PNAS

RESEARCH ARTICLE | STATISTICS

OPEN ACCESS

Fundamental limits in structured principal component analysis and how to reach them

Jean Barbier[a,1,2], Francesco Camilli[a,1,2], Marco Mondelli[b], and Manuel Sáenz[c]

Edited by David Donoho, Stanford University, Stanford, CA; received February 5, 2023; accepted June 6, 2023

How do statistical dependencies in measurement noise influence high-dimensional inference? To answer this, we study the paradigmatic spiked matrix model of principal components analysis (PCA), where a rank-one matrix is corrupted by additive noise. We go beyond the usual independence assumption on the noise entries, by drawing the noise from a low-order polynomial orthogonal matrix ensemble. The resulting noise correlations make the setting relevant for applications but analytically challenging. We provide characterization of the Bayes optimal limits of inference in this model. If the spike is rotation invariant, we show that standard spectral PCA is optimal. However, for more general priors, both PCA and the existing approximate message-passing algorithm (AMP) fall short of achieving the information-theoretic limits, which we compute using the replica method from statistical physics. We thus propose an AMP, inspired by the theory of adaptive Thouless–Anderson–Palmer equations, which is empirically observed to saturate the conjectured theoretical limit. This AMP comes with a rigorous state evolution analysis tracking its performance. Although we focus on specific noise distributions, our methodology can be generalized to a wide class of trace matrix ensembles at the cost of more involved expressions. Finally, despite the seemingly strong assumption of rotation-invariant noise, our theory empirically predicts algorithmic performance on real data, pointing at strong universality properties.

high-dimensional inference | structured data | principal components analysis | replica method | approximate message passing

Significance

The assumption of unstructured "noise," i.e., that the complement of what is considered an interesting "signal" in a certain dataset is pure randomness, has pervaded analytical studies in high-dimensional inference. However, this hypothesis is often too simplistic to capture realistic scenarios. We thus need to understand the role of the noise structure, namely the presence of correlations within it. We address this problem in spiked matrix models and provide characterization of the information-theoretic limits to inference. We also propose a message-passing algorithm which is observed through simulations to saturate these limits by optimally capturing the noise statistical dependencies. The resulting picture shows that both signal and noise structure should be exploited by algorithms in order to produce better signal estimation.

The success of inference and learning algorithms depends strongly on the structure of the high-dimensional noisy data they process. Consequently, quantifying how this structure helps algorithms to overcome the curse of dimensionality has become a central topic in statistics and machine learning. Classical examples include sparsity in compressed sensing (1), low-rank structure in matrix recovery (2), or community structure in community detection (3). In all these models, structure is usually assumed only at the signal's level. But the decomposition of the data into "signal" (the component considered of interest) and "noise" (the rest) is often arbitrary and application dependent. For example, in the classification of "dogs/cats," the training images contain a lot of information unrelated to dogs and cats—e.g., on the notions of "inside/outside," "day/night," etc. Yet, this highly structured potential source of information is discarded as random noise (independent, Gaussian, etc.). Most of the research effort has thus focused on understanding how the signal structure alone helps inferring it. In contrast, much less is known about the role of the noise structure and how to exploit it to improve inference.

Given their ubiquitous appearance in the statistics literature, spiked matrix models, which were originally formulated as models for probabilistic principal component analysis (PCA) (4), are now a paradigm in high-dimensional inference. Thanks to their universality features, they, and their generalizations, find numerous applications in other central problems, including community detection (3), group synchronization (5), and submatrix localization or high-dimensional clustering (6). They thus offer the perfect benchmark to quantify the influence of noise structure. In this paper, we focus on the following estimation problem: A statistician needs to extract a rank-one matrix (the spike) $\mathbf{P}^* := \mathbf{X}^*\mathbf{X}^{*\intercal}$, $\mathbf{X}^* \in \mathbb{R}^N$, from the data

$$\mathbf{Y} = \frac{\lambda}{N}\mathbf{P}^* + \mathbf{Z} \in \mathbb{R}^{N\times N}, \qquad [1]$$

with "noise" $\mathbf{Z}$ and signal-to-noise ratio (SNR) $\lambda \geq 0$.

The spectral properties of finite rank perturbations of large random matrices like Eq. **1** were intensively investigated in random matrix theory (see, e.g., refs. 7–9), showing the presence of a threshold phenomenon coined BBP transition (in reference to the authors of ref. 7): When λ is large enough, the top eigenvalue of $\mathbf{Y}$ detaches from the bulk

Author contributions: J.B., F.C., M.M., and M.S. designed research; performed research; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] J.B. and F.C. contributed equally to this work.

[2] To whom correspondence may be addressed: Email jbarbier@ictp.it or fcamilli@ictp.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2302028120/-/DCSupplemental.

Published July 18, 2023.

of eigenvalues. Its corresponding eigenvector has then a nontrivial projection onto the sought ground truth $\mathbf{X}^*$ and can be used as its estimator. The problem has also been approached from the angle of Bayesian inference (10–13). In particular, besides the previous spectral estimator, there exists a whole family of iterative algorithms, known as approximate message passing (AMP), that can be tailored to take further advantage of prior structural information about the signal and noise. AMP algorithms were first proposed for estimation in linear models (14, 15) but have since been applied to a range of statistical estimation problems, including generalized linear models (16, 17) and low-rank matrix estimation (11, 18). An attractive feature of AMP is that its performance in the high-dimensional limit can often be characterized by a succinct recursion called state evolution (19, 20). Using the state evolution analysis, it has been proved that AMP achieves Bayes-optimal performance for some models (11, 16, 18), and a conjecture posits that for a wide range of estimation problems, AMP is optimal among polynomial-time algorithms (21).

The references mentioned above rely on the assumption of independent and identically distributed (i.i.d.) noise, often taken Gaussian $Z_{ij} = Z_{ji} \sim \mathcal{N}(0, 1)$, under which Eq. **1** is the well-known spiked Wigner model (4). This independence, or "absence of structure," in the noise simplifies greatly the analysis. In order to relax this property, we may seek inspiration from the statistical physics literature on disordered systems. An idea that was first brought forth in refs. 22 and 23 for the Sherrington–Kirkpatrick model, and later imported also in high-dimensional inference (24, 25), is that of giving an inhomogeneous variance profile to the noise matrix elements; we mention that this idea in inference is similar to the earlier definition of "spatially coupled systems" (26, 27) in coding theory, see ref. 12 for its use in the present context. The procedure makes the (Z_{ij}) no longer identically distributed, but it leaves them independent. This is an important step toward more structure in the noise. Yet, the independence assumption is a rather strong one. In fact, ref. 25 showed that a broad class of observation models, as long as the independence assumption holds, are information-theoretically equivalent to one with independent Gaussian noise.

One way to go beyond is to consider noises belonging to the wider class of rotationally invariant matrices. Since the appearance of the seminal studies (28–30), there has been a remarkable development in this direction, as evidenced by the rapidly growing number of papers on spin glasses (31–33) and inference (34–37) that take into account structured disorder, including the present one. Indeed, we hereby consider a spiked model in which the noise $\mathbf{Z}$ is drawn from an orthogonal matrix ensemble different from the Gaussian orthogonal ensemble (the only one with independent entries). Intuitively, the presence of dependencies in the noise should be an advantage for an algorithm sharp enough to see patterns within it and use them to retrieve the sought low-rank matrix. Going in that direction, Fan (35) proposed a version of AMP designed for rotationally invariant noises (using earlier ideas of refs. 31 and 32). Furthermore, in a recent work (38), part of the authors analyzed a Bayes estimator and an AMP, both assuming Gaussian noise, whereas the actual noise in the data was drawn from a generic orthogonal matrix ensemble. However, besides intuition and the mentioned studies, to the best of our knowledge, there is little theoretical understanding of the true role played by noise structure in spiked matrix estimation and more generically in inference. In particular, prior to our work, there was no theoretical prediction of optimal performance to benchmark practical inference algorithms.

1. Setting and Main Results

Our analysis focuses on two types of signal's distributions: the factorized prior $dP_X(\mathbf{x}) = \prod_{i\le N} dP_X(x_i)$ and a uniform prior measure over the N-dimensional sphere of radius $\sqrt{N}$. By convention, $\int x^2\, dP_X(x) = 1$, which amounts to rescale λ. The noise matrix $\mathbf{Z}$ is drawn from a trace random matrix ensemble, defined by a certain potential $V : \mathbb{R} \mapsto \mathbb{R}$. V is extended to matrices as follows: if $\mathbf{A} = \mathrm{diag}(a_1, \ldots, a_N)$ then $V(\mathbf{A}) = \mathrm{diag}(V(a_1), \ldots, V(a_N))$. For real symmetric matrices $\mathbf{M} = \mathbf{U}\mathbf{A}\mathbf{U}^\intercal$, with $\mathbf{U}$ orthogonal, $V(\mathbf{M}) = \mathbf{U}V(\mathbf{A})\mathbf{U}^\intercal$. With these notations, we can write the density of the trace ensemble (with normalization constant C_V) as

$$dP_Z(\mathbf{Z}) = C_V \exp\Big(-\frac{N}{2}\mathrm{Tr}V(\mathbf{Z})\Big)\prod_{i\le j} dZ_{ij}. \qquad [2]$$

Instances of such ensembles have a spectral decomposition $\mathbf{Z} = \mathbf{O}\mathbf{D}\mathbf{O}^\intercal$, with $\mathbf{O}$ uniformly distributed over $N \times N$ orthogonal matrices. The distribution of the eigenvalues in the diagonal matrix $\mathbf{D}$, which is independent of $\mathbf{O}$, can be explicitly written, see *SI Appendix*, section 1.2. Only the special case $V(x) = x^2/(2\sigma)$, corresponding to the Gaussian orthogonal ensemble, induces independent (Gaussian distributed) matrix entries. Any other potential generates dependencies among matrix elements and thus structure. For example, if we take $V(x) = x^4/4$, the probability density would be proportional to $\prod \exp(-\frac{N}{8} Z_{ij}Z_{jk}Z_{kl}Z_{li})$, which is clearly not factorizable over matrix entries.

Analyzing the model for a generic potential V is possible through the methodology presented in this paper. Indeed, as discussed in *SI Appendix*, A, this can be done by studying the inference problem whose noise's potential is a polynomial approximation of V. However, if we take a generic polynomial potential V, the higher the order, the more technical and cumbersome our derivations become. Therefore, for the sake of clarity, we focus on a concrete example of nontrivial correction to i.i.d. noise: the quartic matrix potential $V(x) = \mu x^2/2 + \gamma x^4/4$, where μ and γ are two nonnegative real numbers (39). We could have also considered a nonsymmetric potential with a cubic term too, but for simplicity, we restrict ourselves to that case as symmetry slightly simplifies the computations. The noise $\mathbf{Z}$ drawn from the quartic matrix ensemble has a known $N \to \infty$ asymptotic eigenvalue distribution (40)

$$\rho(x)dx = (\mu + 2a^2\gamma + \gamma x^2)\sqrt{4a^2 - x^2}/(2\pi)\, dx, \qquad [3]$$

where $a^2 := (\sqrt{\mu^2 + 12\gamma} - \mu)/(6\gamma)$. In order to have a coherent definition of SNR, we also fix $\int x^2 d\rho(x) = 1$, which implies $\gamma = \gamma(\mu) = (8 - 9\mu + \sqrt{64 - 144\mu + 108\mu^2 - 27\mu^3})/27$. When $\mu = 1, \gamma(1) = 0$ and we recover the pure Wigner case. On the contrary, $(\mu = 0, \gamma(0) = 16/27)$ corresponds to a purely quartic case with unit variance, the "most structured" ensemble in this class. Therefore, μ allows us to interpolate between unstructured and structured noise ensembles.

We emphasize that, although this model may seem rather academic at first sight, we will see that our main assumption, that is, the rotational invariance of the noise, turns out to yield a theory which accurately predicts the empirical performance of algorithms for inference of low-rank matrices hidden in noise coming from real datasets from various application domains. This is probably a consequence of strong universality properties, yet to be understood from a theoretical perspective, along the lines of refs. 41 and 42. We thus argue that our assumptions are in fact rather mild, making our inference algorithms relevant for potential future applications.

We now introduce the Bayesian framework we are going to analyze. Let $\mathbf{P} := \mathbf{x}\mathbf{x}^\intercal$. The posterior measure reads

$$dP_{X|Y}(\mathbf{x} \mid \mathbf{Y}) = \frac{C_V}{P_Y(\mathbf{Y})} dP_X(\mathbf{x}) \exp\Big(-\frac{N}{2} \mathrm{Tr} V\Big(\mathbf{Y} - \frac{\lambda}{N}\mathbf{P}\Big)\Big). \quad [4]$$

The evidence $P_Y(\mathbf{Y})$ is simply the integral of the numerator. We stress that the prior P_X and the likelihood $P_{Y|X}$ match respectively the distribution of the signal and the noise density P_Z, and λ is known. Therefore, we are in the Bayes-optimal setting. Studying the limits of inference in this setting draws a fundamental line between what is information-theoretically possible and what is not in terms of performance of inference.

A main object of interest is the free entropy, which is minus the Shannon entropy of the data: $F_N(\mathbf{Y}) := -H(\mathbf{Y}) = \mathbb{E} \ln P_Y(\mathbf{Y})$. It is related to the mutual information between signal and data through the identity $I(\mathbf{P}^*; \mathbf{Y}) = -F_N(\mathbf{Y}) + \ln C_V - \frac{N}{2}\mathbb{E}\mathrm{Tr} V(\mathbf{Z})$. The relevance of the latter is extensively discussed in *SI Appendix*, section 1.4. Using the form of the observation model in Eq. **1**, it reads

$$-I(\mathbf{P}^*; \mathbf{Y}) = \mathbb{E} \ln \int dP_X(\mathbf{x}) e^{-H_N(\mathbf{x}; \mathbf{Z}, \mathbf{X}^*)} =: \mathbb{E} \ln \mathcal{Z}, \quad [5]$$

where the Hamiltonian linked to the partition function $\mathcal{Z}$ is

$$H_N(\mathbf{x}; \mathbf{Z}, \mathbf{X}^*) := \tfrac{N}{2} \mathrm{Tr}\big[V\big(\mathbf{Z} + \tfrac{\lambda}{N}(\mathbf{P}^* - \mathbf{P})\big) - V(\mathbf{Z})\big]. \quad [6]$$

In this way, the problem is mapped onto a statistical mechanics model with "quenched randomness" $\mathbf{Z}, \mathbf{X}^*$ and "spins" $\mathbf{x}$ with Gibbs–Boltzmann distribution associated with this Hamiltonian (i.e., the posterior). This Hamiltonian is tricky to directly deal with, so a key point will be to "convert" it into a more tractable quadratic form; see Section 1 and *SI Appendix*, section 3.1.

Result 1: Information-Theoretical Limits. Our first result is a variational formula for the mutual information via the celebrated replica method (43) outlined in Section 1: If we let $\boldsymbol{\tau}_* := \mathrm{argmax}\{f_\rho(\boldsymbol{\tau}) : \boldsymbol{\tau} \in \mathbb{R}^{13}, \nabla f_\rho(\boldsymbol{\tau}) = \mathbf{0}\}$, then we have the following low-dimensional expression for the mutual information between hidden spike and the data:

$$\tfrac{1}{N} I(\mathbf{P}^*; \mathbf{Y}) \xrightarrow{N\to\infty} -f_\rho(\boldsymbol{\tau}_*). \quad [7]$$

The argmax is selected and not the argmin as f_ρ is a free entropy (i.e., minus free energy, the free energy being minimized in physics). f_ρ and its derivation are reported in *SI Appendix*, section 3.2. The 13 coupled fixed point equations coming from $\nabla f_\rho = \mathbf{0}$ will reduce to only 2 (*SI Appendix*, Eqs. **79–84**) thanks to special symmetries inherent to the Bayes-optimal nature of our analysis. One of the two remaining order parameters, denoted m^2 and called (squared) "magnetization," quantifies the asymptotic trace inner product between the minimum mean-square error (MMSE) estimator $\int dP_{X|Y}(\mathbf{x} \mid \mathbf{Y})\mathbf{x}\mathbf{x}^\intercal$ and the spike $\mathbf{X}^*\mathbf{X}^{*\intercal}$. It allows us to compute the MMSE as

$$\tfrac{1}{2N^2} \mathbb{E} \|\mathbf{X}^*\mathbf{X}^{*\intercal} - \int dP_{X|Y}(\mathbf{x} \mid \mathbf{Y})\mathbf{x}\mathbf{x}^\intercal\|_{\mathrm{F}}^2 \xrightarrow{N\to\infty} \tfrac{1-m^2}{2}, \quad [8]$$

with m solving the aforementioned system of equations.

Result 2: Optimality of PCA for Rotationally Invariant Priors. The above results hold for a factorized prior $P_X^{\otimes N}$. Nevertheless, if $\mathbf{X}^*$ is uniformly distributed on the sphere, a variational formula analogous to Eq. **7** can still be derived, as shown in *SI Appendix*, section 3.6, and the related MMSE computed. Analytical arguments and numerical experiments show that the latter can be achieved using the naive spectral estimator $C\boldsymbol{v}\boldsymbol{v}^\intercal$ of $\mathbf{P}^*$ obtained from the principal eigenvector $\boldsymbol{v} = \boldsymbol{v}(\mathbf{Y})$ of $\mathbf{Y}$ properly rescaled by a certain factor $C(\lambda, \rho)$; see ref. 9.

Result 3a: Optimal Preprocessing of the Data. Instead of using an AMP with iterates based on $\mathbf{Y}$, we introduce a preprocessing procedure driven by the AdaTAP formalism (32). The end result is an effective quadratic model (i.e., with only pairwise interactions) which is "equivalent" (in a proper sense described below) to the original one, with coupling matrix

$$J(\mathbf{Y}) = \mu\lambda\mathbf{Y} - \gamma\lambda^2\mathbf{Y}^2 + \gamma\lambda\mathbf{Y}^3. \quad [9]$$

This model being quadratic is now solvable using AdaTAP/AMP and possesses the same thermodynamic properties (free entropy, phase transitions, etc.) as well as the same marginal means and variances as the model in Eq. **4** when $N \to \infty$ (and thus equivalent for our purposes). Therefore, to approximate the MMSE estimator, one can simply "preprocess" $\mathbf{Y}$ by applying $J(\mathbf{Y})$ and then efficiently compute the marginals of the resulting quadratic model by AdaTAP/AMP, see next section. AdaTAP allows us to parametrize the free entropy (i.e., log-partition function) of a model with quadratic Hamiltonian, for a given instance of the interaction matrix, in terms of $O(N)$ order parameters, some of which correspond to the sought marginal means $(\langle x_i \rangle)_{i\le N}$ and associated variances. The extremization w.r.t. them yields equations that can be solved iteratively and identified with an AMP algorithm. However, the Hamiltonian in Eq. **6** is not quadratic in $\mathbf{x}$ but can be made so by fixing certain order parameters as outlined in Section 1. The resulting coupling matrix depends on $\mathbf{Y}$ and on the fixed order parameters, whose values are constrained by Bayes optimality (*SI Appendix*, section 1.4). Using these values, for an initial quartic $V(x) = \mu x^2/2 + \gamma x^4/4$, we get the above interaction matrix in Eq. **9** (*SI Appendix*, sections 5.1 and 5.2).

The "cleaning effect" of $J(\mathbf{Y})$ is illustrated in Fig. 1. In general, for a $(K+1)$-order polynomial matrix potential, the preprocessed matrix is a polynomial $J(\mathbf{Y}) = \sum_{k\le K} c_k \mathbf{Y}^k$, with $(c_k)_{k\le K}$ depending on V. For example, for $V(x) = \xi x^6/6$ (with $\xi = 27/80$ to select unit variance), the preprocessing (derived similarly to the quartic case, see *SI Appendix*, section 5.3) is $J_6(x) = \xi\lambda x^5 - \xi\lambda^2 x^4 - \xi\lambda^2 x^2$; it has an effect similar to that in Fig. 1. We point out that the statistics of the noise could be only partially known. This issue can be overcome by learning the (c_k) from the data; see *SI Appendix*, B.

Result 3b: Bayes-Optimal AMP. First, we show in *SI Appendix*, section 4 that existing AMPs (35, 36) do not saturate the MMSE predicted by Eq. **8**. We provide a replica-based theory showing that despite these existing AMPs are aware of the noise structure/statistics, they nevertheless make an implicit mismatched assumption of i.i.d. Gaussian noise: The noise structure is "only" exploited to enforce convergence despite the mismatch, rather than as a source of greater statistical accuracy, in contrast to the proposed AMP we explain now.

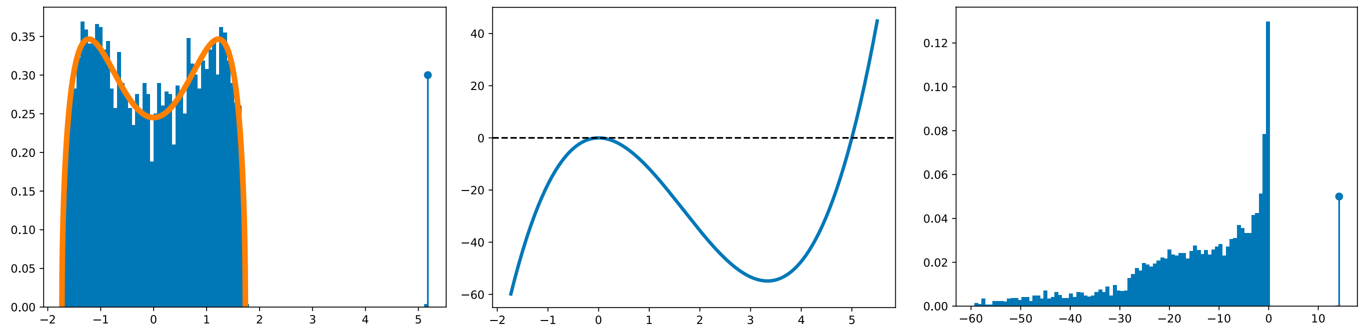

Fig. 1. We have set $\mu = 0, \gamma(0) = 16/27, \lambda = 5, N = 4{,}000$ and generated one instance of the data model Eq. **1**. (*Left*) Histogram of the eigenvalues of **Y**. The leading eigenvalue is emphasized, and the orange curve is the density in Eq. **3**. (*Middle*) Optimal preprocessing function $J(x) = \mu\lambda x - \gamma\lambda^2 x^2 + \gamma\lambda x^3$. (*Right*) Histogram of the eigenvalues of $J(\mathbf{Y})$. The preprocessing J flushes the bulk to the negative axis while pushing only the leading eigenvalue even further from the bulk in the positive direction.

To cure this issue, we employ the preprocessed $J(\mathbf{Y})$ in AMP, which leads to our (conjectural) Bayes-optimal approximate message-passing (BAMP) algorithm with recursion

$$\mathbf{f}^t = J(\mathbf{Y})\mathbf{u}^t - \sum_{i\le t} \mathsf{c}_{t,i}\mathbf{u}^i, \quad \mathbf{u}^{t+1} = g_{t+1}(\mathbf{f}^t), \quad t \ge 1, \qquad \textbf{[10]}$$

with g_{t+1} applied component-wise. For simplicity, we assume to have access to an initialization $\mathbf{u}^1 \in \mathbb{R}^N$ independent of the noise $\mathbf{Z}$ and with a strictly positive correlation with $\mathbf{X}^*$, i.e.,

$$(\mathbf{X}^*, \mathbf{u}^1) \xrightarrow{W_2} (X^*, U_1), \ \ \mathbb{E}[X^* U_1] := \epsilon > 0, \ \ \mathbb{E}[U_1^2] = 1. \qquad \textbf{[11]}$$

This requirement is rather standard in the analysis of AMP algorithms (16, 35, 44). However, as having access to such an initialization is often impractical, recent work (18, 36, 45) has designed AMPs initialized with the top eigenvector $\boldsymbol{v}(\mathbf{Y})$.

By carefully choosing the Onsager coefficients $\{\mathsf{c}_{t,j}\}_{j\in[t]}$, we rigorously obtain BAMP's state evolution characterization.

Theorem 1 (State evolution of BAMP). *Let* $J(\boldsymbol{Y}) = \sum_{i\le K} c_i \boldsymbol{Y}^i$. *Consider the AMP of Eq.* ***10*** *initialized as* Eq. **11**, *with Onsager coefficients* $\{\mathsf{c}_{t,j}\}_{j\in[t]}$ *given in SI Appendix, section 6.2, and where* $(g_{t+1})_{t\ge 1}$ *are* $\mathcal{C}^1$ *and Lipschitz. Then, the following limit holds almost surely for any order 2 pseudo-Lipschitz function*[*] $\psi : \mathbb{R}^{2t+2} \to \mathbb{R}$ *and* $t \ge 1$*:*

$$\frac{1}{N}\sum_{i\le N} \psi(u_i^1, \dots, u_i^{t+1}, f_i^1, \dots, f_i^t, X_i^*) \xrightarrow{N\to\infty} \mathbb{E}\,\psi(U_1, \dots, U_{t+1}, F_1, \dots, F_t, X^*). \qquad \textbf{[12]}$$

Here, for $i \in [t]$, $U_{i+1} = g_{i+1}(F_i)$ *and* $(F_1, \dots, F_t) = \boldsymbol{\mu}_t X^* + (W_1, \dots, W_t)$, *with* $(W_i)_{i\le t}$ *a multivariate Gaussian vector whose covariance as well as* $\boldsymbol{\mu}_t$ *are given in SI Appendix, section 6.2.*

Eq. **12** provides a high-dimensional characterization of our proposed BAMP. A suitable choice of ψ readily gives the MSE of the BAMP iterates. We also note that our result is equivalent to the almost sure convergence in Wasserstein-2 distance of the joint empirical distribution of $(\mathbf{u}^1, \dots, \mathbf{u}^{t+1}, \mathbf{f}^1, \dots, \mathbf{f}^t, \mathbf{X}^*)$ to $(U_1, \dots, U_{t+1}, F_1, \dots, F_t, X^*)$, see corollary 7.21 of ref. 44.

We emphasize that our BAMP algorithm is not the usual AMP of ref. 35, where the data matrix **Y** are just replaced by the preprocessed matrix $J(\mathbf{Y})$. Indeed, tuning the Onsager coefficients $\{\mathsf{c}_{t,i}\}$ entering BAMP requires a type of "multistage" state evolution recursion which is completely different from the one in ref. 35. The acronym we introduce stresses this crucial distinction. While our replica prediction for the MMSE is nonrigorous, the state evolution analysis of BAMP is rigorous. In Section 2, we show that BAMP improves over the AMP in ref. 35 by comparing their fixed points. This improvement is thus a rigorous conclusion, while the conjecture is that BAMP saturates the Bayes-optimal performance.

Finally, the "multistage" state evolution of BAMP suggests a choice of the denoisers in the AMP of ref. 35, which differs from the greedy strategy of ref. 36 (i.e., picking the full posterior mean denoiser at every iteration). The numerical results of Section 2 also show that this denoiser selection—motivated by BAMP—meets the BAMP performance and, hence, the replica prediction of the Bayes-optimal error.

2. Numerical Results and Discussion

BAMP vs the Replica Prediction. The *Left* plot of Fig. 2 considers the quartic ensemble for $\mu = 0$, and the right one refers to the pure power six potential. The signal $\mathbf{X}^*$ has a Rademacher prior $X_i^* \sim \frac{1}{2}(\delta_1 + \delta_{-1})$. The estimators of the spike $\mathbf{X}^*\mathbf{X}^{*\intercal}$ are compared in terms of the MSE achieved at the fixed point, as a function of the SNR λ. All algorithms are run for N = 8,000, they are initialized with $\mathbf{u}^1$ that satisfies Eq. **11**, and the results are averaged over 50 trials; the state evolution recursions and the replica prediction are for $N \to \infty$. In *SI Appendix, section 7.2*, we provide additional numerical results for a sparse Rademacher prior, which display a similar qualitative behavior.

We observe that all algorithms converge rapidly: 10 iterations are sufficient to reach the corresponding fixed points. A few remarks concerning the results displayed in Fig. 2 are now in order. First, in all settings, the fixed point of the BAMP state evolution (red) matches the replica prediction (black). This is a strong numerical evidence supporting our conjecture that the proposed BAMP algorithm is Bayes-optimal. These theoretical curves for $N \to \infty$ are also remarkably close to the MSE achieved by the BAMP algorithm at $N = 8{,}000$.

Second, there is a clear performance gap between our proposed BAMP (red) and the existing AMP algorithms (35, 36) (single-step denoiser in blue, and multistep in ochre). For $V(x) = \xi x^6/6$, the gap is even more evident. As predicted by our theory, the gap is reduced when μ approaches 1 with all curves collapsing for $\mu = 1$; see *SI Appendix, section 7.2*.

[*] A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order 2 if there exists a constant $C > 0$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $\|\psi(\mathbf{x}) - \psi(\mathbf{y})\|_2 \le C(1 + \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)\|\mathbf{x} - \mathbf{y}\|_2$.

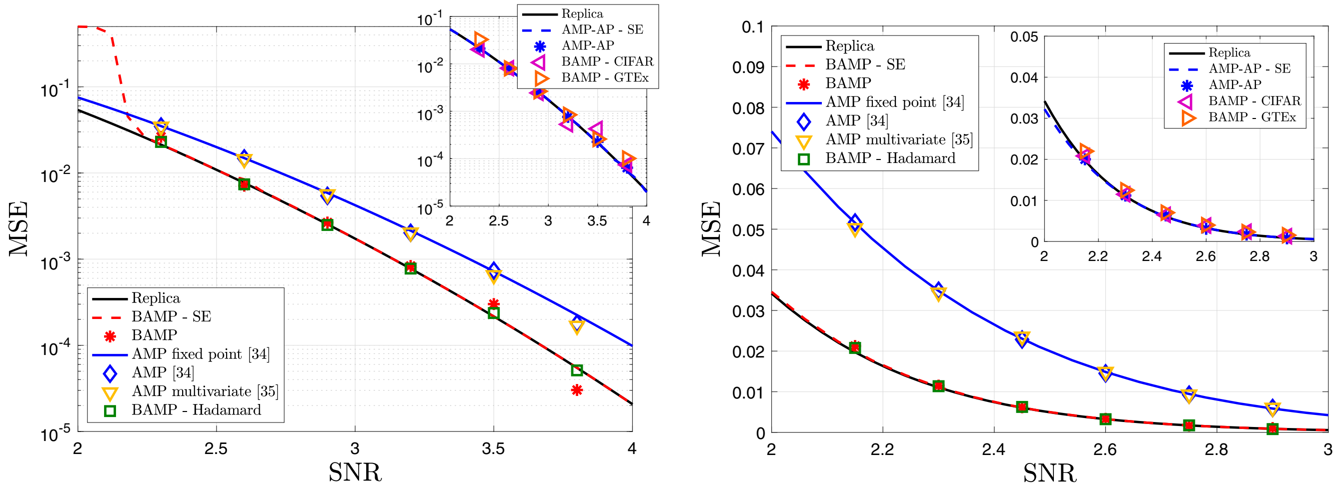

Fig. 2. Quartic potential with $\mu = 0$ (*Left*) and pure power six potential (*Right*). Comparison of the following inference procedures: (*i*) (black) replica prediction of the MMSE, Eq. **8**. (*ii*) (red) Performance of the BAMP algorithm, where g_{t+1} is the single-iterate posterior mean denoiser $g_{t+1}(f) = \mathbb{E}[X^* \mid F_t = f]$. The red line corresponds to the fixed point of the MSE given by the state evolution recursion, and the red stars denote the MSE obtained by running BAMP Eq. **10** with the proper preprocessing. (*iii*) (blue) Performance of the AMP proposed in ref. 35. The blue line corresponds to the fixed point of the MSE obtained with a single-iterate posterior mean as denoiser, and the blue diamonds denote the MSE obtained by running the AMP of ref. 35 with the same denoiser. (*iv*) (ochre squares) MSE obtained by the AMP of ref. 36 (without the preprocessing of **Y**), which employs a full memory posterior mean denoiser: $h_{t+1}(f_1, \ldots, f_t) = \mathbb{E}[X^* \mid (F_1, \ldots, F_t) = (f_1, \ldots, f_t)]$. Finally, (*v*) (green triangles) performance of BAMP when the uniformly distributed matrix **O** (appearing in the spectral decomposition of the noise **Z**) is replaced by the product of the Hadamard–Walsh matrix and a diagonal matrix with i.i.d. Rademacher entries as in ref. 42. In the smaller plots in the *Top-Right* corner, we report the performance of AMP-AP (blue) and of BAMP for our universality experiments involving the CIFAR-10 "plane" class (purple) and the "muscle skeletal" GTEx dataset (orange).

Thirdly, we consider a choice of denoisers in the AMP of ref. 35, which is motivated by our BAMP: If the potential has degree K, every K-th nonlinearity is the full memory posterior mean denoiser, and all the other denoisers are chosen to be the identity. The algorithm is dubbed AMP with alternating posteriors (AMP-AP), and its connection to BAMP is discussed at the end of the 1 section. As evident from the smaller plots in the top right corner, AMP-AP (blue) matches the performance of BAMP and of the replica prediction as well.

Last, BAMP is numerically unstable for low SNR. For the quartic potential and $\lambda = 2.3$, 5 out of 50 trials do not reach the state evolution fixed point (and are thus discarded). Furthermore, BAMP's state evolution detaches from the replica prediction as the SNR gets smaller. Considering an initialization closer to the fixed point mitigates the issue. This instability is likely due to the fact that BAMP's state evolution corresponds to an auxiliary AMP that multiplies the number of iterations (see the 1 section) and which thus amplify errors.

Universality of the Rotational Invariance Assumption. We believe that our results apply beyond the rotational invariance assumption to cases where the eigenbasis of the noise is invariant under more restrictive transformations (such as permutations), or even "quasideterministic." This intuition comes from recent studies (41, 42) showing that, when AMP or its linearized version are used, the class of rotationally invariant matrices leads to the same performance as a much broader class of matrices (with same spectral density). While the existing literature considers a setting different than ours, this still suggests that our predictions should remain true more generally. To confirm this, we plot in Fig. 2 the performance of BAMP when the uniformly distributed matrix **O** (i.e., the noise **Z** eigenbasis) is replaced by i) the product of the Hadamard–Walsh matrix and a diagonal matrix with i.i.d. Rademacher entries, as in ref. 42 (green squares), or ii) the eigenbasis of the covariance matrix for two popular datasets in computer vision and quantitative genetics, i.e., the CIFAR-10 (46) "plane" class and the "muscle skeletal" GTEx dataset (47, 48) (purple and orange markers, respectively, in the *Top-Right* plots). The excellent match clearly supports the universality of our predictions. Additional validations are contained in *SI Appendix*, section 7.2. These results can be understood from the fact that any eigenbasis **O** is typical w.r.t. the Haar measure, so for a fixed instance, as long as **O** is sufficiently independent of the eigenvalues, the universality should hold. This suggests that, in practice, our rotational invariance assumption effectively corresponds to assuming decoupling between eigenbasis and eigenvectors.

3. Methods

Outline of the Replica Computation. The starting point of the replica method is the "replica trick" $\lim_{N\to\infty} \mathbb{E} \ln \mathcal{Z}(\mathbf{Y})/N = \lim_{n\to 0} \lim_{N\to\infty} \ln \mathbb{E}\mathcal{Z}^n(\mathbf{Y})/(Nn)$ that implicitly assumes the commutation of the n, N limits. Another key assumption is to consider $n \in \mathbb{N}$ in the computation and then assume an analytic continuation to n close to 0_+. The expectation is with respect to **Y** or equivalently the independent **O**, **X***; concerning **D**, we only need that its empirical eigenvalue distribution converges weakly to ρ and that it has asymptotically no outliers. When computing $\mathcal{Z}^n$, we get multiple integrals over $(\mathbf{x}_\ell)_{0\le\ell\le n}$, with $\mathbf{x}_0 \equiv \mathbf{X}^*$, and a sum of n Hamiltonians as in Eq. **6** in the exponential. Expanding the exponent, we identify some order parameters: For $1 \le \ell \le n$,

$$v_\ell := \frac{\|\mathbf{x}_\ell\|^2}{N},\ M_{(k)\ell} := \frac{\mathbf{x}_\ell^{\mathsf{T}}\mathbf{Z}^k\mathbf{x}_\ell}{N},\ \kappa_\ell := \frac{\mathbf{x}_\ell^{\mathsf{T}}\mathbf{Z}\mathbf{x}_0}{N},\ m_\ell := \frac{\mathbf{x}_0^{\mathsf{T}}\mathbf{x}_\ell}{N},$$

After fixing these using the Fourier representation of the Dirac delta function, the replicated partition function reads

$$\mathbb{E}\mathcal{Z}^n = \mathbb{E}_{\mathbf{Z},\mathbf{x}_0} \int \prod_{\ell=1}^{n} dP_X(\mathbf{x}_\ell) d\boldsymbol{\tau}_\ell d\hat{\boldsymbol{\tau}}_\ell \, e^{-H_N(\boldsymbol{\tau}_\ell,\hat{\boldsymbol{\tau}}_\ell,\mathbf{x}_\ell;\mathbf{x}_0,\mathbf{Z})},$$

where $H_N(\boldsymbol{\tau},\hat{\boldsymbol{\tau}},\mathbf{x};\mathbf{x}_0,\mathbf{Z}) := Nh(\boldsymbol{\tau},\hat{\boldsymbol{\tau}}) + \mathbf{x}^{\mathsf{T}}\mathbf{J}_1(\boldsymbol{\tau},\hat{\boldsymbol{\tau}},\mathbf{Z})\mathbf{x} + \mathbf{x}^{\mathsf{T}}\mathbf{J}_0(\boldsymbol{\tau},\hat{\boldsymbol{\tau}},\mathbf{Z})\mathbf{x}_0$, and $\boldsymbol{\tau}_\ell := (v_\ell, M_{(1)\ell}, \kappa_\ell, m_\ell)$ with $\hat{\boldsymbol{\tau}}_\ell$ being the Fourier conjugate. The definitions of $(h, \mathbf{J}_1, \mathbf{J}_0)$ can be found in *SI Appendix*, section 3.1. This point is crucial as it allows us to write the n Hamiltonians (one per $\mathbf{x}_\ell$) as at most quadratic functions of $\mathbf{x}_\ell$. Due to the quartic nature of the potential, the original H_N would instead have quartic interactions, or higher-order ones for polynomial V of degree greater than four. Yet, by identifying the proper order parameters, a similar reduction to effective quadratic Hamiltonians would still be possible.

In $\mathbb{E}\mathcal{Z}^n$, the replicas are coupled in the system only through the expectation over the quenched noise, that can be rewritten as an expectation over the Haar distributed noise eigenbasis $\mathbb{E}_{\mathbf{O}}$. The entire computation then boils down to the evaluation of an inhomogeneous log-spherical integral that we introduced and defined as follows: Let the matrices $\mathbf{C}_{\ell\ell'} = \mathrm{diag}((C_{i,\ell\ell'})_{i\le N})$, $\mathbf{C}_i = (C_{i,\ell\ell'})_{\ell,\ell'\le n}$, and vectors $\mathbf{h}_\ell = (h_{i,\ell})_{i\le N}$, $\mathbf{h}_i = (h_{i,\ell})_{\ell\le n}$ all having bounded entries uniformly in N. The sequence $(\mathbf{h}_i \in \mathbb{R}^n, \mathbf{C}_i \in \mathbb{R}^{n\times n})_{i\le N}$ is assumed to have an empirical law tending to that of the random variable $(\mathbf{h} \in \mathbb{R}^n, \mathbf{C} \in \mathbb{R}^{n\times n})$. The inhomogeneous log-spherical integral is defined as

$$\mathcal{I}_N := \frac{1}{N} \ln \mathbb{E}_{\mathbf{O}} \exp\Big(\sum_{\ell,\ell'\le n} (\mathbf{O}\mathbf{x}_\ell)^{\intercal} \mathbf{C}_{\ell\ell'} \mathbf{O}\mathbf{x}_{\ell'} + \sum_{\ell\le n} (\mathbf{O}\mathbf{x}_\ell)^{\intercal} \mathbf{h}_\ell \Big).$$

Its limit depends only on the law of $(\mathbf{C}, \mathbf{h})$ and on the overlaps $q_{\ell\ell'} := \frac{1}{N}\mathbf{x}_\ell^{\intercal}\mathbf{x}_{\ell'}$, $\ell \le \ell'$, that we need to fix with additional Dirac deltas in addition to the previous order parameters. We find that $\lim_{N\to\infty} \mathcal{I}_N$ is expressed by a variational formula; see *SI Appendix*, section 2.1. This integral is a natural generalization of the standard spherical integral (49) and thus may have an interest beyond the present model, in particular, in random matrix theory or spin glasses.

The final ingredient is a replica symmetric ansatz, justified by the strong concentration-of-measure effects taking place in the Bayes-optimal setting (50, 51). It amounts to assume that all order parameters entering the model are independent of the replica index ℓ. Finally, a saddle point yields an extremization over $\mathbb{R}^{13}$ of an effective action. Eqs. **7** and **8** follow directly.

Concerning the reduction from 13 to 2 order parameters (saddle point equations): This is possible thanks to a symmetry arising as a consequence of the Bayes rule which is specific to the Bayes-optimal setting and often called Nishimori identity. It allows to "interchange" the ground-truth signal $\mathbf{X}^*$ with a sample $\mathbf{x}$ from the posterior Eq. **4** inside joint expectations over the posterior and data (see, e.g., ref. 51) and as a consequence to automatically fix the value of most order parameters.

Auxiliary AMP and Onsager Coefficients. The Onsager coefficients $\{\mathsf{c}_{t,i}\}_{i\in[t],t\ge 1}$ are designed so that, conditioned on the signal, the empirical distribution of the iterate $\mathbf{f}^t$ is Gaussian, namely $(\mathbf{f}^1, \dots, \mathbf{f}^t) \xrightarrow{W_2} (F_1, \dots, F_t) := \boldsymbol{\mu}_t X^* + \mathbf{W}_t$, with $\mathbf{W}_t \sim \mathcal{N}(0, \Sigma_t)$ for some mean vector $\boldsymbol{\mu}_t$ and covariance matrix Σ_t. For the AMP in ref. 35, this condition is enforced via the reduction to an *auxiliary* AMP, which also allows to track the iterates of the original algorithm and yields the state evolution parameters, such as $\boldsymbol{\mu}_t$ and Σ_t above. This reduction crucially relies on splitting the matrix $\mathbf{Y}$ that multiplies the iterate $\mathbf{u}^t$, into the rank-one signal plus the noise matrix. In contrast, in Eq. **10**, the iterate is multiplied by the preprocessed matrix $J(\mathbf{Y})$, which cannot be directly split in a similar fashion. Hence, we track all the contributions $(\mathbf{Y}^k\mathbf{u}^t)_{k\le K}$, so that we can split them as $\mathbf{Y}^k\mathbf{u}^t = \mathbf{Y}\mathbf{Y}^{k-1}\mathbf{u}^t = \frac{\lambda}{N}\mathbf{X}^*\langle \mathbf{X}^*, \mathbf{Y}^{k-1}\mathbf{u}^t\rangle + \mathbf{Z}\mathbf{Y}^{k-1}\mathbf{u}^t$.

The key idea is to map the first T iterations of Eq. **10** to the first $K \times T$ iterations of an auxiliary AMP with iterates $(\tilde{\mathbf{z}}^t, \tilde{\mathbf{u}}^t)_{t\in[KT]}$ and denoisers $\{\tilde{h}_{t+1}\}_{t\in[KT]}$,

$$\tilde{\mathbf{z}}^t = \mathbf{Z}\tilde{\mathbf{u}}^t - \sum_{i\le t} \bar{\mathsf{b}}_{t,i}\tilde{\mathbf{u}}^i,\ \tilde{\mathbf{u}}^{t+1} = \tilde{h}_{t+1}(\tilde{\mathbf{z}}^1, \dots, \tilde{\mathbf{z}}^t, \mathbf{u}^1, \mathbf{X}^*), \qquad [13]$$

whose state evolution can instead be deduced from ref. 35. The denoisers $\{\tilde{h}_{t+1}\}_{t\in[KT]}$ of this multistage auxiliary AMP are chosen so that, for $t \in [T]$ and $\ell \in [K]$,

$$\frac{1}{N}\|\tilde{\mathbf{u}}^{K(t-1)+\ell} - \mathbf{Y}^{\ell-1}\mathbf{u}^t\|_2^2 \xrightarrow{N\to\infty} 0. \qquad [14]$$

More specifically, for $t \in [T]$ and $\ell \in \{2, \dots, K\}$, the denoiser $\tilde{h}_{K(t-1)+\ell}$ giving $\tilde{\mathbf{u}}^{K(t-1)+\ell}$ is a linear combination of past iterates $\tilde{\mathbf{u}}^1, \dots, \tilde{\mathbf{u}}^{K(t-1)+\ell-1}$ and of $\tilde{\mathbf{z}}^{K(t-1)+\ell-1}$; furthermore, the coefficients of these linear combinations are chosen to ensure that $\tilde{\mathbf{u}}^{K(t-1)+\ell} \approx \mathbf{Y}^{\ell-1}\mathbf{u}^t$. Hence, by using Eq. **13** with Kt in place of t, one gets $(\mathbf{Y}^\ell\mathbf{u}^t)_{\ell\in[K]}$ from $\tilde{\mathbf{z}}^{Kt}$ and $(\tilde{\mathbf{u}}^{K(t-1)+\ell})_{\ell\in\{2,\dots,K\}}$ (up to an $o_N(1)$). Thus, $J(\mathbf{Y})\mathbf{u}^t$ can be expressed as a linear combination of $(\tilde{\mathbf{u}}^1, \dots, \tilde{\mathbf{u}}^{Kt}, \tilde{\mathbf{z}}^{Kt})$, which in turn is a linear combination of i) the past iterates $\{\mathbf{u}^i\}_{i\in[t]}$, ii) the signal $\mathbf{X}^*$, plus iii) independent Gaussian noise. By inspecting the coefficients of this linear combination, one deduces a) the Onsager coefficients $\{\mathsf{c}_{t,i}\}_{i\in[t],t\ge 1}$ (as the coefficients multiplying the past iterates $\{\mathbf{u}^i\}_{i\in[t]}$), b) the mean μ_t (as the coefficient multiplying the signal $\mathbf{X}^*$), and c) the covariance matrix Σ_t (as the covariance matrix of the remaining noise terms). Finally, by making $\tilde{h}_{Kt+1}$ depend on g_{t+1}, we enforce that $\tilde{\mathbf{u}}^{Kt+1} \approx \mathbf{u}^{t+1}$. The description of the auxiliary AMP is deferred to *SI Appendix*, C.1, and its state evolution follows in *SI Appendix*, C.2.

In summary, the derivation of BAMP's Onsager coefficients involves approximating $\{\mathbf{Y}^k\mathbf{u}^t\}_{k\le K-1}$. This suggests an alternative choice of denoisers leading to the algorithm dubbed AMP-AP: For each batch of K iterations, we pick linear denoisers in the first $K-1$ of them, as this allows to construct $\{\mathbf{Y}^k\mathbf{u}^t\}_{k\le K-1}$; then, at the K-th iteration, we pick the posterior mean using all the past iterates, as this–in principle–allows one to assemble the vectors $\{\mathbf{Y}^k\mathbf{u}^t\}_{k\le K-1}$ to obtain $J(\mathbf{Y})\mathbf{u}^t$ as in BAMP. We note that AMP-AP does not require the coefficients of the polynomial $J(\mathbf{Y})$, but it rather leaves to the posterior mean denoiser to learn them from the data. As such, it provides an efficient alternative to our proposed BAMP.

Data, Materials, and Software Availability. All codes for generating the figures and to test the proposed AMP algorithm have been deposited in GitHub (https://github.com/fcamilli95/Structured-PCA-) (46). The CIFAR-10 dataset is available at (https://www.cs.toronto.edu/~kriz/cifar.html) (48). The GTEx dataset is available at (https://gtexportal.org) (52).

ACKNOWLEDGMENTS. J.B. was funded by the European Union (ERC, CHORAL, project number 101039794). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. M.M. was supported by the 2019 Lopez-Loreta Prize. We would like to thank the reviewers for the insightful comments and, in particular, for suggesting the BAMP-inspired denoisers leading to AMP-AP.

Author affiliations: [a]Quantitative Life Sciences and Mathematics Sections, International Centre for Theoretical Physics, Trieste 34151, Italy; [b]Institute of Science and Technology Austria, Klosterneuburg 3400, Austria; and [c]Centro de Matemática, Universidad de La República, Montevideo 11400, Uruguay

1. D. L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
2. E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).
3. E. Abbe, Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18**, 1–86 (2018).
4. I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327 (2001).
5. A. Perry, A. S. Wein, A. S. Bandeira, A. Moitra, Message-passing algorithms for synchronization problems over compact groups. *Commun. Pure Appl. Math.* **71**, 2275–2322 (2018).
6. T. Lesieur, F. Krzakala, L. Zdeborová, "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel" in *Annual Allerton Conference* (2015).
7. J. Baik, G. B. Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697 (2005).
8. J. Baik, J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models. *J. Multiv. Anal.* **97**, 1382–1408 (2006).
9. F. Benaych-Georges, R. R. Nadakuditi, The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227**, 494–521 (2011).
10. S. B. Korada, N. Macris, Exact solution of the gauge symmetric p-spin glass model on a complete graph. *J. Stat. Phys.* **136**, 205–230 (2009).
11. Y. Deshpande, A Montanari, "Information-theoretically optimal sparse PCA" in *IEEE International Symposium on Information Theory* (2014), pp. 2197–2201.
12. J. Barbier *et al.*, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula" in *Advances in Neural Information Processing Systems* (2016).

13. M. Lelarge, L. Miolane, Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173**, 859–929 (2018).
14. Y. Kabashima, A CDMA multiuser detection algorithm on the basis of belief propagation. *J. Phys. A: Math. General* **36**, 11111 (2003).
15. D. Donoho, A. Maleki, A. Montanari, Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18914–18919 (2009).
16. J. Barbier, F. Krzakala, N. Macris, L. Miolane, L. Zdeborová, Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5451–5460 (2019).
17. S. Rangan, "Generalized approximate message passing for estimation with random linear mixing" in *International Symposium on Information Theory* (2011), pp. 2168–2172.
18. A. Montanari, R. Venkataramanan, Estimation of low-rank matrices via approximate message passing. *Ann. Stat.* **45**, 321–345 (2021).
19. M. Bayati, A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57**, 764–785 (2011).
20. E. Bolthausen, An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Commun. Math. Phys.* **325**, 333–366 (2014).
21. A. Montanari, A. S. Wein, Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. arXiv [Preprint] (2022). http://arxiv.org/abs/2212.06996 (Accessed 3 July 2023).
22. A. Barra, P. Contucci, E. Mingione, D. Tantari, Multi-species mean field spin glasses. Rigorous results. *Ann. Henri Poincaré* **16**, 691–708 (2013).
23. D. Panchenko, The free energy in a multi-species Sherrington–Kirkpatrick model. *Ann. Probab.* **43**, 3494–3513 (2013).
24. D. Alberici, F. Camilli, P. Contucci, E. Mingione, The solution of the deep Boltzmann machine on the Nishimori line. *Commun. Math. Phys.* **387**, 1191–1214 (2021).
25. A. Guionnet, J. Ko, F. Krzakala, L. Zdeborová, Low-rank matrix estimation with inhomogeneous noise. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2208.05918 (Accessed 3 July 2023).
26. A. J. Felstrom, K. S. Zigangirov, Time-varying periodic convolutional codes with low-density parity-check matrix. *IEEE Trans. Inf. Theory* **45**, 2181–2191 (1999).
27. S. Kudekar, T. Richardson, R. Urbanke, Threshold saturation via spatial coupling: Why convolutional LDPC ensembles perform so well over the BEC. *IEEE Trans. Inf. Theory* **57**, 803–834 (2011).
28. E. Marinari, G. Parisi, F. Ritort, Replica field theory for deterministic models: I. Binary sequences with low autocorrelation. *J. Phys. A* **27**, 7615–7645 (1994).
29. E. Marinari, G. Parisi, F. Ritort, Replica field theory for deterministic models: II. A non-random spin glass with glassy behaviour. *J. Phys. A* **27**, 7647–7668 (1994).
30. G. Parisi, M. Potters, Mean-field equations for spin models with orthogonal interaction matrices. *J. Phys. A: Math. General* **28**, 5267 (1999).
31. M. Opper, B. Cakmak, O. Winther, A theory of solving TAP equations for Ising models with general invariant random matrices. *J. Phys. A: Math. Theor.* **49**, 114002 (2016).
32. M. Opper, O. Winther, Adaptive and self-averaging Thouless–Anderson–Palmer mean-field theory for probabilistic modeling. *Phys. Rev. E* **64**, 056131 (2011).
33. A. Maillard *et al.*, High-temperature expansions and message passing algorithms. *J. Stat. Mech.: Theory Exp.* **2019**, 113301 (2019).
34. C. Gerbelot, A. Abbara, F. Krzakala, Asymptotic errors for teacher–student convex generalized linear models (or: how to prove Kabashima's replica formula). *IEEE Trans. Inf. Theory* **69**, 1824–1852 (2023).
35. Z. Fan, Approximate message passing algorithms for rotationally invariant matrices. *Ann. Stat.* **50**, 197–224 (2022).
36. X. Zhong, T. Wang, Z. Fan, Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. arXiv [Preprint] (2021). http://arxiv.org/abs/2110.02318 (Accessed 3 July 2023).
37. R. Venkataramanan, K. Kögler, M. Mondelli, "Estimation in rotationally invariant generalized linear models via approximate message passing" in *International Conference on Machine Learning* (2022), pp. 22120–22144.
38. J. Barbier, T. Hou, M. Mondelli, M. Sáenz, "The price of ignorance: How much does it cost to forget noise structure in low-rank matrix estimation?" in *Advances in Neural Information Processing Systems* (2022).
39. E. Brézin, C. Itzykson, G. Parisi, J. B. Zuber, Planar diagrams. *Comput. Math. Phys.* **59**, 35–51 (1978).
40. M. Potters, J. P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, 2020).
41. R. Dudeja, M. Bakhshizadeh, "Universality of linearized message passing for phase retrieval with structured sensing matrices" in *IEEE Transactions on Information Theory* (2022).
42. R. Dudeja, Y. M. Lu, S. Sen, Universality of approximate message passing with semi-random matrices. arXiv [Preprint] (2022). http://arxiv.org/abs/2204.04281 (Accessed 3 July 2023).
43. M. Mézard, A. Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).
44. O. Y. Feng *et al.*, A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15**, 335–536 (2022).
45. M. Mondelli, R. Venkataramanan, PCA initialization for approximate message passing in rotationally invariant models. *Adv. Neural Inf. Process. Syst.* **34**, 29616–29629 (2021).
46. F. Camilli, M. Mondelli, Structured-PCA-. Codes for Fundamental limits in structured PCA, and how to reach them. https://github.com/fcamilli95/Structured-PCA-. Deposited 13 May 2023.
47. J. Lonsdale *et al.*, The genotype-tissue expression project. *Nat. Genet.* **45**, 580–585 (2013).
48. A. Krizhevsky, V. Nair, G. Hinton, The CIFAR-10 dataset. CIFAR-10. https://www.cs.toronto.edu/~kriz/cifar.html. Accessed 10 May 2023.
49. A. Guionnet, M. Maida, A Fourier view on the R-transform and related asymptotics of spherical integrals. *J. Funct. Anal.* **222**, 435–490 (2005).
50. H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, UK/New York, NY, 2001).
51. J. Barbier, D. Panchenko, Strong replica symmetry in high-dimensional optimal Bayesian inference. *Commun. Math. Phys.* **393**, 1–41 (2022).
52. Broad Institute of MIT and Harvard, GTEx Analysis V8. GTEX Portal. https://gtexportal.org/home/datasets. Accessed 10 May 2023.