The Spiked Matrix Model With Generative Priors

Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, Member, IEEE, and Lenka Zdeborová

Abstract—We investigate the statistical and algorithmic properties of random neural-network generative priors in a simple inference problem: spiked-matrix estimation. We establish a rigorous expression for the performance of the Bayes-optimal estimator in the high-dimensional regime, and identify the statistical threshold for weak-recovery of the spike. Next, we derive a message-passing algorithm taking into account the latent structure of the spike, and show that its performance is asymptotically optimal for natural choices of the generative network architecture. The absence of an algorithmic gap in this case is in stark contrast to known results for sparse spikes, another popular prior for modelling low-dimensional signals, and for which no algorithm is known to achieve the optimal statistical threshold. Finally, we show that linearising our message passing algorithm yields a simple spectral method also achieving the optimal threshold for reconstruction. We conclude with an experiment on a real data set showing that our bespoke spectral method outperforms vanilla PCA.

Index Terms—Signal processing algorithms, principal component analysis (PCA), inverse problems, statistical learning, unsupervised learning.

I. INTRODUCTION

Exploiting structure for efficient signal reconstruction is a central endeavour in modern signal processing. Notable technological advances - such as JPEG and MP3 compression for example - stem from the fact that images and sound admit a sparse representation in wavelet and Fourier bases. In a seminal work, Donoho, Candès and Tao have shown that underparametrised linear systems can be inverted if the signal is assumed to be sparse. This result opened the door for novel sub-Nyquist sampling strategies leveraged by sparsity which are at the heart of Compressed Sensing [1]. But interest in sparse representations reaches far beyond Compressed Sensing, and similar results have been derived for other signal processing tasks, such as sparse coding and sparse principal component analysis (PCA). Despite the remarkable success of these results, they broadly assume the latent sparse representation is given, thus relying on expert knowledge for signal pre-processing.

Recent progress in deep learning has witnessed a surge of interest in neural network-based generative models. Opposed to sparsity, generative networks are trained to learn a latent representation of the structured signal. The expressiveness of neural networks allied with the capacity to capture hierarchical representations led to impressive results in signal modelling, the most notable perhaps being Generative Adversarial Networks (GANs), which can be trained to generate realistic images of human faces [2]. An important and natural question to ask is whether signals from generative models enjoy the same aforementioned interesting properties as sparse signals in reconstruction tasks. Early results in regression-related problems suggest that the latent structure in generative models can be leveraged to improve signal reconstruction [3]–[9], indeed suggesting that [10]:

Generative models are the new sparsity.

In this manuscript we give a further step in this direction by analysing a class of random-neural generative priors in an unsupervised task: rank-one (a.k.a. spiked) matrix factorisation. Given a "data" matrix \( Y \in \mathbb{R}^{n \times p} \), the problem consists in finding two vectors (known as spikes) \( u \in \mathbb{R}^n \) and \( v \in \mathbb{R}^p \) such that \( Y \) can be factorised as \( Y = uv^T + \Delta \xi \), where \( \xi \) is an i.i.d. noise matrix of unit variance. This model is widely studied as a prototype for principal component analysis (PCA), since for small noise (\( \Delta < 1 \)) and Gaussian spikes \( u, v \), the optimal estimator is given by the leading principal component of \( Y \) [11]. Optimality relies on the assumption of unstructured spikes, and no longer hold if one of the spikes is sparse. In a similar spirit to Compressed Sensing, the investigation of sparse spikes in this model resulted into bespoke algorithms widely studied under the umbrella of sparse-PCA, e.g. [12].

An important conclusion of the aforementioned works is the existence of a algorithmic gap for sparse signal reconstruction. In other words, even if signal reconstruction is \( \text{a priori} \) possible, no polynomial-time algorithm is known. For spiked-matrix factorisation, this means that even though the best known sparse-PCA algorithm perform better than vanilla PCA, it doesn’t reach the optimal threshold set by the theoretical, and practically intractable, Bayesian estimator. As we will show, this is in sharp contrast to the class of neural generative models we study, for which we provide a
polynomial time algorithm reaching the optimal theoretical performance, suggesting instead that:

Generative models are better than sparsity.

Before moving to the bulk of the technical analysis, we give a detailed introduction of the model and regime will study, followed by an account of our main contributions.

II. MODEL AND STUDIED REGIME

We will focus on the following two widely studied models in the sparse-PCA literature [13]–[19]:

a) Spiked Wigner model \((\mathbf{w}^\top)\): Consider an unknown vector \((the spike)\) \(\mathbf{v}^* \in \mathbb{R}^p\) drawn from a distribution \(P_v\); we observe a matrix \(Y \in \mathbb{R}^{p \times p}\) with a symmetric noise term \(\xi \in \mathbb{R}^{p \times p}\) and \(\Delta > 0\):

\[
Y = \frac{1}{\sqrt{p}} \mathbf{v}^* \mathbf{v}^\top + \sqrt{\Delta} \xi,
\]

where \(\xi_{ij} \sim \mathcal{N}(0, 1)\) i.i.d. The aim is to find back the hidden spike \(\mathbf{v}^*\) from \(Y\) (up to a global sign).

b) Spiked Wishart (or spiked covariance) model \((\mathbf{w}^\top)\): Consider two unknown vectors \(\mathbf{u}^* \in \mathbb{R}^p\) and \(\mathbf{v}^* \in \mathbb{R}^p\) drawn from distributions \(P_u\) and \(P_v\) and let \(\xi \in \mathbb{R}^{n \times p}\) with \(\xi_{uij} \sim \mathcal{N}(0, 1)\) i.i.d. and \(\Delta > 0\), we observe

\[
Y = \frac{1}{\sqrt{p}} \mathbf{u}^* \mathbf{v}^\top + \sqrt{\Delta} \xi;
\]

the goal is to find back the hidden spikes \(\mathbf{u}^*\) and \(\mathbf{v}^*\) from \(Y \in \mathbb{R}^{n \times p}\).

The noisy high-dimensional limit that we consider in this paper (the thermodynamic limit) is \(p, n \to \infty\) while \(\beta \equiv n/p = \Theta(1)\), and the noise \(\xi\) has a variance \(\Delta = \Theta(1)\). The prior \(P_v\) is representing the spike \(\mathbf{v}\) via a \(k\)-dimensional parametrization with \(\alpha \equiv p/k = \Theta(1)\). In the sparse case, \(k\) is the number of non-zeros components of \(\mathbf{v}^*\), while in generative models \(k\) is the number of latent variables.

A. Considered Generative Models

The simplest non-separable prior \(P_v\) that we consider is the Gaussian model with a covariance matrix \(\Sigma\), that is \(P_v(\mathbf{v}) = \mathcal{N}_p(0, \Sigma)\). This prior is not compressive, yet it captures some structure and can be simply estimated from data via the empirical covariance. We use this prior later to produce Fig. 6.

To exploit the practically observed power of generative models, it would be desirable to consider models (e.g. GANs, variational auto-encoders, restricted Boltzmann machines, or others) trained on datasets of examples of possible spikes. Such training, however, leads to correlations between the weights of the underlying neural networks for which the theoretical part of the present paper does not apply readily. To keep tractability in a closed form, and subsequent theoretical insights, we focus on multi-layer generative models where all the weight matrices \(W^{(l)}\), \(l = 1, \ldots, L\), are fixed, layer-wise independent, i.i.d. Gaussian with zero mean and unit variance. Let \(\mathbf{v} \in \mathbb{R}^p\) be the output of such a generative model

\[
\mathbf{v} = \varphi^{(L)} \left( \frac{1}{\sqrt{k_L}} W^{(L)} \ldots \varphi^{(1)} \left( \frac{1}{\sqrt{k}} W^{(1)} \mathbf{z} \right) \ldots \right).
\]

with \(\mathbf{z} \in \mathbb{R}^k\) a latent variable drawn from separable distribution \(P_{\mathbf{z}}\), with \(\rho_z = \mathbb{E}_{P_{\mathbf{z}}} [\mathbf{z}^2]\). The \(\varphi^{(l)}\) are element-wise activation functions that can be either deterministic or stochastic. It will be useful to define the hidden variables \(\mathbf{h}^{(l+1)} \in \mathbb{R}^{k_{l+1}}\) obtained from the output of layer \(l - 1\). The hidden variable \(\mathbf{h}^{(l+1)} \in \mathbb{R}^{k_{l+1}}\) is then given by

\[
\mathbf{h}^{(l+1)} = \varphi^{(l)} \left( \frac{1}{\sqrt{k_l}} W^{(l)} \mathbf{h}^{(l)} \right) \Leftrightarrow \mathbf{h}^{(l+1)} \sim P_{\text{out}}^{(l)} \left( \cdot \left| \frac{1}{\sqrt{k_l}} W^{(l)} \mathbf{h}^{(l)} \right) \right) \text{ identically.} \quad (4)
\]

The densities \(P_{\text{out}}^{(l)}\) over \(\mathbb{R}^{k_{l+1}}\) parametrise the input/output relationship at each layer of the generative network. Note that since \(\varphi^{(l)}\) act component-wise \(P_{\text{out}}^{(l)}\) is a separable distribution, and factorise in a product of identical \(k_{l+1}\) scalar distributions over \(\mathbb{R}\) which, abusing notation, we will denote by \(P_{\text{out}}^{(l)}\). For instance, a deterministic layer \(l\) with non-linearity \(\varphi^{(l)}\) is fully characterised by the scalar density \(P_{\text{out}}^{(l)}(x | z) = \delta(x - \varphi^{(l)}(z))\).

In the setting considered in this paper the ground-truth spike \(\mathbf{v}^*\) is generated using a ground-truth value of the latent variable \(\mathbf{z}^*\). The spike is then estimated from the knowledge of the data matrix \(Y\), and the known form of the spiked-matrix and of the generative model. In particular the matrices \(W^{(l)} \in \mathbb{R}^{k_{l+1} \times k_l}\) are known, as are the parameters \(\beta, \Delta, P_z, P_u, P_v\) and \(\varphi^{(l)}\). Only the spikes \(\mathbf{v}^*, \mathbf{u}^*\) and the latent vector \(\mathbf{z}^*\) are unknown, and are to be inferred.

For concreteness and simplicity, the generative model that will be analyzed in most examples given in the present paper is the single-layer case of (3) with \(L = 1\). We define the total compression ratio \(\alpha \equiv p/k\). In what follows we will illustrate our results for \(\varphi\) being linear, sign and ReLU functions.

B. Summary of Main Contributions

First, we provide an information-theoretical analysis for the performance of the optimal estimator for the spiked-matrix models (1) and (2). This analysis is based on a rigorous expression for the mutual information between the matrix \(Y\) and a general spike \(\mathbf{v}^*\) from a non-separable distribution \(P_v\) in \(\mathbb{R}^p\), and holds in the afore defined thermodynamic limit. Evaluating this expression on the generative priors discussed in section II-A, we obtain the optimal statistical threshold \(\Delta_c\) below which the spike \(\mathbf{v}^*\) can be reconstructed. On a second moment, we derive an approximate message passing (AMP) algorithm for the models (1) and (2), and show that, for the all the generative architectures analysed, they attain the same performance previously derived for the Bayesian optimal estimator. Next, we propose a simple spectral method derived from our AMP algorithm reaching the same statistical threshold \(\Delta_c\). Finally, we show that this same spectral method can be, in certain cases, rigorously derived from a Random Matrix Theory.

Our main findings are in stark contrast to the known results for sparse-PCA, and therefore it is useful to present them in this context. We draw two main conclusions from the present work:
(i) No algorithmic gap with generative-model priors: Sharp and detailed results are known in the thermodynamic limit (as defined above) when the spike $v^*$ is sampled from a separable distribution $P_v$. A detailed account of several examples can be found in [20]. The main finding for sparse priors $P_v$ is that when the sparsity $\rho = k/p = 1/\alpha$ is large enough then there exist optimal algorithms [14], while for $\rho$ small enough there is a striking gap between statistically optimal performance and the one of best known algorithms [15]. The small-$\rho$ expansion studied in [20] is consistent with the well-known results for exact recovery of the support of $v^*$ [21], [22], which is one of the best-known cases in which gaps between statistical and best-known algorithmic performance were described.

Our analysis of the spiked-matrix models with generative priors reveals that in this case known algorithms are able to obtain (asymptotically) optimal performance even when the dimension is greatly reduced, i.e. $\alpha \gg 1$. Analogous conclusion about the lack of algorithmic gaps was reached for the problem of phase retrieval under a generative prior in [8]. This result suggests that plausibly generative priors are better than sparsity as they lead to algorithmically easier problems.

(ii) Spectral algorithms reaching statistical threshold: Arguably the most basic algorithm used to solve the spiked-matrix model is based on the leading singular vectors of the matrix $Y$. We will refer to this as PCA. Previous work on spiked-matrix models [16], [20] established that in the thermodynamic limit and for separable priors of zero mean PCA reaches the best performance of all known efficient algorithms in terms of the value of noise $\Delta$ below which it is able to provide positive correlation between its estimator and the ground-truth spike. While for sparse priors positive correlation is statistically reachable even for larger values of $\Delta$ [16], [20], no efficient algorithm beating the PCA threshold is known.

In the case of generative priors we find in this paper that other spectral methods improve on the canonical PCA. We design a spectral method, called LAMP, that (under certain assumptions, e.g. zero mean of the spikes) reach the statistically optimal threshold, meaning that for larger values of noise variance no other (even exponential) algorithm is able to reach positive correlation with the spike. Again this is a striking difference with the sparse separable prior, making the generative priors algorithmically more attractive. We demonstrate the performance of LAMP on the spiked-matrix model when the spike is taken to be one of the fashion-MNIST images showing considerable improvement over canonical PCA.

Each of the following sections is dedicated to one of the results above.

III. ANALYSIS OF INFORMATION-THEORETICALLY OPTIMAL ESTIMATION

In this section, we derive a set of fixed point equations, known as state evolution equations, that fully characterize the performance of the optimal estimator for the spike $v^*$. For the sake of concreteness, the results in this Section are given for the Wigner model, and can be fully generalised to the Wishart case.

A. Rigorous Mutual Information

From an optimisation perspective, the problem we want to solve is to find the estimator $v^*$ that minimises the mean square error (MSE)

$$\text{mse}(\Delta) = \mathbb{E}\|\hat{v} - v^*\|^2. \quad (5)$$

Since the information about the generative model $P_v$ of the spike is given, we know that the estimator minimising eq. (5) is given by the mean of the posterior distribution of the spike, i.e. $\hat{v}^{\text{opt}} = \mathbb{E}_{P(v^*|Y)}v$, where $P(v^*|Y)$ is written from Bayes rule as

$$P(v^*|Y) = \frac{1}{P(Y)}P_v(v^*) \prod_{1 \leq i < j \leq p} \frac{1}{\sqrt{2\pi \Delta}} e^{-\frac{1}{2\Delta}(v_{ij} - v_{ij}^*)^2}. \quad (6)$$

The expression above is written in full generality, and for the time being we have not assumed anything about $P_v$. The naive approach of estimating $\hat{v}^{\text{opt}}$ from exact sampling of the posterior is intractable numerically, specially in the large-dimensional limit $p \to \infty$ of interest. However, it is still possible to track the performance of the optimal estimator without direct sampling through the I-MMSE theorem connecting the minimal mean square error (MMSE) to a derivative of the mutual information between the signal and the data [24]. Following this rationale, our first main result is a rigorous expression for the mutual information between the ground-truth spike $v^*$ and the observation $Y$, defined as $I(Y;v^*) = D_{KL}(P_{v^*|Y}||P_vP_Y)$, valid in the thermodynamic limit defined in Section II.

**Theorem III.1 (Mutual Information for the Spiked Wigner Model With Structured Spike):** Informally, assume the spike $v^*$ come from a sequence (of growing dimension $p$) of a generic structured prior $P_v$ on $\mathbb{R}^p$, then

$$\lim_{p \to \infty} \inf_{\rho \geq q_v \geq 0} I(Y;v^*) = \sup_{\rho_v} \left( \inf_{\rho \geq q_v \geq 0} I(v^*;v) \right) = \inf_{\rho \geq q_v \geq 0} i_{\text{opt}}(\Delta, q_v),$$

with $i_{\text{opt}}(\Delta, q_v) = \frac{(\rho_v - q_v)^2}{4\Delta} + \lim_{p \to \infty} I\left( v^*; v + \sqrt{\frac{\Delta}{q_v}} \xi \right)$

(7)

and $\xi$ being a Gaussian vector with zero mean, unit diagonal variance and $\rho_v = \lim_{p \to \infty} \mathbb{E}_{P_v}[v^*v]/p$.

The proof for this theorem is given later in Section VI-A, and instead we draw its consequences. Our theorem connects the asymptotic mutual information of the spiked model with generative prior $P_v$ to the mutual information between $v$ taken from $P_v$ and its noisy version, $I(v^*; v + \sqrt{\Delta/q_v})$. As mentioned before, the mutual information is intimately connected to the performance of the optimal estimator, and one can prove in particular that for the spiked-matrix model [25] the MMSE on the spike $v^*$ is asymptotically given by:

$$\text{MMSE}_v = \rho_v - q_v^*, \quad (8)$$

1This result holds only for sparsity $\rho = O(1)$. A line of works shows that when sparsity $k$ scales slower than linearly with $p$, algorithms more performant than PCA exist [21], [23].
where \( q_\alpha^* \) is the optimizer of the function \( i_{RS}(\Delta, q_\alpha) \). Computing this later mutual information is itself a high-dimensional task, hard in full generality, but it can be done for a range of non-trivial \( P_v \). The simplest tractable case is when the prior \( P_v \) is separable, then it yields back exactly the formula previously known from \([17],[18],[26]\). It can also be computed for the correlated Gaussian generative model, \( P_v(v) = \mathcal{N}(0, \Sigma) \), for which \( I(v; v + \sqrt{\Delta/q_\alpha} \xi) = \text{Tr} (\log (I + q_\alpha \Sigma/\Delta)) / 2 \) is readily known.

More interestingly, the mutual information associated to the multi-layer generative prior with random weights from eq. (3), explicitly written as

\[
P_v(v) = \prod_{l=1}^{L} \frac{k_l}{\sqrt{k_l-1}} \prod_{i=1}^{P} \frac{P^{(L)}(v_{L})}{W_{v,L}} \left( v_{L} \right) \frac{1}{\sqrt{k_l}} \sum_{v_{L-1}}^{k_{L-1}} W^{(L-1)}(v_{L-1}; v_{L}) \right) \right),
\]

(9)
can also be asymptotically computed. Indeed, the corresponding single-layer formula for this mutual information has been derived and proven in \([27]\). For the multi-layer case the mutual information formula has been derived in \([5],[28]\) and proven for the case of two layers in \([29]\). Theorem III.1 together with the results from \([5],[27],[29]\) yields the following formula (see section VI-A for details) for the spiked Wigner model (1) with multi-layer generative prior (3):

\[
i_{RS}(\Delta, q_\alpha) = \frac{q_\alpha^2}{4\Delta} + \frac{1}{4\Delta} q_\alpha^2 + \frac{1}{\alpha} \text{extr}_{\{\hat{q}_l, q_l\}} \left[ \frac{1}{\alpha} \sum_{l=1}^{L} q_\alpha \hat{q}_l q_l + \sum_{l=2}^{L} \alpha_l \Psi^{(l)}_{\text{out}}(\hat{q}_l, q_{l-1}) - \alpha \Psi^{(L+1)}_{\text{out}}(q_\alpha, q_L) - \Psi_z(\hat{q}_z) \right], \tag{10}
\]

where \( \alpha_l = k_l/k_l \) (note that in particular \( \alpha_1 = 1 \)) and the functions \( \Psi_{\text{out}}, \Psi_z \) are defined by:

\[
\Psi_{\text{out}}(x, y) = \mathbb{E}_\xi \left[ Z_{\gamma} \left( x^{1/2}, \xi, x \right) \log \left( Z_{\gamma} \left( x^{1/2}, \xi, x \right) \right) \right],
\]

\[
\Psi_{\text{out}}^{(l)}(x, y) = \mathbb{E}_{\xi, \eta} \left[ Z^{(l)}_{\gamma} \left( x^{1/2}, \xi, y^{1/2}, \eta, \rho_l - y \right) \times \log \left( Z^{(l)}_{\gamma} \left( x^{1/2}, \xi, y^{1/2}, \eta, \rho_l - y \right) \right) \right],
\]

(11)

with \( \xi, \eta \sim \mathcal{N}(0, 1) \) i.i.d., \( \rho_l \) is the second moment of the hidden variable \( h_l \) and \( Z_{\gamma}, Z^{(l)}_{\gamma} \) are the normalizations of the following denoising scalar distributions:

\[
Q_{\gamma}(z; \gamma, \Lambda) = \frac{P_{\gamma}(z)}{Z_{\gamma}(\gamma, \Lambda)} e^{-\frac{1}{2} z^2 + \gamma z},
\]

\[
Q^{(l)}_{\text{out}}(v, x; B, A, \omega, V) = \frac{P^{(l)}(v|x)}{Z^{(l)}_{\text{out}}(B, \omega, V)} e^{-\frac{(v - \omega)^2}{2V} - \frac{(x - \omega)^2}{2V}} \sqrt{2\pi V},
\]

(13)

Result (10) is remarkable in that it connects the asymptotic mutual information of a high-dimensional model with a simple scalar formula that can be easily evaluated. Moreover, it fully characterise the statistical performance of the optimal estimator, allowing us to readily identify the statistical thresholds separating the region between possible and impossible inference of the spike.

We now draw the consequences of eq. (10) for the most common choices of activation.

**B. Optimal Performance and Statistical Thresholds: Phase Diagrams**

Taking the extremization over \( q_\alpha \) and \( (\hat{q}_l, q_{l-1}) \) \( l \leq L \) in eq. (10), we obtain the following system of coupled fixed point equations:

\[
\begin{align*}
q_v &= \Lambda_x (\frac{q_v}{\Delta}, q_L) \\
q_L &= \Lambda_x (\hat{q}_L, q_{L-1}) \\
\vdots
\end{align*}
\]

\[
\begin{align*}
q_l &= \Lambda_x (\hat{q}_l, q_{l-1}) \\
\vdots
\end{align*}
\]

\[
\begin{align*}
q_z &= \Lambda_z (\hat{q}_z) \\
\vdots
\end{align*}
\]

(14)

where we have defined the update functions \( \Lambda_x(x, y) \equiv 2 \partial_{x} \Psi_{\text{out}}(x, y), \Lambda_{\text{out}}(x, y) \equiv 2 \partial_{y} \Psi_{\text{out}}(x, y), \Lambda_z(x) \equiv 2 \partial_{x} \Psi_{z}(x) \) and the layer-wise aspect ratios \( \alpha_l = k_l+1/k_l = \alpha_{l+1}/\alpha_l \). As previously discussed, the fixed point of these equations provide all the information about the performance of the Bayes-optimal estimator through eq. (8).

An important first question that can be answered from eqs. (14) is when does the Bayes-optimal estimator performs better than a random guess from the prior distribution \( P_v \). For instance, we intuitively expect that when the prior is not biased towards a particular direction in \( \mathbb{R}^p \) and for very high noise \( \Delta \gg 1 \) better-than-random estimation is not possible. In terms of fixed points of eqs. (14), this situation corresponds to the existence of the non-informative fixed point \( q_\alpha^* = 0 \) (i.e. maximum MSE, \( \rho_c \), or zero overlap with the spike). Evaluating the right-hand side of eqs. (14) at \( q_v = 0 \), we can see that \( q_\alpha^* = 0 \) is a fixed point if

\[
\mathbb{E}_{P_v}[z] = 0 \quad \text{and} \quad \mathbb{E}_{Q_{\text{out}}^{(l)}}[v] = 0,
\]

(15)

where \( Q_{\text{out}}^{(l)}(v, x) \equiv Q_{\text{out}}^{(l)}(v, x; 0, 0, 0, \rho_l) \) from eq. (13). Note that for multi-layer network with deterministic channels and \( \varphi^{(l)} \equiv \varphi \) for all \( l \), the second condition is equivalent to \( \varphi \) being an odd function.

When the condition (15) holds, \( (q_v, q_L, \hat{q}_L, \ldots, \hat{q}_z, q_z) = (0, 0, \ldots, 0) \) is a fixed point of eq. (14). The numerical stability of this fixed point is determined by whether it is an attractor of the dynamics, and therefore determines a phase transition point \( \Delta_{c} \), defined as the noise below which the fixed point \( \Theta \in \mathbb{R}^{L+1} \) becomes a repeller. The character of the fixed point can be determined by a standard linear stability analysis of the fixed point equations. The transition will then correspond to the value of \( \Delta \) for which the largest eigenvalue of the Jacobian of the eqs. (14) at 0 becomes greater than one.
This Jacobian is given explicitly in (16), shown at the bottom of the page, where we have defined the following shorthand for the second moments of $Q_{\text{out}}(v, x)$:

\[
\begin{align*}
    m^{(l)}_{v,v} &= \left( \mathbb{E}_{Q_{\text{out}}^{(l),o}}(v^2) \right)^2, \\
    m^{(l)}_{v,x} &= \left( \mathbb{E}_{Q_{\text{out}}^{(l),o}}(vx) \right)^2, \\
    m^{(l)}_{x,x} &= \left( \mathbb{E}_{Q_{\text{out}}^{(l),o}}(x^2 - \rho_l^2) \right)^2, \\
    m_{zz} &= \left( \mathbb{E}_{P_z z^2} \right)^2
\end{align*}
\]  

(17)

This result is given in full generality, and it is instructive to compute $\Delta_c$ in specific cases.

First, consider the case of a single-layer generative prior $L = 1$. Fix $P_z(z) = N_z(0, 1)$ and $F_{\text{out}}^{(1)}(v|x) = \delta(v - \varphi(x))$, for $\varphi \in \{\text{linear, sign, relu}\}$. The first two choices of nonlinearity are odd, and therefore in these cases we expect a transition as discussed above. It can be readily computed from the Jacobian eq. (16) and yield $\Delta_c = 1 + \alpha$ for linear activation and $\Delta_c = 1 + \frac{1}{\rho} \alpha$ for sign activation. In both cases, since $\alpha > 0$, it is clear that knowledge of the generative prior improve reconstruction in the sense that the spike can be better reconstructed for larger amplitude of noise $\Delta$. Moreover, the larger $\alpha$ (i.e. the smaller the latent dimension with respect to the signal dimension), the better the reconstruction.

Figure 1 summarizes this discussion. We numerically solve the fixed point eqs. (14) and plot the MMSE obtained from the fixed point in a heat map, for the linear, sign, and relu activations. The white dashed line marks the threshold $\Delta_c$ obtained analytically from the Jacobian in eq. (16). The property that we find the most striking is that in these three evaluated cases, for all values of $\Delta$ and $\alpha$ that we analyzed, we always found that eq. (14) has a unique stable fixed point. Thus we have not identified, in the physics terminology, any first order phase transition. Figure 2 shows some examples of numerical MMSE curves for three nonlinearities discussed and different values of $\alpha$. The fixed point equations were solved iteratively from uncorrelated initial condition, and from initial condition corresponding to the ground truth signal, and found that both lead to the same fixed point.

This observation generalises to deeper $L > 1$ generative priors. Consider $P_z(z) = N_z(0, 1)$ and layer-wise constant activation $F_{\text{out}}^{(l)}(v|x) = \delta(v - \varphi(x))$. For the previous odd activation functions discussed, we find that

**Linear activation:** For $\varphi(x) = x$ the leading eigenvalue of the Jacobian becomes one at

\[
\Delta_c = 1 + \sum_{l=1}^{L} \frac{\alpha_l}{\rho_l^2}.
\]  

(18)

Note in particular that for $L = 1$ and in the limit $\alpha = 0$ we recover the phase transition $\Delta_c = 1$ known from the case with separable prior [20]. For $\alpha > 0$, we have $\Delta_c > 1$ meaning the spike can be estimated more efficiently when its structure is accounted for. In particular, the
we show that the MSE of our algorithm can be tracked exactly in the thermodynamic limit, and that moreover it coincides with the optimal performance discussed in Section III even for large $\alpha$. This result is particularly interesting when compared to the known performance of message passing algorithms for sparse-PCA, for which AMP is not able to reach optimal statistical performance in the small sparsity regime [30].

Approximate message passing algorithms for spiked matrix estimation with separable priors are well known [5], [31], [32]. Our derivation draw on previous works on extending AMP to non-separable priors [5], [31], [32] and we first focus on the more general Wishart model ($uv^\top$). After, we discuss how to get the corresponding result for the Wigner model ($vv^\top$) with a simple change.

### A. Derivation of the Approximate Message Passing Algorithm

Approximate message passing algorithms can be derived systematically for problems that can be written in terms of an acyclic factor graph. The standard idea is to simplify the corresponding Belief Propagation (BP) equations in the limit of a large number of variables. Together with a Gaussian ansatz for the distribution of the BP messages, the expansion of the BP yield a set of $O(k^2)$ simplified equations known as relaxed BP (rBP) equations. The last step to get the corresponding AMP algorithm is to remove the target dependency of the messages that further reduces the number of iterative equations to $O(k)$.

Our derivation is closely related to the derivation of AMP for a series of statistical inference problems with factorised priors, see for example [20] and references therein. In the interest of the reader, instead of repeating the cumbersome steps described above, we rather describe how two known and simple AMP algorithms for independent inference problems can be combined into one for the corresponding structured problem. In particular, this is illustrated for the spiked-matrix estimation with single-layer generative model prior, which can be seen as the combination of a rank-one matrix factorization problem (MF) [20] with a generalised linear model (GLM) [27]. Note that the multi-layer case follows by iterating this procedure.

1) Combining Factor Graphs: Consider the factor graphs for the MF and the GLM problems with separable priors, drawn in Fig. 4. The key idea is to replace the separable prior $P_v$ for the "structured" variable $v$ in the MF model (in green) by a factorised connection channel $P_{\text{out}}$ (see definition [27]) linking the input $v$ with the output factors of the GLM (in red). The resulting factor graph for the structured model is given in Fig. 4, with the same color code.

2) Combining AMP Algorithms: As for the factor graphs, we start by recalling the AMP update equations in the Bayes-optimal case for the two problems in question with separable priors.

\begin{itemize}
  \item [a) AMP equations for the Wishart MF layer (variables $v$ and $u$):] Consider the low-rank matrix factorization model $Y = \frac{1}{\sqrt{D}} u v^\top + \sqrt{\Delta} \xi$ with separable priors $P_u$ and $P_v$ for the variables $u$ and $v$, illustrated in Fig. 4 (upper left). The corresponding Bayes-optimal AMP equations, given in [20],

deepen the generative network for the spike, the easier estimation becomes.

**Sign activation:** For $\varphi(x) = \text{sgn}(x)$ the leading eigenvalue of the Jacobian becomes one at

$$\Delta_c = 1 + \sum_{l=1}^{L} \left(\frac{4}{\alpha^2}\right)^l \frac{\alpha}{\alpha_l}. \tag{19}$$

For $L = 1$ and $\alpha = 0$, $P_v = \text{Bern}(1/2)$, and the transition $\Delta_c = 1$ agrees with the one found for a separable prior distribution [20]. As in the linear case, for $\alpha > 0$, we can estimate the spike for larger values of noise than in the separable case, and depth also improves estimation.

Note that we also didn’t observe first order transitions for deeper networks, at least in the first-to-come-in-mind cases that we have investigated, i.e. deterministic deep networks with $\varphi(l) \equiv \varphi \in \{\text{linear, sign, relu}\}$. However, we do not expect this behaviour to be completely general neither. One can engineer a situation, for instance with a very shifted relu on the last layer, and a very large intermediate layer, so that the spike $v$ becomes effectively sparse with weakly correlated, almost independent, components, thus recovering the classical algorithmic gap [20].

So far we have only discussed the performance of the information theoretic optimal estimator - averting the question of estimating the signal itself. In the next section we close this gap by introducing an approximate message passing (AMP) algorithm for signal reconstruction. Our algorithm has the advantage that its performance can tracked down exactly in the thermodynamic limit, and we will show that in the cases we analysed it exactly follows the same fixed point equations (14) as the ones derived for optimal estimator.

### IV. Approximate Message Passing With Generative Priors

Naive sampling from the high-dimensional posterior distribution is exponentially costly, ruling this approach out from an algorithmic perspective. One should therefore appeal to algorithmically tractable approximations. Approximate Message Passing (AMP) algorithms have proven to be particularly useful for problems defined on random graphs, and successful examples abound in the literature.

In this section we derive and analyze an AMP algorithm tailored for spiked estimation with generative priors. Next,
read:

\[
\begin{align*}
\mathbf{u}^{t+1} &= f_u(B_u^t, A_u^t), \\
\mathbf{v}^{t+1} &= \partial_B f_u(B_u^t, A_u^t), \\
\mathbf{c}^{t+1} &= \partial_B f_u(B_u^t, A_u^t), \\
\mathbf{B}_o^t &= \frac{1}{k} \mathbf{Y}^\top \mathbf{u}^t - \frac{1}{p} \mathbf{1}_p \mathbf{v}^{t-1}, \\
A_v^t &= \frac{1}{k} \mathbf{1}_p^\top \mathbf{1}_p, \\
A_u^t &= \frac{1}{k} \mathbf{1}_p^\top \mathbf{1}_p, \\
P_c(u) &= \frac{1}{Z_c(B, A)} P_c(u) e^{-\frac{1}{k} A_u^t + A_v^t}, \\
P_v(v) &= \frac{1}{Z_v(B, A)} P_v(v) e^{-\frac{1}{k} A_u^t + A_v^t}.
\end{align*}
\]

where the operation \((\cdot)^2\) is taken component-wise. The update functions \(f_u\) and \(f_v\) are respectively the means of the distributions \(Q_u\) and \(Q_v\), defined similarly to eq. (13) as

\[
Q_u(u; B, A) = \frac{1}{Z(u; B, A)} P_u(u) e^{-\frac{1}{k} A_u^t + A_v^t}, \\
Q_v(v; B, A) = \frac{1}{Z(v; B, A)} P_v(v) e^{-\frac{1}{k} A_u^t + A_v^t}.
\]

b) AMP equations for the GLM layer (variable \(z\)): On the other hand, the Bayes-optimal AMP equations for the GLM model \(v^* = \varphi(\frac{1}{\sqrt{k}} W z^*)\) with \(z^* \sim P_z\), given in [27] and illustrated in Fig. 4, read

\[
\begin{align*}
\mathbf{z}^{t+1} &= f_z(\gamma^t, \Lambda^t) \\
\mathbf{c}^{t+1} &= \partial_{\gamma} f_z(\gamma^t, \Lambda^t) \\
\mathbf{g}^t &= f_{out}(v^*, \omega^t, V^t)
\end{align*}
\]

where \(f_z\) is the mean of \(Q_z\) defined in eq. (13) and \(f_{out}\) is the mean of \(V^{-1}(x - \omega)\) with respect to

\[
Q_{out}(x; v^*, \omega, V) = \frac{P_{out}(v^* | x)}{Z_{out}(v^*, \omega, V)} e^{-\frac{(x - \omega)^2}{2V}}
\]

c) Module composition: In principle, composing the AMP equations for the inference problems above is non-trivial and requires a full-blown derivation from the BP equations on the composed factor graph in Fig. 4. Surprisingly, the upshot of this cumbersome computation is rather simple: the AMP equations for the composed model are equivalent to coupling the MF eqs. (20) and the GLM eqs. (22) by replacing \(Q_v(v; B, A)\) and \(Q_{out}(x; \omega, V)\) with the following joint distribution:

\[
Q_{out}(v, x; B, A, \omega, V) = \frac{e^{-\frac{1}{k} A_u^t + A_v^t}}{Z_{out}(B, A, \omega, V)} e^{-\frac{(x - \omega)^2}{2V}}.
\]

The associated update functions \(f_v, f_{out}\) are thus replaced by the mean of \(v\) and \(V^{-1}(x - \omega)\) with respect to this new joint distribution \(Q_{out}\). Replacing the separable distributions \(Q_u\) and \(Q_{out}\) by the joint distribution eq. (24) and corresponding update functions as described above in eq. (20)-(22), we obtain the AMP algorithm for structured model. Additionally, we note that the AMP equations above are also valid for arbitrary weight matrix \(W \in \mathbb{R}^{p \times k}\). In the case of interest where \(W_{il} \sim \mathcal{N}(0, 1)\), using \(\mathbb{E}[W_{il}] = 1\) we can further simplify that leads to the following algorithm Alg. 1.

**Algorithm 1 AMP Algorithm for the Spiked Wishart Model With Single-Layer Generative Prior**

**Input**: vector \(Y \in \mathbb{R}^{n \times p}\) and matrix \(W \in \mathbb{R}^{p \times k}\).

**Initialize to zero**:

\(\mathbf{g}, \mathbf{u}, \mathbf{v}, \mathbf{B}_u, \mathbf{A}_u, (\mathbf{A}_v, \mathbf{A}_u)_{t=0}\)

**Initialize with**:

\(\mathbf{v}^{t=1} = \mathcal{N}(0, \sigma^2), \mathbf{v}^{t=1} = \mathcal{N}(0, \sigma^2),\)

\(\mathbf{z}^{t=1} = \mathcal{N}(0, \sigma^2), \mathbf{v}^{t=1} = \mathbf{1}_n, \mathbf{c}^{t=1} = \mathbf{I}_p, \mathbf{c}^{t=1} = \mathbf{1}_k\).

**repeat**

**Spiked layer**:

\(\mathbf{B}_u^t = \frac{1}{k} \mathbf{Y}^\top \mathbf{v}^t - \frac{1}{p} \mathbf{1}_p \mathbf{1}_p^\top \mathbf{v}^t - (\mathbf{A}_u^t)^{-1}\)

\(\mathbf{A}_v^t = \frac{1}{k} \mathbf{1}_p^\top \mathbf{1}_p - \frac{1}{p} \mathbf{1}_p \mathbf{1}_p^\top \mathbf{v}^t - (\mathbf{A}_v^t)^{-1}\)

**Generative layer**:

\(\mathbf{V}^t = \frac{1}{k} (\mathbf{I}_k \mathbf{c}^{t}_v) \mathbf{1}_p + \mathbf{A}_v^t\)

\(\mathbf{A}^t = \frac{1}{k} \mathbf{1}_p + \mathbf{A}_v^t\)

**Update of the estimated marginals**:

\(\mathbf{u}^{t+1} = f_u(\mathbf{B}_u^t, \mathbf{A}_u^t)\)

\(\mathbf{c}^{t+1} = \partial_{\gamma} f_u(\mathbf{B}_u^t, \mathbf{A}_u^t)\)

\(\mathbf{g}^t = f_{out}(\mathbf{B}_u^t, \mathbf{A}_u^t, \mathbf{A}_v^t, \mathbf{V}^t)\)

**Output**:

\(\mathbf{u}, \mathbf{v}, \mathbf{z}\)

**Algorithm 1**

**B. State Evolution Equations**

Perhaps the most important virtue of AMP-type algorithms is that their asymptotic performance can be tracked exactly via a set of scalar equations called state evolution. The order parameters involved are the average overlap between the estimated signals and the ground truth, and are closely related to the mean square error obtained by the algorithm. This fact has been proven for a range of models including the spiked matrix...
models with separable priors in [33], and with non-separable priors in [32]. Adapting the steps of these works, we now derive the state evolution equations for our structured model. As before, we focus on the derivation for the general Wishart model $uv^T$, from which the Bayes-optimal state evolution equations for the symmetric $vv^T$ can be readily obtained.

Note that the standard derivation starts from the relaxed Belief Propagation (r-BP) equations, which are roughly equivalent to AMP updates up to the Onsager terms containing messages with delayed time indices $(\gamma)^{t-1}$. We briefly recall them below where we introduced the parameters $S_{j+j} = \frac{Y_{j}}{\Delta}$ and $R_{j,j} = \frac{1}{\Delta} + S_{j,j}, \forall j \in [1:n], \mu \in [1:p]$:

1) Relaxed-Belief Propagation Equations:

**Variable u**

$$
\hat{u}_{j+j}^{t+1} = f_u(B_{j+j}^{u,t}, A_{j+j}^{u,t}) \\
\epsilon_{j+j}^{t+1} = \partial f_u(B_{j+j}^{u,t}, A_{j+j}^{u,t}) \\
B_{j+j}^{u,t} = \frac{1}{\sqrt{p_v v}} \sum_{\nu \neq \mu} S_{j}^{\nu,v} \hat{u}_{j}^{t+v} \\
A_{j+j}^{u,t} = \frac{1}{p_v} \sum_{\nu \neq \mu} S_{j}^{\nu,v}(\hat{u}_{j}^{t+v} - R_{j,v}(\hat{u}_{j}^{t} + (\hat{u}_{j}^{t} - j)))^2,
$$

**Variable v**

$$
\hat{v}_{j+j}^{t+1} = f_v(B_{j+j}^{v,t}, A_{j+j}^{v,t}, \omega_{j+j}^{t}, V_{t}) \\
\epsilon_{j+j}^{t+1} = \partial f_v(B_{j+j}^{v,t}, A_{j+j}^{v,t}, \omega_{j+j}^{t}, V_{t}), \\
B_{j+j}^{v,t} = \frac{1}{\sqrt{p_v v}} \sum_{\nu \neq \mu} S_{j}^{\nu,v} \hat{u}_{j}^{t+v} \\
A_{j+j}^{v,t} = \frac{1}{p_v} \sum_{\nu \neq \mu} S_{j}^{\nu,v}(\hat{u}_{j}^{t+v} - R_{j,v}(\hat{u}_{j}^{t} + (\hat{u}_{j}^{t} - j)))^2, \\
\omega_{j+j}^{t} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} W_{j}^{i} \hat{z}_{i+j}^{t}, \quad V_{t} = \frac{1}{k} \sum_{i=1}^{k} W_{j}^{i} \hat{z}_{i+j}^{t}. 
$$

**Variable z**

$$
\hat{z}_{j+j}^{t+1} = f_z(\gamma_{j+j}^{t}, A_{j+j}^{t}) \\
\epsilon_{j+j}^{t} = \partial f_z(\gamma_{j+j}^{t}, A_{j}^{t}) \\
\gamma_{j+j}^{t} = \left( \sum_{\nu \neq \mu} B_{j+j}^{v,t} \right), \quad A_{j+j}^{t} = \left( \sum_{\nu \neq \mu} A_{j+j}^{v,t} \right), \\
B_{j+j}^{v,t} = \frac{1}{\sqrt{k}} \sum_{\nu} S_{j}^{\nu,v} \hat{u}_{j}^{t+v} \\
A_{j+j}^{v,t} = \frac{1}{p} \sum_{\nu \neq \mu} S_{j}^{\nu,v}(\hat{u}_{j}^{t+v} - R_{j,v}(\hat{u}_{j}^{t} + (\hat{u}_{j}^{t} - j)))^2, \\
\omega_{j+j}^{t} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} W_{j}^{i} \hat{z}_{i+j}^{t}, \quad V_{t} = \frac{1}{k} \sum_{i=1}^{k} W_{j}^{i} \hat{z}_{i+j}^{t}.
$$

We take this as our starting point, and refer the curious reader to [20] for more details. The first step is then to define the overlap parameters that measure the reconstruction of our inference problem:

$$
q_{u}^{t} \equiv \mathbb{E}_{u}, \quad \lim_{n \to \infty} \frac{(u)^{t}}{n} = \mathbb{E}_{u}, \quad \lim_{n \to \infty} \frac{u^{t}}{n} \equiv m_{u}, \\
q_{v}^{t} \equiv \mathbb{E}_{v}, \quad \lim_{p \to \infty} \frac{(v)^{t}}{p} = \mathbb{E}_{v}, \quad \lim_{p \to \infty} \frac{v^{t}}{p} \equiv m_{v}, \\
q_{z}^{t} \equiv \mathbb{E}_{z}, \quad \lim_{k \to \infty} \frac{(z)^{t}}{k} = \mathbb{E}_{z}, \quad \lim_{k \to \infty} \frac{z^{t}}{k} \equiv m_{z},
$$

where we used the Nishimori property (see [20]) to obtain the equality between order parameters $q^{*}_x = m^{*}_x$ for $x \in \{v, u, z\}$.

2) Average Distributions: Next, to see how these order parameters come into play, we compute the average distribution of the r-BP messages, taking the average over the random variables $W, \xi$, the planted solutions $v^*, u^*, z^*$ and taking the limit $p \to \infty$. Note that using the BP independence assumption over the messages and keeping only dominant terms in the $1/p$ expansion, the dependency in the target node disappears and yields:

- **Average over $B_{u}, A_{u}$:** Starting with the r-BP equations for the $uv^T$ model in (25) yields

$$
\mathbb{E} [B_{u}^{t}] = \frac{1}{\sqrt{\Delta}} \mathbb{E} \left[ \left( \frac{u^{t} (v)^{t}}{\sqrt{p}} + \sqrt{\Delta} \xi \right)^{t} \right] \xrightarrow{p \to \infty} \frac{q_{u}^{t}}{\Delta} u^{*}, \\
\mathbb{E} [B_{u}^{t}] = \frac{1}{\sqrt{\Delta}} \mathbb{E} \left[ \left( \frac{v^{t} (v^*)^{t}}{\sqrt{p}} + \sqrt{\Delta} \xi \right)^{t} \right] + o(1/p) \xrightarrow{p \to \infty} \frac{q_{v}^{t}}{\Delta} I_{p},
$$

$$
\mathbb{E} [A_{u}^{t}] = \mathbb{E} \left[ \frac{1}{p} \left\| v^{t} \right\|^{2} I_{n} \right] \xrightarrow{p \to \infty} \frac{q_{v}^{t}}{\Delta} I_{p}.
$$

- **Average over $B_{v}, A_{v}$:** Similarly,

$$
\mathbb{E} [B_{v}^{t}] \xrightarrow{p \to \infty} \beta q_{v}^{t} \Delta v^{*}, \quad \mathbb{E} [A_{v}^{t}] \xrightarrow{p \to \infty} \beta q_{v}^{t} \Delta I_{p}, \\
\mathbb{E} [B_{v}^{t}] \xrightarrow{p \to \infty} \beta q_{v}^{t} \Delta I_{p}.
$$

- **Average over $\omega, V$:**

$$
\mathbb{E} [w^{t}] = \mathbb{E} [V] \xrightarrow{p \to \infty} \left( \rho_{v} - \frac{q_{v}^{*}}{2} I_{p} \right), \quad \mathbb{E} [w^{t}] = \mathbb{E} \left[ \frac{1}{k} W z^{t} (z^{t})^{T} W^{T} \right] \xrightarrow{n \to \infty} \frac{q_{v}^{*}}{2} I_{p}.
$$

Wrapping the above equations together, we obtained the distributions of means and variances $B_{u}, A_{u}, B_{v}, A_{v}$ and $\omega, V$:

$$
B_{u} \sim \frac{q_{u}^{*}}{\Delta} u^{*} + \sqrt{\frac{q_{u}^{*}}{\Delta}} \xi_{u}, \quad A_{u} \sim \frac{q_{u}^{*}}{\Delta} I_{n}, \\
B_{v} \sim \beta \frac{q_{v}^{*}}{\Delta} v^{*} + \sqrt{\beta \frac{q_{v}^{*}}{\Delta}} \xi_{v}, \quad A_{v} \sim \beta \frac{q_{v}^{*}}{\Delta} I_{p}, \\
\omega \sim \sqrt{\frac{q_{v}^{*}}{2}} \eta, \quad V \sim \left( \rho_{v} - \frac{q_{v}^{*}}{2} I_{p} \right),
$$

with $\xi_{u} \sim \mathcal{N} (0_{n}, I_{n}), \xi_{v} \sim \mathcal{N} (0_{p}, I_{p}), \eta \sim \mathcal{N} (0_{p}, I_{p})$.

3) State Evolution Equations in the Wishart Model: With the averaged limiting distributions of all the messages, we can now compute the state evolution of the overlaps. Using the definition of the overlaps eq. (26) and distributions in eq. (30), we obtain:

$$
(25)$$
Variable $u$:

$$q_u^{t+1} = E_{u^*} \lim_{n \to \infty} \frac{1}{n} f_u(B_u^t, A_u^t)^T f_u(B_u^t, A_u^t)$$

$$= E_{u^*} \left[ f_u \left( \frac{q_u^t}{\Delta} u^* + \sqrt{\frac{q_u^t}{\Delta}} \xi, \frac{q_u^t}{\Delta} \right)^2 \right]$$

(31)

Variable $v$:

$$q_v^{t+1} = E_{v^*} \lim_{n \to \infty} \frac{1}{n} f_v(B_v^t, A_v^t, \omega^t, V^t)$$

$$= E_{v^*} \left[ f_v \left( \beta q_u^t v^* + \sqrt{\beta q_u^t} \xi, \beta q_u^t, \sqrt{q_v^t} \eta, \rho_z - q_z^t \right)^2 \right]$$

(32)

Variable $z$: Even if the hat overlap doesn’t have as much physical meaning as the standard overlaps that quantify the reconstruction performances, we define it as

$$q_z^t \equiv \alpha E_{v^*} \left[ f_{\text{out}} (B, A, \omega, V)^2 \right],$$

(33)

where $f_{\text{out}}$ is evaluated at:

$$(B, A, \omega, V)$$

$$= \left( \beta q_u^t v^* + \sqrt{\beta q_u^t} \xi, \beta q_u^t, \sqrt{q_v^t} \eta, \rho_z - q_z^t \right)$$

(34)

Variable $z$: Averages are explicitly expressed as a function of the hat overlaps introduced just above:

$$E \left[ \gamma^t \right] \xrightarrow{k \to \infty} q_z^t z^*,$$

$$E \left[ \gamma^t (\gamma^t)^T \right] \xrightarrow{k \to \infty} q_z^t I_k$$

$$E \left[ A^t \right] \xrightarrow{k \to \infty} q_z^t I_k.$$

(35)

And we conclude that at the leading order:

$$\gamma^t \sim q_z^t z^* + \sqrt{q_z^t} \xi,$$

$$A^t \sim q_z^t I_k$$

with $\xi \sim \mathcal{N}(0_k, I_k)$

(36)

From these later equations, we finally obtain

$$q_z^{t+1} = E_{z^*} \lim_{k \to \infty} \frac{1}{k} f_z(\gamma^t, A^t)^T f_z(\gamma^t, A^t)$$

$$= E_{z^*} \left[ f_z \left( \hat{q}_z^t z^* + \sqrt{\hat{q}_z^t} \xi, \rho_z - q_z^t \right)^2 \right]$$

(37)

Finally, equations (31-33, 36-36) constitute the closed set of AMP state evolution equations in the Bayes-optimal case for the Wishart model.

4) State Evolution Equations in the Wigner Model: Finally, similarly to the derivation of the AMP algorithm, the state evolution equations for the Wigner model ($v^T \gamma$) are obtained as a particular case of the above by simply restricting $q_u^t = q_v^t$ and $\beta = 1$. In the end, performing a change of variable, this leaves us with only three coupled equations:

$$q_z^{t+1} = E_{\xi} \left[ Z_{\text{out}} \times f_{\text{out}} \left( \sqrt{q_z^t} \xi, q_z^t \right) \right] = 2 \partial q_z \Psi \left( q_z^t \right),$$

$$q_v^t = \alpha E_{\xi, \eta} \left[ Z_{\text{out}} \times f_{\text{out}} \left( \sqrt{q_v^t} \xi, q_v^t, \sqrt{q_z^t} \eta, \rho_z - q_z^t \right) \right]$$

(38)

$$q_v^{t+1} = E_{\xi, \eta} \left[ Z_{\text{out}} \times f_{\text{out}} \left( \sqrt{q_v^t} \xi, q_v^t, \sqrt{q_z^t} \eta, \rho_z - q_z^t \right) \right]$$

with initialization $q_z^{t=0} = \varepsilon, q_v^{t=0} = \varepsilon$ and a small $\varepsilon > 0$. We notice immediately that (37) are the same equations as the fixed point equations related to the Bayes-optimal estimation (14) with specific time-indices and initialization, but crucially the same fixed points. Thus the analysis of fixed points in Section III-B applies straightforwardly here. In particular, since in all cases analyzed we found the stable fixed point of (14) to be unique, we conclude that our AMP algorithm reaches asymptotically optimal performance in these cases.

We can further check this result by numerically comparing the runs of AMP on finite size instances with the state evolution curves already presented in Fig. 2, also giving an idea of the amplitude of the finite size effects. This experiment is illustrated in Fig. 5, together with a curve for PCA and for LAMP, a spectral method we derive from AMP in the next section. A code for reproducing this experiment is provided in GitHub repository [34].

V. LAMP: A SPECTRAL ALGORITHM FOR GENERATIVE PRIORS

Spectral algorithms are the most popular and simplest methods for solving the spiked matrix estimation problem. For instance, canonical PCA estimates the spike from the leading eigenvector of the matrix $Y$. A classical result from Baik, Ben Arous and Péché (BBP) [11] shows that this eigenvector
is correlated with the signal if and only if the signal-to-noise ratio $\rho_v^2/\Delta > 1$. For sparse separable priors (with $\rho_v^2 = \Theta(1)$), $\Delta_{\text{PCA}} = \rho_v^2$ is also the threshold for AMP and it is conjectured that no polynomial algorithm can improve upon it [20]. In the previous section we have shown that our structured AMP algorithm has a consistently better performance than PCA, and in particular achieve the optimal threshold for better-than-random recovery. This is not a surprise, since different from AMP, vanilla PCA doesn’t take into account the information available from the prior.

Despite all its virtues, AMP is unarguably a convoluted algorithm. It would be desirable to have a simpler spectral algorithm taking into account the structured nature of the prior. In this section we design a spectral algorithm, hereafter named LAMP, matching the AMP recovery threshold. Our derivation follows the strategy pioneered in [35], consisting on the linearization of the AMP equations around the non-informative fixed point. In this section, the discussion is framed on the Wigner model, the Wishart case being a straightforward generalisation.

In order for the $q_v = 0$ expansion to be well-defined, we first need to insure that this is indeed a fixed point. Indeed, this condition was already discussed in Section III for the fixed point equations for the Bayes optimal estimator. Not surprisingly, the same conditions can be obtained independently from the AMP equations by analysing when $\hat{\psi} = 0$ is a fixed point, and are repeated below for convenience.

\[(\hat{\psi}, \hat{\gamma}) = (0, 0) \quad \text{if} \quad C = \{E_{Q_{\text{out}}[v]} = 0 \quad \text{and} \quad E_{P_v}[z] = 0\} \tag{38}\]

That these conditions agree exactly to the ones in eq. (15) is just a rephrasing of the fact that the AMP state evolution eqs. (37) have the same fixed points as the Bayes-optimal estimator.

A. Linearising the AMP Equations

To lighten notation, we denote with $|\cdot|$ quantities that are evaluated at $(B_v, A_v, \omega, V, \gamma, \Lambda) = (0, 0, 0, 1, 0, 0)$, and we linearize the AMP equations in Alg. 1 around the uninformative fixed point

\[
(\hat{\psi}, \hat{c}_v) = (0, \rho_v 1_p), \quad (\hat{\gamma}, \hat{\Lambda}) = (0, 0),
\]

\[
(B_v, A_v) = (0, 0), \quad (\omega, V, g) = (0, 0, 0).
\tag{39}\]

In a scalar formulation, this yields

\[
\delta B_{i,t}^{v,t} = \frac{1}{\Delta} \sum_{j=1}^{p} \frac{Y_{i,j}}{\sqrt{p}} \Delta \hat{y}_j^{v,t} - \frac{1}{\Delta} \left( \sum_{j=1}^{p} \frac{\hat{c}_j^{v,t}}{p} \right) \delta \hat{v}_j^{t-1}
\]

\[
\delta A_{i,t}^{v,t} = \frac{2}{\Delta} \sum_{j=1}^{p} \hat{v}_j^{t} |\delta \hat{v}_j^{t|=s|}
\]

\[
\delta \omega_{i}^{t} = \frac{1}{k} \frac{1}{\sqrt{p}} \sum_{i=1}^{k} W_{ii} \delta \hat{z}_i^{t} - \delta V_i |\delta \hat{v}_i^{t-1}|_{s} - V_i |\hat{g}_i^{t-1}|,
\]

\[
\delta V^{t} = \frac{1}{k} \sum_{i=1}^{k} \delta \hat{c}_i^{v,t},
\]

\[
\delta \gamma_{i}^{t} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} W_{ii} \delta \hat{g}_i^{t} + \delta \Lambda_i |\hat{z}_i^{t}|_{s} + \Lambda_i |\delta \hat{z}_i^{t}|,
\]

\[
\delta \Lambda_{t}^{v} = \frac{2}{k} \sum_{i=1}^{p} \hat{g}_i^{t} |\delta \hat{g}_i^{t}| = 0.
\]

These equations can be simplified and closed over three vectors $\hat{v} \in \mathbb{R}^p$, $\hat{z} \in \mathbb{R}^k$ and $\omega \in \mathbb{R}^p$, where we used the existence condition $C$ that leads to $\delta \omega_{\text{out}} |s| = \delta \omega_{\text{out}} |s| = 0$.

Finally, inserting eq. (44)-(46) in (40), (42), (40), rewriting the partial derivatives of $f_v$, $f_z$ and $f_{\omega}$ at the fixed point $|s|$, as moments of the distributions $P_z$ and $Q_{\text{out}}$ and simplifying the expression using the condition $C$, we finally obtain

\[
\delta \hat{v}_{i+1} = \frac{1}{\sqrt{p}} \rho_v \left( \frac{Y}{\sqrt{p}} \delta \hat{v}_{i} - \rho_v 1_p \delta \hat{v}_{i-1} \right) + \rho_v^{-1} E_{Q_{\text{out}}[v]} 1_p \delta \omega_{i} + \frac{E_{Q_{\text{out}}[v]z|z}}{2} \rho_v \hat{v}_{i} \hat{f}_{i} \hat{f}_{i} \frac{1}{k} \delta \hat{z}_{i},
\]

\[
\delta \gamma_{i+1} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} W_{ii} \delta \hat{g}_{i} + \delta \Lambda_i \hat{z}_i + \Lambda_i |\delta \hat{z}_i|,
\]

\[
\delta \Lambda_{i+1} = \frac{2}{k} \sum_{i=1}^{p} \hat{g}_i |\delta \hat{g}_i| = 0.
\]

Inserting eq. (48)-(49) in (47) and dropping heuristically the time indices, we finally obtain a closed linear equation over $\hat{v} = \gamma^{v+v} \hat{v}$, where the LAMP operator $\gamma^{v+v}$ is given by

\[
\gamma^{v+v} = \frac{1}{\Delta} \left( (a - b) 1_p + \frac{b W W^T}{k} + \frac{c 1_p 1_p^T W W^T}{k} \right) \times
\]

\[
\times \left( \frac{y}{\sqrt{p}} - a 1_p \right),
\]

where the parameters are simply the moments of distributions $P_z$ and $Q_{\text{out}}$

\[
a \equiv E_{Q_{\text{out}}[v]}[v^2] = \rho_v, \quad b \equiv \rho_v^{-1} E_{Q_{\text{out}}[v]}[v] = \rho_v^{-1} E_{Q_{\text{out}}[v]}[v^2],
\]

\[
c \equiv \frac{1}{2} \rho_v^{-3} E_{Q_{\text{out}}[v]}[v^3].
\]

Note that in most of the cases we studied, the parameter $c$, taking into account the skewness of the variable $z$, is zero, simplifying considerably the structured matrix. Moreover, for the
specific examples already discussed in Section III, the LAMP operator $\gamma_p^{vv}$ is very simple. For instance, for Gaussian $z$ and $P_{\text{out}}(v|x) = \delta(v - \text{sign}(x))$, we have $(a, b, c) = (1, 2/\pi, 0)$. Instead, for linear activation we get $(a, b, c) = (1, 1, 0)$. Note that in this last case, the LAMP operator can be written as

$$
\gamma_p^{vv} = \frac{1}{\Delta} K_p \left[ \frac{Y}{\sqrt{p}} - I_p \right]
$$

with

$$
K_p = \frac{[WVV^T]}{k} = \Sigma \approx \frac{1}{n} \sum_\alpha v^\alpha (v^\alpha)^T,
$$

or, in other words, $K_p$ is the covariance matrix of the structured spike $v$. The same observation holds for the sign activation function. Interestingly, the covariance matrix $\Sigma$ can be empirically estimated directly from samples of spikes, without the knowledge of the generative model $(\varphi, W)$ itself, suggesting a simple practical implementation of LAMP. Therefore we finally use a more generic definition for LAMP (linearized-AMP) as the following spectral algorithm:

### Algorithm 2 LAMP Spectral Algorithm

**Input:** Observed matrix $Y \in \mathbb{R}^{p \times p}$, prior $P_v$ on $v \in \mathbb{R}^p$

Take the leading eigenvector $\hat{v} \in \mathbb{R}^p$ of

$$
\Gamma_p^{vv} = K_p \left[ \frac{\hat{v}}{\sqrt{p}} - I_p \right]
$$

with $K_p = \mathbb{E} P_v [vv^T]$. From this perspective, LAMP Alg. 2 can be interpreted as a PCA that takes into account the structure of the prior by incorporating the non-trivial correlations through $K_p$ into the spectral estimation. In particular taking $P_v(\hat{v}) = N_v(0, I_p)$, we obtain $\Gamma_p^{vv} = \frac{1}{\Delta} \left[ \frac{\hat{v}}{\sqrt{p}} - I_p \right]$ and recognize the PCA operator that has been shifted.

Analogously to the state evolution for AMP, the asymptotic performance of both PCA and LAMP can be evaluated in a closed-form for the spiked Wigner model with single-layer generative prior with linear activation. The corresponding expressions are derived in the next section and plotted in Fig. 5 for the three considered algorithms.

### B. State Evolution for LAMP and PCA in the Linear Case

As we have already mentioned in Section IV-B, one of the greatest virtues of AMP is being able to track its asymptotic performance through a set of simple scalar state evolution equations. Interestingly, we can also derive state evolution equations for the LAMP algorithm in the linear case. This allows a direct comparison between the performance of LAMP and the performance of PCA.

For the noiseless linear channel $P_{\text{out}}(v|x) = \delta(v - x)$, the set of eqs. (47-49) are already linear. Hence the LAMP spectral method flows directly from the AMP Alg. (1). As a consequence, this means that the state evolution equations associated to the spectral method are simply dictated by the set of AMP state evolution equations eq. (37).

However, it is worth stressing that as the LAMP returns a normalized estimator, the LAMP MSE is not given by the AMP mean squared error. We now compute the overlaps and mean squared error performed by this spectral algorithm.

Recall that $m_v$ and $q_v$ are the parameters defined in eq. (26), respectively measuring the overlaps between the ground truth $v^*$ and the estimator $\hat{v}$, and the norm of the estimator. In the general case, the MMSE eq. (8) becomes:

$$
\text{MSE}_v = \rho_v + \mathbb{E}_\varphi \lim_{p \to \infty} \frac{1}{p} ||\hat{v}||_2^2 - 2\mathbb{E}_\varphi \lim_{p \to \infty} \frac{1}{p} \hat{v}^\top \hat{v}^*
$$

$$
= \rho_v + q_v - 2m_v,
$$

(54)

However the LAMP spectral method computes the normalized leading eigenvector of the structured matrix $\Gamma_p^{vv}$. Hence the norm of the LAMP estimator is $||\hat{v}||_2^2_{\text{AMP}} = q_v^{\text{LAMP}}$, while the Bayes-optimal AMP estimator is not normalized with $||\hat{v}||_2^2_{\text{AMP}} = q_v^{\text{AMP}} = m_v^{\text{AMP}} \neq 1$, solutions of eq. (37). As the non-normalized LAMP estimator follows AMP state evolutions in the linear case, the overlap with the ground truth is thus given by:

$$
m_v^{\text{LAMP}} = \mathbb{E}_\varphi \lim_{p \to \infty} \frac{1}{p} \hat{v}^\top_{\text{LAMP}} v^* = \mathbb{E}_\varphi \lim_{p \to \infty} \frac{1}{p} \left( \frac{\hat{v}_{\text{LAMP}}}{\|\hat{v}\|_{\text{AMP}}} \right)^\top v^*
$$

$$
= m_v^{\text{AMP}} \left( q_v^{\text{AMP}} \right)^{1/2} = \left( m_v^{\text{AMP}} \right)^{1/2}.
$$

Finally the mean squared error performed by the LAMP method is easily obtained from the optimal overlap reached by the AMP algorithm and yields

$$
\text{MSE}_v^{\text{LAMP}} = \rho_v + 1 - 2 \left( q_v^{\text{AMP}} \right)^{1/2}.
$$

(56)

The respective result for PCA can be obtained from the observation that for the linear case, the $\alpha = 0$ LAMP operator reduces exactly to the matrix $Y$. In other words, in this case LAMP reduces to PCA. In terms of the prior, this is clear since $\alpha = 0$ is equivalent to a separable Gaussian prior, for which the spectral algorithm derived from AMP is exactly given by PCA [20]. Therefore we can simply state that the mean squared error performed by PCA is computed using the optimal overlap reached by AMP at $\alpha = 0$:

$$
\text{MSE}_v^{\text{PCA}} = \rho_v + 1 - 2 \left( q_v^{\text{AMP}} \right)^{1/2}.
$$

(57)

In order to fairly compare PCA, LAMP and AMP in Fig. 5, instead of showing the MSE corresponding to the normalized PCA and LAMP estimators, we rescale these spectral estimators by the optimal normalisation $(\hat{q}_v^{\text{AMP}})^{1/2}$ (obtained from AMP for instance) so that the renormalized MSEs are given by

$$
\text{MSE}_v^{\text{LAMP}} = \rho_v - m_v^{\text{LAMP}}
$$

$$
\text{MSE}_v^{\text{PCA}} = \rho_v - m_v^{\text{PCA}}.
$$

(58)

Therefore in the linear case we simply obtain that LAMP is strictly equivalent to AMP, while PCA is sub-optimal:

$$
\text{MSE}_v^{\text{LAMP}} = \rho_v - q_v^{\text{LAMP}}
$$

$$
\text{MSE}_v^{\text{PCA}} = \rho_v - q_v^{\text{AMP}}|_{\alpha=0}.
$$

(59)

Figure 5 shows good agreement between the state evolution for LAMP and PCA with linear activation (solid lines) and the respective finite instance numerical simulations (points).
C. A Random Matrix Perspective on the Recovery Threshold

Remarkably, the performance of the spectral method based on matrix (53) can be investigated independently of AMP using random matrix theory. An analysis of the random matrix (53) shows that a spectral phase transition for generative prior with linear activation appears at \( \Delta_c = 1 + \alpha \) (as for AMP). This transition is analogous to the well-known BBP transition [11], but for a non-GOE random matrix (53). For the spiked Wigner models with linear generative prior we prove two detailed theorems describing the behavior of the supremum of the bulk spectral density, the transition of the largest eigenvalue and the correlation of the corresponding eigenvector. The theorems counterparts for the linear Wishart model are very similar, and are presented in appendix. We assume in the following that \( \rho_v = 1 \) to simplify the analysis (without any loss of generality). Recall that we have

\[
\Gamma^v_p \equiv \frac{1}{k} WW^\top \left( \frac{1}{\sqrt{\Delta p}} \xi + \frac{1}{\Delta} vv^\top - \frac{1}{\Delta} I_p \right). \tag{60}
\]

Here \( \xi / \sqrt{p} \) is a matrix from the Gaussian Orthogonal Ensemble, i.e. \( \xi \) is a real symmetric matrix with entries drawn independently from a Gaussian distribution with zero mean and variance \( \mathbb{E} \xi_{ij}^2 = (1 + \delta_{ij}) \).

Theorem VI.1 (Bulk of the Spectral Density, Spiked Wigner, Linear Activation): For any \( \alpha, \Delta > 0 \), as \( p \to +\infty \), the spectral measure of \( \Gamma^v_p \) converges almost surely and in the weak sense to a well-defined and compactly supported probability measure \( \mu(\alpha, \Delta) \), and we denote \( \supp \mu \) its support. We separate two cases:

(i) If \( \Delta \leq \frac{1}{4} \), then \( \supp \mu \subseteq \mathbb{R}_- \).

(ii) Assume now \( \Delta > \frac{1}{4} \) and denote \( z_1(\Delta) \equiv -\Delta^{-1} + 2\Delta^{-1/2} > 0 \). Let \( \rho_\Delta \) be the probability measure on \( \mathbb{R} \) with density

\[
\rho_\Delta(dt) = \frac{\sqrt{2}}{\pi} \frac{\Delta}{2} \sqrt{4 - \Delta \left( t + \frac{1}{\Delta} \right)^2} \times 1 \left\{ \left| t + \frac{1}{\Delta} \right| \leq \frac{2}{\sqrt{\Delta}} \right\} dt. \tag{61}
\]

Note that the supremum of the support of \( \rho_\Delta \) is \( z_1(\Delta) \).

The following equation admits a unique solution for \( s \in (-z_1(\Delta)^{-1}, 0) \):

\[
\alpha \int \rho_\Delta(dt) \left( \frac{st}{1 + st} \right)^2 = 1. \tag{62}
\]

We denote this solution as \( s_{\text{edge}}(\alpha, \Delta) \) (or simply \( s_{\text{edge}} \)). The supremum of the support of \( \mu(\alpha, \Delta) \) is denoted \( \lambda_{\text{max}}(\alpha, \Delta) \) (or simply \( \lambda_{\text{max}} \)). It is given by:

\[
\lambda_{\text{max}} = \begin{cases} 
\alpha \int \frac{\rho_\Delta(dt)t}{1 + s_{\text{edge}} t} - \frac{1}{s_{\text{edge}}} & \text{if } \alpha \leq 1, \\
\max \left( 0, \alpha \int \frac{\rho_\Delta(dt)t}{1 + s_{\text{edge}} t} - \frac{1}{s_{\text{edge}}} \right) & \text{if } \alpha > 1.
\end{cases} \tag{63}
\]

As a function of \( \Delta \), \( \lambda_{\text{max}} \) has a unique global maximum, reached exactly at the point \( \Delta = \Delta_c(\alpha) = 1 + \alpha \). Moreover, \( \lambda_{\text{max}}(\alpha, \Delta_c(\alpha)) = 1 \).

Theorem V.2 (Transition of the Largest Eigenvalue and Eigenvector, Spiked Wigner, Linear Activation): Let \( \alpha > 0 \). We denote \( \lambda_1 \geq \lambda_2 \) the first and second eigenvalues of \( \Gamma^v_p \). If \( \Delta \geq \Delta_c(\alpha) \), then as \( p \to \infty \) we have a.s. \( \lambda_1 \to \lambda_{\text{max}} \) and \( \lambda_2 \to \lambda_{\text{max}} \). If \( \Delta \leq \Delta_c(\alpha) \), then as \( p \to \infty \) we have a.s. \( \lambda_1 \to 1 \) and \( \lambda_2 \to \lambda_{\text{max}} \). Further, denoting \( \tilde{v} \) a normalized \( (\|\tilde{v}\|^2 = p) \) eigenvector of \( \Gamma^v_p \) with eigenvalue \( \lambda_1 \), then \( \tilde{v}^\top \tilde{v}^\top/\sqrt{\Delta} \to \epsilon(\Delta) \) a.s., where \( \epsilon(\Delta) = 0 \) for all \( \Delta \geq \Delta_c(\alpha) \), \( \epsilon(\Delta) > 0 \) for all \( \Delta < \Delta_c(\alpha) \) and \( \lim_{\Delta \to 0} \epsilon(\Delta) = 1 \).

Thm. V.1 and Thm. V.2 are illustrated in Fig. 5. The proof gives the value of \( \epsilon(\Delta) \), which turns out to lead to the same MSE as in Fig. 5 in the linear case. The proofs of theorems V.1 and V.2 are given in the following (some proofs and arguments are postponed to appendix), along with the precise arguments used to derive the eigenvalue density, the transition of \( \lambda_1 \) and the computation of \( \epsilon(\Delta) \). These arguments are solely based on random matrix theory. The method of proof of Theorem V.2 is very much inspired by [36]², and allows us to compute numerically the squared correlation \( \epsilon(\Delta) \). It is given, for all \( \Delta < \Delta_c(\alpha) \), as

\[
\epsilon(\Delta) = \frac{1}{\alpha} \frac{S^{(1,2)}(1)^2}{S^{(1,2)}(1)}. \tag{61}
\]

The \( S^{(1,2)} \) and \( S^{(2)} \) functions are defined in the following Lemma VI.2, and formulas are also given that allow to compute them numerically. A Mathematica demonstration notebook is provided in the GitHub repository [34].

In the non-linear case the random matrix analysis is harder to perform. In the matrix \( \Gamma^v_p \), the Wishart matrix \( WW^\top / k \) is replaced by \( a_1 + bWW^\top \) with \( a, b \geq 0 \). It is thus not possible to relate the spectrum of \( \Gamma^v_p \) to the one of a symmetric matrix of the type \( WZW^\top \) with \( W \) a gaussian i.i.d. matrix. Some techniques from free probability could make the computation nevertheless possible, but we leave this analysis for future work.

D. Applying LAMP to Real Data

As we have already remarked, the LAMP operator in eq. (53) only depend on the generative prior through its covariance. An interesting exercise is to apply AMP for real data by simply using the empirical covariance for \( n \) samples of the spikes, \( v^n, \alpha = 1, \ldots, n \).

For illustration, we perform the following experiment: the spikes \( v^* \) are drawn from the standard Fashion-MNIST dataset, and are used to generate the spiked matrix \( Y \) according to eq. (1). We then apply our LAMP algorithm to reconstruct the spikes, repeating this experiment for different values of noise \( \Delta \). In Fig. 6 we compare the reconstruction by LAMP with standard PCA over \( Y \). In principle, we have no theoretical guarantees about the performance of LAMP, since the Fashion-MNIST images are not drawn from the generative class studied above. Nevertheless, it is striking to observe that LAMP outperforms PCA.

²Note that while all the calculations are justified, refinements would be needed in order to be completely rigorous. These refinements would follow exactly some proofs of [37] and [36], so we will refer to them when necessary.
A demonstration notebook illustrating this experiment is provided in the GitHub repository [34].

VI. PROOFS

In this section we present the technical details of the proofs for Theorems VI-A, VI.1, and VI.2.

A. Proof of Theorem III.1

We present a proof of the theorem III.1, for the mutual information of Wigner model eq. (1) with structured prior

$$ Y = \sqrt{\frac{\lambda}{p}} \mathbf{v}^* \mathbf{v}^T + \xi, $$

(64)

where the spike $\mathbf{v}^* \in \mathbb{R}^p$ is drawn from $P_0$. The proof is based on Guerra Interpolation [39], [40]. Some technical parts of the proof will be postponed to appendix, Sec. A.

1) Notations, Free Energies, and Gibbs Average: The mutual information being invariant to reparametrization, we shall work instead inside this section with the following notations:

$$ I(Y; \mathbf{v}) = H(Y) - H(Y|\mathbf{v}). $$

(65)

where $\lambda$ is the signal to noise ratio. Up to the reparametrization, it corresponds to our model with $\Delta = \lambda^{-1}$. Our aim is to compute $I(Y; \mathbf{v})$. While the information theoretic notation is convenient in stating the theorem, it is more convenient to use statistical physics notation and "free energies" for the proof, that relies heavily on concepts from mathematical physics. Let us first translate one into the other. The mutual information between the observation $Y$ and the unknown $\mathbf{v}$ is defined using the entropy as $I(Y; \mathbf{v}) = H(Y) - H(Y|\mathbf{v})$. Using Bayes theorem one obtains $H(Y) = \mathbb{E}_Y\{\log E_p P_Y(Y|\mathbf{v})\}$ and a straightforward computation shows that the mutual information per variable is then expressed as

$$ I(Y; \mathbf{v}) = f_p + \lambda \mathbb{E}[\mathbf{v}^T \mathbf{v}] / 4p, $$

(66)

where, using again statistical physics terms, $f_p = -\mathbb{E}_Y[\log Z_p(Y)] / p$ is the so called free energy density and $Z_p(Y)$ the partition function defined by

$$ Z_p(Y) \equiv \int_{\mathbb{R}^p} dP_p(\mathbf{v}) \exp \left( \sum_{i<j} \left( \frac{\lambda v_i^2 v_j^2}{2p} + \sqrt{\lambda} v_i v_j Y_{ij} \right) \right). $$

(67)

Notice that the sum does not includes the diagonal term in (67). Different conventions can be used depending on whether or not one supposes the diagonal terms to be measured, but this yields only order $1/p$ differences in the free energies, and thus does not affect the limit $p \to \infty$. Correspondingly, we thus define the Hamiltonian:

$$ -H(\mathbf{v}) \equiv \sum_{i<j} \sqrt{\lambda} Y_{ij} v_i v_j - \frac{\lambda}{2p} v_i^2 v_j^2 $$

$$ = \sum_{i<j} \sqrt{\lambda} \xi_{ij} v_i v_j + \frac{\lambda}{p} v_i v_j v_i^* v_j^* - \frac{\lambda}{2p} v_i^2 v_j^2, $$

so that the partition function (67) is associated with the Gibbs-Boltzmann measure $e^{-H} / Z_p(Y)$. Consider now the term $I(\mathbf{v}; \mathbf{v}^* + \mathbf{z})$ that enters the expression to be proven eq. (7). This is the mutual information for another denoising problem, in which we assume one observes a noisy version of the vector $\mathbf{v}^*$, denoted $\tilde{\mathbf{y}}$ such that

$$ \tilde{\mathbf{y}} = \frac{1}{\sigma} \mathbf{v}^* + \mathbf{z}, $$

(68)

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_p, I_p)$ and $\sigma = 1 / \sqrt{\lambda q_v \lambda}$. Again, it is easier to work with free energies. We thus write the corresponding posterior distribution as

$$ P(\mathbf{v}|\tilde{\mathbf{y}}) = \frac{1}{Z_0(\tilde{\mathbf{y}}, \sigma)} P_0(\mathbf{v}) \exp \left( \frac{\|\mathbf{v}\|^2}{2\sigma^2} + \frac{\mathbf{v}^T \tilde{\mathbf{y}}}{\sigma} \right), $$

(69)

where $Z_0(\tilde{\mathbf{y}})$ is the normalization factor. For this denoising problem, the averaged free energy per variables reads

$$ f_p^0(\sigma) \equiv -\frac{1}{p} \mathbb{E}_y[\log Z_0(\tilde{\mathbf{y}}, \sigma)], $$

(70)

and a short computation shows that

$$ I(\mathbf{v}; \mathbf{v}^* + \mathbf{z}) = f_p^0 \left( \frac{1}{\sqrt{\lambda q_v \lambda}} \right) + \frac{\lambda q_v}{2}. $$

Putting all the pieces together, this means that we need to prove the following statement on the free energy $f_p$: the free energy $f_p = -\mathbb{E}_Y[\log Z_p(Y)] / p$ is given, as $p \to \infty$ by

$$ \lim_{p \to \infty} f_p = \min_{\phi_{RS}} \left( \frac{1}{\sqrt{\lambda q_v \lambda}} \right), $$

(71)

This statement is equivalent to theorem III.1, and we shall present a proof for the case where the prior over $\mathbf{v}$ has a "good" limit: we shall assume that the limiting free energy exists and concentrates over the disorder, and that the distribution over each $v_i$ is bounded. These hypothesis will be explicitly given when needed.

Finally, it will be useful to consider Gibbs averages, and to work with $r$ copies of the same system. For any $g : (\mathbb{R}^p)^{r+1} \to \mathbb{R}$, we define the Gibbs average as

$$ \left\langle g(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(r)}, \mathbf{v}^*) \right\rangle \equiv \frac{\int g(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(r)}, \mathbf{v}^*) \prod_{i=1}^r e^{-H(v_i^{(i)})} dP_0(v_i^{(i)})}{\left( \int e^{-H(v_i^{(i)})} dP_0(v_i^{(i)}) \right)^r}. $$

(72)
This is the average of $g$ with respect to the posterior distribution of $r$ copies $v^{(1)}, \cdots, v^{(r)}$ of $v^*$. The variables $\{v^{(i)}\}_{i=1}^{r}$ are called replicas, and are interpreted as random variables independently drawn from the posterior. When $r = 1$ we simply write $g(v, v^*)$ instead of $g(v^{(1)}, v^*)$. Finally we shall denote the overlaps between two replicas as follows: for $l, l' = 1, \cdots, r$, we let

$$R_{l, l'} \equiv v^{(l)} \cdot v^{(l')} = \frac{1}{p} \sum_{i=1}^{p} v^{(l)}_i v^{(l')}_i.$$  

(73)

A simple but useful consequence of Bayes rule is that the $(r + 1)$-tuples $(v^{(1)}, \cdots, v^{(r+1)})$ and $(v^{(1)}, \cdots, v^{(r)}, v^*)$ have the same upper under the expectation $I(x)$ (see [26] or proposition 16 in [17]). This bears the name of the Nishimori property in the spin glass literature [41].

2) Guerra Interpolation for the Upper Bound: We start by using the Guerra interpolation to prove an exact formula for the free energy. Let $t \in [0, 1]$ and let $q_{e}$ be a non-negative variable. We now consider an interpolating Hamiltonian

$$-H_t(v) \equiv \sum_{i < j} \sqrt{\frac{t \lambda}{p}} \xi_{ij} v_i v_j + \sqrt{\frac{t \lambda}{p}} v_i v_i v_j v_j - \sqrt{\frac{t \lambda}{2p}} v_i^2 v_j^2$$

$$+ \sum_{i=1}^{p} \frac{(1 - t)\lambda q_{e} z_i v_i + (1 - t)\lambda q_{e} v_i v_i^*}{2} - \frac{(1 - t)\lambda q_{e}}{v_i^2}.$$ 

The Gibbs states associated with this Hamiltonian $-H_t$ correspond to an estimation problem given an augmented set of observations

$$\{Y_{ij} = \sqrt{\frac{\lambda}{p}} v_i v_j + \xi_{ij}, \ 1 \leq i < j \leq p, \}
$$

$$\{\tilde{y}_i = \sqrt{(1 - t)\lambda q_{e}} v_i v_i^* + z_i, \ 1 \leq i \leq p.\}$$

Reproducing the argument of [26], we prove using Guerra’s interpolation [39] and the Nishimori property the following the proof is done in appendix A-A.

**Proposition VI.1 (Upper Bound on the Free Energy):** Assume the elements of $v$ are bounded. Then there exists a constant $K > 0$ such that for all $q_{e} \in \mathbb{R}$ we have:

$$f_p \geq f_p^0\left(1/\sqrt{\lambda q_{e}}\right) + \frac{\lambda q_{e}^2}{4} + \frac{K}{p}. \quad (74)$$

3) A Bound of the Franz-Parisi Potential: To attack the lower bound, we shall adapt the argument of [25], that uses the Franz-Parisi potential [42], and this will require additional concentration properties on the prior model. For $v^* \in \mathbb{R}^p$ fixed, $m \in \mathbb{R}$ and $\epsilon > 0$ we follow [25] and define

$$\Phi^\epsilon_p(m, v^*) \equiv -\frac{1}{p} \mathbb{E} \log \int \mathbb{I} \{R_{1,1} \in [m, m + \epsilon]\} e^{-H(v)} dP_e(v).$$

(75)

This is simply the free energy with configurations forced to be at a distance $m$ (to precision $\epsilon$) from the ground truth. Note that since the measure is limited to a subset of configurations, it is clear that $\mathbb{E}_e \Phi^\epsilon_p(m, v^*) \geq f_p$.

We are now going to prove an interpolating bound for the Franz-Parisi Potential:

**Proposition VI.2 (Lower Bound on the Franz-Parisi Potential):** Assume the elements of $v$ are bounded. Then there exists $K > 0$ such that for any $m = q_{e}$ and $\epsilon > 0$ we have

$$\Phi^\epsilon_p(m = q_{e}, v^*) \geq f_p^0\left(1/\sqrt{\lambda q_{e}}\right) + \frac{\lambda q_{e}^2}{4} - \frac{\lambda}{2} \epsilon^2 + \frac{K}{p}. \quad (76)$$

The proof is carried out in appendix A-B.

4) From the Potential to a Lower Bound on the Free Energy: It remains to connect the Franz-Parisi potential to the actual free energy. This is done by proving a Laplace-like result between the free energy and the Franz-Parisi free energy, again following the technique used in the separable case in [25]:

**Proposition VI.3:** There exists $K > 0$ such that for all $\epsilon > 0$, we have

$$f_p \geq \mathbb{E}_e\left[\min_{|t| \leq K/\epsilon} \Phi^\epsilon_p(t, v^*)\right] - \frac{\log(K/\epsilon)}{\sqrt{p}}. \quad (77)$$

Combining this proposition with the bound on the Franz-Parisi potential, we see that

$$f_p \geq \mathbb{E}_e\left[\min_{|q_{e}| \leq K/\epsilon} f_p^0\left(1/\sqrt{\lambda q_{e}}\right) + \frac{\lambda q_{e}^2}{4}\right] - \frac{\log(K/\epsilon)}{\sqrt{p}}. \quad (78)$$

At this point, we need to push the expectation with respect to the spike inside the minimum. This is the only assumption that we are going to require over the generative model: that its free energy concentrates over the distribution of spikes. This finally leads to following result:

**Proposition VI.4 (Laplace principle):** Assume that the free energy $f_p^0(v^*)$ concentrates such that

$$\mathbb{E}\left[f_p^0\left(1/\sqrt{\lambda q_{e}}\right) + \frac{\lambda q_{e}^2}{4}\right] + o\left(\frac{\log p}{\sqrt{p}}\right) \quad (79)$$

for some constant $C$ for all $q_{e}$ in $[0, \rho_{e}]$, then:

$$f_p \geq \min_{q_{e}} \left[f_p^0\left(1/\sqrt{\lambda q_{e}}\right) + \frac{\lambda q_{e}^2}{4}\right] + o\left(\frac{\log p}{\sqrt{p}}\right). \quad (80)$$

which gives us the needed converse bound.

We prove these two crucial propositions in appendix A-C.

5) Main Theorem: We can now combine the upper and lower bound to reach the statement of the main theorem, presented in the main part as theorem III.1:

**Theorem VI.1 (Mutual Information for the Spiked Wigner Model With Structured Spike):** Assume the spikes $v^*$ come from a sequence (of growing dimension) of generic structured prior $P_e$ on $\mathbb{R}^p$, such that

1) The elements of $v$ are bounded by a constant.

2) The free energy $f_p^0\left(1/\sqrt{\lambda q_{e}}\right)$ has a limit $f_0\left(\lambda q_{e}\right)$ for all $q_{e}$ in $[0, \rho_{e}]$ as $p \to \infty$.

3) The free energy $f_p^0\left(1/\sqrt{\lambda q_{e}}\right)$ concentrates such that

$$\mathbb{E}_e\left[f_p^0\left(1/\sqrt{\lambda q_{e}}\right), v^*\right] - \mathbb{E}_e\left[f_0^0\left(1/\sqrt{\lambda q_{e}}\right), v^*\right] \leq C/\sqrt{p}$$

for some constant $C$ for all $q_{e}$ in $[0, \rho_{e}]$ as $p \to \infty$.

then

$$\lim_{p \to \infty} i_p \equiv \lim_{p \to \infty} \frac{I(Y; v^*)}{p} = \inf_{\rho_{e} \geq q_{e} \geq 0} i_{\text{RS}}(\Delta, q_{e}). \quad (81)$$


with
\[ i_{RS}(\Delta, q_v) = \frac{(\rho_v - q_v)^2}{4\Delta} + \lim_{p \rightarrow \infty} \frac{I(\mathbf{v}; \mathbf{v} + \sqrt{\Delta}\mathbf{z})}{p} \]  

(82)

where \( z \) being a Gaussian vector with zero mean, unit diagonal variance and \( \rho_v = \lim_{p \rightarrow \infty} E_{\rho_v} \langle \mathbf{v}\mathbf{v}^T \rangle / p \).

6) Mean-Squared Errors: It remains to deduce the optimal mean squared errors from the mutual information. These are applications of known results which we reproduce here briefly.

It is well-known [43] that the mean squared error is minimized by using the conditional expectation of the signal given the observation, that is the posterior mean. The minimal mean square error is thus given by
\[ \text{Matrix} - \text{MMSE}(Y) = \frac{1}{p^2} \| \mathbf{v}^* \mathbf{v}^*^\top - E[\mathbf{v}\mathbf{v}^T] \|_2^2 \]  

(83)

We can now state the final result, that we prove in appendix A-D.

\textbf{Theorem VI.2:} [Matrix MMSE, from [14, 18, 44]] The matrix-MMSE is asymptotically given by
\[ \lim_{p \rightarrow \infty} \text{Matrix} - \text{MMSE}(Y) = \rho_v^2 - (q_v^*)^2 \]  

(85)

where \( q_v^* \) is the optimizer of the function \( i_{RS}(\Delta, q_v) \).

We now consider the problem of reconstruction the spike itself. In this case the mean square error reads
\[ \text{Vector} - \text{mse}(\hat{X}, Y) = \frac{1}{p} \| \mathbf{v} - \mathbf{v}^* \|_2^2 \]  

and
\[ \text{Vector} - \text{MMSE}(Y) = \frac{1}{p} \| \mathbf{v} - E[\mathbf{v}] \|_2^2 \]  

(86)

Taking the square and averaging, we thus find that the asymptotic vector MMSE reads
\[ \text{Vector} - \text{MMSE}(Y) = \rho_v + \frac{\| E[\mathbf{v}] \|_2^2}{p} - 2 \frac{E[\mathbf{v}\mathbf{v}^T]}{p} \]  

and
\[ = \rho_v - \frac{E[\mathbf{v}\mathbf{v}^T]}{p}. \]  

(87)

Note that the absolute value is necessary here, because if the prior is symmetric, it is impossible to distinguish between \( \mathbf{v}^* \) and \(-\mathbf{v}^*\). If the prior is not symmetric, then the absolute value can be removed.

Note that one could shorten the argument above by avoiding the matrix-MSE, looking instead at other \( Z_q \) invariant performance measures which are directly related to the spike \( \mathbf{v}^* \) and also concentrate at functions of the overlap \( q_v \), such as sup \( |\langle \mathbf{v}, \mathbf{v}^* \rangle| \) or sup \( \frac{\| \mathbf{v}\|_2^2}{\| \mathbf{v}\|_2^2 + \| \mathbf{v}^*\|_2^2} \).

\textbf{B. Proof of Theorem V.1}

We begin by showing item (ii) of Theorem V.1. We use the classical notation \( \mathbb{C}_+ \equiv \{ z \in \mathbb{C} | \text{Im}(z) > 0 \} \).

\textbf{Proof of Theorem V.1 (ii):} We begin by treating the more involved case (ii), that is we assume \( \Delta > \frac{1}{4} \). Fortunately, this is easily done by using Theorem 7 in [45], which applies in our case since it only depends on the free energy and the Franz-Parisi bound, that we have reproduced in the coupled cases in the present section. This proposition states the convergence in probability of the overlaps:

\textbf{Theorem VI.3 (Convergence in Probability of the Overlap, From [45]):} Informally, for the Wigner-Spiked model:
\[ \lim_{p \rightarrow \infty} E(\mathbb{I}(|R_{1,s}| > q_v^* \mid \varepsilon)) \rightarrow 0 \]  

(88)

More precisely, if \( p \geq k \) (so \( \alpha \geq 1 \)) we have \( \text{Sp}(\Gamma_{k}^{w}) = \text{Sp}(\Gamma_{k}^{w}) \cup \{ 0 \}^{p-k} \), and conversely if \( k > p \). These additional zero eigenvalues in the case \( \alpha > 1 \) explain the \( \text{max}(0,\cdot) \) term in the expression of \( \lambda_{\max} \) in Theorem V.1.

For the remainder of the proof we can thus consider \( \Gamma_{k}^{w} \) instead of \( \Gamma_{k}^{w} \) given the remark above. Moreover, for simplicity we will drop the \( \mathbf{v}^* \) exponent in those matrices, and just denote them \( \Gamma_{k}, \Gamma_{p} \). The bulk of \( \Gamma_{k} \) can be studied using standard random matrix theory results. Such matrices were first studied by Marchenko and Pastur in a seminal work [46], which was generalized (and made rigorous) later in [37]. Note finally that by the celebrated results of Wigner [47], the spectral distribution of the matrix \( \xi = \sqrt{\Delta p} - I_{p} / \Delta \) converges in law (and almost surely) as \( p \rightarrow \infty \) to \( \rho_{\Delta} \), given by eq. (61). We can then use Theorem 1.1 of [37], that we recall here for our setting:

\textbf{Theorem VI.1 (Silverstein-Bai):} Let \( p, k \to \infty \) with \( p/k \to \alpha > 0 \). Let \( W \in \mathbb{R}^{p \times k} \) be an i.i.d. Gaussian matrix, whose elements come from the standard Gaussian distribution \( N(0,1) \). Let \( T_{p} \in \mathbb{R}^{p \times p} \) be a random symmetric matrix, independent of \( W \), such that the empirical spectral distribution of \( T_{p} \) converges (almost surely) in law to a measure \( \rho_{T} \). Then, almost surely, the empirical spectral distribution of \( B_{k} \equiv \frac{1}{p} W^\top T_{p} W \) converges in law to a (nonrandom) measure \( \mu_{B} \), whose Stieltjes transform satisfies, for every \( z \in \mathbb{C}_+ \):
\[ g_{\mu_{B}}(z) = - \frac{z - \alpha}{1 + t g_{\mu_{B}}(z)} \]  

(90)

Moreover, for every \( z \in \mathbb{C}_+ \), there is a unique solution to eq. (90) such that \( g_{\mu_{B}}(z) \in \mathbb{C}_+ \). This equation thus characterizes unambiguously the measure \( \mu_{B} \).

Applying Theorem VI.1 to our setting shows that we can define \( \nu(\alpha, \Delta) \) as the limit eigenvalue distribution of \( \Gamma_{k} \), and we denote \( g_{\nu}(z) \) its Stieltjes transform. From the remarks above, \( \mu(\alpha, \Delta) \) and \( \nu(\alpha, \Delta) \) only differ by the addition of
a delta distribution. For instance, if $\alpha \geq 1$:

$$\mu(\alpha, \Delta) = \alpha \nu(\alpha, \Delta) + (1 - \alpha)\delta_0. \quad (91)$$

The main quantity of interest to us is $z_{\text{edge}}$, defined as the supremum of the support of $\nu(\alpha, \Delta)$. If $z_{\text{edge}} \geq 0$, then it will also be the supremum of the support of $\mu(\alpha, \Delta)$, and thus equal to $\lambda_{\max}$. Theorem VI.1 shows that for every $z \in \mathbb{C}_+ \cup (\mathbb{R} \setminus \text{supp } \nu)$, $g_\nu(z)$ is the only solution in $\mathbb{C}_+ \cup \mathbb{R}$ to the following equation:

$$g_\nu(z) = -\left[ z - \alpha \int \rho_\Delta(dt) \frac{t}{1 + tg_\nu(z)} \right]^{-1}. \quad (92)$$

The validity of the equation for $\mathbb{R} \setminus \text{supp } \nu$ (and not only on $\mathbb{C}_+$) follows from the continuity of $g_\nu(z)$ on $\mathbb{C}_+ \cup (\mathbb{R} \setminus \text{supp } \nu)$, a generic property of the Stieltjes transform. It is easy to see that $g_\nu$ induces a strictly increasing diffeomorphism $g_\nu : (z_{\text{edge}}, +\infty) \rightarrow (\lim_{z \to z_{\text{edge}}} g_\nu(z), 0)$, so that we can define its inverse $g_\nu^{-1}$ and from eq. (92), it satisfies for every $s \in (\lim_{z \to z_{\text{edge}}}^+ g_\nu(z), 0)$:

$$g_\nu^{-1}(s) = \frac{1}{s} + \alpha \int \rho_\Delta(dt) \frac{t}{1 + st}. \quad (93)$$

In order to compute $z_{\text{edge}}$ from eq. (92), we use a result of Section 4 of [48], also stated for instance in [49], that describes the form of the support of $\nu(\alpha, \Delta)$. It can be stated in the following way. Recall that since $\Delta > \frac{1}{\mu}$, $z_1(\Delta) > 0$ is the maximum of the support of $\rho_\Delta$. Let $s_{\text{edge}}$ be the unique solution in $(-z_1(\Delta)^{-1}, 0)$ of the equation $(g_\nu^{-1})'(s) = 0$, that is by eq. (93):

$$\alpha \int \rho_\Delta(dt) \left( \frac{st}{1 + st} \right)^2 = 1. \quad (94)$$

Indeed, it is straightforward to show that the left-hand side of eq. (94) tends to 0 as $s \rightarrow 0^-$, tends to $+\infty$ as $s \rightarrow -z_1(\Delta)^{-1}$, and is strictly decreasing and continuous function of $s$. Then (see for instance eq. (2.13) and eq. (2.14) of [49]) $z_{\text{edge}}$ is given by

$$z_{\text{edge}} = \lim_{s \to s_{\text{edge}}^{-}} g_\nu^{-1}(s),$$

$$= -\frac{1}{s_{\text{edge}}} + \alpha \int \rho_\Delta(dt) \frac{t}{1 + s_{\text{edge}}t}. \quad (95)$$

This ends the proof of $(ii)$.

Let us make a final remark that will be useful in our future analysis. Note that $z_1(\Delta) > 1$ for all $\Delta > 1$. Moreover, for all $\Delta > 1$, we have by an explicit computation:

$$\alpha \int \rho_\Delta(dt) \left( \frac{t}{1 - t} \right)^2 = \frac{\alpha}{\Delta - 1}. \quad (96)$$

By the argument above, this yields the following result, that we state as a lemma:

**Lemma VI.1:** Assume $\Delta > 1$. Then:

(i) If $\Delta < \Delta_c(\alpha)$, then $s_{\text{edge}} > -1$.

(ii) If $\Delta = \Delta_c(\alpha)$, then $s_{\text{edge}} = -1$.

(iii) If $\Delta > \Delta_c(\alpha)$, then $s_{\text{edge}} < -1$.

**Proof of Theorem VI.1 (i):** Assume now $\Delta < \frac{1}{\mu}$. Then the support of $\rho_\Delta$ is a subset of $\mathbb{R}_-$. Since $0 \in \mathbb{R}_-$, we can use again the remark we made in the proof of $(ii)$ to study $\Gamma_c$ instead of $\Gamma_p$. Moreover, Theorem VI.1 still applies here so that we have the Silverstein equation (93) for every $s \in \mathbb{C}_+$:

$$g_\nu^{-1}(s) = -\frac{1}{s} + \alpha \int \rho_\Delta(dt) \frac{t}{1 + st}. \quad (97)$$

By the Stieltjes-Perron inversion theorem given in appendix theorem B.1, it is enough to check that for every $z > 0$, there exists a unique $s < 0$ such that $g_\nu^{-1}(s) = z$. Indeed, this will yield $s = g_\nu(z) \in \mathbb{R}$. In particular, $\lim_{s \to 0^+} \Im g_\nu(z + it) = 0$ for every $z > 0$, which will imply $\text{supp } \nu \subseteq \mathbb{R}_-$ and thus $\text{supp } \mu \subseteq \mathbb{R}_-$. Therefore, let $z > 0$. From eq. (93) and the fact that $\text{supp } \rho_\Delta \subseteq \mathbb{R}_-$, we easily obtain:

$$\lim_{s \to -\infty} g_\nu^{-1}(s) = 0,$$

$$\lim_{s \to 0^+} g_\nu^{-1}(s) = +\infty.$$

Moreover, $g_\nu^{-1}(s)$ is a strictly increasing continuous function of $s$, so that the existence and unicity of $s = g_\nu(z) < 0$ is immediate, which ends the proof.

We now prove the final statement of Theorem V.1, on the behavior of $\lambda_{\max}$ with $\Delta$, at fixed $\alpha$.

**Proof of the behavior of $\lambda_{\max}$ with $\Delta$:** Let us make a few remarks:

- We already showed that if $\Delta \leq \frac{1}{\mu}$, then $\lambda_{\max} \leq 0$.
- It is trivial by the form of $\Gamma_p$ that, as $\Delta \rightarrow +\infty$, $\lambda_{\max} \rightarrow 0$.

Let $z_{\text{edge}} = -\frac{1}{s_{\text{edge}}} + \alpha \int \rho_\Delta(dt) \frac{1}{1 + s_{\text{edge}}t}$. Then we know that $\lambda_{\max} = z_{\text{edge}}$ if $\alpha = 1$ and $\lambda_{\max} = \max(0, z_{\text{edge}})$ if $\alpha > 1$. In particular, by the remark above, $z_{\text{edge}} \leq 0$ for $\Delta = \frac{1}{\mu}$ and $z_{\text{edge}} \rightarrow 0^+$ as $\Delta \rightarrow \infty$. It is easy to see that $z_{\text{edge}}$ is a continuous and differentiable function of $\Delta$, so that if we show the two following facts for any $\Delta \geq \frac{1}{\mu}$:

$$\frac{dz_{\text{edge}}}{d\Delta} = 0 \iff \Delta = \Delta_c(\alpha) = 1 + \alpha, \quad (98)$$

$$z_{\text{edge}}(\Delta_c(\alpha)) = 1, \quad (99)$$

this would end the proof as $z_{\text{edge}}$ would necessarily have a unique global maximum, located in $\Delta = \Delta_c(\alpha)$, in which we have $\lambda_{\max} = 1$. We thus prove eq. (98) and eq. (97) in the following.

**Proof of eq. (98):** By the chain rule:

$$\frac{dz_{\text{edge}}}{d\Delta} = \frac{\partial z_{\text{edge}}}{\partial \Delta} + \frac{\partial s_{\text{edge}}}{\partial \Delta} \frac{dz_{\text{edge}}}{ds_{\text{edge}}},$$

using the very definition of $s_{\text{edge}}$, eq. (94), as $z_{\text{edge}} = g_\nu^{-1}(s_{\text{edge}})$. Given the explicit form of $\rho_\Delta$, one can compute easily:

$$\frac{\partial z_{\text{edge}}}{\partial \Delta} = -\frac{\alpha}{2s_{\text{edge}}} \left[ \frac{s_{\text{edge}} + 2s_{\text{edge}}^2 - \Delta}{\sqrt{s_{\text{edge}}^2 - 2s_{\text{edge}}(1 + 2s_{\text{edge}})\Delta + \Delta^2} + 1 \right].$$


It is then simple analysis to see that since \( s_{\text{edge}} < 0 \), \( \frac{\partial s_{\text{edge}}}{\partial \Delta} = 0 \) is equivalent to \( s_{\text{edge}} = -1 \) and \( \Delta > 1 \). Recall that \( s_{\text{edge}} \) is originally defined as a solution to eq. (94):

\[
\alpha \int \rho_\Delta(dt) \left( \frac{s_{\text{edge}}t}{1 + s_{\text{edge}}t} \right)^2 = 1.
\]

Inserting \( s_{\text{edge}} = -1 \) into this equation and using the explicit form of \( \rho_\Delta \) given by eq. (61), and using moreover that \( \Delta > 1 \), this reduces to:

\[
\frac{\alpha}{\Delta - 1} = 1,
\]

which is equivalent to \( \Delta = \Delta_c(\alpha) = 1 + \alpha \).

**Proof of eq. (97):** By Lemma VI.1, we know that for \( \Delta = \Delta_c(\alpha) \) we have \( s_{\text{edge}} = -1 \). Given eq. (61), it is then straightforward to compute:

\[
z_{\text{edge}}(\Delta_c(\alpha)) = -1 + \alpha \int \rho_{\Delta_c}(\alpha)(dt) \frac{t}{1 - t} = 1.
\]

\[\blacksquare\]

**C. Proof of Theorem V.2**

**Transition of the Largest Eigenvalue:** This part is a detailed outline of the proof. Some parts of the calculation are not fully rigorous, however they can be justified more precisely by following exactly the lines of [36] and [48]. We will emphasize when such refinements have to be made. Recall that we have by eq. (60) the following decomposition of \( \Gamma_p^{(0)} \) (that we denote \( \Gamma_p \) for simplicity):

\[
\Gamma_p = \begin{bmatrix}
\frac{1}{k} W W^T & \frac{1}{\sqrt{\Delta p}} \xi - \frac{1}{\Delta} I_p \\
\frac{1}{\sqrt{\Delta p}} \nu & \nu^T
\end{bmatrix} + \frac{1}{k} W W^T \nu \nu^T + \frac{1}{k} W W^T \nu \nu^T + \frac{1}{k} W W^T \nu \nu^T + \cdots \quad (98)
\]

Theorem V.1, along with its proof, already describe in great detail the limit eigenvalue distribution of \( \Gamma_p^{(0)} \). We first note that for any \( \lambda \in \mathbb{R} \) that is not an eigenvalue of \( \Gamma_p^{(0)} \) one can write:

\[
det(\lambda I_p - \Gamma_p) = \det \left( I_p - \left( \lambda I_p - \Gamma_p^{(0)} \right)^{-1} \frac{1}{k} W W^T \nu \nu^T \right) \times \det(\lambda I_p - \Gamma_p^{(0)}).
\]

In particular, this implies immediately that \( \lambda \) is an eigenvalue of \( \Gamma_p \) and not an eigenvalue of \( \Gamma_p^{(0)} \) if and only if 1 is an eigenvalue of \( \left( \lambda I_p - \Gamma_p^{(0)} \right)^{-1} \frac{1}{k} W W^T \nu \nu^T \). Since this is a rank-one matrix, its only non-zero eigenvalue is equal to its trace, so it is equivalent to:

\[
1 = \text{Tr} \left( \left( \lambda I_p - \Gamma_p^{(0)} \right)^{-1} \frac{1}{k} W W^T \nu \nu^T \right).
\]

Recall that by definition, \( \nu \) is constructed as \( \nu = W z / \sqrt{k} \), with \( z \) a standard Gaussian i.i.d. vector in \( \mathbb{R}^k \), independent of \( W \). For any matrix \( A \), we have the classical concentration \( \frac{1}{k} z^T A z = \frac{1}{k} \text{Tr} A \) with high probability as \( k \to \infty \). In eq. (99), this yields at leading order as \( p \to \infty \):

\[
\Delta = \frac{1}{p} \text{Tr} \left( \left( \lambda I_p - \Gamma_p^{(0)} \right)^{-1} \left( \frac{W W^T}{k} \right)^2 \right).
\]

We will prefer to use \( k \times k \) matrices. We use the simple linear algebra identity, for any \( p \times p \) symmetric matrix \( A \), and any integer \( q \geq 1 \):

\[
\text{Tr} \left[ \left( \lambda I_p - \frac{W W^T}{k} A \right)^{-1} \left( \frac{W W^T}{k} \right)^q \right] = \text{Tr} \left[ \left( \lambda I_p - \frac{1}{k} W^T A W \right)^{-1} \left( W^T W \right)^q \right].
\]

This can be derived for instance by expanding both sides in powers of \( \lambda^{-1} \) and using the cyclicity of the trace. Finally, we can state that the eigenvalues of \( \Gamma_p \) that are outside of the spectrum of \( \Gamma_p^{(0)} \) must satisfy, as \( k \to \infty \):

\[
\alpha \Delta = \frac{1}{k} \text{Tr} \left( \left( \lambda I_p - \Gamma_p^{(0)} \right)^{-1} \left( \frac{W W^T}{k} \right)^2 \right),
\]

with

\[
\Gamma_p^{(0)} = \frac{1}{k} W^T \left( \frac{1}{\sqrt{\Delta p}} \xi - \frac{1}{\Delta} I_p \right) W.
\]

We will now make use of two important lemmas, at the core of our analysis. They will also prove to be useful in the eigenvector correlation analysis.

**Lemma VI.2:** Recall that \( \nu \) is the limit eigenvalue distribution of \( \Gamma_p^{(0)} \), that the supremum of its support is \( \lambda_{\text{max}} \), and its Stieltjes transform is \( g_\nu \). For every integer \( r \geq 0 \), we define:

\[
S_k^{(r)}(\lambda) \equiv \frac{1}{k} \text{Tr} \left[ \left( \Gamma_p^{(0)} - \lambda I_p \right)^{-1} \left( \frac{W W^T}{k} \right)^r \right].
\]

For \( r \in \{0, 1, 2, 3\} \)
and every \( \lambda > \lambda_{\text{max}} \), as \( k \to \infty \) \( S_k^{(r)}(\lambda) \) converges almost surely to a well defined limit \( \lambda_k^{(r)}(\lambda) \). This limit is given by:

\[
\begin{align*}
S_0^{(0)}(\lambda) &= g_\nu(\lambda), \\
S_1^{(1)}(\lambda) &= g_\nu(\lambda) \left[ \alpha - (1 + \lambda g_\nu(\lambda)) \right], \\
S_2^{(2)}(\lambda) &= g_\nu(\lambda) \left[ \alpha(1 + \alpha) - (1 + 2\alpha)(1 + \lambda g_\nu(\lambda)) \right] + (1 + \lambda g_\nu(\lambda))^2, \\
S_3^{(3)}(\lambda) &= g_\nu(\lambda) \left[ (1 + 3\alpha^2 + \alpha^3) + (2 + 3\alpha)(1 + \lambda g_\nu(\lambda)) - (1 + 5\alpha + 3\alpha^2)(1 + \lambda g_\nu(\lambda)) \right] - (1 + \lambda g_\nu(\lambda))^3.
\end{align*}
\]

We define similarly for every integer \( r, q \geq 0 \):

\[
S_k^{(r,q)}(\lambda) \equiv \frac{1}{k} \text{Tr} \left[ \left( \Gamma_p^{(0)} - \lambda I_p \right)^{-1} \left( \frac{W W^T}{k} \right)^r \times \left( \Gamma_p^{(0)} - \lambda I_p \right)^{-1} \left( \frac{W W^T}{k} \right)^q \right].
\]

Note that \( S_k^{(r,q)} = S_k^{(q,r)} \) and that \( S_k^{(r,0)}(\lambda) = \partial_\lambda S_k^{(r)}(\lambda) \). For every \( \lambda > \lambda_{\text{max}} \), \( S_k^{(1,1)}(\lambda) \) and \( S_k^{(1,2)}(\lambda) \) converge almost surely (as \( k \to \infty \)) to well-defined limits, that satisfy the

\[\text{The almost sure convergence could probably be extended to all } r \in \mathbb{N}^* \text{ but we will only use these values of } r \text{ in the following.}\]
following equations:
\[
S^{(1,1)}(\lambda) = g_\nu(\lambda) S^{(2)}(\lambda) - [1 + \lambda g_\nu(\lambda)] \partial_\lambda S^{(1)}(\lambda) + \alpha g_\nu(\lambda) \left[ g_\nu(\lambda) + S^{(1)}(\lambda) \right] \times \frac{\rho_{\Delta}(dt)}{(1 + t g_\nu(\lambda))^2} \left[ t \partial_\lambda S^{(1)}(\lambda) - g_\nu(\lambda) \right],
\]
\[
S^{(1,2)}(\lambda) = -[1 + \lambda g_\nu(\lambda)] \left[ S^{(1,1)}(\lambda) + (1 + \alpha) \partial_\lambda S^{(1)}(\lambda) \right] + \alpha g_\nu(\lambda) \left[ (1 + \alpha) g_\nu(\lambda) + S^{(1)}(\lambda) + S^{(2)}(\lambda) \right] \times \frac{\rho_{\Delta}(dt)}{(1 + t g_\nu(\lambda))^2} \left[ t \partial_\lambda S^{(1)}(\lambda) - g_\nu(\lambda) \right] + g_\nu(\lambda) S^{(3)}(\lambda).
\]

**Lemma VI.3**: Let \( \alpha, \Delta > 0 \). We focus mainly on \( S^{(2)}(\lambda) \). We have:

(i) For every \( r \), \( S^{(r)}(\lambda) \) is a strictly increasing function of \( \lambda \), and \( \lim_{\lambda \to -\infty} S^{(r)}(\lambda) = 0 \).

(ii) For every \( \lambda > \lambda_{\text{max}} \), \( S^{(2)}(\lambda) = -\alpha \Delta \) if and only if \( \Delta = \Delta_{\alpha}(\lambda) \) and \( \lambda = 1 \).

(iii) For every \( \Delta > \Delta_{\alpha}(\lambda) \), \( \lim_{\lambda \to \lambda_{\text{max}}} S^{(2)}(\lambda) \in (-\alpha \Delta, 0) \) (it is well defined by monotonicity of \( S^{(2)}(\lambda) \)).

Let us see how item (ii) of Lemma VI.3 and eq. (101) end the proof of the eigenvalue transition. First, note that by the celebrated Weyl’s interlacing inequalities [50], we have:

\[
\lim_{p \to \infty} \inf \lambda_1 \geq \lambda_{\text{max}},
\]
\[
\lim_{p \to \infty} \sup \lambda_2 \leq \lambda_{\text{max}}.
\]

This implies that because the perturbation of the matrix is of rank one, at most one outlier eigenvalue will exist in the limit \( p \to \infty \). By eq. (101), this outlier \( \lambda_1 \) exists if and only if it satisfies, in the large \( p \to \infty \) limit, the equation \( S^{(2)}(\lambda_1) = -\alpha \Delta \). By item (ii) of Lemma VI.3, this is the case only for \( \lambda_1 = 1 \) and \( \Delta = \Delta_{\alpha}(\lambda) \), which ends the proof.

A completely rigorous treatment of the previous arguments requires to state more precisely concentration results. Such a treatment has been made in [36] in a very close case (from which all the arguments transpose), and we refer to it for more details. The proof of Lemma VI.2, which is at the core of our proof, is given in the following. We postpone the proof of Lemma VI.3 to the appendix, see Sec. B-B.

**Proof of Lemma VI.2**: The essence of the computation originates from the derivation of Theorem VI.1 in [37]. Note that \( S^{(0)}(\lambda) \) converges a.s. to the Stieljes transform \( g_\nu(\lambda) \) as \( k \to \infty \) by Theorem VI.1. For every \( 1 \leq i \leq p \), \( w_i \) denotes the \( i \)-th row of \( W \). We denote \( y = \frac{1}{\sqrt{\Delta} p} \Delta^{-1} \Delta_p \). Since \( W \) is independent of \( y \), we can denote \( y_1, \ldots, y_p \) the eigenvalues of \( y \), and their empirical distribution converges a.s. to \( \rho_\Delta \) as we know. We have in distribution:

\[
\Gamma_{k,0}^{(0)} = \frac{1}{k} W^T y W = \frac{\alpha}{p} \sum_{i=1}^{p} y_i w_i w_i^T.
\]

For every \( i \), denote:

\[
\Gamma_{k,i}^{(0)} = \frac{\alpha}{p} \sum_{j \neq i} y_j w_j w_j^T.
\]

Note that \( \Gamma_{k,i}^{(0)} \) is independent of \( w_1 \). We start from the (trivial) decomposition, for every \( \lambda \):

\[
-\frac{1}{\lambda} = \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1} - \frac{1}{k} W^T y W \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1}.
\]

We will make use of the Sherman-Morrison formula that gives the inverse of a matrix perturbed by a rank-one change:

\[
(B + \tau \omega \omega^T)^{-1} = B^{-1} - \frac{\tau}{1 + \tau \omega^T B^{-1} \omega} B^{-1} \omega^T B^{-1},
\]

\[
\omega^T (B + \tau \omega \omega^T)^{-1} = \frac{1}{1 + \tau \omega^T B^{-1} \omega} \omega^T B^{-1}.
\]

Using it in eq. (103) yields:

\[
-\frac{1}{\lambda} = -\frac{\alpha}{\lambda} \frac{1}{p} \sum_{i=1}^{p} w_i w_i^T \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1} - \frac{1}{k} W^T y W \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1} w_i + \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1}.
\]

Taking the trace of eq. (106), using the independence of \( w_i \) and \( \Gamma_{k,i}^{(0)} \), and the concentration \( \frac{1}{k} W^T y W \) with high probability for large \( k \), we obtain the following equation:

\[
-\frac{1}{\lambda} = g_\nu(\lambda) - g_\nu(0) \frac{\alpha}{\lambda} \int \rho_\Delta(dt) \frac{t}{1 + t g_\nu(\lambda)}.
\]

This is exactly the identity in Theorem VI.1! In the following, we will use very similar identities. A completely rigorous derivation of these would, however, require many technicalities to ensure in particular the concentration of all the involved quantities. It would exactly follow the proof of [37], and thus we do not repeat all the technicalities here. We can multiply eq. (106) by \( \frac{W^T W}{k} \), and take the trace:

\[
-\frac{1}{\lambda} \frac{1}{k} \text{Tr} \left[ \frac{W^T W}{k} S_{k}^{(1)}(\lambda) \right] = \frac{1}{k} \text{Tr} \left[ \frac{W^T W}{k} S_{k}^{(1)}(\lambda) \right] + \frac{\alpha}{\lambda} \frac{1}{p} \sum_{i=1}^{p} w_i w_i^T \left( \Gamma_{k,i}^{(0)} - \lambda I_k \right)^{-1} \left( \frac{1}{k} \sum_{j} w_j w_j^T \right) \frac{w_i}{\sqrt{k}}.
\]

In the large \( p, k \) limit, this implies that \( S_{k}^{(1)}(\lambda) \) converges to a well-defined limit \( S^{(1)}(\lambda) \), and this limit satisfies:

\[
-\frac{\alpha}{\lambda} = S^{(1)}(\lambda) - \frac{\alpha}{\lambda} \int \rho_\Delta(dt) \frac{t}{1 + t g_\nu(\lambda)} \left( g_\nu(\lambda) + S^{(1)}(\lambda) \right).
\]

Using finally eq. (107), it is equivalent to:

\[
S^{(1)}(\lambda) = g_\nu(\lambda) \left[ \alpha - (1 + \lambda g_\nu(\lambda)) \right].
\]

Multiplying eq. (106) by \( \left( \frac{W^T W}{k} \right)^2 \) or \( \left( \frac{W^T W}{k} \right)^3 \) yields, by the same analysis:

\[
S^{(2)}(\lambda) = \left[ \alpha (1 + \alpha) - (1 + 2 \alpha)(1 + \lambda g_\nu(\lambda)) + (1 + \lambda g_\nu(\lambda))^2 \right] g_\nu(\lambda),
\]

\[
S^{(3)}(\lambda) = \left[ \alpha + 3 \alpha^2 + \alpha^3 \right] - (1 + 5 \alpha + 3 \alpha^2)(1 + \lambda g_\nu(\lambda)) + (2 + 3 \alpha)(1 + \lambda g_\nu(\lambda))^2 - (1 + \lambda g_\nu(\lambda))^2 \right] g_\nu(\lambda).
\]
The convergence of $S_k^{(1,1)}(\lambda)$ and $S_k^{(1,2)}(\lambda)$ follows from the same analysis, as well as the equations they satisfy. We detail the derivation of the equation on $S_k^{(1,1)}(\lambda)$ and leave the derivation of the second equation for the reader. We multiply eq. (106) by $\frac{W^T W}{k \lambda}$. To simplify the calculations, we make use of concentrations, and denote $F_i = \frac{W^T W}{k} - \frac{1}{p} w_i w_i^T$, which is independent of $w_i$. We obtain at leading order as $p \to \infty$:

$$-\frac{W^T W}{k \lambda} (\Gamma_k^{(0)} - \lambda I_k)^{-1} = -\frac{1}{\lambda p} \sum_{i=1}^p \frac{y_i w_i w_i^T}{1 + y_i g_r(\lambda)} \left( (\Gamma_k^{(0)} - \lambda I_k)^{-1} F_i (\Gamma_k^{(0)} - \lambda I_k)^{-1} \right)$$

We multiply this equation by $(\Gamma_k^{(0)} - \lambda I_k)^{-1}$ and use Sherman-Morrison formula eq. (104):

$$\left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1} = -\frac{1}{\lambda} \frac{p}{p} \frac{y_i w_i w_i^T}{1 + y_i g_r(\lambda)} (\Gamma_k^{(0)} - \lambda I_k)^{-1}$$

Using again the concentration of $\frac{1}{p} w^T A w$ on $\frac{1}{p} \text{Tr}[A]$, this yields the cumbersome expression:

$$-\frac{W^T W}{k \lambda} (\Gamma_k^{(0)} - \lambda I_k)^{-1} = -\frac{1}{\lambda} \frac{p}{p} \frac{y_i w_i w_i^T}{1 + y_i g_r(\lambda)} (\Gamma_k^{(0)} - \lambda I_k)^{-1}$$

Inserting this into the equation in the middle of the left column on this page, along with some trivial algebra yields:

$$S_k^{(1,1)}(\lambda) = g_r(\lambda) S_k^{(2)}(\lambda) + [1 + \lambda g_r(\lambda)] \partial_{\lambda} S_k^{(1)}(\lambda)$$

The proof of correlation of the leading eigenvector is derived on the same principles, and is given in appendix Sec. B-C. All together, this ends the proof of Theorem V.2.

A note on the nature of the transition As was already noticed in some previous works (see for instance a related remark in [36]), the existence of a transition in the largest eigenvalue and the corresponding eigenvector for a large matrix of the type $M + \theta P$ (with $P$ of finite rank and $\theta > 0$) depends on the decay of the asymptotic spectral density of $M$ at the right edge of its bulk. For a power-law decay, there can be either no transition, a transition in the largest eigenvalue and the corresponding eigenvector, or a transition in the largest eigenvalue but not in the corresponding eigenvector. The situation in our setting is somewhat more involved, as both the bulk and the spike depend on the parameter $\Delta$, and they are not independent (they are correlated via the matrix $W$). However, this intuition remains true: if we do not show and use it explicitly, the decay of the density of $\mu(\alpha, \Delta)$ at the right edge is of the type $(\lambda_{\text{max}} - \lambda)^{1/2}$, which is the hidden feature that is responsible for a transition both in the largest eigenvalue and the corresponding eigenvector, which is what we show in Theorem V.1.

APPENDIX A

TECHNICALITIES FOR THE MUTUAL INFORMATION PROOF

A. Proof of Proposition VI.1

The proof is a verbatim reproduction of the argument of [26] for non-factorized prior. We define

$$\varphi(t) \equiv -\frac{1}{p} \mathbb{E} \log \int e^{-H(v)} dP_v(v).$$

A simple calculation based on Gaussian integration by parts (in technical terms, Stein’s lemma) applied on the Gaussian variables $\xi$ and $z$ shows that (see [26] for details)

$$\varphi'(t) = \frac{\lambda}{4} \mathbb{E} \left< (R_{1,2} - q_v)^2 \right>_t - \frac{\lambda}{4} q_v^2 - \frac{\lambda}{4p^2} \sum_{i=1}^p \mathbb{E} \left< v_{i}^2 v_{i}^2 \right>_t$$

We now use the Nishimori property, and the expressions involving the pairs $(v, v^\ast)$ and $(v^{(1)}, v^{(2)})$ become equal. We thus obtain

$$\varphi'(t) = -\frac{\lambda}{4} \mathbb{E} \left< (R_{1,2} - q_v)^2 \right>_t + \frac{\lambda}{4} q_v^2 + \frac{\lambda}{4p^2} \sum_{i=1}^p \mathbb{E} \left< v_{i}^2 v_{i}^2 \right>_t.$$
Observe that the last term is $O(1/p)$ since the variables $v_i$ are bounded. Moreover, the first term is always non-negative so we obtain

$$\varphi'(t) \leq \frac{\lambda}{4}q^2_c + \frac{K}{p}. \quad (110)$$

Since $\varphi(1) = f_p$ and $\varphi(0) = f_p^0(1/\sqrt{q}c)$, integrating over $t$, we obtain for all $q_c \geq 0$, $f_p \leq \varphi_H(\lambda, q_c) + \frac{K}{p}$, and this concludes the proof of the upper bound proposition.

### B. Proof of Proposition VI.2

The proof proceeds very similarly. Let $t \in [0, 1]$ and consider a slightly different interpolating Hamiltonian

$$-H_t(\mathbf{v}) = \sum_{i<j} \frac{\lambda}{l} \xi_{ij} v_i v_j + \frac{\lambda}{l} \nu_i \nu_j v_i v_j - \frac{\lambda}{2p} \nu_i \nu_j v_i v_j + \frac{1}{2} \nu_i \nu_i v_i v_i \geq -\frac{1}{2} \nu_i \nu_i v_i v_i,$$


Notice the subtle change: in front of the term $(1 - t)\nu_i \nu_j v_i v_j$ we replace the $q_c$ from the former section by $m$. We define now

$$\varphi_{e,m}(t) \equiv -\frac{1}{p} \log \int_{\mathbb{R}^p} e^{-H_t(\mathbf{v})} \mathbb{1}\{R_{1,*} \in [m, m + \epsilon]\} dP_\nu(\mathbf{v}). \quad (111)$$

Denoting now the Gibbs average with the additional constraint $\mathbb{1}\{R_{1,*} \in [m, m + \epsilon]\}$ as $\langle \gamma \rangle_{m,e}$, we find when we repeat the former computation:

$$\varphi_{e,m}(t) = \frac{\lambda}{4} \mathbb{E}\langle (R_{1,*} - q_c)_{m,e} \rangle - \frac{\lambda}{4} q^2_c + \frac{\lambda}{2} m^2 - \frac{\lambda}{2} \mathbb{E}\langle (R_{1,*} - m)_{m,e} \rangle + o(1).$$

The trick is now to notice that, by construction, the $\mathbb{E}\langle (R_{1,*} - m)_{m,e} \rangle \leq e^2$ given the overlap restriction, and therefore

$$\varphi_{e,m}(t) \geq \frac{\lambda}{4} \mathbb{E}\langle (R_{1,*} - q_c)_{m,e} \rangle - \frac{\lambda}{4} q^2_c + \frac{\lambda}{2} m^2 - \frac{\lambda}{2} \mathbb{E}\langle (R_{1,*} - m)_{m,e} \rangle - \frac{\lambda}{2} \mathbb{E}\langle (R_{1,*} - m)_{m,e} \rangle + o(1),$$

and

$$\varphi_{e,m}(t) \geq -\frac{\lambda}{4} q^2_c + \frac{\lambda}{2} m^2 - \frac{\lambda}{2} \mathbb{E}\langle (R_{1,*} - m)_{m,e} \rangle + o(1).$$

We now denote

$$f_p^0(\sigma, \mathbf{v}^*) \equiv -\frac{1}{N^2} \mathbb{E}_z[\log Z_0(y, \sigma)], \quad (112)$$

with the previous $f_p^0$ being the expectation $f_p^0(\sigma) \equiv \mathbb{E}_z[f_p^0(\sigma, \mathbf{v}^*)]$. Then, since $\varphi_{e,m}(1) = \Phi_\nu^0(m, \mathbf{v}^*)$ and $\varphi_{e,m}(0) \geq f_p^0(1/\sqrt{q}c)$ (again, this is an obvious consequence of the restriction in the integrand) integrating over $t$, we obtain a bound for the Parisi-Franz potential for any $q_c$ and $m$. Using, in particular, the value $m = q_c$, this yields the final result.

### C. Proof of Two Technical Propositions

#### Proposition VI.3: This is proven in [25], and we briefly repeat the argument here. Let $\epsilon > 0$. Since the prior $P_\nu$ has bounded support, we can grid the set of the overlap values $R_{1,*}$ by $2K/\epsilon$ many intervals of size $\epsilon$ for some $K > 0$. This allows the following discretisation, where $l$ runs over the finite range $\{-K/\epsilon, \ldots, K/\epsilon\}$:

$$-f_p = \frac{1}{p} \mathbb{E} \log \sum_l \int_{\mathbb{R}^p} \mathbb{1}\{R_{1,*} \in [l\epsilon, (l + 1)\epsilon]\} e^{-H(\mathbf{v})} dP_\nu(\mathbf{v}). \quad (113)$$

One can show that each term $X_l = \frac{1}{p} \log Z_l$ individually concentrates around its expectation with respect to the random variable $\xi$. This follows from the following lemma

**Lemma A.1:** [from [25]] There exists a constant $K > 0$ such that for all $\gamma \geq 0$ and all $l$,

$$\mathbb{E}_\xi e^{\gamma(2K - \mathbb{E}_\xi[X_l])} \leq \frac{\gamma K}{\sqrt{p}} e^{K^2/2p}. \quad (115)$$

This is a direct consequence of the Tsirelson-Ibragimov-Sudakov inequality [51], see [25], Lemma 7. Given that all $X_l$ concentrates, the expectation of the maximum concentrates as well:

$$\mathbb{E}_\xi \max_l (X_l - \mathbb{E}_\xi[X_l]) \leq \frac{\gamma K}{\sqrt{p}} e^{K^2/2p}. \quad (116)$$

Therefore, inserting the above estimates into (113), we obtain

$$-f_p \leq \mathbb{E}_\nu \max_l \mathbb{E}_\xi X_l + \frac{\log(K/\epsilon)}{\sqrt{p}} + \frac{\log(K/\epsilon)}{p} \frac{\gamma K}{\sqrt{p}}.$$

We set $\gamma = \sqrt{p}$ and obtain

$$\mathbb{E}_\xi \max_l (X_l - \mathbb{E}_\xi[X_l]) \leq \frac{\log(K/\epsilon)}{\sqrt{p}}. \quad (116)$$

Therefore, inserting the above estimates into (113), we obtain

$$-f_p \leq \mathbb{E}_\nu \max_l \mathbb{E}_\xi X_l + \frac{\log(K/\epsilon)}{\sqrt{p}} + \frac{\log(K/\epsilon)}{p} \frac{\gamma K}{\sqrt{p}}.$$
so that finally
\[ f_p \geq \mathbb{E}_\nu \min_i \Phi_i(Ie, \nu^*) - \frac{\log(K/\epsilon)}{\sqrt{p}}, \]
for some constant \( K \).

Proposition VI.4: Here we need to pay attention to the fact that the prior is not separable, and thus at this point the proof differs from form \([25]\). We wish to push the expectation with respect to \( \nu^* \) inside the minimum. We start by using again \( q_v = le \) and defining the following random (in \( \nu^* \) variable):
\[ \tilde{X}_i = -\left( f^0_p \left( 1/\sqrt{\lambda_i}, \nu^* \right) + \frac{\lambda_i^2}{4} \right) \]
and start from Proposition VI.3:
\[ -f_p \leq \mathbb{E}_\nu \left[ \max_{|i| \leq K/\epsilon} \tilde{X}_i \right] + \frac{\lambda^2}{2} + \frac{\log(K/\epsilon)}{\sqrt{p}}. \]  

We now wish to push the max inside. We proceed as follow:
\[ \mathbb{E}_\nu \left[ \max_{|i| \leq K/\epsilon} \left( \tilde{X}_i - \mathbb{E}[\tilde{X}_i] \right) \right] \leq \mathbb{E}_\nu \left[ \sum_i \left( \tilde{X}_i - \mathbb{E}[\tilde{X}_i] \right) \right] = \sum_i \mathbb{E}_\nu \left[ \left( \tilde{X}_i - \mathbb{E}[\tilde{X}_i] \right) \right] \leq \sum_i \frac{C}{\sqrt{p}} = \frac{K}{\epsilon \sqrt{p}} \]
Inserting this in eq.(118) we find that
\[ -f_p \leq \max_{|i| \leq K/\epsilon} \mathbb{E}_\nu \tilde{X}_i + \frac{\lambda^2}{2} + \frac{K'}{\epsilon \sqrt{p}} + \frac{\log(K/\epsilon)}{\sqrt{p}}, \]
and therefore,
\[ f_p \geq \min_{|i| \leq K/\epsilon} \left[ -\mathbb{E}_\nu \tilde{X}_i - \frac{\lambda^2}{2} - \frac{K'}{\epsilon \sqrt{p}} - \frac{\log(K/\epsilon)}{\sqrt{p}} \right], \]
so that choosing finally \( \epsilon = p^{-1/4} \) we reach
\[ f_p \geq \min_{|i| \leq K/\epsilon} \left[ f^0_p \left( 1/\sqrt{\rho_v} \right) + \frac{\lambda_i^2}{4} \right] + o \left( \frac{\log p}{\sqrt{p}} \right). \]
Therefore, the claim is proven.

D. Proof of Theorem VI.2

This is a simple application of the I-MMSE theorem \([24]\), that has been used in this context multiple-times (see e.g \([14], [18], [44]\)). Indeed, the I-MMSE theorem states that, denoting \( \lambda = \Delta^{-1} \):
\[ \frac{d}{d\lambda} \frac{I}{p} = \frac{1}{4} \text{Matrix} - \text{MMSE}(Y) \]
We thus need to compute the derivative of the mutual information:
\[ \frac{d}{d\lambda} i_{RS}(q^*_v, \Delta = 1/\lambda) = \partial_{\lambda} i_{RS}(q^*_v, \Delta = 1/\lambda) + \partial_{q_v} i_{RS}(q_v, \Delta = 1/\lambda)|_{q^*_v} \partial_{\lambda}(q^*_v) \]
\[ = \partial_{\lambda} i_{RS}(q^*_v, \Delta = 1/\lambda) \] (124)
where we used \( \partial_{q_v} i_{RS}(q, \Delta = 1/\lambda)|_{q^*_v} = 0 \). Denoting then
\[ I(\lambda_v, q_v) = \lim_{p \to \infty} \frac{1}{p} (\nu^* + \sqrt{\nu^*}) \] we find
\[ \partial_{\lambda} i_{RS}(q^*_v, \Delta = 1/\lambda) = \frac{(\rho_v - q^*_v)^2}{4} + \partial_{\lambda} I(\lambda_v, q_v)|_{q^*_v} \] (125)
We now use the fact that the derivative of the replica mutual information is zero at \( q^* \). This implies
\[ \frac{\lambda}{2} (\rho_v - q^*_v) = \partial_{q_v} I(\lambda_v, q_v)|_{q^*_v} = \frac{\lambda}{q^*_v} \partial_{\lambda} I(\lambda_v, q_v)|_{q^*_v} \] (126)
so that
\[ \partial_{\lambda} i_{RS}(q^*_v, \Delta = 1/\lambda) = \frac{(\rho_v - q^*_v)^2}{4} + \frac{1}{2} (\rho_v - q^*_v) q^*_v \]
\[ = \frac{1}{4} (\rho_v^2 - (q^*_v)^2) \] (127)
which proves the claim.

APPENDIX B

TECHNICALITIES FOR THE RANDOM MATRIX ANALYSIS

A. Reminder on the Stieltjes Transform

Let \( C_+ = \{ z \in C, \text{Im} z > 0 \} \). For any probability measure \( \nu \) on \( \mathbb{R} \), and any \( z \in C \setminus \text{supp} \nu \), we can define the Stieltjes transform of \( \nu \) as:
\[ g_{\nu}(z) = \mathbb{E}_{\nu} \frac{1}{X - z}, \]
Note that \( g_{\nu}(z) \) is a one-to-one mapping of \( C_+ \) on itself. The Stieltjes transform has proven to be a very useful tool from random matrix theory. One of its important features, that we will use to compute the bulk density (see Fig. 5 of the main material) is the Stieltjes-Perron inversion formula, that we state here (see Theorem X.6.1 of [52]):

Theorem B.1 (Stieltjes-Perron): Assume that \( \nu \) has a continuous density on \( \mathbb{R} \) with respect to the Lebesgue measure. Then:
\[ \forall x \in \mathbb{R}, \quad \frac{d\nu}{dx} = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \text{Im} g_{\nu}(x + i\epsilon). \]
Informally, one has to think that the knowledge of the Stieltjes transform above the real line uniquely determines the measure \( \nu \). Consider a (random) symmetric matrix \( M \) of size \( n \), with real eigenvalues \( \{ \lambda_i \} \). Then the empirical spectral measure of \( M \) is defined as:
\[ \nu_n \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}. \] (128)
For the random matrix ensembles considered here, the (random) probability measure \( \nu_n \) will converge almost surely and in the weak sense to a deterministic probability measure \( \nu \) as \( n \to \infty \). In this case, we will call \( \nu \) the asymptotic spectral measure of \( M \).
B. Proof of Lemma VI.3

Proof: Point (i) is trivial by definition of $S^{(c)}_k(\lambda)$ and the almost sure convergence proven in Lemma VI.2. We turn to points (ii) and (iii). Let us denote the following function:

$$T^{(2)}(s) = s \left[ (1 + \alpha) - (1 + 2\alpha) (1 + sg^{-1}_{p}(s)) + (1 + sg^{-1}_{p}(s))^2 \right].$$

By Lemma VI.2, we have $T^{(2)}(s) = S^{(2)}(g^{-1}_{p}(s))$ so $T^{(2)}(s) < 0$ for $s \in (s_{\text{edge}}, 0)$ by negativity of $S^{(2)}(\lambda)$ (as the trace of a negative matrix). Therefore, point (ii) is equivalent to:

$$\forall s \in (s_{\text{edge}}, 0), \quad T^{(2)}(s) = -\alpha \Delta \Leftrightarrow s = g_{p}(1) \quad \text{and} \quad \Delta \leq \Delta_c(\alpha),$$

while point (iii) means that for every $\Delta > \Delta_c(\alpha)$,

$$\forall s \in (s_{\text{edge}}, 0), \quad T^{(2)}(s) > -\alpha \Delta.$$ (129)

The condition $s > s_{\text{edge}}$ arises naturally as the counterpart of $z \geq \lambda_{\text{max}}$. Recall that by Theorem V.1, we have $\lambda_{\text{max}} \leq 1$ for all $\Delta$. As $g^{-1}_{p}(s)$ is here completely explicit by eq. (93), and recalling the form of $\rho_{\Delta}$ in eq. (61), it is easy to show by an explicit computation the following identity:

$$\forall s \neq -1, \quad T^{(2)}(s) = -\alpha \Delta + \alpha \left[ g^{-1}_{p}(s) - 1 \right] \times \left[ s - \Delta - 2s\Delta + \sqrt{s^2 - 2s(1+s)\Delta + \Delta^2} \right],$$

$$T^{(2)}(-1) = \begin{cases} -\alpha(1+\alpha) & \text{if } \Delta \geq 1, \\ -\alpha\Delta(1+\alpha) & \text{if } \Delta \leq 1. \end{cases}$$

It is then easy to see that the only possible solution to $T(s) = -\alpha \Delta$ with $s \in (s_{\text{edge}}, 0)$ is $s = g_{p}(1)$, if $g_{p}(1) \neq -1$. However, by Lemma VI.1, for any $\Delta > \Delta_c(\alpha)$ we have $s_{\text{edge}} < -1$. Moreover, in this case, one computes very easily (all expressions are explicit) $g^{-1}_{p}(-1) = 1$. Given the identity above, there is therefore no solution to $T^{(2)}(s) = -\alpha \Delta$ in $(s_{\text{edge}}, 0)$. By continuity of $T^{(2)}(s)$, and since $\lim_{s \to -\infty} T^{(2)}(s) = 0$, this implies $T^{(2)}(s) > -\alpha \Delta$ for $s \in (s_{\text{edge}}, 0)$, which proves point (iii).

Assume now $\Delta \leq \Delta_c(\alpha)$. Note that the case $\Delta = \Delta_c(\alpha)$ is easy, as $s_{\text{edge}} = -1$ is the unique solution to $T^{(2)}(s) = -\alpha(1+\alpha)$. For $\Delta < \Delta_c(\alpha)$, by Lemma VI.1 we obtain $-1 < s_{\text{edge}}$. In particular, $g_{p}(1) > s_{\text{edge}} > -1$, and we thus have that $s = g_{p}(1)$ is a solution (and the only one) to $T^{(2)}(s) = -\alpha \Delta$ by the identity shown above. This shows (ii) and ends the proof of Lemma VI.3.

C. Proof of Correlation of the Leading Eigenvector

We now turn to the study of the leading eigenvector. Let $\tilde{v}$ be an eigenvector associated with the largest eigenvalue $\lambda_1$, normalized such that $\|\tilde{v}\|^2 = p$. Then we have:

$$(\lambda_1 I_p - \Gamma^{(0)}_p) \tilde{v} = \frac{1}{\Delta} W W^\top \tilde{v} \tilde{v} = p \tilde{v}. \quad \text{(131)}$$

By normalization of $\tilde{v}$, we obtain:

$$\tilde{v} = \sqrt{p} \frac{\left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1} W W^\top}{\sqrt{\text{Tr} W W^\top (\lambda I_p - \Gamma^{(0)}_p)^{-1}}} \tilde{v},$$

where for convenience we have defined the shorthand:

$$A = \left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1} \left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1}. \quad \text{(132)}$$

Therefore:

$$\frac{1}{p^2} \left| \tilde{v}^\top \tilde{v} \right|^2 = 1 + \frac{\left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1} W W^\top}{\sqrt{\text{Tr} W W^\top (\lambda I_p - \Gamma^{(0)}_p)^{-1}}} \tilde{v}^\top A \left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1} \left( \lambda I_p - \Gamma^{(0)}_p \right)^{-1} \tilde{v}. \quad \text{(133)}$$

The numerator is equal to $[\alpha^{-1} S^{(2)}_k(\lambda_1)]^2$, using the $S^{(c)}$ functions that we introduced in Lemma VI.2. We turn to the denominator. Recall that we can write $\Gamma^{(0)}_p = W W^\top M/k$, with a symmetric matrix $M$ that is independent of $W$. For any $z$ large enough, we can expand:

$$\text{Tr} \left\{ \left( z I_p - \Gamma^{(0)}_p \right)^{-1} \left( z I_p - \Gamma^{(0)}_p \right)^{-1} \left( W W^\top W \right)^3 \right\} = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} z^{-a-b-2} \text{Tr} \left\{ \left( W W^\top W \right)^{a} \left( W W^\top \right)^{b} \times (W W^\top)^3 \right\},$$

$$\sum_{a=0}^{\infty} \sum_{b=0}^{\infty} z^{-a-b-2} \text{Tr} \left\{ \left( W W^\top W \right)^{a} \left( W W^\top \right)^{b} \times (W W^\top)^3 \right\} = k S^{(1,2)}_k(z), \quad \text{(a)}$$

where in (a) we used the cyclicity of the trace. Given the last statement of Theorem V.1, we know $\lim_{p \to \infty} \lambda_1 \geq \lambda_{\text{max}}$, so we can use the above calculation to write:

$$\epsilon(\Delta) = \lim_{\lambda \to \lambda_1} \lim_{k \to \infty} \frac{1}{\alpha} \frac{S^{(2)}_k(\lambda)}{S^{(1,2)}_k(\lambda)}.$$ (134)

As in the eigenvalue transition proof, to make this fully rigorous one would need to use more precisely the concentration results, and follow exactly the lines of [36]. We now use the (already proven) transition of the leading eigenvalue, that gives us the value of $\lambda_1$. 

For $\Delta < \Delta_c(\alpha)$, we know that $\lambda_1$ converges almost surely to 1. Consequently, we have in this case:

$$\epsilon(\Delta) = \frac{1}{\alpha} \frac{\left| S^{(1)}(1) \right|^2}{S^{(2)}(1)}.$$  

By Lemma VI.3, we know that $S^{(2)}(1) = -\alpha \Delta$. Moreover, by Theorem V.1, $\lambda_{max} < 1$. This implies that $S^{(1,2)}(1) \in (0, +\infty)$. Indeed, 1 is out of the bulk of $\nu(\alpha, \Delta)$, so $g_\nu(1) \in (-\infty, 0)$ and by the relations shown in Lemma VI.2, all the transforms $S^{(r)}(1)$ and $S^{(r,q)}(1)$ will be finite. Note that $S^{(1,2)}(1) > 0$ by positivity of the matrices involved. This implies that for every $\Delta < \Delta_c(\alpha)$, $\epsilon(\Delta) > 0$. 

For $\Delta = \Delta_c(\alpha)$, we have $\lambda_{max} = 1$ and $\lim_{\Delta \downarrow \Delta_c(\alpha)} S^{(2)}(1) = -\alpha \Delta$ as we have shown. For every $r, q$, let us define the functions $T^{(r)}(s)$ and $T^{(r,q)}(s)$ by $S^{(r)}(\lambda) = T^{(r)}[g_\nu(\lambda)]$ and $S^{(r,q)}(\lambda) = T^{(r,q)}[g_\nu(\lambda)]$. By Lemma VI.2 and the chain rule, we have:

$$\forall s \in (s_{edge}, 0),$$

$$T^{(1,2)}(s) = sT^{(3)}(s) - \left[ 1 + s g_\nu^{-1}(s) \right] \times$$

$$\times \left[ T^{(1,1)}(s) + (1 + \alpha) \frac{\partial T^{(1)}(s)}{\partial g_\nu^{-1}(s)} \right]$$

$$+ \alpha s \left[ (1 + \alpha) s + T^{(1)}(s) + T^{(2)}(s) \right] \times$$

$$\times \int \frac{\rho_\Delta(dt)}{(1 + ts)} \left[ \frac{\partial T^{(1)}(s)}{\partial g_\nu^{-1}(s)} - s \right].$$

Recall that $g_\nu^{-1}(s)$ is explicit by eq. (93) and $s_{edge} = \lim_{\lambda \to \lambda_{max}} g_\nu(\lambda)$. It moreover satisfies (cf Theorem V.1) $\partial_s g_\nu^{-1}(s_{edge}) = 0$. For $\Delta = \Delta_c(\alpha)$, by Lemma VI.1 we have $g_\nu(1) = -1 = s_{edge}$. It is then only trivial algebra to verify from eq. (135) and the remaining relations of Lemma VI.2 that $T^{(1,2)}(-1) = +\infty$, which implies $\epsilon(\Delta_c(\alpha)) = 0$. 

We investigate here the $\Delta \to 0$ limit. In this limit, we know from eq. (134) and the analysis in the case $\Delta < \Delta_c(\alpha)$ above that:

$$\lim_{\Delta \to 0} \epsilon(\Delta) = \lim_{\Delta \to 0} \frac{\alpha \Delta^2}{S^{(2)}(1)}.$$  

This yields $\lim_{\Delta \to 0} \epsilon(\Delta) = 1$.

Finally, we consider $\Delta > \Delta_c(\alpha)$. By eq. (134) and item (iii) of Lemma VI.3, to obtain $\epsilon(\Delta) = 0$ we only need to prove that $\lim_{\Delta \to \Delta_{max}} S^{(1,2)}(\lambda) = +\infty$. Equivalently, we must show $\lim_{s \to s_{edge}} T^{(1,2)}(s) = +\infty$. Recall that $\partial_s g_\nu^{-1}(s_{edge}) = 0$ and that since $s_{edge}$ is finite, all $T^{(r)}(s_{edge})$ for $r = 0, 1, 2, 3$ are finite as well by Lemma VI.2. It thus only remains to check that $\lim_{s \to s_{edge}} T^{(1,2)}(s) \partial_s g_\nu^{-1}(s) > 0$. This would imply that $\lim_{s \to s_{edge}} T^{(1,2)}(s) = +\infty$. We put this statement as a lemma, actually stronger than what we need:

**Lemma B.1:** For every $\alpha > 0$ and $\Delta > 1$, we have

$$\liminf_{s \to s_{edge}} T^{(1,2)}(s) \partial_s g_\nu^{-1}(s) > 0.$$  

We prove this for every $\Delta > 1$, while only the case $\Delta > 1 + \alpha$ is needed in our analysis. As already argued, this lemma ends the proof.

**Proof of Lemma B.1:** The idea is to lower bound $S^{(1,2)}(\lambda)$ by $\partial_\lambda g_\nu(\lambda)$, for every $\lambda > \lambda_{max}$. We separate three cases:

- First, assume $\alpha > 1$. Then $W^TW/k$ is now full rank. In particular, by the classical results of [46], its lowest eigenvalue, denoted $\zeta_{min}$ converges almost surely to $(1 - \alpha^{-1/2})^2$, the left edge of the Marchenko-Pastur distribution. Moreover, for any two symmetric positive square matrices $A$ and $B$, we know that $\text{Tr} [AB] \geq 0$. Indeed, there exists a positive square root of $A$, and $\text{Tr} [AB] = \text{Tr} [A^{1/2}BA^{1/2}] \geq 0$. This implies immediately that if $a_0$ is the smallest eigenvalue of $A$, then $\text{Tr} [AB] \geq a_0 \text{Tr} [B]$, as $A - a_0 I$ is positive. We can use this to write, for any $\lambda > \lambda_{max}$:

$$S^{(1,2)}(\lambda) = \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1} \frac{W^TW}{k} \right]$$

$$\times \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1} \frac{W^TW}{k} ,$$

$$\geq \zeta_{min} \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1} \frac{W^TW}{k} \right]$$

$$\geq \zeta_{min} \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-2} \right].$$

Taking the limit $k \to \infty$ in this last inequality, we obtain:

$$S^{(1,2)}(\lambda) \geq \left( 1 - \alpha^{-1/2} \right)^6 \partial_\lambda g_\nu(\lambda).$$  

Taking the limit $\lambda \to \lambda_{max}$ (or equivalently $s \to s_{edge}$) yields

$$\liminf_{s \to s_{edge}} T^{(1,2)}(s) \partial_s g_\nu^{-1}(s) \geq \left( 1 - \alpha^{-1/2} \right)^6 > 0.$$  

(137)

- Now assume $\alpha < 1$. We do the same reasoning, as $W^TW/k$ is now full rank, and it smallest eigenvalue, also denoted $\zeta_{min}$ converges a.s. as $k \to \infty$ to $(1 - \sqrt{\alpha})^2$. We know (see the beginning of the current proof of the eigenvector correlation) that we can rewrite $S^{(1,2)}(\lambda)$ as the trace of a $p \times p$ matrix:

$$S^{(1,2)}(\lambda) = \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} \right)^\top - \lambda I_k \right]^{-1} \times$$

$$\times \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1} \frac{W^TW}{k} \right]$$

$$\geq \zeta_{min} \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} \right)^\top - \lambda I_k \right]^{-1} \times$$

$$\geq \zeta_{min} \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-2} \right].$$

Taking the limit $k \to \infty$ in this last inequality, we obtain:

$$S^{(1,2)}(\lambda) \geq \left( 1 - \alpha^{-1/2} \right)^6 \partial_\lambda g_\nu(\lambda).$$  

Taking the limit $\lambda \to \lambda_{max}$ (or equivalently $s \to s_{edge}$) yields

$$\liminf_{s \to s_{edge}} T^{(1,2)}(s) \partial_s g_\nu^{-1}(s) \geq \left( 1 - \alpha^{-1/2} \right)^6 > 0.$$  

(137)
\[
\times \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-1},
\]

\[
\geq \zeta_{\min} \frac{1}{k} \text{Tr} \left[ \left( \Gamma_k^{(0)} - \lambda I_k \right)^{-2} \right],
\]
in which the last inequality comes from \( \text{Tr} [AA^T] \geq \text{Tr} [A^2] \) for any positive square matrix \( A \). Once again, taking the limit \( k \to \infty \), and then the limit \( \lambda \to \lambda_{\text{max}} \), this yields

\[
\liminf_{s \to s_{\text{edge}}} T^{(1,2)}(s) \partial_s g_{-1}^{-1}(s) \geq \left( 1 - \alpha/2 \right)^6 > 0.
\]

(138)

Finally, we treat the \( \alpha = 1 \) case. In this case, we cannot use easy bounds as in the two previous cases as the support of the Marchenko-Pastur distribution touches 0. However, recall that everything is explicit here: \( \rho_\Delta \) is given by eq. (61), \( g_{-1}^{-1}(s) \) is given by eq. (93) and Lemma VI.2 gives all the \( T^{(r)} \) and \( T^{(r,q)} \) in terms of \( g_{-1}^{-1} \) and \( \rho_\Delta \). We can moreover use what we proved in Theorem V.1:

\[
\partial_s g_{-1}^{-1}(s_{\text{edge}}) = \frac{1}{s^2} - \alpha \int \rho_\Delta(dt) \frac{t^2}{(1 + ts_{\text{edge}})^2} = 0.
\]

This can be used to simplify the term \( \partial_s T^{(1)}(s) \) and the term \( \int \rho_\Delta(dt) \frac{t^2}{(1 + ts)^2} \). Some heavy but straightforward algebra yields from these relations that the following limit is finite, and is given by:

\[
\lim_{s \to s_{\text{edge}}} T^{(1,2)}(s) \partial_s g_{-1}^{-1}(s) = h(s_{\text{edge}}),
\]

with

\[
\begin{align*}
\Gamma_{\Delta} &= \frac{h_1(s)^2}{4s^4}, \\
h_1(s) &= -\Delta + \sqrt{\Delta^2 + s^2 - 2\Delta(2s + 1)s} + s, \\
h_2(s) &= 3\Delta - 3\sqrt{\Delta^2 + s^2 - 2\Delta(2s + 1)s} + s(4s - 3),
\end{align*}
\]

It is then very simple algebra (solving quadratic equations and using \( \Delta > 1 \)) to see that there is no real negative solution to \( h(s) = 0 \), and that \( h(s) > 0 \) for all \( s \in (-\infty, 0) \). This implies that \( h(s_{\text{edge}}) > 0 \), which ends the proof.

\section{D. A Note on Non-Linear Activation Functions}

We consider here a non-linear activation function, in the spiked Wigner model or the spiked Wishart model. In these models, the spectral method with a non-linear activation function consists in taking the largest eigenvalue and the corresponding eigenvector of the matrix \( \Gamma_p^{uv} \) (for the spiked Wigner model) or \( \Gamma_p^{uv} \) (for the spiked Wishart model). These matrices are given by:

\[
\begin{align*}
\Gamma_p^{uv} &= \frac{1}{\Delta} \left( (a - b)I_p + b \frac{W W^\top}{k} + c \frac{1_p 1_p^\top W^\top}{\sqrt{k}} \right) \\
\times \frac{Y}{\sqrt{p}} - a I_M, \\
\Gamma_p^{uv} &= \frac{1}{\Delta} \left( (a - b)I_p + b \frac{W W^\top}{k} + c \frac{1_p 1_p^\top W^\top}{\sqrt{k}} \right) \\
\times \frac{1}{a + \frac{\Delta}{\sigma^2} - d \beta I_p}
\end{align*}
\]

in these equations, \( a, b, c \) are coefficients that depend on the non-linearity. In the linear case, \( c = 0 \) and \( a = b = 1 \). Let us now assume for instance a non-linearity such that \( a, b \neq 0 \) and \( c = 0 \). Both \( \Gamma_p^{uv} \) and \( \Gamma_p^{uv} \) can be represented as

\[
\Gamma_p = \left[ (a - b)I_p + b \frac{W W^\top}{k} - M, \right.
\]

(139)

in which \( M \) is a symmetric (non necessarily positive or negative) matrix, independent of \( W \). In order to perform the same analysis we made in the case of a linear activation function, we need in particular to be able to characterize the bulk of such matrices. Although this might be doable with more refined techniques, this does not seem to come as a direct consequence of the analysis of Silverstein and Bai [46], [48]. Indeed, one cannot write that the eigenvalues of \( \Gamma_p \) are identical, up to 0 eigenvalues, to the ones of a matrix of the type

\[
\frac{1}{k} W^\top M' W,
\]

which are the types of matrices covered by the analysis of Bai and Silverstein. Moreover, it is not immediate to use results of free probability [53] in this context. Indeed, \( \Gamma_p \) in eq. (139) is the product of two matrices that are asymptotically free, but \( M \) is not positive, which prevents a priori the use of the classical results on the \( S \)-transform of a product of two asymptotically free matrices. Writing \( \Gamma_p \) as the sum of \( (a - b)M \) and \( b(W W^\top)M/k \) does not yield any obvious results either, as these two matrices are not asymptotically free. For this reason, and although there might exist techniques to study the bulk of the matrix of eq. (139) and the transition in its largest eigenvalue, this is left for future work.

\section{ACKNOWLEDGMENT}

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. They would like to thank Google Cloud for providing them access to their platform through the Research Credits Application program. They also thank A. El Alaoui for insightful discussions about the proof of the Bayes optimal performance, and R. Monasson for his insightful lecture series that inspired partly this work.

\section{REFERENCES}

[1] D. L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[51] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.

[52] N. Dunford and J. T. Schwartz, *Linear Operators. 2, Spectral Theory: Self Adjoint Operators in Hilbert Space*. Hoboken, NJ, USA: Interscience Publishers, 1967.

[53] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free Random Variables*, no. 1. Providence, RI, USA: AMS, 1992.

Benjamin Aubin received the M.Eng. degree from École Polytechnique in 2016, and the M.Sc. degree in theoretical physics from École Normale Supérieure in 2017. He is currently pursuing the Ph.D. degree with the Institute of Theoretical Physics, University Paris-Saclay. His main interest focuses on the theoretical understanding of artificial neural-networks from a statistical physics perspective.

Bruno Loureiro received the B.Sc. degree in mathematics and physics from King’s College London in 2013, and the M.A.St. degree in applied mathematics and the Ph.D. degree in physics from the University of Cambridge, in 2014 and 2018, respectively. He then moved to the University of Cambridge. From 2018 to 2020, he was a Post-Doctoral Fellow with the Institut de Physique Théorique in Saclay. In 2020, he joined the Information, Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne, where he has been a Post-Doctoral Fellow since. His research interests lie on the crossroads between machine learning, information theory, and physics.

Antoine Maillard received the B.S. and M.S. degrees in theoretical physics and mathematics from École Normale Supérieure, Paris, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with École Normale Supérieure with Prof. Florent Krzakala and Prof. Lenka Zdeborová. In 2018, he obtained a Jean-Pierre Aguilar Ph.D. Scholarship from the Fondation CFM pour la recherché. His research interests are at the intersection of high-dimensional statistics, statistical physics, information theory, and random matrix theory.

Florent Krzakala (Member, IEEE) received the Ph.D. degree in physics from University Paris-Sorbonne in 2002. He is currently a Professor of physics and of electrical engineering with École Polytechnique Fédérale de Lausanne. He held positions in Rome La Sapienza, ESPCI Paristech and École Normale Supérieure before moving to Lausanne in 2020. He was the holder of the CFM-ENS DataScience Chair, of a PRAIRIE Institute Chair and of an ERC. He is also a Funder and Scientific Advisor of the startup Lighton. His research interests include statistical physics, machine learning, probability, statistics, signal processing, computer science, and computational optics.

Lenka Zdeborová received the Ph.D. degree in physics from University Paris-Sud and from Charles University, Prague, in 2008. She is currently a Professor of physics and of computer science with École Polytechnique Fédérale de Lausanne. She spent two years in the Los Alamos National Laboratory as the Director’s Postdoctoral Fellow. From 2010 to 2020, she was a Researcher with CNRS working in the Institute of Theoretical Physics in CEA Saclay, France. In 2014, she was awarded the CNRS Bronze Medal, in 2016 Philippe Meyer Prize in theoretical physics and an ERC Starting Grant, in 2018 the Irène Joliot-Curie Prize. She is an editorial board member of the *Journal of Physics A: Physical Review E, Physical Review X, SIMODS, Machine Learning: Science and Technology*, and *Information and Inference*. Her expertise is in applications of methods developed in statistical physics, such as advanced mean field methods, replica method and related message-passing algorithms, to problems in machine learning, signal processing, inference and optimization.