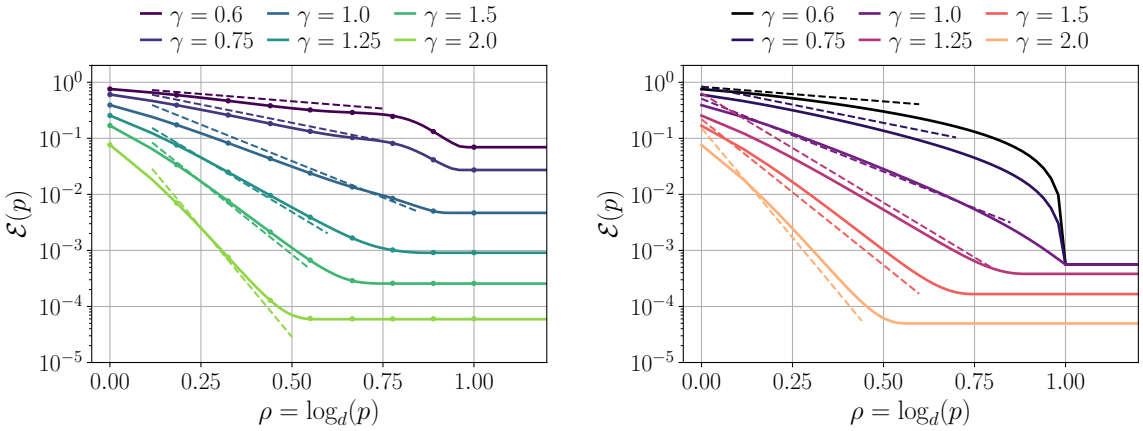


Figure 6: (Left) Test error scaling behavior in the under-regularized noiseless regime for different values of γ as a function of the width p, for fixed $d = 400$, $\alpha = 1.9$ and $\ell = -0.5$. (Right) Test error scaling behavior in the over-regularized noiseless regime for different values of γ as a function of the width p, for fixed $d = 400$, $\alpha = 2.5$ and $\ell = 0.1$.