Spectral Phase Transitions in Non-Linear Wigner Spiked Models

Alice Guionnet[*], Justin Ko[†], Florent Krzakala[‡], Pierre Mergny[§], Lenka Zdeborová[¶]

Abstract

We study the asymptotic behavior of the spectrum of a random matrix where a non-linearity is applied entry-wise to a Wigner matrix perturbed by a rank-one spike with independent and identically distributed entries. In this setting, we show that when the signal-to-noise ratio scale as $N^{\frac{1}{2}(1-1/k_\star)}$, where $k_\star$ is the first non-zero generalized information coefficient of the function, the non-linear spike model effectively behaves as an equivalent spiked Wigner matrix, where the former spike before the non-linearity is now raised to a power $k_\star$. This allows us to study the phase transition of the leading eigenvalues, generalizing part of the work of Baik, Ben Arous and Peché to these non-linear models.

Contents

---

[*]UMPA, ENS Lyon and CNRS, France. Email: `aguionnet@ens-lyon.fr`

[†]UMPA, ENS Lyon and CNRS, France and Department of Statistics and Actuarial Science, University of Waterloo, Canada. Email: `justin.ko@uwaterloo.ca`

[‡]IdePHICS laboratory, École Fédérale Polytechnique de Lausanne, Switzerland. Email: `florent.krzakala@epfl.ch`

[§]IdePHICS laboratory, École Fédérale Polytechnique de Lausanne, Switzerland. Email: `pierre.mergny@epfl.ch`

[¶]SPOC laboratory, École Fédérale Polytechnique de Lausanne, Switzerland. Email: `lenka.zdeborova@epfl.ch`

1 Introduction and Notations

1.1 Introduction

Random Matrix Theory (RMT)[4, 41, 50, 53] has its roots in Wishart's 1928 statistical investigations [56] and Wigner's 1950s work on nuclear models [55]. Since then, its influence has spread to a variety of fields, including high-energy physics [30, 54], spin glass models [25, 6], and number theory [33]. At its core, RMT delves into the intricate high-dimensional and spectral nuances of select random matrices. Notably, seminal RMT insights determined that the spectrum of matrices first explored by Wishart and Wigner aligns with the semi-circular law and the Marčenko-Pastur distribution, respectively.

While Wigner and Wishart matrices can be considered as pure noise random matrices, one of the most studied models beyond these two examples is the *spiked models*, in which a fixed rank matrix, playing the role of a signal, is added to the former Wigner or Wishart matrix. In its simplest form, for example, when one studies the rank-one perturbation of a Wigner matrix (see Sec. 2.2 for the quantitative details), one can show [9, 44, 10, 51, 45] that in the high-dimensional regime, the highest eigenvalue in absolute value λ_1 of this new random matrix undergoes a phase transition, depending on the strength γ of the rank-one perturbation, from a regime where it sticks to the edge of the limiting spectral distribution, to the one where it pops out of the bulk. Similarly, the associated eigenvector behaves as a random vector uniformly sampled on the sphere before the transition and becomes partially aligned with the vector of the rank-one perturbation after the transition.

This phenomenon is now commonly referred to as the 'BBP phase transition' after the names of the three authors [9] who studied fluctuations around the deterministic limit for this type of model. This BBP phase transition phenomenon has been generalized and applied to a variety of different settings; in particular, we refer the reader to [19, 27, 17] for other studies of the fluctuations in similar spike models, to [43, 16] for the study of the large deviation of the top eigenvalue and eigenvector whenever the noise is Gaussian, to [14, 15] for finite-rank perturbation of unitary invariant matrices and to [37, 12, 26, 36, 35, 38, 11] for the study of these spike models from the information-theoretic point of view.

A second family of models that has been extensively studied in the past decade due to its connection with high-dimensional statistical problems when both the number of samples and the dimension diverge is the family of *nonlinear (random matrix) models*, where one study the property of an *entry-wise* non-linearity function $f(\cdot)$ applied to a given random matrix, and we refer in particular to the works [32, 31, 13, 52, 29] for their relation to kernel methods and to [47, 46, 42, 28, 39, 49, 21, 48] for the relation to the spectrum of one-layer neural network at initialization (or random features).

This paper aims to consider a mixture of these two important families of models by establishing the spectral properties of non-linear matrix models applied to rank-one perturbation of Wigner matrices

$$Y_{ij}^{(f)} := \frac{1}{\sqrt{N}}\left[f\left(Z_{ij} + \frac{\gamma}{\sqrt{N}}x_i x_j\right) - \mathbb{E}_Z f(Z)\right], \tag{1.1}$$

for 'arbitrary' non-linearity functions f, and 'arbitrary' distributions of Z_{ij} and $\mathbf{x} = (x_1, \dots, x_N)$. Studying the spectrum of these matrices, we will see that the most interesting regime is when the *signal-to-noise ratio* (SNR) γ scaled with the system size N.

Matrices of the form (1.1) are also important in theoretical machine learning. They can be seen as kernel methods applied to spiked models, but they also appear in the studies of gradient descent in deep neural networks where Z corresponds to the pre-activations at initialization, f is the non-linear activation function, and the low-rank perturbation originates from the dynamics of the early steps of training with gradient descent such as in [7, 22, 23].

A special case of the problem we study was considered by [57], who studied the information-theoretic aspects of recovering a low-rank signal x from unsigned observations of the matrix, corresponding to $f(x) = |x|$, but they have not studied the spectrum of such matrices.

Our main goal is to study the asymptotic behavior of the leading eigenvalue of the matrix $Y^{(f)}$, namely the largest eigenvalue in absolute value of the matrix $Y^{(f)}$: depending on the sign of γ and f, it can be the smallest or the largest eigenvalue of $Y^{(f)}$.
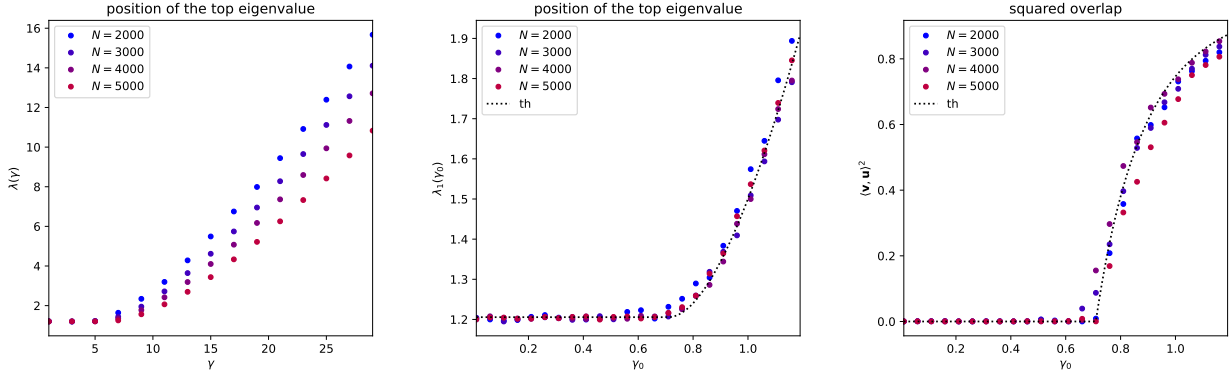
Figure 1: Illustration of the main results for the matrix model defined by Eq. (1.2) with non-linearity $f_1(x) = |x|$ and Gaussian noise for which the information index is $k_\star = 2$. **(Left)** The position of the leading eigenvalue as a function of the SNR γ for different matrix sizes N. As N increases, one observes a shift of the curves towards the right, suggesting that one should rescale the SNR with N. **(Middle)** The same data plotted against the rescaled SNR $\gamma_0 = \gamma N^{-1/4}$ given by Eq. (1.7). The dotted black curves represent the theoretical values given by a BBP-like theorem Eq. (1.9). **(Right)** The squared overlap of the leading eigenvector with the vector $\mathbf{u} \equiv \mathbf{x}^{k_\star}/\|\mathbf{x}^{k_\star}\|$ where $\mathbf{x}$ is the signal vector as a function of the *rescaled* SNR γ_0 given by Eq. (1.7) and different sizes N. The dotted black curves represent the theoretical values given by a BBP-like theorem Eq. (1.10).

1.2 Motivating Examples

As an empirical motivation of the non-trivial behavior one can obtain in this setting, let's consider x_i and Z_{ij} standard independent Gaussian random variables and construct the two $(N \times N)$ symmetric matrices $\boldsymbol{Y}^{(f_1)} := (Y_{ij}^{(f_1)})_{1 \leqslant i,j \leqslant N}$ and $\boldsymbol{Y}^{(f_2)} := (Y_{ij}^{(f_2)})_{1 \leqslant i,j \leqslant N}$ with

$$Y_{ij}^{(f_a)} := \frac{1}{\sqrt{N}}\left[f_a\left(Z_{ij} + \frac{\gamma}{\sqrt{N}}x_i x_j\right) - \mathbb{E}_Z f_a(Z)\right] \quad \text{for } a = 1, 2\,, \tag{1.2}$$

with $f_1(x) := |x|$ and $f_2(x) := x^3 - 3x$ (one has $\mathbb{E}_Z f_1(Z) = \sqrt{2/\pi}$ and $\mathbb{E}_Z f_2(Z) = 0$ in these cases). In the leftmost plot of Fig. 1 and Fig. 2, we have plotted the empirical position of the top eigenvalue λ_1 of each matrix, as a function of γ, for different values of the size N. Unlike the well-known result for the usual 'linear' setting ($f(x) := x$), as one increases the dimension N, one observes a *shift towards the right* of the curves. This empirical phenomenon suggests rescaling the constant $\gamma \equiv \gamma(N)$ with N to get a deterministic limit in the large N limit. An empirical fit further suggests a scaling of the form $\gamma = N^\alpha$ with $\alpha \approx 0.25$ for $\boldsymbol{Y}^{(f_1)}$ and $\alpha \approx 0.33$ for $\boldsymbol{Y}^{(f_2)}$ and a natural question is to give a theoretical explanation for these scaling.

1.3 Summary of the Main Results

We now give a broad overview of the main results obtained in this work and explain quantitatively the behavior of leading eigenvalues and eigenvectors observed in the motivating examples in a more general framework. A careful statement of these main results with precise conditions on the assumptions will be stated in Section 2. Motivated by the examples above, we seek to understand what happens to the spectrum of a matrix

$$Y_{ij}^{(f)} := \frac{1}{\sqrt{N}}\left[f\left(Z_{ij} + \frac{\gamma}{\sqrt{N}}x_i x_j\right) - \mathbb{E}_Z f(Z)\right], \tag{1.3}$$

for 'arbitrary' non-linearity functions f, and 'arbitrary' distributions of Z_{ij} and $\mathbf{x} = (x_1, \ldots, x_N)$. For the sake of simplicity, suppose that the law of Z_{ij} has a smooth density function w_Z that vanishes at infinity, and that the entries of $\mathbf{x}$ are independent with exponential decay. Both of these assumptions
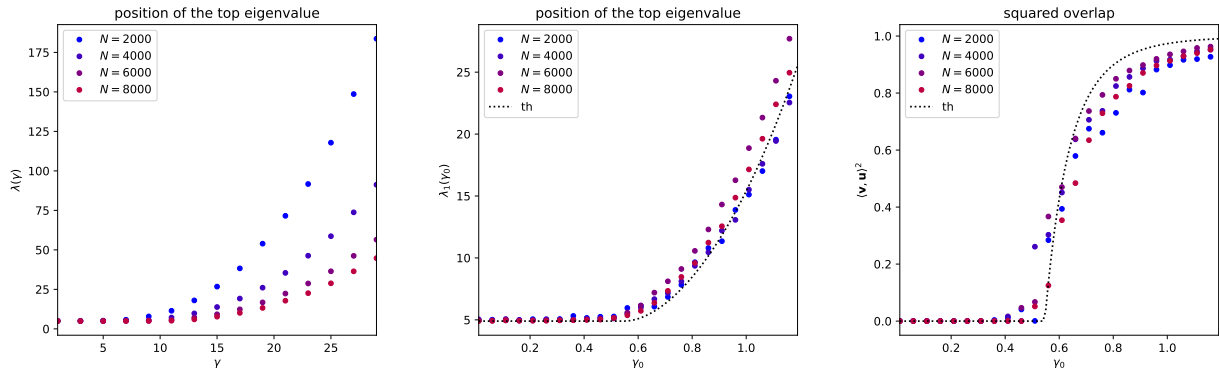
Figure 2: The same as Fig. 1 for the non-linearity $f_2(x) = x^3 - 3x$ and Gaussian noise for which the information index $k_\star = 3$ leading to a rescaling of the SNR $\gamma_0 = \gamma N^{-1/3}$.

are weakened and stated in more generality in Section 2, but the statements of the following results become slightly less explicit without densities.

The relevant scaling for the SNR γ at which the leading eigenvalue is of order one depends on certain statistics that build on the non-linearity f and the law of the noise w_Z. We define the *k-th information coefficient* associated with the non-linearity f and density w_Z of Z by

$$\vartheta_k(f) := (-1)^k \int_{\mathbb{R}} f(x) w_Z^{(k)}(x) \mathrm{d}x. \tag{1.4}$$

where $w_Z^{(k)}$ is the k-th derivative of the density. We then define the associated *information index* $k_\star$ by the first non-zero component of the information coefficient as

$$k_\star := \inf \{1 \leqslant k \leqslant k_0 \text{ such that } \vartheta_k(f) \neq 0\}. \tag{1.5}$$

Whenever the entries Z_{ij} are standard Gaussian random variables, the generalized information coefficients $\vartheta_k(f)$ appearing in this paper reduce to the classical *Hermite coefficients*, that is, the scalar product of f with the k-th Hermite polynomial in the L^2 space weighted by the Gaussian measure. These Hermite coefficients appear naturally in various high-dimensional learning problems, for example [5], where the associated critical index $k_\star$ dictates in a certain manner the *rate* of the learning, in a similar fashion as the relevant scaling of the SNR in this problem, depends on $k_\star$. In Ref. [5], these coefficients are called 'information coefficients' rather than 'Hermite coefficients', hence the name 'generalized information coefficients' used here to denote the more general framework where the noise is not necessarily Gaussian. It will be interesting to know if one can reinterpret these results in terms of the spectral properties of a non-linear rank-one perturbation model.

A consequence of the first main result of our paper indicates that the relevant scaling α of the *signal-to-noise ratio* is given by

$$\gamma(N) = \gamma_0 N^{\frac{1}{2}(1 - \frac{1}{k_\star})}. \tag{1.6}$$

Furthermore, under this scaling and for large N, we show that the non-linear model studied in this paper behaves, up to a vanishing error term, as the usual 'Wigner matrix plus rank-one perturbation' model, with a 'new' rank-one perturbation aligned with the vector $\mathbf{x}^{k_\star} = (x_1^{k_\star}, \ldots, x_N^{k_\star})$.

4

Theorem 1.1 (*Rank-one Equivalence for the Non-Linear Model (Informal)*)

> Suppose that Z has a smooth density w_Z vanishing exponentially fast at infinity and that $\mathbf{x}$ has independent entries with exponential tails. Assume that f is a measurable function such that $\mathbb{E}[f(Z)^4] < \infty$. Moreover, assume that we are in the scaling regime for the SNR where there exists a positive constant γ_0 such that
>
> $$\gamma(N) = \gamma_0 N^{\frac{1}{2}\left(1 - \frac{1}{k_\star}\right)}. \tag{1.7}$$
>
> Then for all $N \geqslant 1$, we have
>
> $$\boldsymbol{Y}^{(f)} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\, \boldsymbol{W} + \gamma_0 \frac{\vartheta_{k_\star}(f)}{k_\star!} \left[\frac{\mathbf{x}^{k_\star}}{\sqrt{N}} \left(\frac{\mathbf{x}^{k_\star}}{\sqrt{N}} \right)^{\mathsf{T}} \right] + \boldsymbol{E}, \tag{1.8}$$
>
> where $\boldsymbol{W}$ is a centered and normalized Wigner matrix with entries with bounded fourth moment and the operator norm of $\boldsymbol{E}$ goes to 0 as $N \to \infty$. Moreover, $\boldsymbol{W}$ and $\mathbf{x}$ are independent.

The rank-one equivalence at the relevant scaling allows us to obtain the limiting behavior of both the leading eigenvalue and associated eigenvector of this non-linear model through classical methods. In particular, the matrix $\boldsymbol{Y}^{(f)}$ exhibits the usual BBP transition [9].

Corollary 1.1 (*The BBP Transition (Informal)*)

> Under the same Hypothesis of Theorem *1.1*, if we denote by λ_1 and $\boldsymbol{v}_1$ the leading eigenvalue and eigenvector of $\boldsymbol{Y}^{(f)}$, we have:
>
> $$\lambda_1 \xrightarrow[N \to \infty]{\text{a.s.}} \mathsf{l}\left(\gamma_0^{k_\star} \frac{\vartheta_{k_\star}(f)}{k_\star!} m_{2k_\star}, \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2} \right), \tag{1.9}$$
>
> and
>
> $$\left\langle \boldsymbol{v}_1, \frac{\mathbf{x}^{k_\star}}{\|\mathbf{x}^{k_\star}\|} \right\rangle^2 \xrightarrow[N \to \infty]{\text{a.s.}} \mathsf{m}\left(\gamma_0^{k_\star} \frac{\vartheta_{k_\star}(f)}{k_\star!} m_{2k_\star}, \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2} \right). \tag{1.10}$$
>
> where $\mathsf{l}(.,.)$ and $\mathsf{m}(.,.)$ are the classical quantities for the leading eigenvalue and squared overlap for rank-one perturbations of Wigner matrices as defined in Eq. (2.6).

Putting these results into action for the Gaussian noise and the non-linear functions $f_1(x) := |x|$ and $f_2(x) := x^3 - 3x$ in Sec. 1.2, one has respectively $k_\star = 2$ and $k_\star = 3$ such that the relevant scaling of Eq. (2.11) writes $\gamma(N) = \gamma_0 N^\alpha$, with $\alpha = 1/4$ and $\alpha = 1/3$ respectively, in accordance with the scaling of the SNR found numerically. In the middle plots of Fig. 1 and Fig. 2, we then plot the leading eigenvalues as a function of the rescaled SNR γ_0 to see that the empirical data points collapse to the theoretical curve from the BBP transition of Corrollary 2.2. In the rightmost plots of Fig. 1 and Fig. 2, we then plot the overlap between the leading eigenvector and the signal again the rescaled SNR γ_0, again seeing a collapse of the theoretical prediction of the BBP transition of Corollary 2.2 with again a perfect agreement.

A major advantage of our approach is that our decomposition is quite robust, and our result can be trivially extended to deal with variants and generalizations of such a model, such as finite rank deformations of Wigner matrices, rank-one deformations of variance-profile Wigner matrices, and rank-one deformations of rectangular matrices (see Section 3).

Organization of the paper - The rest of this article is organized as follows: in Sec. 2, we introduce the main non-linear model studied in this work and define formally the important coefficients $\vartheta_k(f)$. We then introduce the main result of our paper, summarized in Theorem 2.2 from which Theorem 1.1 and Corollary 1.1 are merely special cases under stronger assumptions. Next, in Sec. 3, we generalize our result to higher-rank perturbations and non-symmetric matrices.

The remaining sections of the paper are devoted to the proof of our result. We give a broad overview

of the proof here, and defer to Sec. 4 for a detailed overview of the proof strategy. Our approach uses a rank-one equivalence of the non-linear model to obtain the spectral phase transitions, in comparison to other approaches which involve computing moments or using Stieltjes transforms. From the rank-one equivalence, we can read off the critical scaling as a function of $k_\star$ and apply classical results for perturbations of Wigner matrices to obtain the phase transitions at this critical scaling. Furthermore, the proof of the rank-one equivalence is given entrywise, so it can be easily adapted to prove similar equivalences in for higher rank and non-symmetric matrices as stated in Sec. 3.

The rank-one equivalence is first proved in a simpler case when the non-linearity is smooth, in Sec. 5 through an application of Taylor's theorem and generic bounds on the operator norms of random matrices with independent entries to control the error. In Sec. 6, we prove that the results extend to non-smooth non-linearities provided that there exists a sequence of smooth approximating functions which preserves the information index. Lastly, in Sec. 7, we show that when the noise has a smooth density function, we can explicitly construct this sequence of index preserving smooth approximating functions through a truncation and smoothing argument, showing that we can relax the assumptions on the function f.

1.4 Notations

We recall that a sequence of events $(E_N)_{N\in\mathbb{N}}$ occurs **with overwhelming probability** if for any fixed $A > 0$, there exists a constant C_A independent of N such that $\mathbb{P}(E_N) \geqslant 1 - C_A N^{-A}$ for every $N \in \mathbb{N}$. Similarly, a sequence of events $(E_N)_{N\in\mathbb{N}}$ occurs **with high probability** if for some $c > 0$ independent of N, there exists constants $c, C > 0$ independent of N such that $\mathbb{P}(E_N) \geqslant 1 - C N^{-c}$. We say that the law $\mathbb{P}(\cdot)$ of a random variable X has **stretched exponential tails** if there exists a constant $\alpha > 0$ and universal constants $C, c > 0$ such that for every $M > 0$, $\mathbb{P}(|X| > M) \leqslant Ce^{-cM^\alpha}$. We call α an exponent of the stretched exponential decay. Equality in law between two random variables A and B is denoted by $A \overset{\mathcal{D}}{=} B$.

A function f is **locally Lipschitz with polynomial constant** if for every $L > 0$, there exists a constant $C(L)$ such that

$$\sup_{|x|,|y|\leqslant L}\left|\frac{f(x) - f(y)}{x - y}\right| \leqslant C(L), \tag{1.11}$$

and $C(L)$ is bounded by a polynomial, that is there exists a $\beta \in \mathbb{N}$ and a finite constant C such that for every real number L, $|C(L)| \leqslant C|L|^\beta$. We recall that if $C(L) \equiv C$ is independent of L and the supremum is over the set of definition of f, then f is **Lipschitz**.

For $\mathbf{x} := (x_1, \ldots, x_N)$ and k an integer, we denote by $\mathbf{x}^k := (x_1^k, \ldots, x_N^k)$ the vector where each entry is raised to the power k. For $(N \times M)$ matrices (or vectors if $M = 1$) $\mathbf{A}$ and $\mathbf{B}$, the **Hadamard product** is denoted by $(\mathbf{A} \odot \mathbf{B})_{ij} := A_{ij}B_{ij}$ and the **Hadamard power** by $(\mathbf{A}^{\odot k})_{ij} := A_{ij}^k$. The **leading eigenvalue and eigenvectors** of a symmetric matrix, refer to the eigenvalue with the largest absolute value and its associated eigenvector.

2 Main Results

2.1 Model and Assumptions for the Non-Linear Rank-One Perturbation

We consider the symmetric matrix $\mathbf{Y}^{(f)}$ with entries given by

$$Y_{ij} := \frac{1}{\sqrt{N}}\left[f\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i x_j\right) - \mathbb{E}f(Z_{ij})\right], \tag{2.1}$$

where $\mathbf{x} = (x_1, \ldots, x_N)^\top$ is the *signal vector*, $\mathbf{Z} = (Z_{ij})_{1\leqslant i,j\leqslant N}$ is a symmetric *noise matrix*, f is the *non-linearity* and $\gamma(N)$ is the *signal-to-noise ratio* (SNR) which, in this work, might depend on the dimension N. Subtracting the constant terms $\mathbb{E}f(Z_{ij})$ avoids the appearance of a (non-informative)

Perron-Frobenius mode. We now list the assumptions needed for our main result to hold, where some of them involve an integer $k_\star$ defined in the next section.

For the signal vector:

(**H1**) We assume the entries of the vector $\mathbf{x}$ to be independent and identically distributed (iid), that is with law $\pi_X^{\otimes N}$ and the distribution $\pi_X(\cdot)$ has *stretched exponential tails* and is assumed to be independent of the dimension N. For simplicity, we also assume the distribution to be non-trivial, that is $\pi_X \neq \delta_0$.

We also denote by $m_k := \mathbb{E}_{x \sim \pi_X}[x^k]$, the k-th moment associated to this distribution.

For the noise matrix:

(**H2**) We assume the entries Z_{ij} are *iid* and with distribution μ_Z for $i \leqslant j$, where μ_Z has *stretched exponential tails* and is independent of the dimension N.

We denote by I the support of the distribution of Z and by $\{P_k\}_{k \in \mathbb{N}}$ the family of orthogonal polynomials associated to μ_Z, which always exists since μ_Z admits moments of any order. The case where $\mathrm{d}\mu_Z(x) = \mathrm{d}\mu_G(x) := w_G(x)\mathrm{d}x := (2\pi)^{-1/2}\mathrm{e}^{-x^2/2}\mathrm{d}x$ is the Gaussian distribution is of particular importance and in this case, we simply denote $Z \equiv G$. We recall that for the Gaussian measure, the associated orthogonal polynomials are the famous *Hermite polynomials* $\mathrm{He}_k(x)$ defined as the solution of the differential-recurrence relation $\mathrm{He}_{n+1} = x\mathrm{He}_n - \mathrm{He}_n'$ with initial conditions $\mathrm{He}_0 = 0$ and $\mathrm{He}_1 = 1$.

For the non-linearity:

For a function $f \in C^{k_0}$ for some $k_0 \geqslant 1$ and $f^{(k)} \in L^1(\mathrm{d}\mu_Z)$ for $k \leqslant k_0$, we define for $k \leqslant k_0$ the **k-th (generalized) information coefficient** of f as

$$\vartheta_k(f) = \mathbb{E}_Z f^{(k)}(Z) \, ; \tag{2.2}$$

When $k = 0$, this coefficient simplifies to $\vartheta_0(f) = \mathbb{E}_Z f(Z)$ and in particular $\vartheta_0(f^2) - \vartheta_0(f)^2$ is the variance of the random variable $f(Z)$ for $f \in L^2(\mathrm{d}\mu_Z)$. We define the **(generalized) information index** $k_\star$ as:

$$k_\star := \inf\{1 \leqslant k \leqslant k_0 \text{ such that } \vartheta_k(f) \neq 0\} \, . \tag{2.3}$$

Next, we either assume that f *is sufficiently smooth and bounded* in the following sense:

(**H̃3**) $f \in C^{k_\star+1}$ such that $f^{(k)} \in L^4(\mu_Z)$ for $k \leqslant k_\star$ and $\|f^{(k_\star+1)}\|_\infty < \infty$.

or in the case that f is not smooth, that

(**H3**) (a) $f \in L^4(\mu_Z)$ and f is locally Lipschitz with polynomial constant,

and we can find a sequence $(f_t)_t$ of smooth functions $f_t \in L^4(\mu_Z)$ such that

  (b) $f_t \xrightarrow[t \to \infty]{} f$ in $L^4(\mu_Z)$.

  (c) $f_t \in C^\infty$ has compact support.

  (d) There exists $k_\star$ such that $f_t^{(k)} \in L^4(\mu_Z)$ for $0 \leqslant k \leqslant k_\star$, $\|f_t^{(k_\star+1)}\|_\infty < \infty$ for all t and

  (e) $\vartheta_k(f_t) \xrightarrow[t \to \infty]{} 0$ for all $k < k_\star$ and $\vartheta_{k_\star}(f_t)$ converges as t goes to infinity towards a non-vanishing quantity that we denote $\vartheta_{k_\star}(f)$. In this case, we define the information index of f to be this index $k_\star$.

Note that while (**H̃3**) is stronger than the more general assumption (**H3**), it allows us to obtain a quantitative estimate of the error rate in the main theorem (Thm. 2.2) of this paper, while this rate will be lost in the more general setting of (**H3**).

Assumption (**H3**) is the most general form of our assumption on f which does not need the non-linearity f to be smooth. It is, in particular, satisfied - and this is proven in Sec. 7 - in the critical case where f satisfies (**H3**)a and *the noise distribution is sufficiently smooth and bounded*, that is, if we have the following joint assumptions on f and μ_Z.

(**H4**) (a) The distribution μ_Z admits a density $w_Z := \frac{\mathrm{d}\mu_Z}{\mathrm{d}z}$, with support on the whole real line. Moreover, there exists an integer number ℓ such that its k-th derivative $w_Z^{(k)}$ exists almost everywhere for all $k \leqslant \ell + 1$ and satisfies $w_Z^{(k)} \in L^2(\mathrm{d}x) \cap L^\infty(\mathrm{d}x)$, for all $k \leqslant \ell$.

(b) f is a measurable function such that $\int |f(x) w_Z^{(k)}(x)| \, \mathrm{d}x < \infty$ for $k \leqslant \ell + 1$.

(c) We then set, for $k \leq \ell$,

$$\vartheta_k(f) := (-1)^k \int_{\mathbb{R}} f(x) w_Z^{(k)}(x) \mathrm{d}x \, . \tag{2.4}$$

and we assume that ℓ is larger than $k_\star$, the smallest index k so that $\vartheta_k(f)$ does not vanish. In this case, we define the information index of f to be this index $k_\star$.

Note that whenever f is sufficiently smooth, Eq. (2.4) simply corresponds to the integration by part of Eq. (2.2).

Remark 2.1. The motivating example of $f_1(x) = |x|$ and Z is Gaussian in Figure 1 satisfies Hypothesis (**H3**) because the Gaussian noise satisfies Hypothesis (**H4**).

We conclude this section on the model with two remarks relating the information exponent $\vartheta_k(f)$ to orthogonal polynomials.

Remark 2.2 (Gaussian Noise and Hermite coefficients)**.** For Gaussian noise, the density $w_G(x)$ satisfies (**H4**) and using Stein's identities, $w_G^{(k)}(x) = (-1)^k \mathrm{He}_k(x) w_G(x)$, one can write the information coefficient in terms of the *Hermite coefficients* of f, namely $\vartheta_k(f) = \langle f, \mathrm{He}_k \rangle_{L^2(\mu_G)}$. These coefficients appear naturally in several recent studies related to RMT.

Remark 2.3 (Information coefficients and orthogonal decomposition)**.** Under additional assumptions on f_t and μ_Z, one can differentiate term by term the decomposition of f_t in the orthogonal polynomials basis $\{\mathrm{P}_k\}_{k \in \mathbb{N}}$ associated to μ_Z, to write down its k-th information coefficient as $\vartheta_k(f_t) = \sum_{n=k}^\infty \frac{\langle f_t, \mathrm{P}_n \rangle}{\langle \mathrm{P}_n, \mathrm{P}_n \rangle} \mathbb{E}_Z \mathrm{P}_n^{(k)}(Z)$ and by Hypothesis (**H3**)b, this yields the following formula:

$$\vartheta_k(f) = \sum_{n=k}^\infty \frac{\langle f, \mathrm{P}_n \rangle}{\langle \mathrm{P}_n, \mathrm{P}_n \rangle} \mathbb{E}_Z \mathrm{P}_n^{(k)}(Z) \, . \tag{2.5}$$

2.2 Reminder on the BBP Phase Transition for the Linear Model

Before jumping to the main result of this paper, let us first recall the classical result for the linear model, as the result for the non-linear model will be similar.

We say that a matrix $\boldsymbol{W}$ is a **centered and normalized Wigner matrix with bounded fourth moment** if up to the symmetry constraint, $W_{ij} = W_{ji}$, the entries $\sqrt{N} W_{ij}$ are independent and distributed according to a law $\mathbb{P}(.)$ which is independent of N, with mean zero and variance one, and admits a finite fourth moment. For such matrices, it is well known that as $N \to \infty$ the limiting spectral distribution is given by the semi-circular distribution with support between -2 and 2 (see for example [4]) and the operator norm converges to 2 since the fourth moment is bounded [8].

Similarly, if we introduce the two deterministic piecewise functions

$$
\mathsf{l}(\tilde{\gamma}, \sigma) := 2\sigma \operatorname{sign}(\tilde{\gamma}) \, \mathbb{1}_{(|\tilde{\gamma}|<\sigma)} + \left(\tilde{\gamma} + \frac{\sigma^2}{\tilde{\gamma}}\right) \mathbb{1}_{(|\tilde{\gamma}|\geqslant\sigma)} \quad \text{and} \quad \mathsf{m}(\tilde{\gamma}, \sigma) := \left(1 - \frac{\sigma^2}{\tilde{\gamma}^2}\right) \mathbb{1}_{(|\tilde{\gamma}|\geqslant\sigma)}, \tag{2.6}
$$

where $\mathbb{1}_{(x\geqslant c)}$ is the indicator function, that is $\mathbb{1}_{(x\geqslant c)} = 1$ if $x \geqslant c$ and is null otherwise, we have the following result for the rank-one perturbation of such matrices:

Theorem 2.1 (*[45, 51] Rank-one perturbation of Wigner matrices*)

Let $\boldsymbol{W}$ be a centered and normalized Wigner matrix with bounded fourth moment, $\mathbf{u} \in \mathbb{R}^N$ such that $\|\mathbf{u}\| \to 1$ as $N \to \infty$, $\sigma > 0$ and $\tilde{\gamma}$ real, then if we denote by λ_1 and $\boldsymbol{v}_1$ the leading eigenvalue and corresponding eigenvector of $\sigma\boldsymbol{W} + \tilde{\gamma}\mathbf{u}\mathbf{u}^\top$, we have:

$$
\lambda_1 \xrightarrow[N\to\infty]{\text{a.s.}} \mathsf{l}(\tilde{\gamma}, \sigma), \tag{2.7}
$$

and

$$
\langle \boldsymbol{v}_1, \mathbf{u} \rangle^2 \xrightarrow[N\to\infty]{\text{a.s.}} \mathsf{m}(\tilde{\gamma}, \sigma). \tag{2.8}
$$

where $\mathsf{l}(.,.)$ and $\mathsf{m}(.,.)$ are defined in Eq. (2.6).

Note that, up to re-scaling of $\tilde{\gamma}$ and λ_1, one can always set $\sigma = 1$ without loss of generality. Yet, forecasting the result of the non-linear case of the next section, it will be convenient to consider the general setting $\sigma > 0$. We refer the readers to the works of Capitaine et al. [19] and Diaconu [24] for, respectively, the study of the fluctuations around these deterministic limits and the case where the fourth moment is infinite.

2.3 Phase Transition for the Non-Linear Model

We are now ready to state the main result of the paper:

Theorem 2.2 (*Rank-one Equivalence for the non-linear model*)

Suppose that f, μ_Z and π_X satisfy hypotheses **(H1)**, **(H2)** and **(H3)**. Let $k_\star$ be the generalized information index associated with f and suppose $\gamma(N)$ satisfies

$$
\frac{Q(\log(N))|\gamma(N)|^{k_\star+1}}{N^{\frac{k_\star}{2}}} \to 0
$$

for every polynomial Q. Then, for every $\varepsilon > 0$ and all N sufficiently large depending on ε, with high probability we have the following decomposition:

$$
\boldsymbol{Y}^{(f)} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\, \boldsymbol{W} + \mathbf{P} + \boldsymbol{E}, \tag{2.9}
$$

where

1. $\boldsymbol{W}$ is a centered and normalized Wigner matrix with bounded fourth moment,

2. $\mathbf{P}$ is the rank-one matrix:

$$
\mathbf{P} := \frac{\gamma(N)^{k_\star}}{N^{\frac{k_\star}{2}-\frac{1}{2}}} \frac{\vartheta_{k_\star}(f)}{k_\star!} \left[\frac{\mathbf{x}^{k_\star}}{\sqrt{N}}\left(\frac{\mathbf{x}^{k_\star}}{\sqrt{N}}\right)^\mathsf{T}\right], \tag{2.10}
$$

3. $\boldsymbol{E}$ is a symmetric matrix with operator norm bounded by ε with high probability.

Moreover, $\boldsymbol{W}$ and $\mathbf{P}$ are independent.

Remark 2.4. If f is sufficiently smooth and satisfies (**H̃3**) then one can get explicitly the error rate $\|\boldsymbol{E}\|_{\mathrm{op}} = O\left(\frac{Q(\log(N))|\gamma(N)|^{k_\star+1}}{N^{\frac{k_\star}{2}}}\right)$ (see Lemma 5.1). The function $Q(.)$ in this error rate is explicitly given in Sec. 4 and only depends on the stretched exponential decay of the distributions π_X and μ_Z, and in particular can be set to be a constant function whenever the two have a finite support.

Note that the matrix $\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\,\boldsymbol{W}$ corresponds to the case without rank-one perturbation ($\gamma(N) = 0$). Thus, Theorem 2.2 states that in the high-dimensional regime, the non-linear model effectively behaves as a *linear model* with the original rank-one perturbation $\gamma(N)(\mathbf{x}/\sqrt{N})(\mathbf{x}/\sqrt{N})^\top$ replaced by the matrix $\mathbf{P}$ of Eq. (2.10). This phenomenon, where one can effectively lift the non-linearity, has been observed in other matrix models, see for example Refs. [31, 32, 21, 46, 49].

Lastly, let's remark that, unlike many standard results in RMT, this result is *non-universal* by nature as the coefficients $\vartheta_k(f)$ depend on both the non-linearity function f *and* the noise distribution μ_Z, for example replacing μ_Z by a different distribution $\mu_{Z'}$ might lead to a different scaling of the SNR (see Cor. 2.1 bellow).

As a consequence of this theorem, it follows that the correct scaling of the SNR $\gamma(N)$ is such that the non-zero eigenvalue of the matrix $\mathbf{P}$ of Eq. (2.10) is of order one, which leads to the following result.

Corollary 2.1 (*Relevant Scaling*)

Under the same hypothesis as in Thm. 2.2, the relevant scaling of the SNR is given

$$\gamma(N) = \gamma_0 N^{\frac{1}{2}\left(1 - \frac{1}{k_\star}\right)}. \tag{2.11}$$

*Furthermore, if f satisfies (**H̃3**) we have $\|\boldsymbol{E}\|_{\mathrm{op}} = O\left(Q(\log(N))\,N^{-\frac{1}{2k_\star}}\right) \to 0$ at this scale.*

Note that $\gamma(N)$ is independent of N only if $k_\star = 1$. In particular and as a sanity check if f is the identity map, one retrieves the usual BBP setting with $\mathbf{u} \equiv \mathbf{x}/\sqrt{Nm_2}$. Let's also remark that whenever the noise is symmetric ($Z \overset{\mathcal{D}}{=} -Z$) and f is even, one can easily check that $\vartheta_1(f) = 0$ and hence $k_\star \geqslant 2$, leading to a N-dependent scaling of the SNR for the rank-one perturbation to survive in the large dimensional regime. Note also the bound for the error term in Cor. 2.1 gets significantly worse as one increases the critical index $k_\star$, implying that, in practice, one needs to *push to larger values of N* for the non-linear matrix $\boldsymbol{Y}^{(f)}$ to effectively 'look like' a rank-one perturbation of a Wigner matrix, see Fig. 1 and Fig. 2 for illustrations.

As $\boldsymbol{Y}^{(f)}$ behaves as the standard rank-one perturbation of a Wigner matrix, one can get the limiting position of the possible outlier and the overlap of the top eigenvector thanks to Thm. 2.1. Note that $\mathbf{x}^{k_\star}/\sqrt{N}$ is not normalized to have a unit norm but since the x_i are i.i.d with stretched exponential tails, one has $\|\mathbf{x}^{k_\star}\|^2/N \xrightarrow[N\to\infty]{\text{a.s}} m_{2k_\star}$, leading to the following result.

Corollary 2.2 (*The BBP Transition*)

Under the same Hypothesis of Theorem 2.2 and under the relevant scaling of Eq. (2.11), if we denote by λ_1 and $\boldsymbol{v}_1$ the leading eigenvalue and eigenvector of $\boldsymbol{Y}^{(f)}$, we have:

$$\lambda_1 \xrightarrow[N\to\infty]{\text{a.s.}} \mathsf{I}\left(\gamma_0^{k_\star}\frac{\vartheta_{k_\star}(f)}{k_\star!}\,m_{2k_\star}\,,\,\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\right), \tag{2.12}$$

and

$$\left\langle \boldsymbol{v}_1, \frac{\mathbf{x}^{k_\star}}{\|\mathbf{x}^{k_\star}\|}\right\rangle^2 \xrightarrow[N\to\infty]{\text{a.s.}} \mathsf{m}\left(\gamma_0^{k_\star}\frac{\vartheta_{k_\star}(f)}{k_\star!}\,m_{2k_\star}\,,\,\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\right). \tag{2.13}$$

where $\mathsf{I}(.,.)$ and $\mathsf{m}(.,.)$ are defined in Eq. (2.6).

After the phase transition, the leading eigenvector $\boldsymbol{v}_1$ becomes partially aligned with $\mathbf{x}^{k_\star}$ and *not* the

10

original vector $\mathbf{x}$. In particular, whenever $k_\star$ is even, even after the transition, the matrix $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ loses all information on the sign of all the entries of x_ix_j.

Let's conclude this section regarding the assumptions on the signal vector $\mathbf{x}$. One should expect to lift the assumption (**H1**) to extend the result to the case where the entries of $\mathbf{x}$ are independent but not necessarily identically distributed, for example only being identically distributed (and different from zero) inside each block $B_n(N)$ of a partition of N conditioned to $|B_n(N)|/N \to \rho_n$[1]. Similarly, one can hope to replace the global independence condition by a *weakly dependence condition*, for example by replacing $\mathbf{x}/\sqrt{N}$ by a vector $\boldsymbol{\sigma}$ drawn uniformly over $\mathbb{S}^{N-1}$, since for large N, $\boldsymbol{\sigma}$ behaves as a standard Gaussian vector normalized by its norm. However, one **cannot** replace $\mathbf{x}/\sqrt{N}$ with *any* vector $\mathbf{u}$ with a fixed norm. Indeed, a crucial condition for the result to hold is that the vector $\mathbf{x}/\sqrt{N}$ is *delocalized*, with each entry carrying approximately the same weight of order $O(N^{-1/2})$ and replacing $\mathbf{x}/\sqrt{N}$ by a *localized* vector (say for example $\mathbf{e}_1 = (1,0,\ldots,0)$) will lead to a different scaling.

3 Variants and Generalizations

In this section, we extend our result for natural variants and generalizations of Thm. 2.2.

3.1 Rank-K Deformation of Wigner Matrices

A natural extension of the previous setting is to replace the rank-one perturbation inside the non-linearity of Eq. (2.1) by a rank-K matrix where K is an integer independent of N, that is we consider the matrix $\boldsymbol{Y}_K^{(f)}$ with entries $Y_{K,ij}$ given by:

$$Y_{K,ij} := \frac{1}{\sqrt{N}}\left[f\left(Z_{ij} + \sum_{l=1}^K \frac{\gamma_l(N)}{\sqrt{N}}x_{i,l}x_{j,l}\right) - \mathbb{E}f(Z_{ij})\right]. \tag{3.1}$$

For $l = 1,\ldots,K$, we suppose the column vectors independent and distributed according to $\mathbf{x}_l := (x_{1,l},\ldots,x_{N,l}) \sim \pi_{X,l}^{\otimes N}(.)$ where the distributions $\pi_{X,l}$ might be different for different index l. Let's remark that, in general, the family of vectors $\{\mathbf{x}_l\}_{l=1,\ldots,K}$ does not form an orthogonal basis[2]. In the following, Hyp. (**H1**) has to be understood for each $\pi_{X,l}$. The result of this generalization is as follows.

Theorem 3.1 (*Small-Rank Equivalence For Non-Linear Rank-K Perturbation*)

*Suppose that Hypothesis (**H1**), (**H2**) and (**H̃3**) hold, then with high probability, the matrix model $\boldsymbol{Y}_K^{(f)}$ has the following decomposition:*

$$\boldsymbol{Y}_K^{(f)} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\,\boldsymbol{W} + \mathbf{P}_K + \boldsymbol{E}\,, \tag{3.2}$$

where

1. *$\boldsymbol{W}$ is centered and normalized Wigner matrix with entries with a bounded fourth moment,*

2. *$\mathbf{P}_K$ is the finite-rank matrix:*

$$\mathbf{P}_K := \frac{\vartheta_{k_\star}(f)}{N^{\frac{k_\star}{2}}k_\star!}\left[\sum_{l=1}^K \gamma_l(N)\mathbf{x}_l\mathbf{x}_l^\mathsf{T}\right]^{\odot k_\star}, \tag{3.3}$$

3. *$\boldsymbol{E}$ is a symmetric matrix with operator norm bounded by $\frac{Q(\log(N))|\gamma(N)|^{k_\star+1}}{N^{\frac{k_\star}{2}}}$ with high probability, with $Q(.)$ bounded by a polynomial function and $\gamma(N) := \max_l |\gamma_l(N)|$.*

Furthermore, $\boldsymbol{W}$ and $\mathbf{P}_K$ are independent.

---

[1]note that one should replace $m_{2k_\star}$ by the appropriate limit of $\|\mathbf{x}^{k_\star}\|^2/N$ in this case.

[2]however if one further assumes one of the K distributions (say $\pi_{X,1}(.)$ without loss of generality) to have zero mean then for any $l \neq 1$ we have $\langle \mathbf{x}_1/\sqrt{N}, \mathbf{x}_l/\sqrt{N}\rangle = 0$ with high probability, such that two different normalized vectors are almost surely orthogonal in the large N limit.

As before, the relevant scaling of each SNR $\gamma_l(N)$ is given by $\gamma_l(N) = O(N^{\frac{1}{2}(1-\frac{1}{k_\star})})$.

Remark 3.1 (rank of $\mathbf{P}_K$). For $K \geqslant 2$, if for each $l = 1,\ldots,K$, one has the relevant scaling $\gamma_l(N) = \gamma_{0,l} N^{\frac{1}{2}(1-\frac{1}{k_\star})}$, the rank of the matrix $\mathbf{P}_K$ is - unlike the case $K = 1$ - always higher than the one of the original perturbation matrix. In general, the rank of $\mathbf{P}_K$ is given as the number of different terms in its expression, that is $\mathrm{rank}(\mathbf{P}_K) = \binom{K+k_\star-1}{K-1} \geqslant K$. For example, in the case $K = k_\star = 2$ one has:

$$\mathbf{P}_{K=k_\star=2} = \frac{\vartheta_{k_\star}(f)}{k_\star!} \left[\gamma_{0,1}^2 \frac{\mathbf{x}_1^2}{\sqrt{N}} \left(\frac{\mathbf{x}_1^2}{\sqrt{N}}\right)^\intercal + \gamma_{0,2}^2 \frac{\mathbf{x}_2^2}{\sqrt{N}} \left(\frac{\mathbf{x}_2^2}{\sqrt{N}}\right)^\intercal + 2\gamma_{0,1}\gamma_{0,2} \frac{\mathbf{x}_1 \odot \mathbf{x}_2}{\sqrt{N}} \left(\frac{\mathbf{x}_1 \odot \mathbf{x}_2}{\sqrt{N}}\right)^\intercal \right], \quad (3.4)$$

which, in general, is of rank three and contains a 'cross-term' of the form $\mathbf{x}_1 \odot \mathbf{x}_2$.

3.2 Variance-Profile Wigner Matrices

In this section, we consider a generalization of our main result to inhomogeneous matrices. We refer to the series of work [2, 3, 1] for a study of the spectral property of these models when there is no spike.

Let $\boldsymbol{\Delta} = (\Delta_{ij})_{i,j\leqslant N} \in \mathbb{R}^{N\times N}$ be a (sequence of) **block-constant variance profile matrix**, that is given $n \geqslant 1$, there exists a partition of $[N] := \{1,\ldots,N\}$ given by

$$[N] = \bigsqcup_{s=1}^{n} I_s \tag{3.5}$$

such that the $\Delta_{i,j}$ are constant in the groups $I_s \times I_t$ for $s,t \in \{1,\ldots,n\}$

$$\Delta_{ij} = \Delta_{st}, \quad \text{for } i \in I_s, \, j \in I_t \tag{3.6}$$

and $(\Delta_{st})_{s,t\leqslant n}$ are independent of N. Furthermore, the proportions of indices in each group converges in the limit

$$\frac{|I_s|}{N} \to \rho_s \in (0,1) \quad \text{for all} \quad s \leqslant n. \tag{3.7}$$

The decomposition of $\boldsymbol{Y}^{(f)}$ in Theorem 2.2 implies the following.

Corollary 3.1 (*Small Rank Equivalence with Variance Profiles*)

Under the same Hypothesis of Theorem 2.2 and if $\boldsymbol{\Delta}$ is a variance profile matrix, we have with high probability

$$\boldsymbol{\Delta} \odot \boldsymbol{Y}^{(f)} = \boldsymbol{\Delta} \odot \left(\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2} \, \boldsymbol{W} \right) + \boldsymbol{\Delta} \odot \mathbf{P} + \boldsymbol{E}, \tag{3.8}$$

where $\boldsymbol{W}$, $\mathbf{P}$ and $\boldsymbol{E}$ are as in Theorem 2.2.

3.3 Rank-One Deformation of Rectangular Matrices

In this section, we consider the following 'rectangular' counterpart of the main result.

We say that a $(N \times M)$ matrix $\boldsymbol{X}$ is an **i.i.d standard random rectangular matrix** (with entries with bounded fourth moment) if all its entries are i.i.d from a centered distribution, independent of N, with variance one and bounded fourth moment. We recall that a **Wishart matrix $\boldsymbol{W}$** is given in terms of i.i.d standard random rectangular matrix by $\boldsymbol{W} := \left(\frac{\boldsymbol{X}}{\sqrt{M}}\right)\left(\frac{\boldsymbol{X}}{\sqrt{M}}\right)^\top$ and in the double scaling where $N, M \to \infty$ with fixed aspect ratio $q = N/M$, the empirical distribution of $\boldsymbol{W}$ converges to the *Marčenko-Pastur distribution* [40].

Similar to the rank-one deformation of Wigner matrices, the rank-one deformation of Wishart matrices of the form

$$\boldsymbol{W}' := \left(\sigma \frac{\boldsymbol{X}}{\sqrt{M}} + \gamma \mathbf{u} \mathbf{v}^\top \right) \left(\sigma \frac{\boldsymbol{X}}{\sqrt{M}} + \gamma \mathbf{u} \mathbf{v}^\top \right)^\top , \tag{3.9}$$

has been studied extensively in the literature, and the behaviors of the leading eigenvalue and leading eigenvector of $\boldsymbol{W}'$ also exhibit a phase transition depending on the value of the constants γ and q, see for example Ref. [9, 44].

Our goal is to relate the natural non-linear extension of this model, to the usual 'linear' setting. Namely, we consider the matrix

$$\boldsymbol{Y}^{(f)} := \frac{1}{M} \left(f[\boldsymbol{Z} + \frac{\gamma(N)}{\sqrt{N}} \mathbf{u} \mathbf{v}^\top] - \mathbb{E} f(Z) \mathbf{J}_{(N,M)} \right) \left(f[\boldsymbol{Z} + \frac{\gamma(N)}{\sqrt{N}} \mathbf{u} \mathbf{v}^\top] - \mathbb{E} f(Z) \mathbf{J}_{(N,M)} \right)^\top , \tag{3.10}$$

where the function f is applied entry-wise, $\boldsymbol{Z} = (Z_{ij})_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M}$ with $Z_{ij} \overset{\text{i.i.d}}{\sim} \mu_Z$, $\mathbf{u} = (u_1, \dots, u_N)$ and $\mathbf{v} = (v_1, \dots, v_M)$ with $u_i \overset{\text{i.i.d}}{\sim} \pi_U$ and $v_i \overset{\text{i.i.d}}{\sim} \pi_V$ and $\mathbf{J}_{(N,M)}$ is the matrix of size $(N \times M)$ with entries equal to one. We assume the same set of assumptions as in Sec. 2.1 (where Hyp. (**H1**) has to be understood for both the distribution π_U and π_V and $\boldsymbol{Z}$ is not symmetric anymore), we also assume $N \leqslant M$ without loss of generality. To study the eigenvalues of $\boldsymbol{Y}^{(f)}$, we consider the symmetrized matrix $\overline{\boldsymbol{Y}^{(f)}} \in \mathbb{R}^{(N+M) \times (N+M)}$ defined by

$$\overline{\boldsymbol{Y}^{(f)}} := \frac{1}{\sqrt{M}} \begin{bmatrix} \mathbf{0} & f[\boldsymbol{Z} + \frac{\gamma(N)}{\sqrt{N}} \mathbf{u} \mathbf{v}^\top] - \mathbb{E} f(Z) \mathbf{J}_{(N,M)} \\ f[\boldsymbol{Z} + \frac{\gamma(N)}{\sqrt{N}} \mathbf{u} \mathbf{v}^\top]^\top - \mathbb{E} f(Z) \mathbf{J}_{(M,N)} & \mathbf{0} \end{bmatrix} , \tag{3.11}$$

The matrix $\overline{\boldsymbol{Y}^{(f)}}$ has $2N$ non-zero eigenvalues coming by pair $\{+\lambda_i, -\lambda_i\}_{1 \leqslant i \leqslant N}$ where the λ_i's are given as the square root of the N eigenvalues of $\boldsymbol{Y}^{(f)}$[3]. One can express this new matrix $\overline{\boldsymbol{Y}^{(f)}}$ as

$$\overline{\boldsymbol{Y}^{(f)}} = \boldsymbol{\Delta} \odot \tilde{\boldsymbol{Y}}_f \quad \text{with } \boldsymbol{\Delta} := \sqrt{1 + \frac{N}{M}} \begin{bmatrix} \mathbf{0} & \mathbf{J}_{(N,M)} \\ \mathbf{J}_{(M,N)} & \mathbf{0} \end{bmatrix} \tag{3.12}$$

$$\text{and } \tilde{\boldsymbol{Y}}_f := \frac{1}{\sqrt{N+M}} \left(f[\tilde{\boldsymbol{Z}} + \frac{\tilde{\gamma}(N+M)}{\sqrt{N+M}} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top] - \mathbb{E} f(Z) \mathbf{J}_{(N+M, N+M)} \right) , \tag{3.13}$$

where $\tilde{\boldsymbol{Z}}$ is now a $((N + M) \times (N + M))$ *symmetric* matrix with iid entries distributed according to μ_Z, $\tilde{\gamma}(N + M) := \gamma(N)\sqrt{1 + M/N}$, $\tilde{\mathbf{x}} := [\mathbf{u}, \mathbf{v}] \in \mathbb{R}^{N+M}$. Note that even though the entries of $\tilde{\mathbf{x}}$ are only identically distributed inside each block $B_1 := (1, \dots, N)$ and $B_2 := (N + 1, \dots, N + M)$ but not globally, a quick check of the proof indicates that Corollary 3.1 still holds. As a result, one can get the asymptotic decomposition of $\overline{\boldsymbol{Y}^{(f)}}$, and hence of $\boldsymbol{Y}^{(f)}$, as the spectrum of the two matrices are in one-to-one correspondence, up to the trivial $M - N$ zero eigenvalues of $\overline{\boldsymbol{Y}^{(f)}}$. Eventually the factors $\sqrt{1 + N/M}$ cancel out and one gets the following result for the original matrix $\boldsymbol{Y}^{(f)}$.

---

[3]the reader not familiar with this classical result may find it by simply looking at the characteristic polynomial of $\overline{\boldsymbol{Y}^{(f)}}$ and express it in terms of the one of $\boldsymbol{Y}^{(f)}$.

Corollary 3.2 (*Rank 1 Equivalence for Rectangular Matrices*)

*Assume that Hypotheses (**H1**), (**H2**), and (**H̃3**) hold, for $\boldsymbol{Y}^{(f)}$ given by Eq. (3.10). Then, in the scaling limit where $M(N) = \lfloor N/q \rfloor$ with fixed aspect ratio q, we have with high probability the following decomposition:*

$$\boldsymbol{Y}^{(f)} \overset{\mathcal{D}}{=} \left(\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2} \frac{\boldsymbol{X}}{\sqrt{M}} + \mathbf{P} \right) \left(\sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2} \frac{\boldsymbol{X}}{\sqrt{M}} + \mathbf{P} \right)^{\top} + \boldsymbol{E}, \tag{3.14}$$

where

1. *$\boldsymbol{X}$ is i.i.d standard random rectangular matrix,*

2. *$\mathbf{P}$ is the rank-one matrix*

$$\mathbf{P} = \frac{\gamma(N)^{k_\star}}{N^{\frac{k_\star}{2} - \frac{1}{2}}} \frac{\vartheta_{k_\star}(f)}{k_\star!} \left[\frac{\mathbf{u}^{k_\star}}{\sqrt{N}} \left(\frac{\mathbf{v}^{k_\star}}{\sqrt{M}} \right)^{\top} \right], \tag{3.15}$$

3. *$\boldsymbol{E}$ is a symmetric matrix with operator norm bounded by $\frac{Q(\log(N))|\gamma(N)|^{k_\star + 1}}{N^{\frac{k_\star}{2}}}$ with high probability.*

Furthermore, $\boldsymbol{X}$ and $\mathbf{P}$ are independent. In particular, it follows that the perturbation $\mathbf{P}$ is of non-trivial order precisely when $\gamma = O(N^{\frac{1}{2}\left(1 - \frac{1}{k_\star}\right)})$ and $\|\boldsymbol{E}\|_{\mathrm{op}} = N^{-\frac{1}{2k_\star}} \to 0$ at this scale.

Remark 3.2. Corollary 3.2 can be generalized to non-smooth f satisfying Hypothesis (**H3**)a with a sequence of smooth approximations (**H3**). We will lose the explicit rate on $\|\boldsymbol{E}\|_{\mathrm{op}}$ as in Theorem 2.2.

4 Outline of the Proof

The main idea of the proof is to notice that $\frac{\gamma(N)}{\sqrt{N}} x_i x_j$ is of order $\frac{\gamma(N)}{\sqrt{N}}$ and hence if $\gamma(N)$ does not grow too fast with N, this term is small compared to $Z_{ij} \approx O(1)$, such that one can hope to perform and control the expansion of $f\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}} x_i x_j\right)$ around $f(Z_{ij})$ in terms of powers of $x_i x_j$.

This is done in practice by comparing the result to the case where one adds additional smoothness and boundedness assumptions to the model for which such expansion can be easily controlled and then one obtains the general case by approximations of the former case.

Precisely, we first obtain in Sec. 5 the main result of this paper in the case where we add to the assumptions (**H1**), (**H2**) and (**H̃3**) the following assumptions

(**H̃5**) The entries of $\mathbf{x}$ and $\boldsymbol{Z}$ are bounded;

as a result of Taylor's theorem.

Next, in Sec. 6 we detail how to remove the assumption (**H̃5**) and replace the assumption (**H̃3**) by the more general assumption (**H3**) using concentration and tail inequalities.

Eventually, in Sec. 7, we prove that if f satisfies (**H3**)a and the noise satisfies the smooth conditions (**H4**), we can explicitly construct a sequence $(f_t)_t$ of *cut-smoothed-and-regularized* approximations of f such that (**H3**) is satisfied.

5 Analysis for Smooth Functions

In this section, we describe the main intuition behind the derivation of the rank-one equivalence.

Lemma 5.1 (*Rank-one Equivalence for the non-linear model*)

*Suppose that Hypotheses **(H1)**, **(H2)**, **(H̃3)** and **(H̃5)** hold, then with high probability we have the following decomposition:*

$$\boldsymbol{Y}^{(f)} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\, \boldsymbol{W} + \mathbf{P} + \boldsymbol{E}\,, \tag{5.1}$$

where $\boldsymbol{W}, \mathbf{P}, \boldsymbol{E}$ are as in Thm. 2.2. Furthermore, for $Q(x) = x^{2k_\star+2}$, $\boldsymbol{E}$ has an operator norm bounded by

$$\frac{Q(\log(N))}{N^{\frac{1}{2}}} |\gamma(N)|^{1+\frac{1}{k_\star}} \tag{5.2}$$

with high probability.

Proof. Under Hypothesis **(H̃5)**, there exists a constant $L \geqslant 1$ such that $\max_{1 \leqslant i \leqslant N} |x_i| \leqslant L$ almost surely. Although it is not needed in this proof, we will write our error terms explicitly with respect to the constant L, which will be useful in future computations.

We start by doing Taylor expansion of our function around Z_{ij}, since we assume f is $C^{k_\star+1}$ in **(H̃3)**,

$$Y_{ij} = \frac{1}{\sqrt{N}}(f(Z_{ij}) - \mathbb{E}_Z f(Z)) + \sum_{k=1}^{k_\star} \frac{1}{\sqrt{N}} \frac{f^{(k)}(Z_{ij})}{k!} \left[\frac{\gamma(N)}{\sqrt{N}} x_i x_j\right]^k + \frac{f^{(k_\star+1)}(\xi_{ij})}{(k_\star+1)!} \left[\frac{\gamma(N)}{\sqrt{N}} x_i x_j\right]^{k_\star+1}, \tag{5.3}$$

where $\xi_{ij} \in [Z_{ij} - |\frac{\gamma}{\sqrt{N}} x_i x_j|, Z_{ij} + |\frac{\gamma}{\sqrt{N}} x_i x_j|]$. We now approximate each term in the Taylor expansion

1. **Constant Term**: The first term leads to the matrix

$$\frac{1}{\sqrt{N}} [f(\boldsymbol{Z}) - E_Z f(Z)\mathbf{I}] := \frac{1}{\sqrt{N}} [(f(Z_{ij}) - \mathbb{E}_Z f(Z))]_{1 \leqslant i,j \leqslant N}\,, \tag{5.4}$$

where the entries $(f(Z_{ij}) - \mathbb{E}_Z f(Z))$ are i.i.d (up to the symmetry), of mean zero, variance $\vartheta_0(f^2) - \vartheta_0(f)^2$ and have a bounded fourth moment since $f \in L^4(\mu_Z)$. This term gives rise to the scaled matrix $\boldsymbol{W}$ in Theorem 2.2.

2. **Sub Critical Terms** $(k < k_\star)$**:** The non-critical terms admit the following decomposition

$$\frac{\gamma(N)^k}{N^{\frac{k}{2}-\frac{1}{2}}} \left[\frac{f^{(k)}(Z_{ij})}{k!}\left[\frac{x_i^k}{\sqrt{N}} \frac{x_j^k}{\sqrt{N}}\right] - \frac{\mathbb{E}_Z f^{(k)}(Z)}{k!}\left[\frac{x_i^k}{\sqrt{N}} \frac{x_j^k}{\sqrt{N}}\right] + \frac{\mathbb{E}_Z f^{(k)}(Z)}{k!}\left[\frac{x_i^k}{\sqrt{N}} \frac{x_j^k}{\sqrt{N}}\right]\right] \tag{5.5}$$

where the first term can be controlled using the fact that the matrix

$$\frac{1}{\sqrt{N}k!} \left[f^{(k)}(\boldsymbol{Z}) - E_Z f^{(k)}(Z)\mathbf{I}\right] := \frac{1}{\sqrt{N}} \left[\frac{1}{k!}(f^{(k)}(Z_{ij}) - \mathbb{E}_Z f^{(k)}(Z))\right]_{1 \leqslant i,j \leqslant N}\,, \tag{5.6}$$

is again a Wigner matrix, which has a bounded operator norm since $f^{(k)} \in L^4(\mu_Z)$ thanks to the additional assumption **(H̃3)**. Since the entries of x are uniformly bounded thanks to the additional assumption **(H̃5)**, the diagonal matrix $\mathbf{D}(\mathbf{x}^k) := \mathrm{Diag}(x_1^k, \ldots, x_N^k)$ is also bounded in operator norm which leads to the following bound:

$$\frac{\gamma(N)^k}{N^{\frac{k}{2}}k!} \left\|\left[\frac{f^{(k)}(Z_{ij})}{\sqrt{N}}\left[x_i^k x_j^k\right] - \frac{\mathbb{E}_Z f^{(k)}(Z)}{\sqrt{N}}\left[x_i^k x_j^k\right]\right]_{ij}\right\|_{\mathrm{op}}$$
$$= \frac{\gamma(N)^k}{N^{\frac{k}{2}}k!} \left\|\mathbf{D}(\mathbf{x}^k)\left(\frac{f^{(k)}(\boldsymbol{Z})}{\sqrt{N}} - \frac{\mathbb{E}_Z f^{(k)}(Z)}{\sqrt{N}}\right)\mathbf{D}(\mathbf{x}^k)\right\|_{\mathrm{op}} \leqslant O\left(\frac{Q(L)|\gamma(N)|^k}{N^{\frac{k}{2}}}\right)\,, \tag{5.7}$$

with high probability. Hence the matrix coming from the first term in (5.5) can be absorbed in the error matrix $\boldsymbol{E}$. We recall that if f is smooth, we have $\vartheta_k(f) = \mathbb{E}_Z f^{(k)}(Z)$ which is null by definition of $k_\star$ and since $k < k_\star$. Hence the second term in (5.5) vanishes.

3. **Critical Terms $k_\star$:** The reasoning for the critical term is identical: we have again Eq. (5.5) with $k = k_\star$, but now the remainder term is non-zero and given by:

$$\frac{\gamma(N)^{k_\star}}{N^{\frac{k_\star}{2}-\frac{1}{2}}} \mathbb{E}_Z \frac{f^{(k_\star)}(Z)}{k_\star!}\left[\frac{x_i^{k_\star}}{\sqrt{N}}\frac{x_j^{k_\star}}{\sqrt{N}}\right] = \frac{\gamma(N)^{k_\star}}{N^{\frac{k_\star}{2}-\frac{1}{2}}}\frac{\vartheta_{k_\star}(f)}{k_\star!}\left[\frac{x_i^{k_\star}}{\sqrt{N}}\frac{x_j^{k_\star}}{\sqrt{N}}\right], \tag{5.8}$$

which is nothing else than the (ij) entry of our matrix $\mathbf{P}$ in Theorem 2.2.

4. **Remainder Terms:** The last terms can be written in matrix form as

$$\boldsymbol{R} := \frac{\gamma(N)^{k_\star+1}}{(k_\star+1)! N^{\frac{k_\star}{2}+1}}\mathbf{D}(\mathbf{x}^{k_\star+1})\left(f^{(k_\star+1)}(\boldsymbol{\xi})\right)\mathbf{D}(\mathbf{x}^{k_\star+1}). \tag{5.9}$$

By our assumption (**H̃3**), we have $\|f^{(k_\star+1)}\|_\infty \leqslant C$. We can now obtain a L^2 bound on the spectrum for some universal constant $\bar{C}$ that only depends on C

$$\|\boldsymbol{R}\|_{\mathrm{F}}^2 := \mathrm{Tr}(\boldsymbol{R}^2) \leqslant \bar{C}Q(L)^2 N^2 \frac{\gamma(N)^{2k_\star+2}}{((k_\star+1)!)^2 N^{k_\star+2}}, \tag{5.10}$$

which goes to zero if $\frac{|\gamma(N)|^{2k_\star+2}}{N^{k_\star}} \to 0$. Since

$$\|\boldsymbol{R}\|_{\mathrm{op}} \leqslant \|\boldsymbol{R}\|_{\mathrm{F}} \leqslant O\left(\frac{Q(L)|\gamma(N)|^{k_\star+1}}{N^{\frac{k_\star}{2}}}\right), \tag{5.11}$$

we can absorb this into the error term $\boldsymbol{E}$.

$\square$

6 Analysis for Functions Satisfying Hypothesis (H3)

In this section, we generalize the result in Section 5 to general (not necessarily smooth) functions f satisfying Hypothesis (**H3**)a which admit a sequence of smooth approximations (**H3**). We also remove the conditions that Z and x have bounded support in Hypothesis (**H̃5**), but satisfy Hypothesis (**H1**) and Hypothesis (**H2**).

6.1 Bounding the Tails

We begin by weakening the assumptions in Hypothesis (**H̃5**). Given $L > 1$, consider the following smooth cutoff function

$$\chi_L(x) = \begin{cases} 1 & |x| \leqslant L-1 \\ 0 & |x| > L \end{cases}, \tag{6.1}$$

and interpolated such that $\chi_L(x)$ is smooth for $|x| \in [L-1, L]$. We can further assume that this smooth approximation satisfies, for some integer number K large enough,

$$\sup_{0 \leqslant k \leqslant K} \|\chi_L^{(k)}\|_\infty < \infty. \tag{6.2}$$

Such a cutoff function can be generated by taking a mollification of the indicator function. We define the following truncated versions of f and x,

$$f_L(x) = f(x)\chi_L(x), \tag{6.3}$$

and

$$x_i^L = x_i \mathbb{1}_{(|x_i| \leqslant L)} + L \mathbb{1}_{(|x_i| > L)}. \tag{6.4}$$

We define the following matrix

$$\boldsymbol{Y}_L^{(f_L)} = \frac{1}{\sqrt{N}}\left[f_L\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i^L x_j^L\right) - \mathbb{E}f_L(Z_{ij})\right]. \tag{6.5}$$

The main result of this section is the following estimate which states that the truncated matrix $\boldsymbol{Y}_L^{(f_L)}$ and $\boldsymbol{Y}^{(f)}$ are equivalent with high probability.

Lemma 6.1 (*Truncation Lemma*)

> *Suppose that Hypothesis (**H1**) and Hypothesis (**H2**) and Hypothesis (**H3**)a hold. Let $\gamma = O((\log(N))^{2k_\star+1}N^{\frac{1}{2}(1-\frac{1}{k_\star})})$. Assume that the exponents of the stretched exponential decay of π_X and μ_Z are bounded below by some $\alpha > 0$ and assume that $L \geq L(N) = (\frac{3}{c}\log(N))^{\frac{1}{\alpha}}$ for some small enough constant c. We have*
>
> $$\mathbb{P}(\|\boldsymbol{Y}^{(f)} - \boldsymbol{Y}_L^{(f_L)}\|_\infty \neq 0) \leqslant (N + N^2)e^{-cL^\alpha}.$$
>
> *goes to zero polynomially fast in $1/N$.*

Proof. Notice that

$$
\begin{aligned}
\{\|\boldsymbol{Y}^{(f)} - \boldsymbol{Y}_L^{(f_L)}\|_\infty \neq 0\} &= \{(\boldsymbol{Y}^{(f)})_{ij} = (\boldsymbol{Y}_L^{(f_L)})_{ij} \ \forall i, j\}^C \\
&= \bigcup_{i \leqslant N}\{x_i \neq x_i^L\} \cup \bigcup_{i,j \leqslant N}\{f(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i^L x_j^L) \neq f_L(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i^L x_j^L\}
\end{aligned}
$$

By the union bound, it follows that

$$\mathbb{P}(\|\boldsymbol{Y}^{(f)} - \boldsymbol{Y}_L^{(f_L)}\|_\infty \neq 0) \leqslant N\mathbb{P}(x_1 \neq x_1^L) + N^2\mathbb{P}\left(f(Z_{12} + \frac{\gamma(N)}{\sqrt{N}}x_1^L x_2^L) \neq f_L(Z_{12} + \frac{\gamma(N)}{\sqrt{N}}x_1^L x_2^L)\right) \tag{6.6}$$

$$\leqslant N\mathbb{P}(x_1 \geqslant L(N)) + N^2\mathbb{P}\left(|Z_{12}| \geqslant L(N) - \frac{\gamma(N)L(N)^2}{\sqrt{N}}\right). \tag{6.7}$$

Next, notice that $\frac{\gamma(N)L(N)^2}{\sqrt{N}} \leqslant \frac{L(N)}{2}$ for N sufficiently large. By the stretched exponential bound on the tails in Hypothesis (**H1**) and Hypothesis (**H2**), it follows that

$$\mathbb{P}(\|\boldsymbol{Y}^{(f)} - \boldsymbol{Y}_L^{(f_L)}\|_\infty \neq 0) \leqslant (N + N^2)e^{-cL(N)^\alpha} \tag{6.8}$$

which goes to zero for our choice of L. $\qquad\square$

In particular, moving forward it suffices to study the spectrum of $\boldsymbol{Y}_L^{(f_L)}$ instead of $\boldsymbol{Y}^{(f)}$.

6.2 Smooth Approximations

By Hypothesis (**H3**)b and Hypothesis (**H3**)e, we can find a sequence of smooth f_t converging to f in $L^4(\mu_Z)$ such that

$$\vartheta_k(f_t) \xrightarrow[t \to \infty]{} \vartheta_k(f) \tag{6.9}$$

for all $k \leqslant k_\star$. We can construct a perturbation of the smoothing function such that the critical exponent is preserved, instead of being arbitrarily close. We define

$$g(x) := \sum_{l=0}^{k_\star-1} c_l(f_t)x^l, \tag{6.10}$$

where the coefficients $c_l(f_t)$ are chosen to satisfy

$$\mathbb{E}_Z g^{(k)}(Z) = \sum_{l=k}^{k_\star-1} \frac{l!}{(l-k)!}c_l(f_t)\mathbb{E}[Z^{l-k}] = -\vartheta_k(f_t), \tag{6.11}$$

for all $k \leqslant k_\star - 1$. Such coefficients can be found recursively by setting

$$c_{k_\star-1}(f_t) = \frac{-1}{(k_\star-1)!}\vartheta_{k_\star-1}(f_t),$$

17

$$c_k(f_t) = \frac{1}{k!}\left(-\vartheta_k(f_t) - \sum_{l=k+1}^{k_\star-1} \frac{l!}{(l-k)!} c_l(f_t)\mathbb{E}[Z^{l-k}]\right) \text{ for } k < k_\star - 1, \tag{6.12}$$

(the system of equations we have to solve is an upper triangular matrix with non-zero diagonal so such coefficients exist). In particular, the constants $c_l(f_t)$ are continuous functions of the $\vartheta_k(f_t)$. It follows that

$$f_t^{\text{pert}}(x) := f_t(x) + g(x), \tag{6.13}$$

preserves the critical exponent by linearity

$$\vartheta_k(f_t^{\text{pert}}) = 0 \quad \text{for} \quad k < k_\star \quad \text{and} \quad \vartheta_{k_\star}(f_t^{\text{pert}}) \neq 0. \tag{6.14}$$

We now prove the main result of this subsection, which is an approximation theorem, which states that the spectrums of the smoothed random matrix $\boldsymbol{Y}_L^{(f_t^{\text{pert}})}$ is a good approximation for the spectrum of $\boldsymbol{Y}_L^{(f_L)}$.

Lemma 6.2 (*Approximation Lemma*)

*Let f satisfy Hypothesis (**H3**)c and Hypothesis (**H3**)b and suppose that $\gamma(N)$ is such that $\frac{\gamma(N)L(N)^2}{\sqrt{N}} \to 0$. For every fixed $t \geqslant 0$ and N sufficiently large depending on N, we have with high probability that*

$$\|\boldsymbol{Y}_L^{(f_L)} - \boldsymbol{Y}_L^{(f_t^{\text{pert}})}\|_{\text{op}} \leqslant O(\|f - f_t^{\text{pert}}\|_{L^4(\mu_Z)}^2 + \|f - f_t^{\text{pert}}\|_{L^4(\mu_Z)}^4) \leqslant o_t(1) + o_N(1),$$

where $\lim_{t\to\infty} o_t(1) = 0$ and $\lim_{t\to\infty}\lim_{N\to\infty} o_N(1)$.

Proof. We first fix an x and do the computation conditionally on x.

Step 1: Notice that f_L is globally Lipschitz with constant $Q(L(N))$ and f_t is globally Lipschitz with constant $\|f_t'\|_\infty < \infty$ because it is compactly supported and smooth by Hypothesis (**H3**)c. The operator norm is convex and a Lipschitz functions of its entries, so Talagrand's concentration inequality implies that

$$\mathbb{P}_Z(|\|\boldsymbol{Y}_L^{(f_L)} - \boldsymbol{Y}_L^{(f_t^{\text{pert}})}\|_{\text{op}} - \mathbb{E}_Z\|\boldsymbol{Y}_L^{(f_L)} - \boldsymbol{Y}_L^{(f_t^{\text{pert}})}\|_{\text{op}}| \geqslant t) \leqslant e^{-\frac{t^2 N}{C_t(L)^2}}. \tag{6.15}$$

The constant $C_t(L) = C(Q(L(N)) + \|f_t'\|_\infty)^2$ is independent of x and only depends on the Lipschitz constants of f and f_t^{pert}, and is at most a polynomial in L. Therefore, the upperbound will tend to 0 as $N \to \infty$. In particular, with high probability the operator norm is close to its expected value.

Step 2: We define $\tilde{f}(Z) = f_L(Z + \frac{\gamma(N)}{\sqrt{N}} x_i x_j)$ and $\tilde{f}_t^{\text{pert}}(Z) = f_t^{\text{pert}}(Z + \frac{\gamma(N)}{\sqrt{N}} x_i x_j)$ to be shifted functions. We will show that

$$\mathbb{E}_Z\|\boldsymbol{Y}_L^{(f_L)} - \boldsymbol{Y}_L^{(f_t^{\text{pert}})}\|_{\text{op}} \leqslant O(\|\tilde{f} - \tilde{f}_t^{\text{pert}}\|_{L^4(\mu_Z)}^4 + \|f_L - f_t^{\text{pert}}\|_{L^4(\mu_Z)}^4). \tag{6.16}$$

Conditionally on $\mathbf{x}$, the matrix $\boldsymbol{Y}_L^{(f_L)} - \boldsymbol{Y}_L^{(f_t^{\text{pert}})}$ has independent entries, we can apply Latala's Theorem ([34] or [20, Lemma 6.3]) which bounds the expectation of the operator norm of a matrix with independent entries by the L^2 norm of its rows and the L^4 norm of its entries. Integrating with respect to the noise variables conditionally on $\mathbf{x}$

$$\frac{1}{N}\mathbb{E}\left(f_L\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}} x_i x_j\right) - \mathbb{E}f_L(Z_{ij}) - f_t^{\text{pert}}\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}} x_i x_j\right) + \mathbb{E}f_t^{\text{pert}}(Z_{ij})\right)^2 \tag{6.17}$$

$$\leqslant \frac{2}{N}\mathbb{E}\left(f_L\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}} x_i x_j\right) - f_t^{\text{pert}}\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}} x_i x_j\right)\right)^2 + \frac{2}{N}\mathbb{E}\left(\mathbb{E}f_L(Z_{ij}) - \mathbb{E}f_t^{\text{pert}}(Z_{ij})\right)^2 \tag{6.18}$$

18

$$\leqslant \frac{2}{N}\|\tilde{f}_L - \tilde{f}_t^{\text{pert}}\|^2_{L^2(\mu_Z)} + \frac{2}{N}\|f_L - f_t^{\text{pert}}\|^2_{L^1(\mu_Z)}. \tag{6.19}$$

A similar computation for the fourth moment implies that

$$\frac{1}{N^2}\mathbb{E}\left(f_L\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i x_j\right) - \mathbb{E}f_L(Z_{ij}) - f_t\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i x_j\right) - \mathbb{E}f_t(Z_{ij})\right)^4$$

$$\leqslant \frac{8}{N^2}\mathbb{E}\left(f_L\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i x_j\right) - f_t\left(Z_{ij} + \frac{\gamma(N)}{\sqrt{N}}x_i x_j\right)\right)^4 + \frac{8}{N^2}\mathbb{E}\left(\mathbb{E}f_L(Z_{ij}) - \mathbb{E}f_t(Z_{ij})\right)^4$$

$$\leqslant \frac{8}{N^2}\|\tilde{f}_L - \tilde{f}_t^{\text{pert}}\|^4_{L^4(d\mu_{\mathcal{Z}})} + \frac{8}{N^2}\|f_L - f_t\|^4_{L^1(d\mu_{\mathcal{Z}})}.$$

Therefore, Latala's Theorem implies (6.16). Since $L^4(\mu_Z) \subseteq L^2(\mu_Z) \subseteq L^1(\mu_Z)$ and our assumption on the integrability of f, it suffices to control the fourth moment which gives us (6.16).

Step 3: We now control the norm in the right hand side of (6.16). By the classical inequality

$$(a+b)^4 \leqslant 2^4(a^4 + b^4), \tag{6.20}$$

there exists a universal constant $C_{k_\star}$ such that

$$\|f_L - f_t^{\text{pert}}\|^4_{L^4(\mu_Z)} = C_{k_\star}\left(\|f_L - f_t\|^4_{L^4(\mu_Z)} + \sum_{k<k_\star} c_k^4(f_t)\|x^k\|^4_{L^4(\mu_Z)}\right). \tag{6.21}$$

The second term is small order because $c_k^4(f_t)$ goes to zero as $\vartheta_\ell(f_t)$ go to zero for $\ell < k_\star$ by (6.12) and Z has finite moments by Hypothesis (**H2**).

To control the first term, the triangle inequality implies that

$$\|f_L - f_t^{\text{pert}}\|^4_{L^4(\mu_Z)} \leqslant 2^4\|f - f_L\|^4_{L^4(\mu_Z)} + 2^4\|f - f_t\|^4_{L^4(\mu_Z)}. \tag{6.22}$$

The first term tends to 0 as $N \to \infty$ while the second term goes to 0 by Hypothesis (**H3**)b.

Next, we need to control

$$\|\tilde{f}_L - \tilde{f_t^{\text{pert}}}\|^4_{L^4(\mu_Z)} \leqslant 2^4\|\tilde{f} - \tilde{f}_L\|^4_{L^4(\mu_Z)} + 2^4\|\tilde{f} - \tilde{f}_t\|^4_{L^4(\mu_Z)}. \tag{6.23}$$

Notice that for any $\varepsilon > 0$ and all N sufficiently large $c = \frac{\gamma(N)}{\sqrt{N}}x_i^L x_j^L \in [-\varepsilon, \varepsilon]$ almost surely because $x_i^L x_j^L \in [-L(N)^2, L(N)^2]$ so $|\frac{\gamma(N)}{\sqrt{N}}x_i^L x_j^L| \to 0$ for all N sufficiently large by our Theorem assumptions. Since f is locally Lipschitz with polynomial constant, there exists a polynomial P with bounded degree such that

$$\mathbb{E}_Z(\tilde{f})^4 \leqslant 2^4\mathbb{E}_Z f^4 + 2^4\mathbb{E}(f(Z+c) - f(Z))^4 \leqslant 2^4\mathbb{E}_Z f^4 + 2^4 c^4\mathbb{E}P(|Z+c|) < \infty$$

because $f \in L^4$ and Hypothesis (**H2**) implies that all moments of Z are finite. This bound is also holds almost surely in c. Therefore, the first term of (6.23) goes to zero. To control the second term, notice that

$$\mathbb{E}\|f(Z+c) - f_t(Z+c)\|^4_{L^4(\mu_Z)} \tag{6.24}$$

$$\leqslant 3^4(\mathbb{E}\|f(Z+c) - f(Z)\|^4_{L^4(\mu_Z)} + \mathbb{E}\|f_t(Z+c) - f_t(Z)\|^4_{L^4(\mu_Z)} + \mathbb{E}\|f(Z) - f_t(Z)\|^4_{L^4(\mu_Z)}) \tag{6.25}$$

The first and second term are clearly bounded by $O(\varepsilon^4)$ because f is locally Lipschitz with polynomial constant and f_t is Lipschitz because it is smooth and compactly supported by Hypothesis (**H3**)c. The third term tends to zero by the convergence Hypothesis (**H3**)b.

We, therefore, conclude that

$$\|f - f_t^{\text{pert}}\|^4_{L^4(\mu_Z)} = o_t(1) + o_N(1), \tag{6.26}$$

uniformly for all c. All the bounds are held almost surely in x and are independent of x, so our results hold unconditionally as well, so our smooth approximations are exact in the limit. $\qquad\square$

6.3 A Smooth Approximation of the Main Theorem

We now combine the arguments to prove Theorem 2.2.

Proof of Theorem 2.2. By the truncation Lemma 6.1 we have that with high probability that $\boldsymbol{Y}^{(f)}$ and $\boldsymbol{Y}_L^{(f_L)}$ have the same spectrum. The approximation Lemma 6.2, we also have that with high probability,

$$\boldsymbol{Y}_L^{(f_L)} \overset{\mathcal{D}}{=} \boldsymbol{Y}_L^{(f_t^{\mathrm{pert}})} + \boldsymbol{E}_t + \boldsymbol{E}_{(N)}, \tag{6.27}$$

where $\|\boldsymbol{E_t}\|_{\mathrm{op}} \leqslant o_t(1)$, $\|\boldsymbol{E_{(N)}}\|_{\mathrm{op}} \leqslant o_N(1)$ when N is sufficiently large depending on t.

For every $t \geqslant 1$, the function f_t^{pert} satisfies the conditions of Hypothesis (**H̃3**) because of Hypothesis (**H3**)b. We can conclude that with high probability we have the following decomposition:

$$\boldsymbol{Y}_L^{(f_t^{\mathrm{pert}})} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\,\boldsymbol{W} + \mathbf{P} + \boldsymbol{E}_{(N)}. \tag{6.28}$$

We conclude by Lemma 6.1 that with high probability and all N sufficiently large depending on t that

$$\boldsymbol{Y}^{(f)} \overset{\mathcal{D}}{=} \sqrt{\vartheta_0(f^2) - \vartheta_0(f)^2}\,\boldsymbol{W} + \mathbf{P} + \boldsymbol{E}_{(N)} + \boldsymbol{E}_t \tag{6.29}$$

where the operator norm of $\|\boldsymbol{E}_{(N)}\|_{\mathrm{op}} \to 0$ as $N \to \infty$ for any fixed ε. In particular, given any $\varepsilon > 0$, we may fix a t such that $\|\boldsymbol{E}_t\|_{\mathrm{op}} \leqslant \varepsilon/2$ and then take N sufficiently large depending on ε such that $\|\boldsymbol{E}_{(N)}\|_{\mathrm{op}} < \varepsilon/2$ to conclude the proof. □

Remark 6.1. If we assume sufficient regularity on the function f and are able to find a sequence of approximating f_t^{pert} with uniformly bounded $k_\star + 1$ derivatives for all n, then an explicit bound of the operator norm $\boldsymbol{E}$ in terms of N can be computed.

We end this section with a remark that our proof naturally generalizes to the cases introduced in Section 3.

Proof of Theorem 3.1, Corollary 3.1, Corollary 3.2. It is easy to see that the proof of Theorem 2.2 also holds entry wise, so the proofs of the generalizations are identical to the simple case. We sketch the essential modifications. The critical observation is that the Taylor expansion done in Eq. (5.3) holds if the function f depends on the index, that is $f = f_{ij}$, and in the presence of a higher order spike. The bound in (5.7) also holds for f_{ij} in the setting with a (bounded) variance profile by Latala's theorem. The same construction of smooth approximators also holds in this setting, so Section 6.

Corollary 3.2 naturally follows from Corollary 3.1 and the Girko Hermitization trick in equation (3.11).

□

7 Approximations in the Case of Smooth Density for the Noise

In this section, we show that under the assumptions that f is locally Lipschitz and the noise Z has a sufficiently smooth density given by Hypothesis (**H4**), we explicitly construct a sequence f_t satisfying Hypothesis (**H3**).

7.1 Smooth Truncation $f_{\delta,M}$

We now construct a sequence of smooth approximations of f satisfying the Hypothesis (**H3**). We first truncate the domain of f to a compact set and then smooth this truncated function. We will show that the smoothing can be made exact by taking the smoothing parameter to 0 and then taking the truncation to infinity.

For $M > 0$, we define the following truncation

$$f_M(x) := f(x)\,\mathbb{1}_{(|x| \leqslant M)}. \tag{7.1}$$

Next, we consider a standard smoothing of the non-linearity. We define the mollifier $\eta_\delta \in C^\infty(\mathbb{R})$ by

$$\eta(x) := \frac{1}{\int_{-1}^1 \exp(\frac{1}{y^2-1})\,\mathrm{d}y} \exp\left(\frac{1}{x^2-1}\right) \mathbb{1}_{(|x|<1)} \quad \text{and} \quad \eta_\delta(x) = \frac{1}{\delta}\eta\left(\frac{x}{\delta}\right). \tag{7.2}$$

Let $f_{\delta,M}$ be its smoothing with a mollifier,

$$f_{\delta,M} = f_M * \eta_\delta. \tag{7.3}$$

We recall the following standard facts about the mollifiers, see [18, Proposition 4.18 and Theorem 4.22]:

Lemma 7.1

If $f \in L^p(\mathrm{d}x)$, then $f_{\delta,M}$ is a smooth function and $f_{\delta,M} \to f_M$ in $L^p(\mathrm{d}x)$. Furthermore,

$$supp(f_{\delta,M}) \subseteq \overline{supp(f_M) + supp(\eta_\delta)} \subseteq [-M-\delta, M+\delta]. \tag{7.4}$$

Now suppose that the density $w_Z(z)$ of Z satisfies Hypothesis (**H3**). We define

$$\vartheta_k(f) = (-1)^k \int f(x) w_Z^{(k)}(x)\mathrm{d}x. \tag{7.5}$$

We first prove that $\vartheta_k(f_{\delta,M})$ converges to $\vartheta_k(f)$.

Lemma 7.2

If Hypothesis (**H3**)a and Hypothesis (**H4**) holds, then

$$|\vartheta_k(f) - \vartheta_k(f_{\delta,M})| = o_M(\delta) + o_M(1). \tag{7.6}$$

for all $k \leqslant k_\star < \infty$. The $o_M(\delta)$ term goes to 0 as $\delta \to 0$ uniformly for all M.

Proof. Since $f_{\delta,M}$ is smooth, we can integrate by parts to conclude that

$$\vartheta_k(f_{\delta,M}) = \mathbb{E}_Z f_{\delta,M}^{(k)}(Z) = (-1)^k \int f_{\delta,M}(x) w_Z^{(k)}(x)\mathrm{d}x. \tag{7.7}$$

In particular,

$$|\vartheta_k(f_{\delta,M}) - \vartheta_k(f)| = \left| \int (f_{\delta,M}(x) - f(x)) w_Z^{(k)}(x)\mathrm{d}x \right| \leqslant \int |f_{\delta,M}(x) - f(x)|\,|w_Z^{(k)}(x)|\mathrm{d}x. \tag{7.8}$$

We decompose this integral into two regions

$$\int |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)|\mathrm{d}x$$
$$= \int_{|x|\leqslant M} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)|\mathrm{d}x + \int_{|x|>M} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)|\,\mathrm{d}x. \tag{7.9}$$

On the first region, we can use Holder's inequality and use the assumption (**H4**)a to see that

$$\int_{|x|\leqslant M} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)|\,\mathrm{d}x \tag{7.10}$$

$$\leqslant \left(\int |f_{\delta,M}(x) - f(x)|^2 \mathrm{d}x \right)^{1/2} \left(\int_{|x|\leqslant M} |w_Z^{(k)}(x)|^2 \mathrm{d}x \right)^{1/2} \tag{7.11}$$

$$\leqslant C_k \|f_{\delta,M}(x) - f_M(x)\|_{L^2(\mathrm{d}x)} = o_M(\delta), \tag{7.12}$$

where there error goes to 0 for every fixed M. On the second region, since $f_{\delta,M}$ is supported on a subset of $[-M - \delta, M + \delta]$, we have

$$
\begin{aligned}
&\int_{|x|>M} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)| \, \mathrm{d}x \\
&\leqslant \int_{|x|>M+\delta} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)| \, \mathrm{d}x + \int_{M<|x|<M+\delta} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)| \, \mathrm{d}x.
\end{aligned}
\tag{7.13}
$$

The first term tends to 0 as $M \to \infty$ uniformly for all δ since f is integrable so the integral of its tail converges. For fixed M, we have using the assumption that $w_Z^{(k)}$ is uniformly bounded

$$
\sup_{M<|x|<M+\delta} |f_{\delta,M}(x) - f(x)||w_Z^{(k)}(x)| \leqslant C(M) \,,
\tag{7.14}
$$

so the second integral vanishes if we send $\delta \to 0$. Therefore, we have

$$
\lim_{M\to\infty} \lim_{\delta\to 0} |\vartheta_k(f_{\delta,M}) - \vartheta_k(f)| = 0 \,.
\tag{7.15}
$$

$\square$

We now prove that there exists a parametrization $(\delta(t), M(t))$ such that $(\delta(t), M(t)) \to (0, \infty)$ and $f_t = f_{\delta(t),M(t)}$ satisfies Hypothesis **(H3)**.

Proposition 7.1

> *There exists a sequence $(\delta(t), M(t)) \to (0, \infty)$ such that $f_t = f_{\delta(t),M(t)}$ satisfies Hypothesis **(H3)**.*

Proof. We prove each property separately.

*Hypothesis **(H3)**b:* We have

$$
\|f - f_{\delta,M}\|_{L^4(\mu_Z)}^4 = \mathbb{E}_Z[(f(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(|Z|\leqslant M)}] + \mathbb{E}_Z[(f(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(|Z|>M)}] \,.
\tag{7.16}
$$

On the set $|Z| \leqslant M$ the fact that the density $\frac{d\mathbb{P}_Z}{dz} = w_Z$ is bounded by Hypothesis **(H2)** implies that

$$
\begin{aligned}
\mathbb{E}_Z[(f(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(|Z|\leqslant M)}] &= \mathbb{E}_Z[(f_M(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(|Z|\leqslant M)}]
\tag{7.17} \\
&\leqslant \mathbb{E}_Z[(f_M(Z) - f_{\delta,M}(Z))^4]
\tag{7.18} \\
&\leqslant C\|f_M - f_{\delta,M}\|_{L^4(\mathrm{d}x)}^4 = o_M(\delta) \,,
\tag{7.19}
\end{aligned}
$$

because f_M is compactly supported so $f_{\delta,M} \to f_M$ in $L^4(\mathrm{d}x)$.

On the set, $|Z| > M$, we have $f_{\delta,M}(Z) = 0$ for $Z \in [-M - \delta, M + \delta]^c$ so using the Lipschitz property of f

$$
\begin{aligned}
&\mathbb{E}_Z[(f(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(|Z|>M)}]
\tag{7.20} \\
&= \mathbb{E}_Z[(f(Z) - f_{\delta,M}(Z))^4 \, \mathbb{1}_{(M\leqslant|Z|\leqslant M+\delta))}] + \mathbb{E}_Z[(f(Z))^4 \, \mathbb{1}_{(|Z|\geqslant M+\delta)}]
\tag{7.21} \\
&\leqslant \sup_{M\leqslant x\leqslant M+\delta} |f(x) - f_{\delta,M}(x)|\mathbb{P}(M \leqslant |Z| \leqslant M + \delta) + o(M) = o_M(\delta) + o(M) \,,
\tag{7.22}
\end{aligned}
$$

where $o_M(\delta) \to 0$ as $\delta \to 0$ for every fixed M. In the third line, we used the fact that $f \in L^4(\mu_Z)$ so its tails are integrable by the dominated convergence theorem and the fact that f is Lipschitz. We can control our error by first sending $\delta \to 0$ then $M \to \infty$. In particular, we can jointly define a sequence by taking δ sufficiently small depending on M, which gives us our sequence $(\delta(t), M(t))$.

*Hypothesis **(H3)**c:* This is immediate by Lemma 7.1.

*Hypothesis (**H3**)d:* We will prove that $\|f_{M,\delta}^{(k_*+1)}\| \leqslant C(M,\delta)$ for some constant that does not depend on N. By Young's inequality and the fact that the perturbation term is a polynomial of degree less than $k_\star$,

$$\left\|\frac{\mathrm{d}^{k^\star+1}}{\mathrm{d}x^{k^\star+1}}\left(f_M * \eta_\delta + \sum_{l=0}^{k^*-1} c_l(f_{\delta,M})x^l\right)\right\|_\infty \leqslant \|f_M * \eta_\delta^{(k_\star+1)}\|_\infty \leqslant \|f_M\|_{L^1(\mathrm{d}x)}\|\eta_\delta^{(k_\star+1)}\|_\infty. \tag{7.23}$$

We have

$$\|f_M\|_{L^1(\mathrm{d}x)} \leqslant 2M\left(\sup_{x\in[-M,M]} f(x)\right) \tag{7.24}$$

and

$$\|\eta_\delta^{(k_\star+1)}\|_\infty \leqslant \frac{1}{\delta^{k_\star+1}}\|\eta^{(k_\star+1)}\|_\infty. \tag{7.25}$$

Both of these constants are independent of N, so we have control of the remainder term for free.

By the same argument, we also get that the fourth moment assumption is automatically satisfied for $k < k_\star$ because

$$\left\|\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(f_M * \eta_\delta + \sum_{l=0}^{k^*-1} c_l(f_{\delta,M})x^l\right)\right\|_{L^4(\mu_Z)}$$
$$= \|f_M * \eta_\delta^{(k)} + \sum_{l\geqslant k} c_l(f_{\delta,M})x^l\|_{L^4(\mu_Z)}$$
$$\leqslant C_{k_\star,k}\left(\|f_M^4\|_{L^1(\mathrm{d}x)}\|\eta_\delta^{(k)}\|_{L^4(\mu_Z)} + \sum_{l\geqslant k} c_l(f_{\delta,M})\|x^l\|_{L^4(\mu_Z)}\right).$$

This quantity can clearly be bounded independently of N, because we assumed control of the moments of Z in Hypothesis (**H4**). And in particular, the fourth moment condition is satisfied automatically if we assume sufficient integrability on Z.

*Hypothesis (**H3**)e:* This is immediate by Lemma 7.2. We can take a modify our sequence by possible taking δ even smaller depending on M such that Hypothesis (**H3**)b is satisfied as well. $\qquad\square$

References

[1] Oskari H. Ajanki, László Erdos, and Torben Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, Probability Theory and Related Fields **173** (2019), no. 1, 293–373.

[2] Oskari H. Ajanki, László Erdos, and Torben Krüger, *Universality for general Wigner-type matrices*, Probability Theory and Related Fields **169** (2017), no. 3, 667–727.

[3] Oskari H. Ajanki, Torben Krüger, and László Erdos, *Singularities of Solutions to Quadratic Vector Equations on the Complex Upper Half-Plane*, Communications on Pure and Applied Mathematics **70** (2017), no. 9, 1672–1705, Publisher: John Wiley & Sons, Ltd.

[4] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge Studies in Advanced Mathematics, vol. 118, Cambridge University Press, Cambridge, 2010. MR 2760897

[5] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath, *Online stochastic gradient descent on non-convex losses from high-dimensional inference*, The Journal of Machine Learning Research **22** (2021), no. 1, 4788–4838.

[6] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý, *Random Matrices and Complexity of Spin Glasses*, Communications on Pure and Applied Mathematics **66** (2013), no. 2, 165–201, Publisher: John Wiley & Sons, Ltd.

[7] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang, *High-dimensional asymptotics of feature learning: How one gradient step improves the representation*, Advances in Neural Information Processing Systems **35** (2022), 37932–37946.

[8] Z. D. Bai and Y. Q. Yin, *Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix*, The Annals of Probability **16** (1988), no. 4, 1729–1741.

[9] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, The Annals of Probability **33** (2005), no. 5, 1643 – 1697.

[10] Jinho Baik and Jack W. Silverstein, *Eigenvalues of large sample covariance matrices of spiked population models*, Journal of Multivariate Analysis **97** (2006), no. 6, 1382–1408.

[11] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu, *Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization*, IEEE Transactions on Information Theory **64** (2018), no. 7, 4872–4894.

[12] Jean barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová, *Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula*, Advances in Neural Information Processing Systems (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[13] Florent Benaych-Georges and Romain Couillet, *Spectral analysis of the Gram matrix of mixture models*, ESAIM: Probability and Statistics **20** (2016), 217–237 (en), Publisher: EDP-Sciences.

[14] Florent Benaych-Georges and Raj Rao Nadakuditi, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Advances in Mathematics **227** (2011), no. 1, 494–521.

[15] ———, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, Journal of Multivariate Analysis **111** (2012), 120–135.

[16] Giulio Biroli and Alice Guionnet, *Large deviations for the largest eigenvalues and eigenvectors of spiked random matrices*, 2019.

[17] Alex Bloemendal and Bálint Virág, *Limits of spiked random matrices I*, Probability Theory and Related Fields **156** (2013), no. 3, 795–825.

[18] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer New York, 2010.

[19] M. Capitaine, C. Donati-Martin, and D. Féral, *Central limit theorems for eigenvalues of deformations of Wigner matrices*, Annales de l'I.H.P. Probabilités et statistiques **48** (2012), no. 1, 107–133 (en). MR 2919200

[20] Nicholas A. Cook, Raphael Ducatez, and Alice Guionnet, *Full large deviation principles for the largest eigenvalue of sub-Gaussian Wigner matrices*, 2023.

[21] Cosme Louart, Zhenyu Liao, and Romain Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability **28** (2018), no. 2, 1190–1248.

[22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi, *Neural networks can learn representations with gradient descent*, Conference on Learning Theory, PMLR, 2022, pp. 5413–5452.

[23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan, *Learning two-layer neural networks, one (giant) step at a time*, arXiv preprint arXiv:2305.18270 (2023).

[24] Simona Diaconu, *Finite Rank Perturbations of Heavy-Tailed Wigner Matrices*, 2022.

[25] E Marinari, G Parisi, and F Ritort, *Replica field theory for deterministic models. II. A non-random spin glass with glassy behaviour*, Journal of Physics A: Mathematical and General **27** (1994), no. 23, 7647.

[26] F. Krzakala, J. Xu, and L. Zdeborová, *Mutual information in rank-one matrix estimation*, 2016 IEEE Information Theory Workshop (ITW), September 2016, Journal Abbreviation: 2016 IEEE Information Theory Workshop (ITW), pp. 71–75.

[27] Florent Benaych-Georges, Alice Guionnet, and Mylène Maida, *Fluctuations of the Extreme Eigenvalues of Finite Rank Deformations of Random Matrices*, Electronic Journal of Probability **16** (2011), no. none, 1621–1662.

[28] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová, *Generalisation error in learning with random features and the hidden manifold model*, International Conference on Machine Learning, PMLR, 2020, pp. 3452–3462.

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

[30] C. Itzykson and J.-B. Zuber, *The planar approximation. II*, Journal of Mathematical Physics **21** (2008), no. 3, 411–421.

[31] Noureddine El Karoui, *On information plus noise kernel random matrices*, The Annals of Statistics **38** (2010), no. 5, 3191 – 3216.

[32] _____, *The spectrum of kernel random matrices*, The Annals of Statistics **38** (2010), no. 1, 1 – 50.

[33] J. P. Keating and N. C. Snaith, *Random matrix theory and $\zeta(1/2 + it)$*, Communications in Mathematical Physics **214** (2000), no. 1, 57–89.

[34] Rafał Latała, *Some estimates of norms of random matrices*, Proceedings of the American Mathematical Society **133** (2005), no. 5, 1273–1282.

[35] Marc Lelarge and Léo Miolane, *Fundamental limits of symmetric low-rank matrix estimation*, Conference on Learning Theory, PMLR, 2017, pp. 1297–1301.

[36] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová, *Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering*, 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2016, pp. 601–608.

[37] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, *Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2015, pp. 680–687.

[38] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, *Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications*, J. Stat. Mech. Theory Exp. (2017), no. 7, 073403, 86. MR 3683819

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

[40] V A Marčenko and Leonid A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mathematics of The Ussr-sbornik **1** (1967), 457–483.

[41] Madan Lal Mehta, *Random matrices*, Elsevier, 2004.

[42] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Communications on Pure and Applied Mathematics **75** (2022), no. 4, 667–766.

[43] Mylène Maida, *Large deviations for the largest eigenvalue of rank one deformations of Gaussian ensembles*, Electronic Journal of Probability **12** (2007), no. none, 1131–1150.

[44] Debashis Paul, *Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model*, Statistica Sinica **17** (2007), no. 4, 1617–1642, Publisher: Institute of Statistical Science, Academia Sinica.

[45] S. Péché, *The largest eigenvalue of small rank perturbations of Hermitian random matrices*, Probability Theory and Related Fields **134** (2005), no. 1, 127–173.

[46] S. Péché, *A note on the Pennington-Worah distribution*, Electronic Communications in Probability **24** (2019), no. none, 1 − 7.

[47] Jeffrey Pennington and Pratik Worah, *Nonlinear random matrix theory for deep learning*, Advances in Neural Information Processing Systems (I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[48] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra, *Optimality and suboptimality of pca i: Spiked random matrix models*, The Annals of Statistics **46** (2018), no. 5, 2416–2451.

[49] Vanessa Piccolo and Dominik Schröder, *Analysis of one-hidden-layer neural networks via the resolvent method*, Advances in Neural Information Processing Systems (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, eds.), vol. 34, Curran Associates, Inc., 2021, pp. 5225–5235.

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

[51] David Renfrew and Alexander Soshnikov, *On Finite Rank Deformations of Wigner Matrices II: Delocalized Perturbations*, Random Matrices: Theory and Applications **02** (2013), no. 01, 1250015.

[52] Romain Couillet and Florent Benaych-Georges, *Kernel spectral clustering of large dimensional data*, Electronic Journal of Statistics **10** (2016), no. 1, 1393–1454.

[53] Terence Tao, *Topics in random matrix theory*, vol. 132, American Mathematical Society, 2023.

[54] Jacobus Verbaarschot, *Spectrum of the qcd dirac operator and chiral random matrix theory*, Phys. Rev. Lett. **72** (1994), 2531–2533.

[55] Eugene Paul Wigner, *On the distribution of the roots of certain symmetric matrices*, Annals of Mathematics **67** (1958), 325.

[56] John Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika **20A** (1928), no. 1/2, 32–52.

[57] Yunzhen Yao, Liangzu Peng, and Manolis C. Tsakiris, *Unsigned matrix completion*, 2021 IEEE International Symposium on Information Theory (ISIT), 2021, pp. 2250–2255.