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.

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 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 central problems, including community detection (3), group synchronization (5), and universality features, they, and their generalizations, find numerous applications in other analysis (PCA) (4), are now a paradigm in high-dimensional inference. Thanks to their help 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 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) \( \mathbf{P}^* := \mathbf{X}' \mathbf{X}^{\star \top}, \mathbf{X}^* \in \mathbb{R}^N \), from the data

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

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 \( \lambda \) is large enough, the top eigenvalue of \( \mathbf{Y} \) detaches from the bulk
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 $Z_{ij} = Z_{ij} \sim \mathcal{N}(0, 1)$, 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 ensemble, 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 generally 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 $dP_\mathbf{X}(\mathbf{x}) = \prod_{i \leq N} dP_\mathbf{X}(x_i)$ and a uniform prior measure over the $N$-dimensional sphere of radius $\sqrt{N}$. By convention, $\int x^2 \, dP_\mathbf{X}(x) = 1$, which amounts to rescale $\lambda$. The noise matrix $\mathbf{Z}$ is drawn from a trace random matrix ensemble, defined by a certain potential $V : \mathbb{R} \rightarrow \mathbb{R}$. $V$ is extended to matrices as follows: if $\mathbf{A} = \text{diag}(a_1, \ldots, a_N)$ then $V(\mathbf{A}) = \text{diag}(V(a_1), \ldots, V(a_N))$. For real symmetric matrices $\mathbf{M} = \mathbf{U} \Lambda \mathbf{U}^\top$, with $\mathbf{U}$ orthogonal, $V(\mathbf{M}) = UV(\Lambda)U^\top$. With these notations, we can write the density of the trace ensemble (with normalization constant $C_V$) as

$$dP_\mathbf{Z}(\mathbf{Z}) = C_V \exp \left(-\frac{N}{2} \text{Tr} V(\mathbf{Z})\right) \prod_{i \leq N} dZ_{ij}. \quad [2]$$

Instances of such ensembles have a spectral decomposition $\mathbf{Z} = \mathbf{O} \mathbf{D} \mathbf{O}^\top$, 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 $V(x) = x^4/(2\sigma)$, 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 $\prod \exp(-N \frac{1}{2} Z_{ij} Z_{jk} Z_{kl} Z_{il})$, 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 $V(x) = \mu x^2 + 2 + 2 \gamma x^4/4$, where $\mu$ and $\gamma$ 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 $N \rightarrow \infty$ asymptotic eigenvalue distribution (40)

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

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 $\gamma = \gamma(\mu) = (8 - 9\mu + \sqrt{64 - 144\mu + 108\mu^2 - 27\mu^2})/27$. 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, $\mu$ 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 \( P := x x^T \). The posterior measure reads

\[
dP_{X|Y}(x | Y) = \frac{C_Y}{P_Y(Y)} dP_X(x) \exp \left( - \frac{N}{2} \text{Tr}(V - \frac{\lambda}{N} P) \right).
\]

[4]

The evidence \( P_Y(Y) \) 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 \( \lambda \) 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: \( F_N(Y) := -H(Y) = \mathbb{E} \ln P_Y(Y) \). It is related to the mutual information between signal and data through the identity \( I(P^*; Y) = -F_N(Y) + \ln C_Y - \frac{N}{2} \mathbb{E} \text{Tr}V(Z) \). 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

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

where the Hamiltonian linked to the partition function \( Z \) is

\[
H_N(x; Z; X^*) := \frac{N}{2} \text{Tr}[V(Z + \frac{\lambda}{N} (P^* - P)) - V(Z)].
\]

[6]

In this way, the problem is mapped onto a statistical mechanics model with “quenched randomness” \( Z, X^* \) and “spins” \( x \) 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 \( r_* := \text{argmax}_r f_B(r) : r \in \mathbb{R}^{13}, \nabla f_B(r) = 0 \), then we have the following low-dimensional expression for the mutual information between hidden spike and the data:

\[
\frac{1}{N} I(P^*; Y) \xrightarrow{N \to \infty} -f_B(r_*).
\]

[7]

The argmax is selected and not the argmin as \( f_B \) is a free entropy (i.e., minus free energy, the free energy being minimized in physics). \( f_B \) and its derivation are reported in SI Appendix, section 3.2. The 13 coupled fixed point equations coming from \( \nabla f_B = 0 \) 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 \( \int dP_{X|Y}(x | Y) x x^T \) and the spike \( X^* X^T \). It allows us to compute the MMSE as

\[
\frac{1}{2N} \mathbb{E} \|X^* X^T \|_F - \int dP_{X|Y}(x | Y) x x^T \|_F^{2-N \to \infty} \frac{1-m^2}{2},
\]

[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 \( X^* \) 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 naïve spectral estimator \( C_Y V^* \) of \( P^* \) obtained from the principal eigenvector \( v = v(Y) \) of \( Y \) 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 \( Y \), 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

\[
J(Y) = \mu \lambda Y - \gamma \lambda^2 Y^2 + \gamma \lambda Y^3.
\]

[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” \( Y \) by applying \( J(Y) \) 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 \( \langle x_i \rangle_{i \leq N} \) 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 \( x \) but can be made so by fixing certain order parameters as outlined in Section 1. The resulting coupling matrix depends on \( Y \) 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 \( \bar{V}(x) = \mu x^2/2 + \gamma x^4/4 \), we get the above interaction matrix in Eq. 9 (SI Appendix, sections 5.1 and 5.2).

The “cleaning effect” of \( J(Y) \) is illustrated in Fig. 1. In general, for a \((K + 1)\)-order polynomial matrix potential, the preprocessed matrix is a polynomial \( J(Y) = \sum_{k \leq K} c_k Y^k \), with \( (c_k)_{k \leq K} \) 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 \( f_B(x) = \xi 2^3 3^5 - \xi 2^4 3^4 - \xi 2^2 4^2; \) 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.
Theorem 1 (State evolution of BAMP). Let \( J(Y) = \sum_{i \leq K} c_i Y_i \). Consider the AMP of Eq. 10 initialized as Eq. 11, with Onsager coefficients \( \{ c_{ij} \}_{i \in [t]} \) given in SI Appendix, section 6.2, and where \( (g_{i+1})_{i \geq 1} \) are \( C^1 \) and Lipschitz. Then, the following limit holds almost surely for any order 2 pseudo-Lipschitz function* \( \psi : \mathbb{R}^{2t+2} \rightarrow \mathbb{R} \) and \( t \geq 1 \):

\[
\frac{1}{N} \sum_{i \leq N} \psi(u^1_i, \ldots, u^i_{i+1}, f^1_i, \ldots, f^i_{i+1}, X^*_{i}) \xrightarrow{N \to \infty} \mathbb{E} \psi(U_1, \ldots, U_{i+1}, F_1, \ldots, F_t, X^*). \tag{12}
\]

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

Eq. 12 provides a high-dimensional characterization of our proposed BAMP. A suitable choice of \( \psi \) 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 \( (u^1_i, \ldots, u^i_{i+1}, f^1_i, \ldots, f^i_{i+1}, X^*_{i}) \) to \( (U_1, \ldots, U_{i+1}, F_1, \ldots, F_t, X^*_{i}) \), 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 \( J(Y) \). Indeed, tuning the Onsager coefficients \( \{ c_{ij} \} \) 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 \( X^* \) has a Rademacher prior \( X^* \sim \frac{1}{2}(\delta_1 + \delta_{-1}) \). The estimators of the spike \( X^* X^*^\top \) are compared in terms of the MSE achieved at the fixed point, as a function of the SNR \( \lambda \). All algorithms are run for \( N = 8000 \), they are initialized with \( u^1_i \) 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 \( N = 8000 \).

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 ocher). For \( V(x) = \xi x^6/6 \), the gap is even more evident. As predicted by our theory, the gap is reduced when \( \mu \) approaches 1 with all curves collapsing for \( \mu = 1 \); see SI Appendix, section 7.2. [12]
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 $g_{i+1}(f) = \mathbb{E}[X^i | f_i = f]$. 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) (circles) MSE obtained by the AMP of ref. 36 (without the preprocessing of $Y$), which employs a full memory posterior mean denoiser: $h_{i+1}(f_1, \ldots, f_i) = \mathbb{E}[X^i | (f_1, \ldots, f_i) = (f_1, \ldots, f_i)]$. 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).

**Universalis 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” $\lim_{n \to \infty} \mathbb{E} \ln \mathbb{P}(Y|N) = \lim_{n \to 0} \mathbb{E} \ln \mathbb{P}(Y|N)$ that implicitly assumes the commutation of the $n, N$ limits. Another key assumption is to consider $\mathbf{D}$, the $n$-body potential, as Gaussian. This point is typical w.r.t. the Haar measure, so for a fixed instance, as long as $\mathbf{O}$ 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.

*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” $\lim_{n \to \infty} \mathbb{E} \ln \mathbb{P}(Y|N) = \lim_{n \to 0} \mathbb{E} \ln \mathbb{P}(Y|N)$ that implicitly assumes the commutation of the $n, N$ limits. Another key assumption is to consider $\mathbf{D}$, the $n$-body potential, as Gaussian. This point is typical w.r.t. the Haar measure, so for a fixed instance, as long as $\mathbf{O}$ 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.
quadratic functions of $x_t$. Due to the quartic nature of the potential, the original $H_N$ would instead have quadratic 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 $\mathbb{R}^n$, 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 $\xi \in \mathbb{C}$. 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 $C_{\ell}(x) = \text{diag}(C_{\ell}(x)^T)$, $C = (C_{\ell}(x))_{\ell \neq \ell'}$, and vectors $h = (h_{\ell/2})_{\ell \in \mathbb{N}}$, $h = (h_{\ell/2})_{\ell \in \mathbb{N}}$ all having bounded entries uniformly in $N$. The sequence $(b_{\ell}, b_{\ell/2})$, $C \in \mathbb{R}^{n \times n}$ is assumed to have an empirical law tending to that of the random variable $(h \in \mathbb{R}^n, C \in \mathbb{R}^{n \times n})$. The inhomogeneous log-spherical integral is defined as

$$I_N := \frac{1}{N} \ln \mathbb{E} \exp \left( \sum_{\ell \neq \ell'} \langle O_{\ell} X_{\ell'} \rangle \right).$$

Its limit depends only on the law of $(C, h)$ and on the overlaps $q_{\ell \ell'} := \frac{1}{N} x_{\ell}^T x_{\ell'}$, $\ell \leq \ell'$, that we need to fix with additional Dirac deltas in addition to the previous order parameters. We find that $\lim_{N \to \infty} I_N$ 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 $\ell$. Finally, a saddle point yields an extremization over $\mathbb{R}^{n3}$ 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 the “replica formula” in BAMP. We note that AMP-AP does not require the coefficients of the past iterates $\{u_{\ell,i}\}_{i \geq 1}$ (as the coefficient multiplying the signal $x$, and c) the covariance matrix $\Sigma_1$ (as the covariance matrix of the remaining noise terms). Finally, by making $\tilde{h}_{k+1}$ depend on $g_{k+1}$, we enforce that $\tilde{u}_{k+1} \approx u_{k+1}$. 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 $(\tilde{Y}^t u^t)_{k \leq K-1}$. This suggests an alternative choice of denoiser 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 $(\tilde{Y}^t u^t)_{k \leq K-1}$; 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 $(\tilde{Y}^t u^t)_{k \leq K-1}$ to obtain $\tilde{Y}^K u^K$ as in BAMP. We note that AMP-AP does not require the coefficients of the polynomial $J(Y)$, 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.

### Auxiliary AMP and Onsager Coefficients

The Onsager coefficients $\{c_{\ell,i}\}_{i \geq 1}$, $\ell \geq 1$ are designed so that, conditioned on the signal, the empirical distribution of the iterate $f = (f_1, \ldots, f_t)$, $w_t \rightarrow (F_1, \ldots, F_t) := \mu_k x_t + W_t$, with $W_t \sim \mathcal{N}(0, \Sigma_1)$ for some mean vector $\mu_k$ and covariance matrix $\Sigma_1$. 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 $\mu_k$ and $\Sigma_1$ above. This reduction crucially relies on splitting the matrix $Y$ that multiplies the iterate $u^t$, into the rank-one signal plus the noise matrix. In contrast, in Eq. 10, the iterate is multiplied by the preprocessed matrix $J(Y)$, which cannot be directly split in a similar fashion. Here, we track all the contributions $(\tilde{Y}^t u^t)_{k \leq K}$, so that we can split them as $\tilde{Y}^t u^t = \tilde{Y}^t(1-t) + \frac{1}{2} x^T X^T + Y^{k-1} u^t + Z^{k-1} u^t$.

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 $(\tilde{z}^t, \tilde{u}^t)_{t \in \mathbb{K}}$ and denoisers $(\tilde{h}^t, 1)_{t \in \mathbb{K}}$, where $\tilde{z}^t = \tilde{z}^t - \sum_{\ell \leq \ell'} \tilde{b}_{\ell} \tilde{u}_{\ell} + \tilde{u}_{\ell}^{t-1} = \tilde{h}_{\ell+1} (\tilde{z}^t, \ldots, \tilde{u}^t, \tilde{X}^t, \tilde{X}^t), \quad (13)$

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

$$N \tilde{u}_{\ell}^{t} K(t-1) + \tilde{Y}^{t-1} u_{\ell}^{t} N \rightarrow 0. \quad (14)$$

More specifically, for $t \in \mathbb{K}$ and $\ell \in \{2, \ldots, K\}$, the denoiser $\tilde{h}_{\ell} = \tilde{h}_{\ell}(t-1) + \tilde{e}$ giving $\tilde{u}_{\ell} = \tilde{u}_{\ell}(t-1) + \tilde{e}$ is a linear combination of past iterates $\tilde{u}_{\ell}^{t-1}, \ldots, \tilde{u}_{\ell}^{t-K+1}$ and of $\tilde{X}^{t-1} + \tilde{Y}^{t-1} \tilde{u}_{\ell}^{t-1}$, furthermore, the coefficients of these linear combinations are chosen to ensure that $\tilde{u}_{\ell}^{t-1} + \tilde{e} \approx \tilde{u}_{\ell}^{t-1}$, $\tilde{Y}^{t} \tilde{u}_{\ell}^{t} \approx \tilde{u}_{\ell}^{t}$. Hence, by using Eq. 13 with $Kt$ in place of $t$, one gets $(\tilde{Y}^t u^t)_{t \in \mathbb{K}}$ from $\tilde{X}^t$ and $(\tilde{u}^{(t+1)} - \tilde{u}^{(t)} + \tilde{Y}^{(t-1)} \tilde{u}^{(t-1)}, \ldots, \tilde{u}^{(t+2-k)}, \ldots, \tilde{u}^{(t+1-k)})$ up to an $o(1))$. Thus, $J(Y) u^t$ can be expressed as a linear combination of $(\tilde{u}^{(t+1)} u^{(t)}, \tilde{u}^{(t+2)}, \tilde{u}^{(t)})$, which in turn is a linear combination of i) the past iterates $(\tilde{u}^{(t+1)}, \ldots, \tilde{u}^{(t+1-k)})$, ii) the signal $X$, and iii) independent Gaussian noise. By inspecting the coefficients of this linear combination, one deduces a) the Onsager coefficients $\{c_{\ell,i}\}_{i \geq 1}$ (as the coefficients multiplying the past iterates $(\tilde{u}^{(t+1)}, \ldots, \tilde{u}^{(t+1-k)})$, b) the mean $\mu_k$ (as the coefficient multiplying the signal $X$), and c) the covariance matrix $\Sigma_1$ (as the covariance matrix of the remaining noise terms). Finally, by making $\tilde{h}_{k+1}$ depend on $g_{k+1}$, we enforce that $\tilde{u}_{k+1} \approx u_{k+1}$. The description of the auxiliary AMP is deferred to SI Appendix, C.1, and its state evolution follows in SI Appendix, C.2.