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

Vittorio Erba   VITTORIO.ERBA@EPFL.CH
Statistical Physics of Computation laboratory, Lausanne, Switzerland

Emanuele Troiani   EMANUELE.TROIANI@EPFL.CH
Statistical Physics of Computation laboratory, Lausanne, Switzerland

Lenka Zdeborová   LENKA.ZDEBOROVA@EPFL.CH
Statistical Physics of Computation laboratory, Lausanne, Switzerland

Florent Krzakala   FLORENT.KRZAKALA@EPFL.CH
Information, Learning and Physics laboratory, Lausanne, Switzerland

Abstract

We study the high-dimensional asymptotics of empirical risk minimization (ERM) in over-parametrized two-layer neural networks with quadratic activations trained on synthetic data. We derive sharp asymptotics for both training and test errors by mapping the ℓ_2-regularized learning problem to a convex matrix sensing task with nuclear norm penalization. This reveals that capacity control in such networks emerges from a low-rank structure in the learned feature maps. Our results characterize the global minima of the loss and yield precise generalization thresholds, showing how the width of the target function governs learnability. This analysis bridges and extends ideas from spin-glass methods, matrix factorization, and convex optimization and emphasizes the deep link between low-rank matrix sensing and learning in quadratic neural networks.

1. Introduction

Modern machine learning relies heavily on training highly over-parameterized neural networks, which often generalize well despite having far more parameters than data points [51]. While it is known that large non-linear networks can approximate many functions [10], it remains unclear what these models actually learn in practice, and why training succeeds so often. In particular, we lack a precise understanding of how the structure of the data and the target function affects learnability, and how many samples are needed. Developing such an understanding remains a central theoretical challenge.

A promising route toward addressing these questions is to study models that go beyond the linear case—already well understood, eg. [24, 26, 43]—while remaining simple enough for rigorous analysis. Two-layer networks with quadratic activations are then the next natural candidate that captures some nonlinear behavior while still allowing for mathematical treatment [2, 18, 31]. On top of that working with synthetic data helps isolate the core mechanisms of learning and generalization [3, 11, 35, 36, 49], free from the confounding factors of real-world datasets.

In this paper, we thus study learning by empirical risk minimization with quadratic networks from data that is also generated by a quadratic network. Consider a dataset $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ where the data $\boldsymbol{x}^\mu \in \mathbb{R}^d$ are standard Gaussian $\boldsymbol{x}^\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ (though our results allow for some universality) for $\mu = 1, \ldots, n$, and the labels $y^\mu \in \mathbb{R}$ are generated by an unknown *target* function $f^\star(\mathbf{x})$. We

aim at learning this unknown function using a quadratic neural network $\hat{f}(x; W)$ with a number of hidden unit $m \geq d$:

$$\hat{y} = \hat{f}(\boldsymbol{x}; \hat{W}) := \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \sigma_k \left(\frac{\hat{\mathbf{w}}_k \cdot \boldsymbol{x}}{\sqrt{d}} \right), \tag{1}$$

where $\sigma_k(u) = u^2 - ||\boldsymbol{w}_k||^2/d$ is the (centered) quadratic activation, and where we collect the first-layer weights $\boldsymbol{w}_k \in \mathbb{R}^d$ for $k = 1, \ldots, m$ in the matrix $W \in \mathbb{R}^{m \times d}$. We learn W by empirical risk minimization of the square loss with ℓ_2 regularization (or equivalently, weight decay):

$$\hat{W} = \arg\min \mathcal{L}(W), \quad \text{where} \quad \mathcal{L}(W) := \sum_{\mu=1}^{n} \left(y^\mu - \hat{f}(\boldsymbol{x}^\mu; W) \right)^2 + \lambda \|W\|_F^2. \tag{2}$$

Given the structure of the model, the considered quadratic neural network can represent any centered positive semi-definite quadratic form of the input data—but no more. In particular, functions involving higher-order nonlinearities cannot be captured and are effectively treated as noise by the learner. To focus on the regime where generalization is possible, we therefore choose a target function $f^\star$ that lies within the expressivity class of the model (1) (we will also refer to the model as the *student* while thinking about the target as the *teacher*):

$$y^\mu = f^*(\boldsymbol{x}^\mu; W^*) + \sqrt{\Delta}\xi^\mu, \quad \text{with} \quad f^*(\boldsymbol{x}; W^*) := \frac{1}{\sqrt{m^*}} \sum_{k=1}^{m^*} \sigma_k \left(\frac{\mathbf{w}_k^\star \cdot \boldsymbol{x}}{\sqrt{d}} \right), \tag{3}$$

where $\xi^\mu \sim \mathcal{N}(0, 1)$ is an additional Gaussian label noise. As will shall see, our result will depend on $W^\star$ only through the spectral density of $S^\star = (W^\star)^T W^\star / \sqrt{m^\star d} \in \mathbb{R}^{d \times d}$, which needs to have a well-defined limit as $d \to \infty$.

We will work in the high-dimensional limit $d \to \infty$ with extensive-width target and quadratically many samples, i.e. the joint limit $d, n, m^*, m \to +\infty$ with

$$\alpha = n/d^2 = \mathcal{O}(1), \quad \kappa_* = m^*/d = \mathcal{O}(1), \quad \kappa = m/d = \mathcal{O}(1). \tag{4}$$

Our contributions. We provide an exact characterization of training and generalization in over-parameterized two-layer neural networks with quadratic activations, in the high-dimensional limit with Gaussian data. Our main result (Theorem 1) gives closed-form expressions for the training loss, generalization error, and spectral properties of the global minima of the regularized empirical risk (2), in the regime $\kappa = m/d \geq 1$, $\kappa^* = \mathcal{O}(1)$, and $\lambda > 0$.

Our solution to this problem connects three previously distinct lines of research that turn out to be related: the geometry and training dynamics of quadratic networks in teacher-student setups [2, 18, 33, 47], recent advances in high-dimensional Bayesian analysis of networks with extensive width [29, 31], and the role of implicit regularization in matrix factorization [21] and its connection to matrix compressed sensing and nuclear norm regularization [16, 46].

Concretely, we map the non-linear estimation problem for W in (2) to the linear one of estimating the matrix $S = W^\top W / \sqrt{md} \in \mathbb{R}^{d \times d}$, where, remarkably, the ℓ_2 regularization on W translates into a nuclear norm regularization on S. This reveals that the learning dynamics implicitly favor solutions $\hat{f}$ corresponding to narrow neural networks. We then study this equivalent matrix model by rigorous tools based on approximate message passing [7, 19] and their relation to convex optimization [28]. Our main result is Theorem 1, an analytical prediction for the test error achieved by the ERM, where

importantly the error does not depend on the value of κ as long as $\kappa > 1$, thus describing potentially massively over-parameterized models. Theorem 1 allows us to answer a range of questions as a function of κ^* (see Section A).

1. Location of the interpolation threshold: How many samples n are needed to have only a unique global minimum of (2) in the limit $\lambda = 0^+$, both in the noiseless and noisy case?

2. Generalization performance: How many samples n are needed to achieve perfect generalization with zero label noise?

3. Low-rank limit: What is the generalization performance in the limit $\kappa^* \ll 1$ corresponding to low-rank target functions?

In this paper, we provide new sharp results focusing on the extensive width case, encompassing the full range of target widths and focusing in particular on the setting $0 < \kappa^* < 1$, previously unexplored.

Finally, we remark that while (2) is *a priori* non-convex, it has no non-global local minima [2], implying that if a gradient-based algorithm converges to a minimum at all, it must converge to a global one. Hence, our main result Theorem 1 provides *a closed-form characterization of the behavior at convergence to minima of gradient-based algorithms* for any strictly positive regularization λ.

2. Setting and notations

We consider a dataset of n samples $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$, with $\boldsymbol{x}^\mu \in \mathbb{R}^d$ and $y^\mu \in \mathbb{R}$, constructed as in (1). We remark that the assumption of Gaussian data could be relaxed in the same spirit as in [50, Assumption 2.2], and denote by $\mathbb{E}$ the average of the training set. We will assume that the empirical spectral density of the matrix $S^* = (W^*)^T W^* / \sqrt{m^* d} \in \mathbb{R}^{d \times d}$ converges to a limiting distribution μ^* with finite first and second moment as $d \to \infty$ with $\kappa^* = m^*/d = \mathcal{O}(1)$, and call Q^* its second moment. As our running example, we will focus on the *Marchenko-Pastur (MP) target* case, where the weights $\boldsymbol{w}^*$ are such that the limiting distribution satisfies $\mu^*(x) = \sqrt{\kappa^*} \mu_{\mathrm{M.P.}}(\sqrt{\kappa^*} x)$, where $\mu_{\mathrm{M.P.}}$ is the Marchenko-Pastur distribution [32] with parameter κ^* (the asymptotic spectral distribution of $A^T A/m$ where $A \in \mathbb{R}^{m \times d}$ has i.i.d. zero-mean unit-variance components), and $Q^* = 1 + \kappa^*$. For example, this is the case for $\boldsymbol{w}^*$ with i.i.d. components extracted from a distribution with zero mean and unit variance, but we stress that our results hold also for deterministic targets, as well as targets with different spectral distributions.

We learn the dataset by empirical risk minimization on the loss (2), and unless stated otherwise, in this paper we will always consider learning with $m \geq d$, typically for $m > m^*$. We will measure the performance of the empirical risk estimator using the test error on the labels

$$e_{\text{test}}(W) = \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \left(f^*(\boldsymbol{x}; W^*) - \hat{f}(\boldsymbol{x}; W) \right)^2, \tag{5}$$

where the average is over a new test sample $\boldsymbol{x}$ with the same distribution as the training samples (notice that for the sake of the test error we do not add any label noise on the test label). We will always consider the high-dimensional limit (4).

3. Main theorem

Our main technical result is the characterization of the properties of the global minima of the empirical loss (2) in the high dimensional limit in terms of training and test error.

Theorem 1 (Asymptotics of ERM (2), informal) *Consider the setting of Section 2 with $\kappa \geq 1$. Define $\tilde{\lambda} = \sqrt{\kappa}\lambda$ and $\mu_\delta^* = \mu^* \boxplus \mu_{s.c.,\delta}$, where $\boxplus$ is the free convolution and $\mu_{s.c.,\delta} = \sqrt{4\delta^2 - x^2}/(2\pi\delta^2)$ the semicircle distribution of radius 2δ for $\delta > 0$. Call $(\bar{\delta}, \bar{\epsilon}) \in \mathbb{R}_+^2$ the unique solution of*

$$\begin{cases} 4\alpha\delta - \frac{\delta}{\epsilon} = \partial_1 J(\delta, \tilde{\lambda}\epsilon) \\ Q^* + \frac{\Delta}{2} + 2\alpha\delta^2 - \frac{\delta^2}{\epsilon} = (1 - \epsilon\tilde{\lambda}\partial_2)J(\delta, \tilde{\lambda}\epsilon) \end{cases} \quad \text{where } J(a, b) = \int_b^{+\infty} dx\, \mu_a^*(x)\, (x - b)^2 . \quad (6)$$

Then, for all values of $\alpha, \kappa_*, \lambda > 0$, $\Delta \geq 0$ and $\kappa \geq 1$ any global minimum $\hat{W}/\sqrt[4]{md}$ of (2) satisfies

$$\lim_{d \to \infty} \mathbb{E}e_{\text{test}}(\hat{W}) = 2\alpha\bar{\delta}^2 - \frac{\Delta}{2}, \qquad \lim_{d \to \infty} d^{-2}\mathbb{E}\mathcal{L}(\hat{W}) = \frac{\bar{\delta}^2}{4\bar{\epsilon}^2} - \frac{\tilde{\lambda}}{2}\partial_2 J(\delta, \tilde{\lambda}\epsilon) . \quad (7)$$

Moreover, the empirical singular value density of $\hat{W}$ satisfies

$$\lim_{d \to \infty} \mathbb{E}\frac{1}{d}\sum_{i=1}^d \delta(x - \sigma_i) = F_{\bar{\delta}}(\tilde{\lambda}\bar{\epsilon})\delta(x) + I(x > 0)\left[2x\mu_{\bar{\delta}}(x^2 + \tilde{\lambda}\bar{\epsilon})\right] \quad (8)$$

where $\{\sigma_i\}_{i=1}^d$ are the singular values of $\hat{W}$, I is the indicator function and F_δ is the c.d.f. of μ_δ^*.

Sketch of the proof. The proof, detailed in Appendix B, proceeds as follows. We use a reduction of a regularized version of the minimization (2) to the one of a positive semi-definite matrix estimation:

$$\hat{S} = \arg\min_{S \succeq 0} \tilde{\mathcal{L}}(S), \quad \text{with } \tilde{\mathcal{L}}(S) := \sum_{\mu=1}^n \text{Tr}\left[X(\boldsymbol{x}^\mu)(S - S^*)\right]^2 + \sqrt{md}\left(\lambda\,\text{Tr}(S) + \tau\|S\|_F^2\right) \quad (9)$$

where we defined $S^* = W^*(W^*)^T/\sqrt{m^*d}$ and $X(\boldsymbol{x}) = (\boldsymbol{x}\boldsymbol{x}^T - \mathbb{I}_d)/\sqrt{d}$, and where we have added a Frobenius norm penalty of amplitude τ. Notice that (9) is strongly convex for $\lambda, \tau > 0$, so that $\hat{S}$ is unique. Under the mapping $S(W) = WW^T/\sqrt{md}$, we have $\hat{S} = S(\hat{W})$ for any (Frobenius regularized) solution of (2), due to the uniqueness of $\hat{S}$. The second part is to use the Gaussian universality principle, that allows to replace each matrix $(\boldsymbol{x}\boldsymbol{x}^T - \mathbb{I}_d)/\sqrt{d}$ by a random Wigner matrix, following closely the steps of [31, 50]. This in turns reduces the problem to a rank-penalized matrix recovery problem with $\text{GOE}(d)$ sensing matrices:

$$\hat{S} = \arg\min_{S \succeq 0} \tilde{\mathcal{L}}_G(S), \quad \text{with } \tilde{\mathcal{L}}(S) := \sum_{\mu=1}^n \text{Tr}\left[Z^\mu(S - S^*)\right]^2 + \sqrt{md}\left(\lambda\,\text{Tr}(S) + \tau\|S\|_F^2\right) \quad (10)$$

The problem can then be studied in various way. For instance with the heuristic, but powerful, replica method technique from statistical physics (see for instance [30]). To provide a rigorous approach, we use instead a suitable Approximate Message Passing (AMP) algorithm with non-separable prior [7, 19], designed in such a way as its (unique) non-trivial fixed point is also the fixed point of projected gradient descent on (9) [5, 28, 40], which by the convexity of (9) coincide with the unique $\hat{S}$. We
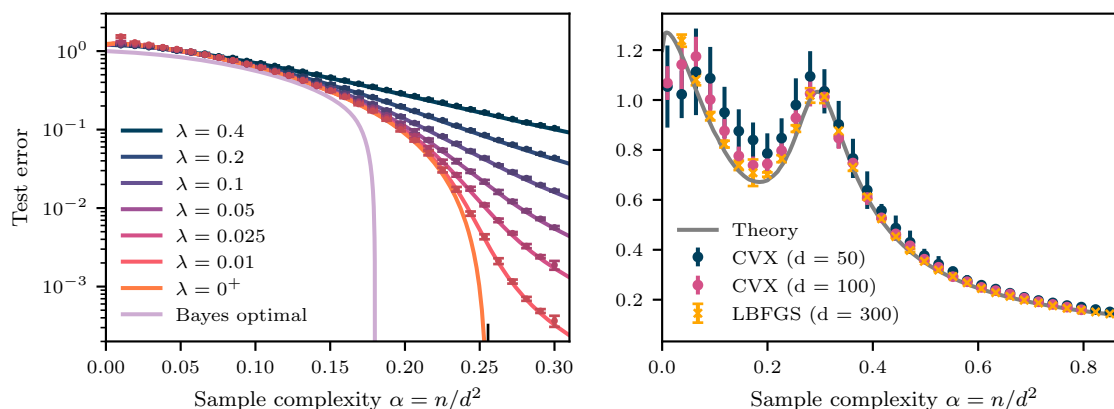
4

Figure 1: Left: Test error of simulations of vanilla GD (crosses, error bars are the standard deviation over 16 realizations of the target/training set at $d = 300$) compared with the results of Theorem 1 (lines) as a function of the number of samples $n = \alpha d^2$, noiseless case $\Delta = 0$. We observe a perfect match, particularly striking in the regime of small test error. The purple line is the Bayes-optimal performance [31]. Right: Test error of simulations of GD run with LBFGS on (2) (yellow dots, $d = 300$) and of a convex solver run on the equivalent convex matrix problem (9) (blue dots $d = 50$, purple $d = 100$ dots), for $\Delta = 0.5$ and $\lambda = 0.02$ and as a function of the number of samples $n = \alpha d^2$. Error bars are the standard deviation over 16 realizations of the target/training set, compared with the result of Theorem 1 (gray line).

then write the state evolution of this AMP algorithm which gives us all the relevant characteristics of the minimizer. Simple manipulations of the state evolution equations, in the limit $\tau \to 0$, lead to the characterization (6). □

Theorem 1 provides an asymptotic result for train and test error, as well as a characterization of the singular values of the optimal neural network weights. It is of independent interest for the matrix compressed sensing problem (10) and extends directly to matrix problems with *any* convex loss and spectral regularization under convex constraints.

A remarkable observation is that the ℓ_2 regularization over the weights W naturally translates to a nuclear norm regularization in the equivalent matrix problem (the convexification of the minimum rank regularization), naturally favoring model weights configurations with an effective lower width (i.e., implementable with fewer hidden units): a weight decay in W thus implies a low nuclear norm of the matrix S.

Theorem 1 also implies that the properties of the global minima do not depend on the network width m, as long as $m \geq d$. This means that a neural network with width $\kappa \gg 1$ will achieve the same test error as a much narrower network with $\kappa = 1$, when trained on the same data, provided that the regularization strength is appropriately matched. Theorem 1 thus described both mildly and massively over-parametrized models since for very large κ the number of learnable parameters will be not only much larger than the target function, but also can be much larger than the number of samples.

5

Overview of Appendix A. In Appendix A, we present the main consequences of Theorem 1. We first analyze the learning curves of the empirical risk minimizer in both noiseless and noisy settings. We then define and characterize the interpolation threshold $\alpha_{\text{inter}}(\kappa, \Delta)$, providing an exact formula and its limits for vanishing and large label noise. Next, we derive the strong recovery threshold $\alpha_{\text{strong}}(\kappa)$ in the noiseless setting, identifying when test error vanishes. Finally, we analyze the $\kappa^* \to 0$ limit, providing a closed-form expression for the test error in this regime and highlighting the presence of a BBP-type phase transition.

Numerical implementation. We provide the code for the solution of the equations in Theorem 1 and the experiments at https://github.com/SPOC-group/OverparametrisedNet.

Acknowledgments

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 3(3):224–294, September 2014. ISSN 2049-8764, 2049-8772.

[2] Yossi Arjevani, Joan Bruna, Joe Kileel, Elzbieta Polak, and Matthew Trager. Geometry and optimization of shallow polynomial networks. *arXiv preprint arXiv:2501.06074*, 2025.

[3] Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

[4] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

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

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] 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.

[8] Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[9] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. On double-descent in uncertainty quantification in overparametrized models. In *International Conference on Artificial Intelligence and Statistics*, pages 7089–7125. PMLR, 2023.

[10] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[11] Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1262–1262. PMLR, 2024.

[12] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36:54754–54768, 2023.

[13] Jairo RL de Almeida and David J Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.

[14] C De Dominicis and Imre Kondor. Eigenvalues of the stability matrix for parisi solution of the long-range spin-glass. *Physical Review B*, 27(1):606, 1983.

[15] 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.

[16] 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.

[17] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A Unifying Tutorial on Approximate Message Passing. *Foundations and Trends® in Machine Learning*, 15 (4):335–536, 2022. ISSN 1935-8237, 1935-8245.

[18] David Gamarnik, Eren C. Kızıldağ, and Ilias Zadik. Stationary Points of Shallow Neural Networks with Quadratic Activation Function, July 2020.

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

[20] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[21] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

[22] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[23] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[24] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[25] Florent Krzakala and Lenka Zdeborová. Statistical physics methods in optimization and machine learning. *Lecture Notes*, 2024.

[26] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

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

[28] 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.

[29] Antoine Maillard and Afonso S. Bandeira. Exact threshold for approximate ellipsoid fitting of random points, October 2023.

[30] Antoine Maillard and Dmitriy Kunisky. Fitting an ellipsoid to random points: Predictions using the replica method, June 2024.

[31] 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.

[32] VA Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.(NS)*, 72(114):4, 1967.

[33] 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.

[34] Simon Martin, Francis Bach, and Giulio Biroli. On the Impact of Overparameterization on the Training of a Shallow Neural Network in High Dimensions. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 3655–3663. PMLR, April 2024.

[35] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[36] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

[37] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

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

[39] Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

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

[41] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[42] Samet Oymak and Babak Hassibi. Sharp MSE Bounds for Proximal Denoising. *Foundations of Computational Mathematics*, 16(4):965–1029, August 2016. ISSN 1615-3375, 1615-3383.

[43] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

[44] Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

[45] 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.

[46] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[47] Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 13445–13455. Curran Associates, Inc., 2020.

[48] Matteo Vilucchio, Yatin Dandi, 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.

[49] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

[50] Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental Limits of Matrix Sensing: Exact Asymptotics, Universality, and Applications, March 2025. arXiv:2503.14121 [stat].

[51] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021.

THE NUCLEAR ROUTE: SHARP ASYMPTOTICS OF ERM IN OVERPARAMETERIZED QUADRATIC NETWORKS

Appendix A. Main consequences of Theorem 1

In this section we focus specifically on the Marchenko-Pastur target case (see Section 2, details on the computation of μ_δ^* in Result 1 are given in Appendix F). We believe that our results generalize (qualitatively for the learning curves, and as is for the thresholds and the low-rank target limit) to any target such that μ^* has a finite spectral gap, i.e. such that $\lambda_{\min} = \min\{x \in \mathrm{supp}(\mu^*)|x > 0\}$ satisfies $\lambda_{\min} > 0$, with mass in zero equal to $\max(0, 1 - \kappa^*)$ (i.e. rank of the associated target matrix $\lfloor \min(\kappa^*, 1)d \rfloor$) and second moment $Q^* = 1 + \kappa^*$.

Learning curves.   In Figure 2 we show the test and train error of any global minimum of the loss function (2) given by Theorem 1, both the noiseless ($\Delta = 0$) and noisy ($\Delta = 0.5$) scenarios. Details on the numerical experiments are given in Appendix F.

In the noiseless case (Figure 2 left), we plot the test error as functions of the sample ratio α for a finite regularization value ($\lambda = 0.4$, blue line), in the limit of vanishing regularization ($\lambda \to 0^+$, yellow line), and for the optimal choice of regularization (dashed line). These are contrasted with the Bayes-optimal performance as derived in [31] (purple line), clearly illustrating the effect of regularization on the perfect recovery threshold, marked by the vertical gray line as per Corollary 1.

In the noisy case (Figure 2 center and right), we show both the test and train losses as functions of the sample ratio α for multiple values of λ (solid lines), including the minimum regularization interpolator limit ($\lambda \to 0^+$, yellow line) and the optimally regularized scenario (dashed line). Notably, we highlight the specific region of α where the non-regularized training loss reaches zero (vertical gray line, as per Result 1), which coincides with the emergence of a cusp – the double descent peak [6, 41] – in the test error as λ decreases.

Interpolation threshold.   We define the interpolation threshold $\alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ as the value of the sample ratio at which the set of PSD matrices S for which the loss (9) is zero for $\lambda = 0$ (we call those the interpolators of the training dataset) shrinks to a single point. For $\alpha > \alpha_{\mathrm{inter}}(\kappa^*, \Delta)$, (9) thus has a single global minimum with positive training loss if $\Delta > 0$, or with both zero training and test loss if $\Delta = 0$. For $\alpha < \alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ instead, the non-regularized $\lambda = 0$ loss admits many global minima (all PSD interpolators), while the limit $\lambda \to 0^+$ has still a single global minimum with zero training loss, and minimum value of the regularization norm. While we stated Theorem 1 for $\lambda > 0$ it also holds, at least at an heuristic level, for $\alpha > \alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ and $\lambda = 0$ by directly plugging $\lambda = 0$ in (6). The theorem *does not hold* for $\lambda = 0$ and $\alpha < \alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ while it can be adapted to the limit $\lambda \to 0^+$ by a careful rescaling, which we comment upon in Appendix C. A direct consequence of Theorem 1 is a theoretical characterization of the position of the interpolation threshold $\alpha_{\mathrm{inter}}(\kappa^*, \Delta)$.

Result 1 (Interpolation threshold)   *Consider the setting of Section 2 for a Marchenko-Pastur target. Then, the interpolation threshold $\alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ satisfies*

$$\alpha_{\mathrm{inter}}(\kappa^*, \Delta) = \frac{1}{4\bar{\delta}} J^{(1,0)}\left(\bar{\delta}, 0\right) + o_d(1) \quad where \quad Q^* + \frac{\Delta}{2} = J\left(\bar{\delta}, 0\right) - \frac{\bar{\delta}}{2} \partial_1 J^{(1,0)}\left(\bar{\delta}, 0\right), \quad (11)$$

where ∂_1 denotes derivative w.r.t. the first argument, and J was defined in (6). Additionally, we have

$$\lim_{\Delta \to 0^+} \alpha_{\mathrm{inter}}(\kappa^*, \Delta) = \frac{1}{4} \begin{cases} 1 + 2\kappa_* - \kappa_*^2 & if\ 0 < \kappa^* < 1 \\ 2 & if\ \kappa^* \geq 1 \end{cases}, \quad \lim_{\Delta \to +\infty} \alpha_{\mathrm{inter}}(\kappa^*, \Delta) = \frac{1}{4}. \quad (12)$$
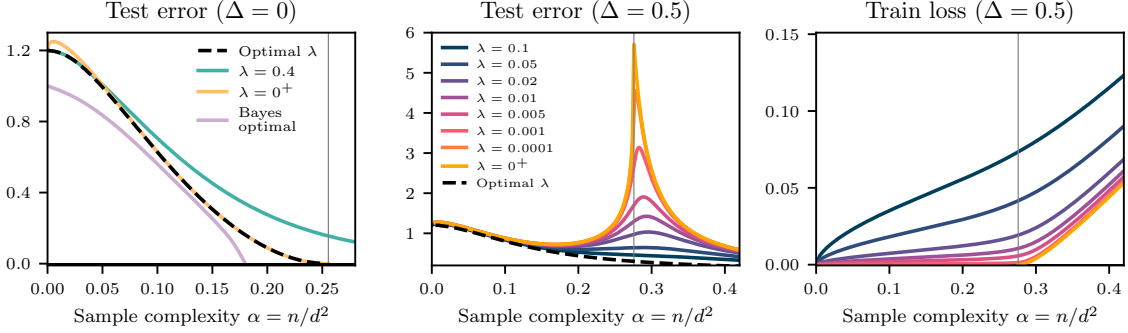
11

Figure 2: (Left) The test error of any global minimum of (2) (Theorem 1) in the noiseless case $\Delta = 0$ for finite regularization $\lambda = 0.4$ (blue line), in the limit $\lambda \to 0^+$ (yellow line) and for optimal regularization (dashed line). We compare with the Bayes-optimal performance [31] (purple line), and highlight the strong recovery threshold (vertical gray line, see Corollary 1). (Center, Right) The test and train loss (2) in the noisy case $\Delta = 0.5$ for several values of the regularization λ (solid lines), $\lambda \to 0^+$ (yellow line) and for optimal regularization (dashed line). We highlight the region of sample ratio α where non-regularized training loss goes to zero (before the vertical grey line, from Result 1), which coincides with the development of a cusp in the test error as λ decreases.

Result 1 is derived derived by considering the solution of the minimization problem (6) and setting $\lambda = 0$ directly at large α – where there is a unique minimum of (9) – and then looking for the smallest value of α consistent with a unique minimum assumption. Equivalently, this is when the asymptotics training loss reach zero (see Appendix D). We plot (12) in Figure 3. As discussed above, the interpolation threshold does not depend on the width of the student.

It is interesting to contrast our results with the *ellipsoid-fitting problem*, where one seeks a PSD matrix S such that a dataset of points $\{x^\mu\}_{\mu=1}^n$ lies on the surface of the associated ellipsoid. This problem, recently studied in a high-dimensional limit similar to (4) [29, 30], predicts that fitting is possible as long as $\alpha < 1/4$. We recover this same threshold, $\alpha_{\text{inter}} = 1/4$, in two extreme cases: when noise dominates ($\Delta \to +\infty$) or when the target is vanishingly small ($\Delta = 0, \kappa^* \to 0^+$). In all other scenarios, however, the interpolation threshold is strictly larger, reflecting the structural advantage of the data over random labeling. Unlike linear ridge regression or kernel methods, the interpolation threshold here is nontrivial and intricately linked to data structure. Notably, it is not simply determined by the ratio of samples to parameters. Despite the effective number of parameters scaling as $\mathcal{O}(d^2/2)$, interpolation can occur well before this count if the target function is sufficiently narrow ($0 < \kappa^* < 1$). This effect arises from the PSD constraint in the equivalent matrix problem (9), marking a fundamental departure from classic kernel theory.

We note that special cases this threshold has been considered in the literature, for $\kappa^* \geq 1$ and $\Delta = 0$ in [18], and our result agrees with these works. An analogue of this threshold (still for $\Delta = 0$) has been considered in [47] for the case of low-width target $m^* = 1$, where the interpolation for the model with $\kappa \geq 1$ happens at $n = 2d$. The authors conjecture without a solid theoretical justification that for generic m^* the interpolation happens at $n = d(m^* + 1) - m^*(m^* + 1)/2$,
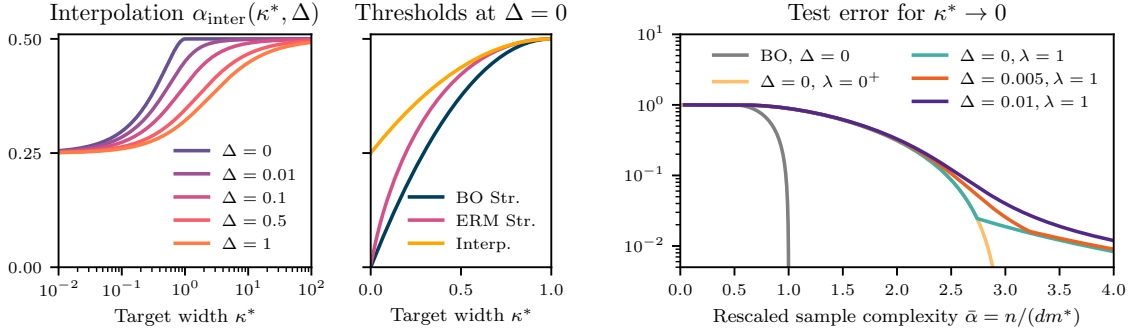
Figure 3: (Left) Interpolation threshold $\alpha_{\mathrm{inter}}(\kappa^*, \Delta)$ as a function of κ^* for several values of label noise Δ (Result 1). Notice the convergence to the $1/4$ random-label-fitting threshold for very narrow targets $\kappa^* \ll 1$ and large label noise $\Delta \gg 1$. (Center) Comparison between interpolation threshold (Result 1, $\Delta = 0$) and strong recovery threshold (Corollary 1) of the global minima of (2), with the BO strong recovery threshold [31]. Minimal regularization interpolators of (2) reach perfect recovery well before the interpolator set shrinks to a singleton on the target weights (the effect is more pronounced for very small ranks of the target function $\kappa^* \ll 1$. (Right) The test error of any global minimum of (2) in the limit $\kappa^* \to 0$ (Result 2) for several values of regularization $\lambda = \bar{\lambda}/\sqrt{\kappa^*}$ and label noise Δ, compared with the Bayes-optimal [31].

which, however, is not compatible with our results. We conclude that their conjecture is incorrect in the regime $m^* = \mathcal{O}_d(d)$, as it predicts interpolation for values lower than the rigorous $1/4$ random labels threshold [29, 30] as well.

Strong recovery threshold. A different phenomenon can be studied in the noiseless case $\Delta = 0$ and $\lambda \to 0^+$, i.e. what we call the *strong recovery*: the value of the sample ratio $\alpha_{\mathrm{strong}}(\kappa^*) = n_{\mathrm{strong}}/d^2$ such that, for $\alpha > \alpha_{\mathrm{strong}}$ the *test error* of the global minima of (2) is zero, and for $\alpha < \alpha_{\mathrm{strong}}$ is strictly positive at $\Delta = 0$. Notice the difference with the previous paragraph, where the interpolation peak was concerning the training error vanishing in the $\Delta > 0$ case, while we here consider the perfect test error at $\Delta = 0$. This limit was studied in the context of matrix model (9) in the minimal-norm interpolation setting [1, 15, 42]. Here we re-derive their result independently as a consequence of Theorem 1.

Corollary 1 (Strong recovery threshold) *Consider the setting of Section 2 for Marchenko-Pastur targets with $\Delta = 0$, $\lambda \to 0^+$ and $\kappa \geq 1$. Define for $x \in [-2, 2]$ the incomplete moments of the semicircle distribution*

$$M_{\mathrm{s.c.}}^{(k)}(x) = \int_{-2}^{x} dx \, \mu_{\mathrm{s.c.}}(x) \, x^k = \frac{1}{2\pi} \int_{-2}^{x} dx \, \sqrt{4 - x^2} \, x^k \,, \tag{13}$$

and call $\bar{c}$ the solution of the equation

$$M_{\mathrm{s.c.}}^{(1)}(c) - c M_{\mathrm{s.c.}}^{(0)}(c) + \frac{c}{1 - \kappa^*} = 0 \,, \tag{14}$$

for $0 < \kappa^* < 1$. Then,

$$
\alpha_{\text{strong}} = \begin{cases} \frac{1}{2} - \frac{1}{2}(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(c) - c M_{\text{s.c.}}^{(1)}(c) \right) & \text{if} \quad 0 < \kappa^* < 1 \\ \frac{1}{2} & \text{if} \quad \kappa^* \geq 1 \end{cases} + o_d(1). \tag{15}
$$

Informally, the strong recovery threshold presented in Result 1 is derived by expanding (6) for $\delta \to 0^+$ within an appropriate scaling ansatz, as the test error is proportional to δ^2 (Theorem 1, noiseless case $\Delta = 0$), and finding the value $\alpha = \alpha_{\text{strong}}$ such that the expansion is consistent. We provide all the details in Appendix D. We plot (15) in Figure 3, comparing with the strong recovery threshold of the BO estimator [31] and with the interpolation threshold Result 1 for $\Delta = 0$.

Small target rank limit. Finally, we consider the small target width limit $\kappa^* \to 0^+$ for κ bounded away from zero. This is an example of large over-parametrization, as the ratio between student and target widths m/m^* diverges. In the limit $\kappa^* \to 0$, i.e. $m^* \text{l} d$, we define the rescaled sample ratio $\bar{\alpha} = \alpha/\kappa^* = n/(m^*d)$. The natural baseline to compare this limit against is given by the analogue limit for the test error of the BO estimator given in [31].

Result 2 (Test error in the $\kappa^* \to 0$ limit) *Consider the setting of Section 2. Consider the limit $\kappa^* \to 0^+$ for any fixed $\kappa > 0$, $\bar{\alpha} = \alpha/\kappa^*$ and $\bar{\lambda} = \lambda\sqrt{\kappa^*}$. Then, if $\Delta = 0$ and $\bar{\lambda} \to 0^+$ the test error satisfies*

$$
\lim_{\kappa^* \to 0^+} \lim_{d \to \infty} \mathbb{E}\, e_{\text{test}}(\hat{S}) = \begin{cases} 1 & \text{if} \quad 0 < \bar{\alpha} \leq 1/2, \\ \frac{2}{9}\bar{\alpha}(4 - \sqrt{6\bar{\alpha} - 2})^2 & \text{if} \quad 1/2 < \bar{\alpha} \leq 3 \\ 0 & \text{if} \quad \bar{\alpha} \geq 3 \end{cases} \tag{16}
$$

In Appendix E we provide analogous expressions, albeit less explicit, for generic regularization $\bar{\lambda} > 0$ and noise $\Delta \geq 0$. We just state that for all $\bar{\lambda} > 0$ and $\Delta \geq 0$, and for $\alpha < \bar{\alpha}_{\text{weak}}(\bar{\lambda}, \Delta)$ the test error equals one, where

$$
\bar{\alpha}_{\text{weak}}(\bar{\lambda}, \Delta) = \max\left(\frac{1 + \Delta/2}{2}, \frac{\bar{\lambda} - 2(1 + \Delta/2)}{4} \right), \tag{17}
$$

and that that for all $\bar{\lambda} > 0$ and $\Delta > 0$ and for $\bar{\alpha} \to +\infty$ he have

$$
\lim_{\kappa^* \to 0^+} \lim_{d \to \infty} \mathbb{E}\, e_{\text{test}}(\hat{S}) = \frac{3\Delta}{2\alpha}(1 + o_\alpha(1)). \tag{18}
$$

Result 2 provides a closed-form expression for the test error in the $\kappa \to 0$ limit. It is derived by a direct expansion of (6) at small κ, and it involves a BBP-like phase transition [4] in the spectrum of $\hat{W}$. We notice that Result 2 holds for any extensive student width $\kappa > 0$ due to the fact that the rank of the global minima of (2) is vanishing as $\mathcal{O}(\kappa^*)$.

Figure 3 right shows the test error in the $\kappa^* \to 0$ limit for a selection of parameters. We see that the solution at $\lambda, \Delta \neq 0$ as a non-trivial behavior, tracking closely but not exactly the $\lambda \to 0^+$, $\Delta \neq 0$ curve until an angular point is reached. We observe that for $\bar{\alpha} < \bar{\alpha}_{\text{weak}}(\bar{\lambda}, \Delta)$, the empirical risk minimizer achieves BO performance, even though that is given by the trivial prior-average estimator. Moreover, for $\Delta = 0$ and $\bar{\lambda} \to 0^+$, the empirical risk minimizer achieves strong recovery for $\bar{\alpha} = 3$ in accordance with Corollary 1. Finally, we remark that for fixed $\bar{\lambda} > 0$ and $\Delta > 0$, the test error goes to zero as $1/\alpha$, which is the Bayes-optimal rate.

Appendix B. Proof of Theorem 1

In this section we discuss theorem 1 for the minima of

$$\hat{W} = \arg\min \mathcal{L}(W), \quad \text{where} \quad \mathcal{L}(W) := \sum_{\mu=1}^{n} \left(y^{\mu} - \hat{f}(\boldsymbol{x}^{\mu}; W) \right)^2 + \lambda \|W\|_F^2 + \tau \|WW^T\|_F^2 \,.$$

(19)

which, due to the added regularization, is strongly convex for $\tau, \lambda > 0$ and hence it admits a unique global minimum.

The proof of Theorem 1 goes as follows:

1. In Section B.1, we show how to map the original learning problem into an equivalent matrix estimation problem. This, in turn, can be reformulated as a vector Generalized Linear Model (GLM) with a non-separable regularization.

2. In Section B.2, we use Gaussian universality and show that this model is asymptotically equivalent to a Gaussian sensing problem, leveraging recent results on universality in high-dimensional estimation.

3. The resulting Gaussian model can then be analyzed using the Generalized Approximate Message Passing (GAMP) algorithm. We briefly recall the relevant properties of GAMP in Section B.3. To apply GAMP rigorously, we exploit the fact that its fixed points coincide with the solutions of the associated convex optimization problem (Section B.3.1). These fixed points are characterized by a set of deterministic equations known as state evolution (Section B.3.2). With appropriate initialization (see Corollary 2), GAMP converges to the correct fixed point (Theorem 5), which corresponds to the global minimizer of our original problem. Hence, the fixed points of the associated state evolution equations describe the properties of the global minima of (1). Notice that this discussion is generic (i.e. allows for generic convex losses and regularizations), as long as the global minimum of the associated matrix problem is unique.

4. Finally, in Appendix B.4, we describe the state evolution equations that precisely characterizes the asymptotic behavior of the minimizer of the empirical risk (2), which is precisely equivalent to the results given in Theorem 1 in the limit $\tau \to 0^+$.

B.1. Mapping to a matrix model

The mapping of the data part of the loss is straightforward. For the ℓ_2 regularization on the network weights we have

$$\sum_{k=1}^{m} \|\boldsymbol{w}_k\|_2^2 = \sum_{i,k=1}^{d,m} W_{ki}^2 = \sqrt{md} \sum_{i,k=1}^{d,m} \frac{W_{ki}^2}{\sqrt{md}} = \sqrt{md} \sum_{i=1}^{d} S(W)_{ii} = \sqrt{md} \operatorname{Tr}(S(W)) \,.$$

(20)

The same goes for the part proportional to τ in (19), that become the Frobenius norm of $\hat{S}$. This leads to the equivalent problem:

$$\hat{S} = \arg\min_{S \succeq 0} \tilde{\mathcal{L}}(S), \quad \text{with} \quad \tilde{\mathcal{L}}(S) := \sum_{\mu=1}^{n} \operatorname{Tr} \left[\frac{\mathbf{x}_\mu \mathbf{x}_\mu^T - \mathbb{I}_d}{\sqrt{d}} (S - S^*) \right]^2 + \sqrt{md} \left(\lambda \operatorname{Tr}(S) + \tau \|S\|_F^2 \right) \,.$$

(21)

15

B.2. Data universality

The universality of the minimal error (or "ground state energy" in the physics jargon) follows directly from a rich line of work on universality of empirical risk minimization [20, 22, 29, 34, 39, 50]. These results establish that, under broad conditions, the asymptotic behavior of the test and training errors becomes equivalent to those with a Gaussian sensing matrix. Given the maturity of this theory, we will refer directly to these foundational works and provide a concise sketch of the main arguments for completeness.

The universality follows directly from Proposition 2.1 in [29], who derived it in the related context of ellipsoid fitting. This was adapted to the "planted" case in [50, Section 3.1, Lemma 3.3]. The proof strategy is to consider an interpolating model where each data matrix is given by $U(t) = \cos(t)(X_\mu X_\mu^T - \mathbb{I}_d)/\sqrt{d} + \sin(t)G_\mu$ in $t \in [0, 2\pi]$, from the original model at $t = 0$ to the one with a $\mathrm{GOE}(d)$ random matrix G_μ at $t = \pi/2$, and to show that the "time" t does not change the expected loss. We refer to these works (and to [39]) for the detailed proofs, and directly state the universality result:

Theorem 2 (Gaussian Universality of the loss, from [31, 50]) *Let $\Psi : \mathbb{R} \to \mathbb{R}$ be a bounded differentiable functions with bounded derivative, then for any finite $\alpha = n/d^2$, the minimum of the training loss $\mathrm{GS}(\{\Phi_\mu\}_{\mu=1}^n) := \arg\min_S \tilde{\mathcal{L}}(S)(\{\Phi_\mu\}_{\mu=1}^n)$ is universal with respect to the input data Φ_μ:*

$$\lim_{d \to \infty} \left| \mathbb{E}_{\{\boldsymbol{x}_\mu\}} \psi \left[GS_d \left(\left\{ \frac{\boldsymbol{x}_\mu \boldsymbol{x}_\mu^T - \mathbb{I}_d}{\sqrt{d}} \right\}_{\mu=1}^n \right) \right] - \mathbb{E}_{\{G_\mu\}^{i.i.d.}_\sim GOE(d)} \psi \left[GS_d(\{G_\mu\}_{\mu=1}^n) \right] \right| = 0. \quad (22)$$

if the matrices $(\{X_\mu X_\mu^T - \mathbb{I}_d\})/\sqrt{d}$ respects the so-called one-dimensional CLT property (Assumption 2.2, [50]).

The required one-dimensional CLT [12, 20, 39] property is a point-wise normality of the projection of the operator, proven in the case of Gaussian $\boldsymbol{x}^\mu$ in [29, Lemma 4.8].

The universality argument can be readily adapted to other quantity as well, in order to show that not only the loss, but also other observables such as the overlaps are universal. We will not repeat the full proof, and directly use Theorem 3 in [12] that can be applied *mutatis mutandis*[1]:

Theorem 3 (Universality of the overlaps [12], informal) *For any bounded-Lipchitz function $\Phi : \mathbb{R}^{d^2} \times \mathbb{R}^{d^2} \to \mathbb{R}$, we have:*

$$\mathbb{E} \left[\Phi \left(\hat{S}, S^* \right) \right] \xrightarrow[\substack{n,d \to +\infty \\ n/d^2 = \alpha > 0}]{} \mathbb{E}_{\mathcal{G}} \left[\Phi \left(\hat{S}, S^* \right) \right]$$

where $\hat{S} = \arg\min_S \tilde{\mathcal{L}}(S)$, and S^* determines the target function.

---

1. The argument is a classic approach of large deviation, or Legendre transform. This is done by applying the universality to a *perturbed* version of the loss, where —to use the parlance of physics— a source term, or a regularization, such as $\lambda_S \mathrm{Tr}[\hat{S}^T S^*]$, is added. Due to the convexity of the loss in λ_S the derivative with respect to λ_S must converges to the same exact quantity in both the original and the equivalent model, since they converge point-wise to the same limit.

This directly can be applied to the overlap $\text{Tr}[\hat{S}^T \hat{S}]$ and $\text{Tr}[\hat{S}^T S^*]$, which are thus universal, as well as to the test error. Notice indeed that

$$e_{\text{test}}(\hat{w}) = \frac{1}{2}\mathbb{E}_x \left(f_{m^*}(x; w^*) - f_m(x; \hat{w})\right)^2 = \frac{1}{d}||S(w) - S(w^*)||_F^2, \tag{23}$$

leading to an expression for the test error as a function of the overlaps. Finally, note that one can apply the same argument to show the universality of any finite $k \in \mathbb{Z}_+$ moment $d^{-1} \text{Tr}[\hat{S}^k]$ (hinging crucially on the PSD constraint to claim that all such perturbations are convex). With mild regularity assumptions (for e.g. compact support, which is satisfied by the Marchenko-Pastur target case studied in detail in the main text), this implies the universality of the spectral distribution of $\hat{S}$ as well.

B.3. Generalized Approximate Message Passing algorithm and state evolution equations

We review here how to use GAMP to solve asymptotics convex optimization problems. Later, in section B.4.1 we show indeed that the present PSD matrix optimization problem can be written as a vector one, with a non-separable denoiser.

We will work in a vector setting, where one has sensing matrices $A \in \mathbb{R}^{N \times D} \sim \mathcal{N}(0, 1/D)$ (i.e. N samples in D dimension) with rows a^μ. Notice that in this scaling $||u||^2$ should be always of order $\mathcal{O}(D)$.

The generic form of the "rectangular" GAMP algorithm reads [23]:

$$\begin{aligned} u^{t+1} &= A^T g_t(v^t) + d_t e_t(u^t), \\ v^t &= A e_t(u^t) - b_t g_{t-1}(v^{t-1}), \end{aligned} \tag{24}$$

with initialization $u^{t=0}$ and the convention that $g_{t-1}(\cdot) = 0$. Here $v \in \mathbb{R}^N$, $u \in \mathbb{R}^D$, $g : \mathbb{R}^N \to \mathbb{R}^N$ and $e : \mathbb{R}^D \to \mathbb{R}^D$.

The terms b_t, d_t are usually called *Onsager's reaction terms* [37], and are tuned in such a way as to erase time-to-time correlations between the iterates (u^t, v^t) and the sensing matrix A, so that from a statistical point of view of (u^t, v^t) it is as if the matrix A was resampled independently at each time-step (conditioned on the previous iterates $\{(u^s, v^s)\}_{s=1}^{t-1}$). One has

$$d_t = -\frac{1}{D}\nabla \cdot g_t(v^t) \quad \text{and} \quad b_t = \frac{1}{D}\nabla \cdot e_t(u^t), \tag{25}$$

where $\nabla \cdot f = \sum_{i=1}^d \partial_i f_i$ denotes the divergence of a function $f : \mathbb{R}^d \to \mathbb{R}^d$. We will see later that in high-dimension $D \gg 1$, $N = \mathcal{O}(D)$, the Onsager's term concentrate and do not depend on the iterates u^t and v^t anymore.

B.3.1. GAMP FOR CONVEX OPTIMIZATION

Using AMP for studying the minimum of a loss is now a classical approach that has been used in many context, e.g. [5, 28, 40, 45, 48]. This is also discussed in detail the pedagogical review [17, Section 4.4] and in the lecture notes [25, Section 12.7.4].

The argument goes as follows: consider the convex optimization problem

$$\underset{u \in \mathcal{C}}{\arg\min} \sum_{\mu=1}^N \ell(u \cdot a, u^* \cdot a) + R(u), \tag{26}$$

with ℓ, R convex in their first argument, and $\mathcal{C} \subseteq \mathbb{R}^D$ a convex set. Define the functions

$$\bar{g}(r, s, \bar{b}) = \arg\min_{z \in \mathbb{R}} \left[\ell(z, s) + \frac{1}{2\bar{b}}(z - r)^2 \right] \quad \text{and} \quad g(r, s, \bar{b}) = \frac{\bar{g}_t(r, s) - r}{\bar{b}},$$

$$e(\boldsymbol{r}, \bar{d}) = \arg\min_{\boldsymbol{z} \in \mathcal{C}} \left[R(\boldsymbol{z}) + \frac{\bar{d}}{2} ||\boldsymbol{z} - \bar{d}^{-1}\boldsymbol{r}||^2 \right] . \tag{27}$$

Consider the GAMP algorithm

$$\boldsymbol{u}^{t+1} = A^T g(\boldsymbol{v}^t, \boldsymbol{y}, b^t) + d_t e(\boldsymbol{u}^t, d^t),$$
$$\boldsymbol{v}^t = Ae(\boldsymbol{u}^t, d^t) - b_t g(\boldsymbol{v}^{t-1}, \boldsymbol{y}, b^{t-1}), \tag{28}$$

where g is applied component-wise to $\boldsymbol{v}$ and $\boldsymbol{y}$. Then, [17, Proposition 4.4] guarantees that *fixed points of* (28) *are solutions of* (26) (by a minor adaptation, as they consider separable regularization, but all their arguments generalize directly). As long as ℓ, R are convex in their first argument, state evolution follows automatically for this choices of g and e.

Additionally, we remark that if the loss is *strictly convex*, then the optimization problem (26) has only a single global minimum, so that AMP will have a single fixed point coinciding with this global minimum.

B.3.2. STATE EVOLUTION

Thanks to the conditional Gaussianity discussed in Appendix B.3, one can track the statistics of the iterated $\boldsymbol{u}, \boldsymbol{v}$ of GAMP through a set of so-called *state evolution equations*. The main difference with respect to previous approaches in our case is that the denoising function e is not separable. However, [7, 19] guarantees that the state evolution that allows to track the performance of AMP remains corrects (under regularity assumptions that are automatically satisfied by e and g being proximal operators).

The state evolution associated to the GAMP

$$\boldsymbol{u}^{t+1} = A^T g_t(\boldsymbol{v}^t, \boldsymbol{y}, b^t) + d_t e_t(\boldsymbol{u}^t, d^t),$$
$$\boldsymbol{v}^t = Ae_t(\boldsymbol{u}^t, d^t) - b_t g_{t-1}(\boldsymbol{v}^{t-1}, \boldsymbol{y}, b^{t-1}) \tag{29}$$

(where again here we consider g applied component-wise to $\boldsymbol{v}$ and $\boldsymbol{y}$) is derived by considering that $y^\mu = \phi(\boldsymbol{u}^* \cdot \boldsymbol{a}^\mu)$ where ϕ is a possibly random non-linearity, and taking into account the evolution of the iterates $\boldsymbol{u}$ and $\boldsymbol{v}$ separately in the parallel and orthogonal directions to $\boldsymbol{u}^*$ (and similarly with $\boldsymbol{v}$ and $\boldsymbol{v}^* = A^T \boldsymbol{u}^*$). Using [7, 19] we have:

Theorem 4 (Theorem 1 in [7], informal) *Define*

$$\begin{cases} m_u^{t+1} &= \frac{N}{D}\mathbb{E}_u^t[\partial_{z^*} g(\omega, \phi(z^*), b^t)] = \frac{N}{D}\mathbb{E}_u^t[z^* g(\omega, \phi(z^*), b^t)] \\ q_u^{t+1} &= \frac{N}{D}\mathbb{E}_u^t[(g(\omega, \phi(z^*), b^t)^2] \\ d^{t+1} &= -\frac{N}{D}\mathbb{E}_u^t \partial_\omega g(\omega, \phi(z^*), b^t) \\ m_v^t &= \frac{1}{D}\mathbb{E}_v^t \left[\left\langle \boldsymbol{u}^* ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t\boldsymbol{u}^*, d^t) \right\rangle \right] \\ q_v^t &= \frac{1}{D}\mathbb{E}_v^t \left[\left\langle e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t\boldsymbol{u}^*, d^t) ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t\boldsymbol{u}^*, d^t) \right\rangle \right] \\ b^t &= \frac{1}{D}\mathbb{E}_v^t \nabla \cdot e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t\boldsymbol{u}^*, d^t) \end{cases} \tag{30}$$

18

where the average $\mathbb{E}_u^t$ is over the randomness in ϕ as well as over $(\omega, z^*) \in \mathbb{R}^2$ distributed as

$$\begin{bmatrix} z^* \\ \omega \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \|\boldsymbol{u}^*\|^2/D & m^t \\ m^t & q^t \end{bmatrix} \right), \tag{31}$$

and $\mathbb{E}_v^t$ is over $\boldsymbol{z}$ i.i.d. Gaussian $\mathcal{N}(0, \mathbb{I}_D)$.

Then the AMP vectors $\boldsymbol{u}^{t+1}$ and $\boldsymbol{v}^t$ converges weakly to their Gaussian version: $\mathbf{U}^t = m_u^t \boldsymbol{u}^* + \sqrt{q_u^t}\boldsymbol{z}$ (with $\boldsymbol{z}$ random Gaussian as above) and $\mathbf{V}^t = \sqrt{q_v^t - m_v^t}\boldsymbol{w} + m_v^t A^T \boldsymbol{u}^*$ (with $\boldsymbol{w} \sim \mathcal{N}(0, \mathbb{I}_N)$), in the sense that, for any deterministic sequence $\phi_n : (\mathbb{R}^D \times \mathbb{R}^N)^t \times \mathbb{R}^N \to \mathbb{R}$ of uniformly pseudo-Lipschitz functions of order k,

$$\phi_n(\boldsymbol{u^0}, \boldsymbol{v^0}, \boldsymbol{u^1}, \boldsymbol{v^1}, \dots, \boldsymbol{v^{t-1}}, \boldsymbol{u^t}) \overset{\mathbb{P}}{\simeq} \mathbb{E}_{\mathbf{U},\mathbf{V}}\left[\phi_{\mathbf{n}}\left(\mathbf{U^0}, \mathbf{V^0}, \mathbf{U^1}, \mathbf{V^1}, \dots, \mathbf{V^{t-1}}, \mathbf{U^t} \right) \right]. \tag{32}$$

An immediate corollary we shall use is the following:

Corollary 2 (Fixed point initialization) *Consider a fixed point of the state evolution equations (30) (we denote the fixed point quantities by* fp*). If one initializes an AMP sequence in the fixed point, i.e. by using $\mathbf{u^0} = m_u^{\mathrm{fp}} \boldsymbol{u}^* + \sqrt{q_u^{\mathrm{fp}}}\boldsymbol{z}$ (with $\boldsymbol{z}$ random Gaussian as above), then in probability for any time $t > 0$ and as $d \to \infty$, the sequence of iterates $\boldsymbol{U}^t$ and $\boldsymbol{V}^t$ remains in their fixed point.*

We then use the following theorem, that ensure with the initialization of 2, then the GAMP equation converges to their fixed point:

Theorem 5 (Convergence of GAMP, Lemma 7 from [28]) *Consider the GAMP iteration with e, g as in (27), where all free parameters are initialized at any fixed point of the state evolution equations. If the associated loss (26) is strongly convex, then GAMP converges to the fixed point of the loss under study:*

$$\lim_{t\to\infty} \lim_{d\to\infty} \frac{1}{\sqrt{d}}\|\boldsymbol{u}^t - \boldsymbol{u}^{\mathrm{fp}}\|_F = 0, \quad \lim_{t\to\infty} \lim_{d\to\infty} \frac{1}{\sqrt{d}}\|\boldsymbol{v}^t - \boldsymbol{v}^{\mathrm{fp}}\|_F = 0 \tag{33}$$

Thus, state evolution allows to compute all scalar observables (such as overlaps, errors, etc) on the iterates, which at convergence and under the setting of Appendix B.3.1, are equivalent to the global minimum of (26).

B.4. State evolution for Theorem 1

Now that we have reviewed how to use GAMP to study the fixed point of an optimization problem, it remains to show that the present matrix problem can be mapped to such an equivalent vectorized problem. Consider (9), and write it in the following form:

$$\hat{S} = \underset{S \in \mathcal{C}}{\arg\min} \, \tilde{\mathcal{L}}(S), \quad \text{with } \tilde{\mathcal{L}}(S) := \sum_{\mu=1}^n \ell\left(\mathrm{Tr}[X^\mu S], \mathrm{Tr}[X^\mu S^*]\right) + R(S), \tag{34}$$

where $S \in \mathrm{Sym}_d$ is a symmetric $d \times d$ matrix, $\mathcal{C} \subseteq \mathbb{R}^D$ is convex and ℓ, R are convex in their first argument.

B.4.1. MAPPING MATRIX-GLM TO VECTOR-GLM

We consider the mapping from vec : $\mathrm{Sym}_d \to \mathbb{R}^{d(d+1)/2}$ (which conveniently maps the Frobenius scalar product in Sym_d given by $\langle A; B\rangle = \mathrm{Tr}(AB)$ to the standard Euclidean scalar product in $\mathbb{R}^{d(d+1)/2}$) given by

$$\mathrm{vec}\,(A)_{(ab)} = \left\langle b^{(ab)}; A \right\rangle = \sqrt{2 - \delta_{ab}} A_{ab}, \tag{35}$$

under the choice of orthonormal basis

$$b_{ij}^{(aa)} = \delta_{ia}\delta_{ja}, \quad b_{ij}^{(ab)} = \frac{\delta_{ia}\delta_{jb} + \delta_{ib}\delta_{ja}}{\sqrt{2}}. \tag{36}$$

Here (ab) stands for the ordered pair of $1 \le a \le b \le d$, and we denote A_{ij} as the i, j entry of a matrix A, while as $A_{(ab)}$ the component of matrix A onto the basis element $b^{(ab)}$. Let us denote $d(d+1)/2 = D$ (we will often use $D \approx d^2/2$ as we will be interested in the leading order in d).

Under this mapping, we have that $\boldsymbol{w} = \sqrt{d/2}\mathrm{vec}\,(S)$ satisfies

$$||\boldsymbol{w}||^2 = \frac{d}{2}\,\mathrm{Tr}[S^2] = \mathcal{O}(D), \tag{37}$$

for any S with asymptotically well-defined spectral density. In particular in state evolution $\boldsymbol{u}^* = \mathrm{vec}\,(S^*)$ is such that $||\boldsymbol{u}^*||^2/D = Q^*$. Moreover, we can define the correctly normalized sensing vectors

$$A_{(ab)}^\mu := \sqrt{\frac{1}{d+1}}\mathrm{vec}\,(X^\mu)_{(ab)} = \sqrt{\frac{d(2 - \delta_{ab})}{2D}} X_{ab}^\mu \sim \mathcal{N}(0, 1/D), \tag{38}$$

for all $1 \le a \le b \le d$, where we used that $X \sim \mathrm{GOE}(d)$, giving

$$\mathrm{Tr}[X^\mu S] = \sum_{(ab)} \mathrm{vec}\,(X^\mu)_{(ab)}\mathrm{vec}\,(S)_{(ab)} = \sqrt{2\frac{d+1}{d}}\sum_{(ab)} A_{(ab)}^\mu \boldsymbol{w}_{(ab)} \approx \sqrt{2}\sum_{(ab)} A_{(ab)}^\mu \boldsymbol{w}_{(ab)}. \tag{39}$$

where $\alpha = n/d^2$.

Thus, we can pick

$$\tilde{\ell}(a, b) = \ell(\sqrt{2}a, \sqrt{2}b), \qquad \tilde{R}(\boldsymbol{w}) = R(\mathrm{mat}\,(\boldsymbol{w})/\sqrt{d/2}) \quad \text{and} \quad \phi(z) = z + \sqrt{\Delta/2}\,\xi \tag{40}$$

where mat () denotes the inverse of vec (), ξ is a standard Gaussian (we are restricting here to the Gaussian noise function ϕ), and similarly we define $\tilde{\mathcal{C}}$ by a rescaling of $\mathcal{C}$. Finally, we define the equivalent vector problem in dimension $D = d(d+1)/2$ with $N = n$ samples

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \tilde{\mathcal{C}}} \tilde{\mathcal{L}}_{\mathrm{vec}}(\boldsymbol{w}), \quad \text{with} \quad \tilde{\mathcal{L}}_{\mathrm{vec}}(\boldsymbol{w}) := \sum_{\mu=1}^N \tilde{\ell}\left(\boldsymbol{a}^\mu \cdot \boldsymbol{w}, \phi(\boldsymbol{a}^\mu \cdot \boldsymbol{w}^*)\right) + \tilde{R}(\boldsymbol{w}). \tag{41}$$

In particular, the GAMP iteration described in Appendix B.3 for this loss solves (9) modulo the bijection vec (), with generic convex loss, regularization and constraint set $\mathcal{C}$.

B.4.2. GAMP DENOISER g FOR THE SQUARE LOSS

We directly specialize to the case of the square loss, but all this can be generalized to generic losses, see [27]. For $\ell(a, b) = (a - b)^2$ we have

$$\bar{g}(r, s, \bar{b}) = \arg\min_{z \in \mathbb{R}} \left[\tilde{\ell}(z, s) + \frac{1}{2\bar{b}}(z - r)^2 \right] = \frac{r + 4\bar{b}s}{1 + 4\bar{b}} \quad \text{and} \quad g(r, s, \bar{b}) = \frac{s - r}{\bar{b} + 1/4}. \quad (42)$$

The associated state evolution equations read (recall that $\phi(z^*) = z^* + \sqrt{\Delta/2}\xi$, including the average over the noise ξ in the activation ϕ in $\mathbb{E}_u$)

$$\begin{cases} d^{t+1} & = -\frac{N}{D}\mathbb{E}_u^t \partial_\omega g(\omega, \phi(z^*), b^t) = 2\alpha \frac{1}{b^t + 1/4} \\ m_u^{t+1} & = \frac{N}{D}\mathbb{E}_u^t [\partial_{z^*} g(\omega, \phi(z^*), b^t)] = 2\alpha \frac{1}{b^t + 1/4} \\ q_u^{t+1} & = \frac{N}{D}\mathbb{E}_u^t [(g(\omega, \phi(z^*), b^t)^2] = 2\alpha \frac{Q^* - 2m + q + \Delta/2}{(b^t + 1/4)^2} \end{cases} \quad (43)$$

where we used that $N/D = 2n/d^2 = 2\alpha$.

B.4.3. GAMP DENOISER e FOR SPECTRAL REGULARIZATION, INCLUDING PSD CONSTRAINTS AND NUCLEAR/FROBENIUS REGULARIZATION

Spectral denoising function e. We have (calling $\Gamma = \mathrm{mat}\left(\sqrt{2/d}r\right)$ and $k = \bar{d}$ to avoid confusion with the dimension d)

$$\begin{aligned} e(r, k) &= \arg\min_{z \in \tilde{\mathcal{C}}} \left[\tilde{R}(z) + \frac{k}{2}||z - k^{-1}r||^2 \right] \\ &= \arg\min_{z \in \tilde{\mathcal{C}}} \left[R(\sqrt{2/d}\mathrm{mat}\,(z)) + \frac{k}{2}||z - k^{-1}r||^2 \right] \\ &= \sqrt{\frac{d}{2}}\mathrm{vec}\left(\arg\min_{Z \in \mathcal{C}} \left[R(Z) + \frac{d^2 k}{4}||Z - k^{-1}\Gamma||_F^2 \right] \right) \end{aligned} \quad (44)$$

Now, assume that both $\mathcal{C}$ and R are spectral, meaning that they do not depend on the eigenvectors of Z. Then, if we consider the spectral decompositions $\Gamma = ODO^T$, $e(r, k) = ULU^T$, and we parametrize $Z = VTV^T$, then for each T

$$U = \arg\max_{V \in O(d)} \mathrm{Tr}(VLV^T ODO^T), \quad (45)$$

which, by Von Neumann's trace inequality [38] is given by the rotation V such that $U^T O$ aligns the eigenvalues of T and D in decreasing order, which if we take the convention that all spectral decomposition are given with eigenvalues sorted decreasingly, gives $U = O$. Thus, the problem reduces to a spectral minimization

$$L = \arg\min_{T \in \mathcal{C}} \left[d^{-2}R(T) + \frac{k}{4}\sum_{i=1}^{d} T_i^2 - \frac{1}{2}\sum_{i=1}^{d} T_i D_i \right], \quad (46)$$

where here we are abusing the notation by writing T_i to mean T_{ii} as T is really a diagonal $d \times d$ matrix, and similarly for D. Notice that this expression is still general.

Spectral denoising function e for the case of the main text. We now specialize to the case of the main text. If $\mathcal{C}$ is the set of PSD matrices and $R(T) = d^2 (\tilde\lambda \sum_i T_i + \tau/2 \sum_i T_i^2)$, then

$$L_i = \frac{1}{2\tau + k} \text{ReLU}\left(D_i - 2\tilde\lambda\right). \tag{47}$$

In the end, this means that if

$$\boldsymbol{r} = \text{vec}\left(\sqrt{d/2}\, ODO^T\right) \tag{48}$$

with eigenvalues D_i sorted decreasingly, then

$$e(\boldsymbol{r}, k) = \sqrt{\frac{d}{2}} \text{vec}\left(O \frac{1}{2\tau + k} \text{ReLU}\left(D - 2\tilde\lambda\right) O^T\right) \tag{49}$$

where here ReLU is applied component-wise to the diagonal matrix $D - 2\tilde\lambda \mathbb{I}_d$.

State evolution equations. Now we need to derive explicit expressions for the associated state evolution equations

$$\begin{cases} m_v^t &= \frac{1}{D} \mathbb{E}_v^t \left[\left\langle \mathbf{x_0}\, ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\right\rangle\right] \\ q_v^t &= \frac{1}{D} \mathbb{E}_v^t \left[\left\langle e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\, ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\right\rangle\right] \\ b^t &= \frac{1}{D} \mathbb{E}_v^t \sum_{i=1}^{D} \partial_i e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)_i \end{cases} . \tag{50}$$

We revert to the case of general $\mathcal{C}$ and R for this, assuming that R is strictly convex over $\mathcal{C}$ for simplicity. It is useful to consider the following function

$$\begin{aligned} \Psi(k, q_u, m_u) &= \frac{1}{D} \mathbb{E}_v^t \min_{\boldsymbol{w} \in \tilde{\mathcal{C}}} \left[\tilde{R}(\boldsymbol{w}) + \frac{k}{2}||\boldsymbol{w}||^2 - \sqrt{q_u}\langle \boldsymbol{w}\, ; \boldsymbol{z}\rangle - m_u \langle \boldsymbol{w}\, ; \boldsymbol{u}^*\rangle\right] \\ &= -\frac{1}{D} \lim_{\beta \to +\infty} \frac{1}{\beta} \mathbb{E}_v^t \log \int_{\boldsymbol{w} \in \tilde{\mathcal{C}}} d\boldsymbol{w}\, e^{-\beta\left(\tilde{R}(\boldsymbol{w}) + \frac{k}{2}||\boldsymbol{w}||^2 - \sqrt{q_u}\langle \boldsymbol{w}\, ; \boldsymbol{z}\rangle - m_u \langle \boldsymbol{w}\, ; \boldsymbol{u}^*\rangle\right)} \\ &= -\frac{1}{D} \lim_{\beta \to +\infty} \mathbb{E}_v^t \frac{1}{\beta} \log \mathcal{Z}_\beta(k, q_u, m_u, \boldsymbol{z}, \boldsymbol{x}_0), \end{aligned} \tag{51}$$

where the limiting procedure and all following integration/derivation/limit exchanges are well defined by the strict convexity of R, and $\mathcal{Z}_\beta$ was defined in the last line. Then we have immediately

$$\begin{aligned} -\partial_{m_u} \Psi(k, q_u, m_u) &= \frac{1}{D} \mathbb{E}_v^t \left[\left\langle \boldsymbol{u}^*\, ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\right\rangle\right], \\ \partial_k \Psi(k, q_u, m_u) &= \frac{1}{D} \mathbb{E}_v^t \left[\left\langle e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\, ; e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \mathbf{u}^*, d^t)\right\rangle\right]. \end{aligned} \tag{52}$$

Moreover (dropping the omnipresent β limit and average $\mathbb{E}$, as well as the inputs of $\mathcal{Z}$ for readability)

$$-2\partial_{q_u}\Psi(k, q_u, m_u) =$$

$$= \frac{1}{D}\mathbb{E}_v^t \frac{1}{\sqrt{q_u}} \int_{\boldsymbol{w}\in\tilde{\mathcal{C}}} d\boldsymbol{w} \sum_i w_i z_i \frac{e^{-\beta\left(\tilde{R}(\boldsymbol{w}) + \frac{k}{2}||\boldsymbol{w}||^2 - \sqrt{q_u}\langle\boldsymbol{w};\boldsymbol{z}\rangle - m_u\langle\boldsymbol{w};\boldsymbol{u}^*\rangle\right)}}{\mathcal{Z}_\beta}$$

$$= \frac{1}{D}\mathbb{E}_v^t \frac{1}{\sqrt{q_u}} \int_{\boldsymbol{w}\in\tilde{\mathcal{C}}} d\boldsymbol{w} \sum_i w_i \partial_{z_i} \frac{e^{-\beta\left(\tilde{R}(\boldsymbol{w}) + \frac{k}{2}||\boldsymbol{w}||^2 - \sqrt{q_u}\langle\boldsymbol{w};\boldsymbol{z}\rangle - m_u\langle\boldsymbol{w};\boldsymbol{u}^*\rangle\right)}}{\mathcal{Z}_\beta}$$

$$= \frac{1}{D}\mathbb{E}_v^t \sum_i \partial_{\sqrt{q_u}z_i + m_u u_i^*} \int_{\boldsymbol{w}\in\tilde{\mathcal{C}}} d\boldsymbol{w} w_i \frac{e^{-\beta\left(\tilde{R}(\boldsymbol{w}) + \frac{k}{2}||\boldsymbol{w}||^2 - \sqrt{q_u}\langle\boldsymbol{w};\boldsymbol{z}\rangle - m_u\langle\boldsymbol{w};\boldsymbol{u}^*\rangle\right)}}{\mathcal{Z}_\beta} \qquad (53)$$

$$= \frac{1}{D}\mathbb{E}_v^t \sum_i \partial_{r_i} e(\boldsymbol{r}, k)_i|_{\boldsymbol{r} = \sqrt{q_u}z_i + m_u u_i^*}$$

$$= \frac{1}{D}\mathbb{E}_v^t \sum_{i=1}^{D} \partial_i e(\sqrt{q_u^t}\boldsymbol{z} + m_u^t \boldsymbol{u}^*, d^t)_i$$

where in the third line we used Stein's lemma, so that

$$\begin{cases} m_v^t &= -\partial_{m_u}\Psi(d^t, q_u^t, m_u^t) \\ q_v^t &= \partial_k\Psi(d^t, q_u^t, m_u^t) \\ b^t &= -2\partial_{q_u}\Psi(d^t, q_u^t, m_u^t) \end{cases} . \qquad (54)$$

We just need to compute Ψ for the sake of the state evolutions.

State evolution equations for the case of the main text. If $\mathcal{C}$ is the set of PSD matrices and $R(T) = d^2(\tilde{\lambda}\sum_i T_i + \tau/2\sum_i T_i^2)$, then

$$\Psi(k, q_u, m_u) = 2\mathbb{E}_v^t\left[\sum_i L_i + \tau/2\sum_i L_i^2 + \frac{k}{4}\sum_i L_i^2 - \frac{m_u}{2}\sum_i L_i D_i\right], \qquad (55)$$

where D_i is the i-th eigenvalue of the matrix $S^* + \sqrt{q_u}/m_u Z$, and L_i is determined by

$$L_i = \frac{m_u}{2\tau + k}\text{ReLU}\left(D_i - 2\tilde{\lambda}/m_u\right), \qquad (56)$$

giving

$$\Psi(k, q_u, m_u) = 2\mathbb{E}_v^t\left[\tilde{\lambda}\sum_i L_i + \tau/2\sum_i L_i^2 + \frac{k}{4}\sum_i L_i^2 - \frac{m_u}{2}\sum_i L_i D_i\right]$$

$$= -\frac{m_u^2}{2(k + 2\tau)}\mathbb{E}_v^t \sum_i \text{ReLU}(D_i - 2\tilde{\lambda}/m_u)^2$$

$$\approx -\frac{m_u^2}{2(k + 2\tau)}\int_{2\tilde{\lambda}/m_u} dx\, \mu^*_{\sqrt{q_u}/m_u}(x)(x - 2\tilde{\lambda}/m_u)^2 \qquad (57)$$

$$= -\frac{m_u^2}{2(k + 2\tau)}J(\sqrt{q_u}/m_u, 2\tilde{\lambda}/m_u),$$

where in the last step we used that at leading order the spectrum of $S^* + \sqrt{q_u}/m_u Z$ concentrates (by assumption in 2) to $\mu^*_{\sqrt{q_u}/m_u}$ defined in Theorem 1.

B.4.4. FINAL FORM OF THE STATE EVOLUTION AND OBSERVABLES

Collecting the state evolution equations derived in Appendices B.4.2 and B.4.3, calling $m_v = m$, $q_v = q$, $b = \Sigma$, $m_u =$, $q_u = \hat{q}$ and $d = \hat{\Sigma}$ we get the system of equations

$$
\begin{cases}
\hat{\Sigma}^{t+1} &= \frac{2\alpha}{\Sigma^t + 1/4} \\
{}^{t+1} &= \frac{2\alpha}{\Sigma^t + 1/4} \\
\hat{q}^{t+1} &= 2\alpha \frac{Q^* - 2m + q + \Delta/2}{(\Sigma^t + 1/4)^2} \\
m^t &= \partial_t \frac{(^t)^2}{2(\hat{\Sigma}^t + 2\tau)} J(\sqrt{\hat{q}^t}/^t, 2\tilde{\lambda}/^t) \\
q^t &= \frac{(^t)^2}{2(\hat{\Sigma}^t + 2\tau)^2} J(\sqrt{\hat{q}^t}/^t, 2\tilde{\lambda}/^t) \\
\Sigma^t &= \frac{(^t)^2}{\hat{\Sigma}^t + 2\tau} \partial_{\hat{q}^t} J(\sqrt{\hat{q}^t}/^t, 2\tilde{\lambda}/^t)
\end{cases}
\tag{58}
$$

where J is defined in Theorem 1. This reduces to the equations presented in in Theorem 1 at its fixed point and under the change of variable $\delta = \sqrt{\hat{q}}/$, $\epsilon = 2/$.

The test error then is given by

$$
e_{\text{test}} = Q_0 - 2m + q = 2\alpha\delta^2 - \Delta/2 \,.
\tag{59}
$$

The expression train loss instead van be derived by using the fact that the residuals in the data part of the loss can be computed from state evolution as $\hat{q}/16$, and by evaluating the regularization at the spectral density μ_δ^* (which is just the Tr of a shift and rescaling of this spectral distribution) giving the expression in Theorem 1 in the limit $\tau \to 0^+$.

Notice that these state evolution can be also heuristically applied in the case $\tilde{\lambda} = 0$ as long as the problem has a unique global minimum. This can be checked by having $\Sigma < +\infty$, as Σ is a proxy for the inverse of the curvature of the loss at the global minimum [9], which becomes flat when $\Sigma = +\infty$. Notice also that Σ can be used to guarantee, at the heurisitc level, that the minimum of the global loss is unique even without the need of the regularization $\tau > 0$.

B.5. Replicon condition for Generalized Approximate Message Passing fixed points

For completeness, we discuss in this section the pointwise convergence of the GAMP algorithm when initialized at a fixed point of the state evolution equations. While the convex nature of the problem ensure this convergence [28], we discussed here the explicit criterion that can also be directly checked within the state evolution formalism. As noted by [8], it hinges on a stability condition—known in statistical physics as the *replicon* or de Almeida-Thouless (AT) condition [13, 14]—which can be directly and easily checked from the state evolution equations.

We thus provide the replicon condition on point-wise convergence of GAMP for generic non-separable denoisers (see for instance [48] for the separable case). Consider again the recursion (24)

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= A^T g_t(\boldsymbol{v}^t) + d_t e_t(\boldsymbol{u}^t), \\
\boldsymbol{v}^t &= A e_t(\boldsymbol{u}^t) - b_t g_{t-1}(\boldsymbol{v}^{t-1}),
\end{aligned}
\tag{60}
$$

with the b_t, d_t given in (25). We want to assess linear stability. To this end, we suppose that the iteration is initialized at a fixed point $(\boldsymbol{u}, \boldsymbol{v})$ (and that at this point we also assume that the nonlinearities g, e and the Onsager's terms are constant in time), consider a Gaussian perturbation

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \epsilon \mathbb{I}_d)$ of $\boldsymbol{u}$, and compute whether the iteration converges back to the fixed point after such perturbation. Notice that we can avoid perturbing the Onsager's terms, as in the high-dimensional limit they are independent on the iterates. We have (at first order in ϵ)

$$
\begin{aligned}
\boldsymbol{v}^{\mathrm{new}} &= Ae(\boldsymbol{u} + \boldsymbol{\epsilon}) - bg(\boldsymbol{v}) = \boldsymbol{v} + A(\nabla e(\boldsymbol{u}))\boldsymbol{\epsilon}\,, \\
\boldsymbol{u}^{\mathrm{new}} &= A^T g(\boldsymbol{v} + A(\nabla e(\boldsymbol{u}))\boldsymbol{\epsilon}) + d\, e(\boldsymbol{u} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{u} + A^T \nabla(g(\boldsymbol{v}))A(\nabla e(\boldsymbol{u}))\boldsymbol{\epsilon} - D^{-1}(\nabla \cdot g(\boldsymbol{v}))(\nabla e(\boldsymbol{u}))\boldsymbol{\epsilon}\,, \\
u_k^{\mathrm{new}} &= u_k + \sum_{i,j=1}^{D} \sum_{\mu,\nu=1}^{N} A_{\nu k} A_{\mu j} \partial_\mu g_\nu \partial_i e_j \epsilon_i - \frac{1}{N} \sum_{\mu=1}^{N} \partial_\mu g_\mu \sum_{i=1}^{D} \partial_i e_k \epsilon_i \\
&= u_k + \sum_{i,j=1}^{D} \sum_{\mu,\nu=1}^{N} (A_{\nu k} A_{\mu j} - D^{-1} \delta_{\mu\nu} \delta_{kj}) \partial_\mu g_\nu \partial_i e_j \epsilon_i\,,
\end{aligned}
\tag{61}
$$

where in the second to last step we used the explicit form of the Onsager's term d, and in the last step we suppressed the dependencies of the functions g, e on $\boldsymbol{u}, \boldsymbol{v}$ for clarity and passed in components notation.

The L2 norm of the perturbation $\boldsymbol{u}^{\mathrm{new}} - \boldsymbol{u}$ equals (on average over the initial perturbation $\boldsymbol{\epsilon}$)

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}(u_k^{\mathrm{new}} - u_k)^2 &= \mathbb{E}_{\boldsymbol{\epsilon}} \sum_{i,j=1}^{D} \sum_{\mu,\nu=1}^{N} (A_{\nu k} A_{\mu j} - N^{-1} \delta_{\mu\nu} \delta_{kj}) \partial_\mu g_\nu \partial_i e_j \epsilon_i \\
&\quad \times \sum_{i',j'=1}^{D} \sum_{\mu',\nu'=1}^{N} (A_{\nu' k} A_{\mu' j'} - N^{-1} \delta_{\mu'\nu'} \delta_{kj'}) \partial_{\mu'} g_{\nu'} \partial_{i'} e_{j'} \epsilon_{i'} \\
&= N^{-2} \sum_{i,j,j'=1}^{D} \sum_{\mu,\nu,\mu',\nu'=1}^{N} \delta_{\nu\nu'} \delta_{jj'} \delta_{\mu\mu'} \partial_\mu g_\nu \partial_i e_j \partial_{\mu'} g_{\nu'} \partial_{i'} e_{j'} \\
&\approx \frac{1}{D} \sum_{i,j=1}^{D} (\partial_i e_j)^2 \times \frac{1}{D} \sum_{\mu,\nu=1}^{N} (\partial_\mu g_\nu)^2\,,
\end{aligned}
\tag{62}
$$

where we used that $\mathbb{E}_{\boldsymbol{\epsilon}} \epsilon_i \epsilon_{i'} = \delta_{ii'}$, and the law of large number to substitute

$$
\begin{aligned}
&(A_{\nu k} A_{\mu j} - D^{-1} \delta_{\mu\nu} \delta_{kj})(A_{\nu' k} A_{\mu' j'} - N^{-1} \delta_{\mu'\nu'} \delta_{kj'}) \\
&\approx D^{-2} \delta_{\nu\nu'} \delta_{jj'} \delta_{\mu\mu'}\,,
\end{aligned}
\tag{63}
$$

where in the last step we only kept the term that will contribute to leading order (equality at elading order is denoted by $\approx$. This is the generalization to non-separable functions g, e of [48, Definition 1].

Thus, the linear stability criterion for GAMP is

$$
\frac{N}{D} \times \frac{1}{D} \sum_{i,j=1}^{D} (\partial_i e_j)^2 \times \frac{1}{N} \sum_{\mu,\nu=1}^{N} (\partial_\mu g_\nu)^2 < 1\,.
\tag{64}
$$

In particular, in the case of independent observations we will have $\partial_\mu g_\nu = \delta_{\mu\nu} \partial_\mu g_\mu$, and in the case of separable denoiser one would have $\partial_i e_j = \partial_{ij} \partial_i e_i$. Notice also that this criterion is independent of

the value of the iterates u, v at the fixed points, as it concentrates onto its state-evolution-predicted value.

A fixed point of the state evolution describes a stable fixed point of GAMP if

$$\int_\epsilon dx\, \mu_\delta^*(x) \int_\epsilon dy\, \mu_\delta^*(y) \left(\frac{\mathrm{ReLU}(x - \epsilon) - \mathrm{ReLU}(y - \epsilon)}{x - y} \right)^2 < 2\alpha. \tag{65}$$

We have checked numerically that this condition was satisfied in our fixed point (in accord with Theorem 5).

Appendix C. Minimal norm interpolation limit $\lambda \to 0^+$

Consider (6). We would like to give a prescription to manipulate this equations in order to describe minimal regularization interpolators. We first remark that the limit $\lambda \to 0^+$ of (6) as written describes the case of non-regularized minimization, and thus Theorem 1 will be valid in this case only for $\alpha > \alpha_{\mathrm{interp}}$, when only one interpolator of the training dataset exists.

Consider now the rescaling of (9) given by $\tilde\lambda^{-1}\tilde{\mathcal{L}}$. For any fixed positive $\tilde\lambda$, this rescaling does not alter the solution of the minimization problem. In the limit $\tilde\lambda \to 0^+$ instead, the loss first imposes that the global minimum is an interpolator of the training dataset, and then minimizes the regularization within the interpolator set. Thus, rescaling the loss by $\tilde\lambda^{-1}$ and *then* sending $\tilde\lambda \to 0^+$ will describe minimal regularization interpolators, as long as $\alpha < \alpha_{\mathrm{interp}}$ in the noisy case (after that threshold no interpolator exists, and the minimization problem is ill-defined) of for all α in the noiseless case (as in this case there exists always one interpolator).

We practically achieve the loss rescaling in (6) by renaming $\tilde\lambda\epsilon \to \epsilon$, and obtaining the new system of equations

$$\begin{cases} 4\alpha\delta - \tilde\lambda \frac{\delta}{\epsilon} = \partial_1 J(\delta, \epsilon) \\ Q^* + \frac{\Delta}{2} + 2\alpha\delta^2 - \tilde\lambda \frac{\delta^2}{\epsilon} = (1 - \epsilon\partial_2)J(\delta, \epsilon) \end{cases}, \tag{66}$$

equivalent to (6) for all $\tilde\lambda > 0$, and then taking $\tilde\lambda \to 0^+$.

Thus, to summarize the learning curves at $\tilde\lambda = 0^+$ are found as follows.

- In the noiseless case $\Delta = 0$, we solve (66) with $\tilde\lambda = 0$ for all values of α.

- In the noisy case $\Delta > 0$, we solve (66) with $\tilde\lambda = 0$ for all values of $\alpha < \alpha_{\mathrm{inter}}$, and we solve (6) for $\tilde\lambda = 0$ for all values of $\alpha > \alpha_{\mathrm{inter}}$.

Appendix D. Derivation of Result 1 and Corollary 1

D.1. Prerequisites

For both Result 1 and Corollary 1 we will need the following results on the spectral density $\mu^* \boxplus \mu_{\mathrm{s.c.},\delta}$, where $\boxplus$ is the free convolution and $\mu_{\mathrm{s.c.},\delta} = \sqrt{4 - x^2}/(2\pi\delta^2)$ the semicircle distribution of radius 2δ for $\delta > 0$. Here $\mu^*(x) = \sqrt{\kappa^*}\mu_{\mathrm{M.P.}}(\sqrt{\kappa^*}x)$, where $\mu_{\mathrm{M.P.}}$ is the Marchenko-Pastur distribution [32] with parameter κ^* (the asymptotic spectral distribution of $A^T A/m$ where $A \in \mathbb{R}^{m \times d}$ has i.i.d. zero-mean unit-variance components), and $Q^* = 1 + \kappa^*$.

We start by recalling that

$$J(\delta, s) = \int_s^{+\infty} dx\, \mu_\delta^*(x)\, (x-s)^2 = Q^* + \delta^2 - 2sM^* + s^2 - \int_{-\infty}^s dx\, \mu_\delta^*(x)\, (x-s)^2, \quad (67)$$

where

$$Q^* = \int dx\, \mu^*(x)\, x^2, \quad M^* = \int dx\, \mu^*(x)\, x, \quad (68)$$

and where we used that

$$\int dx\, \mu_\delta^*(x)\, x = \int dx\, \mu^*(x)\, x + \int dx\, \mu_{\text{s.c.},\delta}\, x = M^*$$
$$\int dx\, \mu_\delta^*(x)\, x^2 = \int dx\, \mu^*(x)\, x^2 + \int dx\, \mu_{\text{s.c.},\delta}\, x^2 = Q^* + \delta^2 \quad (69)$$

by additivity of the mean and variance.

Now we use [31, Appendix D.3] to state that for any fixed x, for $\delta \to 0^+$, we have at leading order

$$\delta\sqrt{1-\kappa^*}\, \mu_\delta^*(x\delta\sqrt{1-\kappa^*}) \approx (1-\kappa^*)\mu_{\text{s.c.}}(x). \quad (70)$$

This will allow us to compute the last integral in (67) when $\delta \ll 1$ for some values of s. Indeed, when $\delta \ll 1$ and $0 < \kappa^* < 1$ the spectral density μ_δ^* is composed by a small semicircle around the origin with mass $1 - \kappa^*$, and an approximately μ^*-shaped bulk gapped away from the origin. So as long as s lies in the gap between the two bulks, only this semicircle will contribute to (67). For $\kappa^* > 1$ instead, the bulk at the origin is not present leading to a zero contribution from all the integrals. The case $\kappa^* = 1$ is more delicate, as there the spectral distribution of the target may be non-gapped (for example, in the narrow target case). We recover this case by a limiting procedure from the two sides.

D.2. Derivation of Result 1

To find the interpolation threshold, we consider (6) for $\tilde{\lambda} = 0$. To do this, we need to make sure that we are in the $\alpha > \alpha_{\text{inter}}$ region, where only one interpolator (noiseless case) or no interpolator (noisy case) exists, giving a single minimum of the training loss. This can be checked by having

$$\Sigma = \alpha\epsilon - \frac{1}{4} < +\infty, \quad (71)$$

where Σ was defined in (58). The interpolation threshold will be exactly at the point where $\Sigma \propto \epsilon$ diverges.

Noisy case. Consider the (6) with $\tilde{\lambda} = 0$. We have

$$\begin{cases} 4\alpha\delta - \frac{\delta}{\epsilon} = \partial_1 J(\delta, 0) \\ Q^* + \frac{\Delta}{2} + 2\alpha\delta^2 - \frac{\delta^2}{\epsilon} = J(\delta, 0) \end{cases}. \quad (72)$$

We now consider the limit $\epsilon \to +\infty$, obtaining

$$\begin{cases} 4\alpha\delta = \partial_1 J(\delta, 0) \\ Q^* + \frac{\Delta}{2} + 2\alpha\delta^2 = J(\delta, 0) \end{cases} \implies \begin{cases} \alpha_{\text{inter}} = \frac{1}{4\delta}\partial_1 J(\delta, 0) \\ Q^* + \frac{\Delta}{2} = J(\delta, 0) - \frac{\delta}{2}\partial_1 J(\delta, 0) \end{cases}, \quad (73)$$

i.e. an equation for δ, which can be plugged in to directly find α. This gives the first part of Result 1. Notice that the value of δ found will be finite, giving also an analytic prediction of the height of the cusp of the test error at interpolation.

For $\Delta \to +\infty$, we see that a consistent solution is given by $\delta \to +\infty$. In that case, μ_δ^* is, at leading order, a semicircle with radius 2δ, so that

$$J(\delta, 0) \approx \frac{1}{2}\delta^2 \quad \text{and} \quad \partial_1 J(\delta, 0) \approx \delta, \tag{74}$$

implying that for large Δ the interpolation threshold converges to $1/4$.

Noiseless case. Here we can advance more. In the noiseless case $\Delta \to 0$, we also know that after interpolation we have zero test error, hence $\delta = 0$. Thus, we can expand $J(\delta, 0)$ for small δ. Using (67) and (70) we have

$$
\begin{aligned}
\partial_\delta J(\delta, 0) &= \partial_\delta (Q^* + \delta^2) - \partial_\delta \int_{-\infty}^0 dx\, \mu_\delta^*(x)\, x^2 \\
&= 2\delta - \partial_\delta \left[\delta^2 (1 - \kappa^*) \int_{-\infty}^0 dy\, \delta\sqrt{1 - \kappa^*} \mu_\delta^*(\delta\sqrt{1 - \kappa^*}y)\, y^2 \right] \\
&\approx 2\delta - \partial_\delta \left[\delta^2 (1 - \kappa^*)^2 \int_{-2}^0 dy\, \mu_{\text{s.c.}}(y)\, y^2 \right] \\
&\approx \delta \left[2 - (1 - \kappa^*)^2 \right],
\end{aligned}
\tag{75}
$$

implying that the equation for δ is satisfied for $\delta = 0$, and that

$$\alpha_{\text{inter}} = \frac{1 + 2\kappa_* - \kappa_*^2}{4}. \tag{76}$$

Let us notice that the $\delta = 0$ solution here is valid for all $\alpha > \alpha_{\text{inter}}$, so we could have worked with (6) with $\tilde{\lambda} = 0$ and finite ϵ, and computed ϵ explicitly as a function of α. Doing this shows that $\epsilon < +\infty$ for all $\alpha > \alpha_{\text{inter}}$, and that $\epsilon \to +\infty$ exactly at the threshold. This concludes the derivation of Result 1.

D.3. Derivation of Corollary 1

Consider first the case $0 < \kappa^* < 1$. We start from the stationary conditions for (6)

$$
\begin{cases}
4\alpha\delta - \frac{\delta}{\epsilon} = \partial_\delta J\left(\delta, \tilde{\lambda}\epsilon\right) \\
Q^* + 2\alpha\delta^2 + \frac{\Delta}{2} - \frac{\delta^2}{\epsilon} = \left[(1 - s\partial_s)J(\delta, s)\right]_{s=\epsilon\tilde{\lambda}}
\end{cases}
\tag{77}
$$

To compute the strong recovery, i.e. the value of α at which the test error is zero, we look for the value of α that is consistent with a solution of (6) in the limit $\delta \to 0^+$. We assume the scaling

$$\epsilon = \delta\sqrt{1 - \kappa^*}c \tag{78}$$

and verify it self-consistently. We have that (70) implies at leading order

$$
\partial_\delta \int_{-\infty}^{\tilde{\lambda}\epsilon} dx\, \mu_\delta^*(x)\, (x - \tilde{\lambda}\epsilon)^2
$$

$$
= \partial_\delta \delta^2 (1 - \kappa^*) \int_{-\infty}^{\tilde{\lambda}\epsilon/\delta\sqrt{1-\kappa^*}} dy\, \delta\sqrt{1-\kappa^*}\mu_\delta^*(\delta\sqrt{1-\kappa^*}y) \left(y - \frac{\tilde{\lambda}\epsilon}{\delta\sqrt{1-\kappa^*}} \right)^2
$$

$$
\approx \partial_\delta \delta^2 (1 - \kappa^*)^2 \int_{-2}^{\tilde{\lambda}\epsilon/\delta\sqrt{1-\kappa^*}} dy\, \mu_{\text{s.c.}}(y) \left(y - \frac{\tilde{\lambda}\epsilon}{\delta\sqrt{1-\kappa^*}} \right)^2 \tag{79}
$$

$$
= 2\delta(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(\tilde{\lambda}c) - 2\tilde{\lambda}c M_{\text{s.c.}}^{(1)}(\tilde{\lambda}c) + \tilde{\lambda}^2 c^2 M_{\text{s.c.}}^{(0)}(\tilde{\lambda}c) \right)
$$

$$
+ 2\delta(1 - \kappa^*)^2 \left(\tilde{\lambda}c M_{\text{s.c.}}^{(1)}(\tilde{\lambda}c) - \tilde{\lambda}^2 c^2 M_{\text{s.c.}}^{(0)}(\tilde{\lambda}c) \right)
$$

$$
= 2\delta(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(\tilde{\lambda}c) - \tilde{\lambda}c M_{\text{s.c.}}^{(1)}(\tilde{\lambda}c) \right),
$$

where we defined the incomplete k-th moment of the semi-circle distribution as

$$
M_{\text{s.c.}}^{(k)}(x) = \int_{-2}^{x} dy\, \mu_{\text{s.c.}}(y)\, y^k. \tag{80}
$$

Similarly

$$
\left[(1 - s\partial_s) \int_{-\infty}^{s} dx\, \mu_\delta^*(x)\, (x - s)^2 \right]_{s=\epsilon\tilde{\lambda}} = \int_{-\infty}^{\tilde{\lambda}\epsilon} dx\, \mu_\delta^*(x)\, (x - \tilde{\lambda}\epsilon)^2
$$

$$
+ 2\epsilon\tilde{\lambda} \int_{-\infty}^{\tilde{\lambda}\epsilon} dx\, \mu_\delta^*(x)\, (x - \epsilon\tilde{\lambda})
$$

$$
= \int_{-\infty}^{\tilde{\lambda}\epsilon} dx\, \mu_\delta^*(x)\, (x^2 - \tilde{\lambda}\epsilon^2) \tag{81}
$$

$$
\approx \delta^2 (1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(\tilde{\lambda}c) - \tilde{\lambda}^2 c^2 M_{\text{s.c.}}^{(0)}(\tilde{\lambda}c) \right).
$$

This leads to the equations (at leading order in $\delta \to 0^+$)

$$
\begin{cases}
\alpha = \frac{1}{2} + \frac{\tilde{\lambda}}{4c\sqrt{1-\kappa^*}\delta} - \frac{1}{2}(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(c) - c M_{\text{s.c.}}^{(1)}(c) \right) \\
\alpha = \frac{1}{2} - \frac{\Delta}{4\delta^2} + \frac{\tilde{\lambda}}{2c\sqrt{1-\kappa^*}\delta} - \frac{1}{2}c^2(1 - \kappa^*) - \frac{1}{2}(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(c) - c^2 M_{\text{s.c.}}^{(0)}(c) \right)
\end{cases} \tag{82}
$$

where we renamed $c\tilde{\lambda} \to c$ (which will allow to be in the interpolation limit for $\tilde{\lambda} = 0$ as detailed in Appendix C). Define

$$
\bar{\lambda} = \frac{\tilde{\lambda}}{4\delta\sqrt{1-\kappa^*}} \quad \text{and} \quad \bar{\Delta} = \frac{\Delta}{4\delta^2}, \tag{83}
$$

that is the small regularization limit and noiseless limit at the critical scaling. Then the equations read

$$
\begin{cases}
\frac{\bar{\lambda}}{c} - \bar{\Delta} = \frac{1}{2}(1 - \kappa^*)^2 \left(c M_{\text{s.c.}}^{(1)}(c) - c^2 M_{\text{s.c.}}^{(0)}(c) + \frac{c^2}{1-\kappa^*} \right) \\
\alpha_{\text{strong}} = \frac{1}{2} + \frac{\bar{\lambda}}{c} - \frac{1}{2}(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(c) - c M_{\text{s.c.}}^{(1)}(c) \right)
\end{cases}, \tag{84}
$$

giving one equation for c, and one for α_{strong} given c. By computing the next-to-leading order, one could access the speed at which $\alpha \to \alpha_{\text{strong}}$ in δ, invert it and find the critical scalings of $\tilde{\lambda}$ and Δ as a function of $\alpha_{\text{strong}} - \alpha$. In the case $\tilde{\lambda} = \Delta = 0$ we have the simpler equation

$$
\begin{cases}
c = (1 - \kappa^*) \left(c M_{\text{s.c.}}^{(0)}(c) - M_{\text{s.c.}}^{(1)}(c) \right) \\
\alpha_{\text{strong}} = \frac{1}{2} - \frac{1}{2}(1 - \kappa^*)^2 \left(M_{\text{s.c.}}^{(2)}(c) - c M_{\text{s.c.}}^{(1)}(c) \right)
\end{cases}
\tag{85}
$$

which modulo conversions gives the same α_{strong} as [15, 12].

For $\kappa^* > 1$, the same equations hold with all M contributions set to zero (see discussion above), giving (again for $\tilde{\lambda} = \Delta = 0$)

$$
\alpha_{\text{strong}} = \frac{1}{2} \, .
\tag{86}
$$

This shows Corollary 1.

Appendix E. Derivation of Result 2

We start again from (6)

$$
\begin{cases}
4\alpha\delta - \frac{\delta}{\epsilon} = \partial_\delta J\left(\delta, \tilde{\lambda}\epsilon\right) \\
Q^* + 2\alpha\delta^2 + \frac{\Delta}{2} - \frac{\delta^2}{\epsilon} = J(\delta, \tilde{\lambda}\epsilon) - \epsilon\tilde{\lambda} J^{(0,1)}(\delta, \epsilon\tilde{\lambda}) \, .
\end{cases}
\tag{87}
$$

We redefine $\tilde{\lambda}\epsilon \to \epsilon$ and consider the limit $\kappa_* \to 0$ under the scaling ansatz $\alpha = \bar{\alpha}\kappa_*$, $\delta = d\kappa_*^{-1/2}$, $\epsilon = c\kappa_*^{-1/2}$ and $\tilde{\lambda} = \bar{\lambda}\kappa_*^{1/2}$. Notice that $Q^* = 1 + \kappa^* \approx 1$ in the limit. The equations read

$$
\begin{cases}
4\bar{\alpha}d - \bar{\lambda}\frac{d}{c} = \partial_d J\left(\delta, \epsilon\right) \\
1 + 2\bar{\alpha}d^2 + \frac{\Delta}{2} - \bar{\lambda}\frac{d^2}{c} = (1 - c\partial_c) J(d\kappa_*^{-1/2}, c\kappa_*^{-1/2})
\end{cases}
\tag{88}
$$

Then

$$
\begin{aligned}
J\left(d\kappa_*^{-1/2}, c\kappa_*^{-1/2}\right) &= \int_{c\kappa_*^{-1/2}}^{+\infty} dx\, \mu^*_{d\kappa_*^{-1/2}}(x)\, (x - c\kappa_*^{-1/2})^2 \\
&= \kappa_*^{-1} \int_c^{+\infty} dy\, \kappa_*^{-1/2}\, \mu^*_{d\kappa_*^{-1/2}}(y\kappa_*^{-1/2})\, (y - c)^2 \, .
\end{aligned}
\tag{89}
$$

We now use the fact that at leading order for small κ^*

$$
\nu(y) \approx \kappa_*^{-1/2}\, \mu^*_{d\kappa_*^{-1/2}}(y\kappa_*^{-1/2})
\tag{90}
$$

is given by a semi-circle distribution of radius $2d$ centered at zero with mass $1 - \kappa^*$, plus a Dirac's delta spike at $y = 1 + d^2$ with mass κ^* if $d^2 \leq 1$, see [31, Appendix E.1].

We then have the following cases in which the spike does not contribute.

- No spike $d^2 > 1$ and $c > 2d$, or spike $0 < d^2 < 1$ and $c > 1 + d^2 > 2d$. Then the integral J does not contribute to the equations. Thus

$$
\begin{cases}
4\bar{\alpha}d = \frac{d}{c}\bar{\lambda} \\
1 + 2\bar{\alpha}d^2 + \frac{\Delta}{2} - \frac{d^2}{c}\bar{\lambda} = 0
\end{cases}
\implies
\begin{cases}
c = \frac{\bar{\lambda}}{4\bar{\alpha}} \\
d^2 = \frac{1 + \Delta/2}{2\bar{\alpha}}
\end{cases}
,
\tag{91}
$$

30

under the condition

$$
\begin{cases} d^2 > 1 \iff \bar{\alpha} < \frac{1+\Delta/2}{2} \\ c > 2d \iff \bar{\alpha} < \frac{\bar{\lambda}^2}{32(1+\Delta/2)} \end{cases} \implies \bar{\alpha} < \min\left(\frac{1+\Delta/2}{2}, \frac{\bar{\lambda}^2}{32(1+\Delta/2)}\right), \quad (92)
$$

or

$$
\begin{cases} d^2 \le 1 \iff \bar{\alpha} \ge \frac{1+\Delta/2}{2} \\ c > 1 + d^2 \iff \bar{\alpha} < \frac{\bar{\lambda}-2(1+\Delta/2)}{4} \end{cases} \implies \frac{1+\Delta/2}{2} \le \bar{\alpha} < \frac{\bar{\lambda}-2(1+\Delta/2)}{4}. \quad (93)
$$

- No spike $d^2 > 1$ and bulk integral contribution $c \to (2d)^-$. In this case c tends to $2d$ with the appropriate rate in order to have order $\mathcal{O}(1)$ contribution from the bulk integral in the equations. We have, calling $2d - c = dt$ i.e. $c = d(2-t)$ for some $t > 0$

$$
\kappa_*^{-1} \int_{(2-t)d}^{2d} dy \, \frac{1}{d} \mu_{\text{s.c.}}\left(\frac{y}{d}\right) (y - d(2-t))^2 = d^2 \kappa_*^{-1} \int_{2-t}^{2} dy \, \mu_{\text{s.c.}}(y) (y - 2 + t)^2
$$
$$
= d^2 \kappa_*^{-1} \frac{16 t^{7/2}}{105\pi} + O\left(\kappa_*^{-1} t^{9/2}\right) \quad (94)
$$

giving

$$
J^{\text{bulk}} = \frac{16 d^2 t^{7/2}}{105\pi\kappa_*} \approx 0
$$
$$
\partial_d J^{\text{bulk}} = \frac{16 d t^{5/2}}{15\pi\kappa_*} \approx T \quad (95)
$$
$$
\partial_c J^{\text{bulk}} = -\frac{8 d t^{5/2}}{15\pi\kappa_*} \approx -\frac{T}{2}
$$

from which we see that the bulk integral contributes to non-diverging order if $t^{5/2} = \mathcal{O}(\kappa_*)$, giving the equations

$$
\begin{cases} 4\bar{\alpha}d - \bar{\lambda}\frac{1}{2} = T \\ 1 + 2\bar{\alpha}d^2 + \frac{\Delta}{4} - \bar{\lambda}\frac{d}{2} = dT \end{cases} \implies \begin{cases} T = 2\sqrt{2\bar{\alpha}(1+\Delta/2)} - \frac{1}{2}\bar{\lambda} \\ d^2 = \frac{1+\Delta/2}{2\bar{\alpha}} \end{cases} \quad (96)
$$

under the conditions

$$
\begin{cases} d^2 > 1 \implies \bar{\alpha} < \frac{1+\Delta/2}{2\bar{\alpha}} \\ T > 0 \implies \bar{\alpha} > \frac{\bar{\lambda}^2}{32(1+\Delta/2)} \end{cases} \implies \frac{\bar{\lambda}^2}{32(1+\Delta/2)} < \bar{\alpha} < \frac{1+\Delta/2}{2\bar{\alpha}} \quad (97)
$$

Thus, we start finding that

$$
\text{MSE} = 1 \quad \text{for} \quad \bar{\alpha} < \max\left(\frac{1+\Delta/2}{2}, \frac{\bar{\lambda}-2(1+\Delta/4)}{4}\right). \quad (98)
$$

We now consider the case in which the spike contributes.

31

- If there is the spike $d^2 < 1$ and it is the only contribution $2d < c < 1 + d^2$, then the equations read

$$\begin{cases} 4\bar{\alpha}d - \frac{d}{c}\bar{\lambda} = 4d(1 + d^2 - c) \\ 1 + 2\bar{\alpha}d^2 - \frac{d^2}{c}\bar{\lambda} + \frac{\Delta}{2} = \left(1 + d^2 - c\right)^2 + 2c\left(1 + d^2 - c\right) \end{cases} \tag{99}$$

This system can be solved explicitly under the conditions above, giving quite an unmanageable expression. It reduces to $\delta = 0$ for $\Delta = \bar{\lambda} = 0$.

- If there is the spike $d^2 < 1$ and it contributes along with the bulk (see discussion above and (95)) $c \to (2d)^-$, then the equations read

$$\begin{cases} 4\bar{\alpha}d - \frac{1}{2}\bar{\lambda} = 4d(1 - d)^2 + T \\ 1 + 2\bar{\alpha}d^2 - \frac{d}{2}\bar{\lambda} + \frac{\Delta}{4} = (1 - d)^4 + 4d(1 - d)^2 + dT \end{cases} \tag{100}$$

These give again unmanageable expressions, but for $\Delta = 0$ it gives

$$\text{MSE} = \frac{2}{9}\bar{\alpha}\left(4 - \sqrt{6\bar{\alpha} - 2}\right)^2 . \tag{101}$$

One can check by explicit solution that the last two cases lead to test error strictly smaller than one. The last two cases can be easily solve numerically, allowing to plot the curves in Figure 3.

Finally, the large α behavior can be found by considering the last case

$$\begin{cases} 4\bar{\alpha}d - \frac{1}{2}\bar{\lambda} = 4d(1 - d)^2 + T \\ 1 + 2\bar{\alpha}d^2 - \frac{d}{2}\bar{\lambda} + \frac{\Delta}{4} = (1 - d)^4 + 4d(1 - d)^2 + dT \end{cases} \tag{102}$$

under the scaling ansatz

$$d = \frac{1}{\sqrt{\alpha}}\left(d_0 + \frac{d_1}{\alpha} + \dots\right) T = \sqrt{\alpha}\left(T_0 + \frac{T_1}{\sqrt{\alpha}} + \dots\right) \tag{103}$$

and solving the equations perturbatively gives

$$d_0 = \frac{\sqrt{\Delta}}{2} \quad \text{and} \quad d_1 = \frac{3\sqrt{\Delta}}{4} \tag{104}$$

giving the large α scaling presented in Result 2.

This completes the derivation of Result 2.

Appendix F. Details of the numerical implementation of Theorem 1

All the code for reproducing the experiments and the figures can be found in the public repository https://github.com/SPOC-group/OverparametrisedNet. The numerical implementation of (6) reduces, at its core, to two sub-tasks: the numerical evaluation of μ_a^*, the integration involved in computing $J(a, b)$ and its partial derivatives, and the numerical solution of the equations themselves.

Computing μ_a^*. Recall that the spectral density μ_a^* is the free convolution of a rescaled Marchenko-Pastur μ^* distribution, and that of a semicircle distribution with radius $2a$. In particular, $\mu^*(x) = \sqrt{\kappa^*}\mu_{\text{M.P.}}(\sqrt{\kappa^*}x)$, where $\mu_{\text{M.P.}}$ is the Marchenko-Pastur distribution with parameter κ^* (the asymptotic spectral distribution of $A^T A/m$ where $A \in \mathbb{R}^{m \times d}$ has i.i.d. zero-mean unit-variance components). This free convolution can be computed using standard random matrix theory [44] tools. We report here the procedure as described in [31], and summarize here only the steps without providing a derivation.

To compute the spectral density under consideration, first we fix the noise level a and the location x at which we want to compute $\mu_a^*(x)$. Then, we solve the following cubic equation (looking for the solution with largest imaginary part, if all solutions are real then the spectral density at that location is zero)

$$\left(\frac{1}{\sqrt{\kappa^*}}a\right)g^3 - \left(\frac{z}{\sqrt{\kappa^*}}+a\right)g^2 + \left(z + \frac{1}{\sqrt{\kappa^*}} - \sqrt{\kappa^*}\right)g - 1 = 0\,, \tag{105}$$

where $z = x - i\tau$ with τ a small positive constant. The solution $g = g(z)$ of the equation is the so-called Stieltjes transform of $\mu^*(x)$, i.e.

$$g(z) = \int dx\, \frac{\mu(x)}{x - z}\,, \tag{106}$$

which can be inverted by the Stieltjes–Perron inversion formula, i.e.

$$\mu_a^*(x) = \lim_{\tau \to 0^+} \frac{1}{\pi} \operatorname{Im} g_{\mu_Y}(x - i\tau)\,. \tag{107}$$

In practice, the limit is done by using $\tau = 10^{-12}$ directly in (105).

Additionally, by imposing that the discriminant of (105) is zero we obtain an equation for x (setting $\tau = 0$) that allows to precisely evaluate the boundaries of each of the bulks of μ_a^* (the distribution has two bulks in the low-noise setting, and one single bulk otherwise, with the splitting threshold depending non-trivially on κ^* and a).

Computing $J(a, b)$. We use the estimate of the bulk boundaries to set-up an efficient integration scheme to finally compute $J(a, b)$. We first compute the bulk boundaries, then intersect the support of the bulks with the integration region $(b, +\infty)$, and finally for each bulk integrate $\mu_a^*(x)(x - b)^2$ using `quad` from the Scipy library in Python. This explicit use of the bulk boundaries allows precise evaluation of $J(a, b)$ even in the low-noise regime, where one of the bulks has very small support with quite large values for the spectral density.

Finally, we compute the partial derivatives of $J(a, b)$ by centered finite differences, with step size equal to 10^{-8}.

Numerical solution of the equations. To solve numerically (6), we find it convenient to revert to the state evolution equations presented in Appendix B.4, equation (58). We follow the prescribed iteration, possibly damping the updates with a factor as small as 10^{-2}, and declaring convergence when the Euclidean distance between two subsequent updates falls below a prescribed tolerance of at least 10^{-4}. It is also convenient to adopt a planting scheme, where the equations for a given value of $\tilde{\lambda}, \Delta, \kappa^*$ are solved for one initial value of α (we found that a value around the mid-point between $\alpha = 0$ and $\alpha = \alpha_{\text{strong}}$ works best), and then the rest of the learning curve as a function of α is computed by successively using the solution at a value α as the initialization for that of $\alpha + \delta\alpha$, where $\delta\alpha$ is a small step-size.

Appendix G. Details of the experiments

For Figure 1 right and $d \leq 100$ we used the CVXPY package in Python. All the other experiments are realized in PyTorch. We instantiate the student weights randomly as centered Gaussian variables with standard deviation 10^{-3} and the target ones with variance 1, appropriately changing the functional form of the target. In Figure 1 right we optimize using LBFGS for the sake of efficiency. For Figure 1 left we optimize using the GD iteration

$$
\boldsymbol{w}_k^{t+1} = \boldsymbol{w}_k^t - 2\lambda\eta \frac{\boldsymbol{w}_k^t}{\sqrt{md}} - \eta \sum_{\mu=1}^{n} \nabla_{\boldsymbol{w}_k} \left[y_\mu - \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \sigma_k \left(\frac{\boldsymbol{w}_k^t \cdot \boldsymbol{x}_\mu}{\sqrt{d}} \right) \right]^2 \tag{108}
$$

with $\eta = 20$.