Submitted to the Annals of Applied Probability

A RANDOM MATRIX APPROACH TO NEURAL NETWORKS

By Cosme Louart, Zhenyu Liao, and Romain Couillet*

CentraleSupélec, University of Paris–Saclay, France.

This article studies the Gram random matrix model $G = \frac{1}{T}\Sigma^{\mathsf{T}}\Sigma$, $\Sigma = \sigma(WX)$, classically found in the analysis of random feature maps and random neural networks, where $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p\times T}$ is a (data) matrix of bounded norm, $W \in \mathbb{R}^{n\times p}$ is a matrix of independent zero-mean unit variance entries, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous (activation) function — $\sigma(WX)$ being understood entry-wise. By means of a key concentration of measure lemma arising from non-asymptotic random matrix arguments, we prove that, as n, p, T grow large at the same rate, the resolvent $Q = (G + \gamma I_T)^{-1}$, for $\gamma > 0$, has a similar behavior as that met in sample covariance matrix models, involving notably the moment $\Phi = \frac{T}{n}\mathrm{E}[G]$, which provides in passing a deterministic equivalent for the empirical spectral measure of G. Application-wise, this result enables the estimation of the asymptotic performance of single-layer random neural networks. This in turn provides practical insights into the underlying mechanisms into play in random neural networks, entailing several unexpected consequences, as well as a fast practical means to tune the network hyperparameters.

1. Introduction. Artificial neural networks, developed in the late fifties (Rosenblatt, 1958) in an attempt to develop machines capable of brain-like behaviors, know today an unprecedented research interest, notably in its applications to computer vision and machine learning at large (Krizhevsky, Sutskever and Hinton, 2012; Schmidhuber, 2015) where superhuman performances on specific tasks are now commonly achieved. Recent progress in neural network performances however find their source in the processing power of modern computers as well as in the availability of large datasets rather than in the development of new mathematics. In fact, for lack of appropriate tools to understand the theoretical behavior of the non-linear activations and deterministic data dependence underlying these networks, the discrepancy between mathematical and practical (heuristic) studies of neural networks has kept widening. A first salient problem in harnessing neural networks lies in their being completely designed upon a deterministic training dataset

*Couillet's work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

MSC 2010 subject classifications: Primary 60B20; secondary 62M45

$X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$, so that their resulting performances intricately depend first and foremost on X. Recent works have nonetheless established that, when smartly designed, mere randomly connected neural networks can achieve performances close to those reached by entirely data-driven network designs (Rahimi and Recht, 2007; Saxe et al., 2011). As a matter of fact, to handle gigantic databases, the computationally expensive learning phase (the so-called backpropagation of the error method) typical of deep neural network structures becomes impractical, while it was recently shown that smartly designed single-layer random networks (as studied presently) can already reach superhuman capabilities (Cambria et al., 2015) and beat expert knowledge in specific fields (Jaeger and Haas, 2004). These various findings have opened the road to the study of neural networks by means of statistical and probabilistic tools (Choromanska et al., 2015; Giryes, Sapiro and Bronstein, 2015). The second problem relates to the non-linear activation functions present at each neuron, which have long been known (as opposed to linear activations) to help design universal approximators for any input-output target map (Hornik, Stinchcombe and White, 1989).

In this work, we propose an original random matrix-based approach to understand the end-to-end regression performance of single-layer random artificial neural networks, sometimes referred to as extreme learning machines (Huang, Zhu and Siew, 2006; Huang et al., 2012), when the number T and size p of the input dataset are large and scale proportionally with the number n of neurons in the network. These networks can also be seen, from a more immediate statistical viewpoint, as a mere linear ridge-regressor relating a *random feature map* $\sigma(WX) \in \mathbb{R}^{n \times T}$ of explanatory variables $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$ and target variables $y = [y_1, \ldots, y_T] \in \mathbb{R}^{d \times T}$, for $W \in \mathbb{R}^{n \times p}$ a randomly designed matrix and $\sigma(\cdot)$ a non-linear $\mathbb{R} \to \mathbb{R}$ function (applied component-wise). Our approach has several interesting features both for theoretical and practical considerations. It is first one of the few known attempts to move the random matrix realm away from matrices with independent or linearly dependent entries. Notable exceptions are the line of works surrounding kernel random matrices (El Karoui, 2010; Couillet and Benaych-Georges, 2016) as well as large dimensional robust statistics models (Couillet, Pascal and Silverstein, 2015; El Karoui, 2013; Zhang, Cheng and Singer, 2014). Here, to alleviate the non-linear difficulty, we exploit concentration of measure arguments (Ledoux, 2005) for non-asymptotic random matrices, thereby pushing further the original ideas of (El Karoui, 2009; Vershynin, 2012) established for simpler random matrix models. While we believe that more powerful, albeit more computational intensive, tools (such as an appropriate adaptation of the Gaussian tools advocated in (Pastur

and Ŝerbina, 2011)) cannot be avoided to handle advanced considerations in neural networks, we demonstrate here that the concentration of measure phenomenon allows one to fully characterize the main quantities at the heart of the single-layer regression problem at hand.

In terms of practical applications, our findings shed light on the already incompletely understood extreme learning machines which have proved extremely efficient in handling machine learning problems involving large to huge datasets (Huang et al., 2012; Cambria et al., 2015) at a computationally affordable cost. But our objective is also to pave to path to the understanding of more involved neural network structures, featuring notably multiple layers and some steps of learning by means of backpropagation of the error.

Our main contribution is twofold. From a theoretical perspective, we first obtain a key lemma, Lemma 1, on the concentration of quadratic forms of the type $\sigma(w^\mathsf{T}X)A\sigma(X^\mathsf{T}w)$ where $w = \varphi(\tilde{w})$, $\tilde{w} \sim \mathcal{N}(0, I_p)$, with $\varphi : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ Lipschitz functions, and $X \in \mathbb{R}^{p\times T}$, $A \in \mathbb{R}^{n\times n}$ are deterministic matrices. This *non-asymptotic* result (valid for all n, p, T) is then exploited under a simultaneous growth regime for n, p, T and boundedness conditions on $\|X\|$ and $\|A\|$ to obtain, in Theorem 1, a deterministic approximation $\bar{Q}$ of the resolvent $\mathrm{E}[Q]$, where $Q = (\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T)^{-1}$, $\gamma > 0$, $\Sigma = \sigma(WX)$, for some $W = \varphi(\tilde{W})$, $\tilde{W} \in \mathbb{R}^{n\times p}$ having independent $\mathcal{N}(0, 1)$ entries. As the resolvent of a matrix (or operator) is an important proxy for the characterization of its spectrum (see e.g., (Pastur and Ŝerbina, 2011; Akhiezer and Glazman, 1993)), this result therefore allows for the characterization of the asymptotic spectral properties of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$, such as its limiting spectral measure in Theorem 2.

Application-wise, the theoretical findings are an important preliminary step for the understanding and improvement of various statistical methods based on random features in the large dimensional regime. Specifically, here, we consider the question of linear ridge-regression from random feature maps, which coincides with the aforementioned single hidden-layer random neural network known as extreme learning machine. We show that, under mild conditions, both the *training* E_{train} and *testing* E_{test} mean-square errors, respectively corresponding to the regression errors on known input-output pairs $(x_1, y_1), \ldots, (x_T, y_T)$ (with $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}^d$) and unknown pairings $(\hat{x}_1, \hat{y}_1), \ldots, (\hat{x}_{\hat{T}}, \hat{y}_{\hat{T}})$, almost surely converge to deterministic limiting values as n, p, T grow large at the same rate (while d is kept constant) for every fixed ridge-regression parameter $\gamma > 0$. Simulations on real image datasets are provided that corroborate our results.

These findings provide new insights into the roles played by the activation function $\sigma(\cdot)$ and the random distribution of the entries of W in random feature maps as well as by the ridge-regression parameter γ in the neural network performance. We notably exhibit and prove some peculiar behaviors, such as the impossibility for the network to carry out elementary Gaussian mixture classification tasks, when either the activation function or the random weights distribution are ill chosen.

Besides, for the practitioner, the theoretical formulas retrieved in this work allow for a fast offline tuning of the aforementioned hyperparameters of the neural network, notably when T is not too large compared to p. The graphical results provided in the course of the article were particularly obtained within a 100- to 500-fold gain in computation time between theory and simulations.

The remainder of the article is structured as follows: in Section 2, we introduce the mathematical model of the system under investigation. Our main results are then described and discussed in Section 3, the proofs of which are deferred to Section 5. Section 4 discusses our main findings. The article closes on concluding remarks on envisioned extensions of the present work in Section 6. The appendix provides some intermediary lemmas of constant use throughout the proof section.

Reproducibility: Python 3 codes used to produce the results of Section 4 are available at https://github.com/Zhenyu-LIAO/RMT4ELM

Notations: The norm $\|\cdot\|$ is understood as the Euclidean norm for vectors and the operator norm for matrices, while the norm $\|\cdot\|_F$ is the Frobenius norm for matrices. All vectors in the article are understood as column vectors.

2. System Model. We consider a ridge-regression task on random feature maps defined as follows. Each input data $x \in \mathbb{R}^p$ is multiplied by a matrix $W \in \mathbb{R}^{n\times p}$; a non-linear function $\sigma : \mathbb{R} \to \mathbb{R}$ is then applied entry-wise to the vector Wx, thereby providing a set of n random features $\sigma(Wx) \in \mathbb{R}^n$ for each datum $x \in \mathbb{R}^p$. The output $z \in \mathbb{R}^d$ of the linear regression is the inner product $z = \beta^{\mathsf{T}}\sigma(Wx)$ for some matrix $\beta \in \mathbb{R}^{n\times d}$ to be designed.

From a neural network viewpoint, the n neurons of the network are the virtual units operating the mapping $W_{i\cdot}x \mapsto \sigma(W_{i\cdot}x)$ ($W_{i\cdot}$ being the i-th row of W), for $1 \leq i \leq n$. The neural network then operates in two phases: a training phase where the regression matrix β is learned based on a known

input-output dataset pair (X, Y) and a testing phase where, for β now fixed, the network operates on a new input dataset $\hat{X}$ with corresponding unknown output $\hat{Y}$.

During the training phase, based on a set of known input $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p\times T}$ and output $Y = [y_1, \ldots, y_T] \in \mathbb{R}^{d\times T}$ datasets, the matrix β is chosen so as to minimize the mean square error $\frac{1}{T}\sum_{i=1}^{T} \|z_i - y_i\|^2 + \gamma\|\beta\|_F^2$, where $z_i = \beta^\mathsf{T}\sigma(Wx_i)$ and $\gamma > 0$ is some regularization factor. Solving for β, this leads to the explicit *ridge-regressor*

$$\beta = \frac{1}{T}\Sigma\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T\right)^{-1} Y^\mathsf{T}$$

where we defined $\Sigma \equiv \sigma(WX)$. This follows from differentiating the mean square error along β to obtain $0 = \gamma\beta + \frac{1}{T}\sum_{i=1}^{T} \sigma(Wx_i)(\beta^\mathsf{T}\sigma(Wx_i) - y_i)^\mathsf{T}$, so that $(\frac{1}{T}\Sigma\Sigma^\mathsf{T} + \gamma I_n)\beta = \frac{1}{T}\Sigma Y^\mathsf{T}$ which, along with $(\frac{1}{T}\Sigma\Sigma^\mathsf{T} + \gamma I_n)^{-1}\Sigma = \Sigma(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T)^{-1}$, gives the result.

In the remainder, we will also denote

$$Q \equiv \left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T\right)^{-1}$$

the *resolvent* of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$. The matrix Q naturally appears as a key quantity in the performance analysis of the neural network. Notably, the mean-square error E_{train} on the training dataset X is given by

$$E_{\text{train}} = \frac{1}{T}\left\|Y^\mathsf{T} - \Sigma^\mathsf{T}\beta\right\|_F^2 = \frac{\gamma^2}{T}\operatorname{tr} Y^\mathsf{T} Y Q^2. \tag{1}$$

Under the growth rate assumptions on n, p, T taken below, it shall appear that the random variable E_{train} concentrates around its mean, letting then appear $\mathrm{E}[Q^2]$ as a central object in the asymptotic evaluation of E_{train}.

The testing phase of the neural network is more interesting in practice as it unveils the actual performance of neural networks. For a test dataset $\hat{X} \in \mathbb{R}^{p\times\hat{T}}$ of length $\hat{T}$, with unknown output $\hat{Y} \in \mathbb{R}^{d\times\hat{T}}$, the test mean-square error is defined by

$$E_{\text{test}} = \frac{1}{T}\left\|\hat{Y}^\mathsf{T} - \hat{\Sigma}^\mathsf{T}\beta\right\|_F^2$$

where $\hat{\Sigma} = \sigma(W\hat{X})$ and β is the same as used in (1) (and thus only depends on (X, Y) and γ). One of the key questions in the analysis of such an elementary neural network lies in the determination of γ which minimizes E_{test}

(and is thus said to have good *generalization* performance). Notably, small γ values are known to reduce E_{train} but to induce the popular *overfitting* issue which generally increases E_{test}, while large γ values engender both large values for E_{train} and E_{test}.

From a mathematical standpoint though, the study of E_{test} brings forward some technical difficulties that do not allow for as a simple treatment through the present concentration of measure methodology as the study of E_{train}. Nonetheless, the analysis of E_{train} allows at least for heuristic approaches to become available, which we shall exploit to propose an asymptotic deterministic approximation for E_{test}.

From a technical standpoint, we shall make the following set of assumptions on the mapping $x \mapsto \sigma(Wx)$.

ASSUMPTION 1 (Subgaussian W). *The matrix W is defined by*

$$W = \varphi(\tilde{W})$$

(understood entry-wise), where $\tilde{W}$ has independent and identically distributed $\mathcal{N}(0,1)$ entries and $\varphi(\cdot)$ is λ_φ-Lipschitz.

For $a = \varphi(b) \in \mathbb{R}^\ell$, $\ell \geq 1$, with $b \sim \mathcal{N}(0, I_\ell)$, we shall subsequently denote $a \sim \mathcal{N}_\varphi(0, I_\ell)$.

Under the notations of Assumption 1, we have in particular $W_{ij} \sim \mathcal{N}(0,1)$ if $\varphi(t) = t$ and $W_{ij} \sim \mathcal{U}(-1,1)$ (the uniform distribution on $[-1,1]$) if $\varphi(t) = -1 + 2\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2} dx$ (φ is here a $\sqrt{2/\pi}$-Lipschitz map).

We further need the following regularity condition on the function σ.

ASSUMPTION 2 (Function σ). *The function σ is Lipschitz continuous with parameter λ_σ.*

This assumption holds for many of the activation functions traditionally considered in neural networks, such as sigmoid functions, the rectified linear unit $\sigma(t) = \max(t, 0)$, or the absolute value operator.

When considering the interesting case of simultaneously large data and random features (or neurons), we shall then make the following growth rate assumptions.

ASSUMPTION 3 (Growth Rate). *As* $n \to \infty$,

$$0 < \liminf_n \min\{p/n, T/n\} \leq \limsup_n \max\{p/n, T/n\} < \infty$$

while $\gamma, \lambda_\sigma, \lambda_\varphi > 0$ *and* d *are kept constant. In addition,*

$$\limsup_n \|X\| < \infty$$
$$\limsup_n \max_{ij} |Y_{ij}| < \infty.$$

3. Main Results.

3.1. *Main technical results and training performance.* As a standard preliminary step in the *asymptotic* random matrix analysis of the expectation $\mathrm{E}[Q]$ of the resolvent $Q = (\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T)^{-1}$, a convergence of quadratic forms based on the row vectors of Σ is necessary (see e.g., (Marčenko and Pastur, 1967; Silverstein and Bai, 1995)). Such results are usually obtained by exploiting the independence (or linear dependence) in the vector entries. This not being the case here, as the entries of the vector $\sigma(X^\mathsf{T}w)$ are in general not independent, we resort to a concentration of measure approach, as advocated in (El Karoui, 2009). The following lemma, stated here in a *non-asymptotic* random matrix regime (that is, without necessarily resorting to Assumption 3), and thus of independent interest, provides this concentration result. For this lemma, we need first to define the following key matrix

$$\Phi = \mathrm{E}\left[\sigma(w^\mathsf{T}X)^\mathsf{T}\sigma(w^\mathsf{T}X)\right] \tag{2}$$

of size $T \times T$, where $w \sim \mathcal{N}_\varphi(0, I_p)$.

LEMMA 1 (Concentration of quadratic forms). *Let Assumptions 1–2 hold. Let also* $A \in \mathbb{R}^{T\times T}$ *such that* $\|A\| \leq 1$ *and, for* $X \in \mathbb{R}^{p\times T}$ *and* $w \sim \mathcal{N}_\varphi(0, I_p)$, *define the random vector* $\sigma \equiv \sigma(w^\mathsf{T}X)^\mathsf{T} \in \mathbb{R}^T$. *Then,*

$$P\left(\left|\frac{1}{T}\sigma^\mathsf{T}A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right) \leq Ce^{-\frac{cT}{\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}\min\left(\frac{t^2}{t_0^2}, t\right)}$$

for $t_0 \equiv |\sigma(0)| + \lambda_\varphi\lambda_\sigma\|X\|\sqrt{\frac{p}{T}}$ *and* $C, c > 0$ *independent of all other parameters. In particular, under the additional Assumption 3,*

$$P\left(\left|\frac{1}{T}\sigma^\mathsf{T}A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right) \leq Ce^{-cn\min(t,t^2)}$$

for some $C, c > 0$.

Note that this lemma partially extends concentration of measure results involving quadratic forms, see e.g., (Rudelson et al., 2013, Theorem 1.1), to non-linear vectors.

With this result in place, the standard resolvent approaches of random matrix theory apply, providing our main theoretical finding as follows.

THEOREM 1 (Asymptotic equivalent for $\mathrm{E}[Q]$). *Let Assumptions 1–3 hold and define* $\bar{Q}$ *as*

$$\bar{Q} \equiv \left(\frac{n}{T} \frac{\Phi}{1+\delta} + \gamma I_T \right)^{-1}$$

where δ *is implicitly defined as the unique positive solution to* $\delta = \frac{1}{T} \operatorname{tr} \Phi \bar{Q}$. *Then, for all* $\varepsilon > 0$, *there exists* $c > 0$ *such that*

$$\left\| \mathrm{E}[Q] - \bar{Q} \right\| \leq c n^{-\frac{1}{2}+\varepsilon}.$$

As a corollary of Theorem 1 along with a concentration argument on $\frac{1}{T} \operatorname{tr} Q$, we have the following result on the spectral measure of $\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma$, which may be seen as a non-linear extension of (Silverstein and Bai, 1995) for which $\sigma(t) = t$.

THEOREM 2 (Limiting spectral measure of $\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma$). *Let Assumptions 1–3 hold and, for* $\lambda_1, \ldots, \lambda_T$ *the eigenvalues of* $\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma$, *define* $\mu_n = \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{\delta}_{\lambda_i}$. *Then, for every bounded continuous function* f, *with probability one*

$$\int f d\mu_n - \int f d\bar{\mu}_n \to 0.$$

where $\bar{\mu}_n$ *is the measure defined through its Stieltjes transform* $m_{\bar{\mu}_n}(z) \equiv \int (t-z)^{-1} d\bar{\mu}_n(t)$ *given, for* $z \in \{ w \in \mathbb{C},\ \Im[w] > 0 \}$, *by*

$$m_{\bar{\mu}_n}(z) = \frac{1}{T} \operatorname{tr} \left(\frac{n}{T} \frac{\Phi}{1+\delta_z} - z I_T \right)^{-1}$$

with δ_z *the unique solution in* $\{ w \in \mathbb{C},\ \Im[w] > 0 \}$ *of*

$$\delta_z = \frac{1}{T} \operatorname{tr} \Phi \left(\frac{n}{T} \frac{\Phi}{1+\delta_z} - z I_T \right)^{-1}.$$

Note that $\bar{\mu}_n$ has a well-known form, already met in early random matrix works (e.g., (Silverstein and Bai, 1995)) on sample covariance matrix models. Notably, $\bar{\mu}_n$ is also the deterministic equivalent of the empirical

spectral measure of $\frac{1}{T}P^\mathsf{T}W^\mathsf{T}WP$ for any deterministic matrix $P \in \mathbb{R}^{p\times T}$ such that $P^\mathsf{T}P = \Phi$. As such, to some extent, the results above provide a consistent asymptotic *linearization* of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$. From standard spiked model arguments (see e.g., (Benaych-Georges and Nadakuditi, 2012)), the result $\|\mathrm{E}[Q]-\bar{Q}\| \to 0$ further suggests that also the eigenvectors associated to isolated eigenvalues of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$ (if any) behave similarly to those of $\frac{1}{T}P^\mathsf{T}W^\mathsf{T}WP$, a remark that has fundamental importance in the neural network performance understanding.

However, as shall be shown in Section 3.3, and contrary to empirical covariance matrix models of the type $P^\mathsf{T}W^\mathsf{T}WP$, Φ explicitly depends on the distribution of W_{ij} (that is, beyond its first two moments). Thus, the aforementioned *linearization* of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$, and subsequently the deterministic equivalent for μ_n, are not universal with respect to the distribution of zero-mean unit variance W_{ij}. This is in striking contrast to the many linear random matrix models studied to date which often exhibit such universal behaviors. This property too will have deep consequences in the performance of neural networks as shall be shown through Figure 3 in Section 4 for an example where inappropriate choices for the law of W lead to network failure to fulfill the regression task.

For convenience in the following, letting δ and Φ be defined as in Theorem 1, we shall denote

$$\Psi = \frac{n}{T}\frac{\Phi}{1+\delta}. \tag{3}$$

Theorem 1 provides the central step in the evaluation of E_{train}, for which not only $\mathrm{E}[Q]$ but also $\mathrm{E}[Q^2]$ needs be estimated. This last ingredient is provided in the following proposition.

PROPOSITION 1 (Asymptotic equivalent for $\mathrm{E}[QAQ]$). *Let Assumptions 1–3 hold and $A \in \mathbb{R}^{T\times T}$ be a symmetric non-negative definite matrix which is either Φ or a matrix with uniformly bounded operator norm (with respect to T). Then, for all $\varepsilon > 0$, there exists $c > 0$ such that, for all n,*

$$\left\| \mathrm{E}[QAQ] - \left(\bar{Q}A\bar{Q} + \frac{\frac{1}{n}\operatorname{tr}\left(\Psi\bar{Q}A\bar{Q}\right)}{1-\frac{1}{n}\operatorname{tr}\Psi^2\bar{Q}^2}\bar{Q}\Psi\bar{Q} \right) \right\| \leq cn^{-\frac{1}{2}+\varepsilon}.$$

As an immediate consequence of Proposition 1, we have the following result on the training mean-square error of single-layer random neural networks.

THEOREM 3 (Asymptotic training mean-square error). *Let Assumptions 1–3 hold and $\bar{Q}$, Ψ be defined as in Theorem 1 and (3). Then, for all $\varepsilon > 0$,*

$$n^{\frac{1}{2}-\varepsilon}\left(E_{\text{train}} - \bar{E}_{\text{train}}\right) \to 0$$

almost surely, where

$$E_{\text{train}} = \frac{1}{T}\left\|Y^{\mathsf{T}} - \Sigma^{\mathsf{T}}\beta\right\|_F^2 = \frac{\gamma^2}{T}\operatorname{tr} Y^{\mathsf{T}} Y Q^2$$
$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T}\operatorname{tr} Y^{\mathsf{T}} Y \bar{Q}\left[\frac{\frac{1}{n}\operatorname{tr}\Psi\bar{Q}^2}{1-\frac{1}{n}\operatorname{tr}(\Psi\bar{Q})^2}\Psi + I_T\right]\bar{Q}.$$

Since $\bar{Q}$ and Φ share the same orthogonal eigenvector basis, it appears that E_{train} depends on the alignment between the right singular vectors of Y and the eigenvectors of Φ, with weighting coefficients

$$\left(\frac{\gamma}{\lambda_i+\gamma}\right)^2\left(1+\lambda_i\frac{\frac{1}{n}\sum_{j=1}^T \lambda_j(\lambda_j+\gamma)^{-2}}{1-\frac{1}{n}\sum_{j=1}^T \lambda_j^2(\lambda_j+\gamma)^{-2}}\right),\ 1\le i\le T$$

where we denoted $\lambda_i = \lambda_i(\Psi)$, $1 \le i \le T$, the eigenvalues of Ψ (which depend on γ through $\lambda_i(\Psi) = \frac{n}{T(1+\delta)}\lambda_i(\Phi)$). If $\liminf_n n/T > 1$, it is easily seen that $\delta \to 0$ as $\gamma \to 0$, in which case $E_{\text{train}} \to 0$ almost surely. However, in the more interesting case in practice where $\limsup_n n/T < 1$, $\delta \to \infty$ as $\gamma \to 0$ and E_{train} consequently does not have a simple limit (see Section 4.3 for more discussion on this aspect).

Theorem 3 is also reminiscent of applied random matrix works on empirical covariance matrix models, such as (Bai and Silverstein, 2007; Kammoun et al., 2009), then further emphasizing the strong connection between the non-linear matrix $\sigma(WX)$ and its linear counterpart $W\Phi^{\frac{1}{2}}$.

As a side note, observe that, to obtain Theorem 3, we could have used the fact that $\operatorname{tr} Y^{\mathsf{T}} Y Q^2 = -\frac{\partial}{\partial\gamma}\operatorname{tr} Y^{\mathsf{T}} Y Q$ which, along with some analyticity arguments (for instance when extending the definition of $Q = Q(\gamma)$ to $Q(z)$, $z \in \mathbb{C}$), would have directly ensured that $\frac{\partial\bar{Q}}{\partial\gamma}$ is an asymptotic equivalent for $-\mathrm{E}[Q^2]$, without the need for the explicit derivation of Proposition 1. Nonetheless, as shall appear subsequently, Proposition 1 is also a proxy to the asymptotic analysis of E_{test}. Besides, the technical proof of Proposition 1 quite interestingly showcases the strength of the concentration of measure tools under study here.

3.2. *Testing performance.* As previously mentioned, harnessing the asymptotic testing performance E_{test} seems, to the best of the authors' knowledge, out of current reach with the sole concentration of measure arguments used for the proof of the previous main results. Nonetheless, if not fully effective, these arguments allow for an intuitive derivation of a deterministic equivalent for E_{test}, which is strongly supported by simulation results. We provide this result below under the form of a yet unproven claim, a heuristic derivation of which is provided at the end of Section 5.

To introduce this result, let $\hat{X} = [\hat{x}_1, \ldots, \hat{x}_{\hat{T}}] \in \mathbb{R}^{p \times \hat{T}}$ be a set of input data with corresponding output $\hat{Y} = [\hat{y}_1, \ldots, \hat{y}_{\hat{T}}] \in \mathbb{R}^{d \times \hat{T}}$. We also define $\hat{\Sigma} = \sigma(W\hat{X}) \in \mathbb{R}^{p \times \hat{T}}$. We assume that $\hat{X}$ and $\hat{Y}$ satisfy the same growth rate conditions as X and Y in Assumption 3. To introduce our claim, we need to extend the definition of Φ in (2) and Ψ in (3) to the following notations: for all pair of matrices (A, B) of appropriate dimensions,

$$\Phi_{AB} = \mathrm{E}\left[\sigma(w^\mathsf{T}A)^\mathsf{T}\sigma(w^\mathsf{T}B)\right]$$
$$\Psi_{AB} = \frac{n}{T}\frac{\Phi_{AB}}{1+\delta}$$

where $w \sim \mathcal{N}_\varphi(0, I_p)$. In particular, $\Phi = \Phi_{XX}$ and $\Psi = \Psi_{XX}$.

With these notations in place, we are in position to state our claimed result.

CONJECTURE 1 (Deterministic equivalent for E_{test}). *Let Assumptions 1–2 hold and $\hat{X}, \hat{Y}$ satisfy the same conditions as X, Y in Assumption 3. Then, for all $\varepsilon > 0$,*

$$n^{\frac{1}{2}-\varepsilon}\left(E_{\text{test}} - \bar{E}_{\text{test}}\right) \to 0$$

almost surely, where

$$E_{\text{test}} = \frac{1}{\hat{T}}\left\|\hat{Y}^\mathsf{T} - \hat{\Sigma}^\mathsf{T}\beta\right\|_F^2$$
$$\bar{E}_{\text{test}} = \frac{1}{\hat{T}}\left\|\hat{Y}^\mathsf{T} - \Psi_{X\hat{X}}^\mathsf{T}\bar{Q}Y^\mathsf{T}\right\|_F^2 + \frac{\frac{1}{n}\operatorname{tr} Y^\mathsf{T}Y\bar{Q}\Psi\bar{Q}}{1-\frac{1}{n}\operatorname{tr}(\Psi\bar{Q})^2}\left[\frac{1}{\hat{T}}\operatorname{tr}\Psi_{\hat{X}\hat{X}} - \frac{1}{\hat{T}}\operatorname{tr}(I_T + \gamma\bar{Q})(\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q})\right].$$

While not immediate at first sight, one can confirm (using notably the relation $\Psi\bar{Q} + \gamma\bar{Q} = I_T$) that, for $(\hat{X}, \hat{Y}) = (X, Y)$, $\bar{E}_{\text{train}} = \bar{E}_{\text{test}}$, as expected.

In order to evaluate practically the results of Theorem 3 and Conjecture 1, it is a first step to be capable of estimating the values of Φ_{AB} for various $\sigma(\cdot)$ activation functions of practical interest. Such results, which call for completely different mathematical tools (mostly based on integration tricks), are provided in the subsequent section.

3.3. *Evaluation of* Φ_{AB}. The evaluation of $\Phi_{AB} = \mathrm{E}[\sigma(w^\mathsf{T}A)^\mathsf{T}\sigma(w^\mathsf{T}B)]$ for arbitrary matrices A, B naturally boils down to the evaluation of its individual entries and thus to the calculus, for arbitrary vectors $a, b \in \mathbb{R}^p$, of

$$\Phi_{ab} \equiv \mathrm{E}[\sigma(w^\mathsf{T}a)\sigma(w^\mathsf{T}b)] = (2\pi)^{-\frac{p}{2}} \int \sigma(\varphi(\tilde{w})^\mathsf{T}a)\sigma(\varphi(\tilde{w})^\mathsf{T}b)e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w}. \tag{4}$$

The evaluation of (4) can be obtained through various integration tricks for a wide family of mappings $\varphi(\cdot)$ and activation functions $\sigma(\cdot)$. The most popular activation functions in neural networks are sigmoid functions, such as $\sigma(t) = \mathrm{erf}(t) \equiv \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du$, as well as the so-called rectified linear unit (ReLU) defined by $\sigma(t) = \max(t, 0)$ which has been recently popularized as a result of its robust behavior in deep neural networks. In physical artificial neural networks implemented using light projections, $\sigma(t) = |t|$ is the preferred choice. Note that all aforementioned functions are Lipschitz continuous and therefore in accordance with Assumption 2.

Despite their not abiding by the prescription of Assumptions 1 and 2, we believe that the results of this article could be extended to more general settings, as discussed in Section 4. In particular, since the key ingredient in the proof of all our results is that the vector $\sigma(w^\mathsf{T}X)$ follows a concentration of measure phenomenon, induced by the Gaussianity of $\tilde{w}$ (if $w = \varphi(\tilde{w})$), the Lipschitz character of σ and the norm boundedness of X, it is likely, although not necessarily simple to prove, that $\sigma(w^\mathsf{T}X)$ may still concentrate under relaxed assumptions. This is likely the case for more generic vectors w than $\mathcal{N}_\varphi(0, I_p)$ as well as for a larger class of activation functions, such as polynomial or piece-wise Lipschitz continuous functions.

In anticipation of these likely generalizations, we provide in Table 1 the values of Φ_{ab} for $w \sim \mathcal{N}(0, I_p)$ (i.e., for $\varphi(t) = t$) and for a set of functions $\sigma(\cdot)$ not necessarily satisfying Assumption 2. Denoting $\Phi \equiv \Phi(\sigma(t))$, it is interesting to remark that, since $\arccos(x) = -\arcsin(x) + \frac{\pi}{2}$, $\Phi(\max(t, 0)) = \Phi(\frac{1}{2}t) + \Phi(\frac{1}{2}|t|)$. Also, $[\Phi(\cos(t)) + \Phi(\sin(t))]_{a,b} = \exp(-\frac{1}{2}\|a - b\|^2)$, a result reminiscent of (Rahimi and Recht, 2007).[1] Finally, note that $\Phi(\mathrm{erf}(\kappa t)) \to$

[1]It is in particular not difficult to prove, based on our framework, that, as $n/T \to \infty$, a random neural network composed of $n/2$ neurons with activation function $\sigma(t) = \cos(t)$

$\sigma(t)$	Φ_{ab}
t	$a^\mathsf{T} b$
$\max(t,0)$	$\frac{1}{2\pi}\|a\|\|b\| \left(\angle(a,b) \operatorname{acos}(-\angle(a,b)) + \sqrt{1-\angle(a,b)^2} \right)$
$\|t\|$	$\frac{2}{\pi}\|a\|\|b\| \left(\angle(a,b) \operatorname{asin}(\angle(a,b)) + \sqrt{1-\angle(a,b)^2} \right)$
$\operatorname{erf}(t)$	$\frac{2}{\pi} \operatorname{asin}\left(\frac{2a^\mathsf{T} b}{\sqrt{(1+2\|a\|^2)(1+2\|b\|^2)}} \right)$
$1_{\{t>0\}}$	$\frac{1}{2} - \frac{1}{2\pi} \operatorname{acos}(\angle(a,b))$
$\operatorname{sign}(t)$	$\frac{2}{\pi} \operatorname{asin}(\angle(a,b))$
$\cos(t)$	$\exp(-\frac{1}{2}(\|a\|^2 + \|b\|^2)) \cosh(a^\mathsf{T} b)$
$\sin(t)$	$\exp(-\frac{1}{2}(\|a\|^2 + \|b\|^2)) \sinh(a^\mathsf{T} b)$.

TABLE 1
Values of Φ_{ab} *for* $w \sim \mathcal{N}(0, I_p)$, $\angle(a,b) \equiv \frac{a^\mathsf{T} b}{\|a\|\|b\|}$.

$\Phi(\operatorname{sign}(t))$ as $\kappa \to \infty$, inducing that the extension by continuity of $\operatorname{erf}(\kappa t)$ to $\operatorname{sign}(t)$ propagates to their associated kernels.

In addition to these results for $w \sim \mathcal{N}(0, I_p)$, we also evaluated $\Phi_{ab} = \mathrm{E}[\sigma(w^\mathsf{T} a)\sigma(w^\mathsf{T} b)]$ for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $w \in \mathbb{R}^p$ a vector of independent and identically distributed entries of zero mean and moments of order k equal to m_k (so $m_1 = 0$); w is not restricted here to satisfy $w \sim \mathcal{N}_\varphi(0, I_p)$. In this case, we find

$$\begin{aligned} \Phi_{ab} &= \zeta_2^2 \left[m_2^2 \left(2(a^\mathsf{T} b)^2 + \|a\|^2 \|b\|^2 \right) + (m_4 - 3m_2^2)(a^2)^\mathsf{T} (b^2) \right] + \zeta_1^2 m_2 a^\mathsf{T} b \\ &\quad + \zeta_2 \zeta_1 m_3 \left[(a^2)^\mathsf{T} b + a^\mathsf{T} (b^2) \right] + \zeta_2 \zeta_0 m_2 \left[\|a\|^2 + \|b\|^2 \right] + \zeta_0^2 \end{aligned} \tag{5}$$

where we defined $(a^2) \equiv [a_1^2, \ldots, a_p^2]^\mathsf{T}$.

It is already interesting to remark that, while classical random matrix models exhibit a well-known universality property — in the sense that their limiting spectral distribution is independent of the moments (higher than two) of the entries of the involved random matrix, here W —, for $\sigma(\cdot)$ a polynomial of order two, Φ and thus μ_n strongly depend on $\mathrm{E}[W_{ij}^k]$ for $k = 3, 4$. We shall see in Section 4 that this remark has troubling consequences. We will notably infer (and confirm via simulations) that the studied neural network may provably fail to fulfill a specific task if the W_{ij} are Bernoulli with zero mean and unit variance but succeed with possibly high performance if the W_{ij} are standard Gaussian (which is explained by the disappearance or not of the term $(a^\mathsf{T} b)^2$ and $(a^2)^\mathsf{T}(b^2)$ in (5) if $m_4 = m_2^2$).

and $n/2$ neurons with activation function $\sigma(t) = \sin(t)$ implements a Gaussian difference kernel.

4. Practical Outcomes. We discuss in this section the outcomes of our main results in terms of neural network application. The technical discussions on Theorem 1 and Proposition 1 will be made in the course of their respective proofs in Section 5.

4.1. *Simulation Results.* We first provide in this section a simulation corroborating the findings of Theorem 3 and suggesting the validity of Conjecture 1. To this end, we consider the task of classifying the popular MNIST image database (LeCun, Cortes and Burges, 1998), composed of grayscale handwritten digits of size 28×28, with a neural network composed of $n = 512$ units and standard Gaussian W. We represent here each image as a $p = 784$-size vector; 1 024 images of sevens and 1 024 images of nines were extracted from the database and were evenly split in 512 training and test images, respectively. The database images were jointly centered and scaled so to fall close to the setting of Assumption 3 on X and $\hat{X}$ (an admissible preprocessing intervention). The columns of the output values Y and $\hat{Y}$ were taken as unidimensional ($d = 1$) with $Y_{1j}, \hat{Y}_{1j} \in \{-1, 1\}$ depending on the image class. Figure 1 displays the simulated (averaged over 100 realizations of W) versus theoretical values of E_{train} and E_{test} for three choices of Lipschitz continuous functions $\sigma(\cdot)$, as a function of γ.

Note that a perfect match between theory and practice is observed, for both E_{train} and E_{test}, which is a strong indicator of both the validity of Conjecture 1 and the adequacy of Assumption 3 to the MNIST dataset.

We subsequently provide in Figure 2 the comparison between theoretical formulas and practical simulations for a set of functions $\sigma(\cdot)$ which do not satisfy Assumption 2, i.e., either discontinuous or non-Lipschitz maps. The closeness between both sets of curves is again remarkably good, although to a lesser extent than for the Lipschitz continuous functions of Figure 1. Also, the achieved performances are generally worse than those observed in Figure 1.

It should be noted that the performance estimates provided by Theorem 3 and Conjecture 1 can be efficiently implemented at low computational cost in practice. Indeed, by diagonalizing Φ (which is a marginal cost independent of γ), $\bar{E}_{\text{train}}$ can be computed for all γ through mere vector operations; similarly $\bar{E}_{\text{test}}$ is obtained by the marginal cost of a basis change of $\Phi_{\hat{X}X}$ and the matrix product $\Phi_{X\hat{X}}\Phi_{\hat{X}X}$, all remaining operations being accessible through vector operations. As a consequence, the simulation durations to generate the aforementioned theoretical curves using the linked Python script were

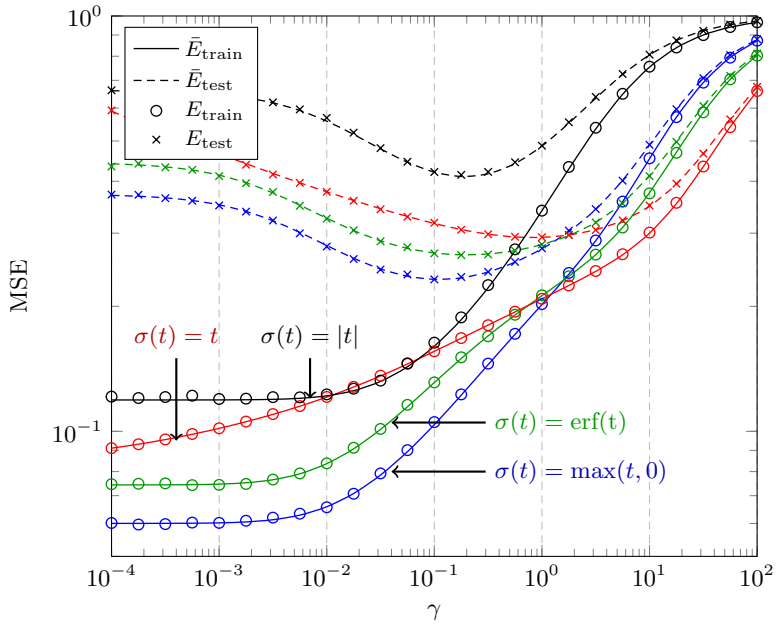

FIG 1. *Neural network performance for Lipschitz continuous $\sigma(\cdot)$, $W_{ij} \sim \mathcal{N}(0, 1)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.*

found to be 100 to 500 times faster than to generate the simulated network performances. Beyond their theoretical interest, the provided formulas therefore allow for an efficient offline tuning of the network hyperparameters, notably the choice of an appropriate value for the ridge-regression parameter γ.

4.2. *The underlying kernel.* Theorem 1 and the subsequent theoretical findings importantly reveal that the neural network performances are directly related to the Gram matrix Φ, which acts as a deterministic kernel on the dataset X. This is in fact a well-known result found e.g., in (Williams, 1998) where it is shown that, as $n \to \infty$ alone, the neural network behaves as a mere kernel operator (this observation is retrieved here in the subsequent Section 4.3). This remark was then put at an advantage in (Rahimi and Recht, 2007) and subsequent works, where random feature maps of the type $x \mapsto \sigma(Wx)$ are proposed as a computationally efficient proxy to evaluate kernels $(x, y) \mapsto \Phi(x, y)$.

As discussed previously, the formulas for $\bar{E}_{\text{train}}$ and $\bar{E}_{\text{test}}$ suggest that good performances are achieved if the dominant eigenvectors of Φ show a good alignment to Y (and similarly for $\Phi_{X\hat{X}}$ and $\hat{Y}$). This naturally drives

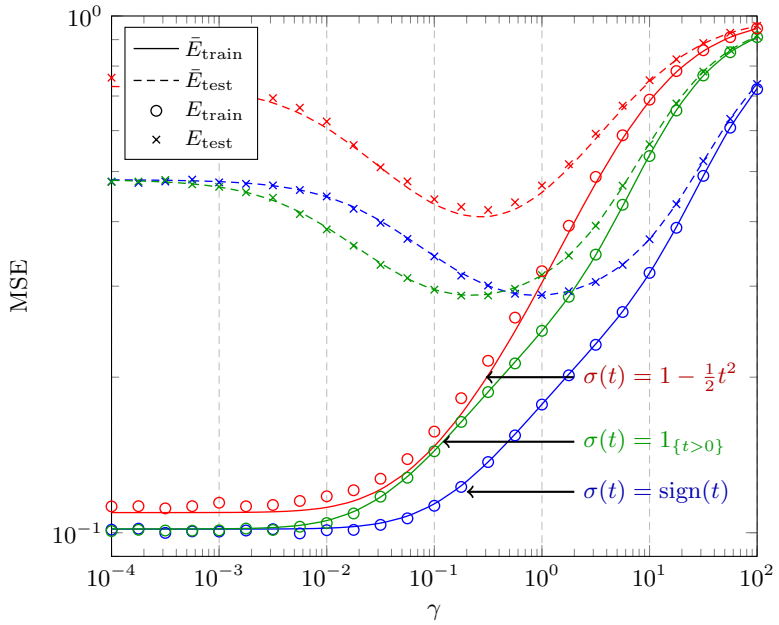

FIG 2. *Neural network performance for $\sigma(\cdot)$ either discontinuous or non Lipschitz, $W_{ij} \sim \mathcal{N}(0,1)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.*

us to finding a priori simple regression tasks where ill-choices of Φ may annihilate the neural network performance. Following recent works on the asymptotic performance analysis of kernel methods for Gaussian mixture models (Couillet and Benaych-Georges, 2016; Zhenyu Liao, 2017; Mai and Couillet, 2017) and (Couillet and Kammoun, 2016), we describe here such a task.

Let $x_1, \ldots, x_{T/2} \sim \mathcal{N}(0, \frac{1}{p}C_1)$ and $x_{T/2+1}, \ldots, x_T \sim \mathcal{N}(0, \frac{1}{p}C_2)$ where C_1 and C_2 are such that $\operatorname{tr} C_1 = \operatorname{tr} C_2$, $\|C_1\|, \|C_2\|$ are bounded, and $\operatorname{tr}(C_1 - C_2)^2 = O(p)$. Accordingly, $y_1, \ldots, y_{T/2+1} = -1$ and $y_{T/2+1}, \ldots, y_T = 1$. It is proved in the aforementioned articles that, under these conditions, it is theoretically possible, in the large p, T limit, to classify the data using a kernel least-square support vector machine (that is, with a training dataset) or with a kernel spectral clustering method (that is, in a completely unsupervised manner) with a non-trivial limiting error probability (i.e., neither zero nor one). This scenario has the interesting feature that $x_i^\mathsf{T} x_j \to 0$ almost surely for all $i \neq j$ while $\|x_i\|^2 - \frac{1}{p}\operatorname{tr}(\frac{1}{2}C_1 + \frac{1}{2}C_2) \to 0$, almost surely, irrespective of the class of x_i, thereby allowing for a Taylor expansion of the non-linear kernels as early proposed in (El Karoui, 2010).

Transposed to our present setting, the aforementioned Taylor expansion allows for a consistent approximation $\tilde{\Phi}$ of Φ by an *information-plus-noise* (spiked) random matrix model (see e.g., (Loubaton and Vallet, 2010; Benaych-Georges and Nadakuditi, 2012)). In the present Gaussian mixture context, it is shown in (Couillet and Benaych-Georges, 2016) that data classification is (asymptotically at least) only possible if $\tilde{\Phi}_{ij}$ explicitly contains the quadratic term $(x_i^\mathsf{T} x_j)^2$ (or combinations of $(x_i^2)^\mathsf{T} x_j$, $(x_j^2)^\mathsf{T} x_i$, and $(x_i^2)^\mathsf{T}(x_j^2)$). In particular, letting $a, b \sim \mathcal{N}(0, C_i)$ with $i = 1, 2$, it is easily seen from Table 1 that only $\max(t, 0)$, $|t|$, and $\cos(t)$ can realize the task. Indeed, we have the following Taylor expansions around $x = 0$:

$$\begin{aligned}
\operatorname{asin}(x) &= x + O(x^3) \\
\sinh(x) &= x + O(x^3) \\
\operatorname{acos}(x) &= \frac{\pi}{2} - x + O(x^3) \\
\cosh(x) &= 1 + \frac{x^2}{2} + O(x^3) \\
x \operatorname{acos}(-x) + \sqrt{1 - x^2} &= 1 + \frac{\pi x}{2} + \frac{x^2}{2} + O(x^3) \\
x \operatorname{asin}(x) + \sqrt{1 - x^2} &= 1 + \frac{x^2}{2} + O(x^3)
\end{aligned}$$

where only the last three functions (only found in the expression of Φ_{ab} corresponding to $\sigma(t) = \max(t, 0)$, $|t|$, or $\cos(t)$) exhibit a quadratic term.

More surprisingly maybe, recalling now Equation (5) which considers non-necessarily Gaussian W_{ij} with moments m_k of order k, a more refined analysis shows that the aforementioned Gaussian mixture classification task will fail if $m_3 = 0$ and $m_4 = m_2^2$, so for instance for $W_{ij} \in \{-1, 1\}$ Bernoulli with parameter $\frac{1}{2}$. The performance comparison of this scenario is shown in the top part of Figure 3 for $\sigma(t) = -\frac{1}{2}t^2 + 1$ and $C_1 = \operatorname{diag}(I_{p/2}, 4I_{p/2})$, $C_2 = \operatorname{diag}(4I_{p/2}, I_{p/2})$, for $W_{ij} \sim \mathcal{N}(0, 1)$ and $W_{ij} \sim \text{Bern}$ (that is, Bernoulli $\{(-1, \frac{1}{2}), (1, \frac{1}{2})\}$). The choice of $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ with $\zeta_1 = 0$ is motivated by (Couillet and Benaych-Georges, 2016; Couillet and Kammoun, 2016) where it is shown, in a somewhat different setting, that this choice is optimal for class recovery. Note that, while the test performances are overall rather weak in this setting, for $W_{ij} \sim \mathcal{N}(0, 1)$, E_{test} drops below one (the amplitude of the $\hat{Y}_{ij}$), thereby indicating that non-trivial classification is performed. This is not so for the Bernoulli $W_{ij} \sim \text{Bern}$ case where E_{test} is systematically greater than $|\hat{Y}_{ij}| = 1$. This is theoretically explained by the fact that, from Equation (5), Φ_{ij} contains structural information about the data classes through the term $2m_2^2(x_i^\mathsf{T} x_j)^2 + (m_4 - 3m_2^2)(x_i^2)^\mathsf{T}(x_j^2)$ which in-

duces an information-plus-noise model for Φ as long as $2m_2^2+(m_4-3m_2^2) \neq 0$, i.e., $m_4 \neq m_2^2$ (see (Couillet and Benaych-Georges, 2016) for details). This is visually seen in the bottom part of Figure 3 where the Gaussian scenario presents an isolated eigenvalue for Φ with corresponding structured eigenvector, which is not the case of the Bernoulli scenario. To complete this discussion, it appears relevant in the present setting to choose W_{ij} in such a way that $m_4-m_2^2$ is far from zero, thus suggesting the interest of heavy-tailed distributions. To confirm this prediction, Figure 3 additionally displays the performance achieved and the spectrum of Φ observed for $W_{ij} \sim$ Stud, that is, following a Student-t distribution with degree of freedom $\nu = 7$ normalized to unit variance (in this case $m_2 = 1$ and $m_4 = 5$). Figure 3 confirms the large superiority of this choice over the Gaussian case (note nonetheless the slight inaccuracy of our theoretical formulas in this case, which is likely due to too small values of p, n, T to accommodate W_{ij} with higher order moments, an observation which is confirmed in simulations when letting ν be even smaller).

4.3. *Limiting cases.* We have suggested that Φ contains, in its dominant eigenmodes, all the usable information describing X. In the Gaussian mixture example above, it was notably shown that Φ may completely fail to contain this information, resulting in the impossibility to perform a classification task, even if one were to take infinitely many neurons in the network. For Φ containing useful information about X, it is intuitive to expect that both $\inf_\gamma \bar{E}_{\text{train}}$ and $\inf_\gamma \bar{E}_{\text{test}}$ become smaller as n/T and n/p become large. It is in fact easy to see that, if Φ is invertible (which is likely to occur in most cases if $\liminf_n T/p > 1$), then

$$\lim_{n\to\infty} \bar{E}_{\text{train}} = 0$$
$$\lim_{n\to\infty} \bar{E}_{\text{test}} - \frac{1}{\hat{T}} \left\| \hat{Y}^\mathsf{T} - \Phi_{\hat{X}X} \Phi^{-1} Y^\mathsf{T} \right\|_F^2 = 0$$

and we fall back on the performance of a classical kernel regression. It is interesting in particular to note that, as the number of neurons n becomes large, the effect of γ on E_{test} flattens out. Therefore, a smart choice of γ is only relevant for small (and thus computationally more efficient) neuron layers. This observation is depicted in Figure 4 where it is made clear that a growth of n reduces E_{train} to zero while E_{test} saturates to a non-zero limit which becomes increasingly irrespective of γ. Note additionally the interesting phenomenon occurring for $n \leq T$ where too small values of γ induce important performance losses, thereby suggesting a strong importance of proper choices of γ in this regime.

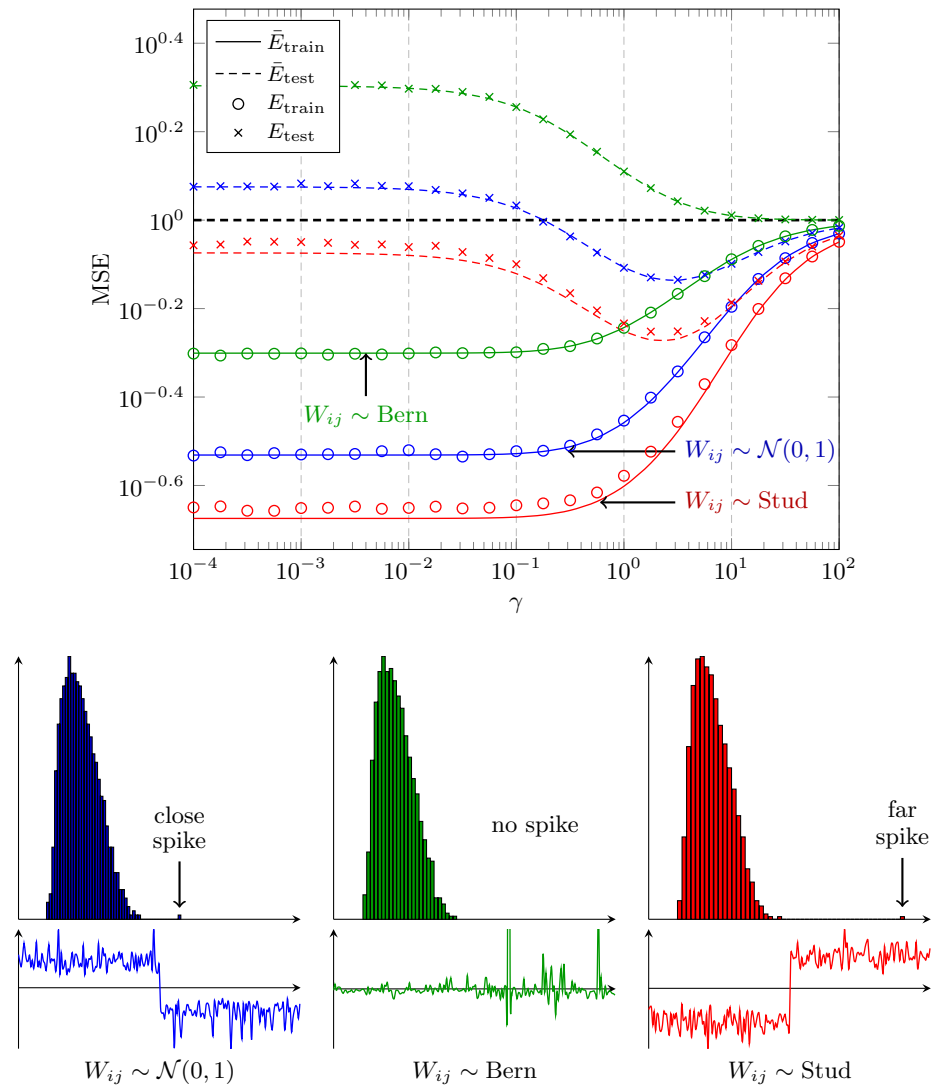

FIG 3. *(Top) Neural network performance for* $\sigma(t) = -\frac{1}{2}t^2 + 1$*, with different* W_{ij}*, for a 2-class Gaussian mixture model (see details in text),* $n = 512$*,* $T = \hat{T} = 1024$*,* $p = 256$*. (Bottom) Spectra and second eigenvector of* Φ *for different* W_{ij} *(first eigenvalues are of order* n *and not shown; associated eigenvectors are provably non informative).*

Of course, practical interest lies precisely in situations where n is not too large. We may thus subsequently assume that $\limsup_n n/T < 1$. In this case, as suggested by Figures 1–2, the mean-square error performances achieved as $\gamma \to 0$ may predict the superiority of specific choices of $\sigma(\cdot)$ for

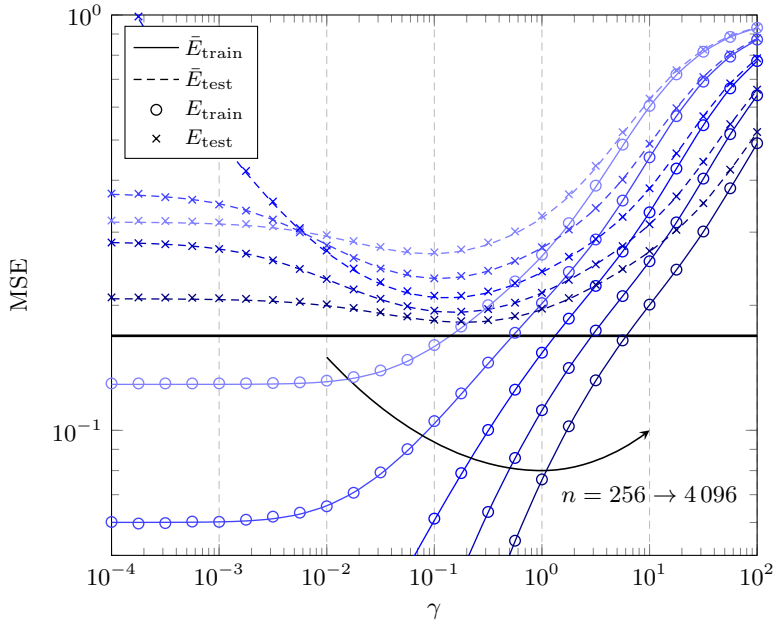

FIG 4. *Neural network performance for growing n (256, 512, 1 024, 2 048, 4 096) as a function of γ, $\sigma(t) = \max(t, 0)$; 2-class MNIST data (sevens, nines), $T = \hat{T} = 1024$, $p = 784$. Limiting ($n = \infty$) $\bar{E}_{\text{test}}$ shown in thick black line.*

optimally chosen γ. It is important for this study to differentiate between cases where $r \equiv \text{rank}(\Phi)$ is smaller or greater than n. Indeed, observe that, with the spectral decomposition $\Phi = U_r \Lambda_r U_r^\mathsf{T}$ for $\Lambda_r \in \mathbb{R}^{r \times r}$ diagonal and $U_r \in \mathbb{R}^{T \times r}$,

$$\delta = \frac{1}{T} \operatorname{tr} \Phi \left(\frac{n}{T} \frac{\Phi}{1+\delta} + \gamma I_T \right)^{-1} = \frac{1}{T} \operatorname{tr} \Lambda_r \left(\frac{n}{T} \frac{\Lambda_r}{1+\delta} + \gamma I_r \right)^{-1}$$

which satisfies, as $\gamma \to 0$,

$$\begin{cases} \delta & \to \frac{r}{n-r} & , \ r < n \\ \gamma\delta & \to \Delta = \frac{1}{T} \operatorname{tr} \Phi \left(\frac{n}{T} \frac{\Phi}{\Delta} + I_T \right)^{-1} & , \ r \geq n. \end{cases}$$

A phase transition therefore exists whereby δ assumes a finite positive value in the small γ limit if $r/n < 1$, or scales like $1/\gamma$ otherwise.

As a consequence, if $r < n$, as $\gamma \to 0$, $\Psi \to \frac{n}{T}(1 - \frac{r}{n})\Phi$ and $\bar{Q} \sim \frac{T}{n-r} U_r \Lambda_r^{-1} U_r^\mathsf{T} + \frac{1}{\gamma} V_r V_r^\mathsf{T}$, where $V_r \in \mathbb{R}^{T \times (n-r)}$ is any matrix such that $[U_r \ V_r]$ is orthogonal, so that $\Psi \bar{Q} \to U_r U_r^\mathsf{T}$ and $\Psi \bar{Q}^2 \to U_r \Lambda_r^{-1} U_r^\mathsf{T}$; and thus,

$\bar{E}_{\text{train}} \to \frac{1}{T} \operatorname{tr} Y V_r V_r^\mathsf{T} Y^\mathsf{T} = \frac{1}{T}\|Y V_r\|_F^2$, which states that the residual training error corresponds to the energy of Y not captured by the space spanned by Φ. Since E_{train} is an increasing function of γ, so is $\bar{E}_{\text{train}}$ (at least for all large n) and thus $\frac{1}{T}\|Y V_r\|_F^2$ corresponds to the lowest achievable asymptotic training error.

If instead $r > n$ (which is the most likely outcome in practice), as $\gamma \to 0$, $\bar{Q} \sim \frac{1}{\gamma}(\frac{n}{T}\frac{\Phi}{\Delta} + I_T)^{-1}$ and thus

$$\bar{E}_{\text{train}} \overset{\gamma \to 0}{\longrightarrow} \frac{1}{T} \operatorname{tr} Y Q_\Delta \left[\frac{\frac{1}{n}\operatorname{tr} \Psi_\Delta Q_\Delta^2}{1 - \frac{1}{n}\operatorname{tr}(\Psi_\Delta Q_\Delta)^2} \Psi_\Delta + I_T \right] Q_\Delta Y^\mathsf{T}$$

where $\Psi_\Delta = \frac{n}{T}\frac{\Phi}{\Delta}$ and $Q_\Delta = (\frac{n}{T}\frac{\Phi}{\Delta} + I_T)^{-1}$.

These results suggest that neural networks should be designed both in a way that reduces the rank of Φ while maintaining a strong alignment between the dominant eigenvectors of Φ and the output matrix Y.

Interestingly, if X is assumed as above to be extracted from a Gaussian mixture and that $Y \in \mathbb{R}^{1\times T}$ is a classification vector with $Y_{1j} \in \{-1, 1\}$, then the tools proposed in (Couillet and Benaych-Georges, 2016) (related to spike random matrix analysis) allow for an explicit evaluation of the aforementioned limits as n, p, T grow large. This analysis is however cumbersome and outside the scope of the present work.

5. Proof of the Main Results. In the remainder, we shall use extensively the following notations:

$$\Sigma = \sigma(WX) = \begin{bmatrix} \sigma_1^\mathsf{T} \\ \vdots \\ \sigma_n^\mathsf{T} \end{bmatrix}, \quad W = \begin{bmatrix} w_1^\mathsf{T} \\ \vdots \\ w_n^\mathsf{T} \end{bmatrix}$$

i.e., $\sigma_i = \sigma(w_i^\mathsf{T} X)^\mathsf{T}$. Also, we shall define $\Sigma_{-i} \in \mathbb{R}^{(n-1)\times T}$ the matrix Σ with i-th row removed, and correspondingly

$$Q_{-i} = \left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma - \frac{1}{T}\sigma_i\sigma_i^\mathsf{T} + \gamma I_T \right)^{-1}.$$

Finally, because of exchangeability, it shall often be convenient to work with the generic random vector $w \sim \mathcal{N}_\varphi(0, I_T)$, the random vector σ distributed as any of the σ_i's, the random matrix Σ_- distributed as any of the Σ_{-i}'s, and with the random matrix Q_- distributed as any of the Q_{-i}'s.

5.1. *Concentration Results on* Σ. Our first results provide concentration of measure properties on functionals of Σ. These results unfold from the following concentration inequality for Lipschitz applications of a Gaussian vector; see e.g., (Ledoux, 2005, Corollary 2.6, Propositions 1.3, 1.8) or (Tao, 2012, Theorem 2.1.12). For $d \in \mathbb{N}$, consider μ the canonical Gaussian probability on $\mathbb{R}^d$ defined through its density $d\mu(w) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|w\|^2}$ and $f : \mathbb{R}^d \to \mathbb{R}$ a λ_f-Lipschitz function. Then, we have the said *normal concentration*

$$\mu\left(\left\{\left|f - \int f d\mu\right| \geq t\right\}\right) \leq C e^{-c\frac{t^2}{\lambda_f^2}} \tag{6}$$

where $C, c > 0$ are independent of d and λ_f. As a corollary (see e.g., (Ledoux, 2005, Proposition 1.10)), for every $k \geq 1$,

$$\mathrm{E}\left[\left|f - \int f d\mu\right|^k\right] \leq \left(\frac{C\lambda_f}{\sqrt{c}}\right)^k.$$

The main approach to the proof of our results, starting with that of the key Lemma 1, is as follows: since $W_{ij} = \varphi(\tilde{W}_{ij})$ with $\tilde{W}_{ij} \sim \mathcal{N}(0,1)$ and φ Lipschitz, the normal concentration of $\tilde{W}$ transfers to W which further induces a normal concentration of the random vector σ and the matrix Σ, thereby implying that Lipschitz functionals of σ or Σ also concentrate. As pointed out earlier, these concentration results are used in place for the independence assumptions (and their multiple consequences on convergence of random variables) classically exploited in random matrix theory.

Notations: In all subsequent lemmas and proofs, the letters $c, c_i, C, C_i > 0$ will be used interchangeably as positive constants independent of the key equation parameters (notably n and t below) and may be reused from line to line. Additionally, the variable $\varepsilon > 0$ will denote any small positive number; the variables c, c_i, C, C_i may depend on ε.

We start by recalling the first part of the statement of Lemma 1 and subsequently providing its proof.

LEMMA 2 (Concentration of quadratic forms). *Let Assumptions 1–2 hold. Let also* $A \in \mathbb{R}^{T\times T}$ *such that* $\|A\| \leq 1$ *and, for* $X \in \mathbb{R}^{p\times T}$ *and*

$w \sim \mathcal{N}_\varphi(0, I_p)$, define the random vector $\sigma \equiv \sigma(w^\mathsf{T} X)^\mathsf{T} \in \mathbb{R}^T$. Then,

$$P\left(\left|\frac{1}{T}\sigma^\mathsf{T} A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right) \leq Ce^{-\frac{cT}{\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}\min\left(\frac{t^2}{t_0^2}, t\right)}$$

for $t_0 \equiv |\sigma(0)| + \lambda_\varphi\lambda_\sigma\|X\|\sqrt{\frac{p}{T}}$ and $C, c > 0$ independent of all other parameters.

Proof. The layout of the proof is as follows: since the application $w \mapsto \frac{1}{T}\sigma^\mathsf{T} A\sigma$ is "quadratic" in w and thus not Lipschitz (therefore not allowing for a natural transfer of the concentration of w to $\frac{1}{T}\sigma^\mathsf{T} A\sigma$), we first prove that $\frac{1}{\sqrt{T}}\|\sigma\|$ satisfies a concentration inequality, which provides a high probability $O(1)$ bound on $\frac{1}{\sqrt{T}}\|\sigma\|$. Conditioning on this event, the map $w \mapsto \frac{1}{\sqrt{T}}\sigma^\mathsf{T} A\sigma$ can then be shown to be Lipschitz (by isolating one of the σ terms for bounding and the other one for retrieving the Lipschitz character) and, up to an appropriate control of concentration results under conditioning, the result is obtained.

Following this plan, we first provide a concentration inequality for $\|\sigma\|$. To this end, note that the application $\psi : \mathbb{R}^p \to \mathbb{R}^T$, $\tilde{w} \mapsto \sigma(\varphi(\tilde{w})^\mathsf{T} X)^\mathsf{T}$ is Lipschitz with parameter $\lambda_\varphi\lambda_\sigma\|X\|$ as the combination of the λ_φ-Lipschitz function $\varphi : \tilde{w} \mapsto w$, the $\|X\|$-Lipschitz map $\mathbb{R}^n \to \mathbb{R}^T$, $w \mapsto X^\mathsf{T} w$ and the λ_σ-Lipschitz map $\mathbb{R}^T \to \mathbb{R}^T$, $Y \mapsto \sigma(Y)$. As a Gaussian vector, $\tilde{w}$ has a normal concentration and so does $\psi(\tilde{w})$. Since the Euclidean norm $\mathbb{R}^T \to \mathbb{R}$, $Y \mapsto \|Y\|$ is 1-Lipschitz, we thus have immediately by (6)

$$P\left(\left|\left\|\frac{1}{\sqrt{T}}\sigma(w^\mathsf{T} X)\right\| - \mathrm{E}\left[\left\|\frac{1}{\sqrt{T}}\sigma(w^\mathsf{T} X)\right\|\right]\right| \geq t\right) \leq Ce^{-\frac{cTt^2}{\|X\|^2\lambda_\sigma^2\lambda_\varphi^2}}$$

for some $c, C > 0$ independent of all parameters.

Finally, using again the Lipschitz character of $\sigma(w^\mathsf{T} X)$,

$$\left|\left\|\sigma(w^\mathsf{T} X)\right\| - \left\|\sigma(0)1_T^\mathsf{T}\right\|\right| \leq \left\|\sigma(w^\mathsf{T} X) - \sigma(0)1_T^\mathsf{T}\right\| \leq \lambda_\sigma \|w\| \cdot \|X\|$$

so that, by Jensen's inequality,

$$\begin{aligned}\mathrm{E}\left[\left\|\frac{1}{\sqrt{T}}\sigma(w^\mathsf{T} X)\right\|\right] &\leq |\sigma(0)| + \lambda_\sigma \mathrm{E}\left[\frac{1}{\sqrt{T}}\|w\|\right]\|X\| \\ &\leq |\sigma(0)| + \lambda_\sigma\sqrt{\mathrm{E}\left[\frac{1}{T}\|w\|^2\right]}\|X\|\end{aligned}$$

with $\mathrm{E}[\|\varphi(\tilde{w})\|^2] \leq \lambda_\varphi^2 \mathrm{E}[\|\tilde{w}\|^2] = p\lambda_\varphi^2$ (since $\tilde{w} \sim \mathcal{N}(0, I_p)$). Letting $t_0 \equiv |\sigma(0)| + \lambda_\sigma \lambda_\varphi \|X\| \sqrt{\frac{p}{T}}$, we then find

$$P\left(\left\| \frac{1}{\sqrt{T}} \sigma(w^\mathsf{T} X) \right\| \geq t + t_0 \right) \leq Ce^{-\frac{cTt^2}{\lambda_\varphi^2 \lambda_\sigma^2 \|X\|^2}}$$

which, with the remark $t \geq 4t_0 \Rightarrow (t - t_0)^2 \geq t^2/2$, may be equivalently stated as

$$\forall t \geq 4t_0,\ P\left(\left\| \frac{1}{\sqrt{T}} \sigma(w^\mathsf{T} X) \right\| \geq t \right) \leq Ce^{-\frac{cTt^2}{2\lambda_\varphi^2 \lambda_\sigma^2 \|X\|^2}}. \tag{7}$$

As a side (but important) remark, note that, since

$$\begin{aligned} P\left(\left\| \frac{\Sigma}{\sqrt{T}} \right\|_F \geq t\sqrt{T} \right) &= P\left(\sqrt{\sum_{i=1}^n \left\| \frac{\sigma_i}{\sqrt{T}} \right\|^2} \geq t\sqrt{T} \right) \\ &\leq P\left(\max_{1 \leq i \leq n} \left\| \frac{\sigma_i}{\sqrt{T}} \right\| \geq \sqrt{\frac{T}{n}} t \right) \\ &\leq nP\left(\left\| \frac{\sigma}{\sqrt{T}} \right\| \geq \sqrt{\frac{T}{n}} t \right) \end{aligned}$$

the result above implies that

$$\forall t \geq 4t_0,\ P\left(\left\| \frac{\Sigma}{\sqrt{T}} \right\|_F \geq t\sqrt{T} \right) \leq Cne^{-\frac{cT^2t^2}{2n\lambda_\varphi^2 \lambda_\sigma^2 \|X\|^2}}$$

and thus, since $\|\cdot\|_F \geq \|\cdot\|$, we have

$$\forall t \geq 4t_0,\ P\left(\left\| \frac{\Sigma}{\sqrt{T}} \right\| \geq t\sqrt{T} \right) \leq Cne^{-\frac{cT^2t^2}{2n\lambda_\varphi^2 \lambda_\sigma^2 \|X\|^2}}$$

Thus, in particular, under the additional Assumption 3, with high probability, the operator norm of $\frac{\Sigma}{\sqrt{T}}$ cannot exceed a rate $\sqrt{T}$.

REMARK 1 (Loss of control of the structure of Σ). *The aforementioned control of $\|\Sigma\|$ arises from the bound $\|\Sigma\| \leq \|\Sigma\|_F$ which may be quite loose (by as much as a factor $\sqrt{T}$). Intuitively, under the supplementary Assumption 3, if $\mathrm{E}[\sigma] \neq 0$, then $\frac{\Sigma}{\sqrt{T}}$ is "dominated" by the matrix $\frac{1}{\sqrt{T}} \mathrm{E}[\sigma] 1_T^\mathsf{T}$, the operator norm of which is indeed of order $\sqrt{n}$ and the bound is tight. If*

$\sigma(t) = t$ and $\mathrm{E}[W_{ij}] = 0$, we however know that $\|\frac{\Sigma}{\sqrt{T}}\| = O(1)$ (Bai and Silverstein, 1998). One is tempted to believe that, more generally, if $\mathrm{E}[\sigma] = 0$, then $\|\frac{\Sigma}{\sqrt{T}}\|$ should remain of this order. And, if instead $\mathrm{E}[\sigma] \neq 0$, the contribution of $\frac{1}{\sqrt{T}}\mathrm{E}[\sigma]1_T^\mathsf{T}$ should merely engender a single large amplitude isolate singular value in the spectrum of $\frac{\Sigma}{\sqrt{T}}$ and the other singular values remain of order $O(1)$. These intuitions are not captured by our concentration of measure approach.

Since $\Sigma = \sigma(WX)$ is an entry-wise operation, concentration results with respect to the Frobenius norm are natural, where with respect to the operator norm are hardly accessible.

Back to our present considerations, let us define the probability space $\mathcal{A}_K = \{w,\ \|\sigma(w^\mathsf{T}X)\| \leq K\sqrt{T}\}$. Conditioning the random variable of interest in Lemma 2 with respect to $\mathcal{A}_K$ and its complementary $\mathcal{A}_K^c$, for some $K \geq 4t_0$, gives

$$P\left(\left|\frac{1}{T}\sigma(w^\mathsf{T}X)A\sigma(w^\mathsf{T}X)^\mathsf{T} - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right)$$
$$\leq P\left(\left\{\left|\frac{1}{T}\sigma(w^\mathsf{T}X)A\sigma(w^\mathsf{T}X)^\mathsf{T} - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right\}, \mathcal{A}_K\right) + P(\mathcal{A}_K^c).$$

We can already bound $P(\mathcal{A}_K^c)$ thanks to (7). As for the first right-hand side term, note that on the set $\{\sigma(w^\mathsf{T}X), w \in \mathcal{A}_K\}$, the function $f : \mathbb{R}^T \to \mathbb{R} :\ \sigma \mapsto \sigma^\mathsf{T}A\sigma$ is $K\sqrt{T}$-Lipschitz. This is because, for all $\sigma, \sigma + h \in \{\sigma(w^\mathsf{T}X), w \in \mathcal{A}_K\}$,

$$\|f(\sigma + h) - f(\sigma)\| = \left\|h^\mathsf{T}A\sigma + (\sigma + h)^\mathsf{T}Ah\right\| \leq K\sqrt{T}\,\|h\|.$$

Since conditioning does not allow for a straightforward application of (6), we consider instead $\tilde{f}$, a $K\sqrt{T}$-Lipschitz continuation to $\mathbb{R}^T$ of $f_{\mathcal{A}_K}$, the restriction of f to $\mathcal{A}_K$, such that all the radial derivative of $\tilde{f}$ are constant in the set $\{\sigma, \|\sigma\| \geq K\sqrt{T}\}$. We may thus now apply (6) and our previous results to obtain

$$P\left(\left|\tilde{f}(\sigma(w^\mathsf{T}X)) - \mathrm{E}[\tilde{f}(\sigma(w^\mathsf{T}X))]\right| \geq KTt\right) \leq e^{-\frac{cTt^2}{\|X\|^2\lambda_\sigma^2\lambda_\varphi^2}}.$$

Therefore,

$$P\left(\left\{\left|f(\sigma(w^\mathsf{T}X)) - \mathrm{E}[\tilde{f}(\sigma(w^\mathsf{T}X))]\right| \geq KTt\right\}, \mathcal{A}_K\right)$$
$$= P\left(\left\{\left|\tilde{f}(\sigma(w^\mathsf{T}X)) - \mathrm{E}[\tilde{f}(\sigma(w^\mathsf{T}X))]\right| \geq KTt\right\}, \mathcal{A}_K\right)$$
$$\leq P\left(\left|\tilde{f}(\sigma(w^\mathsf{T}X)) - \mathrm{E}[\tilde{f}(\sigma(w^\mathsf{T}X))]\right| \geq KTt\right) \leq e^{-\frac{cTt^2}{\|X\|^2\lambda_\sigma^2\lambda_\varphi^2}}.$$

Our next step is then to bound the difference $\Delta = |\mathrm{E}[\tilde{f}(\sigma(w^\mathsf{T}X))] - \mathrm{E}[f(\sigma(w^\mathsf{T}X))]|$. Since f and $\tilde{f}$ are equal on $\{\sigma, \|\sigma\| \leq K\sqrt{T}\}$,

$$\Delta \leq \int_{\|\sigma\|\geq K\sqrt{T}} \left(|f(\sigma)| + |\tilde{f}(\sigma)| \right) d\mu_\sigma(\sigma)$$

where μ_σ is the law of $\sigma(w^\mathsf{T}X)$. Since $\|A\| \leq 1$, for $\|\sigma\| \geq K\sqrt{T}$, $\max(|f(\sigma)|, |\tilde{f}(\sigma)|) \leq \|\sigma\|^2$ and thus

$$\begin{aligned}
\Delta &\leq 2\int_{\|\sigma\|\geq K\sqrt{T}} \|\sigma\|^2 d\mu_\sigma = 2\int_{\|\sigma\|\geq K\sqrt{T}} \int_{t=0}^{\infty} \mathbb{1}_{\|\sigma\|^2\geq t} dt d\mu_\sigma \\
&= 2\int_{t=0}^{\infty} P\left(\left\{\|\sigma\|^2 \geq t\right\}, \mathcal{A}_K^c\right) dt \\
&\leq 2\int_{t=0}^{K^2T} P(\mathcal{A}_K^c) dt + 2\int_{t=K^2T}^{\infty} P(\|\sigma(w^\mathsf{T}X)\|^2 \geq t) dt \\
&\leq 2P(\mathcal{A}_K^c)K^2T + 2\int_{t=K^2T}^{\infty} Ce^{-\frac{ct}{2\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}} dt \\
&\leq 2CTK^2 e^{-\frac{cTK^2}{2\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}} + \frac{2C\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}{c} e^{-\frac{cTK^2}{2\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}} \\
&\leq \frac{6C}{c}\lambda_\varphi^2\lambda_\sigma^2\|X\|^2
\end{aligned}$$

where in last inequality we used the fact that for $x \in \mathbb{R}$, $xe^{-x} \leq e^{-1} \leq 1$, and $K \geq 4t_0 \geq 4\lambda_\sigma\lambda_\varphi\|X\|\sqrt{\frac{p}{T}}$. As a consequence,

$$P\left(\left\{\left|f(\sigma(w^\mathsf{T}X)) - \mathrm{E}[f(\sigma(w^\mathsf{T}X))]\right| \geq KTt + \Delta\right\}, \mathcal{A}_K\right) \leq Ce^{-\frac{cTt^2}{\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}}$$

so that, with the same remark as before, for $t \geq \frac{4\Delta}{KT}$,

$$P\left(\left\{\left|f(\sigma(w^\mathsf{T}X)) - \mathrm{E}[f(\sigma(w^\mathsf{T}X))]\right| \geq KTt\right\}, \mathcal{A}_K\right) \leq Ce^{-\frac{cTt^2}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}}.$$

To avoid the condition $t \geq \frac{4\Delta}{KT}$, we use the fact that, probabilities being lower than one, it suffices to replace C by λC with $\lambda \geq 1$ such that

$$\lambda Ce^{-c\frac{Tt^2}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}} \geq 1 \quad \text{for } t \leq \frac{4\Delta}{KT}.$$

The above inequality holds if we take for instance $\lambda = \frac{1}{C}e^{\frac{18C^2}{c}}$ since then $t \leq \frac{4\Delta}{KT} \leq \frac{24C\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}{cKT} \leq \frac{6C\lambda_\varphi\lambda_\sigma\|X\|}{c\sqrt{pT}}$ (using successively $\Delta \geq \frac{6C}{c}\lambda_\varphi^2\lambda_\sigma^2\|X\|^2$

and $K \geq 4\lambda_\sigma\lambda_\varphi\|X\|\sqrt{\frac{p}{T}})$ and thus

$$\lambda C e^{-\frac{cTt^2}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}} \geq \lambda C e^{-\frac{18C^2}{cp}} \geq \lambda C e^{-\frac{18C^2}{c}} \geq 1.$$

Therefore, setting $\lambda = \max(1, \frac{1}{C}e^{\frac{C'^2c}{2}})$, we get for every $t > 0$

$$P\left(\left\{\left|f(\sigma(w^\mathsf{T}X) - \mathrm{E}[f(\sigma(w^\mathsf{T}X)]\right| \geq KTt\right\}, \mathcal{A}_K\right) \leq \lambda C e^{-\frac{cTt^2}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}}$$

which, together with the inequality $P(\mathcal{A}_K^c) \leq Ce^{-\frac{cTK^2}{2\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}}$, gives

$$P\left(\left|f(\sigma(w^\mathsf{T}X) - \mathrm{E}[f(\sigma(w^\mathsf{T}X)]\right| \geq KTt\right) \leq \lambda C e^{-\frac{Tct^2}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}} + Ce^{-\frac{cTK^2}{2\lambda_\varphi^2\lambda_\sigma^2\|X\|^2}}.$$

We then conclude

$$\begin{aligned} &P\left(\left|\frac{1}{T}\sigma(w^\mathsf{T}X)A\sigma(w^\mathsf{T}X)^\mathsf{T} - \frac{1}{T}\operatorname{tr}(\Phi A)\right| \geq t\right) \\ &\leq (\lambda+1)Ce^{-\frac{cT}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}\min(t^2/K^2, K^2)} \end{aligned}$$

and, with $K = \max(4t_0, \sqrt{t})$,

$$P\left(\left|\frac{1}{T}\sigma(w^\mathsf{T}X)A\sigma(w^\mathsf{T}X)^\mathsf{T} - \frac{1}{T}\operatorname{tr}(\Phi A)\right| \geq t\right) \leq (\lambda+1)Ce^{-\frac{cT\min\left(\frac{t^2}{16t_0^2}, t\right)}{2\|X\|^2\lambda_\varphi^2\lambda_\sigma^2}}.$$

Indeed, if $4t_0 \leq \sqrt{t}$ then $\min(t^2/K^2, K^2) = t$, while if $4t_0 \geq \sqrt{t}$ then $\min(t^2/K^2, K^2) = \min(t^2/16t_0^2, 16t_0^2) = t^2/16t_0^2$. □

As a corollary of Lemma 2, we have the following control of the moments of $\frac{1}{T}\sigma^\mathsf{T}A\sigma$.

Corollary 1 (Moments of quadratic forms). *Let Assumptions 1–2 hold. For $w \sim \mathcal{N}_\varphi(0, I_p)$, $\sigma \equiv \sigma(w^\mathsf{T}X)^\mathsf{T} \in \mathbb{R}^T$, $A \in \mathbb{R}^{T\times T}$ such that $\|A\| \leq 1$, and $k \in \mathbb{N}$,*

$$\mathrm{E}\left[\left|\frac{1}{T}\sigma^\mathsf{T}A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right|^k\right] \leq C_1\left(\frac{t_0\eta}{\sqrt{T}}\right)^k + C_2\left(\frac{\eta^2}{T}\right)^k$$

with $t_0 = |\sigma(0)| + \lambda_\sigma\lambda_\varphi\|X\|\sqrt{\frac{p}{T}}$, $\eta = \|X\|\lambda_\sigma\lambda_\varphi$, and $C_1, C_2 > 0$ independent of the other parameters. In particular, under the additional Assumption 3,

$$\mathrm{E}\left[\left|\frac{1}{T}\sigma^\mathsf{T}A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right|^k\right] \leq \frac{C}{n^{k/2}}$$

PROOF. We use the fact that, for a nonnegative random variable Y, $\mathrm{E}[Y] = \int_0^\infty P(Y > t)dt$, so that

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{T}\sigma^\mathsf{T} A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right|^k\right] \\
&= \int_0^\infty P\left(\left|\frac{1}{T}\sigma^\mathsf{T} A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right|^k > u\right) du \\
&= \int_0^\infty kv^{k-1} P\left(\left|\frac{1}{T}\sigma^\mathsf{T} A\sigma - \frac{1}{T}\operatorname{tr}\Phi A\right| > v\right) dv \\
&\leq \int_0^\infty kv^{k-1} Ce^{-\frac{cT}{\eta^2}\min\left(\frac{v^2}{t_0^2}, v\right)} dv \\
&\leq \int_0^{t_0} kv^{k-1} Ce^{-\frac{cTv^2}{t_0^2\eta^2}} dv + \int_{t_0}^\infty kv^{k-1} Ce^{-\frac{cTv}{\eta^2}} dv \\
&\leq \int_0^\infty kv^{k-1} Ce^{-\frac{cTv^2}{t_0^2\eta^2}} dv + \int_0^\infty kv^{k-1} Ce^{-\frac{cTv}{\eta^2}} dv \\
&= \left(\frac{t_0\eta}{\sqrt{cT}}\right)^k \int_0^\infty kt^{k-1} Ce^{-t^2} dt + \left(\frac{\eta^2}{cT}\right)^k \int_0^\infty kt^{k-1} Ce^{-t} dt
\end{aligned}$$

which, along with the boundedness of the integrals, concludes the proof. □

Beyond concentration results on functions of the vector σ, we also have the following convenient property for functions of the matrix Σ.

LEMMA 3 (Lipschitz functions of Σ). *Let $f : \mathbb{R}^{n\times T} \to \mathbb{R}$ be a λ_f-Lipschitz function with respect to the Froebnius norm. Then, under Assumptions 1–2,*

$$P\left(\left|f\left(\frac{\Sigma}{\sqrt{T}}\right) - \mathrm{E}f\left(\frac{\Sigma}{\sqrt{T}}\right)\right| > t\right) \leq Ce^{-\frac{cTt^2}{\lambda_\sigma^2\lambda_\varphi^2\lambda_f^2\|X\|^2}}$$

for some $C, c > 0$. In particular, under the additional Assumption 3,

$$P\left(\left|f\left(\frac{\Sigma}{\sqrt{T}}\right) - \mathrm{E}f\left(\frac{\Sigma}{\sqrt{T}}\right)\right| > t\right) \leq Ce^{-cTt^2}.$$

PROOF. Denoting $W = \varphi(\tilde{W})$, since $\mathrm{vec}(\tilde{W}) \equiv [\tilde{W}_{11}, \cdots, \tilde{W}_{np}]$ is a Gaussian vector, by the normal concentration of Gaussian vectors, for g a λ_g-

Lipschitz function of W with respect to the Frobenius norm (i.e., the Euclidean norm of $\mathrm{vec}(W)$), by (6),

$$P\left(|g(W)-\mathrm{E}[g(W)]|>t\right)=P\left(\left|g(\varphi(\tilde{W}))-\mathrm{E}[g(\varphi(\tilde{W}))]\right|>t\right)\leq Ce^{-\frac{ct^2}{\lambda_g^2\lambda_\varphi^2}}$$

for some $C,c>0$. Let's consider in particular $g:W\mapsto f(\Sigma/\sqrt{T})$ and remark that

$$\begin{aligned}|g(W+H)-g(W)|&=\left|f\left(\frac{\sigma((W+H)X)}{\sqrt{T}}\right)-f\left(\frac{\sigma(WX)}{\sqrt{T}}\right)\right|\\&\leq\frac{\lambda_f}{\sqrt{T}}\left\|\sigma((W+H)X)-\sigma(WX)\right\|_F\\&\leq\frac{\lambda_f\lambda_\sigma}{\sqrt{T}}\left\|HX\right\|_F\\&=\frac{\lambda_f\lambda_\sigma}{\sqrt{T}}\sqrt{\mathrm{tr}\,HXX^\mathsf{T}H^\mathsf{T}}\\&\leq\frac{\lambda_f\lambda_\sigma}{\sqrt{T}}\sqrt{\|XX^\mathsf{T}\|}\|H\|_F\end{aligned}$$

concluding the proof. □

A first corollary of Lemma 3 is the concentration of the Stieltjes transform $\frac{1}{T}\mathrm{tr}\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma-zI_T\right)^{-1}$ of μ_n, the empirical spectral measure of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$, for all $z\in\mathbb{C}\setminus\mathbb{R}^+$ (so in particular, for $z=-\gamma$, $\gamma>0$).

Corollary 2 (Concentration of the Stieltjes transform of μ_n). *Under Assumptions 1–2, for $z\in\mathbb{C}\setminus\mathbb{R}^+$,*

$$\begin{aligned}&P\left(\left|\frac{1}{T}\mathrm{tr}\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma-zI_T\right)^{-1}-\mathrm{E}\left[\frac{1}{T}\mathrm{tr}\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma-zI_T\right)^{-1}\right]\right|>t\right)\\&\leq Ce^{-\frac{\mathrm{cdist}(z,\mathbb{R}^+)^2Tt^2}{\lambda_\sigma^2\lambda_\varphi^2\|X\|^2}}\end{aligned}$$

for some $C,c>0$, where $\mathrm{dist}(z,\mathbb{R}^+)$ is the Hausdorff set distance. In particular, for $z=-\gamma$, $\gamma>0$, and under the additional Assumption 3

$$P\left(\left|\frac{1}{T}\mathrm{tr}\,Q-\frac{1}{T}\mathrm{tr}\,\mathrm{E}[Q]\right|>t\right)\leq Ce^{-cnt^2}.$$

PROOF. We can apply Lemma 3 for $f : R \mapsto \frac{1}{T}\operatorname{tr}(R^\mathsf{T}R - zI_T)^{-1}$, since we have

$$
\begin{aligned}
&|f(R+H) - f(R)| \\
&= \left| \frac{1}{T}\operatorname{tr}((R+H)^\mathsf{T}(R+H) - zI_T)^{-1}((R+H)^\mathsf{T}H + H^\mathsf{T}R)(R^\mathsf{T}R - zI_T)^{-1} \right| \\
&\leq \left| \frac{1}{T}\operatorname{tr}((R+H)^\mathsf{T}(R+H) - zI_T)^{-1}(R+H)^\mathsf{T}H(R^\mathsf{T}R - zI_T)^{-1} \right| \\
&+ \left| \frac{1}{T}\operatorname{tr}((R+H)^\mathsf{T}(R+H) - zI_T)^{-1}H^\mathsf{T}R(R^\mathsf{T}R - zI_T)^{-1} \right| \\
&\leq \frac{2\|H\|}{\operatorname{dist}(z, \mathbb{R}^+)} \leq \frac{2\|H\|_F}{\operatorname{dist}(z, \mathbb{R}^+)}
\end{aligned}
$$

where, for the second to last inequality, we successively used the relations $|\operatorname{tr} AB| \leq \sqrt{\operatorname{tr} AA^\mathsf{T}}\sqrt{\operatorname{tr} BB^\mathsf{T}}$, $|\operatorname{tr} CD| \leq \|D\| \operatorname{tr} C$ for nonnegative definite C, and $\|(R^\mathsf{T}R - zI_T)^{-1}R^\mathsf{T}R\| \leq 1$, $\|(R^\mathsf{T}R - zI_T)^{-1}\| \leq \operatorname{dist}(z, \mathbb{R}^+)^{-1}$, for $z \in \mathbb{C} \setminus \mathbb{R}^+$, and finally $\|\cdot\| \leq \|\cdot\|_F$. □

Lemma 3 also allows for an important application of Lemma 2 as follows.

LEMMA 4 (Concentration of $\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma$). *Let Assumptions 1–3 hold and write* $W^\mathsf{T} = [w_1, \ldots, w_n]$. *Define* $\sigma \equiv \sigma(w_1^\mathsf{T}X)^\mathsf{T} \in \mathbb{R}^T$ *and, for* $W_-^\mathsf{T} = [w_2, \ldots, w_n]$ *and* $\Sigma_- = \sigma(W_-X)$, *let* $Q_- = (\frac{1}{T}\Sigma_-^\mathsf{T}\Sigma_- + \gamma I_T)^{-1}$. *Then, for* $A, B \in \mathbb{R}^{T\times T}$ *such that* $\|A\|, \|B\| \leq 1$

$$
P\left(\left| \frac{1}{T}\sigma^\mathsf{T}AQ_-B\sigma - \frac{1}{T}\operatorname{tr}\Phi A\mathrm{E}[Q_-]B \right| > t \right) \leq Ce^{-cn\min(t^2,t)}
$$

for some $C, c > 0$ *independent of the other parameters.*

PROOF. Let $f : R \mapsto \frac{1}{T}\sigma^\mathsf{T}A(R^\mathsf{T}R + \gamma I_T)^{-1}B\sigma$. Reproducing the proof of Corollary 2, conditionally to $\frac{1}{T}\|\sigma\|^2 \leq K$ for any arbitrary large enough $K > 0$, it appears that f is Lipschitz with parameter of order $O(1)$. Along with (7) and Assumption 3, this thus ensures that

$$
\begin{aligned}
&P\left(\left| \frac{1}{T}\sigma^\mathsf{T}AQ_-B\sigma - \frac{1}{T}\sigma^\mathsf{T}A\mathrm{E}[Q_-]B\sigma \right| > t \right) \\
&\leq P\left(\left| \frac{1}{T}\sigma^\mathsf{T}AQ_-B\sigma - \frac{1}{T}\sigma^\mathsf{T}A\mathrm{E}[Q_-]B\sigma \right| > t,\ \frac{\|\sigma\|^2}{T} \leq K \right) + P\left(\frac{\|\sigma\|^2}{T} > K \right) \\
&\leq Ce^{-cnt^2}
\end{aligned}
$$

for some $C, c > 0$. We may then apply Lemma 1 on the bounded norm matrix $A\mathrm{E}[Q_-]B$ to further find that

$$\begin{aligned}
&P\left(\left|\frac{1}{T}\sigma^\mathsf{T} AQ_-B\sigma - \frac{1}{T}\operatorname{tr}\Phi A\mathrm{E}[Q_-]B\right| > t\right)\\
&\leq P\left(\left|\frac{1}{T}\sigma^\mathsf{T} AQ_-B\sigma - \frac{1}{T}\sigma^\mathsf{T} A\mathrm{E}[Q_-]B\sigma\right| > \frac{t}{2}\right)\\
&+ P\left(\left|\frac{1}{T}\sigma^\mathsf{T} A\mathrm{E}[Q_-]B\sigma - \frac{1}{T}\operatorname{tr}\Phi A\mathrm{E}[Q_-]B\right| > \frac{t}{2}\right)\\
&\leq C' e^{-c' n \min(t^2, t)}
\end{aligned}$$

which concludes the proof. □

As a further corollary of Lemma 3, we have the following concentration result on the training mean-square error of the neural network under study.

COROLLARY 3 (Concentration of the mean-square error). *Under Assumptions 1–3,*

$$P\left(\left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQ^2 - \frac{1}{T}\operatorname{tr} Y^\mathsf{T} Y\mathrm{E}\left[Q^2\right]\right| > t\right) \leq Ce^{-cnt^2}$$

for some $C, c > 0$ independent of the other parameters.

PROOF. We apply Lemma 3 to the mapping $f : R \mapsto \frac{1}{T}\operatorname{tr} Y^\mathsf{T} Y(R^\mathsf{T} R + \gamma I_T)^{-2}$. Denoting $Q = (R^\mathsf{T} R + \gamma I_T)^{-1}$ and $Q^H = ((R+H)^\mathsf{T}(R+H) + \gamma I_T)^{-1}$, remark indeed that

$$\begin{aligned}
&|f(R+H) - f(R)|\\
&= \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} Y((Q^H)^2 - Q^2)\right|\\
&\leq \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} Y(Q^H - Q)Q^H\right| + \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQ(Q^H - Q)\right|\\
&= \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQ^H((R+H)^\mathsf{T}(R+H) - R^\mathsf{T} R)QQ^H\right|\\
&+ \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQQ^H((R+H)^\mathsf{T}(R+H) - R^\mathsf{T} R)Q\right|\\
&\leq \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQ^H(R+H)^\mathsf{T} HQQ^H\right| + \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQ^H H^\mathsf{T} RQQ^H\right|\\
&+ \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQQ^H(R+H)^\mathsf{T} RQ\right| + \left|\frac{1}{T}\operatorname{tr} Y^\mathsf{T} YQQ^H H^\mathsf{T} RQ\right|.
\end{aligned}$$

As $\|Q^H(R+H)^\mathsf{T}\| = \sqrt{\|Q^H(R+H)^\mathsf{T}(R+H)Q^H\|}$ and $\|RQ\| = \sqrt{\|QR^\mathsf{T}RQ\|}$ are bounded and $\frac{1}{T}\operatorname{tr} Y^\mathsf{T}Y$ is also bounded by Assumption 3, this implies

$$|f(R+H) - f(R)| \leq C\|H\| \leq C\|H\|_F$$

for some $C > 0$. The function f is thus Lipschitz with parameter independent of n, which allows us to conclude using Lemma 3. □

The aforementioned concentration results are the building blocks of the proofs of Theorem 1–3 which, under all Assumptions 1–3, are established using standard random matrix approaches.

5.2. *Asymptotic Equivalents.*

5.2.1. *First Equivalent for* $\mathrm{E}[Q]$. This section is dedicated to a first characterization of $\mathrm{E}[Q]$, in the "simultaneously large" n, p, T regime. This preliminary step is classical in studying resolvents in random matrix theory as the direct comparison of $\mathrm{E}[Q]$ to $\bar{Q}$ with the implicit δ may be cumbersome. To this end, let us thus define the intermediary deterministic matrix

$$\tilde{Q} = \left(\frac{n}{T}\frac{\Phi}{1+\alpha} + \gamma I_T\right)^{-1}$$

with $\alpha \equiv \frac{1}{T}\operatorname{tr}\Phi\mathrm{E}[Q_-]$, where we recall that Q_- is a random matrix distributed as, say, $(\frac{1}{T}\Sigma^\mathsf{T}\Sigma - \frac{1}{T}\sigma_1\sigma_1^\mathsf{T} + \gamma I_T)^{-1}$.

First note that, since $\frac{1}{T}\operatorname{tr}\Phi = \mathrm{E}[\frac{1}{T}\|\sigma\|^2]$ and, from (7) and Assumption 3, $P(\frac{1}{T}\|\sigma\|^2 > t) \leq Ce^{-cnt^2}$ for all large t, we find that $\frac{1}{T}\operatorname{tr}\Phi = \int_0^\infty t^2 P(\frac{1}{T}\|\sigma\|^2 > t)dt \leq C'$ for some constant C'. Thus, $\alpha \leq \|\mathrm{E}[Q_-]\|\frac{1}{T}\operatorname{tr}\Phi \leq \frac{C'}{\gamma}$ is uniformly bounded.

We will show here that $\|\mathrm{E}[Q] - \tilde{Q}\| \to 0$ as $n \to \infty$ in the regime of Assumption 3. As the proof steps are somewhat classical, we defer to the appendix some classical intermediary lemmas (Lemmas 5–7). Using the resolvent identity, Lemma 5, we start by writing

$$\begin{aligned}
\mathrm{E}[Q] - \tilde{Q} &= \mathrm{E}\left[Q\left(\frac{n}{T}\frac{\Phi}{1+\alpha} - \frac{1}{T}\Sigma^\mathsf{T}\Sigma\right)\right]\tilde{Q} \\
&= \mathrm{E}[Q]\frac{n}{T}\frac{\Phi}{1+\alpha}\tilde{Q} - \mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}\Sigma\right]\tilde{Q} \\
&= \mathrm{E}[Q]\frac{n}{T}\frac{\Phi}{1+\alpha}\tilde{Q} - \frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q\sigma_i\sigma_i^\mathsf{T}\right]\tilde{Q}
\end{aligned}$$

which, from Lemma 6, gives, for $Q_{-i} = (\frac{1}{T}\Sigma^\mathsf{T}\Sigma - \frac{1}{T}\sigma_i\sigma_i^\mathsf{T} + \gamma I_T)^{-1}$,

$$\begin{aligned}\mathrm{E}[Q] - \tilde{Q} &= \mathrm{E}[Q]\frac{n}{T}\frac{\Phi}{1+\alpha}\tilde{Q} - \frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q_{-i}\frac{\sigma_i\sigma_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}\right]\tilde{Q} \\ &= \mathrm{E}[Q]\frac{n}{T}\frac{\Phi}{1+\alpha}\tilde{Q} - \frac{1}{1+\alpha}\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q_{-i}\sigma_i\sigma_i^\mathsf{T}\right]\tilde{Q} \\ &\quad + \frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[\frac{Q_{-i}\sigma_i\sigma_i^\mathsf{T}\left(\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right)}{(1+\alpha)(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)}\right]\tilde{Q}.\end{aligned}$$

Note now, from the independence of Q_{-i} and $\sigma_i\sigma_i^\mathsf{T}$, that the second right-hand side expectation is simply $\mathrm{E}[Q_{-i}]\Phi$. Also, exploiting Lemma 6 in reverse on the rightmost term, this gives

$$\begin{aligned}\mathrm{E}[Q] - \tilde{Q} &= \frac{1}{T}\sum_{i=1}^n \frac{\mathrm{E}[Q-Q_{-i}]\Phi}{1+\alpha}\tilde{Q} \\ &\quad + \frac{1}{1+\alpha}\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q\sigma_i\sigma_i^\mathsf{T}\tilde{Q}\left(\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right)\right]. \qquad (8)\end{aligned}$$

It is convenient at this point to note that, since $\mathrm{E}[Q] - \tilde{Q}$ is symmetric, we may write

$$\begin{aligned}\mathrm{E}[Q] - \tilde{Q} &= \frac{1}{2}\frac{1}{1+\alpha}\left(\frac{1}{T}\sum_{i=1}^n \left(\mathrm{E}[Q-Q_{-i}]\Phi\tilde{Q} + \tilde{Q}\Phi\mathrm{E}[Q-Q_{-i}]\right)\right. \\ &\quad \left. + \frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[\left(Q\sigma_i\sigma_i^\mathsf{T}\tilde{Q} + \tilde{Q}\sigma_i\sigma_i^\mathsf{T}Q\right)\left(\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right)\right]\right). \qquad (9)\end{aligned}$$

We study the two right-hand side terms of (9) independently.

For the first term, since $Q - Q_{-i} = -Q\frac{1}{T}\sigma_i\sigma_i^\mathsf{T}Q_{-i}$,

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^n \frac{\mathrm{E}[Q-Q_{-i}]\Phi}{1+\alpha}\tilde{Q} &= \frac{1}{1+\alpha}\frac{1}{T}\mathrm{E}\left[Q\frac{1}{T}\sum_{i=1}^n \sigma_i\sigma_i^\mathsf{T}Q_{-i}\right]\Phi\tilde{Q} \\ &= \frac{1}{1+\alpha}\frac{1}{T}\mathrm{E}\left[Q\frac{1}{T}\sum_{i=1}^n \sigma_i\sigma_i^\mathsf{T}Q\left(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i\right)\right]\Phi\tilde{Q}\end{aligned}$$

where we used again Lemma 6 in reverse. Denoting $D = \mathrm{diag}(\{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i\}_{i=1}^n)$, this can be compactly written

$$\frac{1}{T}\sum_{i=1}^n \frac{\mathrm{E}[Q-Q_{-i}]\Phi}{1+\alpha}\tilde{Q} = \frac{1}{1+\alpha}\frac{1}{T}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\right]\Phi\tilde{Q}.$$

Note at this point that, from Lemma 7, $\|\Phi\tilde{Q}\| \leq (1+\alpha)\frac{T}{n}$ and

$$\left\| Q\frac{1}{\sqrt{T}}\Sigma^{\mathsf{T}} \right\| = \sqrt{\left\| Q\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma Q \right\|} \leq \gamma^{-\frac{1}{2}}.$$

Besides, by Lemma 4 and the union bound,

$$P\left(\max_{1\leq i\leq n} D_{ii} > 1+\alpha+t \right) \leq Cne^{-cn\min(t^2,t)}$$

for some $C, c > 0$, so in particular, recalling that $\alpha \leq C'$ for some constant $C' > 0$,

$$\begin{aligned}
\mathrm{E}\left[\max_{1\leq i\leq n} D_{ii} \right] &= \int_0^{2(1+C')} P\left(\max_{1\leq i\leq n} D_{ii} > t \right) dt + \int_{2(1+C')}^{\infty} P\left(\max_{1\leq i\leq n} D_{ii} > t \right) dt \\
&\leq 2(1+C') + \int_{2(1+C')}^{\infty} Cne^{-cn\min((t-(1+C'))^2, t-(1+C'))} dt \\
&= 2(1+C') + \int_{1+C'}^{\infty} Cne^{-cnt} dt \\
&= 2(1+C') + e^{-Cn(1+C')} = O(1).
\end{aligned}$$

As a consequence of all the above (and of the boundedness of α), we have that, for some $c > 0$,

$$\frac{1}{T}\left\| \mathrm{E}\left[Q\frac{1}{T}\Sigma^{\mathsf{T}} D\Sigma Q \right] \Phi\tilde{Q} \right\| \leq \frac{c}{n}. \tag{10}$$

Let us now consider the second right-hand side term of (9). Using the relation $ab^{\mathsf{T}} + ba^{\mathsf{T}} \preceq aa^{\mathsf{T}} + bb^{\mathsf{T}}$ in the order of Hermitian matrices (which unfolds from $(a-b)(a-b)^{\mathsf{T}} \succeq 0$), we have, with $a = T^{\frac{1}{4}} Q\sigma_i(\frac{1}{T}\sigma_i^{\mathsf{T}} Q_{-i}\sigma_i - \alpha)$ and $b = T^{-\frac{1}{4}}\tilde{Q}\sigma_i$,

$$\begin{aligned}
&\frac{1}{T}\sum_{i=1}^{n} \mathrm{E}\left[\left(Q\sigma_i\sigma_i^{\mathsf{T}}\tilde{Q} + \tilde{Q}\sigma_i\sigma^{\mathsf{T}} Q \right) \left(\frac{1}{T}\sigma_i^{\mathsf{T}} Q_{-i}\sigma_i - \alpha \right) \right] \\
&\preceq \frac{1}{\sqrt{T}}\sum_{i=1}^{n} \mathrm{E}\left[Q\sigma_i\sigma_i^{\mathsf{T}} Q \left(\frac{1}{T}\sigma_i^{\mathsf{T}} Q_{-i}\sigma_i - \alpha \right)^2 \right] + \frac{1}{T\sqrt{T}}\sum_{i=1}^{n} \mathrm{E}\left[\tilde{Q}\sigma_i\sigma_i^{\mathsf{T}}\tilde{Q} \right] \\
&= \sqrt{T}\mathrm{E}\left[Q\frac{1}{T}\Sigma^{\mathsf{T}} D_2^2 \Sigma Q \right] + \frac{n}{T\sqrt{T}}\tilde{Q}\Phi\tilde{Q}
\end{aligned}$$

where $D_2 = \text{diag}(\{\frac{1}{T}\sigma_i^\mathsf{T} Q_{-i}\sigma_i - \alpha\}_{i=1}^n)$. Of course, since we also have $-aa^\mathsf{T} - bb^\mathsf{T} \preceq ab^\mathsf{T} + ba^\mathsf{T}$ (from $(a+b)(a+b)^\mathsf{T} \succeq 0$), we have symmetrically

$$\begin{aligned}&\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[\left(Q\sigma_i\sigma_i^\mathsf{T}\tilde{Q} + \tilde{Q}\sigma_i\sigma^\mathsf{T}Q\right)\left(\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right)\right]\\ &\succeq -\sqrt{T}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D_2^2\Sigma Q\right] - \frac{n}{T\sqrt{T}}\tilde{Q}\Phi\tilde{Q}.\end{aligned}$$

But from Lemma 4,

$$\begin{aligned}P\left(\|D_2\| > tn^{\varepsilon-\frac{1}{2}}\right) &= P\left(\max_{1\le i\le n}\left|\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right| > tn^{\varepsilon-\frac{1}{2}}\right)\\ &\le Cne^{-c\min(n^{2\varepsilon}t^2, n^{\frac{1}{2}+\varepsilon}t)}\end{aligned}$$

so that, with a similar reasoning as in the proof of Corollary 1,

$$\left\|\sqrt{T}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D_2^2\Sigma Q\right]\right\| \le \sqrt{T}\mathrm{E}\left[\|D_2\|^2\right] \le Cn^{\varepsilon'-\frac{1}{2}}$$

where we additionally used $\|Q\Sigma\| \le \sqrt{T}$ in the first inequality.

Since in addition $\left\|\frac{n}{T\sqrt{T}}\tilde{Q}\Phi\tilde{Q}\right\| \le Cn^{-\frac{1}{2}}$, this gives

$$\left\|\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[\left(Q\sigma_i\sigma_i^\mathsf{T}\tilde{Q} + \tilde{Q}\sigma_i\sigma_i^\mathsf{T}Q\right)\left(\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i - \alpha\right)\right]\right\| \le Cn^{\varepsilon-\frac{1}{2}}.$$

Together with (9), we thus conclude that

$$\left\|\mathrm{E}[Q] - \tilde{Q}\right\| \le Cn^{\varepsilon-\frac{1}{2}}.$$

Note in passing that we proved that

$$\|\mathrm{E}\left[Q - Q_-\right]\| = \frac{T}{n}\left\|\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q - Q_{-i}\right]\right\| = \left\|\frac{1}{n}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\right]\right\| \le \frac{c}{n}$$

where the first equality holds by exchangeability arguments.

In particular,

$$\alpha = \frac{1}{T}\operatorname{tr}\Phi\mathrm{E}[Q_-] = \frac{1}{T}\operatorname{tr}\Phi\mathrm{E}[Q] + \frac{1}{T}\operatorname{tr}\Phi(\mathrm{E}[Q_-] - \mathrm{E}[Q])$$

where $|\frac{1}{T}\operatorname{tr}\Phi(\mathrm{E}[Q_-]-\mathrm{E}[Q])|\leq\frac{c}{n}$. And thus, by the previous result,

$$\left|\alpha-\frac{1}{T}\operatorname{tr}\Phi\tilde{Q}\right|\leq Cn^{-\frac{1}{2}+\varepsilon}\frac{1}{T}\operatorname{tr}\Phi.$$

We have proved in the beginning of the section that $\frac{1}{T}\operatorname{tr}\Phi$ is bounded and thus we finally conclude that

$$\left\|\alpha-\frac{1}{T}\operatorname{tr}\Phi\tilde{Q}\right\|\leq Cn^{\varepsilon-\frac{1}{2}}.$$

5.2.2. *Second Equivalent for* $\mathrm{E}[Q]$. In this section, we show that $\mathrm{E}[Q]$ can be approximated by the matrix $\bar{Q}$, which we recall is defined as

$$\bar{Q}=\left(\frac{n}{T}\frac{\Phi}{1+\delta}+\gamma I_T\right)^{-1}$$

where $\delta>0$ is the unique positive solution to $\delta=\frac{1}{T}\operatorname{tr}\Phi\bar{Q}$. The fact that $\delta>0$ is well defined is quite standard and has already been proved several times for more elaborate models. Following the ideas of (Hoydis, Couillet and Debbah, 2013), we may for instance use the framework of so-called standard interference functions (Yates, 1995) which claims that, if a map $f:[0,\infty)\to(0,\infty)$, $x\mapsto f(x)$, satisfies $x\geq x'\Rightarrow f(x)\geq f(x')$, $\forall a>1, af(x)>f(ax)$ and there exists x_0 such that $x_0\geq f(x_0)$, then f has a unique fixed point (Yates, 1995, Th 2). It is easily shown that $\delta\mapsto\frac{1}{T}\operatorname{tr}\Phi\bar{Q}$ is such a map, so that δ exists and is unique.

To compare $\tilde{Q}$ and $\bar{Q}$, using the resolvent identity, Lemma 5, we start by writing

$$\tilde{Q}-\bar{Q}=(\alpha-\delta)\tilde{Q}\frac{n}{T}\frac{\Phi}{(1+\alpha)(1+\delta)}\bar{Q}$$

from which

$$\begin{aligned}|\alpha-\delta|&=\left|\frac{1}{T}\operatorname{tr}\Phi\left(\mathrm{E}[Q_-]-\bar{Q}\right)\right|\\&\leq\left|\frac{1}{T}\operatorname{tr}\Phi\left(\tilde{Q}-\bar{Q}\right)\right|+cn^{-\frac{1}{2}+\varepsilon}\\&=|\alpha-\delta|\frac{1}{T}\operatorname{tr}\frac{\Phi\tilde{Q}\frac{n}{T}\Phi\bar{Q}}{(1+\alpha)(1+\delta)}+cn^{-\frac{1}{2}+\varepsilon}\end{aligned}$$

which implies that

$$|\alpha-\delta|\left(1-\frac{1}{T}\operatorname{tr}\frac{\Phi\tilde{Q}\frac{n}{T}\Phi\bar{Q}}{(1+\alpha)(1+\delta)}\right)\leq cn^{-\frac{1}{2}+\varepsilon}.$$

It thus remains to show that

$$\limsup_n \frac{1}{T}\operatorname{tr}\frac{\Phi\tilde{Q}\frac{n}{T}\Phi\bar{Q}}{(1+\alpha)(1+\delta)}<1$$

to prove that $|\alpha-\delta|\leq cn^{\varepsilon-\frac{1}{2}}$. To this end, note that, by Cauchy–Schwarz's inequality,

$$\frac{1}{T}\operatorname{tr}\frac{\Phi\tilde{Q}\frac{n}{T}\Phi\bar{Q}}{(1+\alpha)(1+\delta)}\leq\sqrt{\frac{n}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2\cdot\frac{n}{T(1+\alpha)^2}\frac{1}{T}\operatorname{tr}\Phi^2\tilde{Q}^2}$$

so that it is sufficient to bound the limsup of both terms under the square root strictly by one. Next, remark that

$$\delta=\frac{1}{T}\operatorname{tr}\Phi\bar{Q}=\frac{1}{T}\operatorname{tr}\Phi\bar{Q}^2\bar{Q}^{-1}=\frac{n(1+\delta)}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2+\gamma\frac{1}{T}\operatorname{tr}\Phi\bar{Q}^2.$$

In particular,

$$\frac{n}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2=\frac{\delta\frac{n}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)\frac{n}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2+\gamma\frac{1}{T}\operatorname{tr}\Phi\bar{Q}^2}\leq\frac{\delta}{1+\delta}.$$

But at the same time, since $\|(\frac{n}{T}\Phi+\gamma I_T)^{-1}\|\leq\gamma^{-1}$,

$$\delta\leq\frac{1}{\gamma T}\operatorname{tr}\Phi$$

the limsup of which is bounded. We thus conclude that

$$\limsup_n \frac{n}{T(1+\delta)^2}\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2<1. \tag{11}$$

Similarly, α, which is known to be bounded, satisfies

$$\alpha=(1+\alpha)\frac{n}{T(1+\alpha)^2}\frac{1}{T}\operatorname{tr}\Phi^2\tilde{Q}^2+\gamma\frac{1}{T}\operatorname{tr}\Phi\tilde{Q}^2+O(n^{\varepsilon-\frac{1}{2}})$$

and we thus have also

$$\limsup_n \frac{n}{T(1+\alpha)^2}\frac{1}{T}\operatorname{tr}\Phi^2\tilde{Q}^2<1$$

which completes to prove that $|\alpha - \delta| \leq cn^{\varepsilon - \frac{1}{2}}$.

As a consequence of all this,

$$\|\tilde{Q} - \bar{Q}\| = |\alpha - \delta| \cdot \left\| \frac{\tilde{Q} \frac{n}{T} \Phi \bar{Q}}{(1+\alpha)(1+\delta)} \right\| \leq cn^{-\frac{1}{2}+\varepsilon}$$

and we have thus proved that $\|\mathrm{E}[Q] - \bar{Q}\| \leq cn^{-\frac{1}{2}+\varepsilon}$ for some $c > 0$.

From this result, along with Corollary 2, we now have that

$$\begin{aligned} &P\left(\left| \frac{1}{T} \operatorname{tr} Q - \frac{1}{T} \operatorname{tr} \bar{Q} \right| > t \right) \\ &\leq P\left(\left| \frac{1}{T} \operatorname{tr} Q - \frac{1}{T} \operatorname{tr} \mathrm{E}[Q] \right| > t - \left| \frac{1}{T} \operatorname{tr} \mathrm{E}[Q] - \frac{1}{T} \operatorname{tr} \bar{Q} \right| \right) \\ &\leq C' e^{-c' n (t - cn^{-\frac{1}{2}+\varepsilon})} \leq C' e^{-\frac{1}{2} c' n t} \end{aligned}$$

for all large n. As a consequence, for all $\gamma > 0$, $\frac{1}{T} \operatorname{tr} Q - \frac{1}{T} \operatorname{tr} \bar{Q} \to 0$ almost surely. As such, the difference $m_{\mu_n} - m_{\bar{\mu}_n}$ of Stieltjes transforms $m_{\mu_n} : \mathbb{C} \setminus \mathbb{R}^+ \to \mathbb{C}$, $z \mapsto \frac{1}{T} \operatorname{tr}(\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma - z I_T)^{-1}$ and $m_{\bar{\mu}_n} : \mathbb{C} \setminus \mathbb{R}^+ \to \mathbb{C}$, $z \mapsto \frac{1}{T} \operatorname{tr}(\frac{n}{T} \frac{\Phi}{1+\delta_z} - z I_T)^{-1}$ (with δ_z the unique Stieltjes transform solution to $\delta_z = \frac{1}{T} \operatorname{tr} \Phi (\frac{n}{T} \frac{\Phi}{1+\delta_z} - z I_T)^{-1}$) converges to zero for each z in a subset of $\mathbb{C} \setminus \mathbb{R}^+$ having at least one accumulation point (namely $\mathbb{R}^-$), almost surely so (that is, on a probability set $\mathcal{A}_z$ with $P(\mathcal{A}_z) = 1$). Thus, letting $\{z_k\}_{k=1}^{\infty}$ be a converging sequence strictly included in $\mathbb{R}^-$, on the probability one space $\mathcal{A} = \cap_{k=1}^{\infty} \mathcal{A}_k$, $m_{\mu_n}(z_k) - m_{\bar{\mu}_n}(z_k) \to 0$ for all k. Now, m_{μ_n} is complex analytic on $\mathbb{C} \setminus \mathbb{R}^+$ and bounded on all compact subsets of $\mathbb{C} \setminus \mathbb{R}^+$. Besides, it was shown in (Silverstein and Bai, 1995; Silverstein and Choi, 1995) that the function $m_{\bar{\mu}_n}$ is well-defined, complex analytic and bounded on all compact subsets of $\mathbb{C} \setminus \mathbb{R}^+$. As a result, on $\mathcal{A}$, $m_{\mu_n} - m_{\bar{\mu}_n}$ is complex analytic, bounded on all compact subsets of $\mathbb{C} \setminus \mathbb{R}^+$ and converges to zero on a subset admitting at least one accumulation point. Thus, by Vitali's convergence theorem (Titchmarsh, 1939), with probability one, $m_{\mu_n} - m_{\bar{\mu}_n}$ converges to zero everywhere on $\mathbb{C} \setminus \mathbb{R}^+$. This implies, by (Bai and Silverstein, 2009, Theorem B.9), that $\mu_n - \bar{\mu}_n \to 0$, vaguely as a signed finite measure, with probability one, and, since $\bar{\mu}_n$ is a probability measure (again from the results of (Silverstein and Bai, 1995; Silverstein and Choi, 1995)), we have thus proved Theorem 2.

5.2.3. *Asymptotic Equivalent for* $\mathrm{E}[QAQ]$*, where* A *is either* Φ *or symmetric of bounded norm.* The evaluation of the second order statistics of

the neural network under study requires, beside $\mathrm{E}[Q]$, to evaluate the more involved form $\mathrm{E}[QAQ]$, where A is a symmetric matrix either equal to Φ or of bounded norm (so in particular $\|\bar{Q}A\|$ is bounded). To evaluate this quantity, first write

$$\begin{aligned}\mathrm{E}[QAQ] &= \mathrm{E}\left[\bar{Q}AQ\right] + \mathrm{E}\left[(Q-\bar{Q})AQ\right] \\ &= \mathrm{E}\left[\bar{Q}AQ\right] + \mathrm{E}\left[Q\left(\frac{n}{T}\frac{\Phi}{1+\delta} - \frac{1}{T}\Sigma^{\mathsf{T}}\Sigma\right)\bar{Q}AQ\right] \\ &= \mathrm{E}\left[\bar{Q}AQ\right] + \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q\Phi\bar{Q}AQ\right] - \frac{1}{T}\sum_{i=1}^{n}\mathrm{E}\left[Q\sigma_i\sigma_i^{\mathsf{T}}\bar{Q}AQ\right].\end{aligned}$$

Of course, since QAQ is symmetric, we may write

$$\begin{aligned}\mathrm{E}[QAQ] = \frac{1}{2}\Bigg(&\mathrm{E}\left[\bar{Q}AQ + QA\bar{Q}\right] + \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q\Phi\bar{Q}AQ + QA\bar{Q}\Phi Q\right] \\ &-\frac{1}{T}\sum_{i=1}^{n}\mathrm{E}\left[Q\sigma_i\sigma_i^{\mathsf{T}}\bar{Q}AQ + QA\bar{Q}\sigma_i\sigma_i^{\mathsf{T}}Q\right]\Bigg)\end{aligned}$$

which will reveal more practical to handle.

First note that, since $\left\|\mathrm{E}[Q]-\bar{Q}\right\| \leq Cn^{\varepsilon-\frac{1}{2}}$ and A is such that $\|\bar{Q}A\|$ is bounded, $\|\mathrm{E}[\bar{Q}AQ]-\bar{Q}A\bar{Q}]\| \leq \|\bar{Q}A\|\|\mathrm{E}[Q]-\bar{Q}\| \leq C'n^{\varepsilon-\frac{1}{2}}$, which provides an estimate for the first expectation. We next evaluate the last right-hand side expectation above. With the same notations as previously, from exchangeability arguments and using $Q = Q_- - Q\frac{1}{T}\sigma\sigma^{\mathsf{T}}Q_-$, observe that

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{n}\mathrm{E}\left[Q\sigma_i\sigma_i^{\mathsf{T}}\bar{Q}AQ\right] &= \frac{n}{T}\mathrm{E}\left[Q\sigma\sigma^{\mathsf{T}}\bar{Q}AQ\right] \\ &= \frac{n}{T}\mathrm{E}\left[\frac{Q_-\sigma\sigma^{\mathsf{T}}\bar{Q}AQ}{1+\frac{1}{T}\sigma^{\mathsf{T}}Q_-\sigma}\right] \\ &= \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q_-\sigma\sigma^{\mathsf{T}}\bar{Q}AQ\right] \\ &\quad + \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q_-\sigma\sigma^{\mathsf{T}}\bar{Q}AQ\frac{\delta-\frac{1}{T}\sigma^{\mathsf{T}}Q_-\sigma}{1+\frac{1}{T}\sigma^{\mathsf{T}}Q_-\sigma}\right]\end{aligned}$$

which, reusing $Q = Q_- - Q\frac{1}{T}\sigma\sigma^\mathsf{T}Q_-$, is further decomposed as

$$
\begin{aligned}
&\frac{1}{T}\sum_{i=1}^n \mathrm{E}\left[Q\sigma_i\sigma_i^\mathsf{T}\bar{Q}AQ\right] \\
&= \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-\right] - \frac{n}{T^2}\frac{1}{1+\delta}\mathrm{E}\left[\frac{Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-\sigma\sigma^\mathsf{T}Q_-}{1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma}\right] \\
&+ \frac{n}{T}\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-\frac{\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma}{(1+\delta)\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)}\right] \\
&- \frac{n}{T^2}\mathrm{E}\left[\frac{Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-\sigma\sigma^\mathsf{T}Q_-\left(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)}{(1+\delta)\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)^2}\right] \\
&= \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q_-\Phi\bar{Q}AQ_-\right] - \frac{n}{T}\frac{1}{1+\delta}\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}Q_-\frac{\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma}{1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma}\right] \\
&+ \frac{n}{T}\mathrm{E}\left[Q_-\frac{\sigma\sigma^\mathsf{T}\left(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)}{(1+\delta)\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)}\bar{Q}AQ_-\right] \\
&- \frac{n}{T}\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}Q_-\frac{\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma\left(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)}{(1+\delta)\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)^2}\right] \\
&\equiv Z_1 + Z_2 + Z_3 + Z_4
\end{aligned}
$$

(where in the previous to last line, we have merely reorganized the terms conveniently) and our interest is in handling $Z_1 + Z_1^\mathsf{T} + Z_2 + Z_2^\mathsf{T} + Z_3 + Z_3^\mathsf{T} + Z_4 + Z_4^\mathsf{T}$. Let us first treat term Z_2. Since $\bar{Q}AQ_-$ is bounded, by Lemma 4, $\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma$ concentrates around $\frac{1}{T}\operatorname{tr}\Phi\bar{Q}AE[Q_-]$; but, as $\|\Phi\bar{Q}\|$ is bounded, we also have $|\frac{1}{T}\operatorname{tr}\Phi\bar{Q}AE[Q_-] - \frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}| \leq cn^{\varepsilon-\frac{1}{2}}$. We thus deduce, with similar arguments as previously, that

$$
\begin{aligned}
-Q_-\sigma\sigma^\mathsf{T}Q_-Cn^{\varepsilon-\frac{1}{2}} &\preceq Q_-\sigma\sigma^\mathsf{T}Q_-\left[\frac{\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma}{1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma} - \frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{1+\delta}\right] \\
&\preceq Q_-\sigma\sigma^\mathsf{T}Q_-Cn^{\varepsilon-\frac{1}{2}}
\end{aligned}
$$

with probability exponentially close to one, in the order of symmetric matrices. Taking expectation and norms on both sides, and conditioning on the

aforementioned event and its complementary, we thus have that

$$\begin{aligned}&\left\|\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}Q_-\frac{\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma}{1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma}\right]-\mathrm{E}\left[Q_-\Phi Q_-\right]\frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{1+\delta}\right\|\\&\leq\|\mathrm{E}\left[Q_-\Phi Q_-\right]\|\,Cn^{\varepsilon-\frac{1}{2}}+C'ne^{-cn^{\varepsilon'}}\\&\leq\|\mathrm{E}\left[Q_-\Phi Q_-\right]\|\,C''n^{\varepsilon-\frac{1}{2}}\end{aligned}$$

But, again by exchangeability arguments,

$$\begin{aligned}\mathrm{E}[Q_-\Phi Q_-]=\mathrm{E}[Q_-\sigma\sigma^\mathsf{T}Q_-]&=\mathrm{E}\left[Q\sigma\sigma^\mathsf{T}Q\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)^2\right]\\&=\frac{T}{n}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D^2\Sigma Q\right]\end{aligned}$$

with $D=\operatorname{diag}(\{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_-\sigma_i\})$, the operator norm of which is bounded as $\mathrm{O}(1)$. So finally,

$$\left\|\mathrm{E}\left[Q_-\sigma\sigma^\mathsf{T}Q_-\frac{\frac{1}{T}\sigma^\mathsf{T}\bar{Q}AQ_-\sigma}{1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma}\right]-\mathrm{E}\left[Q_-\Phi Q_-\right]\frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{1+\delta}\right\|\leq Cn^{\varepsilon-\frac{1}{2}}.$$

We now move to term $Z_3+Z_3^\mathsf{T}$. Using the relation $ab^\mathsf{T}+ba^\mathsf{T}\preceq aa^\mathsf{T}+bb^\mathsf{T}$,

$$\begin{aligned}&\mathrm{E}\left[(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma)\frac{Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-+Q_-A\bar{Q}\sigma\sigma^\mathsf{T}Q_-}{(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma)^2}\right]\\&\preceq\sqrt{n}\mathrm{E}\left[\frac{(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma)^2}{(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma)^4}Q_-\sigma\sigma^\mathsf{T}Q_-\right]+\frac{1}{\sqrt{n}}\mathrm{E}\left[Q_-A\bar{Q}\sigma\sigma^\mathsf{T}\bar{Q}AQ_-\right]\\&=\sqrt{n}\frac{T}{n}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D_3^2\Sigma Q\right]+\frac{1}{\sqrt{n}}\mathrm{E}\left[Q_-A\bar{Q}\Phi\bar{Q}AQ_-\right]\end{aligned}$$

and the symmetrical lower bound (equal to the opposite of the upper bound), where $D_3=\operatorname{diag}((\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)/(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i))$. For the same reasons as above, the first right-hand side term is bounded by $Cn^{\varepsilon-\frac{1}{2}}$. As for the second term, for $A=I_T$, it is clearly bounded; for $A=\Phi$, using $\frac{n}{T}\frac{\bar{Q}\Phi}{1+\delta}=I_T-\gamma\bar{Q}$, $\mathrm{E}[Q_-A\bar{Q}\Phi\bar{Q}AQ_-]$ can be expressed in terms of $\mathrm{E}[Q_-\Phi Q_-]$ and $\mathrm{E}[Q_-\bar{Q}^k\Phi Q_-]$ for $k=1,2$, all of which have been shown to be bounded (at most by Cn^ε). We thus conclude that

$$\left\|\mathrm{E}\left[\left(\delta-\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)\frac{Q_-\sigma\sigma^\mathsf{T}\bar{Q}AQ_-+Q_-A\bar{Q}\sigma\sigma^\mathsf{T}Q_-}{(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma)^2}\right]\right\|\leq Cn^{\varepsilon-\frac{1}{2}}.$$

Finally, term Z_4 can be handled similarly as term Z_2 and is shown to be of norm bounded by $Cn^{\varepsilon-\frac{1}{2}}$.

As a consequence of all the above, we thus find that

$$\begin{aligned}\mathrm{E}\left[QAQ\right] = \bar{Q}A\bar{Q} + \frac{n}{T}\frac{\mathrm{E}\left[Q\Phi\bar{Q}AQ\right]}{1+\delta} - \frac{n}{T}\frac{\mathrm{E}\left[Q_-\Phi\bar{Q}AQ_-\right]}{1+\delta} \\ + \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{(1+\delta)^2}\mathrm{E}[Q_-\Phi Q_-] + O(n^{\varepsilon-\frac{1}{2}}).\end{aligned}$$

It is attractive to feel that the sum of the second and third terms above vanishes. This is indeed verified by observing that, for any matrix B,

$$\begin{aligned}\mathrm{E}\left[QBQ\right] - \mathrm{E}\left[Q_-BQ\right] &= \frac{1}{T}\mathrm{E}\left[Q\sigma\sigma^\mathsf{T}Q_-BQ\right] \\ &= \frac{1}{T}\mathrm{E}\left[Q\sigma\sigma^\mathsf{T}QBQ\left(1+\frac{1}{T}\sigma^\mathsf{T}Q_-\sigma\right)\right] \\ &= \frac{1}{n}\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma QBQ\right]\end{aligned}$$

and symmetrically

$$\mathrm{E}\left[QBQ\right] - \mathrm{E}\left[QBQ_-\right] = \frac{1}{n}\mathrm{E}\left[QBQ\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\right]$$

with $D = \operatorname{diag}(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)$, and a similar reasoning is performed to control $\mathrm{E}[Q_-BQ]-\mathrm{E}[Q_-BQ_-]$ and $\mathrm{E}[QBQ_-]-\mathrm{E}[Q_-BQ_-]$. For B bounded, $\|\mathrm{E}[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma QBQ]\|$ is bounded as $O(1)$, and thus $\|\mathrm{E}[QBQ]-\mathrm{E}[Q_-BQ_-]\|$ is of order $O(n^{-1})$. So in particular, taking A of bounded norm, we find that

$$\mathrm{E}\left[QAQ\right] = \bar{Q}A\bar{Q} + \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{(1+\delta)^2}\mathrm{E}[Q_-\Phi Q_-] + O(n^{\varepsilon-\frac{1}{2}}).$$

Take now $B=\Phi$. Then, from the relation $AB^\mathsf{T}+BA^\mathsf{T} \preceq AA^\mathsf{T}+BB^\mathsf{T}$ in the order of symmetric matrices,

$$\begin{aligned}&\left\|\mathrm{E}\left[Q\Phi Q\right] - \frac{1}{2}\mathrm{E}\left[Q_-\Phi Q + Q\Phi Q_-\right]\right\| \\ &= \frac{1}{2n}\left\|\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\Phi Q + Q\Phi Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\right]\right\| \\ &\leq \frac{1}{2n}\left(\left\|\mathrm{E}\left[Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\frac{1}{T}\Sigma^\mathsf{T}D\Sigma Q\right]\right\| + \left\|\mathrm{E}\left[Q\Phi Q\Phi Q\right]\right\|\right).\end{aligned}$$

The first norm in the parenthesis is bounded by Cn^ε and it thus remains to control the second norm. To this end, similar to the control of $\mathrm{E}[Q\Phi Q]$, by writing $\mathrm{E}[Q\Phi Q\Phi Q] = \mathrm{E}[Q\sigma_1\sigma_1^\mathsf{T} Q\sigma_2\sigma_2^\mathsf{T} Q]$ for σ_1, σ_2 independent vectors with the same law as σ, and exploiting the exchangeability, we obtain after some calculus that $\mathrm{E}[Q\Phi Q]$ can be expressed as the sum of terms of the form $\mathrm{E}[Q_{++}\frac{1}{T}\Sigma_{++}^\mathsf{T} D\Sigma_{++}Q_{++}]$ or $\mathrm{E}[Q_{++}\frac{1}{T}\Sigma_{++}^\mathsf{T} D\Sigma_{++}Q_{++}\frac{1}{T}\Sigma_{++}^\mathsf{T} D_2\Sigma_{++}Q_{++}]$ for D, D_2 diagonal matrices of norm bounded as $O(1)$, while Σ_{++} and Q_{++} are similar as Σ and Q, only for n replaced by $n+2$. All these terms are bounded as $O(1)$ and we finally obtain that $\mathrm{E}[Q\Phi Q\Phi Q]$ is bounded and thus

$$\left\| \mathrm{E}\left[Q\Phi Q\right] - \frac{1}{2}\mathrm{E}\left[Q_-\Phi Q + Q\Phi Q_-\right] \right\| \leq \frac{C}{n}.$$

With the additional control on $Q\Phi Q_- - Q_-\Phi Q_-$ and $Q_-\Phi Q - Q_-\Phi Q_-$, together, this implies that $\mathrm{E}[Q\Phi Q] = \mathrm{E}[Q_-\Phi Q_-] + O_{\|\cdot\|}(n^{-1})$. Hence, for $A = \Phi$, exploiting the fact that $\frac{n}{T}\frac{1}{1+\delta}\Phi\bar{Q}\Phi = \Phi - \gamma\bar{Q}\Phi$, we have the simplification

$$\begin{aligned}
\mathrm{E}\left[Q\Phi Q\right] &= \bar{Q}\Phi\bar{Q} + \frac{n}{T}\frac{\mathrm{E}\left[Q\Phi\bar{Q}\Phi Q\right]}{1+\delta} - \frac{n}{T}\frac{\mathrm{E}\left[Q_-\Phi\bar{Q}\Phi Q_-\right]}{1+\delta} \\
&\quad + \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2}\mathrm{E}[Q_-\Phi Q_-] + O_{\|\cdot\|}(n^{\varepsilon-\frac{1}{2}}) \\
&= \bar{Q}\Phi\bar{Q} + \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2}\mathrm{E}[Q\Phi Q] + O_{\|\cdot\|}(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}$$

or equivalently

$$\mathrm{E}\left[Q\Phi Q\right]\left(1 - \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2}\right) = \bar{Q}\Phi\bar{Q} + O_{\|\cdot\|}(n^{\varepsilon-\frac{1}{2}}).$$

We have already shown in (11) that $\limsup_n \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2} < 1$ and thus

$$\mathrm{E}\left[Q\Phi Q\right] = \frac{\bar{Q}\Phi\bar{Q}}{1 - \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2}} + O_{\|\cdot\|}(n^{\varepsilon-\frac{1}{2}}).$$

So finally, for all A of bounded norm,

$$\mathrm{E}\left[QAQ\right] = \bar{Q}A\bar{Q} + \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi\bar{Q}A\bar{Q}}{(1+\delta)^2}\frac{\bar{Q}\Phi\bar{Q}}{1 - \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi^2\bar{Q}^2}{(1+\delta)^2}} + O(n^{\varepsilon-\frac{1}{2}})$$

which proves immediately Proposition 1 and Theorem 3.

5.3. *Derivation of* Φ_{ab}.

5.3.1. *Gaussian* w. In this section, we evaluate the terms Φ_{ab} provided in Table 1. The proof for the term corresponding to $\sigma(t) = \mathrm{erf}(t)$ can be already be found in (Williams, 1998, Section 3.1) and is not recalled here. For the other functions $\sigma(\cdot)$, we follow a similar approach as in (Williams, 1998), as detailed next.

The evaluation of Φ_{ab} for $w \sim \mathcal{N}(0, I_p)$ requires to estimate

$$\mathcal{I} \equiv (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(w^\mathsf{T} a)\sigma(w^\mathsf{T} b) e^{-\frac{1}{2}\|w\|^2} dw.$$

Assume that a and b and not linearly dependent. It is convenient to observe that this integral can be reduced to a two-dimensional integration by considering the basis $e_1, \ldots, e_p$ defined (for instance) by

$$e_1 = \frac{a}{\|a\|}, \quad e_2 = \frac{\frac{b}{\|b\|} - \frac{a^\mathsf{T} b}{\|a\|\|b\|}\frac{a}{\|a\|}}{\sqrt{1 - \frac{(a^\mathsf{T} b)^2}{\|a\|^2\|b\|^2}}}$$

and $e_3, \ldots, e_p$ any completion of the basis. By letting $w = \tilde{w}_1 e_1 + \ldots + \tilde{w}_p e_p$ and $a = \tilde{a}_1 e_1$ ($\tilde{a}_1 = \|a\|$), $b = \tilde{b}_1 e_1 + \tilde{b}_2 e_2$ (where $\tilde{b}_1 = \frac{a^\mathsf{T} b}{\|a\|}$ and $\tilde{b}_2 = \|b\|\sqrt{1 - \frac{(a^\mathsf{T} b)^2}{\|a\|^2\|b\|^2}}$), this reduces $\mathcal{I}$ to

$$\mathcal{I} = \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \sigma(\tilde{w}_1 \tilde{a}_1)\sigma(\tilde{w}_1 \tilde{b}_1 + \tilde{w}_2 \tilde{b}_2) e^{-\frac{1}{2}(\tilde{w}_1^2 + \tilde{w}_2^2)} d\tilde{w}_1 d\tilde{w}_2.$$

Letting $\tilde{w} = [\tilde{w}_1, \tilde{w}_2]^\mathsf{T}$, $\tilde{a} = [\tilde{a}_1, 0]^\mathsf{T}$ and $\tilde{b} = [\tilde{b}_1, \tilde{b}_2]^\mathsf{T}$, this is conveniently written as the two-dimensional integral

$$\mathcal{I} = \frac{1}{2\pi} \int_{\mathbb{R}^2} \sigma(\tilde{w}^\mathsf{T} \tilde{a})\sigma(\tilde{w}^\mathsf{T} \tilde{b}) e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w}.$$

The case where a and b would be linearly dependent can then be obtained by continuity arguments.

The function $\sigma(t) = \max(t, 0)$. For this function, we have

$$\mathcal{I} = \frac{1}{2\pi} \int_{\min(\tilde{w}^\mathsf{T}\tilde{a}, \tilde{w}^\mathsf{T}\tilde{b}) \geq 0} \tilde{w}^\mathsf{T} \tilde{a} \cdot \tilde{w}^\mathsf{T} \tilde{b} \cdot e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w}.$$

Since $\tilde{a} = \tilde{a}_1 e_1$, a simple geometric representation lets us observe that

$$\left\{\tilde{w} \mid \min(\tilde{w}^\mathsf{T}\tilde{a}, \tilde{w}^\mathsf{T}\tilde{b}) \geq 0\right\} = \left\{r\cos(\theta)e_1 + r\sin(\theta)e_2 \mid r \geq 0,\ \theta \in [\theta_0 - \frac{\pi}{2}, \frac{\pi}{2}]\right\}$$

where we defined $\theta_0 \equiv \arccos\left(\frac{\tilde{b}_1}{\|\tilde{b}\|}\right) = -\arcsin\left(\frac{\tilde{b}_1}{\|\tilde{b}\|}\right) + \frac{\pi}{2}$. We may thus operate a polar coordinate change of variable (with inverse Jacobian determinant equal to r) to obtain

$$\begin{aligned}\mathcal{I} &= \frac{1}{2\pi}\int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{\mathbb{R}^+}(r\cos(\theta)\tilde{a}_1)\left(r\cos(\theta)\tilde{b}_1 + r\sin(\theta)\tilde{b}_2\right) re^{-\frac{1}{2}r^2}d\theta dr \\ &= \tilde{a}_1\frac{1}{2\pi}\int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos(\theta)\left(\cos(\theta)\tilde{b}_1 + \sin(\theta)\tilde{b}_2\right)d\theta\int_{\mathbb{R}^+} r^3e^{-\frac{1}{2}r^2}dr.\end{aligned}$$

With two integration by parts, we have that $\int_{\mathbb{R}^+} r^3e^{-\frac{1}{2}r^2}dr = 2$. Classical trigonometric formulas also provide

$$\begin{aligned}\int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos(\theta)^2d\theta &= \frac{1}{2}(\pi-\theta_0) + \frac{1}{2}\sin(2\theta_0) \\ &= \frac{1}{2}\left(\pi - \arccos\left(\frac{\tilde{b}_1}{\|\tilde{b}\|}\right) + \frac{\tilde{b}_1}{\|\tilde{b}\|}\frac{\tilde{b}_2}{\|\tilde{b}\|}\right) \\ \int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos(\theta)\sin(\theta)d\theta &= \frac{1}{2}\sin^2(\theta_0) = \frac{1}{2}\left(\frac{\tilde{b}_2}{\|\tilde{b}\|}\right)^2\end{aligned}$$

where we used in particular $\sin(2\arccos(x)) = 2x\sqrt{1-x^2}$. Altogether, this is after simplification and replacement of $\tilde{a}_1$, $\tilde{b}_1$ and $\tilde{b}_2$,

$$\mathcal{I} = \frac{1}{2\pi}\|a\|\|b\|\left(\sqrt{1-\angle(a,b)^2} + \angle(a,b)\arccos(-\angle(a,b))\right).$$

It is worth noticing that this may be more compactly written as

$$\mathcal{I} = \frac{1}{2\pi}\|a\|\|b\|\int_{-1}^{\angle(a,b)}\arccos(-x)dx.$$

which is minimum for $\angle(a,b) \to -1$ (since $\arccos(-x) \geq 0$ on $[-1,1]$) and takes there the limiting value zero. Hence $\mathcal{I} > 0$ for a and b not linearly dependent.

For a and b linearly dependent, we simply have $\mathcal{I} = 0$ for $\angle(a,b) = -1$ and $\mathcal{I} = \frac{1}{2}\|a\|\|b\|$ for $\angle(a,b) = 1$.

The function $\sigma(t) = |t|$. Since $|t| = \max(t,0) + \max(-t,0)$, we have

$$\begin{aligned}|w^\mathsf{T}a|\cdot|w^\mathsf{T}b| &= \max(w^\mathsf{T}a,0)\max(w^\mathsf{T}b,0) + \max(w^\mathsf{T}(-a),0)\max(w^\mathsf{T}(-b),0) \\ &\quad + \max(w^\mathsf{T}(-a),0)\max(w^\mathsf{T}b,0) + \max(w^\mathsf{T}a,0)\max(w^\mathsf{T}(-b),0).\end{aligned}$$

Hence, reusing the results above, we have here

$$\mathcal{I} = \frac{\|a\|\|b\|}{2\pi}\left(4\sqrt{1-\angle(a,b)^2} + 2\angle(a,b)\operatorname{acos}(-\angle(a,b)) - 2\angle(a,b)\operatorname{acos}(\angle(a,b))\right).$$

Using the identity $\operatorname{acos}(-x) - \operatorname{acos}(x) = 2\operatorname{asin}(x)$ provides the expected result.

The function $\sigma(t) = 1_{t\geq 0}$. With the same notations as in the case $\sigma(t) = \max(t, 0)$, we have to evaluate

$$\mathcal{I} = \frac{1}{2\pi}\int_{\min(\tilde{w}^\mathsf{T}\tilde{a},\tilde{w}^\mathsf{T}\tilde{b})\geq 0} e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w}.$$

After a polar coordinate change of variable, this is

$$\mathcal{I} = \frac{1}{2\pi}\int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta \int_{\mathbb{R}^+} re^{-\frac{1}{2}r^2} dr = \frac{1}{2} - \frac{\theta_0}{2\pi}$$

from which the result unfolds.

The function $\sigma(t) = \operatorname{sign}(t)$. Here it suffices to note that $\operatorname{sign}(t) = 1_{t\geq 0} - 1_{-t\geq 0}$ so that

$$\begin{aligned}\sigma(w^\mathsf{T}a)\sigma(w^\mathsf{T}b) = {} & 1_{w^\mathsf{T}a\geq 0}1_{w^\mathsf{T}b\geq 0} + 1_{w^\mathsf{T}(-a)\geq 0}1_{w^\mathsf{T}(-b)\geq 0} \\ & - 1_{w^\mathsf{T}(-a)\geq 0}1_{w^\mathsf{T}b\geq 0} - 1_{w^\mathsf{T}a\geq 0}1_{w^\mathsf{T}(-b)\geq 0}\end{aligned}$$

and to apply the result of the previous section, with either (a, b), $(-a, b)$, $(a, -b)$ or $(-a, -b)$. Since $\arccos(-x) = -\arccos(x) + \pi$, we conclude that

$$\mathcal{I} = (2\pi)^{-\frac{p}{2}}\int_{\mathbb{R}^p} \operatorname{sign}(w^\mathsf{T}a)\operatorname{sign}(w^\mathsf{T}b)e^{-\frac{1}{2}\|w\|^2} dw = 1 - \frac{2\theta_0}{\pi}.$$

The functions $\sigma(t) = \cos(t)$ and $\sigma(t) = \sin(t)$.. Let us first consider $\sigma(t) = \cos(t)$. We have here to evaluate

$$\begin{aligned}\mathcal{I} &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \cos\left(\tilde{w}^\mathsf{T}\tilde{a}\right)\cos\left(\tilde{w}^\mathsf{T}\tilde{b}\right) e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w} \\ &= \frac{1}{8\pi}\int_{\mathbb{R}^2} \left(e^{\imath\tilde{w}^\mathsf{T}\tilde{a}} + e^{-\imath\tilde{w}^\mathsf{T}\tilde{a}}\right)\left(e^{\imath\tilde{w}^\mathsf{T}\tilde{b}} + e^{-\imath\tilde{w}^\mathsf{T}\tilde{b}}\right) e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w}\end{aligned}$$

which boils down to evaluating, for $d \in \{\tilde{a} + \tilde{b}, \tilde{a} - \tilde{b}, -\tilde{a} + \tilde{b}, -\tilde{a} - \tilde{b}\}$, the integral

$$e^{-\frac{1}{2}\|d\|^2}\int_{\mathbb{R}^2} e^{-\frac{1}{2}\|\tilde{w}-\imath d\|^2} d\tilde{w} = (2\pi)e^{-\frac{1}{2}\|d\|^2}.$$

Altogether, we find

$$\mathcal{I} = \frac{1}{2}\left(e^{-\frac{1}{2}\|a+b\|^2} + e^{-\frac{1}{2}\|a-b\|^2}\right) = e^{-\frac{1}{2}(\|a\|+\|b\|^2)}\cosh(a^\mathsf{T} b).$$

For $\sigma(t) = \sin(t)$, it suffices to appropriately adapt the signs in the expression of $\mathcal{I}$ (using the relation $\sin(t) = \frac{1}{2i}(e^t + e^{-t})$) to obtain in the end

$$\mathcal{I} = \frac{1}{2}\left(e^{-\frac{1}{2}\|a+b\|^2} + e^{-\frac{1}{2}\|a-b\|^2}\right) = e^{-\frac{1}{2}(\|a\|+\|b\|^2)}\sinh(a^\mathsf{T} b)$$

as desired.

5.4. *Polynomial $\sigma(\cdot)$ and generic w.* In this section, we prove Equation 5 for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $w \in \mathbb{R}^p$ a random vector with independent and identically distributed entries of zero mean and moment of order k equal to m_k. The result is based on standard combinatorics. We are to evaluate

$$\Phi_{ab} = \mathrm{E}\left[\left(\zeta_2(w^\mathsf{T} a)^2 + \zeta_1 w^\mathsf{T} a + \zeta_0\right)\left(\zeta_2(w^\mathsf{T} b)^2 + \zeta_1 w^\mathsf{T} b + \zeta_0\right)\right].$$

After development, it appears that one needs only assess, for say vectors $c, d \in \mathbb{R}^p$ that take values in $\{a, b\}$, the moments

$$\begin{aligned}
\mathrm{E}[(w^\mathsf{T} c)^2 (w^\mathsf{T} d)^2] &= \sum_{i_1 i_2 j_1 j_2} c_{i_1} c_{i_2} d_{j_1} d_{j_2} \mathrm{E}[w_{i_1} w_{i_2} w_{j_1} w_{j_2}] \\
&= \sum_{i_1} m_4 c_{i_1}^2 d_{i_1}^2 + \sum_{i_1 \neq j_1} m_2^2 c_{i_1}^2 d_{j_1}^2 + 2 \sum_{i_1 \neq i_2} m_2^2 c_{i_1} d_{i_1} c_{i_2} d_{i_2} \\
&= \sum_{i_1} m_4 c_{i_1}^2 d_{i_1}^2 + \left(\sum_{i_1 j_1} - \sum_{i_1 = j_1}\right) m_2^2 c_{i_1}^2 d_{j_1}^2 \\
&\quad + 2\left(\sum_{i_1 i_2} - \sum_{i_1 \neq i_2}\right) m_2^2 c_{i_1} d_{i_1} c_{i_2} d_{i_2} \\
&= m_4 (c^2)^\mathsf{T} (d^2) + m_2^2 (\|c\|^2 \|d\|^2 - (c^2)^\mathsf{T} (d^2)) \\
&\quad + 2 m_2^2 \left((c^\mathsf{T} d)^2 - (c^2)^\mathsf{T} (d^2)\right) \\
&= (m_4 - 3m_2^2)(c^2)^\mathsf{T} (d^2) + m_2^2 \left(\|c\|^2 \|d\|^2 + 2(c^\mathsf{T} d)^2\right) \\
\mathrm{E}\left[(w^\mathsf{T} c)^2 (w^\mathsf{T} d)\right] &= \sum_{i_1 i_2 j} c_{i_1} c_{i_2} d_j \mathrm{E}[w_{i_1} w_{i_2} w_j] = \sum_{i_1} m_3 c_{i_1}^2 d_{i_1} = m_3 (c^2) d \\
\mathrm{E}\left[(w^\mathsf{T} c)^2\right] &= \sum_{i_1 i_2} c_{i_1} c_{i_2} \mathrm{E}[w_{i_1} w_{i_2}] = m_2 \|c\|^2
\end{aligned}$$

where we recall the definition $(a^2) = [a_1^2, \ldots, a_p^2]^\mathsf{T}$. Gathering all the terms for appropriate selections of c, d leads to (5).

5.5. *Heuristic derivation of Conjecture 1.* Conjecture 1 essentially follows as an aftermath of Remark 1. We believe that, similar to Σ, $\hat{\Sigma}$ is expected to be of the form $\hat{\Sigma} = \hat{\Sigma}^\circ + \hat{\bar{\sigma}} 1_{\hat{T}}^\mathsf{T}$, where $\hat{\bar{\sigma}} = \mathrm{E}[\sigma(w^\mathsf{T}\hat{X})]^\mathsf{T}$, with $\|\frac{\hat{\Sigma}^\circ}{\sqrt{T}}\| \leq n^\varepsilon$ with high probability. Besides, if $X, \hat{X}$ were chosen as constituted of Gaussian mixture vectors, with non-trivial growth rate conditions as introduced in (Couillet and Benaych-Georges, 2016), it is easily seen that $\bar{\sigma} = c1_p + v$ and $\hat{\bar{\sigma}} = c1_p + \hat{v}$, for some constant c and $\|v\|, \|\hat{v}\| = O(1)$.

This subsequently ensures that $\Phi_{X\hat{X}}$ and $\Phi_{\hat{X}\hat{X}}$ would be of a similar form $\Phi^\circ_{X\hat{X}} + \bar{\sigma}\hat{\bar{\sigma}}^\mathsf{T}$ and $\Phi^\circ_{\hat{X}\hat{X}} + \hat{\bar{\sigma}}\hat{\bar{\sigma}}^\mathsf{T}$ with $\Phi^\circ_{X\hat{X}}$ and $\Phi^\circ_{\hat{X}\hat{X}}$ of bounded norm. These facts, that would require more advanced proof techniques, let envision the following heuristic derivation for Conjecture 1.

Recall that our interest is on the test performance E_{test} defined as

$$E_{\text{test}} = \frac{1}{\hat{T}} \left\| \hat{Y}^\mathsf{T} - \hat{\Sigma}^\mathsf{T}\beta \right\|_F^2$$

which may be rewritten as

$$\begin{aligned} E_{\text{test}} &= \frac{1}{\hat{T}} \operatorname{tr}\left(\hat{Y}\hat{Y}^\mathsf{T}\right) - \frac{2}{T\hat{T}} \operatorname{tr}\left(YQ\Sigma^\mathsf{T}\hat{\Sigma}\hat{Y}^\mathsf{T}\right) + \frac{1}{T^2\hat{T}} \operatorname{tr}\left(YQ\Sigma^\mathsf{T}\hat{\Sigma}\hat{\Sigma}^\mathsf{T}\Sigma QY^\mathsf{T}\right) \\ &\equiv Z_1 - Z_2 + Z_3. \end{aligned} \tag{12}$$

If $\hat{\Sigma} = \hat{\Sigma}^\circ + \hat{\bar{\sigma}} 1_{\hat{T}}^\mathsf{T}$ follows the aforementioned claimed operator norm control, reproducing the steps of Corollary 3 leads to a similar concentration for E_{test}, which we shall then admit. We are therefore left to evaluating $\mathrm{E}[Z_2]$ and $\mathrm{E}[Z_3]$.

We start with the term $\mathrm{E}[Z_2]$, which we expand as

$$\begin{aligned}
\mathrm{E}[Z_2] &= \frac{2}{T\hat{T}}\mathrm{E}\left[\mathrm{tr}(YQ\Sigma^\mathsf{T}\hat{\Sigma}\hat{Y}^\mathsf{T})\right] = \frac{2}{T\hat{T}}\sum_{i=1}^n\left[\mathrm{tr}(YQ\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{Y}^\mathsf{T})\right] \\
&= \frac{2}{T\hat{T}}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(\frac{YQ_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{Y}^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}\right)\right] \\
&= \frac{2}{T\hat{T}}\frac{1}{1+\delta}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(YQ_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{Y}^\mathsf{T}\right)\right] \\
&+ \frac{2}{T\hat{T}}\frac{1}{1+\delta}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(YQ_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{Y}^\mathsf{T}\right)\frac{\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}\right] \\
&= \frac{2n}{T\hat{T}}\frac{1}{1+\delta}\mathrm{tr}\left(Y\mathrm{E}[Q_-]\Phi_{X\hat{X}}\hat{Y}^\mathsf{T}\right) + \frac{2}{T\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ\Sigma^\mathsf{T}D\hat{\Sigma}\hat{Y}^\mathsf{T}\right)\right] \\
&\equiv Z_{21} + Z_{22}
\end{aligned}$$

with $D = \mathrm{diag}(\{\delta - \frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i\})$, the operator norm of which is bounded by $n^{\varepsilon-\frac{1}{2}}$ with high probability. Now, observe that, again with the assumption that $\hat{\Sigma} = \hat{\Sigma}^\circ + \bar{\sigma}1_{\hat{T}}^\mathsf{T}$ with controlled $\hat{\Sigma}^\circ$, Z_{22} may be decomposed as

$$\begin{aligned}
\frac{2}{T\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ\Sigma^\mathsf{T}D\hat{\Sigma}\hat{Y}^\mathsf{T}\right)\right] &= \frac{2}{T\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ\Sigma^\mathsf{T}D\hat{\Sigma}^\circ\hat{Y}^\mathsf{T}\right)\right] \\
&+ \frac{2}{T\hat{T}}\frac{1}{1+\delta}1_{\hat{T}}^\mathsf{T}\hat{Y}^\mathsf{T}\mathrm{E}\left[YQ\Sigma^\mathsf{T}D\bar{\sigma}\right].
\end{aligned}$$

In the display above, the first right-hand side term is now of order $O(n^{\varepsilon-\frac{1}{2}})$. As for the second right-hand side term, note that $D\bar{\sigma}$ is a vector of independent and identically distributed zero mean and variance $O(n^{-1})$ entries; while note formally independent of $YQ\Sigma^\mathsf{T}$, it is nonetheless expected that this independence "weakens" asymptotically (a behavior several times observed in linear random matrix models), so that one expects by central limit arguments that the second right-hand side term be also of order $O(n^{\varepsilon-\frac{1}{2}})$.

This would thus result in

$$\begin{aligned}
\mathrm{E}[Z_2] &= \frac{2n}{T\hat{T}}\frac{1}{1+\delta}\mathrm{tr}\left(Y\mathrm{E}[Q_-]\Phi_{X\hat{X}}\hat{Y}^\mathsf{T}\right) + O(n^{\varepsilon-\frac{1}{2}}) \\
&= \frac{2n}{T\hat{T}}\frac{1}{1+\delta}\mathrm{tr}\left(Y\bar{Q}\Phi_{X\hat{X}}\hat{Y}^\mathsf{T}\right) + O(n^{\varepsilon-\frac{1}{2}}) \\
&= \frac{2}{\hat{T}}\mathrm{tr}\left(Y\bar{Q}\Psi_{X\hat{X}}\hat{Y}^\mathsf{T}\right) + O(n^{\varepsilon-\frac{1}{2}})
\end{aligned}$$

where we used $\|\mathrm{E}[Q_-] - \bar{Q}\| \leq Cn^{\varepsilon-\frac{1}{2}}$ and the definition $\Psi_{X\hat{X}} = \frac{n}{T}\frac{\Phi_{X\hat{X}}}{1+\delta}$.

We then move on to $\mathrm{E}[Z_3]$ of Equation (12), which can be developed as

$$\begin{aligned}
\mathrm{E}[Z_3] &= \frac{1}{T^2\hat{T}}\mathrm{E}\left[\mathrm{tr}\left(YQ\Sigma^\mathsf{T}\hat{\Sigma}\hat{\Sigma}^\mathsf{T}\Sigma QY^\mathsf{T}\right)\right] \\
&= \frac{1}{T^2\hat{T}}\sum_{i,j=1}^n \mathrm{E}\left[\mathrm{tr}\left(YQ\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{\sigma}_j\sigma_j^\mathsf{T}QY^\mathsf{T}\right)\right] \\
&= \frac{1}{T^2\hat{T}}\sum_{i,j=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}\frac{\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right] \\
&= \frac{1}{T^2\hat{T}}\sum_{i=1}^n\sum_{j\neq i} \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i}\frac{\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right] \\
&+ \frac{1}{T^2\hat{T}}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\hat{\sigma}_i^\mathsf{T}\hat{\sigma}_i\sigma_i^\mathsf{T}Q_{-i}}{(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)^2}Y^\mathsf{T}\right)\right] \equiv Z_{31} + Z_{32}.
\end{aligned}$$

In the term Z_{32}, reproducing the proof of Lemma 1 with the condition $\|\hat{X}\|$ bounded, we obtain that $\frac{\hat{\sigma}_i^\mathsf{T}\hat{\sigma}_i}{\hat{T}}$ concentrates around $\frac{1}{\hat{T}}\mathrm{tr}\,\Phi_{\hat{X}\hat{X}}$, which allows us to write

$$\begin{aligned}
Z_{32} &= \frac{1}{T^2\hat{T}}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\,\mathrm{tr}(\Phi_{\hat{X}\hat{X}})\sigma_i^\mathsf{T}Q_{-i}}{(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)^2}Y^\mathsf{T}\right)\right] \\
&+ \frac{1}{T^2\hat{T}}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\left(\hat{\sigma}_i^\mathsf{T}\hat{\sigma}_i - \mathrm{tr}\,\Phi_{\hat{T}}\right)\sigma_i^\mathsf{T}Q_{-i}}{(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)^2}Y^\mathsf{T}\right)\right] \\
&= \frac{1}{T^2}\frac{\mathrm{tr}(\Phi_{\hat{X}\hat{X}})}{\hat{T}}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-i}\sigma_i\sigma_i^\mathsf{T}Q_{-i}}{(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-i}\sigma_i)^2}Y^\mathsf{T}\right)\right] \\
&+ \frac{1}{T^2}\sum_{i=1}^n \mathrm{E}\left[\mathrm{tr}\left(YQ\sigma_i\left(\frac{\hat{\sigma}_i^\mathsf{T}\hat{\sigma}_i - \mathrm{tr}\,\Phi_{\hat{T}}}{\hat{T}}\right)\sigma_i^\mathsf{T}QY^\mathsf{T}\right)\right] \\
&\equiv Z_{321} + Z_{322}
\end{aligned}$$

with $D = \mathrm{diag}(\{\frac{1}{\hat{T}}\sigma_i^\mathsf{T}\hat{\sigma}_i - \frac{1}{\hat{T}}\mathrm{tr}\,\Phi_{\hat{T}\hat{T}}\}_{i=1}^n)$ and thus Z_{322} can be rewritten as

$$Z_{322} = \frac{1}{T}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q\Sigma^\mathsf{T}}{\sqrt{T}}D\frac{\Sigma Q}{\sqrt{T}}Y^\mathsf{T}\right)\right] = O(n^{\varepsilon-\frac{1}{2}})$$

while for Z_{321}, following the same arguments as previously, we have

$$\begin{aligned}
Z_{321} &= \frac{1}{T^2}\frac{\operatorname{tr}\Phi_{\hat{X}\hat{X}}}{\hat{T}}\sum_{i=1}^{n}\mathrm{E}\left[\operatorname{tr}\left(Y\frac{Q_{-i}\sigma_i\sigma_i^{\mathsf{T}}Q_{-i}}{(1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-i}\sigma_i)^2}Y^{\mathsf{T}}\right)\right]\\
&= \frac{1}{T^2}\frac{\operatorname{tr}\Phi_{\hat{X}\hat{X}}}{\hat{T}}\sum_{i=1}^{n}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\operatorname{tr}\left(YQ_{-i}\sigma_i\sigma_i^{\mathsf{T}}Q_{-i}Y^{\mathsf{T}}\right)\right]\\
&+ \frac{1}{T^2}\frac{\operatorname{tr}\Phi_{\hat{X}\hat{X}}}{\hat{T}}\sum_{i=1}^{n}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\operatorname{tr}\left(YQ\sigma_i\sigma_i^{\mathsf{T}}QY^{\mathsf{T}}\right)\left((1+\delta)^2-(1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-i}\sigma_i)^2\right)\right]\\
&= \frac{1}{T^2}\frac{\operatorname{tr}\Phi_{\hat{X}\hat{X}}}{\hat{T}}\sum_{i=1}^{n}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\operatorname{tr}\left(YQ_{-i}\Phi_XQ_{-i}Y^{\mathsf{T}}\right)\right]\\
&+ \frac{1}{T^2}\frac{\operatorname{tr}\Phi_{\hat{X}\hat{X}}}{\hat{T}}\sum_{i=1}^{n}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\operatorname{tr}\left(YQ\Sigma^{\mathsf{T}}D\Sigma QY^{\mathsf{T}}\right)\right]\\
&= \frac{n}{T^2}\mathrm{E}\left[\operatorname{tr}\left(YQ_-\Phi_XQ_-Y^{\mathsf{T}}\right)\right]\frac{\operatorname{tr}(\Phi_{\hat{X}\hat{X}})}{\hat{T}(1+\delta)^2}+O(n^{\varepsilon-\frac{1}{2}})
\end{aligned}$$

where $D=\operatorname{diag}(\{(1+\delta)^2-(1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-i}\sigma_i)^2\}_{i=1}^n)$.

Since $\mathrm{E}[Q_-AQ_-]=\mathrm{E}[QAQ]+O_{\|\cdot\|}(n^{\varepsilon-\frac{1}{2}})$, we are free to plug in the asymptotic equivalent of $\mathrm{E}[QAQ]$ derived in Section 5.2.3, and we deduce

$$\begin{aligned}
Z_{32} &= \frac{n}{T^2}\mathrm{E}\left[\operatorname{tr}Y\left(\bar{Q}\Phi_X\bar{Q}+\frac{\bar{Q}\Psi_X\bar{Q}\cdot\frac{1}{n}\operatorname{tr}\left(\Psi_X\bar{Q}\Phi_X\bar{Q}\right)}{1-\frac{1}{n}\operatorname{tr}\left(\Psi_X^2\bar{Q}^2\right)}\right)Y^{\mathsf{T}}\right]\frac{\operatorname{tr}(\Phi_{\hat{X}\hat{X}})}{\hat{T}(1+\delta)^2}\\
&= \frac{\frac{1}{n}\operatorname{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^{\mathsf{T}}\right)}{1-\frac{1}{n}\operatorname{tr}\left(\Psi_X^2\bar{Q}^2\right)}\frac{1}{\hat{T}}\operatorname{tr}(\Psi_{\hat{X}\hat{X}})+O(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}$$

The term Z_{31} of the double sum over i and j ($j\neq i$) needs more efforts. To handle this term, we need to remove the dependence of both σ_i and σ_j in Q in sequence. We start with j as follows:

$$\begin{aligned}
Z_{31} &= \frac{1}{T^2\hat{T}}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{E}\left[\operatorname{tr}\left(YQ\sigma_i\hat{\sigma}_i^{\mathsf{T}}\frac{\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q_{-j}}{1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j}Y^{\mathsf{T}}\right)\right]\\
&= \frac{1}{T^2\hat{T}}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{E}\left[\operatorname{tr}\left(YQ_{-j}\sigma_i\hat{\sigma}_i^{\mathsf{T}}\frac{\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q_{-j}}{1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j}Y^{\mathsf{T}}\right)\right]\\
&- \frac{1}{T^3\hat{T}}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{E}\left[\operatorname{tr}\left(Y\frac{Q_{-j}\sigma_j\sigma_j^{\mathsf{T}}Q_{-j}\sigma_i\hat{\sigma}_i^{\mathsf{T}}}{1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j}\frac{\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q_{-j}}{1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j}Y^{\mathsf{T}}\right)\right]\\
&\equiv Z_{311}-Z_{312}
\end{aligned}$$

where in the previous to last inequality we used the relation

$$Q = Q_{-j} - \frac{Q_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}.$$

For Z_{311}, we replace $1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j$ by $1+\delta$ and take expectation over w_j

$$\begin{aligned}
Z_{311} &= \frac{1}{T^2\hat{T}}\sum_{i=1}^n\sum_{j\neq i}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\sigma_i\hat{\sigma}_i^\mathsf{T}\frac{\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right]\\
&= \frac{1}{T^2\hat{T}}\sum_{j=1}^n\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right]\\
&= \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^n\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}Y^\mathsf{T}\right)\right]\\
&+ \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^n\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}(\delta-\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j)}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right]\\
&\equiv Z_{3111}+Z_{3112}.
\end{aligned}$$

The idea to handle Z_{3112} is to retrieve forms of the type $\sum_{j=1}^n d_j\hat{\sigma}_j\sigma_j^\mathsf{T} = \hat{\Sigma}^\mathsf{T}D\Sigma$ for some D satisfying $\|D\| \leq n^{\varepsilon-\frac{1}{2}}$ with high probability. To this end, we use

$$\begin{aligned}
Q_{-j}\frac{\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}}{T} &= Q_{-j}\frac{\Sigma^\mathsf{T}\hat{\Sigma}}{T} - Q_{-j}\frac{\sigma_j\hat{\sigma}_j^\mathsf{T}}{T}\\
&= Q\frac{\Sigma^\mathsf{T}\hat{\Sigma}}{T} + \frac{Q\sigma_j\sigma_j^\mathsf{T}Q}{1-\frac{1}{T}\sigma_j^\mathsf{T}Q\sigma_j}\frac{\Sigma^\mathsf{T}\hat{\Sigma}}{T} - Q_{-j}\frac{\sigma_j\hat{\sigma}_j^\mathsf{T}}{T}
\end{aligned}$$

and thus Z_{3112} can be expanded as the sum of three terms that shall be

studied in order:

$$\begin{aligned}
Z_{3112} &= \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\Sigma_{-j}^{\mathsf{T}}\hat{\Sigma}_{-j}\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q_{-j}(\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)}{1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j}Y^{\mathsf{T}}\right)\right]\\
&= \frac{1}{T\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ\frac{\Sigma^{\mathsf{T}}\hat{\Sigma}}{T}\hat{\Sigma}^{\mathsf{T}}D\Sigma QY^{\mathsf{T}}\right)\right]\\
&+ \frac{1}{T\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q\sigma_j\sigma_j^{\mathsf{T}}Q\Sigma^{\mathsf{T}}\hat{\Sigma}\hat{\sigma}_j(\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)\sigma_j^{\mathsf{T}}Q}{T(1-\frac{1}{T}\sigma_j^{\mathsf{T}}Q\sigma_j)}Y^{\mathsf{T}}\right)\right]\\
&- \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(YQ\sigma_j\hat{\sigma}_j^{\mathsf{T}}\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q(\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)(1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)Y^{\mathsf{T}}\right)\right]\\
&\equiv Z_{31121}+Z_{31122}-Z_{31123}.
\end{aligned}$$

where $D=\mathrm{diag}(\{\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j\}_{i=1}^n)$. First, Z_{31121} is of order $O(n^{\varepsilon-\frac{1}{2}})$ since $Q\frac{\Sigma^{\mathsf{T}}\hat{\Sigma}}{T}$ is of bounded operator norm. Subsequently, Z_{31122} can be rewritten as

$$Z_{31122}=\frac{1}{\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ\frac{\Sigma^{\mathsf{T}}D\Sigma}{T}QY^{\mathsf{T}}\right)\right]=O(n^{\varepsilon-\frac{1}{2}})$$

with here

$$D=\mathrm{diag}\left\{\frac{\left(\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j\right)\left(\frac{1}{T}\mathrm{tr}\left(Q_{-j}\frac{\Sigma_{-j}^{\mathsf{T}}\hat{\Sigma}_{-j}}{T}\Phi_{\hat{X}X}\right)+\frac{1}{T}\mathrm{tr}\left(Q_{-j}\Phi\right)\frac{1}{T}\mathrm{tr}\,\Phi_{\hat{X}\hat{X}}\right)}{(1-\frac{1}{T}\sigma_j^{\mathsf{T}}Q\sigma_j)(1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)}\right\}_{i=1}^{n}.$$

The same arguments apply for Z_{31123} but for

$$D=\mathrm{diag}\left\{\frac{\mathrm{tr}\,\Phi_{\hat{X}\hat{X}}}{T}(\delta-\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)(1+\frac{1}{T}\sigma_j^{\mathsf{T}}Q_{-j}\sigma_j)\right\}_{i=1}^{n}$$

which completes to show that $|Z_{3112}|\leq Cn^{\varepsilon-\frac{1}{2}}$ and thus

$$\begin{aligned}
Z_{311} &= Z_{3111}+O(n^{\varepsilon-\frac{1}{2}})\\
&= \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\Sigma_{-j}^{\mathsf{T}}\hat{\Sigma}_{-j}\hat{\sigma}_j\sigma_j^{\mathsf{T}}Q_{-j}Y^{\mathsf{T}}\right)\right]+O(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}$$

It remains to handle Z_{3111}. Under the same claims as above, we have

$$\begin{aligned}
Z_{3111} &= \frac{1}{T\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\frac{\Sigma_{-j}^{\mathsf{T}}\hat{\Sigma}_{-j}}{T}\Phi_{\hat{X}X}Q_{-j}Y^{\mathsf{T}}\right)\right] \\
&= \frac{1}{T\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\frac{\sigma_i\hat{\sigma}_i^{\mathsf{T}}}{T}\Phi_{\hat{X}X}Q_{-j}Y^{\mathsf{T}}\right)\right] \\
&= \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^{\mathsf{T}}}{1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}Q_{-ij}Y^{\mathsf{T}}\right)\right] \\
&\quad - \frac{1}{T^3\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^{\mathsf{T}}}{1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}\frac{Q_{-ij}\sigma_i\sigma_i^{\mathsf{T}}Q_{-ij}}{1+\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-ij}\sigma_i}Y^{\mathsf{T}}\right)\right] \\
&\equiv Z_{31111} - Z_{31112}
\end{aligned}$$

where we introduced the notation $Q_{-ij} = (\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma - \frac{1}{T}\sigma_i\sigma_i^{\mathsf{T}} - \frac{1}{T}\sigma_j\sigma_j^{\mathsf{T}} + \gamma I_T)^{-1}$. For Z_{31111}, we replace $\frac{1}{T}\sigma_i^{\mathsf{T}}Q_{-ij}\sigma_i$ by δ, and take the expectation over w_i,

as follows

$$
\begin{aligned}
Z_{31111} &= \frac{1}{T^2\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}Q_{-ij}Y^\mathsf{T}\right)\right]\\
&= \frac{1}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}\Phi_{\hat{X}X}Q_{-ij}Y^\mathsf{T}\right)\right]\\
&+ \frac{1}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}(\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i)}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}Q_{-ij}Y^\mathsf{T}\right)\right]\\
&= \frac{n^2}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_{--}\Phi_{X\hat{X}}\Phi_{\hat{X}X}Q_{--}Y^\mathsf{T}\right)\right]\\
&+ \frac{1}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\sigma_i\hat{\sigma}_i^\mathsf{T}\left(\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i\right)\Phi_{\hat{X}X}Q_{-j}Y^\mathsf{T}\right)\right]\\
&+ \frac{1}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\sigma_i\hat{\sigma}_i^\mathsf{T}\Phi_{\hat{X}X}\frac{Q_{-j}\frac{1}{T}\sigma_i\sigma_i^\mathsf{T}Q_{-j}}{1-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-j}\sigma_i}Y^\mathsf{T}\left(\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i\right)\right)\right]\\
&= \frac{n^2}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_{--}\Phi_{X\hat{X}}\Phi_{\hat{X}X}Q_{--}Y^\mathsf{T}\right)\right]\\
&+ \frac{1}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\Sigma_{-j}^\mathsf{T}D\hat{\Sigma}_{-j}\Phi_{\hat{X}X}Q_{-j}Y^\mathsf{T}\right)\right]\\
&+ \frac{n}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^{n}\mathrm{E}\left[YQ_{-j}\Sigma_{-j}^\mathsf{T}D'\Sigma_{-j}Q_{-j}Y^\mathsf{T}\right]+O(n^{\varepsilon-\frac{1}{2}})\\
&= \frac{n^2}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_{--}\Phi_{X\hat{X}}\Phi_{\hat{X}X}Q_{--}Y^\mathsf{T}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})
\end{aligned}
$$

with Q_{--} having the same law as Q_{-ij}, $D=\mathrm{diag}(\{\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i\}_{i=1}^n)$ and $D'=\mathrm{diag}\left\{\frac{(\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i)\frac{1}{T}\mathrm{tr}\left(\Phi_{\hat{X}X}Q_{-ij}\Phi_{X\hat{X}}\right)}{(1-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-j}\sigma_i)(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i)}\right\}_{i=1}^n$, both expected to be of order $O(n^{\varepsilon-\frac{1}{2}})$. Using again the asymptotic equivalent of $\mathrm{E}[QAQ]$ devised in Section 5.2.3, we then have

$$
\begin{aligned}
Z_{31111} &= \frac{n^2}{T^2\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_{--}\Phi_{X\hat{X}}\Phi_{\hat{X}X}Q_{--}Y^\mathsf{T}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})\\
&= \frac{1}{\hat{T}}\mathrm{tr}\left(Y\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}Y^\mathsf{T}\right)+\frac{1}{\hat{T}}\mathrm{tr}\left(\Psi_X\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}\right)\frac{\frac{1}{n}\mathrm{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\mathrm{tr}(\Psi_X^2\bar{Q}^2)}\\
&+ O(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}
$$

Following the same principle, we deduce for Z_{31112} that

$$\begin{aligned}
Z_{31112} &= \frac{1}{T^3\hat{T}}\frac{1}{1+\delta}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}\frac{Q_{-ij}\sigma_i\sigma_i^\mathsf{T}Q_{-ij}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}Y^\mathsf{T}\right)\right]\\
&= \frac{1}{T^3\hat{T}}\frac{1}{(1+\delta)^3}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-ij}\sigma_i\sigma_i^\mathsf{T}Q_{-ij}Y^\mathsf{T}\right)\frac{1}{T}\mathrm{tr}\left(\Phi_{\hat{X}X}Q_{-ij}\Phi_{X\hat{X}}\right)\right]\\
&+ \frac{1}{T^3\hat{T}}\frac{1}{(1+\delta)^3}\sum_{j=1}^{n}\sum_{i\neq j}\mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\sigma_iD_i\sigma_i^\mathsf{T}Q_{-j}Y^\mathsf{T}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})\\
&= \frac{n^2}{T^3\hat{T}}\frac{1}{1+\delta}\mathrm{E}\left[\mathrm{tr}\left(YQ_{--}\Phi_XQ_{--}Y^\mathsf{T}\right)\frac{1}{T}\mathrm{tr}\left(\Phi_{\hat{X}X}Q_{--}\Phi_{X\hat{X}}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})\\
&= \frac{1}{\hat{T}}\mathrm{tr}\left(\Psi_{\hat{X}X}\bar{Q}\Psi_{X\hat{X}}\right)\frac{\frac{1}{n}\mathrm{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\mathrm{tr}(\Psi_X^2\bar{Q}^2)}+O(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}$$

with $D_i = \frac{1}{T}\mathrm{tr}\left(\Phi_{\hat{X}X}Q_{-ij}\Phi_{X\hat{X}}\right)\left[(1+\delta)^2-(1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i)^2\right]$, also believed to be of order $O(n^{\varepsilon-\frac{1}{2}})$. Recalling the fact that $Z_{311} = Z_{3111}+O(n^{\varepsilon-\frac{1}{2}})$, we can thus conclude for Z_{311} that

$$\begin{aligned}
Z_{311} &= \frac{1}{\hat{T}}\mathrm{tr}\left(Y\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}Y^\mathsf{T}\right)+\frac{1}{\hat{T}}\mathrm{tr}\left(\Psi_X\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}\right)\frac{\frac{1}{n}\mathrm{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\mathrm{tr}(\Psi_X^2\bar{Q}^2)}\\
&- \frac{1}{\hat{T}}\mathrm{tr}\left(\Psi_{\hat{X}X}\bar{Q}\Psi_{X\hat{X}}\right)\frac{\frac{1}{n}\mathrm{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\mathrm{tr}(\Psi_X^2\bar{Q}^2)}+O(n^{\varepsilon-\frac{1}{2}}).
\end{aligned}$$

As for Z_{312}, we have

$$\begin{aligned}
Z_{312} &= \frac{1}{T^3\hat{T}}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}\frac{\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right]\\
&= \frac{1}{T^3\hat{T}}\sum_{j=1}^{n}\mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}\frac{\hat{\sigma}_j\sigma_j^\mathsf{T}Q_{-j}}{1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j}Y^\mathsf{T}\right)\right].
\end{aligned}$$

Since $Q_{-j}\frac{1}{T}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}$ is expected to be of bounded norm, using the concen-

tration inequality of the quadratic form $\frac{1}{T}\sigma_j^\mathsf{T} Q_{-j}\frac{\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}}{T}\hat{\sigma}_j$, we infer

$$Z_{312} = \frac{1}{T\hat{T}}\sum_{j=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}Y^\mathsf{T}}{(1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j)^2}\right)\left(\frac{1}{T^2}\mathrm{tr}\left(Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\Phi_{\hat{X}X}\right)+O(n^{\varepsilon-\frac{1}{2}})\right)\right]$$
$$= \frac{1}{T\hat{T}}\sum_{j=1}^n \mathrm{E}\left[\mathrm{tr}\left(Y\frac{Q_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}Y^\mathsf{T}}{(1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j)^2}\right)\left(\frac{1}{T^2}\mathrm{tr}\left(Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\Phi_{\hat{X}X}\right)\right)\right]+O(n^{\varepsilon-\frac{1}{2}}).$$

We again replace $\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j$ by δ and take expectation over w_j to obtain

$$Z_{312} = \frac{1}{T\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^n \mathrm{E}\left[\mathrm{tr}\left(YQ_{-j}\sigma_j\sigma_j^\mathsf{T}Q_{-j}Y^\mathsf{T}\right)\frac{1}{T^2}\mathrm{tr}\left(Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\Phi_{\hat{X}X}\right)\right]$$
$$+\frac{1}{T\hat{T}}\frac{1}{(1+\delta)^2}\sum_{j=1}^n \mathrm{E}\left[\frac{\mathrm{tr}(YQ_{-j}\sigma_j D_j\sigma_j^\mathsf{T}Q_{-j}Y^\mathsf{T})}{(1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j)^2}\frac{1}{T^2}\mathrm{tr}\left(Q_{-j}\Sigma_{-j}^\mathsf{T}\hat{\Sigma}_{-j}\Phi_{\hat{X}X}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})$$
$$= \frac{n}{T\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_-\Phi_X Q_-Y^\mathsf{T}\right)\frac{1}{T^2}\mathrm{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right)\right]$$
$$+\frac{1}{T\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ\Sigma^\mathsf{T}D\Sigma QY^\mathsf{T}\right)\frac{1}{T^2}\mathrm{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right)\right]+O(n^{\varepsilon-\frac{1}{2}})$$

with $D_j = (1+\delta)^2-(1+\frac{1}{T}\sigma_j^\mathsf{T}Q_{-j}\sigma_j)^2 = O(n^{\varepsilon-\frac{1}{2}})$, which eventually brings the second term to vanish, and we thus get

$$Z_{312} = \frac{n}{T\hat{T}}\frac{1}{(1+\delta)^2}\mathrm{E}\left[\mathrm{tr}\left(YQ_-\Phi_X Q_-Y^\mathsf{T}\right)\frac{1}{T^2}\mathrm{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right)\right]+O(n^{\varepsilon-\frac{1}{2}}).$$

For the term $\frac{1}{T^2}\mathrm{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right)$ we apply again the concentration inequality to get

$$\frac{1}{T^2}\mathrm{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right) = \frac{1}{T^2}\sum_{i\neq j}\mathrm{tr}\left(Q_{-j}\sigma_i\hat{\sigma}_i^\mathsf{T}\Phi_{\hat{X}X}\right)$$
$$= \frac{1}{T^2}\sum_{i\neq j}\mathrm{tr}\left(\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}\right)$$
$$= \frac{1}{T^2}\frac{1}{1+\delta}\sum_{i\neq j}\mathrm{tr}\left(Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}\Phi_{\hat{X}X}\right)+\frac{1}{T^2}\frac{1}{1+\delta}\sum_{i\neq j}\mathrm{tr}\left(\frac{Q_{-ij}\sigma_i\hat{\sigma}_i^\mathsf{T}(\delta-\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i)}{1+\frac{1}{T}\sigma_i^\mathsf{T}Q_{-ij}\sigma_i}\Phi_{\hat{X}X}\right)$$
$$= \frac{n-1}{T^2}\frac{1}{1+\delta}\mathrm{tr}\left(\Phi_{\hat{X}X}\mathrm{E}[Q_{--}]\Phi_{X\hat{X}}\right)+\frac{1}{T^2}\frac{1}{1+\delta}\mathrm{tr}\left(Q_{-j}\Sigma_{-j}^\mathsf{T}D\hat{\Sigma}_{-j}\Phi_{\hat{X}X}\right)+O(n^{\varepsilon-\frac{1}{2}})$$

with high probability, where $D = \mathrm{diag}(\{\delta - \frac{1}{T}\sigma_i^\mathsf{T} Q_{-ij}\sigma_i\}_{i=1}^n)$, the norm of which is of order $O(n^{\varepsilon-\frac{1}{2}})$. This entails

$$\frac{1}{T^2}\operatorname{tr}\left(Q_-\Sigma_-^\mathsf{T}\hat{\Sigma}_-\Phi_{\hat{X}X}\right) = \frac{n}{T^2}\frac{1}{1+\delta}\operatorname{tr}\left(\Phi_{\hat{X}X}\mathrm{E}[Q_{--}]\Phi_{X\hat{X}}\right) + O(n^{\varepsilon-\frac{1}{2}})$$

with high probability. Once more plugging the asymptotic equivalent of $\mathrm{E}[QAQ]$ deduced in Section 5.2.3, we conclude for Z_{312} that

$$Z_{312} = \frac{1}{\hat{T}}\operatorname{tr}\left(\Psi_{\hat{X}X}\bar{Q}\Psi_{X\hat{X}}\right)\frac{\frac{1}{n}\operatorname{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X^2\bar{Q}^2)} + O(n^{\varepsilon-\frac{1}{2}})$$

and eventually for Z_{31}

$$\begin{aligned}Z_{31} = {} & \frac{1}{\hat{T}}\operatorname{tr}\left(Y\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}Y^\mathsf{T}\right) + \frac{1}{\hat{T}}\operatorname{tr}\left(\Psi_X\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}\right)\frac{\frac{1}{n}\operatorname{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X^2\bar{Q}^2)} \\ & - \frac{2}{\hat{T}}\operatorname{tr}\left(\Psi_{\hat{X}X}\bar{Q}\Psi_{X\hat{X}}\right)\frac{\frac{1}{n}\operatorname{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X^2\bar{Q}^2)} + O(n^{\varepsilon-\frac{1}{2}}).\end{aligned}$$

Combining the estimates of $\mathrm{E}[Z_2]$ as well as Z_{31} and Z_{32}, we finally have the estimates for the test error defined in (12) as

$$\begin{aligned}E_{\text{test}} = {} & \frac{1}{\hat{T}}\left\|\hat{Y}^\mathsf{T} - \Psi_{X\hat{X}}^\mathsf{T}\bar{Q}Y^\mathsf{T}\right\|_F^2 \\ & + \frac{\frac{1}{n}\operatorname{tr}\left(Y\bar{Q}\Psi_X\bar{Q}Y^\mathsf{T}\right)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X^2\bar{Q}^2)}\left[\frac{1}{\hat{T}}\operatorname{tr}\Psi_{\hat{X}\hat{X}} + \frac{1}{\hat{T}}\operatorname{tr}\left(\Psi_X\bar{Q}\Psi_{X\hat{X}}\Psi_{\hat{X}X}\bar{Q}\right) - \frac{2}{\hat{T}}\operatorname{tr}\left(\Psi_{\hat{X}\hat{X}}\bar{Q}\Psi_{X\hat{X}}\right)\right] \\ & + O(n^{\varepsilon-\frac{1}{2}}).\end{aligned}$$

Since by definition, $\bar{Q} = (\Psi_X + \gamma I_T)^{-1}$, we may use

$$\Psi_X\bar{Q} = (\Psi_X + \gamma I_T - \gamma I_T)(\Psi_X + \gamma I_T)^{-1} = I_T - \gamma\bar{Q}$$

in the second term in brackets to finally retrieve the form of Conjecture 1.

6. Concluding Remarks. This article provides a possible direction of exploration of random matrices involving entry-wise non-linear transformations (here through the function $\sigma(\cdot)$), as typically found in modelling neural networks, by means of a concentration of measure approach. The main advantage of the method is that it leverages the concentration of an initial random vector w (here a Lipschitz function of a Gaussian vector) to transfer concentration to all vector σ (or matrix Σ) being Lipschitz functions of

w. This induces that Lipschitz functionals of σ (or Σ) further satisfy concentration inequalities and thus, if the Lipschitz parameter scales with n, convergence results as $n \to \infty$. With this in mind, note that we could have generalized our input-output model $z = \beta^{\mathsf{T}}\sigma(Wx)$ of Section 2 to

$$z = \beta^{\mathsf{T}}\sigma(x; \mathcal{W})$$

for $\sigma : \mathbb{R}^p \times \mathcal{P} \to \mathbb{R}^n$ with $\mathcal{P}$ some probability space and $\mathcal{W} \in \mathcal{P}$ a random variable such that $\sigma(x; \mathcal{W})$ and $\sigma(X; \mathcal{W})$ (where $\sigma(\cdot)$ is here applied column-wise) satisfy a concentration of measure phenomenon; it is not even necessary that $\sigma(X; \mathcal{W})$ has a *normal* concentration so long that the corresponding concentration function allows for appropriate convergence results. This generalized setting however has the drawback of being less explicit and less practical (as most neural networks involve linear maps Wx rather than non-linear maps of $\mathcal{W}$ and x).

A much less demanding generalization though would consist in changing the vector $w \sim \mathcal{N}_\varphi(0, I_p)$ for a vector w still satisfying an exponential (not necessarily normal) concentration. This is the case notably if $w = \varphi(\tilde{w})$ with $\varphi(\cdot)$ a Lipschitz map with Lipschitz parameter bounded by, say, $\log(n)$ or any small enough power of n. This would then allow for w with heavier than Gaussian tails.

Despite its simplicity, the concentration method also has some strong limitations that presently do not allow for a sufficiently profound analysis of the testing mean square error. We believe that Conjecture 1 can be proved by means of more elaborate methods. Notably, we believe that the powerful Gaussian method advertised in (Pastur and Ŝerbina, 2011) which relies on Stein's lemma and the Poincaré–Nash inequality could provide a refined control of the residual terms involved in the derivation of Conjecture 1. However, since Stein's lemma (which states that $\mathrm{E}[x\phi(x)] = \mathrm{E}[\phi'(x)]$ for $x \sim \mathcal{N}(0, 1)$ and differentiable polynomially bounded ϕ) can only be used on products $x\phi(x)$ involving the linear component x, the latter is not directly accessible; we nonetheless believe that appropriate ansatzs of Stein's lemma, adapted to the non-linear setting and currently under investigation, could be exploited.

As a striking example, one key advantage of such a tool would be the possibility to evaluate expectations of the type $Z = \mathrm{E}[\sigma\sigma^{\mathsf{T}}(\frac{1}{T}\sigma^{\mathsf{T}}Q_-\sigma - \alpha)]$ which, in our present analysis, was shown to be bounded in the order of symmetric matrices by $\Phi C n^{\varepsilon - \frac{1}{2}}$ with high probability. Thus, if no matrix (such as $\bar{Q}$) pre-multiplies Z, since $\|\Phi\|$ can grow as large as $O(n)$, Z cannot be shown to vanish. But such a bound does not account for the fact that

Φ would in general be unbounded because of the term $\bar{\sigma}\bar{\sigma}^\mathsf{T}$ in the display $\Phi = \bar{\sigma}\bar{\sigma}^\mathsf{T} + \mathrm{E}[(\sigma - \bar{\sigma})(\sigma - \bar{\sigma})^\mathsf{T}]$, where $\bar{\sigma} = \mathrm{E}[\sigma]$. Intuitively, the "mean" contribution $\bar{\sigma}\bar{\sigma}^\mathsf{T}$ of $\sigma\sigma^\mathsf{T}$, being post-multiplied in Z by $\frac{1}{T}\sigma^\mathsf{T} Q_- \sigma - \alpha$ (which averages to zero) disappears; and thus only smaller order terms remain. We believe that the aforementioned ansatzs for the Gaussian tools would be capable of subtly handling this self-averaging effect on Z to prove that $\|Z\|$ vanishes (for $\sigma(t) = t$, it is simple to show that $\|Z\| \leq Cn^{-1}$). In addition, Stein's lemma-based methods only require the differentiability of $\sigma(\cdot)$, which need not be Lipschitz, thereby allowing for a larger class of activation functions.

As suggested in the simulations of Figure 2, our results also seem to extend to non continuous functions $\sigma(\cdot)$. To date, we cannot envision a method allowing to tackle this setting.

In terms of neural network applications, the present article is merely a first step towards a better understanding of the "hardening" effect occurring in large dimensional networks with numerous samples and large data points (that is, simultaneously large n, p, T), which we exemplified here through the convergence of mean-square errors. The mere fact that some standard performance measure of these random networks would "freeze" as n, p, T grow at the predicted regime and that the performance would heavily depend on the distribution of the random entries is already in itself an interesting result to neural network understanding and dimensioning. However, more interesting questions remain open. Since neural networks are today dedicated to classification rather than regression, a first question is the study of the asymptotic statistics of the output $z = \beta^\mathsf{T}\sigma(Wx)$ itself; we believe that z satisfies a central limit theorem with mean and covariance allowing for assessing the asymptotic misclassification rate.

A further extension of the present work would be to go beyond the single-layer network and include multiple layers (finitely many or possibly a number scaling with n) in the network design. The interest here would be on the key question of the best distribution of the number of neurons across the successive layers.

It is also classical in neural networks to introduce different (possibly random) biases at the neuron level, thereby turning $\sigma(t)$ into $\sigma(t + b)$ for a random variable b different for each neuron. This has the effect of mitigating the negative impact of the mean $\mathrm{E}[\sigma(w_i^\mathsf{T} x_j)]$, which is independent of the neuron index i.

Finally, neural networks, despite their having been recently shown to op-

erate almost equally well when taken random in some very specific scenarios, are usually only *initiated* as random networks before being subsequently trained through backpropagation of the error on the training dataset (that is, essentially through convex gradient descent). We believe that our framework can allow for the understanding of at least finitely many steps of gradient descent, which may then provide further insights into the overall performance of deep learning networks.

APPENDIX A: INTERMEDIARY LEMMAS

This section recalls some elementary algebraic relations and identities used throughout the proof section.

LEMMA 5 (Resolvent Identity). *For invertible matrices* A, B, $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$.

LEMMA 6 (A rank-1 perturbation identity). *For* A *Hermitian,* v *a vector and* $t \in \mathbb{R}$, *if* A *and* $A + tvv^\mathsf{T}$ *are invertible, then*

$$\left(A + tvv^\mathsf{T}\right)^{-1} v = \frac{A^{-1}v}{1 + tv^\mathsf{T}A^{-1}v}.$$

LEMMA 7 (Operator Norm Control). *For nonnegative definite* A *and* $z \in \mathbb{C} \setminus \mathbb{R}^+$,

$$\| (A - zI_T)^{-1} \| \leq \operatorname{dist}(z, \mathbb{R}^+)^{-1}$$
$$\|A (A - zI_T)^{-1} \| \leq 1$$

where $\operatorname{dist}(x, \mathcal{A})$ *is the Hausdorff distance of a point to a set. In particular, for* $\gamma > 0$, $\|(A + \gamma I_T)^{-1}\| \leq \gamma^{-1}$ *and* $\|A(A + \gamma I_T)^{-1}\| \leq 1$.

REFERENCES

AKHIEZER, N. I. and GLAZMAN, I. M. (1993). *Theory of linear operators in Hilbert space*. Courier Dover Publications.

BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *The Annals of Probability* **26** 316-345.

BAI, Z. D. and SILVERSTEIN, J. W. (2007). On the signal-to-interference-ratio of CDMA systems in wireless communications. *Annals of Applied Probability* **17** 81-101.

BAI, Z. D. and SILVERSTEIN, J. W. (2009). *Spectral analysis of large dimensional random matrices*, second ed. Springer Series in Statistics, New York, NY, USA.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis* **111** 120–135.

Cambria, E., Gastaldo, P., Bisio, F. and Zunino, R. (2015). An ELM-based model for affective analogical reasoning. *Neurocomputing* **149** 443–455.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B. and LeCun, Y. (2015). The Loss Surfaces of Multilayer Networks. In *AISTATS*.

Couillet, R. and Benaych-Georges, F. (2016). Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics* **10** 1393–1454.

Couillet, R. and Kammoun, A. (2016). Random Matrix Improved Subspace Clustering. In *2016 Asilomar Conference on Signals, Systems, and Computers*.

Couillet, R., Pascal, F. and Silverstein, J. W. (2015). The random matrix regime of Maronna's M-estimator with elliptically distributed samples. *Journal of Multivariate Analysis* **139** 56–78.

Giryes, R., Sapiro, G. and Bronstein, A. M. (2015). Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? *IEEE Transactions on Signal Processing* **64** 3444-3457.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks* **2** 359–366.

Hoydis, J., Couillet, R. and Debbah, M. (2013). Random beamforming over quasi-static and fading channels: a deterministic equivalent approach. *IEEE Transactions on Information Theory* **58** 6392-6425.

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* **70** 489–501.

Huang, G.-B., Zhou, H., Ding, X. and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **42** 513–529.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **304** 78–80.

Kammoun, A., Kharouf, M., Hachem, W. and Najim, J. (2009). A central limit theorem for the sinr at the lmmse estimator output for large-dimensional signals. *IEEE Transactions on Information Theory* **55** 5048–5063.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19** 2362–2405.

El Karoui, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics* **38** 1–50.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* 1097–1105.

LeCun, Y., Cortes, C. and Burges, C. (1998). The MNIST database of handwritten digits.

Ledoux, M. (2005). *The concentration of measure phenomenon* **89**. American Mathematical Soc.

Loubaton, P. and Vallet, P. (2010). Almost sure localization of the eigenvalues in a Gaussian information plus noise model. Application to the spiked models. *Electronic Journal of Probability* **16** 1934–1959.

Mai, X. and Couillet, R. (2017). The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math USSR-Sbornik* **1** 457-483.

Pastur, L. and Ŝerbina, M. (2011). *Eigenvalue distribution of large random matrices*. American Mathematical Society.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems* 1177–1184.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65** 386.

Rudelson, M., Vershynin, R. et al. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab* **18** 1–9.

Saxe, A., Koh, P. W., Chen, Z., Bhand, M., Suresh, B. and Ng, A. Y. (2011). On random weights and unsupervised feature learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* 1089–1096.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* **61** 85–117.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis* **54** 175-192.

Silverstein, J. W. and Choi, S. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis* **54** 295-309.

Tao, T. (2012). *Topics in random matrix theory* **132**. American Mathematical Soc.

Titchmarsh, E. C. (1939). *The Theory of Functions*. Oxford University Press, New York, NY, USA.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *in Compressed Sensing, 210–268, Cambridge University Press*.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation* **10** 1203–1216.

Yates, R. D. (1995). A framework for uplink power control in cellular radio systems. *IEEE Journal on Selected Areas in Communications* **13** 1341-1347.

Zhang, T., Cheng, X. and Singer, A. (2014). Marchenko-Pastur Law for Tyler's and Maronna's M-estimators. *http://arxiv.org/abs/1401.3424*.

Zhenyu Liao, R. C. (2017). A Large Dimensional Analysis of Least Squares Support Vector Machines. *(submitted to) Journal of Machine Learning Research*.