The price of ignorance: how much does it cost to forget noise structure in low-rank matrix estimation?

Jean Barbier[*]
International Center for Theoretical Physics
jbarbier@ictp.it

Tianqi Hou
Huawei
thou@connect.ust.hk

Marco Mondelli[*]
Institute of Science and Technology Austria
marco.mondelli@ist.ac.at

Manuel Sáenz[*]
International Center for Theoretical Physics
msaenz@ictp.it

Abstract

We consider the problem of estimating a rank-1 signal corrupted by structured rotationally invariant noise, and address the following question: *how well do inference algorithms perform when the noise statistics is unknown and hence Gaussian noise is assumed?* While the matched Bayes-optimal setting with unstructured noise is well understood, the analysis of this mismatched problem is only at its premises. In this paper, we make a step towards understanding the effect of the strong source of mismatch which is the noise statistics. Our main technical contribution is the rigorous analysis of a Bayes estimator and of an approximate message passing (AMP) algorithm, both of which incorrectly assume a Gaussian setup. The first result exploits the theory of spherical integrals and of low-rank matrix perturbations; the idea behind the second one is to design and analyze an artificial AMP which, by taking advantage of the flexibility in the denoisers, is able to "correct" the mismatch. Armed with these sharp asymptotic characterizations, we unveil a rich and often unexpected phenomenology. For example, despite AMP is in principle designed to efficiently compute the Bayes estimator, the former is *outperformed* by the latter in terms of mean-square error. We show that this performance gap is due to an incorrect estimation of the signal norm. In fact, when the SNR is large enough, the overlaps of the AMP and the Bayes estimator coincide, and they even match those of optimal estimators taking into account the structure of the noise.

1 Introduction

The estimation of a low-rank matrix from noisy data is a central problem in machine learning, and it appears, e.g., in sparse principal component analysis (PCA) [50, 88], community detection [1, 62], and group synchronization [69]. In this paper, we consider the prototypical task of recovering a symmetric rank-1 matrix $\boldsymbol{X}\boldsymbol{X}^\top$ from noisy observations of the form

$$\boldsymbol{Y} = \frac{\sqrt{\lambda_*}}{N}\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{Z} \in \mathbb{R}^{N\times N}. \tag{1}$$

Here, $\lambda_* > 0$ is the signal-to-noise ratio (SNR) which quantifies the signal strength, and $\boldsymbol{Z} \in \mathbb{R}^{N\times N}$ is a random matrix that captures the noise. A natural estimator of $\boldsymbol{X}$ is given by the principal eigenvector of $\boldsymbol{Y}$. Its performance and, more generally, the behavior of the eigenvalues and eigenvectors of (1) has been studied in exquisite detail in statistics [49, 68] and random matrix theory [7, 8, 21, 22, 26, 40, 52, 6]. Going beyond the spectral estimator given by the principal eigenvector, approximate message

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[*]Equal contribution

passing (AMP) constitutes a popular family of iterative algorithms. The reason for this popularity lies in two especially attractive features: *(i)* AMP algorithms can be tailored to exploit knowledge on the structure of the signal; and *(ii)* under suitable assumptions, their performance in the high-dimensional limit is accurately tracked by a deterministic recursion known as *state evolution* [18, 24, 48]. Using the state evolution machinery, it has been showed that AMP achieves Bayes-optimal performance in some Gaussian models [29, 30, 61, 13], and a bold conjecture from statistical physics is that AMP is optimal in the class of polynomial-time algorithms for a large class of inference problems with random design. AMP algorithms have been recently considered for several practical scenarios, including genetics [86], inpainting [66] and MRI image recovery [76].

A general class of noise models that has received attention in the literature is given by the family of *rotationally invariant matrices*. This is much milder than requiring $\boldsymbol{Z}$ to be Gaussian: it only imposes that the matrix of eigenvectors is uniformly random, and allows for an arbitrary spectrum. Hence, $\boldsymbol{Z}$ can capture the complex correlation structure which often occurs in applications (e.g., recommender systems [20] and bioinformatics [72]). However, estimating noise statistics from data is costly or, for problems involving large-scale datasets (computational genomics is a paradigmatic example [36, 2]), may be impossible. Thus, one natural idea is to simply assume Gaussian statistics for $\boldsymbol{Z}$. The case in which $\boldsymbol{Z}$ is actually Gaussian has been thoroughly studied [53, 28, 55, 54, 12, 61]. Beyond Gaussianity, a rapidly growing literature is focusing on rotationally invariant models assuming *perfect knowledge of the statistics of the structured matrix appearing in the problem* (such as noise in inference, a sensing, data, or coding matrix in regression tasks, weight matrices in neural networks, or a matrix of interactions in spin glass models) [67, 23, 37, 38, 42, 65, 56, 57, 74, 78, 79, 16, 33, 66, 71, 47]. However, despite this impressive progress when the noise statistics is known, low-rank estimation in a *mismatched setting with partial to no knowledge of the statistics of the rotationally invariant noise matrix remains poorly understood*. The goal of this work is thus to shed light into the following fundamental question:

Suppose that the noise statistics is unknown or unreacheable, and hence Gaussian noise is naively assumed. What is the impact of this mismatch on the overall performance of inference methods?

1.1 Summary of contributions

In this paper, we provide rigorous performance guarantees for two inference methods which incorrectly assume Gaussian noise statistics: a Bayes estimator, and an AMP algorithm. Then, by exploiting these sharp analytical characterizations, we describe a number of surprising effects coming from numerical simulations. Our main findings are detailed below.

Theoretical results. *(i)* We give a closed-form expression for the mean-square error (MSE) of the Bayes estimator which samples from the *mismatched* posterior (cf. Theorem 1). Under a certain concentration assumption, we also present an asymptotic result on the overlap of such estimator. *(ii)* We provide a state evolution analysis (cf. Theorem 2) for the *Gaussian* AMP algorithm which is designed for Gaussian noise.

Numerical results. The mismatched Bayes and AMP estimators display surprising behaviors – already for the case of a spherically distributed signal – and the two performance metrics (MSE and overlap) exhibit a remarkably different phenomenology: *(i)* As for the MSE, the Gaussian AMP is *outperformed* by the Bayes estimator. Here, the surprise comes from the fact that AMP algorithms are, in principle, designed to sample from the posterior distribution (and this is often what happens in the matched case). *(ii)* In contrast, when the SNR is large enough, the overlaps of the two mismatched estimators *coincide*, and they even match the overlap of estimators which exploit the noise statistics, namely the optimal spectral method which minimizes the MSE and the correct AMP designed in [37]. *(iii)* Under certain conditions, the MSE of the mismatched Bayes estimator *matches* that of a Gaussian spectral method with no information on the noise structure. *(iv)* The mismatched estimators are *outperformed* – in terms of MSE – by the optimal spectral method and the correct AMP, whose performance coincide. *(v)* Finally, the MSE curves of the Bayes and Gaussian spectral methods exhibit a striking non-monotone behavior.

1.2 Related work

The impact of mismatch. Given the practical relevance of understanding the effect of mismatch in statistical inference, a line of work has approached the issue from various angles. Regression being

probably the most paradigmatic model of inference task, mismatch has been thoroughly studied in this context. In information theory, the trend was initiated in [83, 85]. Studies for M-estimation and robust statistics followed based on statistical mechanics approaches [3], on the analysis of approximate message passing algorithms [25, 77, 32, 43], on Gordon's convex min-max theorem [81, 80] and the leave-one-out method [35, 34]. Concerning the analysis of Bayesian approaches to mismatched linear regression, we refer to [27, 63, 75, 11] or [9] for a review.

In contrast, the problem studied in the present paper, namely low-rank matrix inference with mismatch, has received attention only recently. It is clear that in some way, related questions were analyzed in the aforementioned random matrix theory literature, although from a rather different perspective. Concerning the precise issue of the performance degradation due to mismatch in Bayesian inference, an exception (and inspiration for our work) is the recent paper [71]. However, the authors of [71] focus on mismatch in the signal-to-noise ratio for the Gaussian noise setting; their results can thus be recovered as a special case of the present ones. To the best of our knowledge, this is the first work that considers mismatch in the noise statistics for low-rank matrix estimation.

Approximate message passing. AMP algorithms have been applied to a wide range of inference problems. Examples include estimation in linear models [19, 18, 31], generalized linear models [73, 13, 58, 59], and low-rank matrix recovery with Gaussian noise [17, 29, 41, 51, 55, 61], see also the survey [39]. A general AMP iteration for rotationally invariant matrices has been recently analyzed in [37, 87], and by providing suitable instances of this abstract iteration, AMP algorithms have been developed for low-rank [37, 87, 60] and generalized linear models [82]. Furthermore, an AMP-based method which uses the classical idea of empirical Bayes to reduce the high-dimensional noise in PCA is proposed in [86], which also provides applications to genetics. However, the existing results cannot be applied to the mismatched setting considered in this work, since the Gaussian AMP does not contain the right Onsager corrections. In fact, the algorithm designer assumes Gaussian statistics for the noise and, therefore, constructs an AMP algorithm with the Onsager correction suitable for Gaussian noise. Finally, we remark that the performance of the Gaussian AMP is numerically compared with that of the correct AMP (exploiting the knowledge of noise statistics) in [87]. Our Theorem 2 provides rigorous foundations for such a comparison.

2 Setup of the problem

2.1 Random matrix theory preliminaries

We start with some useful notions of random matrix theory. Given a probability measure ρ of compact support $K \subseteq \mathbb{R}$, we let $H_\rho : \mathbb{R}\backslash K \mapsto \mathbb{R}$ be the *Hilbert transform* of ρ: $H_\rho(z) := \int_K \rho(d\gamma)(z-\gamma)^{-1}$. We define $\overline{\gamma} := \max K$, $\underline{\gamma} := \min K$, $\overline{h} := \lim_{z\downarrow\overline{\gamma}} H_\rho(z)$ and $\underline{h} := \lim_{z\uparrow\underline{\gamma}} H_\rho(z)$, where the limits exist due to the monotonicity of H_ρ but may be infinite. As H_ρ is a bijection between $\mathbb{R}\backslash K$ and its image $(\underline{h}, \overline{h})\backslash\{0\}$, its inverse exists and it is denoted by $K_\rho : (\underline{h}, \overline{h})\backslash\{0\} \mapsto \mathbb{R}\backslash K$. The *R-transform* of ρ is $R_\rho : (\underline{h}, \overline{h})\backslash\{0\} \mapsto \mathbb{R}\backslash K$ given by $R_\rho(x) := K_\rho(x) - 1/x$. The coefficients $\{\bar{\kappa}_k\}_{k\geq 1}$ of the Taylor series of R_ρ are the *free cumulants* associated to ρ, i.e., $R_\rho(x) = \sum_{k=0}^{\infty} \bar{\kappa}_{k+1} x^k$, and they can be computed from the moments of ρ, see e.g. Section 2.5 of [64]. Furthermore, $R'_\rho(x)$ and $H'_\rho(z)$ denote the derivatives of the R-transform and the Hilbert transform of ρ, respectively.

2.2 Model of mismatched low-rank matrix estimation

We consider the problem of estimating a rank-one informative *spike* $\boldsymbol{X}\boldsymbol{X}^\top$ corrupted by an additive symmetric noise matrix $\boldsymbol{Z} \in \mathbb{R}^{N\times N}$ from data $\boldsymbol{Y}$ generated as in (1). The scalar $\lambda_* \in \mathbb{R}_{\geq 0}$ is the signal-to-noise ratio (SNR), and the noise $\boldsymbol{Z}$ can be decomposed as $\boldsymbol{Z} = \boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^\top$, where $\boldsymbol{\Sigma} := \mathrm{diag}(\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^{N\times N}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{Z}$ and $\boldsymbol{O}$ is some orthogonal matrix. We also denote by $\overline{\gamma}_N$ and $\underline{\gamma}_N$ respectively $\max\{\gamma_1, \ldots, \gamma_N\}$ and $\min\{\gamma_1, \ldots, \gamma_N\}$. Similarly, $\overline{\nu}_N$ and $\underline{\nu}_N$ denote the largest and smallest eigenvalues of the data matrix $\boldsymbol{Y}$. Throughout the paper, the main technical assumption will be the following.

Assumption 1. *The signal $\boldsymbol{X}$ has norm $\sqrt{N}$. The random noise matrix $\boldsymbol{Z}$ is rotationally invariant and is independent of $\boldsymbol{X}$. Moreover, the empirical measure $N^{-1}\sum_{i\leq N}\delta_{\gamma_i}$ of eigenvalues of $\boldsymbol{Z}$ converges weakly towards a limiting measure ρ of compact support $K \subseteq \mathbb{R}$. Finally, $\overline{\gamma}_N$ and $\underline{\gamma}_N$ converge a.s. to $\overline{\gamma} = \max K$ and $\underline{\gamma} = \min K$, respectively.*

The noise matrix $\boldsymbol{Z}$ being rotationally invariant means that $\boldsymbol{O}$ is a random orthogonal matrix (i.e., sampled from the Haar measure) independent of $\boldsymbol{\Sigma}$. Under Assumption 1, we have that $\overline{\nu}_N$ and $\underline{\nu}_N$ converge a.s. to finite limits, which we denote by $\overline{\nu}$ and $\underline{\nu}$, respectively. These have an explicit form in terms of ρ and λ_* given in [21, Theorem 2.1] and which we reproduce for convenience in Appendix A. We remark that the assumption on the ground-truth signal is rather mild: $\boldsymbol{X}$ can have any distribution over the sphere of radius $\sqrt{N}$, and it might even be deterministic. For example, $\boldsymbol{X} \in \{-1, 1\}^N$ could be a uniform vector over the centred N-dimensional hyper-cube of sides $2\sqrt{N}$.

For the analysis of the Bayes estimator, we require a second assumption on the asymptotic eigenvalue distribution ρ. This ensures that the limit for the overlap of the spectral estimators with the signal is continuous in the SNR, whenever a small additive Wigner noise of variance $\epsilon \geq 0$ is added to $\boldsymbol{Y}$.

Assumption 2. *Let ρ_ϵ be the spectral density of the free convolution of ρ and a semicircle law of radius $2\epsilon \geq 0$. Let H_{ρ_ϵ} be the associated Hilbert transform and $\overline{\gamma}_\epsilon$ the rightmost point of the support of ρ_ϵ. Then, we assume that $\lim_{z \downarrow \overline{\gamma}_\epsilon} H'_{\rho_\epsilon}(z) = -\infty$ for all $\epsilon \geq 0$.*

For the definition of the free convolution and its link to random matrix theory, we refer the reader to [4, 70]. We remark that Assumption 2 is satisfied by a wide class of random matrices. In particular, by combining Theorem 2.2 of [10] with Proposition 2.4 of [21], it suffices that the support of the limiting spectral measure ρ is *(i)* compact, *(ii)* connected, and *(iii)* it has a proper decay rate at its edges (i.e., ρ is of Jacobi type), see Assumption 2.1 of [10]. These conditions can be easily checked for many random matrix models, including the ones discussed in the examples of Section 4.

Note that, if ρ is the semicircle law of radius 2, the noise is asymptotically equal in distribution to a standard Wigner matrix $\boldsymbol{W}$ with density $\sim \exp(-\frac{N}{4}\mathrm{Tr}\boldsymbol{W}^2)$. As the elements of this type of symmetric matrices are i.i.d. Gaussian, in this case we say that there is *Gaussian noise*. For other limiting ρ, the elements of $\boldsymbol{Z} \neq \boldsymbol{W}$ remain correlated and we say that there is *structured noise*. We denote by $\boldsymbol{W}$ a sequence of standard Wigner matrices while we reserve $\boldsymbol{Z}$ for a generic sequence of rotationally invariant matrices.

Sources of mismatch. In this paper, we consider what happens when there is a *mismatch* between the true noise statistics and the assumptions on the noise statistics made in the inference algorithm. In particular, we study the case in which the noise is assumed to be Gaussian. Gaussianity is in fact the most standard assumption made when precise knowledge of the noise structure is lacking. We also consider the case in which the SNR is estimated incorrectly, i.e., the statistician assumes that the data is generated according to $(\sqrt{\lambda}/N)\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{W}$, where $\lambda \neq \lambda_*$ and λ_* is the correct SNR present in (1). Our goal is to quantify how these sources of ignorance affect the algorithmic performance.

2.3 Three classes of inference procedures for PCA

Mismatched Bayes estimator. The statistician assumes Gaussian noise and that the signal $\boldsymbol{X}$ has no specific structure, i.e., it is uniformly distributed on the N-dimensional sphere $\mathbb{S}^{N-1}(\sqrt{N})$. The mismatched posterior distribution reads, using the Gaussian log-likelihood $-\frac{N}{4}\mathrm{Tr}(\boldsymbol{Y} - \frac{\sqrt{\lambda}}{N}\boldsymbol{x}\boldsymbol{x}^\top)^2$,

$$P_{\mathrm{mis}}(d\boldsymbol{x} \mid \boldsymbol{Y}) = \frac{1}{Z_N(\boldsymbol{Y})} \exp\Big(\frac{\sqrt{\lambda}}{2}\langle \boldsymbol{x}, \boldsymbol{Y}\boldsymbol{x}\rangle\Big)\, \mu_N(d\boldsymbol{x}), \tag{2}$$

where $Z_N(\boldsymbol{Y})$ is the normalization constant and μ_N is the uniform measure over $\mathbb{S}^{N-1}(\sqrt{N})$. In case the signal $\boldsymbol{X}$ lies on the sphere (by Assumption 1) but is not uniformly distributed on it, then this leads to a third source of mismatch. However, the MSE formula of Theorem 1 does *not* depend on it; e.g. if $\boldsymbol{X} \in \{-1, 1\}^N$ the error remains the same. The origin of this is the uniformly spherical prior used in the posterior (2) which is the most uninformative one.

The Bayes estimators we analyze are obtained as the posterior means

$$M_{\mathrm{mis}}(\boldsymbol{Y}) := \int \boldsymbol{x}\boldsymbol{x}^\top P_{\mathrm{mis}}(d\boldsymbol{x} \mid \boldsymbol{Y}), \tag{3}$$

which may not be practical to compute. Notice that, if the noise is Wigner, then the likelihood $\exp(\cdots)$ in (2) is correct. This Bayes-optimal case has already been rigorously characterized (see, for example, [12, 14, 54]), and it will serve us as a base case for comparison with our mismatched setting.

The performance of a sequence of estimators $M_N(\boldsymbol{Y}) \in \mathbb{R}^{N\times N}$ of the spike $\boldsymbol{X}\boldsymbol{X}^\top$ is quantified via the asymptotic mean-square error (MSE):

$$\mathrm{MSE}(M) := \lim_{N\to+\infty} \frac{1}{2N^2}\mathbb{E}\|\boldsymbol{X}\boldsymbol{X}^\top - M_N(\boldsymbol{Y})\|_{\mathrm{F}}^2, \tag{4}$$

where $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm and $\mathbb{E}$ the expectation over the spike $\boldsymbol{X}\boldsymbol{X}^\top$ and the noise $\boldsymbol{Z}$. We also consider another performance measure which is insensitive to the norm of an estimator $\hat{\boldsymbol{x}}$ of $\boldsymbol{X}$, namely, the rescaled overlap:

$$\mathrm{Overlap}(\hat{\boldsymbol{x}}) := \lim_{N\to+\infty} \frac{|\langle \boldsymbol{X}, \hat{\boldsymbol{x}}(\boldsymbol{Y})\rangle|}{\|\hat{\boldsymbol{x}}(\boldsymbol{Y})\| \cdot \|\boldsymbol{X}\|}, \tag{5}$$

where $\|\boldsymbol{X}\| = \sqrt{N}$ due to the spherical constraint. This alignment measure is sometimes also referred to as *cosine distance*. For both AMP and spectral estimators, this overlap will almost surely be deterministic. If the norm of the estimator vanishes, by convention we fix the overlap to zero. Note that the definition (5) is not meaningful for the Bayes case. In fact, if one defines $\hat{\boldsymbol{x}}_{\mathrm{mis}} := \int \boldsymbol{x} P_{\mathrm{mis}}(d\boldsymbol{x} \mid \boldsymbol{Y})$, then this quantity is zero by sign symmetry ($\pm\boldsymbol{x}$ have the same posterior weight). Thus, the correct definition of the overlap for the mismatched Bayes estimator is

$$\mathrm{Overlap}_{\mathrm{mis}} := \lim_{N\to+\infty} \Big(\frac{1}{N}\frac{1}{\|M_{\mathrm{mis}}(\boldsymbol{Y})\|_{\mathrm{F}}} \int P_{\mathrm{mis}}(d\boldsymbol{x} \mid \boldsymbol{Y})\langle \boldsymbol{X}, \boldsymbol{x}\rangle^2\Big)^{1/2}, \tag{6}$$

while for the "one-shot" AMP and spectral estimators we use the former definition (5).

Approximate message passing. We remark that, as opposed to our analysis of the Bayes estimator, the state evolution characterization of AMP (Theorem 2) does not require $\boldsymbol{X}$ to be uniformly distributed on $\mathbb{S}^{N-1}(\sqrt{N})$. The AMP analysis can accomodate more generic (potentially mismatched) prior distributions and, in fact, AMP algorithms are well equipped to exploit structure in the signal. For simplicity, we assume to have access to an initialization $\hat{\boldsymbol{x}}^1 \in \mathbb{R}^N$, which is independent of the noise $\boldsymbol{Z}$ and has a strictly positive correlation with $\boldsymbol{X}$, i.e.,

$$(\boldsymbol{X}, \hat{\boldsymbol{x}}^1) \xrightarrow{W_2} (X, \hat{x}_1), \quad \mathbb{E}[X\,\hat{x}_1] := \epsilon > 0, \quad \mathbb{E}[\hat{x}_1^2] \le 1, \tag{7}$$

where $(\boldsymbol{X}, \hat{\boldsymbol{x}}^1) \xrightarrow{W_2} (X, \hat{x}_1)$ denotes convergence of the joint empirical distribution of $(\boldsymbol{X}, \hat{\boldsymbol{x}}^1)$ to the random variable $(X, \hat{x}_1)$ in Wasserstein-2 (W_2) distance. Then, for $t \ge 1$ the AMP iteration reads

$$\boldsymbol{x}^t = \boldsymbol{Y}\hat{\boldsymbol{x}}^t - \beta_t \hat{\boldsymbol{x}}^{t-1}, \quad \hat{\boldsymbol{x}}^{t+1} = h_{t+1}(\boldsymbol{x}^t), \tag{8}$$

where we assume that $\hat{\boldsymbol{x}}^0 = \boldsymbol{0}$. Here, the function $h_{t+1} : \mathbb{R} \to \mathbb{R}$ is applied component-wise, and it can be chosen in order to exploit prior information about the signal; $\beta_1 = 0$ and, for $t \ge 2$, $\beta_t = \langle h_t'(\boldsymbol{x}^{t-1})\rangle$, where h_t' denotes the derivative of h_t. The AMP estimator of $\boldsymbol{X}$ is $\hat{\boldsymbol{x}}^t$, and the one of the spike $\boldsymbol{X}\boldsymbol{X}^\top$ is $M_{\mathrm{AMP}}^t = \hat{\boldsymbol{x}}^t(\hat{\boldsymbol{x}}^t)^\top$. We refer to this algorithm as *Gaussian AMP*, since this is the AMP that is implemented for Gaussian noise (and known SNR). For a discussion on how (8) can be derived, we refer the interested reader to the review [39]. We note that the initialization (7) is impractical and one can design AMP iterations which are initialized with the eigenvector $\overline{\boldsymbol{v}}_N$ of the data matrix $\boldsymbol{Y}$ associated to the largest eigenvalue $\overline{\nu}_N$, see [61, 60, 87]. We are also able to provide a state evolution result for the Gaussian AMP with spectral initialization, and this is discussed in Appendix C.4.

Spectral estimators. Finally, we consider spectral estimators of the form $C\,\overline{\boldsymbol{v}}_N\overline{\boldsymbol{v}}_N^\top$, where $\overline{\boldsymbol{v}}_N$ is the unit-norm eigenvector associated to the largest eigenvalue $\overline{\nu}_N(\boldsymbol{Y})$ and $C > 0$ is a scaling constant taking into account the spectrum of the data matrix. The asymptotic description of $\overline{\boldsymbol{v}}_N$ and $\overline{\nu}_N$ for additive and multiplicative low-rank perturbations of rotationally invariant matrices has been obtained in [21, 22, 6]. In particular, by Theorem 2.2 of [22], the squared overlap $\langle \boldsymbol{X}, \overline{\boldsymbol{v}}_N\rangle^2/N$ converges a.s. to a constant $C(\rho, \lambda_*)$, whose explicit expression is recalled in Appendix A. We study two variants:

- The *optimal spectral estimator (OptSpec)* is given by $M_{\mathrm{OS}} := C_{\mathrm{OS}}\overline{\boldsymbol{v}}_N\overline{\boldsymbol{v}}_N^\top$, where $C_{\mathrm{OS}}(\rho, \lambda_*) := C(\rho, \lambda_*)$ is the optimal scaling that minimizes the MSE.
- The *Gaussian mismatched spectral estimator (GauSpec)* is given by $M_{\mathrm{GS}} := C_{\mathrm{GS}}\overline{\boldsymbol{v}}_N\overline{\boldsymbol{v}}_N^\top$, where C_{GS} is the optimal scaling if the noise was Gaussian and the SNR was equal to $\lambda > 0$. By letting ρ_{SC} be the standard semi-circle law, we have $C_{\mathrm{GS}}(\lambda) := C(\rho_{\mathrm{SC}}, \lambda) = 1 - 1/\lambda$.

For both OptSpec and GauSpec, the estimator of $\boldsymbol{X}$ is given by $\hat{\boldsymbol{x}}_{\mathrm{GS}} = \hat{\boldsymbol{x}}_{\mathrm{OS}} = \hat{\boldsymbol{x}}_{\mathrm{Spec}} := \sqrt{N}\,\overline{\boldsymbol{v}}_N$. Notice that, while OptSpec requires knowing exactly the statistics of the noise and the SNR, GauSpec

represents the spectral estimator that would be used by a statistician who assumes the noise to be Gaussian and the SNR λ (instead of λ_*). An application of Theorem 2.2 in [22] gives that

$$\mathrm{MSE}(M_{\mathrm{OS}}) = \frac{1}{2}(1 - C^2(\rho, \lambda_*)), \quad \text{and} \quad \mathrm{MSE}(M_{\mathrm{GS}}) = \frac{1}{2}(1 + (1 - 1/\lambda)^2 - 2(1 - 1/\lambda)C(\rho, \lambda_*)).$$

Furthermore, the respective overlaps coincide and are given by $\sqrt{C_{\mathrm{OS}}}$. Note that $\mathrm{MSE}(M_{\mathrm{GS}}) = \mathrm{MSE}(M_{\mathrm{OS}}) + \frac{1}{2}(C(\rho, \lambda_*) - (1 - 1/\lambda))^2$. Thus, as expected, we have $\mathrm{MSE}(M_{\mathrm{OS}}) \leq \mathrm{MSE}(M_{\mathrm{GS}})$. We also remark that, as the noise matrix is rotationally invariant, both the MSE and the overlap of the spectral estimators do not depend on the ground-truth signal distribution.

3 Main results: performance of inference algorithms

3.1 Mismatched Bayes estimator

Consider the following functions of $\lambda, \lambda_* > 0$ and of the asymptotic spectral noise density ρ:

$$\begin{aligned} M(\lambda, \lambda_*) &:= \Big(1 - \frac{1}{\sqrt{\lambda\lambda_*}}\Big)\Big(1 - \frac{1}{\lambda_*} R'_\rho\Big(\frac{1}{\sqrt{\lambda_*}}\Big)\Big)\mathbf{1}\big(\overline{h}\sqrt{\lambda_*} \geq 1 \cap \lambda\lambda_* > 1\big), \\ Q(\lambda, \lambda_*) &:= \Big(1 - \frac{1}{\sqrt{\lambda\lambda_*}}\Big)^2 \mathbf{1}\big(\overline{h}\sqrt{\lambda_*} \geq 1 \cap \lambda\lambda_* > 1\big) + \Big(1 - \frac{\overline{h}}{\sqrt{\lambda}}\Big)^2 \mathbf{1}\big(\overline{h}\sqrt{\lambda_*} < 1 \cap \sqrt{\lambda} > \overline{h}\big), \end{aligned} \tag{9}$$

where $\mathbf{1}(E)$ stands for the indicator function of the event E. The idea is that $Q(\lambda, \lambda_*)$ and $M(\lambda, \lambda_*)$ represent the squared norm of the mismatched Bayes estimator and its inner product with the signal, respectively. This leads to our main result on the mismatched Bayes MSE, which is stated below.

Theorem 1 (Performance of mismatched Bayes estimator)**.** *Consider a spiked model* (1)*. Let Assumptions 1 and 2 hold. Then, the MSE of the mismatched Bayes estimator* (3) *is given by*

$$\mathrm{MSE}(M_{\mathrm{mis}}) = \frac{1}{2}\big(1 + Q(\lambda, \lambda_*) - 2M(\lambda, \lambda_*)\big). \tag{10}$$

As a consequence of the auxiliary lemmas used in the proof of Theorem 1 and presented in Appendix B, we have that $N^{-2}\int P_{\mathrm{mis}}(d\boldsymbol{x} \mid \boldsymbol{Y})\langle \boldsymbol{X}, \boldsymbol{x}\rangle^2$ converges almost surely to $M(\lambda, \lambda_*)$. Furthermore, for noise matrices regularized by arbitrarily small Wigner matrices, by using similar techniques as in [15] we can prove that $\|M_{\mathrm{mis}}\|_{\mathrm{F}}/N$ converges almost surely to $Q^{1/2}(\lambda, \lambda_*)$. We conjecture that this convergence holds even without the Wigner regularization. By assuming this conjecture, we have the following almost sure convergence result for the overlap:

$$\mathrm{Overlap}_{\mathrm{mis}} = \frac{M^{1/2}(\lambda, \lambda_*)}{Q^{1/4}(\lambda, \lambda_*)}, \tag{11}$$

which holds whenever the denominator is non-zero. When comparing the overlaps of different methods in Section 4, we will use (11) for the mismatched Bayes estimator.

Let us highlight the following property of the MSE (10) when there is no mismatch in the SNR:

$$\text{If } \lambda = \lambda_* \text{ and } \overline{h} \geq 1, \text{ then } \mathrm{MSE}(M_{\mathrm{mis}}) = \mathrm{MSE}(M_{\mathrm{GS}}). \tag{12}$$

However, in general $\mathrm{MSE}(M_{\mathrm{mis}})$ and $\mathrm{MSE}(M_{\mathrm{GS}})$ differ. E.g., when the assumed SNR λ is different from the real one λ_*, the norm of M_{mis} incorporates information about λ_* while the one for M_{GS} only depends on λ. Finally, we note that, from a statistical view-point, Assumption 1 can be replaced by the following: $\boldsymbol{X}$ is uniformly distributed on $\mathbb{S}^{N-1}(\sqrt{N})$ and is independent of $\boldsymbol{Z}$, which may be a generic symmetric matrix with converging empirical spectral density.

Sketch of the proof of Theorem 1. The goal is to evaluate the asymptotic values of $\mathbb{E}\|M_{\mathrm{mis}}\|_{\mathrm{F}}^2$ and $\mathrm{tr}\mathbb{E}[M_{\mathrm{mis}}\boldsymbol{X}\boldsymbol{X}^\top]$, from which both the MSE and the mean overlap of the Bayes estimator can be obtained. Our strategy is to first compute the log-moment generating function f_{mis} of the mismatched Bayes model. Then, the above quantities of interest can be accessed by taking derivatives of f_{mis} with respect to appropriate parameters. However, as the noise in the inference model is not Gaussian, $\mathbb{E}\|M_{\mathrm{mis}}\|_{\mathrm{F}}^2$ cannot be computed in the standard way using an "I-MMSE" type of formula [46] (like it is done, e.g., in [71]). To address this issue, we introduce a generalized model for which the noise matrix is given by the original one plus an independent Wigner matrix multiplied by a small parameter ϵ. We then compute the log-moment generating function of this generalized model using the theory of low-rank perturbations of rotationally invariant random matrices [21] and of the low-rank spherical integral [45]. Finally, the desired results are obtained by a limiting argument in $\epsilon \to 0$.

3.2 Approximate message passing

Our main result is a characterization of the iterates (8) in the high-dimensional limit. We show that the joint empirical distribution of $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t, \hat{\boldsymbol{x}}^1, \dots, \hat{\boldsymbol{x}}^{t+1})$ converges (in W_2) to the random vector $(x_1, \dots, x_t, \hat{x}_1, \dots, \hat{x}_{t+1})$. The law of the random vector $(x_1, \dots, x_t, \hat{x}_1, \dots, \hat{x}_{t+1})$ is captured by a deterministic *state evolution (SE)* recursion, which can be expressed via a sequence of vectors $\boldsymbol{\mu}_t = (\mu_1, \dots, \mu_t)$ and matrices $\boldsymbol{\Sigma}_t, \boldsymbol{\Delta}_t, \boldsymbol{B}_t \in \mathbb{R}^{t \times t}$ defined as follows. We start with the initialization

$$\mu_1 = \sqrt{\lambda_*}\epsilon, \quad \boldsymbol{\Sigma}_1 = \bar{\kappa}_2 \mathbb{E}[\hat{x}_1^2], \quad \boldsymbol{\Delta}_1 = \mathbb{E}[\hat{x}_1^2], \quad \boldsymbol{B}_1 = \bar{\kappa}_1, \tag{13}$$

where λ_* is the true SNR (see (1)), and $\{\bar{\kappa}_k\}_{k \geq 1}$ are the free cumulants associated to the asymptotic spectral measure of the noise ρ. For $t \geq 1$, given $\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\Delta}_t, \boldsymbol{B}_t$, let

$$\begin{cases} (z_1, \dots, z_t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t) \text{ and independent of } (X, \hat{x}_1), \\ x_t = z_t + \mu_t X - \bar{\beta}_t \hat{x}_{t-1} + \sum_{i=1}^t (\boldsymbol{B}_t)_{t,i} \hat{x}_i, \quad \hat{x}_{t+1} = h_{t+1}(x_t), \end{cases} \tag{14}$$

where we have defined $\hat{x}_0 = 0$, $\bar{\beta}_1 = 0$, and for $t \geq 2$, $\bar{\beta}_t = \mathbb{E}[h_t'(x_{t-1})]$. Then, the parameter μ_{t+1} is

$$\mu_{t+1} = \sqrt{\lambda_*} \mathbb{E}[X\, \hat{x}_{t+1}]. \tag{15}$$

Next, we compute the matrices $\boldsymbol{\Delta}_{t+1}$ and $\boldsymbol{B}_{t+1}$ as

$$\begin{cases} (\boldsymbol{\Delta}_{t+1})_{i,j} = \mathbb{E}[\hat{x}_i\, \hat{x}_j], \quad 1 \leq i, j \leq t+1, \\ \boldsymbol{B}_{t+1} = \sum_{j=0}^t \bar{\kappa}_{j+1} \bar{\boldsymbol{\Phi}}_{t+1}^j \text{ where } (\bar{\boldsymbol{\Phi}}_{t+1})_{i,j} = 0 \text{ if } i \leq j, \text{ and } (\bar{\boldsymbol{\Phi}}_{t+1})_{i,j} = \mathbb{E}[\partial_j \hat{x}_i] \text{ if } i > j, \end{cases} \tag{16}$$

where $\partial_j \hat{x}_i$ denotes the partial derivative $\partial_{z_j} h_i(z_{i-1} + \mu_{i-1} X - \bar{\beta}_{i-1} \hat{x}_{i-2} + \sum_{k=1}^{i-1} (\boldsymbol{B}_t)_{i-1,k} \hat{x}_k))$. Finally, we define the covariance matrix $\boldsymbol{\Sigma}_{t+1}$ as

$$\boldsymbol{\Sigma}_{t+1} = \sum_{j=0}^{2t} \bar{\kappa}_{j+2} \boldsymbol{\Theta}_{t+1}^{(j)}, \quad \text{with} \quad \boldsymbol{\Theta}_{t+1}^{(j)} = \sum_{i=0}^{j} (\bar{\boldsymbol{\Phi}}_{t+1})^i \boldsymbol{\Delta}_{t+1} (\bar{\boldsymbol{\Phi}}_{t+1}^\top)^{j-i}. \tag{17}$$

Note that the $t \times t$ top left sub-matrices of $\boldsymbol{\Sigma}_{t+1}, \boldsymbol{\Delta}_{t+1}$ and $\boldsymbol{B}_{t+1}$ are given by $\boldsymbol{\Sigma}_t, \boldsymbol{\Delta}_t$ and $\boldsymbol{B}_t$. At this point, we are ready to state our state evolution characterization of the AMP algorithm (8). The result is presented in terms of *pseudo-Lipschitz* test functions. A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order 2, i.e., $\psi \in \mathrm{PL}(2)$, if there is a constant $C > 0$ such that $\|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})\| \leq C(1 + \|\boldsymbol{x}\| + \|\boldsymbol{y}\|)\|\boldsymbol{x} - \boldsymbol{y}\|$. We also make the following assumption on the functions $\{h_{t+1}\}_{t \geq 1}$.

Assumption 3. *The function $h_{t+1} : \mathbb{R} \to \mathbb{R}$ is Lipschitz, and the partial derivatives*

$$\partial_{z_k} h_{t+1}\Big(z_t + \mu_t X - \bar{\beta}_t \hat{x}_{t-1} + \sum_{k=1}^t (\boldsymbol{B}_t)_{t,k} \hat{x}_k\Big)$$

are continuous on a set of probability 1*, under the laws of $(z_1 \dots, z_t)$ and $(\hat{x}_1, \dots, \hat{x}_t)$ given in* (14)*.*

This requirement covers most practically relevant choices of h_{t+1} (e.g., soft-thresholding or ReLU), and it is rather standard in the AMP literature, see e.g. [39, 60, 82, 87].

Theorem 2 (State evolution of Gaussian AMP)**.** *Consider a spiked model* (1) *and the AMP algorithm* (8) *initialized as in* (7)*. Let Assumptions 1 and 3 hold, and let $\psi : \mathbb{R}^{2t+2} \to \mathbb{R}$ be any pseudo-Lipschitz functions of order* 2*. Then, for each $t \geq 1$, we almost surely have that, as $N \to +\infty$,*

$$\frac{1}{N} \sum_{i=1}^N \psi((\boldsymbol{x}^1)_i, \dots, (\boldsymbol{x}^t)_i, (\hat{\boldsymbol{x}}^1)_i, \dots, (\hat{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i) \to \mathbb{E}\psi(x_1, \dots, x_t, \hat{x}_1, \dots, \hat{x}_{t+1}, X), \tag{18}$$

where the random variables on the right are defined in (14)*.*

By using Definition 6.7 and Theorem 6.8 of [84], one readily obtains that (18) is equivalent to the convergence of $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t, \hat{\boldsymbol{x}}^1, \dots, \hat{\boldsymbol{x}}^{t+1}, \boldsymbol{X})$ to $(x_1, \dots, x_t, \hat{x}_1, \dots, \hat{x}_{t+1}, X)$ in W_2 distance. Applying (18) to the pseudo-Lipschitz functions $\psi(\hat{x}_{t+1}, X) = (\hat{x}_{t+1} - X)^2$, $\psi(\hat{x}_{t+1}, X) = \hat{x}_{t+1} \cdot X$ and $\psi(\hat{x}_{t+1}, X) = (\hat{x}_{t+1})^2$, we obtain a high-dimensional characterization of the AMP performance.

Corollary 1. *Consider the setting of Theorem 2. Then, for each $t \geq 1$, we almost surely have that*

$$\mathrm{MSE}(M_{\mathrm{AMP}}^t) := \lim_{N \to +\infty} \frac{1}{2N^2} \mathbb{E}\|\boldsymbol{X}\boldsymbol{X}^\top - \hat{\boldsymbol{x}}^t (\hat{\boldsymbol{x}}^t)^\top\|_{\mathrm{F}}^2 = \frac{1}{2}\Big(1 - 2\,(\mathbb{E}[\hat{x}_t \cdot X])^2 + (\mathbb{E}[(\hat{x}_t)^2])^2\Big),$$

$$\mathrm{Overlap}(\hat{\boldsymbol{x}}^t) := \lim_{N \to +\infty} \frac{|\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^t \rangle|}{\|\hat{\boldsymbol{x}}^t\| \cdot \|\boldsymbol{X}\|} = \frac{|\mathbb{E}[\hat{x}_t \cdot X]|}{\sqrt{\mathbb{E}[(\hat{x}_t)^2]}}.$$

Sketch of the proof of Theorem 2. The key insight is to exploit the flexibility given by the denoisers of the abstract AMP iteration [87] to "correct" the mismatch. These denoisers are denoted by $\{\tilde{h}_t\}_{t\geq 2}$ (see (33) and (34) in Appendix C.1) and they serve solely as a proof technique, hence they do not have an impact on the practicality of the algorithm; in contrast, the denoisers $\{h_t\}_{t\geq 2}$ (see (8)) can be chosen by the algorithm designer to take advantage of the information available about the signal. Thus, we can design an *auxiliary AMP* such that *(i)* it admits a state evolution characterization, and *(ii)* its iterates are close to (8). This requires the delicate induction argument in Appendix C: in Appendix C.1, we define the auxiliary AMP; in Appendix C.2, we prove a state evolution result for it; in Appendix C.3, we show by induction that *(i)* the state evolution parameters of the auxiliary AMP coincide with those defined in this section, and *(ii)* the iterates of the auxiliary AMP are close (in ℓ_2 norm) to the iterates of the AMP (8), thus concluding the proof of Theorem 2.

4 Numerical results and discussion

In all experiments, the density ρ of the eigenvalues of the noise has unit variance and the distribution of the ground-truth signal is uniform on the sphere of radius $\sqrt{N}$. For the sake of simplicity, we decouple the effect of the mismatch in *(i)* the noise statistics and *(ii)* the SNR. More specifically, in the first two examples, we set $\lambda = \lambda_*$, so that only mismatch in the noise statistics is present; and in the last example, the noise is Gaussian, so that there is only SNR mismatch. Additional results when $\boldsymbol{Z}$ is the free convolution of Rademacher and semicircle spectra are reported in Appendix D.3.

Models of mismatch. *(a) Marcenko-Pastur spectrum.* An example of non-symmetric density is the Marcenko-Pastur law $\rho(x) = \sqrt{x(4-x)}/(2\pi x)\,\mathbf{1}(x \in [0,4])$. The R-transform is $R_\rho(x) = 1/(1-x)$, and the results are displayed in Fig. 1a (the same formulas apply if the law is centered).

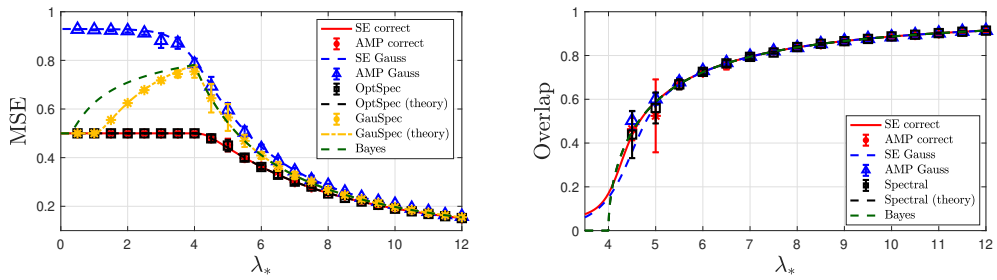
(b) Uniform spectrum. We let ρ be the uniform distribution $\mathcal{U}[-\sqrt{3},\sqrt{3}]$. In this case, we have $R_\rho(x) = \sqrt{3}/\tanh(\sqrt{3}x) - 1/x$, and the results are in Fig. 1b.

(c) Wigner matrix/semicircle spectrum with mismatched SNR. We consider a Gaussian noise matrix $\boldsymbol{W}$ (as assumed by the statistician), but with mismatched SNR by setting $\lambda = 4\lambda_*$, see Fig. 1c.

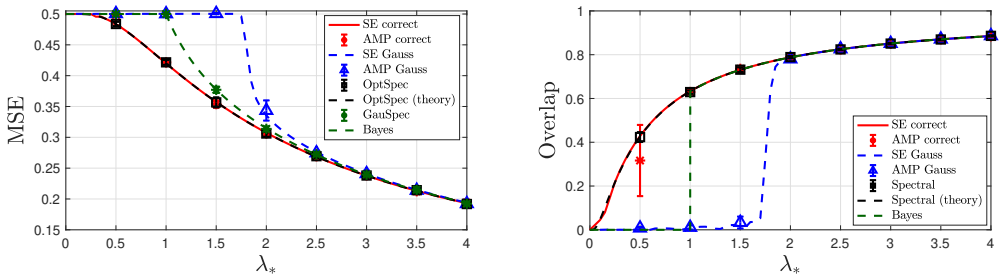
We remark that Assumption 1 is verified for all models. Furthermore, in all the three cases, the support of ρ is a single non-empty interval. Hence, Theorem 2.2 of [10] and Proposition 2.4 of [21] can be used to show that Assumption 2 holds.

Inference algorithms and set-up. The estimators of the spike $\boldsymbol{X}\boldsymbol{X}^\intercal$ are compared in terms of the MSE (4) (plots on the left). The AMP and spectral estimators of $\boldsymbol{X}$ are compared in terms of the overlap (5), and the Bayes estimator of $\boldsymbol{X}$ in terms of the overlap (6) (plots on the right). We evaluate those formulas at $N = 8000$ for the various algorithms, and in the $N \to +\infty$ limit for the theoretical predictions. The *correct AMP* (together with its own SE) is in red. This AMP is *correct* in the sense that the statistician is aware of the noise statistics and, thus, incorporates the right Onsager corrections. The form of the correct AMP and the corresponding state evolution are readily obtained from the results in [37, 87] and, for the reader's convenience, they are recalled in Appendix D.1. As non-linearities, we use the posterior-mean denoising functions of Section 5.1 in [87], and we estimate the SE parameters consistently from data (see Appendix D.2). The *Gaussian AMP* (together with its own SE) is in blue. This is the AMP algorithm (8), which is chosen when the noise statistics or the SNR are unknown (as it would be optimal for Gaussian noise and known SNR). Its state evolution is given by Theorem 2. As non-linearities $\{h_{t+1}\}_{t\geq 1}$, we use again the posterior-mean denoising functions, and estimate consistently the state evolution parameters from data (see Appendix D.2). We note that the denoisers of the Gaussian AMP depend only on a single iterate, while the denoisers of the correct AMP incorporate all the past iterates. In contrast, the state evolution parameters of the Gaussian AMP at time t depend on *all* the past, in order to correct for the mismatch. The *mismatched Bayes estimator* is plotted in green. Its MSE and overlap are given by (10) and (11), respectively. The *optimal spectral estimator (OptSpec)* is plotted in black, and the *Gaussian mismatched spectral estimator (GauSpec)* is plotted in yellow in Fig. 1a and 1c (where it differs from Bayes) or in green in Fig. 1b (where it coincides with Bayes). The performance measures of both variants of spectral estimators are given at the end of Section 2.3. Each experiment is repeated for 10 independent runs. We report the average and error bars at 1 standard deviation.

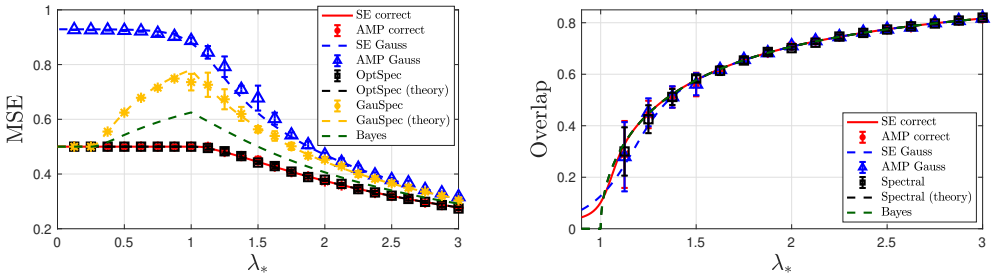
A rich and surprising phenomenology. *(i)* An intriguing phenomenon is that, in terms of MSE, *the Gaussian AMP does not perform as well as the mismatched Bayes estimator*. This effect occurs



(a) Marcenko-Pastur noise.



(b) Uniform eigenvalues distribution in $[-\sqrt{3}, \sqrt{3}]$.



(c) Wigner matrix/semicircle spectrum with mismatched SNR $\lambda = 4\lambda_*$.

Figure 1: MSE (on the left) and overlap (on the right), as a function of the true SNR λ_* in three mismatched settings.

in all our settings, and it is surprising as the Gaussian AMP (8) is precisely designed to efficiently compute the Bayes estimator (3). In the Bayes-optimal case, it does so (outside of its computational gap) [12]. However, if there is a mismatch, *AMP does not compute the (mismatched) posterior mean*. This finding adds to the already existing evidence [5] that AMP is poorly understood, and studying the fundamental reasons behind this behavior is an exciting avenue for future research.

(ii) By comparing MSE[2] and overlap curves, we understand that the discrepancy between Gaussian AMP and Bayes comes from an incorrect estimation of the signal norm. In fact, *at large enough SNR, the overlaps of Bayes and Gaussian AMP match, and they even coincide with the overlaps of optimal algorithms* (correct AMP and OptSpec), which exploit knowledge about the noise structure. Understanding the origin of the MSE discrepancy (namely, a wrong estimate of the signal length) can potentially lead to procedures which correct for this effect and, hence, reduce the MSE.

(iii) When there is no SNR mismatch and $\bar{h} \geq 1$ (cf. (12)), *the MSEs of the Bayes and Gaussian spectral estimators match* (see Fig. 1b). Both have no information about the noise structure and about the signal distribution. Yet, this equality is remarkable given that the spectral estimator is the solution of an optimization problem (it is a "zero temperature" estimator in physics parlance), while the Bayes

[2]The MSE is constructed from the norm $\mathbb{E}\|M(\boldsymbol{Y})\|_{\mathrm{F}}^2$ and the inner product $\mathrm{tr}\mathbb{E}[M(\boldsymbol{Y})\boldsymbol{X}\boldsymbol{X}^\top]$.

estimator aims at computing the mean of a certain posterior distribution (it is a "finite temperature" estimator).

(iv) All mismatched estimators are outperformed by the optimal spectral method and the correct AMP. Both these algorithms take advantage of the noise statistics and achieve the same MSE. This *suggests that the two estimators are Bayes-optimal*. We remark that solving this conjecture would require understanding the information-theoretic limits of low-rank estimation with structured noise.

(v) Finally, a striking phenomenon – first observed in [71] – is that *the Bayes and Gaussian spectral MSE curves may be non-decreasing with the true SNR* λ_*, see Figs. 1a and 1c. Going beyond the analysis in [71], we remark that, for large enough SNR, all estimators yield the same overlap and therefore all "point in the correct direction" given by the leading eigenvector, see also part *(ii)* of this discussion. This links the initial MSE increase to a wrong estimation of the signal's norm, due to an over-confidence in the data quality (recall that the assumed SNR λ is equal to $4\lambda_*$ in Fig. 1c).

Acknowledgements

M. Mondelli was partially supported by the 2019 Lopez-Loreta Prize. The authors acknowledge discussions with A. Krajenbrink, M. Robinson, A. Depope, N. Macris and F. Pourkamali.

References

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Gad Abraham, Yixuan Qiu, and Michael Inouye. Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 2017.

[3] Madhu Advani and Surya Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

[4] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge university press, 2010.

[5] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. Glassy nature of the hard phase in inference problems. *Phys. Rev. X*, 9, Jan. 2019.

[6] Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

[7] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[8] Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[9] Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. Bayesian inference in high-dimensional models, 2021. `arXiv:2101.04491`.

[10] Zhigang Bao, László Erdős, and Kevin Schnelli. On the support of the free additive convolution. *Journal d'Analyse Mathématique*, 142(1):323–348, 2020.

[11] Jean Barbier, Wei-Kuo Chen, Dmitry Panchenko, and Manuel Sáenz. Performance of bayesian linear regression in a model with mismatch, 2021. `arXiv:2107.06936`.

[12] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems*, 2016.

[13] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[14] Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability theory and related fields*, 174(3):1133–1185, 2019.

[15] Jean Barbier and Nicolas Macris. The adaptive interpolation method for proving replica formulas. Applications to the Curie–Weiss and Wigner spike models. *Journal of Physics A: Mathematical and Theoretical*, 52(29):294002, 2019.

[16] Jean Barbier, Nicolas Macris, Antoine Maillard, and Florent Krzakala. The mutual information in random linear estimation beyond i.i.d. matrices. In *International Symposium on Information Theory*, pages 1390–1394, 2018.

[17] Jean Barbier, Nicolas Macris, and Cynthia Rush. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *Advances in Neural Information Processing Systems*, 2020.

[18] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[19] Mohsen Bayati and Andrea Montanari. The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[20] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

[21] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[22] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[23] Bhaswar B. Bhattacharya and Subhabrata Sen. High temperature asymptotics of orthogonal mean-field spin glasses. *Journal of Statistical Physics*, 162(1):63–80, 2016.

[24] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[25] Jelena Bradic. Robustness in sparse high-dimensional linear models: relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, 10(2):3894–3944, 2016.

[26] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1):1–47, 2009.

[27] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

[28] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.

[29] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory*, pages 2197–2201, 2014.

[30] David L. Donoho, Adel Javanmard, and Andrea Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, 2013.

[31] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[32] David L. Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[33] Rishabh Dudeja, Junjie Ma, and Arian Maleki. Information theoretic limits for phase retrieval with subsampled haar sensing matrices. *IEEE Transactions on Information Theory*, 66(12):8002–8045, 2020.

[34] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

[35] Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[36] Barbara E. Engelhardt and Matthew Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics*, 6(9):e1001117, 2010.

[37] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224, 2022.

[38] Zhou Fan and Yihong Wu. The replica-symmetric free energy for ising spin glasses with orthogonally invariant couplings, 2021. `arXiv:2105.02797`.

[39] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on Approximate Message Passing, 2021. `arXiv:2105.02180`.

[40] Delphine Féral and Sandrine Péché. The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in mathematical physics*, 272(1):185–228, 2007.

[41] Alyson K. Fletcher and Sundeep Rangan. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562, 2018.

[42] Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 2019.

[43] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: how to prove Kabashima's replica formula), 2020. `arXiv:2006.06581`.

[44] Alice Guionnet and Jiaoyang Huang. Large deviations asymptotics of rectangular spherical integral, 2021. `arXiv:2106.07146`.

[45] Alice Guionnet and Mylène Maida. A Fourier view on the R-transform and related asymptotics of spherical integrals. *Journal of Functional Analysis*, 222(2):435–490, 2005.

[46] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[47] TianQi Hou, YuHao Liu, Teng Fu, and Jean Barbier. Sparse superposition codes under VAMP decoding with generic rotational invariant coding matrices, 2022. `arXiv:2202.04541`.

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

[49] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[50] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[51] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

[52] Antti Knowles and Jun Yin. The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, 2013.

[53] Satish Babu Korada and Nicolas Macris. Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230, 2009.

[54] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3-4):859–929, 2018.

[55] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[56] Junjie Ma, Ji Xu, and Arian Maleki. Analysis of sensing spectral for signal recovery under a generalized linear model. In *Advances in Neural Information Processing Systems*, 2021.

[57] Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

[58] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Advances in Neural Information Processing Systems*, 2020.

[59] Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[60] Marco Mondelli and Ramji Venkataramanan. PCA initialization for approximate message passing in rotationally invariant models. In *Neural Information Processing Systems*, 2021.

[61] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 45(1):321–345, 2021.

[62] Cristopher Moore. The computer science and physics of community detection: landscapes, phase transitions, and hardness, 2017. `arXiv:1702.00467`.

[63] Sumit Mukherjee and Subhabrata Sen. Variational inference in high-dimensional linear regression, 2021. `arXiv:2104.12232`.

[64] Jonathan Novak. Three lectures on free probability. *Random matrix theory, interacting particle systems, and integrable systems*, 65(309-383):13, 2014.

[65] Manfred Opper, Burak Cakmak, and Ole Winther. A theory of solving tap equations for ising models with general invariant random matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11):114002, 2016.

[66] Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347, 2020.

[67] Giorgio Parisi and Marc Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995.

[68] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[69] Amelia Perry, Alexander S. Wein, Afonso S. Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

[70] Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory*. Cambridge University Press, 2021.

[71] Farzad Pourkamali and Nicolas Macris. Mismatched estimation of rank-one symmetric matrices under gaussian noise, 2021. `arXiv:2107.08927`.

[72] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[73] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *International Symposium on Information Theory*, pages 2168–2172, 2011.

[74] Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

[75] Kolyan Ray and Botond Szabó. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12, 2021.

[76] Subrata Sarkar, Rizwan Ahmad, and Philip Schniter. MRI image recovery using damped denoising vector AMP. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8108–8112, 2021.

[77] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[78] Takashi Takahashi and Yoshiyuki Kabashima. Macroscopic analysis of vector approximate message passing in a model mismatch setting. In *International Symposium on Information Theory*, pages 1403–1408, 2020.

[79] Koujin Takeda, Shinsuke Uda, and Yoshiyuki Kabashima. Analysis of CDMA systems that are characterized by eigenvalue spectrum. *EPL (Europhysics Letters)*, 76(6):1193, 2006.

[80] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[81] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, 2015.

[82] Ramji Venkataramanan, Kevin Kögler, and Marco Mondelli. Estimation in rotationally invariant generalized linear models via approximate message passing, 2021. `arXiv:2112.04330`.

[83] Sergio Verdú. Mismatched estimation and relative entropy. *IEEE Transactions on Information Theory*, 56(8):3712–3720, 2010.

[84] Cédric Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

[85] Tsachy Weissman. The relationship between causal and noncausal mismatched estimation in continuous-time awgn channels. *IEEE Transactions on Information Theory*, 56(9):4256–4273, 2010.

[86] Xinyi Zhong, Chang Su, and Zhou Fan. Empirical bayes PCA in high dimensions, 2020. `arXiv:2012.11676`.

[87] Xinyi Zhong, Tianhao Wang, and Zhou Fan. Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization, 2021. `arXiv:2110.02318`.

[88] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Our model assumptions are clearly explained in Assumptions 1, 2 and 3. To obtain the overlap of the mismatched Bayes estimator, we make a conjecture on the convergence of $\|M_{\mathrm{mis}}\|_{\mathrm{F}}/N$ (see the discussion right after the statement of Theorem 1).
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not see direct paths to applications with negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our model assumptions are clearly explained in Assumptions 1, 2 and 3. To obtain the overlap of the mismatched Bayes estimator, we make a conjecture on the convergence of $\|M_{\mathrm{mis}}\|_{\mathrm{F}}/N$ (see the discussion right after the statement of Theorem 1).
   (b) Did you include complete proofs of all theoretical results? [Yes] Proof sketches are included right after the corresponding statements. The full proofs of Theorem 1 and 2 are in Appendices B and C, respectively.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In Section 4, we describe the experimental setup containing the instructions to reproduce the results of Figure 1 (and of Figure 2 contained in Appendix D.3 as well). Additional details are given in Appendices D.1 and D.2. The code that implements the various algorithms is included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All details of the experimental setup are described in Section 4 and in Appendices D.1 and D.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Each experiment is repeated for 10 independent runs, and we report the average together with the confidence interval at 1 standard deviation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not perform large scale experiments that required to specify this type of details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We are not using existing assets or curating/releasing new assets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Low-rank perturbations of rotationally invariant matrices

In this appendix, we recall some known results concerning low-rank perturbations of rotationally invariant matrices [21, 22]. We present them in a form which is more convenient for our discussion, and we specialize them for rank-1 perturbations. The notations are the same as in Section 3.

The first result characterizes the asymptotic value of the largest eigenvalue of the perturbed symmetric matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times N}$.

Theorem 3 (Theorem 2.1 of [21])**.** *Under Assumption 1, as $N \to +\infty$, the largest eigenvalue $\overline{\nu}_N$ of $\boldsymbol{Y}$ converges almost surely to*

$$\overline{\nu} := K_\rho\Big(\frac{1}{\sqrt{\lambda_*}}\Big)\mathbf{1}\big(\overline{h}\sqrt{\lambda_*} \geq 1\big) + \overline{\gamma}\,\mathbf{1}\big(\overline{h}\sqrt{\lambda_*} < 1\big).$$

This theorem thus implies that, below the phase transition, the presence of the perturbation does not modify the largest eigenvalue of the noise matrix. For the inner product between $\boldsymbol{X}$ and the eigenvector of $\boldsymbol{Y}$ corresponding to the largest eigenvalue, the following result holds.

Theorem 4 (Theorem 2.2 of [21])**.** *Under Assumption 1, as $N \to +\infty$, the eigenvector $\overline{\boldsymbol{v}}_N$ corresponding to the largest eigenvalue $\overline{\nu}_N$ of $\boldsymbol{Y}$ is such that, almost surely,*

$$C(\rho, \lambda_*) := \lim_{N\to+\infty} \frac{\langle \boldsymbol{X}, \overline{\boldsymbol{v}}_N\rangle^2}{N} = \Big(1 - \frac{1}{\lambda_*} R'_\rho\Big(\frac{1}{\sqrt{\lambda_*}}\Big)\Big)\mathbf{1}\big(\overline{h}\sqrt{\lambda_*} \geq 1\big).$$

Below the phase transition the spectral estimators are thus asymptotically uncorrelated with the signal while above the transition the leading eigenvector starts to align in its direction.

B Proofs for the Bayes estimator

It is useful to define a slight generalization of the model (1) that includes a "Wigner regularization" of the rotationally-invariant noise. Namely, in this section, we consider the following generalized model for the data:

$$\boldsymbol{Y}_\epsilon = \frac{\sqrt{\lambda_*}}{N}\boldsymbol{X}\boldsymbol{X}^\intercal + \boldsymbol{Z} + \sqrt{\epsilon}\,\boldsymbol{W}, \tag{19}$$

where, as before, $\boldsymbol{W}$ is a standard Wigner matrix and $\epsilon > 0$ is some constant. We denote by $\langle\cdot\rangle_\epsilon$ the mean w.r.t. the associated mismatched posterior obtained by replacing $\boldsymbol{Y}$ with $\boldsymbol{Y}_\epsilon$ in equation (2):

$$\langle g(\boldsymbol{x})\rangle_\epsilon := \frac{1}{Z_N(\boldsymbol{Y}_\epsilon)} \int g(\boldsymbol{x}) \exp\Big(\frac{\sqrt{\lambda}}{2}\langle \boldsymbol{x}, \boldsymbol{Y}_\epsilon \boldsymbol{x}\rangle\Big)\,\mu_N(d\boldsymbol{x}). \tag{20}$$

At $\epsilon = 0$, the operator $\langle\cdot\rangle_0$ corresponds to the expectation w.r.t. to the original mismatched measure (2). The other notations are analogous to those for the non-regularized model, though we will add a subscript ϵ when it is useful to remark the dependence of a given quantity over ϵ. For example, we denote ρ_ϵ the weak limit of the empirical spectral density of $\boldsymbol{Y}_\epsilon$, R_ϵ its associated R-transform, $\overline{\nu}_\epsilon$ the limit of the maximum eigenvalue of $\boldsymbol{Y}_\epsilon$, $\overline{\gamma}_\epsilon$ the limit of the maximum eigenvalue of the noise matrix $\boldsymbol{Z} + \sqrt{\epsilon}\,\boldsymbol{W}$ and $\overline{h}_\epsilon := \lim_{z\downarrow\overline{\gamma}_\epsilon} H_{\rho_\epsilon}(z)$.

Our first intermediate result, Proposition 1, characterizes the log-spherical integral (which plays the role of log-moment generating function) for the regularized data matrix:

$$I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_\epsilon\Big) := \int \exp\Big(\frac{\sqrt{\lambda}}{2}\langle \boldsymbol{x}, \boldsymbol{Y}_\epsilon \boldsymbol{x}\rangle\Big)\,\mu_N(d\boldsymbol{x}). \tag{21}$$

Although this proposition may be interesting in itself, its main purpose is to allow us to derive the MSE of the mismatched Bayes estimator in terms of its derivatives w.r.t. ϵ and λ_*. This is in fact the reason for introducing the regularization in the first place.

Like in the analysis of mismatched inference with Gaussian noise of [71], we will make use of the following result for the asymptotic value of the rank-1 spherical integral.

Theorem 5 (Theorem 6 of [45])**.** *Define $\overline{h}_\epsilon := \lim_{z\downarrow\overline{\nu}_\epsilon} H_{\rho_\epsilon}(z)$. Assume that Assumption 1 holds. Recall that $\overline{\nu}_\epsilon$ is the limit of the maximum eigenvalue of $\boldsymbol{Y}_\epsilon$ and must be finite. Then,*

$$\lim_{N\to+\infty} \frac{1}{N} \ln I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_\epsilon\Big) = \frac{1}{2}\int_0^{\sqrt{\lambda}} R_{\rho_\epsilon}(t)dt\,\mathbf{1}\big(\sqrt{\lambda} \leq \overline{h}_\epsilon\big) + g_{\lambda,\epsilon}(\overline{\nu}_\epsilon)\,\mathbf{1}\big(\sqrt{\lambda} > \overline{h}_\epsilon\big).$$

We can now state our first intermediate proposition.

Proposition 1 (Spherical integral)**.** *For every $\lambda, \epsilon > 0$, define the function $g_{\lambda,\epsilon} : (\overline{\gamma}_\epsilon, +\infty) \mapsto \mathbb{R}$ as*

$$g_{\lambda,\epsilon}(x) := \frac{\sqrt{\lambda}}{2}x - \frac{1}{2}\int d\rho_\epsilon(\gamma)\ln(x-\gamma) - \frac{1}{2} - \frac{1}{4}\ln\lambda.$$

Then, under Assumption 1, for every $\lambda, \lambda_* > 0$ we almost surely have that

$$\lim_{N\to+\infty}\frac{1}{N}\ln I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_\epsilon\Big) = f_\epsilon(\lambda,\lambda_*) := \begin{cases} g_{\lambda,\epsilon}(K_{\rho_\epsilon}(\frac{1}{\sqrt{\lambda_*}})) & \text{if } \overline{h}_\epsilon\sqrt{\lambda_*} \geq 1 \text{ and } \lambda\lambda_* > 1, \\ g_{\lambda,\epsilon}(\overline{\gamma}_\epsilon) & \text{if } \overline{h}_\epsilon\sqrt{\lambda_*} < 1 \text{ and } \sqrt{\lambda} > \overline{h}_\epsilon, \\ \frac{1}{2}\int_0^{\sqrt{\lambda}} R_{\rho_\epsilon}(t)dt & \text{otherwise.} \end{cases}$$

Moreover we also have that

$$\lim_{N\to+\infty}\frac{1}{N}\mathbb{E}\ln I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_\epsilon\Big) = f_\epsilon(\lambda,\lambda_*).$$

Proof of Proposition 1. By Theorem 3 applied to the generalized model (19), we are almost surely under the hypothesis of Theorem 5. Combining these two results proves the first claim.

To get the result in expectation, it suffices to notice that by Assumption 1 the sequence of values of the largest eigenvalue of $\boldsymbol{Z}$ is a.s. convergent. The same is a.s. true for the Wigner matrix $\boldsymbol{W}$. There thus almost surely exists some bounded $C > 0$ such that

$$\frac{\sqrt{\lambda}}{2}\langle \boldsymbol{x}, \boldsymbol{Y}_\epsilon \boldsymbol{x}\rangle \leq \frac{\sqrt{\lambda}}{2}(C + \sqrt{\lambda_*})N.$$

Then, the last result follows by dominated convergence. □

B.1 Proof of Theorem 1

To establish this result, define the (matrix) *magnetization* and the (matrix) *overlap* according to

$$M_N := \Big(\frac{\langle \boldsymbol{X}, \boldsymbol{x}\rangle}{N}\Big)^2, \qquad Q_N := \Big(\frac{\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle}{N}\Big)^2.$$

respectively. Here, the supra-indices indicate two conditionally (on $\boldsymbol{Y}_\epsilon$) independent samples from the mismatched posterior $P_{\text{mis}}(\cdot \mid \boldsymbol{Y}_\epsilon)$ of the generalized model (19) (with mean (20)). Theorem 1 is the consequence of the asymptotic formulas for the mean magnetization and overlap given in the lemmas below.

Lemma 1. *In the setting considered, for all $\lambda, \lambda_* > 0$ fixed and letting $M(\cdot,\cdot)$ be given by* (9)*, we almost surely have*

$$\lim_{N\to+\infty}\langle M_N\rangle_0 = M(\lambda,\lambda_*).$$

Proof. For this proof we can set $\epsilon = 0$ all along. Note that, in the exponent of the spherical integral (21), the only term in which λ_* appears (when writing $\boldsymbol{Y}_0$ explicitly in terms of $\boldsymbol{X}\boldsymbol{X}^\top$) is

$$\frac{\sqrt{\lambda\lambda_*}}{2N}\langle \boldsymbol{X}, \boldsymbol{x}\rangle^2 = N\frac{\sqrt{\lambda\lambda_*}}{2}M_N.$$

Then, we have that for all $N \geq 1$

$$\frac{d}{d\sqrt{\lambda_*}}\frac{1}{N}\ln I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_0\Big) = \frac{\sqrt{\lambda}}{2}\langle M_N\rangle_0.$$

We also have

$$\frac{d^2}{(d\sqrt{\lambda_*})^2}\frac{1}{N}\ln I_N\Big(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_0\Big) = \frac{\lambda N}{4}\Big(\langle M_N^2\rangle_0 - \langle M_N\rangle_0^2\Big) \geq 0.$$

Thus, by the convexity of $\ln I_N(\frac{\sqrt{\lambda}}{2}, \boldsymbol{Y}_0)$ w.r.t. $\sqrt{\lambda_*}$ and Proposition 1, we get that almost surely

$$\lim_{N\to+\infty}\langle M_N\rangle_0 = \frac{2}{\sqrt{\lambda}}\frac{df_0(\lambda,\lambda_*)}{d\sqrt{\lambda_*}} = 4\sqrt{\frac{\lambda_*}{\lambda}}\frac{df_0(\lambda,\lambda_*)}{d\lambda_*}.$$

Recall that $K_\rho(x) = R_\rho(x) + 1/x$. Then, the derivative of f_0 with respect to λ_* is given by

$$\frac{df_0}{d\lambda_*}(\lambda,\lambda_*) = \frac{1}{4}\Big(\frac{1}{\lambda_*^2} - \Big(\frac{\lambda}{\lambda_*^3}\Big)^{1/2}\Big)\Big(R_\rho'\Big(\frac{1}{\sqrt{\lambda_*}}\Big) - \lambda_*\Big)\mathbf{1}\big(\overline{h}\sqrt{\lambda_*} \geq 1 \cap \lambda\lambda_* > 1\big)$$

with $\overline{h} = \overline{h}_0$. From these two last equations, the result of the lemma follows. □

Lemma 2. *In the setting considered, for all $\lambda, \lambda_* > 0$ fixed and letting $Q(\cdot,\cdot)$ be given by* (9)*, we have*

$$\lim_{N\to+\infty} \mathbb{E}\langle Q_N\rangle_0 = Q(\lambda,\lambda_*).$$

Proof. Note that in the exponent of the spherical integral (21) the only term in which ϵ appears (when writing $\boldsymbol{Y}_\epsilon$ explicitly in terms of $\boldsymbol{W}$) is $\sqrt{\lambda\epsilon}\,\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}/2$. Then, we have that for every $N \geq 1$

$$\frac{d}{d\sqrt{\epsilon}}\frac{1}{N}\mathbb{E}\ln I_N\Big(\frac{\sqrt{\lambda}}{2},\boldsymbol{Y}_\epsilon\Big)\Big|_{\epsilon=0} = \frac{\sqrt{\lambda}}{2N}\mathbb{E}\big\langle \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\big\rangle_0. \tag{22}$$

Furthermore,

$$\frac{d^2}{(d\sqrt{\epsilon})^2}\mathbb{E}\ln I_N\Big(\frac{\sqrt{\lambda}}{2},\boldsymbol{Y}_\epsilon\Big) = \frac{\lambda}{4}\mathbb{E}\Big(\big\langle(\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x})^2\big\rangle_\epsilon - \big\langle\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}\big\rangle_\epsilon^2\Big) \geq 0. \tag{23}$$

Since $\boldsymbol{W}$ is Wigner, its elements are distributed as independent Gaussian random variables (up to symmetry) with variance $1/N$ for the off-diagonal elements and $2/N$ on the diagonal. Using Gaussian integration by parts[3] in the r.h.s. of (22), we get that

$$\frac{d}{d\sqrt{\epsilon}}\frac{1}{N}\mathbb{E}\ln I_N\Big(\frac{\sqrt{\lambda}}{2},\boldsymbol{Y}_\epsilon\Big) = \frac{\lambda\sqrt{\epsilon}}{2}(1-\mathbb{E}\langle Q_N\rangle_\epsilon), \tag{24}$$

and thus

$$\frac{d}{d\epsilon}\frac{1}{N}\mathbb{E}\ln I_N\Big(\frac{\sqrt{\lambda}}{2},\boldsymbol{Y}_\epsilon\Big)\Big|_{\epsilon=0} = \frac{\lambda}{4}(1-\mathbb{E}\langle Q_N\rangle_0). \tag{25}$$

Hence, by the convexity of $\mathbb{E}\ln I_N(\frac{\sqrt{\lambda}}{2},\boldsymbol{Y}_\epsilon)$ w.r.t. ϵ and Proposition 1, we obtain that

$$\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 = 1 - \frac{4}{\lambda}\frac{df_\epsilon(\lambda,\lambda_*)}{d\epsilon}\Big|_{\epsilon=0}. \tag{26}$$

To find the asymptotic value of $\mathbb{E}\langle Q_N\rangle_0$, we are thus left to compute $d_\epsilon f_\epsilon(\lambda,\lambda_*)$. We will then need to consider the three regimes of Proposition 1.

Firstly, if $\overline{h}\sqrt{\lambda_*} \geq 1$ and $\lambda\lambda_* > 1$, then

$$\begin{aligned}\frac{df_\epsilon(\lambda,\lambda_*)}{d\epsilon}\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}g_{\lambda,\epsilon}(K_{\rho_\epsilon}(\lambda_*^{-1/2}))\Big|_{\epsilon=0}\\ &= \frac{\partial g_{\lambda,\epsilon}(x)}{\partial\epsilon}\Big|_{\epsilon=0,x=K_{\rho_0}(\lambda_*^{-1/2})} + \frac{dg_{\lambda,\epsilon}(x)}{dx}\Big|_{\epsilon=0,x=K_{\rho_0}(\lambda_*^{-1/2})}\frac{dK_{\rho_\epsilon}(\lambda_*^{-1/2})}{d\epsilon}\Big|_{\epsilon=0}.\end{aligned}$$

For computing $\partial_\epsilon g_{\lambda,\epsilon}(x)$, note that the only term of $g_{\lambda,\epsilon}$ depending on ϵ is

$$\ell(x,\epsilon) := -\frac{1}{2}\int d\rho_\epsilon(\gamma)\ln(x-\gamma),$$

for which we have $d_x\ell(x,\epsilon) = -\frac{1}{2}H_{\rho_\epsilon}(x)$. Thus, for any appropriate x_0,

$$\ell(x,\epsilon) = -\frac{1}{2}\int_{x_0}^{x} H_{\rho_\epsilon}(y)dy + C(x_0,\epsilon),$$

where $C(x_0,\epsilon)$ is a constant depending only on x_0 and ϵ. Furthermore, by the Dyson Brownian motion characterization of the eigenvalues of the matrix $\boldsymbol{Z} + \sqrt{\epsilon}\boldsymbol{W}$, we have that the limiting Stieltjes transform $H_{\rho_\epsilon}(x)$ is a solution of Burger's equation (see for example [4, Proposition 4.3.10]); that is,

$$\frac{dH_{\rho_\epsilon}(x)}{d\epsilon} = -H_{\rho_\epsilon}(x)\frac{dH_{\rho_\epsilon}(x)}{dx} = -\frac{1}{2}\frac{dH_{\rho_\epsilon}^2(x)}{dx}. \tag{27}$$

[3] The used formula here is $\mathbb{E}Zf(Z) = \sigma^2\mathbb{E}f'(Z)$ for $Z \sim \mathcal{N}(0,\sigma^2)$.

We then have

$$\frac{d\ell(x,\epsilon)}{d\epsilon} = -\frac{1}{2}\int_{x_0}^{x}\frac{dH_{\rho_\epsilon}(y)}{d\epsilon}dy + \frac{dC(x_0,\epsilon)}{d\epsilon} = \frac{1}{4}H_{\rho_\epsilon}^2(x) - \frac{1}{4}H_{\rho_\epsilon}^2(x_0) + \frac{dC(x_0,\epsilon)}{d\epsilon}. \tag{28}$$

As the left hand-side is independent of x_0, $-\frac{1}{4}H_{\rho_\epsilon}^2(x_0) + d_\epsilon C(x_0,\epsilon)$ is some function $c(\epsilon)$ of ϵ alone. This gives that

$$\frac{\partial g_{\lambda,\epsilon}(x)}{\partial\epsilon}\Big|_{\epsilon=0,x=K_{\rho_0}(\lambda_*^{-1/2})} = \frac{1}{4}H_{\rho_0}^2(K_{\rho_0}(\lambda_*^{-1/2})) + c(0) = \frac{1}{4\lambda_*} + c(0). \tag{29}$$

Furthermore,

$$\frac{dg_{\lambda,\epsilon}(x)}{dx} = \frac{\sqrt{\lambda}}{2} - \frac{1}{2}H_{\rho_\epsilon}(x), \tag{30}$$

and $d_\epsilon K_{\rho_\epsilon}(x) = x$ (which follows from the additivity of the R-transform and the fact that the R-transform associated with the asymptotic spectral density of the Wigner part $\sqrt{\epsilon}\boldsymbol{W}$ is ϵx). Then,

$$\frac{dg_{\lambda,\epsilon}(x)}{dx}\Big|_{\epsilon=0,x=K_{\rho_0}(\lambda_*^{-1/2})}\frac{dK_{\rho_\epsilon}(\lambda_*^{-1/2})}{d\epsilon}\Big|_{\epsilon=0} = \frac{1}{2}\Big(\sqrt{\frac{\lambda}{\lambda_*}} - \frac{1}{\lambda_*}\Big). \tag{31}$$

Combining equations (26), (29), and (31) we get that, if $\overline{h}\sqrt{\lambda_*} \geq 1$ and $\lambda\lambda_* > 1$, then

$$\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 = 1 - \frac{2}{\sqrt{\lambda\lambda_*}} + \frac{1}{\lambda\lambda_*} + c(0) = \Big(1 - \frac{1}{\sqrt{\lambda\lambda_*}}\Big)^2 + c(0).$$

We now show that $c(0) = 0$. First note that $c(0)$ does not depend on λ nor λ_*. Assume that $c(0)$ is positive and fix λ_* s.t. $\overline{h}\sqrt{\lambda_*} > 1$. Choose λ large enough so that we simultaneously have $\lambda\lambda_* > 1$ and $(1 - 1/\sqrt{\lambda\lambda_*})^2 + c(0) > 1$. This can be done because $c(0) > 0$. For these values of λ and λ_*, we thus have that $\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 > 1$. However, this is a contradiction, since for every $\lambda, \lambda_* > 0$ and $N \geq 1$ a.s. $0 \leq Q_N \leq 1$. We therefore have that $c(0)$ is non-positive. In a similar way, we can prove that $c(0)$ is non-negative. Then, we conclude that, if $\overline{h}\sqrt{\lambda_*} \geq 1$ and $\lambda\lambda_* > 1$,

$$\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 = \Big(1 - \frac{1}{\sqrt{\lambda\lambda_*}}\Big)^2.$$

Secondly, we will now establish the limit of the overlap when $\overline{h}\sqrt{\lambda_*} < 1$ and $\sqrt{\lambda} > \overline{h}$. In this region,

$$\frac{df_\epsilon(\lambda,\lambda_*)}{d\epsilon}\Big|_{\epsilon=0} = \frac{d}{d\epsilon}g_{\lambda,\epsilon}(\overline{\gamma}_\epsilon)\Big|_{\epsilon=0} = \frac{\partial g_{\lambda,\epsilon}(x)}{\partial\epsilon}\Big|_{\epsilon=0,x=\overline{\gamma}_0} + \frac{\partial g_{\lambda,\epsilon}(x)}{\partial x}\Big|_{\epsilon=0,x=\overline{\gamma}_0}\frac{\partial\overline{\gamma}_\epsilon}{\partial\epsilon}\Big|_{\epsilon=0}. \tag{32}$$

The partial derivatives $\partial_\epsilon g_{\lambda,\epsilon}(x)$ and $\partial_x g_{\lambda,\epsilon}(x)$ are obtained as before. Therefore, we only need to compute $\partial_\epsilon\overline{\gamma}|_{\epsilon=0}$. We will prove that as $\epsilon \to 0$, $\overline{\gamma}_\epsilon = \overline{\gamma}_0 + \overline{h}_0\epsilon + o(\epsilon)$. As a consequence of the large deviation principle of [44, Theorem 2.5], $\overline{\gamma}_\epsilon$ is equal to the supremum of the compact support K_ϵ of ρ_ϵ. By Assumption 2, for all $\epsilon > 0$ we have $\lim_{z\downarrow\overline{\gamma}_\epsilon} H'_{\rho_\epsilon}(z) = -\infty$. This implies that $\lim_{x\downarrow\overline{h}_\epsilon} K'_{\rho_\epsilon}(x) = 0$. Thus, $\overline{h}_\epsilon$ is a solution to the equation $K'_{\rho_0}(x) + \epsilon = 0$ (again, this can be seen by additivity of the R-transform and its link to K_{ρ_ϵ}). As K'_{ρ_0} is analytic and monotonously increasing, the solution is unique and depends smoothly on ϵ. Thus, $\overline{h}_\epsilon = \overline{h}_0 + \mathcal{O}(\epsilon)$, which means that

$$\overline{\gamma}_\epsilon = \lim_{x\downarrow\overline{h}_\epsilon} K_{\rho_\epsilon}(x) = \lim_{x\downarrow\overline{h}_0} K_{\rho_0}(x) + \epsilon\overline{h}_0 + o(\epsilon) = \overline{\gamma}_0 + \epsilon\overline{h}_0 + o(\epsilon),$$

where we used that $\lim_{x\downarrow\overline{h}_0} K'_{\rho_0}(x) = 0$. This proves that

$$\partial_\epsilon\overline{\gamma}_\epsilon|_{\epsilon=0} = \overline{h}_0.$$

Combining this with equations (26), (28), (30), and (32) gives the result in this region.

Finally, by Proposition 1 and (26), if (λ,λ_*) does not verify either of the conditions considered above, then by dominated convergence

$$\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 = 1 - \frac{2}{\lambda}\int_0^{\sqrt{\lambda}}\partial_\epsilon R_{\rho_\epsilon}(x)dx = 1 - \frac{2}{\lambda}\int_0^{\sqrt{\lambda}} x\,dx = 0.$$

This finishes the proof that $\lim_{N\to+\infty}\mathbb{E}\langle Q_N\rangle_0 = Q(\lambda,\lambda_*)$. □

Proof of Theorem 1. In order to establish the formula for the MSE, it suffices to use Lemmas 1 and 2 combined with the observation that

$$\frac{1}{2N^2}\mathbb{E}\|\langle\boldsymbol{x}\boldsymbol{x}^\top\rangle_0 - \boldsymbol{X}\boldsymbol{X}^\top\|_{\mathrm{F}}^2 = \frac{1}{2}\Big(1 + \mathbb{E}\langle Q_N\rangle_0 - 2\mathbb{E}\langle M_N\rangle_0\Big).$$

□

C Proofs for Approximate Message Passing

C.1 Auxiliary AMP

The iterates of the auxiliary AMP are denoted by $\tilde{\boldsymbol{z}}^t, \tilde{\boldsymbol{x}}^t \in \mathbb{R}^N$, and they are computed as follows, for $t \geq 1$:

$$\tilde{\boldsymbol{z}}^t = \boldsymbol{Z}\tilde{\boldsymbol{x}}^t - \sum_{i=1}^{t} \mathsf{b}_{t,i}\tilde{\boldsymbol{x}}^i, \quad \tilde{\boldsymbol{x}}^{t+1} = \tilde{h}_{t+1}(\tilde{\boldsymbol{z}}^1, \dots, \tilde{\boldsymbol{z}}^t, \hat{\boldsymbol{x}}^1, \boldsymbol{X}). \tag{33}$$

The iteration (33) is initialized with $\tilde{\boldsymbol{x}}^1 = \hat{\boldsymbol{x}}^1$, where $\hat{\boldsymbol{x}}^1$ is also the initialization of the true AMP (see (7)). For $t \geq 1$, the functions $\tilde{h}_{t+1} : \mathbb{R}^{t+2} \to \mathbb{R}$ are applied component-wise, and they are recursively defined as

$$\begin{aligned}\tilde{h}_{t+1}(z_1, \dots, z_t, \hat{x}_1, x) = h_{t+1}\Bigg(z_t + (\boldsymbol{B}_t)_{t,1}\hat{x}_1 + \sum_{i=2}^{t}(\boldsymbol{B}_t)_{t,i}\tilde{h}_i\left(z_1, \dots, z_{i-1}, \hat{x}_1, x\right) + \mu_t x \\ - \bar{\beta}_t \tilde{h}_{t-1}\left(z_1, \dots, z_{t-2}, \hat{x}_1, x\right)\Bigg).\end{aligned} \tag{34}$$

The idea is that the choice (34) for the denoisers $\{\tilde{h}_{t+1}\}_{t\geq 1}$ allows to "correct" for the mismatch and compensate for the wrong Onsager corrections in (8). Here, h_{t+1} is the denoiser of the true AMP (see (8)), and $(\mu_t, \bar{\beta}_t, \boldsymbol{B}_t)$ come from the state evolution recursion of the true AMP: μ_1 is given by (13), and, for $t \geq 2$, μ_t is given by (15); $\bar{\beta}_1 = 0$ and, for $t \geq 2$, $\bar{\beta}_t = \mathbb{E}[h_t'(x_{t-1})]$, where the law of x_{t-1} is given by (14); and $\boldsymbol{B}_t$ is defined via (16). We now discuss how to obtain the coefficients $\{\mathsf{b}_{t,i}\}_{i=1}^t$. Let us define the matrix $\boldsymbol{\Phi}_t \in \mathbb{R}^{t\times t}$ as

$$(\boldsymbol{\Phi}_t)_{i,j} = 0, \quad \text{for } i \leq j, \qquad (\boldsymbol{\Phi}_t)_{i,j} = \langle \partial_j \tilde{\boldsymbol{x}}^i \rangle, \quad \text{for } i > j, \tag{35}$$

where, for $j < i$, the vector $\langle \partial_j \tilde{\boldsymbol{x}}^i \rangle \in \mathbb{R}^N$ denotes the partial derivative of $\tilde{h}_i : \mathbb{R}^{i+1} \to \mathbb{R}$ with respect to the j-th input (applied component-wise). Then, the vector $(\mathsf{b}_{t,1}, \dots, \mathsf{b}_{t,t})$ is given by the last row of the matrix $\tilde{\boldsymbol{B}}_t \in \mathbb{R}^{t\times t}$ defined as

$$\tilde{\boldsymbol{B}}_t = \sum_{j=0}^{t-1} \kappa_{j+1}\boldsymbol{\Phi}_t^j, \tag{36}$$

where $\{\kappa_k\}_{k\geq 1}$ denotes the sequence of free cumulants associated to the matrix $\boldsymbol{Z}$.

C.2 State evolution of auxiliary AMP

Using Theorem 2.3 in [87], we provide a state evolution result for the auxiliary AMP (33). In particular, we show in Proposition 2 that the joint empirical distribution of $(\tilde{\boldsymbol{z}}^1, \dots, \tilde{\boldsymbol{z}}^t)$ converges to a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_t)$.

The covariance matrices $\{\tilde{\boldsymbol{\Sigma}}_t\}_{t\geq 1}$ are defined recursively, starting with $\tilde{\boldsymbol{\Sigma}}_1 = \bar{\kappa}_2\mathbb{E}[\hat{x}_1^2]$, where $\hat{x}_1$ is defined in (7). Given $\tilde{\boldsymbol{\Sigma}}_t$, let

$$\begin{aligned}&(\tilde{z}_1, \dots, \tilde{z}_t) \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_t) \text{ and independent of } (X, \hat{x}_1), \\ &\tilde{x}_s = \tilde{h}_s\left(\tilde{z}_1, \dots, \tilde{z}_{s-1}, \hat{x}_1, X\right), \quad \text{for } s \in \{2, \dots, t+1\},\end{aligned} \tag{37}$$

where $\tilde{h}_s$ is defined via (34) and we set $\tilde{x}_1 = \hat{x}_1$. Let $\tilde{\boldsymbol{\Phi}}_{t+1}, \tilde{\boldsymbol{\Delta}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ be matrices with entries given by

$$\begin{aligned}&(\tilde{\boldsymbol{\Phi}}_{t+1})_{i,j} = 0, \quad \text{for } i \leq j, \qquad (\tilde{\boldsymbol{\Phi}}_{t+1})_{i,j} = \mathbb{E}[\partial_j \tilde{x}_i], \quad \text{for } i > j, \\ &(\tilde{\boldsymbol{\Delta}}_{t+1})_{i,j} = \mathbb{E}[\tilde{x}_i\tilde{x}_j], \quad 1 \leq i, j \leq t+1,\end{aligned} \tag{38}$$

where $\partial_j \tilde{x}_i$ denotes the partial derivative $\partial_{z_j}\tilde{h}_i(\tilde{z}_1, \dots, \tilde{z}_{i-1}, \hat{x}_1, X)$. Then, we compute the covariance matrix $\tilde{\boldsymbol{\Sigma}}_{t+1}$ as

$$\tilde{\boldsymbol{\Sigma}}_{t+1} = \sum_{j=0}^{2t} \bar{\kappa}_{j+2}\tilde{\boldsymbol{\Theta}}_{t+1}^{(j)}, \qquad \text{with} \quad \tilde{\boldsymbol{\Theta}}_{t+1}^{(j)} = \sum_{i=0}^{j}(\tilde{\boldsymbol{\Phi}}_{t+1})^i\tilde{\boldsymbol{\Delta}}_{t+1}(\tilde{\boldsymbol{\Phi}}_{t+1}^\top)^{j-i}. \tag{39}$$

It can be verified that the $t \times t$ top left sub-matrix of $\tilde{\boldsymbol{\Sigma}}_{t+1}$ is given by $\tilde{\boldsymbol{\Sigma}}_t$.

Proposition 2 (State evolution for auxiliary AMP)**.** *Consider the auxiliary AMP in* (33) *and the state evolution random variables defined in* (37)*. Let* $\tilde{\psi} : \mathbb{R}^{2t+2} \to \mathbb{R}$ *be any pseudo-Lipschitz functions of order 2. Then for each* $t \geq 1$*, we almost surely have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \tilde{\psi}\left((\tilde{\boldsymbol{z}}^1)_i, \ldots, (\tilde{\boldsymbol{z}}^t)_i, (\tilde{\boldsymbol{x}}^1)_i, \ldots, (\tilde{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i\right) = \mathbb{E}[\tilde{\psi}(\tilde{z}_1, \ldots \tilde{z}_t, \tilde{x}_1, \ldots, \tilde{x}_{t+1}, X)]. \tag{40}$$

Equivalently, as $N \to \infty$*, almost surely:*

$$(\tilde{\boldsymbol{z}}^1, \ldots, \tilde{\boldsymbol{z}}^t, \tilde{\boldsymbol{x}}^1, \ldots, \tilde{\boldsymbol{x}}^{t+1}, \boldsymbol{X}) \xrightarrow{W_2} (\tilde{z}_1, \ldots, \tilde{z}_t, \tilde{x}_1, \ldots, \tilde{x}_{t+1}, X).$$

Proof. The result follows from Theorem 2.3 in [87]. In fact, Assumption 2.1 of [87] holds because of the model assumptions on $\boldsymbol{Z}$, Assumption 2.2(a) holds because $(\boldsymbol{X}, \hat{\boldsymbol{x}}^1) \xrightarrow{W_2} (X, \hat{x}_1)$ from (7), and Assumption 2.2(b) follows from the definition of $\tilde{h}_{t+1}$ in (34) and our Assumption 3. As the auxiliary AMP in (33) is of the standard form for which the state evolution result of Theorem 2.3 in [87] holds, the proof is complete. □

C.3 Proof of Theorem 2

We start by presenting a useful technical lemma.

Lemma 3. *Let* $F : \mathbb{R}^t \to \mathbb{R}$ *be a Lipschitz function, and let* $\partial_k F$ *denote its derivative with respect to the* k*-th argument, for* $1 \leq k \leq t$*. Assume that* $\partial_k F$ *is continuous almost everywhere in the* k*-th argument, for each* k*. Let* $(V_1^{(m)}, \ldots, V_t^{(m)})$ *be a sequence of random vectors in* $\mathbb{R}^t$ *converging in distribution to the random vector* $(V_1, \ldots, V_t)$ *as* $m \to \infty$*. Furthermore, assume that the distribution of* $(V_1, \ldots, V_t)$ *is absolutely continuous with respect to the Lebesgue measure. Then,*

$$\lim_{m\to\infty} \mathbb{E}[\partial_k F(V_1^{(m)}, \ldots, V_t^{(m)})] = \mathbb{E}[\partial_k F(V_1, \ldots, V_t)], \qquad 1 \leq k \leq t. \tag{41}$$

The result was proved for $t = 2$ in [18, Lemma 6]. The proof for $t > 2$ is essentially the same; see also [39, Lemma 7.14].

At this point, we are ready to give the proof of Theorem 2.

Proof of Theorem 2. The first step is to show the equivalence between the state evolution for the true AMP and the corresponding one for the auxiliary AMP. In particular, we prove that, for $t \geq 1$,

$$(\tilde{z}_1, \ldots, \tilde{z}_t) \overset{\mathrm{d}}{=} (z_1, \ldots, z_t), \tag{42}$$

where the random variables on the left are defined in (37), and the random variables on the right are defined in (14). As $(\tilde{z}_1, \ldots, \tilde{z}_t) \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_t)$ and $(z_1, \ldots, z_t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$, (42) is readily implied by

$$\tilde{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t, \qquad \text{for all } t \geq 1. \tag{43}$$

We now show that (43) holds by induction. The base case ($t = 1$) follows from the definitions of $\tilde{\boldsymbol{\Sigma}}_1, \boldsymbol{\Sigma}_1$. Towards induction, assume that $\tilde{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t$ for some $t \geq 1$. By comparing the definition of $\hat{x}_s$ in (14) with the definition of $\tilde{x}_s$ in (37) and by using the choice of $\tilde{h}_s$ in (34) (for $s \in \{2, \ldots, t+1\}$), we readily obtain that $\Delta_{t+1} = \tilde{\Delta}_{t+1}$ and $\bar{\boldsymbol{\Phi}}_{t+1} = \tilde{\bar{\boldsymbol{\Phi}}}_{t+1}$. Hence, by using (17) and (39), we have that $\tilde{\boldsymbol{\Sigma}}_{t+1} = \boldsymbol{\Sigma}_{t+1}$ and the proof of (43) is complete.

The second step is to show that, for any pseudo-Lipschitz function $\psi : \mathbb{R}^{2t+2} \to \mathbb{R}$ of order 2, the following limit holds almost surely for $t \geq 1$:

$$\begin{aligned} \lim_{N\to\infty} \Bigg| &\frac{1}{N} \sum_{i=1}^{N} \psi\left((\boldsymbol{x}^1)_i, \ldots, (\boldsymbol{x}^t)_i, (\hat{\boldsymbol{x}}^1)_i, \ldots, (\hat{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i\right) \\ &- \frac{1}{N} \sum_{i=1}^{N} \psi\left((\boldsymbol{u}^1)_i, \ldots, (\boldsymbol{u}^t)_i, (\tilde{\boldsymbol{x}}^1)_i, \ldots, (\tilde{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i\right) \Bigg| = 0, \end{aligned} \tag{44}$$

where, we have defined for $s \in \{1, \dots, t\}$,

$$\boldsymbol{u}^s = \tilde{\boldsymbol{z}}^s + (\boldsymbol{B}_s)_{s,1}\tilde{\boldsymbol{x}}^1 + \sum_{i=2}^{s}(\boldsymbol{B}_s)_{s,i}\tilde{\boldsymbol{x}}^i + \mu_s \boldsymbol{X} - \bar{\beta}_s \tilde{\boldsymbol{x}}^{s-1}. \tag{45}$$

From here till the end of the argument, all the limits hold almost surely, and we use C to denote a generic positive constant, which can change from line to line and is independent of N. By using that ψ is pseudo-Lipschitz, we have that

$$\begin{aligned}
&\Bigg| \frac{1}{N}\sum_{i=1}^{N} \psi\left((\boldsymbol{x}^1)_i, \dots, (\boldsymbol{x}^t)_i, (\hat{\boldsymbol{x}}^1)_i, \dots, (\hat{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i\right) \\
&\qquad - \frac{1}{N}\sum_{i=1}^{N} \psi\left((\boldsymbol{u}^1)_i, \dots, (\boldsymbol{u}^t)_i, (\tilde{\boldsymbol{x}}^1)_i, \dots, (\tilde{\boldsymbol{x}}^{t+1})_i, (\boldsymbol{X})_i\right) \Bigg| \\
&\le \frac{C}{N}\sum_{i=1}^{N}\Bigg(1 + |(\boldsymbol{X})_i| + 2|(\hat{\boldsymbol{x}}^1)_i| + \sum_{\ell=1}^{t}\Big(|(\boldsymbol{x}^\ell)_i| + |(\boldsymbol{u}^\ell)_i| + |(\hat{\boldsymbol{x}}^{\ell+1})_i| + |(\tilde{\boldsymbol{x}}^{\ell+1})_i|\Big)\Bigg) \\
&\qquad \cdot \Bigg(\sum_{\ell=1}^{t}\Big(|(\boldsymbol{x}^\ell)_i - (\boldsymbol{u}^\ell)_i|^2 + |(\hat{\boldsymbol{x}}^{\ell+1})_i - (\tilde{\boldsymbol{x}}^{\ell+1})_i|^2\Big)\Bigg)^{1/2} \\
&\le C(4t+3)\Bigg[1 + \frac{\|\boldsymbol{X}\|^2}{N} + \sum_{\ell=1}^{t}\Big(\frac{\|\boldsymbol{x}^\ell\|^2}{N} + \frac{\|\boldsymbol{u}^\ell\|^2}{N} + \frac{\|\hat{\boldsymbol{x}}^{\ell+1}\|^2}{N} + \frac{\|\tilde{\boldsymbol{x}}^{\ell+1}\|^2}{N}\Big)\Bigg]^{1/2} \\
&\qquad \cdot \Bigg(\sum_{\ell=1}^{t}\Bigg(\frac{\|\boldsymbol{x}^\ell - \boldsymbol{u}^\ell\|^2}{N} + \frac{\|\hat{\boldsymbol{x}}^{\ell+1} - \tilde{\boldsymbol{x}}^{\ell+1}\|^2}{N}\Bigg)\Bigg)^{1/2},
\end{aligned} \tag{46}$$

where the last step uses twice Cauchy-Schwarz inequality. We now inductively show that as $N \to \infty$: *(i)* each of the terms in the last line of (46) converges to zero, and *(ii)* the terms within the square brackets in (46) all converge to finite, deterministic limits.

Base case ($t = 1$). We have that

$$\begin{aligned}
\boldsymbol{x}^1 - \boldsymbol{u}^1 &= \boldsymbol{Y}\hat{\boldsymbol{x}}^1 - \tilde{\boldsymbol{z}}^1 - (\boldsymbol{B}_1)_{1,1}\tilde{\boldsymbol{x}}^1 - \mu_1 \boldsymbol{X} \\
&= \boldsymbol{Z}\hat{\boldsymbol{x}}^1 + \sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^1\rangle}{N}\boldsymbol{X} - \boldsymbol{Z}\tilde{\boldsymbol{x}}^1 + \mathsf{b}_{1,1}\tilde{\boldsymbol{x}}^1 - (\boldsymbol{B}_1)_{1,1}\tilde{\boldsymbol{x}}^1 - \mu_1\boldsymbol{X} \\
&= \sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^1\rangle}{N}\boldsymbol{X} + \mathsf{b}_{1,1}\tilde{\boldsymbol{x}}^1 - (\boldsymbol{B}_1)_{1,1}\tilde{\boldsymbol{x}}^1 - \mu_1\boldsymbol{X},
\end{aligned} \tag{47}$$

where the first equality uses (8) and (45), the second equality uses (1) and (33), and the third equality uses that $\tilde{\boldsymbol{x}}^1 = \hat{\boldsymbol{x}}^1$. Hence, by triangle inequality,

$$\begin{aligned}
\frac{\|\boldsymbol{x}^1 - \boldsymbol{u}^1\|^2}{N} &\le 2\left(\sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^1\rangle}{N} - \mu_1\right)^2 \frac{\|\boldsymbol{X}\|^2}{N} + 2\left(\mathsf{b}_{1,1} - (\boldsymbol{B}_1)_{1,1}\right)^2 \frac{\|\tilde{\boldsymbol{x}}^1\|^2}{N} \\
&\le C\left(\left(\sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^1\rangle}{N} - \mu_1\right)^2 + \left(\mathsf{b}_{1,1} - (\boldsymbol{B}_1)_{1,1}\right)^2\right),
\end{aligned} \tag{48}$$

where the last inequality uses that $\tilde{\boldsymbol{x}}^1 = \hat{\boldsymbol{x}}^1$ and that $(\boldsymbol{X}, \hat{\boldsymbol{x}}^1)$ converge in W_2 to random variables with finite second moments. By using (7) and recalling that $\mu_1 = \sqrt{\lambda_*}\epsilon$ (cf. (13)), we have

$$\lim_{N\to\infty} \sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^1\rangle}{N} = \sqrt{\lambda_*}\epsilon = \mu_1. \tag{49}$$

Furthermore, note that $(\boldsymbol{B}_1)_{1,1} = \bar{\kappa}_1$ (cf. (13)) and $\mathsf{b}_{1,1} = \kappa_1$ (cf. (36)). Hence, by Assumption 1, $\kappa_1 \to \bar{\kappa}_1$, as $N \to \infty$. Hence, $\mathsf{b}_{1,1} \to (\boldsymbol{B}_1)_{1,1}$ and, by combining this observation with (48) and (49), we obtain that

$$\lim_{N\to\infty} \frac{\|\boldsymbol{x}^1 - \boldsymbol{u}^1\|^2}{N} = 0. \tag{50}$$

Next, by expressing $\hat{\boldsymbol{x}}^2$ via (8) and $\tilde{\boldsymbol{x}}^2$ via (33), (34) and (45), we have that

$$\hat{\boldsymbol{x}}^2 - \tilde{\boldsymbol{x}}^2 = h_2(\boldsymbol{x}^1) - h_2(\boldsymbol{u}^1). \tag{51}$$

Thus, as h_2 is Lipschitz, (50) immediately implies that

$$\lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^2 - \tilde{\boldsymbol{x}}^2\|^2}{N} = 0. \tag{52}$$

An application of the triangle inequality gives that, for any $i \geq 1$,

$$\begin{aligned} \|\boldsymbol{u}^i\| - \|\boldsymbol{x}^i - \boldsymbol{u}^i\| \leq \|\boldsymbol{x}^i\| \leq \|\boldsymbol{u}^i\| + \|\boldsymbol{x}^i - \boldsymbol{u}^i\|, \\ \|\tilde{\boldsymbol{x}}^{i+1}\| - \|\hat{\boldsymbol{x}}^{i+1} - \tilde{\boldsymbol{x}}^{i+1}\| \leq \|\hat{\boldsymbol{x}}^{i+1}\| \leq \|\tilde{\boldsymbol{x}}^{i+1}\| + \|\hat{\boldsymbol{x}}^{i+1} - \tilde{\boldsymbol{x}}^{i+1}\|. \end{aligned} \tag{53}$$

Thus, by using (53) with $i = 1$ and Proposition 2, we obtain that

$$\begin{aligned} \lim_{N\to\infty} \frac{\|\boldsymbol{x}^1\|^2}{N} &= \lim_{N\to\infty} \frac{\|\boldsymbol{u}^1\|^2}{N} = \mathbb{E}[(\tilde{z}_1 + (\boldsymbol{B}_1)_{1,1}\hat{x}_1 + \mu_1 X)^2], \\ \lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^2\|^2}{N} &= \lim_{N\to\infty} \frac{\|\tilde{\boldsymbol{x}}^2\|^2}{N} = \mathbb{E}[(\tilde{x}_2)^2], \end{aligned} \tag{54}$$

which concludes the base step.

Induction step. Assume towards induction that

$$\lim_{N\to\infty} \frac{\|\boldsymbol{x}^j - \boldsymbol{u}^j\|^2}{N} = 0, \quad \text{for } j \in \{1, \dots, t\}, \tag{55}$$

$$\lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^j - \tilde{\boldsymbol{x}}^j\|^2}{N} = 0, \quad \text{for } j \in \{2, \dots, t+1\}, \tag{56}$$

$$\begin{aligned} \lim_{N\to\infty} \frac{\|\boldsymbol{x}^j\|^2}{N} &= \lim_{N\to\infty} \frac{\|\boldsymbol{u}^j\|^2}{N} \\ &= \mathbb{E}\left[\left(\tilde{z}_j + (\boldsymbol{B}_j)_{j,1}\hat{x}_1 + \sum_{i=2}^{j} (\boldsymbol{B}_j)_{j,i}\tilde{x}_i + \mu_j X - \bar{\beta}_j \tilde{x}_{j-1}\right)^2\right], \quad \text{for } j \in \{1, \dots, t\}, \end{aligned} \tag{57}$$

$$\lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^j\|^2}{N} = \lim_{N\to\infty} \frac{\|\tilde{\boldsymbol{x}}^j\|^2}{N} = \mathbb{E}[\tilde{x}_j^2], \quad \text{for } j \in \{2, \dots, t+1\}. \tag{58}$$

We now show that the following limits hold:

$$\lim_{N\to\infty} \frac{\|\boldsymbol{x}^{t+1} - \boldsymbol{u}^{t+1}\|^2}{N} = 0, \tag{59}$$

$$\lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^{t+2} - \tilde{\boldsymbol{x}}^{t+2}\|^2}{N} = 0, \tag{60}$$

$$\begin{aligned} \lim_{N\to\infty} \frac{\|\boldsymbol{x}^{t+1}\|^2}{N} &= \lim_{N\to\infty} \frac{\|\boldsymbol{u}^{t+1}\|^2}{N} \\ &= \mathbb{E}\left[\left(\tilde{z}_{t+1} + (\boldsymbol{B}_{t+1})_{t+1,1}\hat{x}_1 + \sum_{i=2}^{t+1} (\boldsymbol{B}_{t+1})_{t+1,i}\tilde{x}_i + \mu_{t+1} X - \bar{\beta}_{t+1}\tilde{x}_t\right)^2\right], \end{aligned} \tag{61}$$

$$\lim_{N\to\infty} \frac{\|\hat{\boldsymbol{x}}^{t+2}\|^2}{N} = \lim_{N\to\infty} \frac{\|\tilde{\boldsymbol{x}}^{t+2}\|^2}{N} = \mathbb{E}[\tilde{x}_{t+2}^2]. \tag{62}$$

By doing so, we will have proved also the induction step and consequently that (44) holds.

Using similar passages as in (47), we obtain

$$\begin{aligned} \boldsymbol{x}^{t+1} - \boldsymbol{u}^{t+1} &= \boldsymbol{Y}\hat{\boldsymbol{x}}^{t+1} - \beta_{t+1}\hat{\boldsymbol{x}}^t - \tilde{\boldsymbol{z}}^{t+1} - (\boldsymbol{B}_{t+1})_{t+1,1}\tilde{\boldsymbol{x}}^1 - \sum_{i=2}^{t+1} (\boldsymbol{B}_{t+1})_{t+1,i}\tilde{\boldsymbol{x}}^i - \mu_{t+1}\boldsymbol{X} + \bar{\beta}_{t+1}\tilde{\boldsymbol{x}}^t \\ &= \boldsymbol{Z}\hat{\boldsymbol{x}}^{t+1} + \sqrt{\lambda_*}\frac{\langle \boldsymbol{X}, \hat{\boldsymbol{x}}^{t+1}\rangle}{N}\boldsymbol{X} - \beta_{t+1}\hat{\boldsymbol{x}}^t - \boldsymbol{Z}\tilde{\boldsymbol{x}}^{t+1} + \sum_{i=1}^{t+1} \mathsf{b}_{t+1,i}\tilde{\boldsymbol{x}}^i - (\boldsymbol{B}_{t+1})_{t+1,1}\tilde{\boldsymbol{x}}^1 \\ &\quad - \sum_{i=2}^{t+1} (\boldsymbol{B}_{t+1})_{t+1,i}\tilde{\boldsymbol{x}}^i - \mu_{t+1}\boldsymbol{X} + \bar{\beta}_{t+1}\tilde{\boldsymbol{x}}^t. \end{aligned} \tag{63}$$

Hence, by triangle inequality,

$$\begin{aligned}\frac{\|\boldsymbol{x}^{t+1}-\boldsymbol{u}^{t+1}\|^2}{N} \le C\Bigg(&\frac{\|\boldsymbol{Z}\hat{\boldsymbol{x}}^{t+1}-\boldsymbol{Z}\tilde{\boldsymbol{x}}^{t+1}\|^2}{N}+\left(\sqrt{\lambda_*}\frac{\langle \boldsymbol{X},\hat{\boldsymbol{x}}^{t+1}\rangle}{N}-\mu_{t+1}\right)^2\frac{\|\boldsymbol{X}\|^2}{N}\\ &+\sum_{i=1}^{t+1}\left(\mathsf{b}_{t+1,i}-(\boldsymbol{B}_{t+1})_{t+1,i}\right)^2\frac{\|\tilde{\boldsymbol{x}}^i\|^2}{N}+(\bar{\beta}_{t+1}-\beta_{t+1})^2\frac{\|\hat{\boldsymbol{x}}^t\|^2}{N}+(\bar{\beta}_{t+1})^2\frac{\|\tilde{\boldsymbol{x}}^t-\hat{\boldsymbol{x}}^t\|^2}{N}\Bigg)\\ &:= C(T_1+T_2+T_3+T_4+T_5).\end{aligned} \tag{64}$$

Consider the first term. Since $\|\boldsymbol{Z}\|_{\mathrm{op}} \le C$, the induction hypothesis (56) implies that $T_1 \to 0$ as $N \to \infty$.

Consider the second term. The following chain of equalities holds:

$$\lim_{N\to\infty}\sqrt{\lambda_*}\frac{\langle \boldsymbol{X},\hat{\boldsymbol{x}}^{t+1}\rangle}{N}=\lim_{N\to\infty}\sqrt{\lambda_*}\frac{\langle \boldsymbol{X},\tilde{\boldsymbol{x}}^{t+1}\rangle}{N}=\sqrt{\lambda_*}\mathbb{E}[X\,\tilde{x}_{t+1}]=\sqrt{\lambda_*}\mathbb{E}[X\,\hat{x}_{t+1}]=\mu_{t+1}. \tag{65}$$

Here, the first equality uses (56) together with the fact that $\|\boldsymbol{X}\|^2/N = 1$; the second equality follows from Proposition 2; the third equality uses (42) and the definitions of $\hat{x}_{t+1}$ and $\tilde{x}_{t+1}$ in (14) and (37), respectively; and the fourth equality uses the definition of μ_{t+1} in (15). Finally, using (65) and again that $\|\boldsymbol{X}\|^2/N = 1$ gives that $T_2 \to 0$ as $N \to \infty$.

Consider the third term. The following chain of equalities holds, for $1 \le j < i \le (t+1)$,

$$\lim_{N\to\infty}(\boldsymbol{\Phi}_{t+1})_{i,j}=\lim_{N\to\infty}\langle \partial_j\tilde{\boldsymbol{x}}^i\rangle=\mathbb{E}[\partial_j\tilde{x}_i]=\mathbb{E}[\partial_j\hat{x}_i]=(\bar{\boldsymbol{\Phi}}_{t+1})_{i,j} \tag{66}$$

Here, the first equality uses the definition (35); the second equality follows from Lemma 3, as $\tilde{\boldsymbol{x}}^i = \tilde{h}_i(\tilde{\boldsymbol{z}}^1,\dots,\tilde{\boldsymbol{z}}^{i-1},\hat{\boldsymbol{x}}^1,\boldsymbol{X})$ converges in W_2 (and therefore in distribution) to $\tilde{x}_i = \tilde{h}_i(\tilde{z}_1,\dots,\tilde{z}_{i-1},\hat{x}_1,X)$ and h_i satisfies Assumption 3; the third equality uses (42) and the definitions of $\hat{x}_i$ and $\tilde{x}_i$ in (14) and (37), respectively; and the fourth equality uses the definition of $(\bar{\boldsymbol{\Phi}}_{t+1})_{i,j}$ in (16). By Assumption 1, as $N \to \infty$, $\kappa_j \to \bar{\kappa}_j$ for all j. Thus, by combining (66) with the definitions of $\boldsymbol{B}_{t+1}$ and $\tilde{\boldsymbol{B}}_{t+1}$ in (16) and (36), respectively, we conclude that, as $N \to \infty$, $\mathsf{b}_{t+1,i} \to (\boldsymbol{B}_{t+1})_{t+1,i}$ for $i \in \{1,\dots,t+1\}$. By using the induction hypothesis (58), which shows that $\|\tilde{\boldsymbol{x}}^i\|^2/N$ converges to a finite limit, we conclude that $T_3 \to 0$ as $N \to \infty$.

Consider the fourth term. By using the induction hypothesis (55) and (57), together with (46), we obtain that $\boldsymbol{x}^t$ and $\boldsymbol{u}^t$ have the same W_2 limit given by Proposition 2, namely,

$$\boldsymbol{x}^t \xrightarrow{W_2} \tilde{z}_t+(\boldsymbol{B}_t)_{t,1}\hat{x}_1+\sum_{i=2}^{t}(\boldsymbol{B}_t)_{t,i}\tilde{x}_i+\mu_t X-\bar{\beta}_t\tilde{x}_{t-1}.$$

Thus, by recalling that $\beta_{t+1} = \langle h'_{t+1}(\boldsymbol{x}^t)\rangle$, an application of Lemma 3 gives that

$$\lim_{N\to\infty}\beta_{t+1}=\mathbb{E}\left[h'_{t+1}\left(\tilde{z}_t+(\boldsymbol{B}_t)_{t,1}\hat{x}_1+\sum_{i=2}^{t}(\boldsymbol{B}_t)_{t,i}\tilde{x}_i+\mu_t X-\bar{\beta}_t\tilde{x}_{t-1}\right)\right]. \tag{67}$$

Furthermore, by using (42) and recalling the definition of $\bar{\beta}_{t+1}$, we have that the RHS of (67) is equal to $\bar{\beta}_{t+1}$. Hence, by using the induction hypothesis (58), we obtain that $T_4 \to 0$ as $N \to \infty$. Finally, by using the induction hypothesis (56), we conclude that also $T_5 \to 0$ as $N \to \infty$.

As $T_i \to 0$ for $i \in \{1,\dots,5\}$, (64) implies that (59) holds. Next, as h_{t+1} is Lipschitz, (59) immediately implies (60). Then, by using (53) with $i = t+1$ and Proposition 2, we obtain that (61) and (62) hold, thus concluding the inductive proof. The result we have just proved by induction, combined with (46), gives that (44) holds.

Another application of Proposition 2, together with (44), gives that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi\left((\boldsymbol{x}^1)_i,\dots,(\boldsymbol{x}^t)_i,\,(\hat{\boldsymbol{x}}^1)_i,\dots,(\hat{\boldsymbol{x}}^{t+1})_i,\,(\boldsymbol{X})_i\right)=\mathbb{E}[\psi(u_1,\dots,u_t,\tilde{x}_1,\dots,\tilde{x}_{t+1},X)], \tag{68}$$

where we have defined for $s \in \{1,\dots,t\}$,

$$u_s=\tilde{z}_s+(\boldsymbol{B}_s)_{s,1}\tilde{x}_1+\sum_{i=2}^{s}(\boldsymbol{B}_s)_{s,i}\tilde{x}_i+\mu_s X-\bar{\beta}_s\tilde{x}_{s-1}. \tag{69}$$

Finally, by using (42), we have that

$$\mathbb{E}[\psi(u_1,\dots,u_t,\tilde{x}_1,\dots,\tilde{x}_{t+1},X)] = \mathbb{E}[\psi(x_1,\dots,x_t,\hat{x}_1,\dots,\hat{x}_{t+1},X)]. \tag{70}$$

By combining (68) and (70), we obtain that the desired result (18) holds, which concludes the proof. □

C.4 State evolution for Gaussian AMP with spectral initialization

In this appendix, we consider the Gaussian AMP iteration (8) with spectral initialization $\hat{\boldsymbol{x}}_1 = \overline{\boldsymbol{v}}_N$, where $\overline{\boldsymbol{v}}_N$ denotes the eigenvector of the data matrix $\boldsymbol{Y}$ associated to the largest eigenvalue $\overline{\nu}_N$. As for the Gaussian AMP previously analyzed, one can show that the joint empirical distribution of $(\boldsymbol{x}^1,\dots,\boldsymbol{x}^t,\hat{\boldsymbol{x}}^1,\dots,\hat{\boldsymbol{x}}^{t+1},\boldsymbol{X})$ converges (in W_2 distance) to the random vector $(x_1,\dots,x_t,\hat{x}_1,\dots,\hat{x}_{t+1},X)$. The law of this random vector can be captured via a state evolution recursion, which is expressed via a sequence of vectors $\boldsymbol{\mu}_t^{\mathrm{si}} = (\mu_0,\mu_1,\dots,\mu_t)$ and matrices $\boldsymbol{\Sigma}_t^{\mathrm{si}}, \boldsymbol{\Delta}_t^{\mathrm{si}}, \boldsymbol{B}_t^{\mathrm{si}} \in \mathbb{R}^{(t+1)\times(t+1)}$, defined recursively as follows. Here, the superscript si highlights that the state evolution refers to an AMP with *spectral initialization*. We also note that, to ease the notation, we have shifted the initialization index from 1 to 0.

We start with the initialization

$$\mu_0^{\mathrm{si}} = \sqrt{C_{\mathrm{OS}}}, \quad \boldsymbol{\Sigma}_0^{\mathrm{si}} = 1 - C_{\mathrm{OS}}, \quad \boldsymbol{\Delta}_0^{\mathrm{si}} = 1/\lambda_*, \quad \boldsymbol{B}_0^{\mathrm{si}} = R_\rho\left(\frac{1}{\sqrt{\lambda_*}}\right), \tag{71}$$

where λ_* is the true SNR (see (1)), and $R_\rho(\cdot)$ denotes the R-transform of ρ. Inductively, having defined $(\boldsymbol{\mu}_{t-1}^{\mathrm{si}}, \boldsymbol{\Sigma}_{t-1}^{\mathrm{si}}, \boldsymbol{\Delta}_{t-1}^{\mathrm{si}}, \boldsymbol{B}_{t-1}^{\mathrm{si}})$, we define the following joint law

$$\begin{aligned}
&(z_0,,\dots,z_{t-1}) \mid X \sim \mathcal{N}(\boldsymbol{\mu}_{t-1}^{\mathrm{si}}\cdot X, \boldsymbol{\Sigma}_{t-1}^{\mathrm{si}}), \quad \hat{x}_0 = z_0/\sqrt{\lambda_*}, \quad \hat{x}_1 = z_0,\\
&x_s = z_s - \bar{\beta}_s \hat{x}_{s-1} + \sum_{i=0}^{s} (\boldsymbol{B}_s^{\mathrm{si}})_{s,i}\hat{x}_i, \quad \hat{x}_{s+1} = h_{s+1}(x_s), \text{ for } s = 1,\dots,t-1,
\end{aligned} \tag{72}$$

where we have set $\bar{\beta}_1 = 0$, and for $t \geq 2$, $\bar{\beta}_t = \mathbb{E}[h_t'(x_{t-1})]$. Then, the parameter μ_t^{si} is

$$\mu_t^{\mathrm{si}} = \sqrt{\lambda_*}\mathbb{E}[X\,\hat{x}_t]. \tag{73}$$

Next, we decompose the second moment matrix $\boldsymbol{\Delta}_t^{\mathrm{si}}$ of $(\hat{x}_0,\dots,\hat{x}_t)$ into four parts:

$$\boldsymbol{\Delta}_t^{\mathrm{si}} = \bar{\boldsymbol{\Delta}}_t^{\mathrm{si}} + \tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}} + (\tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}})^\top + \hat{\boldsymbol{\Delta}}_t^{\mathrm{si}}, \tag{74}$$

where $\bar{\boldsymbol{\Delta}}_t^{\mathrm{si}}, \tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}}, \hat{\boldsymbol{\Delta}}_t^{\mathrm{si}} \in \mathbb{R}^{(t+1)\times(t+1)}$ are defined as follows:

$$\begin{aligned}
&(\bar{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = \mathbb{E}[\hat{x}_{i-1}\,\hat{x}_{j-1}] \text{ for } i \in \{2,\dots,t+1\}, j \in \{2,\dots,t+1\}, \text{ and } (\bar{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = 0 \text{ otherwise},\\
&(\tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = \mathbb{E}[\hat{x}_{i-1}\,\hat{x}_{j-1}] \text{ for } i = 1, j \in \{2,\dots,t+1\} \text{ and } (\tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = 0 \text{ otherwise},\\
&(\hat{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = \mathbb{E}[\hat{x}_{i-1}\,\hat{x}_{j-1}] \text{ for } i = j = 1, \text{ and } (\hat{\boldsymbol{\Delta}}_t^{\mathrm{si}})_{i,j} = 0 \text{ otherwise}.
\end{aligned} \tag{75}$$

Then, let us define the matrix $\bar{\boldsymbol{\Phi}}_t^{\mathrm{si}} \in \mathbb{R}^{(t+1)\times(t+1)}$ as

$$(\bar{\boldsymbol{\Phi}}_t^{\mathrm{si}})_{i,j} = 0 \text{ if } i \leq j, \text{ and } (\bar{\boldsymbol{\Phi}}_t)_{i,j} = \mathbb{E}[\partial_j \hat{x}_{i-1}] \text{ if } i > j. \tag{76}$$

At this point, we define the two matrices $\boldsymbol{B}_t^{\mathrm{aux}}, \tilde{\boldsymbol{B}}_t^{\mathrm{aux}} \in \mathbb{R}^{(t+1)\times(t+1)}$ as

$$\boldsymbol{B}_t^{\mathrm{aux}} = \sum_{j=0}^{\infty} \bar{\kappa}_{j+1}\bar{\boldsymbol{\Phi}}_t^j, \qquad \tilde{\boldsymbol{B}}_t^{\mathrm{aux}} = \sum_{j=0}^{\infty} \tilde{\kappa}_{j+1}\bar{\boldsymbol{\Phi}}_t^j, \tag{77}$$

where

$$\tilde{\kappa}_s = \sum_{j=0}^{\infty} \bar{\kappa}_{j+s}\left(\frac{1}{\sqrt{\lambda_*}}\right)^j, \qquad \text{for } s \geq 1. \tag{78}$$

Then, the first column of $\boldsymbol{B}_t^{\mathrm{si}}$ consists in the first column of $\tilde{\boldsymbol{B}}_t^{\mathrm{aux}}$, and the remaining columns of $\boldsymbol{B}_t^{\mathrm{si}}$ are equal to the corresponding columns of $\boldsymbol{B}_t^{\mathrm{aux}}$.

Finally, we define the covariance matrix $\boldsymbol{\Sigma}_t^{\mathrm{si}}$ as

$$\boldsymbol{\Sigma}_t^{\mathrm{si}} = \sum_{j=0}^{\infty}\sum_{i=0}^{j}(\bar{\boldsymbol{\Phi}}_t^{\mathrm{si}})^i(\bar{\kappa}_{j+2}\bar{\boldsymbol{\Delta}}_t^{\mathrm{si}} + \tilde{\kappa}_{j+2}\tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}} + \tilde{\kappa}_{j+2}(\tilde{\boldsymbol{\Delta}}_t^{\mathrm{si}})^\top + \hat{\kappa}_{j+2}\hat{\boldsymbol{\Delta}}_t^{\mathrm{si}})((\bar{\boldsymbol{\Phi}}_t^{\mathrm{si}})^\top)^{j-i}, \tag{79}$$

where we have set

$$\hat{\kappa}_s = \sum_{j=0}^{\infty}(j+1)\bar{\kappa}_{j+s}\left(\frac{1}{\sqrt{\lambda_*}}\right)^j, \qquad \text{for } s \geq 1. \tag{80}$$

Having defined the state evolution recursion above, we can prove that (18) holds, where we recall that the AMP algorithm is initialized with the spectral estimate $\hat{\boldsymbol{x}}_1 = \overline{\boldsymbol{v}}_N$. The proof follows steps similar to those detailed in Appendices C.1-C.3 to show Theorem 2. In this case, the auxiliary AMP is given by the AMP with spectral initialization described in Section 3.2 of [87], and the denoisers[4] u_t are taken to be the following:

$$\begin{aligned} u_1(z_0) &= z_0, \\ u_{t+1}(z_0,\dots,z_t) &= h_{t+1}\left(z_t + \sum_{i=0}^{t}(\boldsymbol{B}_t^{\mathrm{si}})_{t,i}u_i(z_0,\dots,z_{i-1}) - \bar{\beta}_t u_{t-1}(z_0,\dots,z_{t-2})\right), \end{aligned} \tag{81}$$

where the matrix $\boldsymbol{B}_t^{\mathrm{si}}$ is obtained via the state evolution recursion described above.

D Implementation details and additional numerical results

D.1 Correct AMP: algorithm and corresponding state evolution

In our experiments, for both the correct and Gaussian AMP, we assume to have access to an initialization $\hat{\boldsymbol{x}}^1 \in \mathbb{R}^N$ s.t. (7) holds. Then, for $t \geq 1$, the *correct AMP* iteration reads

$$\boldsymbol{x}_{\mathrm{c}}^t = \boldsymbol{Y}\hat{\boldsymbol{x}}_{\mathrm{c}}^t - \sum_{i=1}^{t}\mathsf{b}_{t,i}^{\mathrm{c}}\hat{\boldsymbol{x}}_{\mathrm{c}}^i, \quad \hat{\boldsymbol{x}}_{\mathrm{c}}^{t+1} = h_{t+1}(\boldsymbol{x}_{\mathrm{c}}^t,\dots,\boldsymbol{x}_{\mathrm{c}}^1). \tag{82}$$

To obtain the coefficients $\{\mathsf{b}_{t,i}^{\mathrm{c}}\}_{i=1}^t$, we define the matrix $\boldsymbol{\Phi}_t^{\mathrm{c}} \in \mathbb{R}^{t\times t}$ as

$$(\boldsymbol{\Phi}_t^{\mathrm{c}})_{i,j} = 0, \quad \text{for } i \leq j, \qquad (\boldsymbol{\Phi}_t^{\mathrm{c}})_{i,j} = \langle \partial_j \hat{\boldsymbol{x}}_{\mathrm{c}}^i\rangle, \quad \text{for } i > j, \tag{83}$$

where, for $j < i$, the vector $\langle \partial_j \hat{\boldsymbol{x}}_{\mathrm{c}}^i\rangle \in \mathbb{R}^N$ denotes the partial derivative of $h_i : \mathbb{R}^{i-1} \to \mathbb{R}$ with respect to the j-th input (applied component-wise). Then, the vector $(\mathsf{b}_{t,1}^{\mathrm{c}},\dots,\mathsf{b}_{t,t}^{\mathrm{c}})$ is given by the last row of the matrix $\boldsymbol{B}_t^{\mathrm{c}} \in \mathbb{R}^{t\times t}$ defined as

$$\boldsymbol{B}_t^{\mathrm{c}} = \sum_{j=0}^{t-1}\kappa_{j+1}(\boldsymbol{\Phi}_t^{\mathrm{c}})^j, \tag{84}$$

where $\{\kappa_k\}_{k\geq 1}$ denotes the sequence of free cumulants associated to the matrix $\boldsymbol{Y}$. By using the results of [37, 87] (e.g., Theorem 2.3 in [87]), one can obtain a state evolution result for the correct AMP (82). More specifically, we have that

$$(\boldsymbol{x}_{\mathrm{c}}^1,\dots,\boldsymbol{x}_{\mathrm{c}}^t,\, \hat{\boldsymbol{x}}_{\mathrm{c}}^1,\dots,\hat{\boldsymbol{x}}_{\mathrm{c}}^{t+1},\, \boldsymbol{X}) \xrightarrow{W_2} (x_1^{\mathrm{c}},\dots,x_t^{\mathrm{c}},\, \hat{x}_1^{\mathrm{c}},\dots,\hat{x}_{t+1}^{\mathrm{c}},\, X). \tag{85}$$

The law of the random vector $(x_1^{\mathrm{c}},\dots,x_t^{\mathrm{c}},\, \hat{x}_1^{\mathrm{c}},\dots,\hat{x}_{t+1}^{\mathrm{c}})$ is expressed via a sequence of vectors $\boldsymbol{\mu}_t^{\mathrm{c}} = (\mu_1^{\mathrm{c}},\dots,\mu_t^{\mathrm{c}})$ and matrices $\boldsymbol{\Sigma}_t^{\mathrm{c}}, \boldsymbol{\Delta}_t^{\mathrm{c}} \in \mathbb{R}^{t\times t}$ defined recursively as follows. We start with the initialization

$$\mu_1^{\mathrm{c}} = \sqrt{\lambda_*}\epsilon, \quad \boldsymbol{\Sigma}_1^{\mathrm{c}} = \bar{\kappa}_2\mathbb{E}[\hat{x}_1^2], \quad \boldsymbol{\Delta}_1^{\mathrm{c}} = \mathbb{E}[\hat{x}_1^2], \tag{86}$$

where λ_* is the SNR (see (1)), ϵ is given in (7), and $\{\bar{\kappa}_k\}_{k\geq 1}$ are the free cumulants associated to the asymptotic spectral measure of the noise ρ. For $t \geq 1$, given $\boldsymbol{\mu}_t^{\mathrm{c}}, \boldsymbol{\Sigma}_t^{\mathrm{c}}, \boldsymbol{\Delta}_t^{\mathrm{c}}$, let

$$\begin{aligned} &(x_1^{\mathrm{c}},\dots,x_t^{\mathrm{c}}) = (\mu_1^{\mathrm{c}},\dots,\mu_t^{\mathrm{c}})X + (w_1^{\mathrm{c}},\dots,w_t^{\mathrm{c}}), \\ &(w_1^{\mathrm{c}},\dots,w_t^{\mathrm{c}}) \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_t^{\mathrm{c}}) \text{ and independent of } (X,\hat{x}_1), \\ &\hat{x}_s^{\mathrm{c}} = h_s(x_1^{\mathrm{c}},\dots,x_{s-1}^{\mathrm{c}}), \quad \text{for } s \in \{2,\dots,t+1\}, \end{aligned} \tag{87}$$

[4] For the denoisers, we use the notation u_t for consistency with (3.9)-(3.10) in [87].

Then, $\bar{\boldsymbol{\Phi}}^{\mathrm{c}}_{t+1}, \boldsymbol{\Delta}^{\mathrm{c}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ are matrices with entries given by

$$\begin{aligned} &(\bar{\boldsymbol{\Phi}}^{\mathrm{c}}_{t+1})_{i,j} = 0, \quad \text{for } i \le j, \qquad (\bar{\boldsymbol{\Phi}}^{\mathrm{c}}_{t+1})_{i,j} = \mathbb{E}[\partial_j \hat{x}^{\mathrm{c}}_i], \quad \text{for } i > j, \\ &(\boldsymbol{\Delta}^{\mathrm{c}}_{t+1})_{i,j} = \mathbb{E}[\hat{x}^{\mathrm{c}}_i \hat{x}^{\mathrm{c}}_j], \quad 1 \le i, j \le t+1, \end{aligned} \tag{88}$$

where $\partial_j \hat{x}^{\mathrm{c}}_i$ denotes the partial derivative $\partial_{x^{\mathrm{c}}_j} h_i(x^{\mathrm{c}}_1, \dots, x^{\mathrm{c}}_{i-1})$. Finally, we compute μ^{c}_{t+1} and $\boldsymbol{\Sigma}^{\mathrm{c}}_{t+1}$ as

$$\begin{aligned} \mu^{\mathrm{c}}_{t+1} &= \mathbb{E}[X \hat{x}^{\mathrm{c}}_{t+1}], \\ \boldsymbol{\Sigma}^{\mathrm{c}}_{t+1} &= \sum_{j=0}^{2t} \bar{\kappa}_{j+2} \sum_{i=0}^{j} (\bar{\boldsymbol{\Phi}}^{\mathrm{c}}_{t+1})^i \boldsymbol{\Delta}^{\mathrm{c}}_{t+1} ((\bar{\boldsymbol{\Phi}}^{\mathrm{c}}_{t+1})^\top)^{j-i}. \end{aligned} \tag{89}$$

As usual, the $t \times t$ top left sub-matrix of $\boldsymbol{\Sigma}^{\mathrm{c}}_{t+1}$ is given by $\boldsymbol{\Sigma}^{\mathrm{c}}_t$.

D.2 Choice of non-linearities

For the denoiser $h_{t+1}(x_1, \dots, x_t)$ of the correct AMP iteration (82), we use the posterior-mean that takes into account all the past iterates, namely,

$$h_{t+1}(x_1, \dots, x_t) = \mathbb{E}[X \mid (x^{\mathrm{c}}_1, \dots, x^{\mathrm{c}}_t) = (x_1, \dots, x_t)], \tag{90}$$

where $X, x^{\mathrm{c}}_1, \dots, x^{\mathrm{c}}_t$ are the state evolution random variables defined above. If the distribution of $\boldsymbol{X}$ is uniform on the sphere, then $X \sim \mathcal{N}(0,1)$ and the conditional expectation (90) can be simplified as

$$h_{t+1}(x_1, \dots, x_t) = \frac{(\boldsymbol{\mu}^{\mathrm{c}}_t)^\top (\boldsymbol{\Sigma}^{\mathrm{c}}_t)^{-1} \boldsymbol{x}_t}{1 + (\boldsymbol{\mu}^{\mathrm{c}}_t)^\top (\boldsymbol{\Sigma}^{\mathrm{c}}_t)^{-1} \boldsymbol{\mu}^{\mathrm{c}}_t}, \tag{91}$$

where we use the short-hand $\boldsymbol{x}_t = (x_1, \dots, x_t)$. The state evolution parameters $\boldsymbol{\mu}^{\mathrm{c}}_t$ and $\boldsymbol{\Sigma}^{\mathrm{c}}_t$ needed to implement the denoiser h_{t+1} are estimated consistently from the data.

For the denoiser $h_{t+1}(x_t)$ of the Gaussian AMP iteration (8), we use the posterior-mean denoiser that takes into account a single iterate, namely,

$$h_{t+1}(x_t) = \frac{\mu^{\mathrm{G}}_t}{(\mu^{\mathrm{G}}_t)^2 + (\boldsymbol{\Sigma}^{\mathrm{G}}_t)_{t,t}} x_t. \tag{92}$$

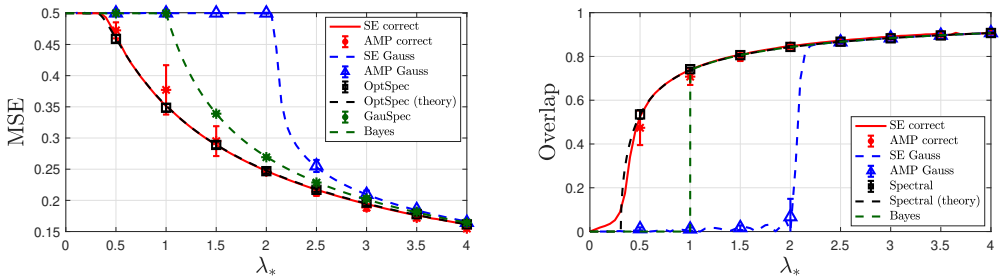
In (92), we use the supra-index G (as opposed to c) to indicate that these quantities correspond to the Gaussian AMP (8) (as opposed to the correct AMP (82)). As usual, the parameters μ^{G}_t and $(\boldsymbol{\Sigma}^{\mathrm{G}}_t)_{t,t}$ are obtained from the data by using the recursion of the correct AMP specialized to the case of Gaussian noise. Let us highlight that this recursion can be implemented also in the mismatched setting, as it depends only on data. However, it does *not* lead to consistent estimates of the state evolution parameters as derived in Theorem 2, because of the mismatch.

D.3 Additional numerical results

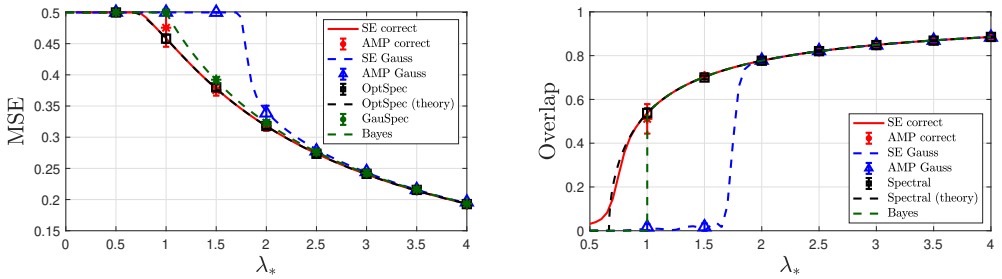
Let the matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be an orthogonal matrix with "Rademacher spectrum" $\rho = \frac{1}{2}(\delta_1 + \delta_{-1})$ (the eigenvalues are i.i.d. uniform ± 1) and $\boldsymbol{W}$ a standard Wigner matrix. Then, the noise is $\boldsymbol{Z}_t := \sqrt{t}\,\boldsymbol{W} + \sqrt{1-t}\,\boldsymbol{A}$ for $t \in [0,1]$. For $t = 1$, this coincides with the pure Wigner case: here, the Bayes estimator is optimal, and the Gaussian AMP is also optimal unless a statistical-to-computational gap is present [28, 12, 54]. In contrast, for $t \in [0,1)$, there is a mismatch and our results give a sharp asymptotic characterization of the Bayes and AMP estimators, cf. Theorem 1 and 2, respectively. By additivity of the R-transform for (asymptotically) free random matrices [70], denoting $R_t(x)$ the R-transform of $\boldsymbol{Z}_t$, we obtain

$$R_t(x) = tx + \frac{\sqrt{4(1-t)x^2 + 1} - 1}{2x}.$$

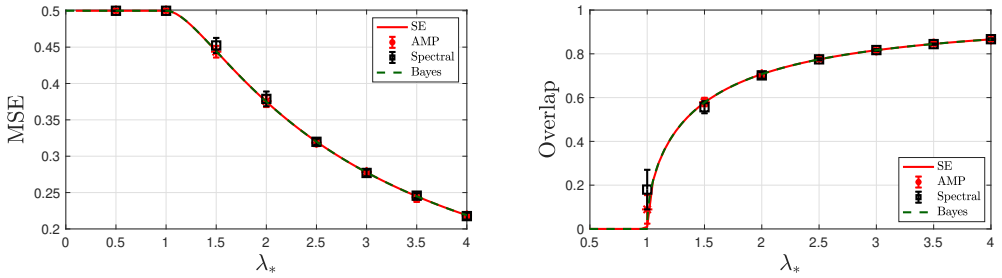
In Fig. 2, we take $t \in \{0.2, 0.5, 1\}$. We note that, for this model, Assumption 1 clearly holds, and we have verified numerically that Assumption 2 holds too. As t goes from 0 to 1, we get closer to a model without mismatch and, therefore, the performance gap between mismatched algorithms (Gaussian AMP, GauSpec, Bayes) and optimal ones (correct AMP, OptSpec) shrinks. As expected, all curves collapse at $t = 1$. The phenomenology described at the end of Section 4 can also be observed in this setting.



(a) $t = 0.2$.



(b) $t = 0.5$.



(c) $t = 1$.

Figure 2: MSE (on the left) and overlap (on the right), as a function of the true SNR λ_*, when the noise spectrum is the free convolution of Rademacher and semicircle spectra.